ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES

UNAM~SIGLO XXI

5

MATEMÁTICAS

FÍSICA

COMPUTACIÓN

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES

UNAM~SIGLO XXI

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES
UNAM-SIGLO XXI

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES
UNAM~SIGLO XXI

VOLUMEN 5

Matemáticas | Física | Computación

Matemáticas

Javier Bracho
(coordinador)

José Luis Abreu León
Michael Barot
Javier Bracho

Física

María Luisa Marquina
(coordinadora)

Raúl Arturo Espejel Morales
María Luisa Marquina Fábrega
Marco Antonio Martínez Negrete
José Luis Morán López
Miguel C. Núñez Cabrera

Computación

Sergio Rajsbaum
(coordinador)

Ernesto Bribiesca Correa
José Galaviz Casas
Sergio Rajsbaum
Francisco Solsona

siglo veintiuno editores

México, 2010

Enciclopedia de conocimientos fundamentales : UNAM-Siglo XXI / coord. Jaime Labastida y Rosaura Ruiz. – México : UNAM ; Siglo XXI, 2010.
v. ; 27 cm.
Incluye bibliografías
Contenido: v. 1. Español, Literatura – v. 2. Filosofía, Ciencias sociales, Arte – v. 3. Historia, Geografía – v. 4. Química, Biología, Ciencias de la salud – v. 5. Matemáticas, Física, Computación.
ISBN 978-607-02-1760-9 (UNAM obra completa)
ISBN 978-607-03-0225-1 (Siglo XXI obra completa)

1. Enciclopedias y diccionarios. I. Labastida, Jaime. II: Ruiz, Rosaura. III. Universidad Nacional Autónoma de México.

036.1-scdd20 Biblioteca Nacional de México

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES UNAM-SIGLO XXI

Coordinación general | Jaime Labastida
Coordinación académica | Rosaura Ruiz
Coordinación operativa | Alfredo Arnaud
Coordinación editorial | Rosanela Álvarez y José María Castro Mussot
Diseño de la enciclopedia | María Luisa Martínez Passarge
Portadas | Ricardo Martínez

VOLUMEN 5

Coordinación editorial | María Oscos y Paloma Zubieta (Matemáticas), María Luisa Martínez Passarge (Física), Ivonne Murillo y Gabriela Parada (Computación)
Formación | Maia Fernández, Gustavo Jasso, María Luisa Martínez Passarge, Alejandro Ordóñez, María Oscos, Gabriela Parada, Paloma Zubieta
Corrección | Homero Alemán, Silvia Arce, Guillermo Bermúdez, Carmen Jiménez, Jessica Juárez, María Oscos, Alejandro Reza, Kenia Salgado, Felipe Sierra, Ricardo Valdés
Ilustración | Maia Fernández, Víctor Daniel Haro Gómez, Alejandro Ordóñez
Asistencia editorial | Paloma Zubieta (Matemáticas), Raúl Espejel (Física), Luis A. Martínez (Computación)

Portada | Ricardo Martínez
Mujer con palma, 1995

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES UNAM-SIGLO XXI

1ª edición | 2010

La coordinación general agradece la colaboración y el apoyo de las siguientes dependencias de la unam: Escuela Nacional Preparatoria, Colegio de Ciencias y Humanidades; Consejo Académico del Bachillerato; Facultad de Filosofía y Letras, Facultad de Ciencias, Facultad de Ciencias Políticas y Sociales, Facultad de Economía, Facultad de Derecho, Facultad de Medicina, Facultad de Química, Facultad de Contaduría y Administración; Instituto de Ecología, Instituto de Biología, Instituto de Geografía, Instituto de Investigaciones Filosóficas, Instituto de Matemáticas, Instituto de Física, Instituto de Investigaciones en Materiales, Instituto de Investigaciones Históricas; Dirección General de Cómputo y de Tecnologías de Información y Comunicación, Dirección General de Divulgación de la Ciencia, Dirección General de Actividades Cinematográficas, Dirección General de Televisión Universitaria, Dirección de Literatura; Centro Universitario de Estudios Cinematográficos; revista *¿Cómo Ves?*, *Gaceta unam*.

isbn unam de la obra: 978-607-02-1760-9
isbn Siglo XXI de la obra: 978-607-03-0225-1
isbn unam vol. 5: 978-607-02-1782-1
isbn Siglo XXI vol. 5: 978-607-03-0241-1

Impreso y hecho en México.

RICARDO MARTÍNEZ

Ciudad de México, 28 de octubre de 1918 | 11 de enero de 2009

Pareja, 2003 | óleo/tela | 100 × 200 cm

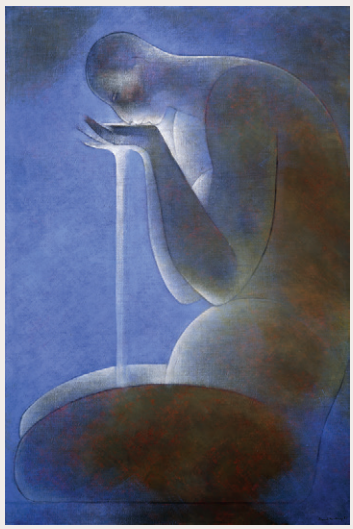
Mujer con palma, 1995 | óleo/tela | 175 × 200 cm

Desde muy joven y a lo largo de su vida Ricardo Martínez nunca dejó su oficio. Lentamente pasó de los paisajes geométricos, bodegones y retratos a la figura humana.

Dotados de un poder monumental que recuerda a la escultura precolombina, sus desnudos —en los que colores, gradaciones y matices logran un todo sinfónico— son ficciones, formas casi abstractas, religiosas, mágicas, no nacidas de la realidad.

A manera de homenaje, los editores de la *Enciclopedia de conocimientos fundamentales* UNAM~*Siglo XXI* se honran en mostrar en sus portadas cinco pinturas de este creador mexicano.

Hombre pensando, 2006 | óleo/tela | 200 × 175 cm



Mujer con agua, 1987 | óleo/tela | 150 × 100 cm

Mujer con niño, 1994 | óleo/tela | 200 × 135 cm

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES

JOSÉ NARRO ROBLES

RECTOR DE LA UNAM

El conocimiento es el camino a la libertad y la justicia. Entre más nociones y valores civiles conforman nuestro bagaje, más amplios serán nuestros horizontes, más diversas nuestras opciones, mayor nuestra posibilidad de elegir y más responsable nuestro comportamiento. En la sociedad contemporánea, el saber se convierte en patrimonio insustituible, en factor de impulso para el desarrollo de un país y en herramienta fundamental para el progreso individual de sus habitantes.

Poseer los fundamentos básicos de cada área y disciplina constituye un valor agregado para el ejercicio profesional y una sólida base para la continuación de estudios superiores. Con esta visión, como parte de su función histórica de transmitir, generar y divulgar las ciencias, las humanidades y las artes, la Universidad Nacional Autónoma de México (UNAM) pone en circulación la *Enciclopedia de conocimientos fundamentales.*

Esta obra adquiere una importancia primordial en tiempos en que los retos que enfrenta la nación en el ámbito educativo son mayúsculos. Más de 33 millones de mexicanos mayores de quince años se encuentran en situación de rezago educativo. Somos un país cuyo nivel promedio de escolaridad apenas rebasa los ocho años de estudio, además de que es considerable el número de jóvenes que desafortunadamente no tiene cabida en el sistema educativo y que tampoco encuentra espacio en el mercado de trabajo.

Una faceta que ejemplifica las insuficiencias del sistema se expresa en el hecho de que sólo dieciocho de cada cien alumnos que ingresan a la educación básica logran concluir los estudios superiores. El resto, 82 por ciento, abandona en algún momento su preparación. El problema es particularmente grave en el tránsito del bachillerato a los estudios profesionales y en los primeros semestres de la licenciatura. En esto radica parte de la trascendencia de esta *Enciclopedia*, elaborada por académicos de bachillerato, licenciatura y posgrado de la UNAM y editada por destacados especialistas de Siglo XXI.

El que dos instituciones de profunda raigambre mexicana, líderes nacionales y regionales en sus ámbitos de acción unan sus esfuerzos y experiencias para hacer posible la *Enciclopedia de conocimientos fundamentales* es la expresión genuina del compromiso que comparten de contribuir a la construcción de un México mejor.

Gracias a esta colaboración, tanto nuestros estudiantes de la Escuela Nacional Preparatoria como del Colegio de Ciencias y Humanidades tendrán a su disposición, en sus respectivos planteles, ejemplares de esta obra, esencial para su formación media superior.

Es además un propósito de la UNAM y de Siglo XXI Editores el que este material esté al alcance del público más amplio y diverso, como una referencia invaluable y fuente básica de los saberes que como mínimo requiere todo individuo en materia de ciencias, de humanidades, de ciencias sociales, de lenguas y de matemáticas.

Representa para nuestra casa de estudios una enorme satisfacción refrendar, mediante la *Enciclopedia de conocimientos fundamentales*, su vocación de servicio a la sociedad a la que se debe, además de contribuir con este legado a la construcción del país democrático, justo y equitativo que todos deseamos y por el que tantas generaciones han luchado.

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES

JAVIER DE LA FUENTE

SECRETARIO DE DESARROLLO INSTITUCIONAL, UNAM

El fortalecimiento de la educación media superior y la divulgación del saber hacia el público en general figuran entre las múltiples prioridades de la Universidad Nacional Autónoma de México (UNAM), pues constituyen un compromiso para enfrentar tanto los rezagos en materia educativa como las exigencias en el ámbito profesional de la sociedad contemporánea. Nuestra máxima casa de estudios contribuye de manera constante a enfrentar los desafíos de nuestro tiempo con estrategias y soluciones concretas. Éste es el sentido y el espíritu de la *Enciclopedia de conocimientos fundamentales*.

Profesores e investigadores de los tres niveles educativos de la UNAM, especialistas en filosofía, ciencias sociales, artes, literatura, lengua española, historia, geografía, química, biología, ciencias de la salud, matemáticas, física y computación, se dieron a la tarea de establecer, de manera conjunta, cuáles serían los saberes indispensables de cada área con los que debe contar todo ciudadano mexicano de nuestro tiempo para enfrentar su realidad cotidiana. A ellos se sumaron destacados asesores de Siglo XXI Editores, muchos de ellos académicos reconocidos de la UNAM, que revisaron, adaptaron y perfeccionaron los contenidos de este proyecto

El resultado de este magnífico esfuerzo académico colegiado y conjunto es la obra que hoy ve la luz. Al abordar un total de trece disciplinas, el material que tiene usted en sus manos resulta esencial tanto para el desarrollo académico como para el ejercicio profesional de estudiantes que inician su formación superior, maestros de educación media superior, y todo ciudadano adulto. En su totalidad, constituye un material invaluable para fomentar el conocimiento interdisciplinario, poner a su alcance y enriquecer su cultura general.

El primer tomo, orientado a las Lenguas, se aproxima a la literatura a través de la lectura, las figuras y los géneros literarios como el mito, el relato, la poesía, el teatro y el ensayo. Plantea además temas específicos respecto al español, en torno a la lengua y la comunicación, los textos narrativos, expositivos, argumentativos, orales y monográficos, así como las nuevas formas de leer y escribir en el siglo actual.

El segundo tomo de esta *Enciclopedia* está dedicado a las Humanidades. Aborda, en el ámbito de la filosofía, temas de razonamiento lógico, conocimiento y verdad, lenguaje, ciencia y tecnología, existencia y libertad, política y sociedad, artes y belleza. En el terreno de las ciencias sociales propone una introducción a la sociología, la antropología, la política, el derecho, la economía y la administración. En cuanto al arte, plantea cuestiones torales sobre el

sentido social de esta actividad, la estética, la creación, la interpretación y la apreciación, complementadas con entrevistas a destacados creadores mexicanos.

El tercer volumen se enfoca a la historia de México, su multiculturalidad, la conquista, la primera y la segunda integraciones planetarias de nuestro país y su organización en el siglo XX. En cuanto a la geografía, aborda la dimensión territorial de los recursos naturales, la organización del territorio, la población en el espacio geográfico, los riesgos naturales y entrópicos, los procesos políticos y el territorio mexicano.

El cuarto tomo está dedicado a las Ciencias. En el dominio de la química, ofrece nociones sobre la historia de esta disciplina, las mezclas y sustancias, los átomos, las moléculas y los iones, el lenguaje de esta ciencia, los enlaces, las reacciones químicas y su energía, la estequimetría, los ácidos y bases, las reacciones de óxido-reducción, la química y el entorno. En materia de biología, aborda su concepto como ciencia, explica sus particularidades en los ámbitos celular, molecular y bioquímico, y define aspectos de la genética, de la evolución, de la ecología y de la relación de esta ciencia con la sociedad. En lo que toca a las ciencias de la salud, plantea una introducción a los conceptos de la salud y la enfermedad, expone las funciones vitales básicas, el inicio de la vida, y las etapas de crecimiento y desarrollo desde la infancia hasta la vejez.

El volumen cinco ofrece conocimientos fundamentales en matemáticas, sus orígenes y su función en la actividad humana, y su expresión en la naturaleza. En materia de física, aborda la mecánica, la electricidad y el magnetismo, la óptica, la física de fluidos y la termodinámica, en una lógica de lo grande a lo pequeño. Finalmente, ofrece nociones básicas de computación referentes a la algorítmica, la programación, la información, la abstracción, las computadoras, las redes, el multimedia y las aplicaciones de esta especialidad.

A los contenidos de cada uno de estos cinco tomos, Siglo XXI Editores ha añadido una antología de textos esenciales y paradigmáticos de autores clásicos en su respectiva especialidad cuya contribución universal constituye hoy una referencia obligada para el desempeño cotidiano, sea cual sea nuestra actividad. Así, el lector tendrá acceso a fragmentos de la obra de Platón, Aristóteles, Galileo, Newton, Descartes, Humboldt, Darwin, Einstein, Octavio Paz, entre muchos otros.

Cada tomo de la *Enciclopedia* cuenta adicionalmente con un DVD, en el que se ofrece material didáctico complementario sustentado en fuentes especializadas de la UNAM, con el fin de ampliar el aprendizaje de sus usuarios. Esta obra combina el uso de herramientas tradicionales con las posibilidades que ofrecen las nuevas tecnologías, para contribuir, con ello, a que alumnos, maestros y ciudadanos en general cuenten con elementos que les permitan insertarse a la nueva sociedad del conocimiento.

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES

ROSAURA RUIZ

COORDINADORA ACADÉMICA DEL PROYECTO ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES, UNAM-SIGLO XXI

Los múltiples programas que la Universidad Nacional Autónoma de México (UNAM) ha concebido y puesto en marcha permiten hacer frente —con un espíritu innovador y con la calidad académica que la distinguen— a los grandes rezagos de la educación media superior y superior del país, y promover el mejoramiento de la calidad educativa en todos sus ámbitos.

La formación integral de los alumnos y de todo individuo, por medio de la adquisición de conocimientos y del desarrollo de habilidades, resulta hoy más que nunca indispensable, tanto para satisfacer los requerimientos de la educación superior como para formar ciudadanos socialmente responsables. Para ello se requiere dotarles de saberes pertinentes para el ejercicio profesional o la continuación de su formación académica y, en ambos casos, para ensanchar su espectro de alternativas de respuesta y solución a los desafíos que plantea la vida cotidiana.

Como antecedente de esta *Enciclopedia*, la UNAM desplegó un ambicioso programa de acciones compartidas encaminado a ofrecer una novedosa propuesta para actualizar los contenidos temáticos de las disciplinas que comparten los dos subsistemas del bachillerato de la UNAM. A partir de la reflexión, la integración de diversos esfuerzos y la reelaboración de los procesos de enseñanza —en función de los cambios y exigencias de nuestra sociedad dentro de un contexto global—, fueron seleccionados los saberes básicos de trece disciplinas, entendidos como herramientas para el desarrollo personal y profesional de todo ciudadano.

El objetivo final de este proyecto ha sido el de contribuir a la formación de ciudadanos críticos, con un pensamiento lógico, capaces de enfrentar problemáticas y de plantear soluciones. Todo ello en el entendido de que una educación sustentada en la acumulación informativa resulta obsoleta en los albores del siglo XXI, y de que es preciso impulsar la apropiación de una cultura general y el desarrollo de habilidades estratégicas para capacitar a hombres y mujeres de modo que aprendan en forma propositiva y significativa a lo largo de la vida.

Los temas considerados en cada una de las disciplinas que conforman esta *Enciclopedia* han sido establecidos y acotados en razón de su relevancia y pertinencia, tanto dentro del contexto académico como en congruencia con las exigencias del entorno contemporáneo; se sustentan así en el avance y desarrollo reciente de cada disciplina y en su potencial como instrumento de transformación.

Se reafirma en este sentido la trascendencia del trabajo colegiado, crítico y plural de los docentes e investigadores que han hecho posible este proyecto, con el que la UNAM contribuye a elevar la calidad e innovar en los procesos de enseñanza-aprendizaje, además de refrendar su compromiso con los jóvenes de nuestro país. La obra resultante de este ejercicio académico —esta *Enciclopedia de conocimientos fundamentales*—, pone énfasis en un proceso formativo sustentado en la profundización, la reflexión, la cabal comprensión y asimilación del conocimiento, en contraste con una perspectiva basada en la acumulación progresiva de información.

Lo que hoy tenemos a la vista es, pues, el resultado del esfuerzo colectivo en el que convergen el compromiso universitario, la experiencia académica, la visión transformadora y la voluntad creativa de quienes participaron en esta *Enciclopedia*.

El que el fruto de este proyecto esté disponible en las bibliotecas y los hogares mexicanos es ya un motivo de orgullo y satisfacción. El que su contenido se convierta en coadyuvante del mejoramiento individual y social de quien se beneficie de él, es la aspiración máxima de todos los que lo hemos hecho posible.

ENCICLOPEDIA DE CONOCIMIENTOS FUNDAMENTALES

JAIME LABASTIDA

DIRECTOR GENERAL SIGLO XXI EDITORES

La *Enciclopedia* que el lector tiene en sus manos es diferente a las que se podría llamar habituales. Lo es en diversos aspectos fundamentales, sin duda alguna. En primer término, tiene carácter temático. Esto significa que tiene un orden distinto al que poseen otras enciclopedias. La nuestra no responde a un orden alfabético. En segundo término, su temática guarda estrecha relación con las disciplinas académicas de la educación media superior: su orden, por consecuencia, lo determina la estructura lógica a la que responden estas disciplinas, que van de lo general y lo básico a lo particular y específico.

Nuestra voz española *enciclopedia* viene de una expresión helena, lo sabe todo mundo, ἐν κύκλος παιδεία, la *educación en círculo*; con otras palabras, *educación total*, *completa*. ¿Una educación total? ¿Un saber o un conjunto de saberes de carácter *universal*? ¿Quién, el día de hoy, pedagogo, científico o filósofo, aspira a tanto? El cúmulo de los conocimientos es ya de tal naturaleza que nadie puede creer que existan nada ni nadie que estén en condiciones de dar (o de poseer) la totalidad de los conocimientos que proporcionan las humanidades, la ciencia y la tecnología en sus avances constantes en las más diversas disciplinas.

Si resulta imposible abarcar la totalidad del conocimiento humano en una publicación de esta naturaleza, ¿qué pretende, pues, una enciclopedia como ésta, propia, en lo fundamental, para los estudiantes de educación media superior o para un público amplio? Ya se ha dicho que se trata de una enciclopedia temática, cuyo orden responde al que tienen las disciplinas científicas y humanísticas del sistema escolar del bachillerato. Ese orden no es arbitrario ni se deriva de una mera convención, como la que posee el alfabeto; no va, pues, desde la A hasta la Z, sino desde nuestra lengua, el español, hasta una técnica actual, el sistema de cómputo. Hegel hacía notar el carácter convencional y arbitrario de las enciclopedias y por esa causa exigió de su *Enciclopedia de las ciencias filosóficas* una estructura que respondiera al sistema, o sea, que fuera lógica, racional.

La Edad Moderna ha producido al menos dos enciclopedias paradigmáticas. Las dos intentaron la síntesis más completa del conocimiento de la época. Es posible que lograran su objetivo: iluminaron las conciencias para siempre. Sin embargo, como dijo Heráclito, *nuevas aguas corren tras las aguas*: el conocimiento no puede estancarse. La *Encyclopaedia Britannica*, pues de ella se habla, se editó por primera vez en 1757. Era una modesta publicación en tres volúmenes, pero poseía el carácter que la haría famosa: conjugaba el texto escrito con

la descripción gráfica de aquello a lo que el texto hacía referencia. El día de hoy, la *Encyclopaedia Britannica* la forman al menos 29 gruesos volúmenes.

La otra enciclopedia paradigmática se debe al talento y la valentía de aquel inmenso filósofo que se llamó Denis Diderot. Es la publicación más importante del siglo XVIII, el siglo llamado de Las Luces. La conocemos todavía con el nombre de la *Gran Enciclopedia*.

La imprenta democratizó la razón e hizo posible la expansión de la cultura y el conocimiento. La sabiduría, que hasta ese momento había sido propiedad privada de unos cuantos y se transmitía de modo oral o, de modo igualmente trabajoso, a través de la copia manuscrita de gruesos volúmenes en los monasterios europeos, de súbito pudo entrar en las casas de todos los hombres. El círculo del conocimiento posible adquirió una dimensión hasta ese momento desconocida y luego, desde el siglo XIX, el hecho de que lo mismo el padre que la madre estuvieran obligados, por la nueva situación económica, a emplearse en actividades productivas, hizo nacer la escuela moderna. Mientras que los hijos de los aristócratas recibían enseñanzas por parte de preceptores privados en sus casas, los hijos del pueblo acudían a las escuelas públicas. Ambos podían estudiar en los nuevos instrumentos: los libros que las imprentas reproducían por miles de ejemplares.

Paideia es voz asociada al niño (*pais*, *paidós*). Es la educación de los niños, desde luego. Produjo, en nuestra lengua, la palabra *pedagogía* que, en sentido amplio, quiere decir *educación* y, ya lo dije, en el caso de la voz *enciclopedia* pretende una educación total y, por lo tanto, imposible.

Pero si a una enciclopedia temática moderna le es imposible abarcar la totalidad de los conocimientos humanos, ¿qué pretende ésta, que la UNAM y Siglo XXI ofrecen a los lectores? La nuestra pone el acento en el *método*: sus autores son conscientes de que tan importante es el *resultado* como el *proceso* que condujo hacia él. Aquí, el acento no está puesto en la *memoria* sino en la *formulación* de problemas, porque *método*, ya se sabe, es una palabra formada a partir de la voz griega *odós*, *camino*. Tan decisivos son el camino como el lugar de la llegada. Saber preguntar, saber indagar, saber establecer dudas, saber organizar los conocimientos, saber que no se sabe, crear, inventar, interrogar al mundo contemporáneo, duro y exigente como pocos, con una pasión que brota —si hemos de creerle a José Gorostiza— de aquella *soledad en llamas* que es la inteligencia, es uno de los propósitos de nuestra *Enciclopedia*.

De allí que los textos de las diversas disciplinas vayan acompañados de antologías o reunión de textos —muchos de ellos clásicos— que no pretenden sino complementar, enriquecer e invitar a los lectores a profundizar en temas, autores, creaciones, teorías, corrientes del pensamiento: la sabiduría actual es una herencia, una acumulación de los siglos anteriores. Antes que respuestas, tenemos dudas y preguntas.

ÍNDICE

MATEMÁTICAS

Los autores 1

Agradecimientos 2

Introducción 3

¿Cómo se escriben las matemáticas? 5

TEMA 1 EL PORQUÉ DE LAS MATEMÁTICAS

1.1 Introducción 9

1.2 La actividad humana 12

1.2.1 Música y matemáticas 13

1.2.2 La Pirámide de los Nichos 15

1.2.3 ¿Cómo se puede medir la profundidad de un pozo? 19

1.3 Las matemáticas de la naturaleza 22

1.3.1 La esfera 23

1.3.2 La balanza 24

1.3.3 La gravedad 26

1.4 Las propias matemáticas 33

1.4.1 Pitágoras, Fermat y Wiles 34

1.4.2 Los números irracionales y el espacio euclidiano 36

1.5 Conclusión 40

TEMA 2 MATEMÁTICAS DE LA ACTIVIDAD HUMANA

2.1 Introducción 42

2.2 Números para contar 43

2.3 Números para medir 47

2.4 Números para expresar lo continuo 52

2.5 ¿Cómo calcular de manera eficiente? 55

2.6 Los números de la computación 57

2.7 Medir lo inalcanzable 60
2.8 La medición de la Tierra 64
2.9 La pirámide truncada 67
2.10 El número π y la cuadratura del círculo 72
2.11 Cilindros, conos y esferas 77
2.12 La cuadratura de la parábola y el método de demostración por inducción 83
2.13 Las cónicas y su uso 87
2.13.1 Secciones de un cono 87
2.13.2 Las cónicas como lugares geométricos 88
2.13.3 La excentricidad 90
2.13.4 La forma de una antena 91
2.13.5 Ecuaciones de segundo grado 93
2.13.6 Cónicas en la física 94
2.14 Probabilidad y estadística, calculando el azar 95
2.14.1 Necesidad de una teoría de la probabilidad 96
2.14.2 El problema con el que se inicia el cálculo de probabilidades 98
2.14.3 El modelo matemático general de la probabilidad 99
2.14.4 Probabilidad condicional, eventos independientes y variables aleatorias 101
2.14.5 El lanzamiento de canicas a una pared y la distribución normal 103
2.14.6 La ley de los grandes números 104
2.14.7 La paradoja del cumpleaños 106
2.14.8 El teorema del límite central 106
2.14.9 La estadística 108

TEMA 3 LAS MATEMÁTICAS EN LA NATURALEZA

3.1 Introducción 114
3.2 La simetría en la naturaleza 115
3.2.1 Algunos objetos simétricos 115
3.2.2 Composición e inversos de simetrías 117
3.2.3 El concepto detrás de la simetría 118
3.2.4 Las simetrías del *Nautilus* 119
3.2.5 Simetría de fórmulas 120
3.2.6 Simetría conceptual 122
3.2.7 Simetría en la física 124
3.3 Espacio, tiempo y movimiento 125
3.3.1 El espacio 126
3.3.2 El continuo espacio-tiempo: los números reales 127
3.3.3 El movimiento 129
3.3.4 La velocidad y el concepto de derivada 130
3.3.5 La integral y el teorema fundamental del cálculo 133
3.3.6 Newton y las leyes de Kepler, una nueva concepción del Universo 136
3.4 Las órbitas celestes 137
3.5. Las ecuaciones que modelan el mundo 143
3.5.1 Primeros ejemplos 143
3.5.2 Un modelo para la radiactividad 144
3.5.3 Modelos para el crecimiento poblacional 147
3.5.4 Un modelo para la propagación de un virus 150

3.6 El campo y los vectores 152
3.6.1 La electrostática y el concepto de campo vectorial 153
3.6.2 Álgebra vectorial 154
3.6.3 Cálculo vectorial 155
3.6.4 El campo electromagnético 158
3.6.5 La física cuántica 160
3.6.6 Albert Einstein y la teoría de la relatividad 163
3.6.7 Conclusión 165

TEMA 4 LAS MATEMÁTICAS DE LAS MATEMÁTICAS

4.1 Introducción 167
4.2 Razón áurea (y Fibonacci) 168
4.2.1 El pentagrama místico 170
4.2.2 La sucesión de Fibonacci 173
4.3 De Königsberg a Google 177
4.3.1 Los puentes de Königsberg 177
4.3.2 Paseos eulerianos 179
4.3.3 La Red y la red 181
4.4 La conquista del infinito 181
4.4.1 Cómo medir lo infinito 181
4.4.2 Diferentes infinitos 183
4.4.3 Números ordinales y números cardinales 185
4.4.4 La base formal de las matemáticas 186
4.5 El desarrollo del álgebra 188
4.5.1 La aritmética 188
4.5.2 El largo nacimiento de la notación algebraica moderna 189
4.5.3 Ecuaciones lineales, cuadráticas, cúbicas y de cuarto grado 191
4.5.4 Nuevos horizontes 192
4.5.5 El álgebra moderna 193
4.5.6 Álgebra lineal 195
4.5.7 Algebraización 196
4.5.8 El gran proyecto de clasificación de los grupos simples 198
4.6 ¿Qué es la geometría hoy? 199
4.6.1 El quinto postulado 200
4.6.2 El toro plano y el plano elíptico 201
4.6.3 El plano proyectivo 203
4.6.4 Los espacios multidimensionales 205
4.6.5 Topología 206
4.7 ¿Cómo fundamentar las matemáticas? 208
4.8 ¿Qué se puede medir? 214
4.9 ¿Qué se puede resolver? 215
4.9.1 Limitación a ecuaciones algebraicas 215
4.9.2 El cálculo con "el número" $\mathbf{i} = \sqrt{-1}$ 217
4.9.3 Representación geométrica de los números complejos 219
4.9.4 La ecuación cuadrática con coeficientes complejos 221
4.9.5 Las ecuaciones de tercer y cuarto grados 222
4.9.6 El teorema fundamental del álgebra 225
4.9.7 Polinomios, raíces y simetrías 226
4.9.8 La teoría de Galois 230

4.10 ¿Qué se puede construir? 234
4.10.1 Delimitación de la pregunta 234
4.10.2 Los problemas clásicos 235
4.10.3 El plano complejo como modo algebraico 236
4.10.4 Descripción alterna del campo de los números construibles 238
4.10.5 Sobre la imposibilidad de resolver los problemas clásicos 241
4.11 ¿Qué se puede demostrar? 244
4.11.1 El sistema axiomático 244
4.11.2 La teoría de conjuntos como base para las matemáticas 246
4.11.3 El programa de Hilbert 247
4.11.4 El teorema de Gödel 249
4.12 Y… ¿si todo quedara descubierto? 253

Bibliografía 256

APÉNDICE MATEMÁTICAS`
Michel Serres 261
Los orígenes de la geometría
Isaac Newton 281
Principios matemáticos de la filosofía natural

FÍSICA
Los autores 289
Agradecimientos 291
Introducción 293

TEMA 1 DESDE LA GRAN EXPLOSIÓN
1.1 La gran teoría 295
1.2 Un brillo variable 296
1.3 Grandes distancias 298
1.4 Grandes velocidades 299
1.5 ¡Es variable! 299
1.6 Recesión galáctica 301
1.7 Se ve lo mismo 303
1.8 El huevo cósmico 304
1.9 La radiación fósil 306
1.10 Una tercera evidencia 309

TEMA 2 MECÁNICA
Introducción 313
2.1 La idea de movimiento 314

2.1.1 El movimiento de los astros 315
2.1.2 Percepción sensorial del movimiento 316
2.1.3 Marcos de referencia del movimiento 319
2.1.4 Sistema Internacional de Unidades 322
2.1.5 Rapidez y velocidad 323
2.2 Movimiento rectilíneo uniforme 324
2.3 Movimiento oscilatorio. Movimiento ondulatorio transversal y longitudinal 324
2.4 Efecto Doppler 328
2.4.1 Efecto Doppler en el agua 329
2.5 De Aristóteles a Galileo: una aportación importante para la ciencia 330
2.5.1 Modelo aristotélico 331
2.5.2 El modelo de Galileo 333
2.6 La aceleración 335
2.7 La medición de la fuerza 336
2.8 La gran aportación de Isaac Newton: la idea de inercia 337
2.9 La relación de la masa, la aceleración y la fuerza. Segunda ley de Newton 338
2.10 La acción y la reacción. Tercera ley de Newton 340
2.11 La ley de la gravitación universal 341
2.11.1 Leyes de Kepler 343
2.11.2 El campo gravitatorio 344
2.12 La cantidad de movimiento lineal 344
2.13 El concepto de trabajo mecánico 344
2.14 La energía: una idea fructífera y alternativa a la fuerza 345
2.15 Las leyes de conservación. La conservación de la cantidad de movimiento o ímpetu 346
2.16 La energía cinética 347
2.16.1 Teorema Trabajo-Energía Cinética 347
2.17 Energía potencial 348

TEMA 3 ELECTRICIDAD Y MAGNETISMO

3.1 Carga eléctrica 350
3.1.1 Conservación de la carga 351
3.1.2 Ley de Coulomb 353
3.1.3 Campo eléctrico 355
3.1.4 Potencial eléctrico 355
3.2 Nociones de circuitos simples 357
3.2.1 Circuitos 357
3.2.2 Potencia eléctrica 359
3.3. Nociones de electromagnetismo 360
3.3.1 Campo magnético 360
3.3.2 Materiales ferromagnéticos, paramagnéticos y diamagnéticos 361
3.3.3 Bobinas, campos magnéticos y corrientes eléctricas 361
3.3.4 Generación de un campo magnético por una corriente eléctrica 362
3.3.5 Generación de una corriente eléctrica por un campo magnético 364
3.3.6 Ley de Faraday 366
3.4 Ondas electromagnéticas. Espectro electromagnético 368

TEMA 4 ÓPTICA

Introducción 370

4.1 Óptica geométrica 371
- 4.1.1 Imágenes en espejos curvos 374
- 4.1.2 Refracción de la luz. Ley de Snell 375
- 4.1.3 Formación de imágenes con una lente delgada biconvexa 377

4.2 Naturaleza de la luz 384

TEMA 5 FÍSICA DE FLUIDOS

Introducción 386

5.1 Nociones de hidrostática 387
- 5.1.1 Presión atmosférica 387
- 5.1.2 Unidad de la presión 388
- 5.1.3 Variación de la presión atmosférica 389
- 5.1.4 Presión hidrostática. Principio de Pascal 389
- 5.1.5 Medición de la presión atmosférica 391
- 5.1.6 Presión debajo de la superficie del agua 392
- 5.1.7 Principio de Arquímedes. Peso relativo o aparente 393
- 5.1.8 Peso aparente o relativo 396

5.2 Nociones de hidrodinámica 398
- 5.2.1 Ecuación de continuidad 398
- 5.2.2 Ecuación de Bernoulli 400

TEMA 6 TERMODINÁMICA

Introducción 405

6.1 ¿Cómo protegernos del frío en invierno y del calor en verano? 407
- 6.1.1 Nociones preliminares sobre temperatura: paredes adiabáticas y diatérmicas. Conductividad térmica 407
- 6.1.2 Energía interna, calor y equilibrio térmico 408
- 6.1.3 ¿Por qué los objetos de metal y madera se sienten a diferente temperatura? 410
- 6.1.4 Noción científica de la temperatura 412
- 6.1.5 Construcción de un termómetro 414
- 6.1.6 Ley cero o de transitividad de la termodinámica 416
- 6.1.7 El termómetro de gas a volumen constante y la lectura "correcta" de la temperatura de un objeto 416
- 6.1.8 Ecuación de estado de un "gas muy diluido" o gas ideal o perfecto 419

6.2 ¿Cómo ahorrar energéticos en el hogar? 426
- 6.2.1 Conservación de energía 426
- 6.2.2 Capacidad térmica 427
- 6.2.3 Primera ley de la termodinámica 432
- 6.2.4 Ahorro de gas 433

6.3 ¿Cómo reducir la contaminación para un desarrollo sustentable? 441
- 6.3.1 Generación de electricidad por combustibles fósiles 442
- 6.3.2 Motores térmicos 442
- 6.3.3 La segunda ley de la termodinámica 455
- 6.3.4 Tarea termodinámica y su eficiencia. Contraste termodinámico y exergía 458
- 6.3.5 La exergía 459
- 6.3.6 Ahorro de exergía 461

6.3.7 Eficiencia de la segunda ley de la termodinámica 461
6.3.8 Desarrollo sustentable 462
6.3.9 Uso lineal y cíclico de los recursos exergéticos 464
6.3.10 Huella ecológica 465
6.3.11 Consumo de recursos energéticos agotables 466
6.3.12 Los energéticos renovables 467
6.3.13 Ecoaldeas y ecomunicipios 468

TEMA 7 LO MÁS PEQUEÑO

Introducción 470

7.1 Concepción de átomo 472
7.1.1 Rayos catódicos 473
7.1.2 El corpúsculo llamado electrón 475
7.1.3 El modelo atómico de Thomson 478
7.1.4 El modelo de Rutherford: el descubrimiento del núcleo atómico 479
7.1.5 La atómica trinidad 481
7.1.6 La búsqueda de nuevas partículas 482

Bibliografía básica 483

APÉNDICE FÍSICA

Aristóteles 487
Obras

Galileo Galilei 494
Diálogo acerca de dos nuevas ciencias

Albert Einstein 498
La relatividad

COMPUTACIÓN

Los autores 503

Agradecimientos 505

Introducción 507

TEMA 1 COMPUTACIÓN

1.1 Introducción: entre el polvo y la divinidad 509

1.2 Problemas 512
1.2.1 El problema de los regalos de Arcadio 513

1.3 Problemas de la vida cotidiana 515
1.3.1 El significado de resolver un problema 515

1.4 Algoritmos: resolviendo el problema 516
1.4.1 Una solución: búsqueda exhaustiva 516
1.4.2 Análisis de la solución de una búsqueda exhaustiva 517
1.4.3 Análisis del caso general de la búsqueda exhaustiva 517

1.5 Crecimiento exponencial 519
1.5.1 Crecimiento exponencial en computación 519
1.5.2 Crecimiento exponencial en la sociedad y en la naturaleza 520
1.5.3 Ejemplos de crecimiento exponencial 522
1.5.4 Árboles 523
1.5.5 Qué tan rápida es una computadora 523
1.5.6 Ejemplos de crecimiento exponencial benéficos 525
1.6 Problemas probablemente difíciles, seguramente difíciles y aun peores 526
1.6.1 Problemas exponenciales 526
1.6.2 Problemas NP-completos 527
1.6.3 Problemas peores que los exponenciales 527
1.7 Resumen 528

TEMA 2 ALGORÍTMICA

2.1 Introducción a la algorítmica 529
2.1.1 Cocinar galletas 530
2.1.2 Recetas de cocina versus algoritmos 531
2.1.3 Tipos de algoritmo 533
2.2 Inducción y gráficas 534
2.2.1 El método de inducción 534
2.2.2 Colorear mapas 535
2.2.3 Gráficas 538
2.2.4 Acertijo de los tróminos 539
2.2.5 Probando funciones por inducción 540
2.3 Recursividad 543
2.3.1 Algoritmo para colorear mapas con seis colores 543
2.3.2 Un problema y un algoritmo: las torres de Hanoi 544
2.4 Búsqueda exhaustiva 553
2.4.1 Coloración de gráficas 554
2.4.2 El problema de ordenamiento 554
2.4.3 La pareja de puntos más cercanos 555
2.5 Divide y vencerás 556
2.5.1 Ordenamiento por inserción 556
2.5.2 Ordenamiento de burbuja 558
2.5.3 Búsqueda binaria 559
2.5.4 Ordenamiento por combinación 560
2.5.5 Ordenamiento rápido 560
2.5.6 Parejas de puntos 561
2.6 Órdenes de crecimiento 562
2.7 Resumen 564

TEMA 3 PROGRAMACIÓN

3.1 Introducción 565
3.1.1 La programación y su importancia 565
3.1.2 Lenguajes de programación 567
3.1.3 Notas acerca de programación y el lenguaje presentado 560
3.1.4 Un primer ejemplo: las torres de Hanoi 569

3.2 Nociones básicas de Scheme 571
3.2.1 El lenguaje de programación Scheme 572
3.2.2 El ambiente de programación DrScheme 572
3.2.3 Metodología de diseño 573
3.2.4 Expresiones primitivas y datos simples 573
3.2.5 Recursividad 580
3.2.6 Ciclos 582
3.2.7 Asignación 582
3.3 Abstracción con datos 585
3.3.1 Definición de estructuras 585
3.3.2 Constructores y selectores 587
3.3.3 Operaciones con listas 589
3.3.4 Recorriendo listas 591
3.3.5 Datos simbólicos 592
3.3.6 Vectores, gráficas y laberintos 595
3.3.7 Construcción de laberintos perfectos 599
3.4 Resumen 603

TEMA 4 INFORMACIÓN

4.1 Los miedos de la futura suegra de Arcadio 604
4.2 Símbolos 605
4.2.1 Símbolos, palabras, mensajes 606
4.2.2 Bits 607
4.3 Representando el mundo mediante bits 609
4.3.1 Representando números 609
4.3.2 Representando imágenes 610
4.3.3 Codificación y el mundo 611
4.3.4 Comprensión en la computadora 613
4.3.5 Comprensión en la naturaleza 614
4.3.6 Expansión en computadoras 616
4.3.7 Códigos detectores y correctores de errores 617
4.4 Medir información 619
4.4.1 Cantidad de información y entropía 621
4.4.2 Codificación eficiente 623
4.4.3 Criptografía básica 625
4.4.4 Protocolos criptográficos 626
4.5 Resumen y conclusiones 629
4.5.1 Tiempo contra espacio 629
4.5.2 Limitaciones de codificación 630
4.5.3 Limitaciones de la criptografía 630
4.5.4 Derechos de autor 631

TEMA 5 ABSTRACCIÓN 633

5.1 La abstracción 635
5.1.1 Los inicios: sentido de número y contar 635
5.1.2 Abstracción: el camino del conocimiento 636
5.1.3 Abstracción en computación 637

5.2 Modelos de cómputo 639
5.2.1 Máquinas de estados finitos 640
5.2.2 La geometría plana, un modelo de cómputo restringido 643
5.2.3 Modelos de computadoras 644
5.2.4 Tesis de Church-Turing 648
5.3 Lógica 648
5.3.1 El sueño de Leibniz 648
5.3.2 Un problema fundamental 649
5.3.3 La limitación inherente de las matemáticas 650
5.3.4 Álgebra booleana 651
5.3.5 Lógica de primer orden 651
5.3.6 Lógica y conocimiento 655
5.4 Análisis de problemas 657
5.4.1 En el banco 657
5.4.2 La visión del computólogo 657
5.4.3 Abstracción en programación 658
5.5 Resumen 659

TEMA 6 COMPUTADORAS

6.1 Problemas de electricidad 660
6.2 Sótano: transistores y funciones de conmutación 662
6.2.1 Transistores 663
6.3 Planta baja: compuertas y circuitos integrados 665
6.3.1 Compuertas elementales 665
6.3.2 Diseño lógico 666
6.3.3 Transistores y compuertas NAND 669
6.4 Primer piso: arquitectura de computadoras 670
6.4.1 Arquitectura de Von Neumann 670
6.4.2 Frecuencia de operación 672
6.4.3 La jerarquía de memoria: la idea del caché 672
6.5 Segundo piso: lenguajes de bajo nivel 675
6.5.1 Lenguaje de máquina 675
6.5.2 Ejecución con cauce segmentado 676
6.5.3 Lenguaje ensamblador 678
6.6 Tercer piso: sistemas operativos 679
6.6.1 Manejo de procesos: planificadores 679
6.6.2 Manejo de memoria 680
6.6.3 Sistemas de archivos (otra abstracción) 681
6.6.4 Interfaz de texto o gráfica 681
6.7 Penthouse: software de aplicación 683
6.8 Resumen 683

TEMA 7 REDES

7.1. Cómputo distribuido 684
7.1.1 Introducción 684
7.1.2 Exclusión mutua 685
7.1.3 Propiedades de la exclusión mutua 686

7.2 Comunicación 686
7.2.1 Otro problema de la vida real 686
7.2.2 Un mismo lenguaje 687
7.2.3 Protocolo 687
7.2.4 Consenso 688
7.2.5 Comunicación uno a uno 690
7.2.6 Comunicación transitoria y persistente 691
7.3 Redes 691
7.3.1 Correo terrestre 691
7.3.2 Redes de computadoras 693
7.3.3 Conmutación de paquetes 694
7.3.4 Comunicación entre homólogos y abstracción 695
7.3.5 Enrutamiento 696
7.3.6 Congestión 696
7.3.7 Transmisión 696
7.4 Internet: red de redes 697
7.4.1 Infraestructura 697
7.4.2 Direccionamiento 699
7.4.3 Idioma 700
7.5 Web 701
7.5.1 Qué es la web 701
7.5.2 Memoria inmediata o caché 702
7.5.3 Estándares 702
7.5.4 La revolución web 703
7.5.5 La web semántica: llevar la web a nuevos niveles 703
7.5.6 Motores de búsqueda o buscadores 704
7.6 Aplicaciones 705
7.6.1 E-mail 705
7.6.2 Mensajería instantánea 707
7.6.3 Acceso remoto 708
7.6.4 Colaboración y software libre 708
7.6.5 Producción por comunes 709
7.7 Resumen 709

TEMA 8 MULTIMEDIA

8.1 Introducción 710
8.2 Texto 711
8.2.1 Diseño de tipos 711
8.3 Sonido 712
8.3.1 Almacenar bits 712
8.4 Imágenes y video 713
8.4.1 Imágenes digitales 713
8.4.2 Adquisición de imágenes 716
8.4.3 Algunas características de las imágenes digitales 716
8.4.4 Segmentación de imágenes 718
8.4.5 Representaciones y descripciones de objetos 719
8.4.6 Reconocimiento de objetos 729

8.5 Animación 730
8.5.1 Realidad virtual inmersiva 732
8.5.2 La animación en el cine 734
8.5.3 Generación de imágenes para la pantalla azul 735
8.6 Resumen 735

TEMA 9 APLICACIONES

9.1 Introducción 736
9.2 Ciencias de la Tierra 737
9.2.1 Cartografía automatizada 737
9.2.2 Geomorfología 738
9.2.3 Climatología 739
9.2.4 Meteorología 740
9.3 Robótica 741
9.3.1 Robots de servicio 741
9.3.2 Los robots en la literatura 744
9.4 Juegos 745
9.4.1 Ajedrez, un juego difícil 746
9.4.2 Go, un juego imposible 748
9.4.3 Hacia una solución 749
9.4.4 Generación de sólidos con voxeles 750
9.5 Bioinformática 751
9.6 La computación en los negocios 753
9.7 La computación y el arte 756
9.8 Resumen 757

Glosario 758
Bibliografía 760

APÉNDICE COMPUTACIÓN

Paul Strathern 765
Turing y la computadora

Similarmente el área BC^2 es igual al área del rectángulo BMLD

Los rectángulos AELM y BMLD suman el cuadrado de la hipotenusa AB

$h = AC$

$\therefore a_{\Delta ABF} = \frac{AC^2}{2}$

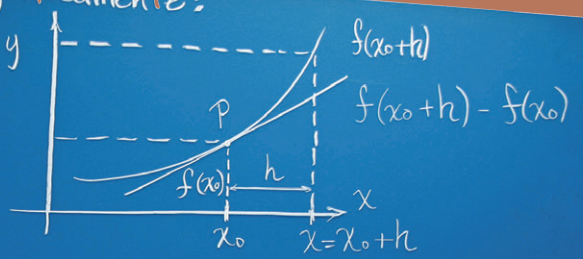
$h = AM$

$a_{\Delta AEC} = \frac{AM \times ML}{2}$

$\measuredangle ADC$

$BD \perp BC$

gráficamente:



Ejemplo:

Calcular la derivada de la función $f(x) = 3x^2$, en el punto $x=2$

se sabe que $x = x_0 + h$

siendo $x_0 = 2$

$$f(2) = \lim_{h\to 0} \frac{f(2+h) - f(2)}{h}$$

$$= \lim_{h\to 0} \frac{3(2+h)^2 - 3(2)^2}{h}$$

$$= \lim_{h\to 0} \frac{3(4+4h+h^2) - 12}{h}$$

$$= \lim_{h\to 0} \frac{12+12h+3h^2 - 12}{h}$$

$$= \lim_{h\to 0} (12+12h) = 12+0 = 12$$

Regla de la cadena

$h(x) = g[f(x)]$

$\Rightarrow h'(x) = g'[f(x)] \cdot f'(x)$

se expresa

$\frac{dh}{dx} = \frac{dg}{df}\frac{df}{dx}$

$\Rightarrow \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$

$\frac{f(x)}{f'(x)}$

$\frac{\sqrt{b^2-4ac}}{2a}$

$D = b^2 - 4ac$

$a=5, b=-3, c=-2$

$D = (-3)^2 - 4(5)(-2)$

$D = 9+40$

$x = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$

$x = \frac{-(-3) \pm \sqrt{(-3)^2-4(5)(-2)}}{2(5)}$

$x_1 = \frac{9+\sqrt{49}}{10}, \quad x_2 = \frac{9-\sqrt{49}}{10}$

$x_1 = \frac{9+7}{10} = \frac{16}{10}$

$x_2 = \frac{9-7}{10} = \frac{2}{10}$

Ecuación 3er grado

$x^3 + ax^2 + bx + c = 0$

$$x = \sqrt[3]{-\frac{q}{2} + \sqrt{\Delta}} + \sqrt[3]{-\frac{q}{2} - \sqrt{\Delta}} - \frac{a}{3}$$

donde

$p = \frac{3b-a^2}{3}, \quad q = \frac{2a^3 - 9ab + 27c}{27}$

$\Delta = \left(\frac{q}{2}\right)^2 + \left(\frac{p}{3}\right)^3$

si

$p = -3uv$

$u = \sqrt[3]{-\frac{q}{2} + \sqrt{\Delta}}$

$v = \sqrt[3]{-\frac{q}{2} - \sqrt{\Delta}}$

$v^3 = -\frac{\left(\frac{p}{3}\right)^3}{\left(\frac{q}{2}\right)}$

$v^3 = -\frac{q}{2} -$

$= \pi \cdot \left(2L\frac{\Delta r}{2} + 2r\frac{\Delta L}{2}\right)$

$= \pi \cdot (L\Delta r + r\Delta L)$

por semejanza de triángulos

$\frac{\Delta r}{\Delta L} = \frac{r}{L}$

$\therefore L\Delta r = r\Delta L$

$\left(L + \frac{\Delta L}{2}\right)$

$\Delta S =$

$\sum \Delta S_i = 2\pi RH$

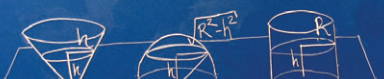
donde H es la altura total cubierta de la esfera

$V_\Delta = \frac{1}{3}\pi r^2 h$

$V_O = \frac{4}{3}\pi R^3$

$V_\square = \pi r^2 h$

$S_E : S_C :: 2:3$



LOS AUTORES

JOSÉ LUIS ABREU LEÓN

Fue profesor en la Facultad de Ciencias, miembro fundador del CIMAT, investigador y director del IIMAS, profesionista autónomo en España especializado en el desarrollo de software educativo, director técnico del proyecto MALTED de la Comisión Europea y director de Desarrollo Tecnológico en el Instituto Latinoamericano de la Comunicación Educativa. Actualmente es técnico académico en el Instituto de Matemáticas de la UNAM. Creador de la herramienta Descartes con la que se desarrollan contenidos educativos interactivos de matemáticas y física en España y México. Sus intereses son el desarrollo de herramientas para la creación de contenidos digitales interactivos y la publicación en web de tales contenidos para la enseñanza, la difusión y la investigación.

MICHAEL BAROT

Estudió matemáticas en la Universidad de Zúrich y obtuvo el doctorado en el Instituto de Matemáticas de la UNAM. Por su tesis doctoral recibió el premio Weizman. Desde 1998 trabaja como investigador en el Instituto de Matemáticas. Es miembro del Sistema Nacional de Investigadores y de la Academia Mexicana de Ciencias. Ha participado en la Maestría en Docencia para la Educación Media Superior (MADEMS), en la elaboración de la serie de videos "Aventuras Matemáticas" y también en el proyecto Ixtli, donde fungió como corresponsable. Tiene tres libros publicados, uno para nivel secundaria, otro de bachillerato y el último de licenciatura.

JAVIER BRACHO

Estudió la licenciatura en matemáticas en la Facultad de Ciencias de la UNAM y obtuvo el doctorado en el Instituto Tecnológico de Massachusetts (MIT). A partir de entonces es profesor de la Facultad de Ciencias e investigador del Instituto de Matemáticas de la UNAM, del cual es actualmente director. Es, desde su inicio, miembro del Sistema Nacional de Investigadores. Recibió la Distinción Universidad Nacional para Jóvenes Académicos en Docencia en Ciencias Exactas en 1993. Tiene dos libros publicados, uno de divulgación y otro de texto, en el Fondo de Cultura Económica, donde también participa en el comité de ciencias.

AGRADECIMIENTOS

En la elaboración de este libro participaron muchas personas a las que debemos agradecer. En las múltiples sesiones de discusión, planeación y diseño conceptual participaron Concepción Ruiz Ruiz-Funes, Emiliano Mora y Paloma Zubieta López. Concha y Paloma también contribuyeron con textos de algunos apartados, y Paloma, además, corrigió el estilo y preparó los textos para su formación; también coordinó las fases finales de producción del libro y, cual hada madrina, juntó los fragmentos para hacerlo realidad. Patricia Covarrubias y Juan Andrés Burgueño contribuyeron en el apartado de estadística. Las maestras Ma. Eugenia Otero Ulibarri y Dora Lidia Rodríguez Zúñiga, del Colegio de Ciencias y Humanidades, junto con los maestros Heriberto Marín Arellano y Emilio Velarde González Baz, de la Escuela Nacional Preparatoria, tuvieron la amabilidad de leer nuestros manuscritos y escuchar nuestras ideas e intenciones; sus críticas y comentarios, siempre atinados y constructivos, influyeron positivamente en el resultado. Les agradecemos el tiempo y dedicación que generosamente nos brindaron.

Por la producción del DVD agradecemos a las siguientes personas: Óscar Escamilla coordinó la producción general y la de los materiales interactivos. El propio Óscar, junto con Carlos Alberto Jaimes, Abraham Pita, Julio Prado, Carlos Serrato y Erika Tovilla, diseñaron y programaron los materiales interactivos en los que Mariana Villada Carbó corrigió el estilo y la redacción. Carlos Alberto Jaimes coordinó la producción de los videos y también creó la mayoría de las ilustraciones y animaciones. Roberto Elier, Mariana Villada y Paloma Zubieta escribieron los guiones. Felipe Bonilla fue el realizador de los videos y estuvo a cargo también de la edición y la posproducción. La voz en los videos es de Pablo Flores. Las fotografías usadas en los fondos son de Yamina del Real. El diseñador gráfico de todo el contenido del disco es Alfonso Pascal. En la asesoría académica de los contenidos de los videos participó Gonzalo Zubieta Badillo.

Con seguridad olvidamos mencionar la generosa contribución de algunas personas que apoyaron la realización de este libro. A ellos también damos las gracias y rogamos disculpen la omisión. Finalmente, reconocemos el excelente trabajo de edición que orquestaron Rosanela Álvarez desde la UNAM y María Oscos desde Siglo XXI.

p. xxvi, Pizarrones de clase en que se desarrollan algunos de los temas centrales en matemáticas | © Yamina del Real y Carlos Jaimes.

INTRODUCCIÓN

LOS AUTORES

Este libro es una suerte de mosaico y caleidoscopio. Mosaico, pues consta de piezas independientes que se ensamblan en un todo, y caleidoscopio porque repite y repite la misma imagen reflejada en tres espejos: la imagen es la creación matemática; los espejos son la actividad humana, la naturaleza y la matemática misma. Ante todo, las matemáticas son una actividad creativa —cada pieza del mosaico da muestra de ello—, surgen de considerar problemas de diversa índole cuyas soluciones van armando, como si fueran piezas en un rompecabezas, una estructura de conocimiento de una consistencia y solidez sorprendente. Esta contundencia reside en que se crean o recrean dentro de cada mente humana que se acerca a ellas, en que se basan en la razón a tal grado que trascienden culturas y épocas históricas y en que, siendo tan abstractas, resultan estar ligadas a la realidad de maneras insólitas y fundamentales. Sin ellas, lo que llamamos ciencia no existiría y nuestra vida cotidiana actual sería impensable.

Es muy común que en una clase de matemáticas surja la pregunta ¿esto para qué sirve? En este libro, además hacemos otra: ¿por qué se creó, de dónde nace? Al responderla, identificamos tres fuentes básicas: la actividad humana, la comprensión de la naturaleza y la matemática misma, que se tratan en los temas 2, 3 y 4, respectivamente. En el tema 1, a manera de introducción, ahondamos en el porqué de esta estructura, en qué queremos decir con ella. Los otros tres temas consisten en apartados independientes y autónomos que se inician desde cero. Al avanzar en la lectura de los apartados, es inevitable que ésta vaya complicándose, pues trata sobre ideas —ideas matemáticas— que requieren de tiempo, reflexión, creatividad y esfuerzo por parte del lector. Sin embargo, si alguien tropieza demasiado o desespera, puede pasar al siguiente apartado o a cualquier otro, pues tampoco están seriados —salvo por el tema 1, concebido como un todo—. Podría decirse que es un libro para "picar" en el que, después de la introducción, se vale dar una probadita por aquí, otra por allá, echarse algún bocadillo completo o bien leerlo de corrido.

No es un libro de texto, pero sí pretende ser un apoyo para el alumno y para el maestro. Este último puede encontrar nuevas ideas para tratar temas en clase, maneras distintas de enfocarlos o ejemplos interesantes para los alumnos. El estudiante podrá descubrir aspectos novedosos que fomenten su curiosidad por la materia. El público en general encontrará una visión amplia y actualizada de la actividad matemática.

El nivel en el que se inicia cada apartado es de bachillerato, pero no necesariamente es el mismo en el que acaba, a veces se vuelve técnico y otras meramente descriptivo y cultural. Las partes más técnicas o difíciles de seguir están marcadas con el icono  , que indica que se requiere de más tiempo y esfuerzo para comprenderlas e, insistimos, pueden posponerse sin menoscabo del entendimiento general. Preferimos incluirlas que omitirlas, pues las matemáticas siempre representan retos intelectuales y no pretendemos esconderlos o disfrazarlos, sino exponerlos tal cual son.

En este libro presentamos a las matemáticas sobre todo como un fenómeno cultural, y para ello escogimos algunos ejemplos de su desarrollo. Tocamos diversos temas de las matemáticas para que, como en la música, al escuchar diversos géneros, comprendamos mejor su variedad y riqueza. En son de broma, quisiéramos que fuera "un viaje a Acapulco": con algo para todos los bolsillos y todos los gustos.

El libro incluye un DVD que se produjo con el mismo espíritu; contiene cápsulas históricas y biográficas, además de unidades educativas interactivas. Algunas tienen relación con partes del texto, pero otras son independientes. Amplían la visión que, a nuestro entender, debiera tener un bachiller de las matemáticas.

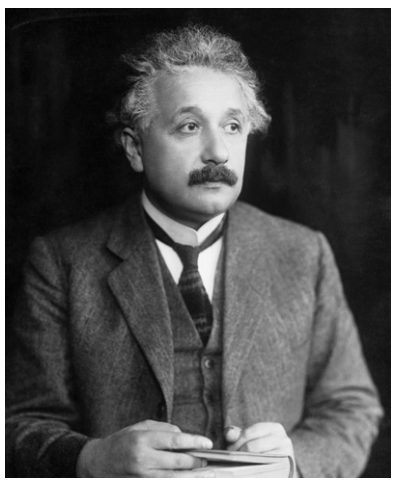
¿CÓMO SE ESCRIBEN LAS MATEMÁTICAS?

Las matemáticas son una ciencia peculiar: usan un lenguaje muy especial al que es necesario acostumbrarse. Este lenguaje, como cualquier otro, requiere ser aprendido; de lo contrario, representa una barrera de entendimiento que impide acercarse, disfrutar y comprender el contenido.

En su libro *Breve historia del tiempo*, el físico británico Stephen Hawkins relata su experiencia con la editorial, donde le dijeron que "por cada fórmula en tu texto, reduces tu público a la mitad". Hawkins intentó no poner una sola fórmula, pero al final, no lo pudo resistir e incluyó la fórmula más famosa de todas:

$$E = mc^2.$$

La fórmula es de Albert Einstein, físico alemán del siglo xx, y es consecuencia de su teoría general de la relatividad.



Einstein en 1921 | © *Latin Stock México.*

Para muchas personas esta fórmula tiene un significado casi religioso. Se sabe que significa algo, mucho, que expresa algo trascendental. Es claro que está escrita en lenguaje matemático. Se ve por el signo de igualdad —las dos rayas paralelas— que es una *ecuación*, donde se igualan los dos lados. Cada vez que aparece el símbolo de igualdad " =", separa dos lados —el izquierdo y el derecho— y los iguala. En la ecuación de Einstein, el lado izquierdo es la simple letra E, que significa la cantidad física *energía*. El lado derecho es más complicado: en él aparecen tres símbolos: m, c y un 2. La letra m significa *masa*, la letra c es la *velocidad de la luz* y el 2 significa que se eleva la velocidad de la luz al cuadrado, es decir, se multiplica por sí misma. Se sobreentiende, además, que cuando dos letras que representan cantidades independientes están juntas, las cantidades se multiplican. Así que la fórmula en palabras se lee:

La energía es igual al producto de la masa por el cuadrado de la velocidad de la luz.

La velocidad de la luz es fija, una constante. En el vacío, la luz avanza 299 792 458 metros cada segundo, es decir, daría casi ocho vueltas a la Tierra en un segundo. Pero ¿de qué masa se está hablando, de qué energía? Lo que expresa esta fórmula es que la energía y la masa de cualquier sistema físico se pueden convertir una en la otra. Como la velocidad de la luz es tan grande, la ecuación dice que muy poca masa equivale a una energía enorme.

En este ejemplo se pueden apreciar varios aspectos:

- Una ecuación tiene dos lados, o como se dice en matemáticas, dos *miembros*.
- Cada uno de los miembros se puede constituir de diferente forma. Puede ser una simple letra o algo más complejo.
- En una ecuación se deben identificar los diferentes ingredientes: las letras usadas, que pueden ser *variables* como la masa y la energía, o *constantes*, como la velocidad de la luz. Sus significados dependen del *contexto* en el que se plantea la ecuación.
- Una ecuación debe interpretarse y su lectura no siempre es fácil. Se requiere de una buena instrucción para lograr la familiaridad con las ecuaciones. A veces, hay que leerla una y otra vez, hay que juntar sus piezas como en un rompecabezas y, entonces, podrá desplegar una belleza similar a la de Einstein.

Sabemos de la dificultad que provoca el uso de las ecuaciones, fórmulas y símbolos en las matemáticas, pero no podemos prescindir de ellos. En la historia, no siempre se usó la simbología actual. Por ejemplo, en 1559, el matemático francés Jean Buteau escribía:

$$I\diamondsuit P\,6\_P\,9\;[\;I\diamondsuit P3\_P\,24 \qquad (1)$$

Lo que hoy se escribe como:

$$x^2 + 6x + 9 = x^2 + 3x + 24$$

Unos sesenta años antes, en 1494, el matemático italiano Luca Pacioli escribía:

$$\text{Trouame.I.n°.che.gi\_to al suo quadrat° facia.12.}$$

Lo que hoy se escribe como:

$$x + x^2 = 12$$

Con buena voluntad se pueden descifrar estas maneras exóticas de denotar el contenido. Por ejemplo, en la ecuación (1) hay que entender los símbolos P como *plus*, es decir, como nuestro "más"; el símbolo $\diamondsuit$ como el cuadrado del número —aludiendo al área de un cuadrado con lado I—; la línea $\_$ como la misma variable I —que corresponde a nuestra x— y el símbolo $[$ como la igualdad. El segundo ejemplo es más cercano a cómo se lee en la actualidad la ecuación en italiano. Estos ejemplos muestran que hace 500 años no había consenso sobre cómo anotar las matemáticas. Fue un proceso largo, un desarrollo de siglos en el cual, poco a poco, se establecieron ciertas convenciones. Por ejemplo, el símbolo de igualdad que usamos hoy día ($=$) lo utilizó por primera vez el matemático inglés Robert Recorde, en 1557.

En los dos ejemplos anteriores de ecuaciones con la notación actual, el contexto es implícito. Estas ecuaciones son diferentes de la de Einstein en que los símbolos no tienen sig-

nificados extra, sino que se sobreentienden. Puesto que sólo aparece la letra x, además de los números y los símbolos de suma e igualdad, interpretamos que x también representa a un número. Plantean, entonces, la pregunta ¿existirá un número que al sustituir en vez de x, cumpla la ecuación? Eso sería resolver la ecuación: a veces se puede y otras no, como veremos más adelante en otras partes de este libro.

Se puede decir que, en gran medida, las matemáticas fueron tan prolíficas a partir del siglo XVIII gracias a una simbología y notación más simple, consensuada entre la comunidad de los matemáticos. Se pudieron expresar y comunicar mejor; además, resulta que una buena notación a veces ayuda a entender. En particular, la física moderna es absolutamente impensable sin el uso de las matemáticas, y eso no quiere decir números, sino conceptos, simbología, notación y métodos involucrados.

El uso del lenguaje simbólico es un cuchillo de doble filo: por un lado, hace extremadamente eficiente la notación y el manejo de conceptos pero, por otro, constituye un obstáculo serio para entenderlos. Sería un error imperdonable pensar que las matemáticas *sólo* son fórmulas. Más bien, las matemáticas son lo que está escondido en las fórmulas y la mejor manera de explicar aquello, es decir, lo escondido, es a veces justo a través de éstas. Cualquier otra manera de intentarlo es más complicada y tortuosa.

La política que adoptamos en este libro es tratar de evitar el formalismo riguroso y procurar transmitir las matemáticas mismas. Hemos hecho un gran esfuerzo por omitir fórmulas innecesarias y por llegar a lo que está en el fondo mediante vías alternas, sin transitar por la simbología. Que esto sólo fue posible en relativamente pocos casos es consecuencia de la complejidad de las matemáticas mismas. Muchas veces la fórmula es el camino de comunicación menos malo y cuando ya se tiene familiaridad con su uso, se convierte en el camino más directo. Insistimos en esta advertencia, pues sabemos de las múltiples dificultades. Este libro no se escribió para *aprender* matemáticas —creemos que esto sólo se puede lograr mediante un intercambio más dinámico y activo—, se escribió para orientar sobre los alcances de esta ciencia, sobre su significado en nuestra cultura y, en particular, sobre su inserción en la tecnología y en la vida del siglo XXI.

EL PORQUÉ DE LAS MATEMÁTICAS

TEMA 1

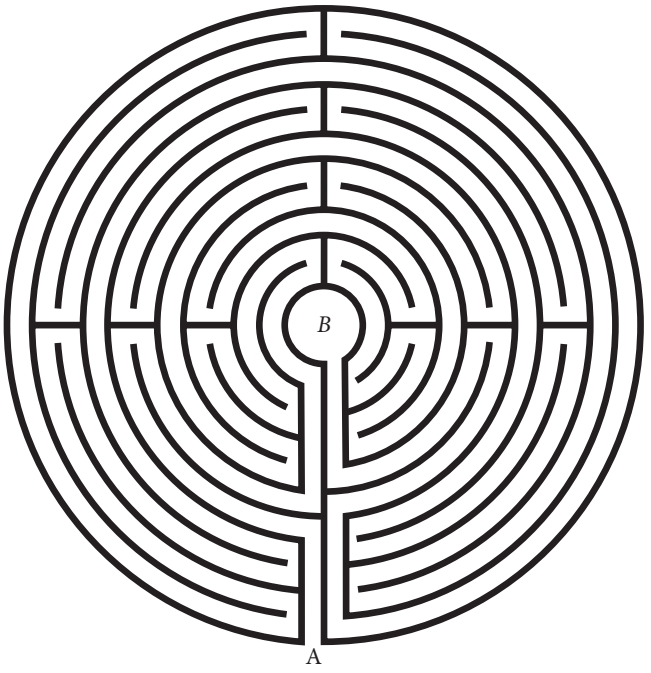


Figura 1.1 Un laberinto medieval donde no es posible perderse. Éste es el origen de los laberintos modernos con bifurcaciones y que presentan un reto para entrar y salir.

1.1 INTRODUCCIÓN

Como se explica en la introducción a la materia, este libro se divide en tres partes que corresponden a las principales fuentes de motivación para la creación matemática: la actividad humana, la naturaleza y las propias matemáticas. El propósito de este primer capítulo es ilustrar, con ejemplos sencillos, dichas fuentes de motivación matemática.

Sin embargo, es conveniente aclarar que hay una fuente de creación matemática que antecede a las otras y sin la cual serían estériles: la irresistible atracción del ser humano por los retos de todo tipo, en especial, los intelectuales. Esta atracción no se limita a un grupo especial o selecto de personas, como los matemáticos, sino que se puede reconocer en cualquier persona, por ejemplo, cuando en un café o en el metro se resuelve como entretenimiento uno de los *sudokus* impresos a diario en los periódicos. ¿Qué es lo que lleva a una persona a intentar, durante largas horas, resolver estos retos? ¿Qué gana con ello? Quienes nunca lo

han hecho les puede parecer extraño, pero lo único que se obtiene es la satisfacción de haberlo resuelto, de lograr algo con la mente.

						7	9	
5			9	8				
	8	4						
		3	4					7
			7		3			
	6	8				2		
							7	
8	2				6	1		
		1	8			9		

Figura 1.2 Las reglas del sudoku *son sencillas: las celdas deben llenarse con las cifras del 1 al 9 de tal manera que no se repita una de ellas en la misma fila, columna o cuadrado pequeño.*

Hace mucho tiempo que los periódicos imprimen juegos de entretenimiento o desafíos mentales, como los crucigramas. Unos cien años atrás estaban de moda los problemas de ajedrez, hace unos decenios el cubo de Rubik y ahora los *sudokus*. Los periódicos están al corriente de que a muchos de sus lectores les encantan estos juegos y de que hay quienes los compran sólo por ellos, por consiguiente, siempre tienen secciones que los contienen.

Antes de los *sudokus* o de otros desafíos ya mencionados, muchas personas dedicaban su tiempo libre a encontrar **cuadrados mágicos,** que son arreglos de números en los que se obtiene la misma suma en cada fila y columna. Podemos ver un ejemplo en el grabado *Melancolía* de Alberto Durero, arriba y a la derecha, bajo la campana.

La figura 1.4 es un cuadrado mágico más grande que el del cuadro de Durero, donde la suma de cada fila y columna es de 65.

Figura 1.3 Melancolía, *de Alberto Durero, 1514* | *© Interfoto, Archivo Digital.*

9	22	20	1	13
21	5	7	24	8
12	19	2	14	18
17	4	25	3	16
6	15	11	23	10

Figura 1.4 Cuadrado mágico de 5×5.

Los cuadrados mágicos eran conocidos desde el año 650 a.C. por los matemáticos chinos y alrededor del siglo VII d.C. por los árabes. Aparecen también en las culturas de India y Persia y, en cada una de ellas, se les atribuyen distintos poderes, como el de atraer la suerte.

Podemos construir fácilmente un cuadrado mágico de tamaño tres por tres. Primero determinamos cuál debe ser la suma de los números en cada fila y columna. Si usamos los números del 1 al 9 la suma es:

$$1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 = 45$$

Como esta suma debe distribuirse en tres renglones, cada uno debe sumar 15. Si en alguna celda ponemos el número más grande, o sea el 9, debemos sumarle solamente 6 en las otras dos celdas de su misma fila y su misma columna. La única forma de sumar 6 con los números restantes son: $1 + 5$ y $2 + 4$, ya que $3 + 3$ repite un número. Usamos estas dos combinaciones para llenar la fila y la columna donde ya habíamos colocado el 9, así que el actual estado de nuestro cuadrado mágico es como sigue:

9	1	5
2		
4		

En las cuatro celdas restantes debemos colocar los cuatro números que faltan: 3, 6, 7 y 8. Buscaremos ahora la ubicación del número 7. No puede estar en la misma fila que el 4 dado que $4 + 7 = 11$ y nos faltarían justo otros 4 para alcanzar la suma 15. De la misma manera, no puede estar en la columna del 1, pues faltarían 7. Por ello, sólo queda un lugar para el 7: debe estar en la columna del 5 y la fila del 2. Ahora, es fácil terminar de rellenar el cuadrado mágico. Se deja este ejercicio para que el lector lo termine.

¿Qué pasos seguimos para construir el cuadrado mágico de 3 por 3? No se probaron muchas distribuciones para ver, si de casualidad, una funcionaba. Tampoco seguimos una estrategia ordenada para evaluar todos los posibles arreglos sin repetir uno, hasta encontrar el que buscábamos. Simplemente, empleamos el razonamiento lógico, ningún otro procedimiento puede llevarnos al resultado. Cuando logramos resolver un problema por deducción lógica, sentimos la misma satisfacción que un niño al poner la última pieza de un rompecabezas.

Usemos este cuadrado mágico para formar un *sudoku*: si intercambiamos filas o columnas obtenemos otros cuadrados mágicos de 3 por 3 que podemos acomodar en el tablero, por ejemplo, de la siguiente manera:

9	1	5	2	6	7	4	8	3
2	6	7	4	8	3	9	1	5
4	8	3	9	1	5	2	6	7
1	5	9	6	7	2	8	3	4
6	7	2	8	3	4	1	5	9
8	3	4	1	5	9	6	7	2
5	9	1	7	2	6	3	4	8
7	2	6	3	4	8	5	9	1
3	4	8	5	9	1	7	2	6

Figura 1.5 Sudoku *que contiene nueve cuadrados mágicos.*

Este *sudoku* terminado consiste en nueve cuadrados mágicos; si además elegimos, por ejemplo, la esquina superior izquierda de cada bloque, obtenemos un arreglo de tres por tres números que, a su vez, forman otro cuadrado mágico. Lo mismo ocurre para cualquier otra celda que elijamos en cada bloque.

Se pueden hacer cuadrados mágicos de todos los tamaños: tres por tres, cuatro por cuatro, cien por cien o tres millones por tres millones. También de uno por uno, que consiste en una única celda, pero como es el ejemplo más sencillo, no presenta mucho interés. No obstante, hay un tamaño imposible para los cuadrados mágicos. ¿Cuál es? Dejamos este reto al lector.

Los cuadrados mágicos no tuvieron mayor trascendencia en la historia ni en la ciencia. En un principio, quizá se emplearon con la esperanza de encontrar algo verdaderamente mágico pero, con el tiempo, se convirtieron en un simple pasatiempo. Sin embargo, cautivaron por igual a hombres y mujeres de distintas épocas y culturas, al evidenciar la fascinación de la humanidad por los retos intelectuales.

Si bien no todas las personas responden a estos desafíos, hay algo en la naturaleza humana que impulsa a plantearse problemas e intentar resolverlos. Este impulso es característico de la actividad matemática, aunque las matemáticas son mucho más que eso según veremos en las siguientes secciones.

1.2 LA ACTIVIDAD HUMANA

Figura 1.6 Mapamundi de gran utilidad para los navegantes, diseñado por el matemático alemán Gerardus Mercator. Hasta la fecha ésta sigue siendo la manera más usada de presentar la superficie de la Tierra en el plano | © *Latin Stock México.*

Cualquier alumno debe haberse preguntado alguna vez en clase: ¿esto para qué me va a servir? ¿Dónde están las matemáticas en la vida cotidiana? Sorprendentemente, basta con usar el teléfono celular o una computadora para estar cerca de las matemáticas... si se aprende a encontrarlas.

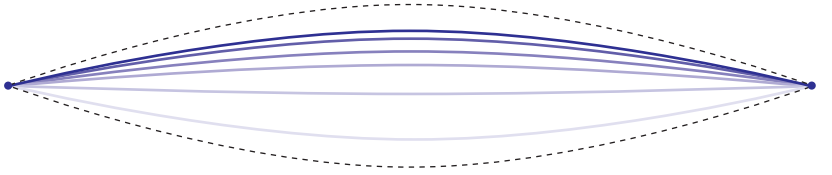
En esta sección presentamos tres problemas que pueden resultar divertidos y, hasta cierto punto, mostrar cómo las matemáticas se usan para resolver problemas de la vida cotidiana. El primero tiene que ver con la música, el segundo con la arqueología y el tercero con medir la profundidad de un pozo.

1.2.1 Música y matemáticas

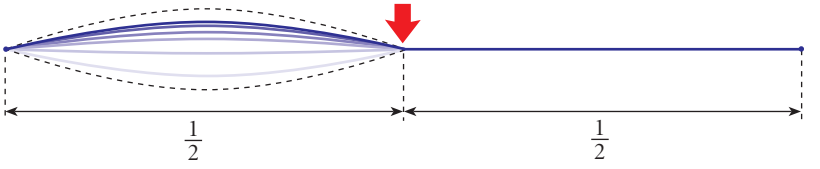
La relación entre la música y las matemáticas era ya conocida por las culturas caldea, egipcia, babilónica y china; sin embargo, hasta la escuela pitagórica, en la Grecia del siglo VI a.C., fue cuando estas actividades humanas quedaron unidas para siempre por la teoría de la cuerda vibrante.

Si se hace vibrar una cuerda, el sonido que produce depende de su longitud, su grosor y su tensión. El tono es más agudo conforme la cuerda se acorta. Lo que se descubrió en la escuela pitagórica es que, al dividir la cuerda, hay proporciones que producen sonidos más agradables que otros. Por ejemplo, si se divide la cuerda justo a la mitad, la vibración tiene un tono de una octava mayor que la producida por la longitud original. Por ello, la octava juega un papel fundamental en nuestra comprensión de la música.

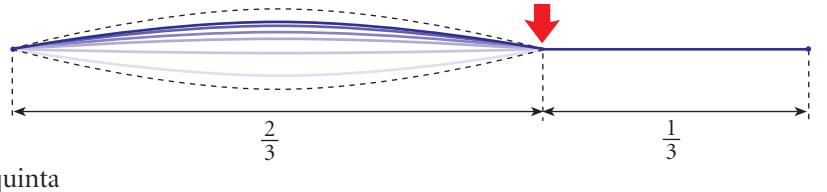
Los pitagóricos establecieron cuatro intervalos o relaciones entre las longitudes de las cuerdas que producían las únicas consonancias permitidas, es decir, aquellos sonidos que podían escucharse simultáneamente con un efecto agradable. Para producir todas las notas musicales sólo se tienen estos cuatro intervalos y sus combinaciones. El papel fundamental de las fracciones en la música era, sin duda, una de las razones por las cuales Pitágoras consideraba que la esencia de la realidad sólo podía expresarse por medio de números.



L unísono



$\frac{1}{2}L$ octava



$\frac{2}{3}L$ quinta

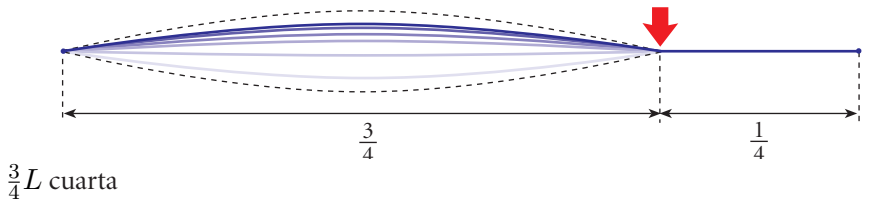


Figura 1.7 Intervalos entre las longitudes de las cuerdas. (incluye todas las imágenes anteriores).

La teoría de la cuerda vibrante se puede resumir de la siguiente manera: si se tensa una cuerda y se hace vibrar, emite un sonido de un tono; si se hace vibrar la mitad de la cuerda, el tono aumentará una octava; si se hace vibrar dos tercios de la cuerda, el tono estará un quinto por encima del que produjo la cuerda entera; si se hace vibrar tres cuartos de la cuerda original, el tono estará una cuarta por arriba del tono original.

Al vibrar, una cuerda o cualquier otro cuerpo transmite su vibración al aire que la rodea; estas alteraciones en la densidad del aire se propagan en forma de ondas y cuando llegan a nuestros oídos, las percibimos como sonido.

Una de las características más importantes de una onda sonora es su frecuencia, que se mide en hertzios —un hertzio equivale a una oscilación por segundo—. El oído de un niño sano percibe sonidos de 12 a 20 000 hertzios pero, al envejecer, este rango disminuye, especialmente para la percepción de sonidos agudos.

Para darnos una idea diremos que un bajo, es decir, la voz humana más grave, canta en el rango de 80 a 300 hertzios y una soprano, con voz muy aguda, alcanza un rango entre 220 a 1 000 hertzios.

Las notas musicales son sonidos puros en los que únicamente está presente una frecuencia. La relación que existe entre la frecuencia de distintos sonidos es muy importante y se muestra en la siguiente tabla:

Do	Re	Mi	Fa	Sol	La	Si	Do
1	$\frac{9}{8}$	$\frac{5}{4}$	$\frac{4}{3}$	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{15}{8}$	2

Figura 1.8 Frecuencia relativa de los intervalos a partir del do.

La razón de las frecuencias se llama intervalo o distancia y es el cociente entre las frecuencias. Por ejemplo, el intervalo entre fa y la es $\frac{4}{3} \div \frac{5}{3} = \frac{4}{5}$, igual al intervalo entre do y mi.

Si silbamos con los labios, el sonido es casi de una sola frecuencia, pero si hablamos o se toca un instrumento se producen varias frecuencias a la vez. Usualmente hay una frecuencia predominante, la principal, y otras secundarias de menor intensidad que tienen cierta relación matemática con la frecuencia principal y se llaman armónicos. La intensidad con la cual se emiten los armónicos hace que los instrumentos musicales suenen diferentes entre sí.

Aunque los pitagóricos nunca hablaron de armónicos, determinaron que las cuerdas de longitudes con razones 1 : 2 y 2 : 3 producían combinaciones de sonidos muy agradables y, a partir de estas proporciones, construyeron una escala musical.

Los pitagóricos fueron los primeros en establecer la música como una disciplina matemática, una de las siete fundamentales que los jóvenes tenían que estudiar en la escuela. Las relaciones que se establecieron hace más de 2 500 años entre las matemáticas y la música están vigentes y siguen presentes en cualquier sala de conciertos. Una vez que un descubri-

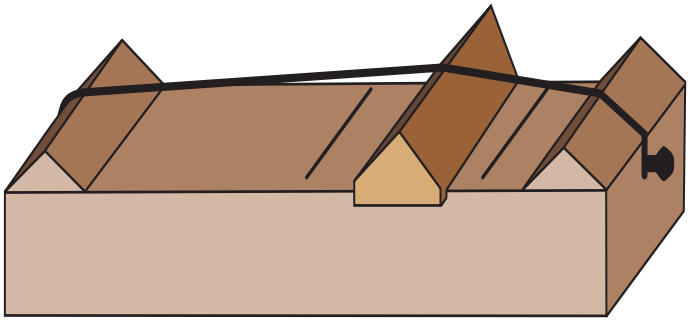


Figura 1.9 Los pitagóricos usaban un instrumento, el monocordio, para producir una escala todavía usada en la actualidad y llamada pitagórica diatónica.

miento matemático se ha establecido, queda para siempre en la vida cotidiana. ¡Que no nos extrañe que los violines se afinen hoy día como los habrían afinado los miembros de la escuela de Pitágoras!

1.2.2 La Pirámide de los Nichos

En marzo de 1785, don Diego Ruiz, recolector de impuestos del gobierno español, buscaba sembradíos clandestinos de tabaco en la selva tropical a las afueras de Papantla, Veracruz. No sabemos si halló lo que buscaba, pero lo que sí encontró fue una pirámide de grandes dimensiones que hoy forma parte del sitio arqueológico conocido como El Tajín. Hasta 1938 empezó la limpieza y reconstrucción del lugar, formado por varios edificios y canchas de juegos de pelota. La peculiaridad del sitio es una pirámide que tiene, en cada uno de sus costados, numerosos nichos a partir de los cuales toma su nombre. Se dice que tiene 365 nichos, tantos como los días del año, por lo que se piensa que la pirámide cumplió alguna función similar a la de un calendario.

Figura 1.10 La foto muestra la pirámide por el costado este, el único que tiene una rampa construida encima de la estructura antigua. A causa de esta rampa, el conteo de los nichos no ha sido fácil, pero varios arqueólogos afirman que la pirámide nunca tuvo 365 nichos | © Latin Stock México.

¿Cómo se pueden acomodar 365 nichos en una pirámide? Esa pregunta se la hicieron con toda seguridad los totonacas antes de empezar con la construcción de la pirámide.

Aunque acomodar nichos en una pirámide no sea un problema de nuestra vida cotidiana, sí lo fue, al menos, para los sacerdotes y sabios encargados de aquella construcción. Si bien las matemáticas del bachillerato son muy pocas veces útiles de manera directa en la vida cotidiana, nos permiten asomarnos de una forma distinta a nuestra cultura: mediante una comprensión más profunda de la tecnología que nos rodea, nos hacen partícipes de los logros de la humanidad.

Volviendo al problema de acomodar nichos, primero hay que observar que 365 no es divisible entre 4, es decir, no es posible acomodar el mismo número de nichos en cada costado tal que en total sumen 365. Pero el número 365 es divisible entre 5: $365 = 5 \cdot 73$. Entonces habría que construir una pirámide con cinco lados, y de ese tipo no hay en México; con seguridad, esto no les hubiera parecido una buena idea.

Si se coloca un nicho en la punta quedan 364 nichos para ser acomodados en los cuatro costados de una pirámide auténtica, como la de El Tajín. Dado que $364 = 4 \cdot 91$, habría 91 nichos en cada lado. La distribución de los nichos en cada lado de la pirámide se puede hacer de muchas maneras; por ejemplo, se podrían acomodar 40 nichos en el nivel más bajo, 30 en el que sigue y 21 en el tercero y último, pero quedaría una pirámide muy ancha con sólo 3 niveles. Además, pasaría que del nivel más ancho al del medio se disminuye el número de nichos en 10, mientras que del medio al último se disminuye en 9. Esto ocasionaría que los nichos en el último nivel tendrían que ser más pequeños que el resto o que la pirámide no tuviera un borde regular. Dicho lo anterior, sería mejor que de nivel en nivel siempre disminuyera el mismo número de nichos. Por ejemplo, podríamos pensar en la sucesión $3, 5, 7, 9, 11, 13, 15, 17$ para el número de nichos en cada nivel. Esto nos daría una pirámide regular con $3 + 5 + 7 + 9 + 11 + 13 + 15 + 17 = 80$ nichos. Si agregamos otro nivel, éste tendrá 19 nichos y se excedería de los 91 que deben ser. Al parecer, la distribución de 365 nichos en una pirámide no es un problema tan sencillo.

Las sucesiones de este estilo se llaman **sucesiones aritméticas** y han sido estudiadas. Entre las más famosas está, por ejemplo, aquella que empieza con uno y aumenta siempre de uno en uno: $1, 2, 3, 4, 5, 6, 7, 8$. Se cuenta que un maestro de matemáticas de Alemania pidió a sus alumnos calcular la suma de los primeros 100 números para tenerlos ocupados por un buen rato; uno de ellos entregó su respuesta después de un minuto, mientras los demás necesitaron casi la hora completa. La gran sorpresa fue que la primera respuesta era de las pocas correctas, resuelta por Carl Friedrich Gauss, un joven que el mundo reconocería después como uno de los matemáticos más influyentes de todos los tiempos.

No es que Gauss calculara muy rápido, sino que usó un truco que él inventó: si se suman los números en pares, tomándolos de los extremos así:

$$(100 + 1) + (99 + 2) + (98 + 3) + \ldots + (53 + 48) + (52 + 49) + (51 + 50),$$

entonces, cada par suma 101 y hay 50 pares. Así que el resultado es:

$$101 \cdot 50 = 5\,050.$$

Figura 1.11 Carl Friedrich Gauss (1777-1855) | © Latin Stock México.

Los números siguientes:

$$\begin{aligned} 1 &= 1 \\ 1+2 &= 3 \\ 1+2+3 &= 6 \\ 1+2+3+4 &= 10 \\ 1+2+3+4+5 &= 15 \\ 1+2+3+4+5+6 &= 21 \end{aligned}$$

se llaman **números triangulares**. También el 91 es triangular:

$$1+2+3+4+5+6+7+8+9+10+11+12+13+14 = 91.$$

Pero una pirámide con los nichos acomodados de esta manera tendría 14 niveles y sería muy empinada, pues se vería así:

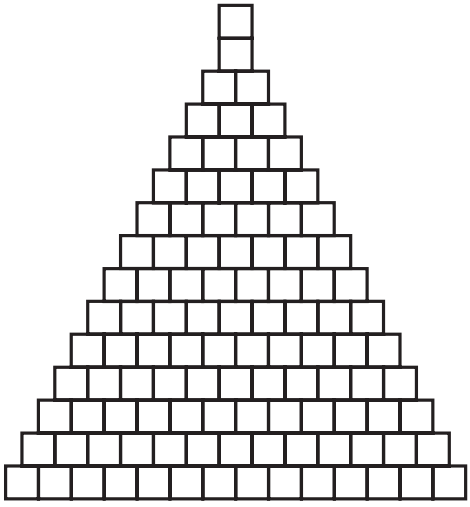


Figura 1.12 Pirámide de 91 nichos en 14 niveles.

No olvidemos que el último nicho se añadió en la cima —el nicho que se quitó de 365 para obtener $364 = 4 \cdot 91$—. Para encontrar una solución mejor, quizá conviene no tratar de trabajar con ejemplos específicos sino con todos los ejemplos a la vez.

Las sucesiones aritméticas no siempre empiezan con el número 1. Si se usa una variable, se puede decir que la sucesión empieza con a. Luego, en vez de aumentar por 1, se puede decir que aumenta siempre en b, otra variable. Y se puede usar una tercera variable para denotar el número de niveles: n. La sucesión, entonces, empieza así:

$$a,\ a+b,\ a+b+b$$

donde el tercer número de nuestra sucesión es $a+2b$, el cuarto es $a+3b$, el quinto es $a+4b$ y el enésimo es $a+(n-1)b$. La sucesión quedaría así:

$$a,\ a+b,\ a+2b,\ a+3b,\ a+4b, \ldots, a+(n-2)b,\ a+(n-1)b.$$

Con el argumento del joven Gauss, se puede calcular la suma de estos números y se obtiene la siguiente fórmula:

$$\frac{(2a+(n-1)b)\cdot n}{2}.$$

El problema consiste en encontrar números a, b y n tales que $\frac{(2a+(n-1)b)\cdot n}{2} = 91$. Si se multiplican ambos lados de esta ecuación por 2 se obtiene que:

$$(2a+(n-1)b)\cdot n = 182.$$

El lado izquierdo de esta ecuación es un producto y uno de sus factores es n. Por lo tanto el lado derecho debe ser divisible entre n. Como $182 = 2 \cdot 91 = 2 \cdot 7 \cdot 13$, se pueden probar diferentes opciones para el valor de n. Si se busca una pirámide con no demasiados niveles hay solamente una opción para asignar un valor a n y ésa es $n = 7$. Entonces $n-1 = 6$ y se puede sustituir en la ecuación original, que queda así:

$$(2a+6b)\cdot 7 = 26 \cdot 7.$$

Entonces $2a+6b$ debe ser igual a 26. Nuevamente hay varias posibles soluciones —que son $a=1, b=4$; $a=4, b=3$; $a=7, b=2$ y $a=10, b=1$—, pero la que más se acerca a nuestro problema es $a=4$ y $b=3$. La sucesión que buscamos es:

$$4,\ 7,\ 10,\ 13,\ 16,\ 19,\ 22$$

y la pirámide tiene entonces la siguiente forma:

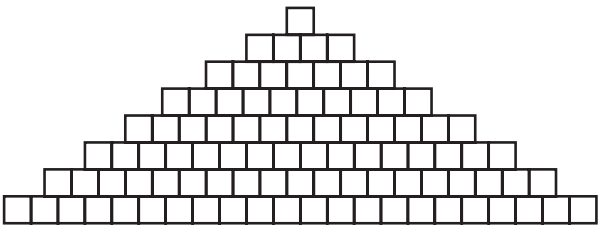


Figura 1.13 Pirámide de 91 nichos en 8 niveles.

Es interesante comparar esta solución con la idea original de los totonacas antes de que hubieran añadido la rampa. La siguiente imagen muestra otro lado de la pidrámide que, además, es el mejor conservado.

Figura 1.14 Otra vista de la Pirámide de los Nichos | © Age Fotostock. Archivo Digital.

Se observa sólo una diferencia con nuestra sucesión: el nivel superior tiene un nicho más. No sabemos cómo llegaron los totonacas a la solución, tal vez simplemente por ensayo y error, al intentar una y otra vez. El análisis que se llevó a cabo aquí requiere de ciertos conceptos y estrategias: necesita generalizar, atacar todas las opciones a la vez al usar **variables**. Esto es un proceso común en matemáticas. George Pólya, un matemático húngaro, decía que a veces es más fácil resolver muchos problemas en conjunto que uno solo a la vez porque al ver la generalidad se puede entender mejor. De esta manera es claro que la matemática ayuda a resolver problemas de nuestra cultura, algunos de importancia crucial, como en este caso.

1.2.3 ¿Cómo se puede medir la profundidad de un pozo?

Un modelo matemático son una o varias fórmulas o ecuaciones, escritas en lenguaje matemático, que describen un proceso o un sistema de procesos. Los modelos matemáticos se usan para predecir o simular o, a veces, sólo para describir lo que pasa en la realidad. Hoy en día, todas las ciencias, incluso las sociales, modelan sus problemas con matemáticas. Por ejemplo, por medio de un modelo matemático se puede tratar de predecir cómo va a mutar un virus, cómo va a cambiar el clima en una región, se puede determinar la altura de un puente e incluso describir algún fenómeno económico y social.

Aquí presentamos un problema muy sencillo pero a la vez muy interesante: cómo medir la profundidad de un pozo. En realidad la idea fundamental no es que el lector entusiasmado vaya midiendo pozos por todos lados, sino que vea cómo, en principio, se hace un modelo matemático.

Para medir la profundidad de un pozo se pueden usar distintas estrategias. Una forma podría ser echar al pozo una cuerda con una piedra amarrada a un extremo. Después de oír que la piedra toca el agua, habría que sacar la cuerda y medirla. No parece ser un método muy práctico porque se necesita una cuerda muy larga. Además, el pozo puede ser profundo y oscuro, de manera que no es posible saber cuándo la cuerda llega al fondo.

Otra posibilidad es usar un cronómetro. Se puede medir el tiempo que transcurre entre que se deja caer la piedra y se escucha el sonido de la piedra al chocar en el agua. ¿Cómo se

relaciona el tiempo que tarda la piedra en llegar al agua con la profundidad del pozo? En otras palabras, ¿cómo relacionar tiempo con distancia? Es claro que cuanto más tiempo transcurre, más profundo es el pozo, pero ¿cómo se puede obtener una medida de profundidad a partir de una medida de tiempo?

Figura 1.15 | © Latin Stock México.

Cuando se suelta una piedra en un pozo, como sólo actúa sobre ella la fuerza de gravedad, decimos que cae en caída libre. La fórmula de caída libre es:

$$d = \frac{1}{2}gt^2.$$

d es la distancia vertical, en este caso la profundidad del pozo.
$g = 9.81\frac{m}{s^2}$ es la constante gravitacional, esto es, la aceleración con la que cae la piedra.
t es el tiempo; que en este caso es el que transcurre entre que se suelta la piedra y ésta golpea el fondo del pozo.

Por ejemplo, en una caída de 3 segundos $(3\ \text{s})$, la fórmula nos daría una profundidad de:

$$d = \frac{1}{2}g\,(3\ \text{s})^2 = \frac{1}{2}\left(9.81\frac{\text{m}}{\text{s}^2}\right)(9\ \text{s})^2 = 44.145\text{m}.$$

Es decir, un poco más de 44 metros. Éste es un primer modelo, el más sencillo, pero no el más exacto.

Como el pozo es oscuro, no se puede ver cuándo la piedra llega al fondo, solamente se escucha. Tomemos en cuenta, además, que el sonido tarda un cierto tiempo en llegar a nosotros; hay que buscar una manera de medir el tiempo que pasa desde que la piedra choca en el agua y el sonido sube por el pozo y llega a nosotros.

Como la velocidad del sonido (v) es: $v = 340\frac{\text{m}}{\text{s}^2}$, el tiempo que el sonido tarda en subir por el pozo puede calcularse con la fórmula:

$$v = \frac{d}{t}$$

de la que puede despejarse la variable t:

$$t = \frac{d}{v}.$$

v es la velocidad; en este caso la del sonido: $340\frac{\text{m}}{\text{s}^2}$.
d es la distancia; en este caso es la que el sonido recorre hasta llegar a nosotros, es decir, la profundidad del pozo.
t es el tiempo; en este caso, en que el sonido tarda en subir por el pozo hasta llegar al oído del observador que escucha.

¡Cuidado! Aquí se ha vuelto a usar la variable t para denotar al tiempo, aunque ahora se trata de un tiempo diferente. En la caída, la variable t representaba el tiempo que tarda la piedra en caer y ahora representa el tiempo que tarda el sonido en subir desde el fondo hasta el escucha. Esto es algo que ocurre con frecuencia en matemáticas: el significado de una variable depende del contexto en el cual se usa. Para evitar confusiones, distinguiremos estos dos tiempos con un subíndice: $t_{\text{caída}}$ para el tiempo de caída y t_{sonido} para el del sonido.

Si el pozo realmente tuviera 44.1 metros de profundidad el sonido tardaría 0.13 segundos en recorrerlo. No es mucho, pero un segundo modelo del problema podría ser restarle este tiempo a los tres segundos que tardamos en escuchar la caída, lo que dejaría $3 - 0.13 = 2.87$ segundos para la caída libre, lo cual resultaría en:

$$d = \frac{1}{2}g\,(2.87\text{ s})^2 = \frac{1}{2}\left(9.81\frac{\text{m}}{\text{s}^2}\right)\left(8.2369\text{ s}^2\right) = 40.4019945\text{ m}.$$

Esta segunda aproximación es más exacta, pero no del todo correcta, pues tomamos el tiempo del sonido para una profundidad que no es la que obtenemos. El problema radica en que no se conoce ni el tiempo de la caída ni el tiempo que tarda en subir el sonido por separado, lo único que se conoce es que el tiempo (t) desde que se suelta la piedra hasta que se escucha el golpe es la suma de ambos:

$$t = t_{\text{caída}} + t_{\text{sonido}} \qquad (1)$$

Las otras fórmulas importantes, que relacionan la distancia con los tiempos, son:

$$d = \frac{1}{2}gt_{\text{caída}}^2 \qquad (2)$$

$$\frac{d}{v} = t_{\text{sonido}} \qquad (3)$$

Lo que se busca es una fórmula que relacione la distancia d directamente con el tiempo total t . Para obtenerla se requiere cierta comodidad con el manejo del álgebra. El lector que no esté familiarizado con ella puede simplemente observar esto como un ejemplo particular. De la ecuación (1) se obtiene que:

$$t_{\text{sonido}} = t - t_{\text{caída}}$$

Ahora se puede sustituir t_{sonido} en la ecuación (3) por la expresión $t - t_{\text{caída}}$, lo que nos da una nueva ecuación:

$$t - t_{\text{caída}} = \frac{d}{v}.$$

De esta ecuación se puede despejar la variable $t_{\text{caída}}$ y se obtiene $t_{\text{caída}} = t - \frac{d}{v}$. Sustituyendo lo anterior en la ecuación (2):

$$d = \frac{1}{2}g\left(t - \frac{d}{v}\right)^2.$$

d es la profundidad del pozo.
$g = 9.81\frac{m}{s^2}$ es la constante gravitatoria.
$v = 340\frac{\text{m}}{\text{s}^2}$ es la velocidad del sonido en el aire.
t es la suma de los tiempos $t_{\text{caída}}$ y t_{sonido}.

Ahora sí, obtuvimos una fórmula general y válida para cualquier pozo. Un tercer y último modelo para nuestro problema original que, admitimos, es una ecuación más complicada que la del primer modelo, donde únicamente se tomó en cuenta el tiempo de caída. Una ecuación compleja es un fenómeno común: entre mejor sea el modelo y más preciso, más complicadas serán las matemáticas involucradas. Así, tenemos que para un tiempo de $t = 3$ segundos se obtiene una profundidad $d = 40.7$ metros.

Resolvamos de postre otro problema relacionado con el anterior. Si en el pozo decidiéramos gritar, ¿cuánto tiempo pasaría hasta que se escuchara el eco desde el fondo del pozo?

Partamos de que, en cualquier pozo, el sonido tarda el mismo tiempo en bajar que en subir. Como $t = \frac{d}{v}$, entonces el sonido tarda $2\frac{d}{v}$, donde d es la profundidad del pozo que equivale a $d = 26.16942$ metros y $v = 340\frac{\text{m}}{\text{s}^2}$ es la velocidad del sonido en el aire.

Entonces, $t = \frac{2(26.16942)}{340} = 0.1539$ segundos. ¡Y ya acabamos!

1.3 LAS MATEMÁTICAS DE LA NATURALEZA

Figura 1.16 Grabado de Ernst Haeckel, biólogo, filósofo, físico y artista alemán que retrató y nombró a miles de especies. En este grabado se aprecia la belleza matemática de las conchas marinas.

En la naturaleza hay fenómenos muy diversos y complejos. Resulta asombroso que el ser humano haya podido desentrañar el mecanismo de algunos de ellos y, además, expresarlos con fórmulas matemáticas muy sencillas. En esta sección veremos ejemplos de fórmulas simples que describen el comportamiento de algunos de los fenómenos más importantes de la naturaleza.

Galileo Galilei, un científico italiano del siglo XVI, descubrió, entre otras cosas, el tipo de trayectoria que sigue una piedra al ser lanzada al aire. Para ello, empleó la matemática situándola en el lugar especial dentro de las ciencias que ocupa todavía en la actualidad. En sus

libros, escritos como diálogos entre tres personajes, describe y desarrolla sus ideas sobre cómo puede entenderse la naturaleza. Discute fenómenos como la balanza, la caída libre de los cuerpos y el péndulo. La siguiente cita aparece en su obra maestra *Il saggiatore* y es de las más importantes y famosas en la historia de la ciencia, pues dice que:

> Sr. Sarsi, las cosas no son así. La filosofía está escrita en ese grandísimo libro que tenemos abierto ante los ojos, quiero decir, el Universo, pero no se puede entender si antes no se aprende a entender la lengua, a conocer los caracteres con los que está escrito. Está escrito en lengua matemática y sus caracteres son triángulos, círculos y otras figuras geométricas, sin las cuales es imposible entender ni una palabra; sin ellos es como girar vanamente en un oscuro laberinto.[1]

Hay que aclarar que, en aquel entonces, *filosofía* abarcaba todos los conocimientos, esto es, representaba lo que hoy llamamos *ciencia*. Si se acepta la visión de Galilei, entonces para comprender el Universo son necesarias las matemáticas. Afortunadamente no es necesario estudiar matemáticas avanzadas para poder entender los asuntos más importantes de la naturaleza. Incluso a veces, las matemáticas requeridas para comprender algunas cosas de la naturaleza y su comportamiento son extremadamente simples, como en el caso de la esfera y sus aplicaciones que se muestra continuación.

1.3.1 La esfera

La esfera es una de las figuras geométricas que podemos contemplar frecuentemente a nuestro alrededor.

Figura 1.17 Algunos objetos esféricos y circulares | © Latin Stock México.

Los objetos de esta ilustración no son esféricos por pura casualidad. La pelota de tenis tiene que ser esférica para que rebote de manera completamente predecible. El balón de futbol americano, al no ser redondo, rebota en el piso de forma impredecible. Sería imposible jugar tenis con una pelota con la forma de un balón de futbol americano. La burbuja de jabón es esférica como consecuencia de la presión uniforme que ejerce sobre sus

[1] Del original (1979, p. 38): "*La filosofia è scritta in questo grandissimo libro che continuamente ci sta aperto innanzi a gli occhi (io dico l'Universo), ma non si può intendere se prima non s'impara a intender la lingua, e conoscer i caratteri, ne' quali è scritto. Egli è scritto in lingua matematica, e i caratteri son triangoli, cerchi, ed altre figure geometriche, senza i quali mezi è impossibile a intenderne umanamente parola; senza questi è un aggirarsi vanamente per un oscuro laberinto.*"

paredes el aire que contiene y por la resistencia también uniforme del material de que está hecha.

El ojo es esférico para que podamos mirar en muchas direcciones sin tener que voltear la cabeza. La esfera es el único cuerpo que puede girar libremente en todas direcciones aunque esté apoyado en varias partes. Este hecho se aprovecha también en la cabeza del fémur, el hueso humano de mayor tamaño, y también en los balines de los rodamientos o baleros.

Debido a la función que desempeñan en la naturaleza, al uso para el que fueron diseñados o a las propiedades de los materiales que los forman, muchos objetos tienen forma esférica. De este modo, la esfera es una forma geométrica, y por lo tanto, matemática, que observamos con frecuencia en el Universo. Es un ente abstracto, un concepto, puesto que muchos objetos de la naturaleza pueden describirse como una esfera. La esfera es uno de esos "caracteres" matemáticos de los que habla Galileo que describen el Universo.

1.3.2 La balanza

Si bien la balanza es una máquina creada por el hombre —que sirve para medir y comparar pesos—, su comportamiento depende de las leyes naturales al aprovechar la gravedad; dicho de otro modo, la balanza nos muestra cómo funciona la gravedad. Por lo anterior, al describir el funcionamiento de la balanza habremos descrito una pequeña parte de la naturaleza.

La balanza más sencilla es una tabla equilibrada sobre un rodillo, como se muestra en la figura 1.3.3.

Supongamos que la tabla y el rodillo son muy ligeros, pero firmes. La balanza mantiene

Figura 1.18 Balanza sencilla.

su estado de equilibrio si ponemos sobre la tabla dos pesas iguales a la misma distancia en ambos lados del rodillo, por ejemplo, de un kilogramo a un metro de distancia cada una:

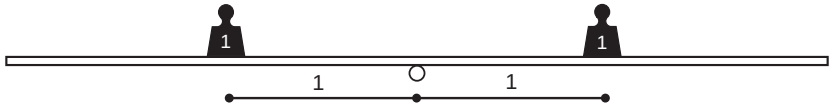


Figura 1.19 Balanza equilibrada.

Hay una ley que describe cómo cambia el peso al cambiar la distancia. Sabemos por nuestra experiencia que cuanto más lejos, menor peso se requiere. Esto es lo que nos dice la intuición, pero nos gustaría encontrar una ley precisa que nos diga qué peso, cuántos gramos habría que poner a dos metros del rodillo para equilibrar el peso de un kilogramo que está a un metro del rodillo. El siguiente razonamiento muestra que tiene que ser medio kilo. Podemos imaginarnos el kilogramo dividido en dos pesas iguales puestas sobre una balanza que, a su vez, reposa en el lugar en donde antes estaba la pesa de un kilo:

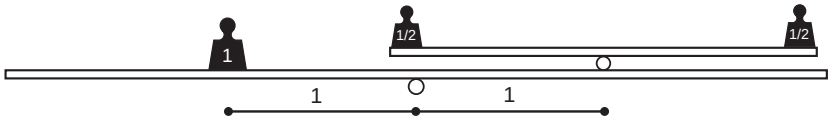


Figura 1.20 La balanza al sustituir el peso derecho por una balanza pequeña.

Como la balanza pequeña está en equilibrio podemos quitar el rodillo y la tabla de arriba —por hipótesis, su peso es despreciable— y apoyar las pesas directamente sobre la tabla grande sin que se altere el equilibrio. Finalmente, la pesa de medio kilo que se apoya justo sobre el rodillo central también puede quitarse.

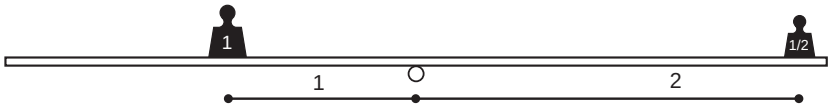


Figura 1.21 Vemos que al quitar la pesa apoyada en el rodillo central, no se influye en el equilibrio de la balanza.

Con ello descubrimos la ley básica de la balanza: al doble de distancia hay que poner la mitad del peso.

La ley general de la balanza se obtiene de este caso al usar consecutivamente el mismo procedimiento de sustituir las pesas por balanzas equilibradas. Por ejemplo, remplacemos ahora la pesa de un kilo del lado izquierdo por una nueva balanza equilibrada con pesas de $\frac{2}{3}$ y $\frac{1}{3}$ de kilo, la primera colocada medio metro a la izquierda y la segunda un metro a la derecha del lugar que ocupaba la pesa de un kilo, como muestra la siguiente imagen:

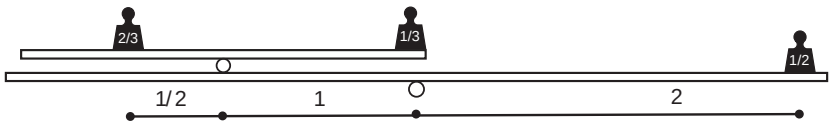


Figura 1.22 La balanza después de sustituir el peso izquierdo por una balanza pequeña.

Al volver a eliminar el rodillo y la tabla que sostiene a estas dos pesas, así como también la que queda sobre el rodillo de la balanza, obtenemos la siguiente balanza en equilibrio:

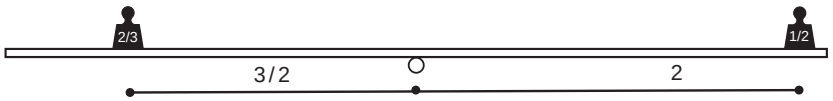


Figura 1.23 La balanza al eliminar el rodillo y la tabla de la balanza pequeña, así como también la pesa del medio.

El lector atento se dará cuenta de que existe una ley detrás de estos números: si la distancia se duplica, el peso se divide a la mitad y si la distancia es $\frac{3}{2}$ de metro entonces el peso es de $\frac{2}{3}$ kilogramo. Podemos escribir esto como una igualdad:

$$\frac{2}{3}kg \cdot \frac{3}{2}m = \frac{1}{2}kg \cdot 2m.$$

De esta manera podríamos argumentar para otros pesos y otras distancias. Y siempre encontraríamos que, cuando hay equilibrio, el producto del peso por la distancia al rodillo es igual en ambos lados. La fórmula matemática:

$$p_1 d_1 = p_2 d_2$$

expresa esta ley de manera muy sencilla: nos dice que el producto del primer peso p_1 por su distancia d_1, al rodillo, es igual al producto del segundo peso p_2 multiplicado por su distancia d_2 hasta el rodillo. En particular, la fórmula se satisface en el caso anterior pues:

$$p_1 = \frac{1}{2} \text{ y } d_1 = 2 \text{ y } p_2 = \frac{2}{3},\ d_2 = \frac{3}{2}.$$

El lenguaje matemático es a la vez muy condensado y muy expresivo. En una sola fórmula se expresan todas las situaciones de equilibrio de una balanza. La fórmula cubre una infinidad de casos.

El ojo no entrenado no puede entender a primera vista la generalidad y la importancia de esta fórmula, al igual que el ojo de una persona no entrenada en la escritura musical no puede entender una partitura y el que no conoce las letras y sus combinaciones para un sistema lingüístico no puede entender algo escrito. Sin embargo, un poco de entrenamiento matemático permite comprender el significado de los símbolos que aparecen en la fórmula, la relación que expresa entre las cantidades representadas por los símbolos. Como queda claro, unas pocas matemáticas nos llevan muy lejos.

Además de la sencillez de la fórmula obtenida, resulta sorprendente que el razonamiento que nos llevó a descubrirla no requirió de experimentos con objetos reales. Fue suficiente hacerlos mentalmente y argumentar de manera lógica. No siempre es posible descubrir las leyes de la naturaleza por medio de experimentos mentales, sin embargo, estos últimos se utilizan muchas veces para acercarnos a ellas.

Mientras el caso de la esfera nos muestra un acercamiento directo a la naturaleza, el de la balanza nos revela una ley más profunda que se expresa mediante una fórmula matemática que llegamos a descubrir mediante razonamientos lógicos.

A continuación, presentaremos el caso de la gravedad que nos lleva a un tipo de fenómenos aún más complejos, en los que no basta la razón para descubrir las leyes naturales sino que hacen falta, además, experimentos que den información numérica que no podríamos obtener por pura lógica, pero en los que, las leyes que los rigen, siguen pudiéndose expresar con fórmulas matemáticas sencillas.

1.3.3 La gravedad

En la sección anterior se llegó a una fórmula matemática para expresar una ley de la naturaleza: la de la balanza. Con ello queda ilustrado que las matemáticas pueden ser útiles para describir fenómenos naturales sencillos. Existen fenómenos mucho más complejos en los cuales las matemáticas no sólo son útiles sino verdaderamente imprescindibles. Uno de ellos es el de la gravedad. Comenzaremos por su aspecto más simple, la caída de los cuerpos.

Sabemos por nuestra experiencia diaria que los cuerpos caen. Si soltamos una taza ésta va a dar al suelo y probablemente se rompa. Desde épocas muy remotas se sabe lo anterior, pero no fue sino hasta 1590 cuando Galileo Galilei describió en términos matemáticos precisos el fenómeno de la **caída libre** —cuyo nombre expresa que nada se interpone en ella— de los cuerpos. Galileo demostró mediante la observación, la experimentación y el razonamiento que, cuando cae un cuerpo, su velocidad aumenta constantemente. Así, la velocidad del cuerpo que cae es proporcional al tiempo que ha transcurrido desde que comenzó a caer. La constante de proporcionalidad se llama la aceleración de la gravedad, se

denota por g, y hoy sabemos que su valor es de $9.81\frac{m}{s^2}$. Esto quiere decir que la velocidad de un cuerpo en caída libre aumenta en $9.81\frac{m}{s^2}$ cada segundo. En fórmula, lo anterior se escribe así:

$$v = gt$$

donde v es la velocidad del cuerpo, t es el tiempo transcurrido desde que comenzó su caída y g es la constante de la aceleración de la gravedad.

Galileo demostró que, como consecuencia de esta ley de movimiento según la cual la velocidad aumenta constantemente con el tiempo, la distancia recorrida por el cuerpo en su caída aumenta de manera cuadrática con respecto al tiempo, es decir:

$$d = \frac{1}{2}gt^2$$

donde d es la distancia que un cuerpo ha recorrido en su caída, t es el tiempo que ha transcurrido desde que comenzó a caer y g es la constante de la aceleración de la gravedad.

El siguiente esquema muestra una pelota que cae de un edificio de 80 metros y una gráfica con las posiciones de la pelota cada dos décimas de segundo.

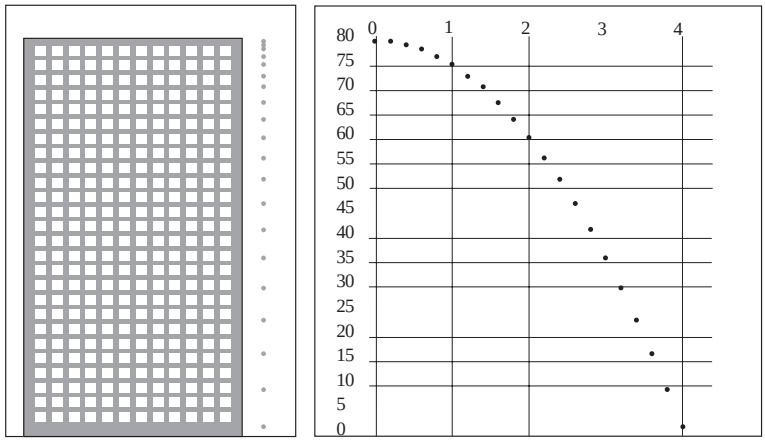


Figura 1.24 A la izquierda se presentan las distintas posiciones de la pelota cada cierta fracción de tiempo ($\frac{1}{4}$ de segundo). A la derecha, se observa la gráfica de la altura contra el tiempo.

La forma que une los puntos de la gráfica es una curva conocida ya por los antiguos griegos a la que llamaron **parábola**. En la gráfica se observa que la pelota cae aproximadamente 5 metros durante el primer segundo, el triple de 5 durante el siguiente segundo, cinco veces esa misma cantidad durante el tercer segundo y siete veces la cantidad durante el cuarto segundo. Es decir, la distancia recorrida sigue la sucesión $1, 3, 5, 7\ldots$ de los números impares. Por tanto, la distancia total que ha caído en cada segundo es la suma de los primeros impares: $1,\ 1+3,\ 1+3+5,\ 1+3+5+7\ldots$ que son los cuadrados de los números enteros $1,\ 4,\ 9,\ 16$, etc. Esta observación podemos escribirla como:

$$d = 4.9 \cdot t^2$$

que corresponde a la fórmula de Galileo $d = \frac{1}{2}gt^2$ ya que $\frac{1}{2}g \cong 4.9$.

Por ejemplo, la distancia que recorre un cuerpo en caída libre en el doble de un lapso de tiempo es cuatro veces la que recorrió en el lapso original. La distancia recorrida en el triple de un lapso de tiempo es igual a nueve veces la distancia original. Esta fórmula representa la ley del movimiento de caída libre. No es una ley lineal sino cuadrática: la velocidad depende linealmente del tiempo, pero la distancia recorrida es proporcional al cuadrado del tiempo.

No es una ley complicada, pero podríamos considerar que es ligeramente más compleja que la que describe a la balanza, pues en aquélla no aparecen cantidades elevadas al cuadrado y en ésta sí.

Galileo demostró también que la trayectoria descrita por un objeto lanzado al aire es una *parábola*, esa curva que habían estudiado dos mil años antes los griegos por razones enteramente diferentes; de hecho, la usaron para intentar resolver el problema —meramente matemático— de la duplicación del cubo, que consistía en encontrar el lado de un cubo que tuviera el doble del volumen de otro cubo dado.

Si regresamos a la parábola como trayectoria de una piedra o bala lanzada, lo sorprendente es que Galilei obtuvo esta descripción no de la observación directa sino mediante una argumentación lógica. Para ello tuvo que introducir la relatividad del movimiento. Su ejemplo es famoso: si dejamos caer una piedra desde una torre, ésta cae por una trayectoria vertical al piso. Pero si dejamos caer la piedra desde un mástil de un barco que navega en una dirección, ¿dónde caerá la piedra?, ¿junto al mástil o detrás de éste porque cae de forma vertical y mientras tanto el barco avanza? El hecho es que cae junto al mástil. La explicación de Galileo es que el marinero que deja caer la piedra cree que cae vertical porque él también se mueve, pero para un espectador en la playa la piedra no sólo cae, sino que al soltarla se impulsa en la dirección del barco. De esta manera, Galileo descubre que las velocidades pueden descomponerse en una componente vertical y otra horizontal.

El movimiento de caída libre no parecía tener nada que ver con el movimiento de los planetas, hasta que uno de los más grandes genios de la historia de la ciencia, Isaac Newton, pensó en relacionarlos. El resultado de esas investigaciones constituye uno de los mayores logros científicos, quizá la hazaña intelectual más importante de todos los tiempos. Lo más sorprendente es que está al alcance de cualquier estudiante de bachillerato dispuesto a familiarizarse con la manipulación de algunas expresiones algebraicas no muy complicadas. Para acercarnos al tema es necesario que repasemos los conocimientos que la humanidad tenía del movimiento planetario en la época de Newton.

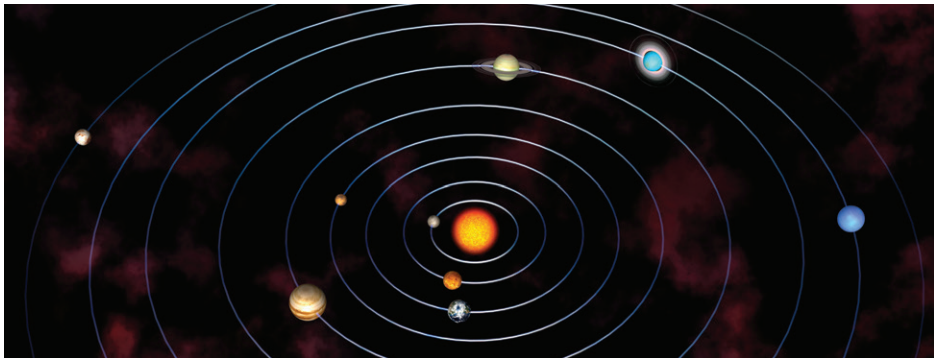
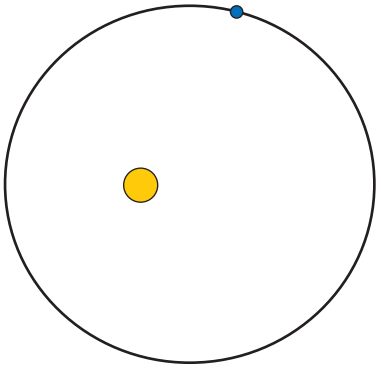


Figura 1.25 Representación de los planetas en órbita alrededor del Sol | © Latin Stock México.

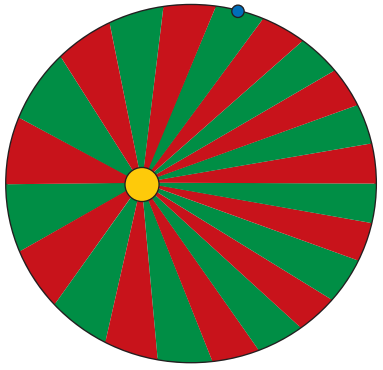
Johannes Kepler fue un contemporáneo de Galileo que dedicó la mayor parte de su vida al estudio de las órbitas de los planetas, es decir, a las trayectorias que describen los planetas alrededor del Sol. Kepler era un hombre profundamente religioso y convencido de que Dios dispuso a los planetas en perfecta armonía describiendo figuras geométricas sencillas. Gracias a las mediciones precisas de Tycho Brahe, después de muchos intentos fallidos con circunferencias y óvalos, le fue posible describir las órbitas celestes mediante la elipse, curva que al igual que la parábola había sido estudiada por los griegos por sus propiedades puramente geométricas. Kepler formuló sus hallazgos principalmente en dos de sus libros: *Astronomia nova* (Nueva astronomía, 1609) y *Harmonices mundi* (La armonía de los mundos, 1619) en los que describe el movimiento de los planetas en tres leyes, llamadas las leyes de Kepler del movimiento planetario.

Leyes de Kepler del movimiento planetario

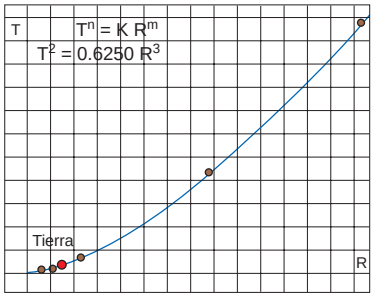
1. *Los planetas giran alrededor del Sol como órbitas elípticas, donde el Sol es uno de los focos.*



2. *Los radios-vectores de los planetas respecto al Sol barren áreas iguales en tiempos iguales.*



3. *Los cuadrados de los periodos de revolución de los planetas son proporcionales a los cubos de los radios medios de sus órbitas.*



Conviene hacer algunas aclaraciones sobre estas leyes. La primera, aunque afirma que las órbitas de los planetas son elípticas, en realidad se trata de elipses con una excentricidad muy pequeña, lo que quiere decir que son casi circulares. La segunda nos indica que los planetas se mueven con mayor velocidad cuando están más cerca del Sol que cuando están más lejos. Finalmente, la tercera, que puede resultar muy extraña, no es más que el descubrimiento de que los radios de las órbitas y sus tiempos de revolución no son arbitrarios, sino que guardan una relación funcional que no es lineal ni cuadrática, más bien algo intermedio: el cuadrado del tiempo es proporcional al cubo del radio:

$$\frac{T^2}{R^3} = K$$

donde K es un número fijo para todo el sistema planetario, el mismo para todos los planetas.

Es importante observar que la tercera ley le llevó a Kepler doce años más que las otras para descubrirla y representa una relación muy precisa entre la velocidad con la que giran los planetas y sus distancias al Sol. ¿Cómo es que se da esta relación aparentemente tan extraña? No hay ninguna relación aparente entre los radios de las órbitas y, por ejemplo, los tamaños de los planetas. ¿Qué hace que sí la haya entre los periodos y los radios? ¿Y por qué es a la vez tan simple y tan extraña?

Kepler propuso un modelo matemático del movimiento de los planetas pero no una explicación. Creía que estas leyes provenían de la mano de Dios y no requerían mayor comprensión.

Así quedaron las cosas por varios años hasta que, en 1687, se publicó el libro *Philosophiae Naturalis Principia Mathematica* de Isaac Newton, matemático, físico y astrónomo inglés. En este libro, Newton explica que se pueden deducir las tres leyes de Kepler a partir de leyes más fundamentales y simples.

La idea principal de Newton es que debe haber una fuerza que mantiene a los planetas en órbitas alrededor del Sol y esa misma fuerza debe ser la responsable del movimiento de caída libre de los cuerpos sobre la Tierra.

Antes de llegar a esta idea Newton, al continuar el trabajo de Galileo, había establecido sus tres famosas leyes:

Leyes del movimiento de Newton

1] *Todo cuerpo permanece en estado de reposo o de movimiento rectilíneo uniforme a menos que haya una fuerza externa que lo modifique.*
2] *Cuando una fuerza actúa sobre un cuerpo, éste adquiere una aceleración constante proporcional a la magnitud de la fuerza e inversamente proporcional a la masa del cuerpo.*
3] *A toda acción corresponde una reacción igual y en sentido contrario.*

En símbolos, la segunda ley de Newton se escribe así:

$$a = \frac{F}{m}$$

donde a es la aceleración que recibe el cuerpo, m es su masa y F es la fuerza que actúa sobre él. Para Newton era claro que si los cuerpos caen es debido a que hay una fuerza que los atrae hacia el suelo y esta fuerza debe ser proporcional a la masa del cuerpo ya que todos los cuerpos, independientemente de su masa, caen con la misma aceleración g. La Tierra atrae a los cuerpos sobre su superficie con una fuerza $F = mg$, donde m es la masa del cuerpo, aunque en forma coloquial decimos "el peso" del cuerpo. Con este razonamiento, Newton define el concepto de la masa de un cuerpo y lo relaciona con su peso por medio de la constante g; a la vez, diferencia ambos conceptos al definir a la masa como una propiedad intrínseca del cuerpo, mientras que considera al peso como una propiedad del cuerpo en relación con la Tierra, es decir, representa la fuerza con la que la Tierra lo atrae.

Al razonar con ayuda de la simetría y apoyado en su tercera ley del movimiento, Newton piensa que si la Tierra atrae a los cuerpos que se encuentran cerca de su superficie con una fuerza proporcional a la masa de los mismos, entonces debe haber otra fuerza, igual y en sentido contrario, con la que cada cuerpo atrae a la Tierra y esta fuerza, lógicamente, también debe ser proporcional a la masa de la Tierra. Dicha fuerza debe ser la responsable de la

caída de los cuerpos hacia la Tierra y también de las trayectorias de los planetas que, en lugar de ser líneas rectas, giran alrededor del Sol porque éste los atrae. Ésta es la gran idea de Newton: hay una fuerza de atracción entre todos los cuerpos, que aumenta proporcionalmente con la masa de cada cuerpo.

A partir de las ideas anteriores y por deducción matemática, Newton logró establecer que la fuerza de atracción entre dos cuerpos de masas m y M que se encuentran a una distancia R debe tener la forma:

$$F = G\frac{Mm}{R^2}$$

donde G es una constante universal, es decir, es la misma para todos los cuerpos. Ésta es la famosa ley de la gravitación universal de Newton. Dice que la fuerza de gravedad con la que se atraen dos cuerpos de masas M y m es proporcional al producto de sus masas e inversamente proporcional al cuadrado de la distancia R que los separa. G se llama la **constante de la gravitación universal** y puede calcularse a partir de la aceleración de la gravedad g y el radio r y la masa m de la Tierra. Esta fórmula expresa una ley de la naturaleza de manera compacta y precisa. Las matemáticas son enormemente económicas: con sólo ocho símbolos —cinco letras, un número, una igualdad y una raya de quebrados— se expresa todo este profundo conocimiento. Al final de esta sección hay una deducción de esta fórmula a partir de la tercera ley de Kepler.

Newton logró demostrar que las tres leyes del movimiento planetario de Kepler son meras consecuencias matemáticas de la ley de la gravitación universal y de sus tres leyes básicas del movimiento de los cuerpos. Este hecho constituye una explicación científica completa del movimiento planetario. Una ley universal que explica los movimientos de todos los cuerpos terrestres y celestes de manera unificada y clara. Probablemente no hay otro descubrimiento científico de mayor trascendencia y belleza que éste. El lector interesado en estas deducciones podrá seguirlas en el tema 3, en la sección sobre "Espacio, tiempo y movimiento".

Para llegar al descubrimiento de la ley de la gravitación universal, Galileo, Kepler, Newton y muchos científicos más tuvieron que luchar contra las ideas preconcebidas que imperaban en aquella época, algunas de carácter religioso y otras apoyadas en antiguas tradiciones, pero todas haciendo una apología irracional de la ignorancia y la intolerancia. A nosotros, seres humanos del siglo XXI, nos parece natural que una misma ley de la naturaleza explique fenómenos terrestres y celestes, pero debemos recordar que en aquel entonces se pensaba que lo que pasaba en la Tierra era algo de naturaleza muy distinta a lo que pasaba en el cielo, el cual se identificaba no con lo material sino con lo divino. Según Aristóteles, el cielo y sus ingredientes como el Sol, la Luna, las estrellas y los planetas, estaban compuestos por una materia completamente distinta a las terrenales. Las ideas de Galileo, Kepler y Newton fueron por tanto muy revolucionarias: la atracción de la gravedad actúa de la misma manera en el cielo y en la Tierra. La misma ley explica, por un lado, el movimiento de los planetas alrededor del Sol, el de la Luna alrededor de la Tierra y la caída de los cuerpos sobre la superficie terrestre y predice que, en cualquier astro celeste, debe haber una fuerza que hace caer a los cuerpos que se encuentran sobre su superficie. La atracción gravitatoria actúa en la Tierra y en el cielo, en el sistema solar y entre estrellas y galaxias, siempre de la misma manera.

Parece haber una relación muy íntima entre la naturaleza y las matemáticas, tal como Galileo Galilei lo hiciera notar en la nota que aparece al principio de esta sección y que escribió muchos años antes del descubrimiento de Newton. Quizá esa idea suya inspiró a Newton e inspira a los científicos, cuya misión es estudiar a la naturaleza al tratar de descu-

brir sus leyes y expresarlas, de preferencia, en lenguaje matemático, porque cuando esto se logra se obtiene un conocimiento que es a la vez profundo y sencillo.

Ésta es una de las grandes fuerzas inspiradoras y generadoras de las matemáticas.

Como la ley de la gravitación universal es uno de los grandes logros culturales, en este primer capítulo quisimos explicar la deducción de la fórmula —a pesar de que el álgebra requerida para ello rebasa el nivel de las matemáticas usadas hasta ahora— que se encuentra en el recuadro a continuación. Ya se mencionó que en los libros de matemáticas hay partes más difíciles de leer; ésta es una de ellas y, por consiguiente, se puede omitir, si el lector siente que no entiende, para reanudar la lectura a partir de la siguiente sección.

¿De dónde sale la fórmula $F = G\frac{Mm}{R^2}$? A continuación se presenta una deducción de la ley de la gravitación universal a partir de la tercera ley de Kepler. Esta deducción requiere de algunas manipulaciones algebraicas a las que tal vez el lector no esté acostumbrado.

Del estudio del movimiento circular uniforme se sabe que para que un cuerpo de masa m se mueva con velocidad constante v en una trayectoria circular de radio R es necesario que actúe sobre él una fuerza —llamada fuerza centrípeta— dirigida hacia el centro de la trayectoria y de magnitud:

$$F = m\frac{v^2}{R}. \qquad (1)$$

La demostración de este hecho podrá consultarse en el tema 3, como ya se mencionó. La velocidad de un planeta cuya órbita es aproximadamente circular de radio R y cuyo periodo de revolución es T, debe ser igual al perímetro de la órbita $2\pi R$ entre T, es decir:

$$v = \frac{2\pi R}{T}$$

por lo tanto, sustituyendo esta expresión para la velocidad en la fórmula (1) obtenemos que la fuerza con la que el Sol atrae al planeta debe ser:

$$F = m\frac{(\frac{2\pi R}{T})^2}{R} = m\frac{4\pi^2 R}{T^2}.$$

Esta última fórmula se obtuvo al sustituir v en la primera, por su expresión $\frac{2\pi R}{T}$, y tras realizar algunas operaciones algebraicas como elevar al cuadrado un producto y simplificar fracciones.[1] En la última igualdad utilizamos la tercera ley de Kepler para sustituir T^2 por su equivalente KR^3. Así, obtenemos:

$$F = m\frac{4\pi^2 R}{KR^3} = m\frac{4\pi^2}{KR^2}$$

[1] La sustitución es un método común al trabajar con ecuaciones: éstas expresan igualdades, y si una parte es igual a otra la podemos remplazar sin alterar la ecuación. Las operaciones algebraicas son sólo cambios de forma que no alteran tampoco la igualdad. El lector que no esté aún familiarizado con estos procedimientos tendrá algunas dificultades para entender todos los detalles en una primera lectura, pero poco a poco se familiarizará con ellos hasta llegar a comprenderlos perfectamente, pues en realidad son simples manipulaciones numéricas con números representados por letras.

lo cual nos dice que la fuerza de gravedad debe depender inversamente del cuadrado de la distancia R entre el planeta y el Sol. Si definimos $G = \frac{4\pi^2}{KM}$ donde M es la masa del Sol, podemos escribir esta igualdad como:

$$F = G\frac{Mm}{R^2}.$$

Que es precisamente la fórmula a la que queríamos llegar.

1.4 LAS PROPIAS MATEMÁTICAS

Figura 1.26 Detalle de La escuela de Atenas*, del pintor italiano Rafael, que muestra a Pitágoras, fundador de la primera escuela de matemáticas, que se enfocaba a los números y sus aplicaciones en el arte y la música | © Latin Stock México.*

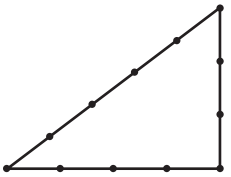
Las matemáticas son una fuente inagotable de preguntas y problemas pues se construyen "a base de" entender, que en cierto sentido, es disipar dudas. Saber dudar, hacerse buenas preguntas, es parte integral del quehacer matemático.

Aunque su materia de trabajo sean ideas abstractas, las matemáticas tienen una enorme solidez basada en su posible comunicación, de una generación a otra, sin alteraciones. Lo que nuestros antepasados entendieron por *"dos"*, independientemente del vocablo que hayan usado para referirse a esta idea, es lo mismo que entendemos ahora. Y cuando dieron nombre al *"cuatro"*, la idea que se usa hoy como estereotipo de lo obvio y fácil *"dos más dos son cuatro"*, les ayudó a entender y formular el concepto *"más"*. Los niños de cada generación reviven este proceso de entender significados y acceder a nuevos niveles de comprensión que compartimos todos. Esa propiedad de las matemáticas de reconstruirse en la mente de cada ser humano se debe a que se arman con una lógica implacable. No son arbitrarias sino naturales y están basadas en el razonamiento. Lo que aquí llamamos "entender", ese momento de iluminación, o respiro, en el que todas las piezas caen en su lugar, ese "¡ah, sí!" que hemos experimentado todos, es lo que solidifica y nos permite compartir, más allá de tiempos y culturas, ese mundo etéreo de las matemáticas.

Entender, en el sentido de razonamiento y no en la acepción autoritaria de "¡entiende, niño!", está muy relacionado con dudar o cuestionar. Cuando entendemos algo es porque resolvimos, aunque no necesariamente de manera explícita, alguna duda. Por consiguiente, es natural que al solucionar un problema o acceder a un nuevo nivel de entendimiento surjan nuevas dudas o preguntas. Trabajar en problemas que surgen de las propias matemáticas, es decir, pensar en ellos, tratar de resolverlos o clarificar las interrogantes asociadas que surgen a su alrededor, ha sido uno de los grandes generadores del conocimiento matemático. Hoy en día, este proceso ocurre cotidianamente en la investigación y es, probablemente, el motor más prolífico de desarrollo que tienen las matemáticas. Veamos un ejemplo de lo anterior, al que puede asociarse el nacimiento de las matemáticas como ciencia.

1.4.1 Pitágoras, Fermat y Wiles

En Babilonia y Egipto se desarrolló el conocimiento para construir edificaciones y trazar ciudades, el cual perdura hasta nuestros días. Uno de los puntos básicos de esta tecnología es poder trazar ángulos rectos, pues da lugar a formas racionales de dividir los espacios, y para lograrlo se usaba el siguiente método. Al trazar un triángulo cuyos lados miden 3, 4 y 5 unidades de longitud —no importa si son metros, pies o una vara cualquiera—, se forma un ángulo recto entre los lados que miden 3 y 4. Este mismo método lo usan en la actualidad nuestros albañiles para encontrar la "escuadra" con más precisión de lo que da una pequeña regleta: toman un "reventón" —así le llaman a los hilos con que se guían para la construcción— de 12 metros $(3+4+5=12)$, lo tensan en los puntos correspondientes y ahí, entre los lados de 3 y 4 unidades, obtienen su esquina perfecta.



Esta manera de trazar triángulos rectángulos se manejaba como conocimiento empírico, es decir, práctico y corroborado por la experiencia, hasta que los griegos exhibieron lo que hacía tan especial a la terna de números 3, 4 y 5. Para esto, deben haberse preguntado algo como: ¿habrá otras ternas de números con la propiedad de formar ángulos rectos?, ¿qué las distingue?

La respuesta a la pregunta sobre qué ternas forman ángulos rectos es sorprendente: tiene que ver con áreas. Si nos fijamos en los tres cuadrados cuyos lados son los números dados, la suma de las áreas de los dos cuadrados chicos tiene que ser igual al área del cuadrado grande. Cualquier albañil babilonio estará de acuerdo en que con 25 losetas cuadradas se pueden cubrir un cuadrado de 5 por 5, o bien uno de 3 por 3 y uno de 4 por 4. Con la notación de nuestros días, podríamos escribir lo anterior como:

$$3^2 + 4^2 = 9 + 16 = 25 = 5^2 .$$

Se sabe que los babilonios ya habían observado esta relación numérica y, además, de que conocían otros ejemplos. Pero que esto no fuera una casualidad, sino la razón profunda de que el método para trazar ángulos rectos funcionara, es un salto a otro nivel de entendimiento. Una pregunta abstracta, intrínsecamente matemática, sobre la relación entre una terna de números (3, 4 y 5) y un hecho geométrico, el que con ellos se construye un ángulo recto, da lugar al célebre teorema de Pitágoras, uno de los primeros en su género y un logro de la humanidad equiparable al descubrimiento del fuego o de la rueda. El descubrimiento de la primera "verdad" general y no trivial, como que dos más dos son cuatro, que no depende de la fe o la experiencia, sino que es producto del razonamiento puro. Una verdad que tiene razón de ser más allá de la cultura o idiosincrasia desde la cual se le contemple. Una verdad que, además, rige a la realidad que nos circunda y sugiere que dicha realidad puede comprenderse por medio de la razón.

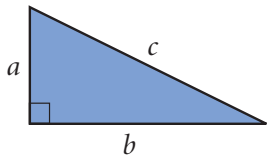


Figura 1.27 En un triángulo rectángulo se les llama **catetos** *a los dos lados que forman el ángulo recto e* **hipotenusa** *al tercer lado.*

El teorema de Pitágoras dice que las ternas de números $a,\ b,\ c$, ordenadas de manera creciente y que corresponden a las longitudes de los lados de triángulos rectángulos son las que cumplen la relación:

$$a^2 + b^2 = c^2. \qquad (1)$$

Pitágoras lo enunció en términos de áreas y para nuestros propósitos no es relevante cuál fue su motivación o argumentación original, lo importante es que dio el gran salto a un enunciado general y demostrable. Una prueba del teorema se obtiene al colocar, de dos maneras diferentes, cuatro triángulos rectángulos con catetos a y b en un cuadrado de lado $a + b$, como en la figura 1.28. Lo que dejan de cubrir estos cuatro triángulos en una de ellas son dos cuadrados de lados a y b, con área total $a^2 + b^2$; y, en la otra, lo que no cubren es un cuadrado cuyo lado es la hipotenusa *c*.

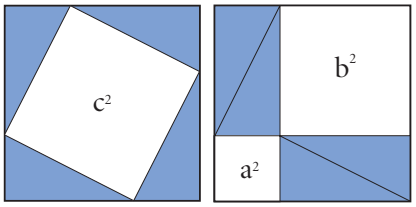


Figura 1.28 Dos maneras de colocar cuatro triángulos rectángulos en un cuadrado cuyo lado es la suma de sus catetos.

De la ecuación (1), al considerar que la terna $3,\ 4,\ 5$ la satisface, surgen dos preguntas interesantes. La primera es si habrá otras ternas de números naturales que la cumplan, además de las obvias que son múltiplos de la original como por ejemplo $6,\ 8,\ 10$, es decir, si hay otros triángulos rectángulos cuyos lados sean enteros. La respuesta es que sí. Por ejemplo, la siguiente "terna pitagórica" primitiva, que quiere decir que no es múltiplo de otra, es $5,\ 12,\ 13$. La escuela que se formó alrededor de Pitágoras y continuó los trabajos después de su muerte, alrededor del año 507 a.C., obtuvo familias infinitas de estas ternas. Demostraron, por ejemplo, que cualquier número impar es parte de una de ellas. Sin embargo, no fue hasta el siglo XIII que Fibonacci encontró un método para listar todas las ternas pitagóricas.

La segunda pregunta que surge de la ecuación (1) es más general y ya no tiene una connotación geométrica tan evidente. No obstante, tiene una historia que ilustra el punto remarcado en esta sección. Si cambiamos el exponente 2 por uno más grande $3,\ 4,\ 5,\ 6, \ldots$, que podemos denotar por n, se obtiene la ecuación

$$a^n + b^n = c^n. \qquad (2)$$

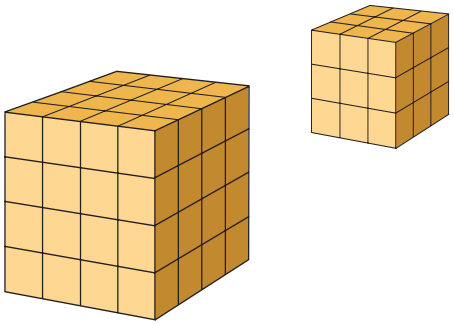


Figura 1.29 Dos cubos de lados enteros. ¿Con sus cubitos se podrá armar otro?

Para $n = 3$, esta ecuación tiene aun un significado geométrico. ¿Se puede partir un cubo formado por cubitos unitarios en dos cubos también formados por cubitos unitarios? Es una pregunta que nos invita a buscar una solución: ¿ $2^3 + 3^3$ será igual a 4^3? No, porque $8 + 27 = 35$ es diferente de 64. Pero quizás haya otros números para los cuales sí es cierto... o quizá no. O bien podemos encontrar una terna de números y resolvemos la pregunta, o bien habrá que demostrar que esto es imposible.

La pregunta más general puede plantearse así: ¿la ecuación $a^n + b^n = c^n$ tiene soluciones enteras?, es decir, ¿existen números a, b, c enteros tales que $a^n + b^n = c^n$? Este problema se atribuye a Pierre de Fermat en el siglo XV, aunque es muy probable que se hubiera planteado antes dado lo natural que es.

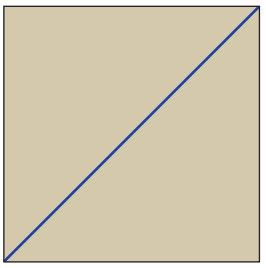
Cuenta la leyenda que Fermat escribió en el margen de un libro que había encontrado una demostración muy elegante de que sólo para $n = 2$ había soluciones enteras, resultado que se refiere a las ternas pitagóricas. Dicha demostración "elegante" nunca apareció, por lo que se hizo famoso el problema como el "Teorema de Fermat". En realidad no era —todavía— un teorema sino una *conjetura*: la intuición profunda de algo que debe ser cierto, acompañada de fuertes indicios de que es cierto, pero que carece de una demostración. Fermat sí dejó una demostración para el caso $n = 4$, es decir, que para $n = 4$ la ecuación (2) no tiene ninguna solución entera. A lo largo de los siglos siguientes se demostró la imposibilidad para otros casos como $n = 3$, 5 y 7. También se demostró para algunas familias infinitas de potencias n. Cada paso añadía fundamento y *glamour* a la conjetura.

No fue sino hasta 1995 cuando se concluyó la demostración del teorema de Fermat. El matemático inglés Andrew Wiles dio los pasos finales y es a quien se atribuye la demostración. Quedará para los siglos venideros como el "teorema de Fermat-Wiles". Pero, como decía Newton, para ver lejos, Wiles se paró sobre los hombros de gigantes. En este caso, los gigantes son innumerables pues se desarrollaron áreas completas de las matemáticas, como la teoría algebraica de los números y la geometría algebraica, cuyas ideas y conceptos son la base de la demostración. Éste es un ejemplo célebre de un problema intrínsecamente matemático, sencillo de plantear, que fue semillero de grandes desarrollos. Ejemplo claro de cómo las propias matemáticas generan preguntas que dan lugar a nuevas matemáticas.

1.4.2 Los números irracionales y el espacio euclidiano

Volviendo a la antigua Grecia, la experiencia de haber descubierto que se pueden obtener resultados generales con demostraciones formales, como el teorema de Pitágoras, dio un impulso enorme a las matemáticas. Se puede decir que con ello se da su banderazo de salida. La recién creada "escuela pitagórica" plantea muchos problemas y avanza en la sistematización de las matemáticas. Para sus miembros, fue como descubrir un continente inexplorado con tesoros y recompensas detrás de cada loma: teoremas y más teoremas, demostraciones y construcciones. Los hizo sentir que habían encontrado el camino del entendimiento y la iluminación por medio de la razón; experimentaban colectivamente el éxtasis de la creatividad intelectual y se constituyeron como una secta. Los pitagóricos llegaron a creer que *la estructura del Universo era aritmética y geométrica*. Y lo creyeron con tal fervor que, cuando un joven de su escuela demostró con sus propios métodos que había una distancia que no se podía expresar mediante números, tomaron una actitud sobrecogedora que revela lo profundo de su convicción. Decidieron que se trataba de un error de Dios y que, como escuela, tenían la obligación de guardarle el secreto. Juraron sobre su vida no revelar nunca ese desliz en la "creación" del Universo.

El argumento del joven que reveló el "fallo de las divinidades" se conoce ahora como el descubrimiento de los números irracionales o, más concretamente, de que la raíz cuadrada de dos es irracional. La concepción que tenían los pitagóricos de la "aritmética" es que los números se expresan como razón entre dos naturales, que son de la forma $\frac{p}{q}$ con p y q enteros; éstos son lo que ahora conocemos como *números racionales*, porque se expresan como una "razón", a pesar de que en México les llamamos "quebrados". Por otro lado, en la geometría surge de manera muy natural una longitud, la de la diagonal de un cuadrado *unitario*, con lado 1.



Como dicha diagonal es la hipotenusa de un triángulo rectángulo con catetos de longitud 1, por el teorema de Pitágoras, al elevar a este número al cuadrado se obtiene 2, pues $1^2 + 1^2 = 1 + 1 = 2$. A este número lo llamamos ahora "raíz de 2" y lo denotamos $\sqrt{2}$. Para los griegos era el lado de un cuadrado con área 2. Suponer que $\sqrt{2}$ se expresa como quebrado o fracción lleva a una contradicción —lo que los pitagóricos explicaron como "un error de Dios"—. Veamos cómo llegaron a esa contradicción.

Si suponemos que $\sqrt{2}$ es una fracción, entonces deberían existir dos números enteros p y q, tales que:

$$\sqrt{2} = \frac{p}{q}.$$

Puede considerarse que p y q no son ambos pares —cada quebrado se puede escribir así, ya que las fracciones $\frac{p}{q}$ pueden simplificarse hasta que el numerador o el denominador no sea par; por ejemplo, cuando simplificamos $\frac{12}{16} = \frac{6}{8} = \frac{3}{4}$, el numerador ya no es par. Al multiplicar ambos lados de la igualdad anterior por q se obtiene la nueva igualdad:

$$\sqrt{2} \cdot q = p.$$

Si ahora se elevan al cuadrado ambos miembros, como $(\sqrt{2} \cdot q)^2 = 2 \cdot q^2$, se observa que:

$$2 \cdot q^2 = p^2.$$

Como el lado izquierdo es par, ya que 2 es uno de sus factores, también lo es el lado derecho, es decir, p^2 es un número par. Obsérvese que el cuadrado de un número impar siempre es impar y que el cuadrado de un número par, además de ser par es divisible por 4. Así que como p^2 es par, debe ser divisible por 4. Por tanto, se sigue que q^2 también es par y, entonces, también q es par. Esto contradice lo que supusimos sobre p y q al principio de nuestro argumento: que no eran ambos pares. Esta contradicción nos indica que la hipótesis de que $\sqrt{2}$ podía escribirse como una fracción, como un quebrado, tiene que ser falsa. En otras palabras, $\sqrt{2}$ no puede ser un número racional y, por definición, es irracional.

De las propias matemáticas, de la interacción de dos de sus disciplinas —la aritmética y la geometría— y con razonamientos matemáticos, surgió un descubrimiento inesperado. Y con él, una andanada de preguntas y problemas: ¿habrá más números irracionales?, ¿cómo los expresamos o trabajamos con ellos?, etc. Éste es otro ejemplo histórico de cómo las matemáticas se motivan a sí mismas para crecer e indagar continuamente.

Por último, veamos un ejemplo de cómo algunos desarrollos matemáticos tienen consecuencias inesperadas y profundas tanto en la comprensión de la naturaleza como en la actividad humana.

El ímpetu que se dio al desarrollo de las matemáticas en la antigua Grecia alcanza un clímax con la aparición, alrededor del año 300 a.C., de *Los elementos de Euclides*. Son una serie de libros en los que se sistematizan los resultados y los conocimientos de geometría que se habían obtenido hasta ese momento, y en los que se establece el **método axiomático**. Nuestras ideas abstractas de punto, línea, plano, distancia, ángulo, área, volumen, etc., se fundamentan en esa obra, que se usó como libro de texto por más de dos milenios. En ella, se delinea el modelo teórico de lo que subyace al espacio en que vivimos, que se llama aún el **espacio euclidiano** y es el escenario de la geometría de nuestro entorno. Suponemos desde entonces que sobre un espacio euclidiano —la esencia geométrica tridimensional mínima, limpia y vacía— está transcurriendo el mundo y que ahí se hospeda y se mueve la materia, el mundo físico con nosotros incluidos. Y esto sigue siendo válido sólo en pequeña escala, porque Einstein, a principios del siglo xx, nos cambió esa idea a gran escala. Pero, en fin, en el espacio euclidiano es donde tiene sentido hablar de planos, y en ellos de triángulos y de círculos.

Emparentadas íntimamente con los círculos están las elipses. Un círculo visto de lado es una elipse; de alguna manera, las elipses son círculos "apachurrados". En muchos casos en los que nuestros ojos ven una elipse —vasos, tazas, llantas, etc. —, nuestra mente lo interpreta como un círculo porque el cerebro "sabe" que al cambiar el punto de vista la figura es, efectivamente, un círculo. Los griegos se dieron a la tarea de estudiarlas y, en particular, de dar una definición precisa de ellas. Encontraron dos. La primera tiene que ver con cómo trazarlas en el plano. Son el lugar geométrico de los puntos cuya suma de sus distancias a dos puntos fijos, llamados focos, es constante. Así que para trazar una elipse en un jardín tendríamos que hacer lo siguiente: se clavan dos estacas —que serían los focos—, se les amarra una cuerda holgada cuya longitud es la constante que se menciona en la definición y, luego, al ir tensándola en todas las direcciones, se dibuja la elipse. A partir de esta definición queda claro que el círculo es un caso límite de elipse que corresponde a cuando los dos focos coinciden en su centro y, entonces, la constante es el doble del radio.

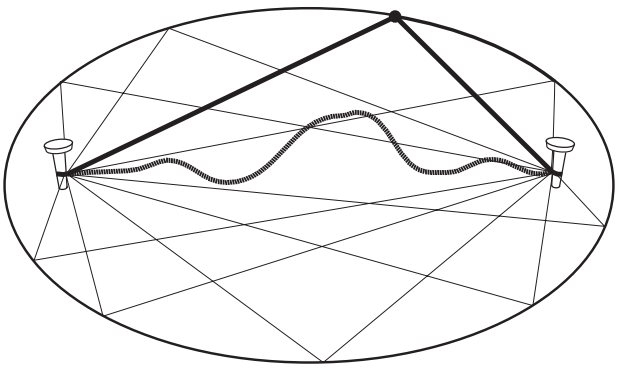


Figura 1.30 "Método del jardinero" para trazar una elipse.

La segunda definición tiene que ver con la percepción de círculos y necesita de la tercera dimensión pues se sale del plano. Si levantamos el centro de un círculo perpendicularmente al plano en el que "vive" —digamos que horizontal— y luego consideramos todas las líneas que van de este punto al círculo, obtenemos un cono circular. Las elipses se obtienen al cortar este cono con planos inclinados cercanos al horizontal —la dirección en la que se inclina el plano, corresponde a aquella en la que se "alarga" el círculo. Nuestra percepción de círculos se explica con el proceso inverso: un cono elíptico cortado por otro plano origina el "círculo" real.

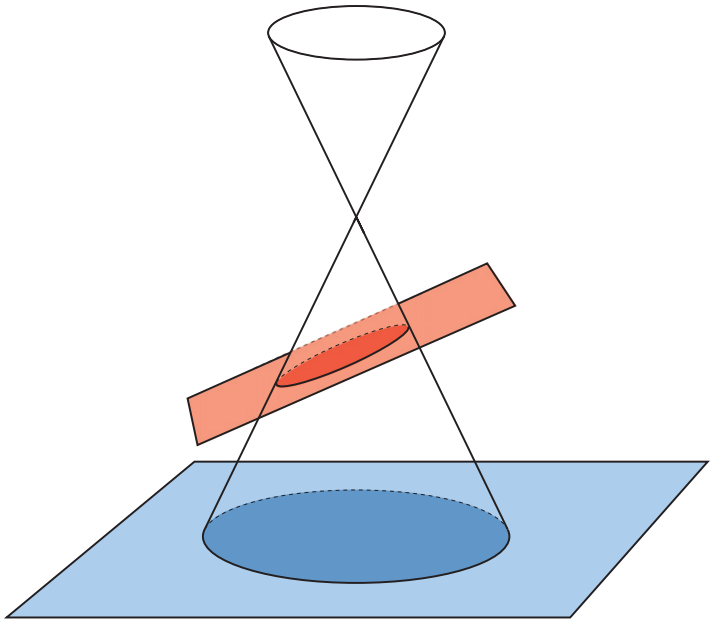


Figura 1.31 Las elipses se obtienen al cortar un cono circular con un plano.

Si inclinamos más de la cuenta el plano —y ya que estamos en la teoría, no hay nada que nos detenga, por lo que sería natural hacerlo—, se obtienen dos nuevos tipos de curvas: las parábolas y las hipérbolas. De aquí el nombre de **secciones cónicas** que dieron los griegos, con la participación destacada de Apolonio de Perga, a estos tres tipos de curvas planas. Obtuvieron descripciones de estas curvas como lugares geométricos en términos de distancias, similares a la que vimos para las elipses. Se maravillaron con los resultados que obtuvieron y que dejaron como legado a la posteridad. Hicieron matemáticas por sí mismas que parecían muy alejadas de la "realidad".

Quién hubiera pensado en ese entonces que, dos milenios después, esas mismas curvas sirvieran para entender fenómenos físicos como la trayectoria de los planetas alrededor del Sol —que son elipses con el Sol en uno de sus focos—, o las trayectorias de las balas o proyectiles —que describen parábolas. O bien, dentro de las propias matemáticas, como cuando Descartes puso coordenadas al plano y las secciones cónicas reaparecen como soluciones de ecuaciones cuadráticas con dos incógnitas. O, más aún, que las hipérbolas serían fundamentales para diseñar los lentes ópticos y las parábolas para las telecomunicaciones del presente. Esa "magia" que se ha dado una y otra vez en la historia, ese hecho insólito de que las matemáticas desarrolladas por sí mismas luego reaparecen y se interconectan con otros fenómenos al parecer independientes —tanto de la naturaleza como de la actividad humana—, es uno de sus grandes misterios. Hace pensar, como lo harían Pitágoras, Platón y luego Galileo, que son la herramienta fundamental para entender nuestro entorno.

1.5 CONCLUSIÓN

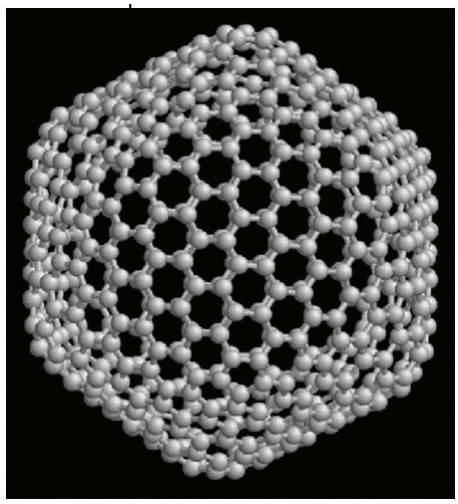


Figura 1.32 Fullereno icosaédrico, un arreglo de 540 puntos que forman hexágonos y pentágonos. Esta forma estructura moléculas que se producen artificialmente y se estudian por sus aplicaciones en la medicina.

Los ejemplos de las secciones anteriores muestran al lector que las fuentes de los problemas matemáticos son variadas, pero pueden clasificarse de manera natural en los tres grandes grupos mencionados: la actividad humana, la naturaleza y las matemáticas mismas.

A pesar de la variedad de sus orígenes, las matemáticas forman un cuerpo de conocimientos perfectamente coherente e interconectado. Los números se relacionan con las figuras geométricas y ambos, con las expresiones algebraicas. Todos los aspectos de las matemáticas se relacionan entre sí y forman una unidad de extraordinaria coherencia, belleza y poder.

Pero ¿qué son las matemáticas? El último capítulo del libro enfrentará esta pregunta que también es un reto para la mente humana, en este caso, de tipo filosófico. De momento, partiendo de los ejemplos presentados en este capítulo, daremos una respuesta tentativa. Las matemáticas son el destilado de la actividad intelectual; aquello que aunque provenga de problemas concretos resulta ser una verdad general necesaria y absoluta que trasciende los casos particulares que motivaron su descubrimiento. Las matemáticas son un producto de la observación, la reflexión y el trabajo humanos, pero sus verdades tienen tal fuerza que parecen ser parte de un mundo absoluto, cerrado y perfecto que existiera independientemente del hombre.

La capacidad que tienen las matemáticas para representar el mundo físico nos inclina a creer que hay una relación íntima, que aún no comprendemos bien, entre la naturaleza y las matemáticas. Sin embargo no podemos afirmar esta idea con total seguridad.

Toda la actividad matemática, independientemente de donde se inicie, se realiza apoyada en la irrefrenable curiosidad del hombre. Es sorprendente la fuerza que da al ser humano esta curiosidad y la cantidad de energía que está dispuesto a emplear para satisfacerla. Parece ser una característica genética que, como casi todas las de una especie, probablemente esté ligada a su capacidad para sobrevivir, lo mismo que la capacidad de oponer el dedo pulgar a los otros, caminar erguido y seguir el instinto sexual. Resolver crucigramas, hacer *sudoku*, jugar ajedrez o *bridge*, leer novelas de misterio y resolver acertijos son algunos ejemplos de actividades intelectuales cotidianas que los seres humanos realizamos por gusto sin que nadie nos obligue. Actividades que requieren esfuerzo, difíciles, pero que pueden proporcionar mucha satisfacción.

En la película de Stanley Kubrick, *2001 Odisea del espacio*, el mono, al darse cuenta de que puede usar un hueso como herramienta para defenderse y atacar, celebra su des-

cubrimiento al golpear a diestra y siniestra con él y, al lanzarlo al aire en su euforia, en pleno vuelo se transforma en una nave espacial. Esta famosa, poética y emocionante escena ilustra la satisfacción del descubrimiento y sugiere que con ello se inicia un proceso de superación, que lo llevará a conquistar el planeta y posiblemente el Universo. Es una metáfora perfecta de la evolución humana y su relación con el descubrimiento y la invención.

Figura 1.33 El descubrimiento del hueso como herramienta en la película de Stanley Kubrick | © Latin Stock México.

Nuevos descubrimientos renuevan la experiencia y repiten la emoción. El fenómeno acaba por convertirse en una característica genética de la especie que le ayuda en su progreso. Con el paso de muchas generaciones, el placer por el descubrimiento se extiende a todos los ámbitos de la vida: el trabajo, el juego, el deporte, la lectura, etc. La innovación produce placer. El descubrimiento produce placer. La superación de un reto produce placer.

El ser humano sabe que al esforzar su intelecto puede resolver problemas y descubrir verdades; dicha actividad le produce una satisfacción extraordinaria, equiparable a la de encontrarse en el pico más alto de una montaña o al placer erótico, cuando esa pulsión, vital al fin y al cabo, lo lleva a la cúspide, al orgasmo. El placer de superar retos es, probablemente, lo que más ha contribuido al desarrollo de las matemáticas y a lo que las matemáticas deben su dinamismo y vitalidad.

MATEMÁTICAS DE LA ACTIVIDAD HUMANA

TEMA

2

Figura 2.1 Niña jugando con ábaco | © Latin Stock México.

2.1 INTRODUCCIÓN

Nuestra vida moderna está llena de avances tecnológicos a tal grado que es difícil imaginarla sin ellos. Menos conocido, en cambio, es el hecho de que gran parte de esta tecnología se basa en desarrollos matemáticos, tanto de siglos pasados como de la actualidad. Hoy día, las matemáticas están escondidas sobre todo en los aparatos electrónicos y por ello es difícil percatarse de la importancia que tienen para el funcionamiento de éstos.

Ésta es una característica de la modernidad y no siempre fue así. Nuestra era actual nace a partir de una "época mecánica", en la cual dominaba la preocupación sobre conceptos de la física que, desde Galilei, se formularon en el lenguaje de las matemáticas. Si retrocedemos todavía más, encontramos que aun las primeras culturas avanzadas, como los babilonios o los egipcios, tenían conocimientos matemáticos que usaron para gran provecho en la resolución de problemas de la vida diaria. Entre estos primeros conocimientos matemáticos es-

tá lo que hoy se conoce como el "teorema de Pitágoras", que sirvió para establecer cuándo un ángulo es recto.

Aunque las matemáticas que se tratan en este tema tienen su origen en la vida diaria, no es fácil encontrarlas en forma cotidiana. En efecto, las matemáticas que se verán aquí no son útiles en el sentido de que no aportan algo para ir de compras al mercado o arreglar un automóvil; al igual que la historia, por ejemplo, nos permiten entender mejor nuestra cultura y, por lo tanto, ser partícipes de ella.

Este tema no presenta un desarrollo histórico sino, más bien, uno conceptual: comienza al revisar el concepto de número y diversas aplicaciones; después versa sobre conceptos más geométricos como la medición de áreas o volúmenes. En seguida, el tema recorre los inicios de un campo que se formalizó sólo hasta el siglo XVIII y que hoy se conoce como el "cálculo integral". Al final, se revisa incluso cómo las matemáticas llegaron a dominar el azar.

2.2 NÚMEROS PARA CONTAR

¿Alguien está haciendo "cuernos"? No; así muestran los chinos el número seis. Hay diferentes convenciones sobre cómo se cuentan los números con las manos. Nosotros usamos el índice y el dedo medio para indicar el dos, mientras los europeos usan el índice y el pulgar. Contar es una de las habilidades que aprendemos desde muy pequeños y nos distingue de muchos animales, aunque no de todos; se sabe, por ejemplo, que los cuervos pueden contar bastante bien hasta 6 o 7.

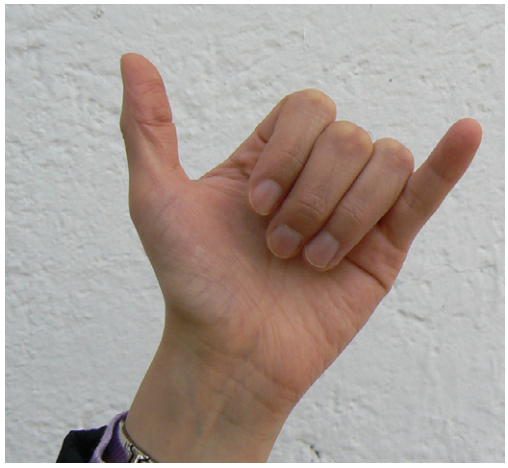


Figura 2.2 Una seña que, según el contexto cultural, puede tener dos significados muy distintos.

Se dice que contar es una de las primeras actividades humanas: aunque todos contamos con los dedos, no se hace igual en todos lados. Precisamente por ello puede parecer sorprendente que, hoy en día, la gran mayoría de las culturas usan sólo los números arábigos para denotar cantidades. ¿Qué es lo que hace tan particularmente especiales a los números arábigos? ¿Por qué se impusieron sobre otras maneras de denotar? Preguntas como esta y otras se responderán en este apartado. Además, queremos evidenciar que la forma en que denotamos los números incide en la facilidad para hacer operaciones como sumar, restar o multiplicar. Descubriremos que la invención del cero juega un papel muy importante en todo lo anterior. El cero no sólo es la expresión numérica de la ausencia o la nada, sino que posibilita el reutilizar las mismas cifras para denotar unidades, decenas, centenas, miles o millones, como veremos más adelante.

El libro *Discorsi* de Galilei fue escrito en el año MDCXXXVIII, como se puede apreciar en la figura 2.3. Si bien ya no estamos familiarizados con esta manera de escribir los números —conocidos como números romanos—, para leer la fecha en la portada del libro hay que saber los valores de cada símbolo:

$$M = 1000,\ D = 500,\ C = 100,\ L = 50,$$
$$X = 10,\quad V = 5,\quad I = 1$$

Entonces, podemos conocer, sin mayor dificultad, que el año de impresión es 1638.

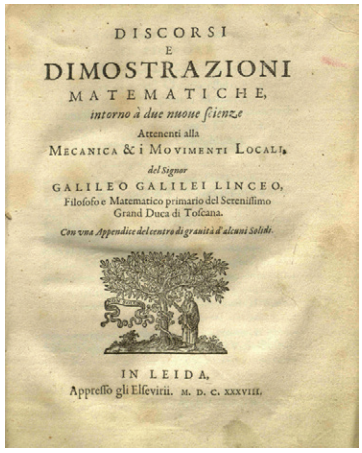
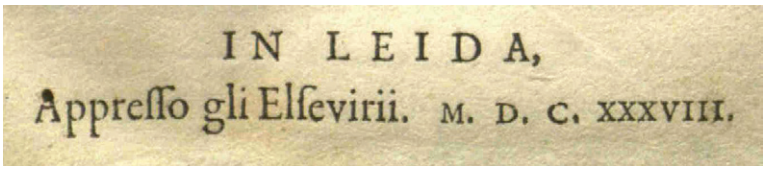

DISCORSI
E
DIMOSTRAZIONI
MATEMATICHE,
intorno à due nuoue scienze
Attenenti alla
MECANICA & i MOVIMENTI LOCALI,
del Signor
GALILEO GALILEI LINCEO,
Filosofo e Matematico primario del Sereniſſimo
Grand Duca di Toſcana.
Con vna Appendice del centro di grauità d alcuni Solidi.
IN LEIDA,
Appreſſo gli Elſeuirii. M. D. C. XXXVIII.

Figura 2.3 Portada de la primera edición de los Discorsi *de Galileo Galilei.*


IN LEIDA,
Appreſſo gli Elſevirii. M. D. C. XXXVIII.

Más difícil es traducir el año de impresión del libro *De revolutionibus orbium coelestium* de Copérnico, que es MDXLIII. Uno podría pensar que es igual a la suma de $1000 + 500 + 10 + 50 + 3 = 1563$ pero, en realidad, fue impreso en 1543. Lo que sucede en este caso es que hay que aplicar una regla para el uso de los números romanos: restar un signo de menor valor cuando antecede a uno de mayor valor. Como el X está antes del L, se resta 10 a 50 y por ello se obtiene la cifra de 40.

Se complica el asunto si tratamos de multiplicar con números romanos. Por ejemplo:

$$\text{XIX por XLVII} = ?$$

¡Claro! X por XLVII es CDLXX, ya que simplemente se cambian los signos al multiplicar por diez. Así, lo que debemos calcular es:

$$\text{CDLXX}-\text{XLVII}+\text{CDLXX}$$

Sin embargo, ahora empezamos con algunas dificultades. Por ejemplo, CD + CD no es CCDD ni CDCD, sino DCCC. Luego, LXX + LXX no es LLXXXX, sino CXL. Los dos resultados anteriores se deben sumar y después restarles XLVII. Entonces, la operación sería:

$$\text{DCCC} + \text{CXL} - \text{XLVII}.$$

Se observa que el segundo sumando es C + XL, del cual se resta XL + VII. De esta manera, podemos quitar el término XL en ambos lados:

$$\text{DCCC} + \text{C} - \text{VII}.$$

Ahora bien, el cien —es decir, C— tiene que prestar X para poder restar VII. Usamos que C = XC + X. Además, X − VII = III. Finalmente, el resultado de la operación es:

$$\text{DCCC} + \text{XC} + \text{III} = \text{DCCCXCIII}.$$

Hicimos este cálculo sin transformar los números a nuestro sistema decimal, es decir, sin usar que XIX = 19 y XLVII = 47. El algoritmo que se enseña en la secundaria nos proporciona el resultado sin problema alguno:

$$19 \times 47 = 893.$$

$$\begin{array}{r} 19 \\ \times\ 47 \\ \hline 133 \\ 76 \\ \hline 893 \end{array}$$

En efecto, $893 = \text{DCCCXCIII}$. ¿Por qué parece más difícil trabajar con números romanos? No es solamente que no estamos tan familiarizados con estos símbolos sino, también, que no hay un algoritmo más o menos sencillo. En nuestro sistema decimal las mismas cifras se vuelven a reutilizar: la cifra 2 puede significar 20, 200 o dos millones en el número $2\ 480\ 225$. En este último número se observa algo peculiar: una de las cifras es un cero, así que tiene que expresar de manera explícita la ausencia de múltiplos de los miles. Si no tuviéramos un cero, el número no se podría escribir y se leería como 248225 que, obviamente, es algo equivocado.

El cero se inventó un par de veces en la historia de la humanidad. Aparece en el siglo III a.C. con la cultura babilónica y, también, con la maya en el siglo I d.C. Nuestra notación del cero se la debemos a la India, que la transmitió a los árabes. Todas estas culturas usaron un sistema posicional —donde se indica el valor de los símbolos según su posición—. En el número 22, el primero y el segundo 2 tienen significados diferentes, aunque se escriba exactamente el mismo símbolo en ambos casos. Esto es una gran ventaja: se usan sólo 10 símbolos. Además, las operaciones con los números expresados en este sistema se pueden explicar razonablemente bien y funcionan para todos los casos, sin importar la cantidad de dígitos.

El sistema posicional también tiene una desventaja: es fácil cambiar los números de lugar. Basta agregar, por ejemplo, un simple 4 antes del 720 para que el cheque ahora diga $4\ 720$, ¡qué buen negocio sería! Por lo anterior, en los cheques y en cada documento importante que implique cantidades se suelen repetir los números en forma escrita: "720 y setecientos veinte".

Otra ventaja del sistema posicional es que resulta fácil escribir números muy grandes. En el sistema romano es muy difícil: se ponían rayas sobre los símbolos para indicar que ahora se trataba de una magnitud mayor. Por ejemplo, $\overline{\text{IV}}$ significaba $4\ 000$ y $\overline{\text{X}} = 10\ 000$. En nuestro sistema no tenemos dificultad para representar un millón o un millón de millones, término que se conoce en español como billón. ¡El sistema no tiene límites! Por ejemplo, la masa del Sol es aproximadamente de:

$$1\ 990\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\,\text{kg}$$

Notemos que usualmente se ponen comas o se incluyen espacios en blanco para indicar grupos de tres y no perder la cuenta. En la escritura romana, la misma cantidad tendría que representarse como:

es decir, como 1990 con nueve rayas encima. Pero aun en el sistema decimal, un número así de grande es difícil de leer y para eso se usa la notación científica:

$$1\ 990\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\ 000\,\text{kg} = 199 \cdot 10^{28}\,\text{kg}$$

es decir, el número 199 seguido de 28 ceros y que se lee como "199 por diez a la veintiocho".

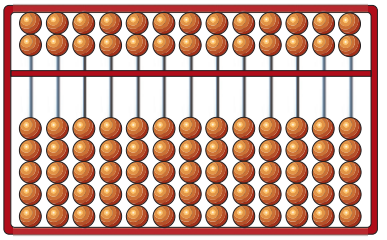


Figura 2.4 El ábaco está basado en el sistema decimal.

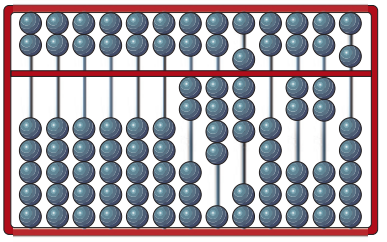


Figura 2.5 Representación del número 2 480 225.

El ábaco que se muestra en la figura 2.4 se basa también en el sistema decimal. Como se observa, tiene varias hileras usualmente divididas en dos partes: una de cinco canicas y otra de sólo dos. Las primeras valen uno —diez, cien, mil, etc. — mientras las de dos valen cinco —cincuenta, quinientos, cinco mil, etcétera.

En la figura 2.5 se muestra el número $2\ 480\ 225$. Alguien experimentado en el uso del ábaco es muy rápido al efectuar sumas y, razonablemente rápido al hacer multiplicaciones. Cabe mencionar que en el fondo, emplea los mismos procedimientos que nos enseñan a nosotros.

Dichos procedimientos son muy naturales. Por ejemplo, si queremos sumar 493 más $2\ 865$, es claro que la suma tendrá: $3+5=8$ unidades, $9+6=15$ decenas, $4+8=12$ centenas y $2+0=2$ miles. Pero 15 decenas son una centena y 5 decenas. Así que se debe aumentar el número de centenas por uno: tenemos 5 decenas y $12+1=13$ centenas. Éstas son a la vez un millar y 3 centenas. Tenemos entonces 3 millares, 3 centenas, 5 decenas y 8 unidades, es decir $3\ 358$. En el procedimiento que se aprende en la primaria, el razonamiento anterior se repite una y otra vez hasta que se hace en automático, de tal manera que ya no haya que pensar en él cada vez.

Con la multiplicación sucede algo similar: se hace de manera muy automatizada. En el fondo, hay una necesidad de hacerlo justo así y no de otra manera. Por ejemplo, si se quiere multiplicar 19 por 47 —el ejemplo que hicimos con los números romanos—, debemos calcular siete veces 19 y luego sumar cuarenta veces 19. Pero 7 por 19 se puede calcular como 7 por 9 y sumarle 7 por 10. De esta forma se obtiene que:

$$7\times 19 = 7\times 9 + 7\times 10 = 63 + 70 = 133. \qquad (1)$$

Por otro lado, cuarenta veces 19 no es otra cosa que multiplicar 19 primero por cuatro y luego por 10. Sin embargo, como multiplicar por diez implica añadir un cero, resulta que:

$$4\times 19 = 4\times 9 + 4\times 10 = 36 + 40 = 76$$

y por lo tanto,

$$40\times 19 = 760. \qquad (2)$$

Al sustituir las operaciones de (1) y (2) tenemos que:

$$19\times 47 = 19\times 40 + 19\times 7 = 760 + 133$$

y esto se resuelve con el procedimiento de la adición: $760+133=893$. Si procedemos a multiplicar siguiendo el esquema aprendido no hacemos otra cosa que seguir esta lógica.

Lo principal es que no importa cuán grandes sean los números, el procedimiento siempre funciona —aunque a veces sea muy tedioso y prefiramos tomar una calculadora de bolsillo para resolverlo. Aunque estas herramientas nos son muy útiles, también es relevante comprender que el mecanismo que gobierna el procedimiento aprendido es una absoluta

necesidad. En el fondo de cada multiplicación siempre están las tablas de multiplicar del uno hasta el nueve. Desde esta perspectiva, no hay nada de sorprendente en el hecho de que todos los niños del planeta tienen que aprenderse las tablas de multiplicar de igual manera.

Si denotáramos nuestros números como los romanos tendríamos serios problemas para hacer cálculos. Ahora entendemos por qué se dice que los progresos científicos de los romanos fueron muy limitados, precisamente porque no usaban un buen sistema de notación.

2.3 NÚMEROS PARA MEDIR

Figura 2.6 Personificación de la justicia* *por Luca Giordano (1634-1705), donde vemos como atributos la espada y la balanza. Esta última simboliza la mesura en el ámbito de la ética y es también una herramienta en los negocios. Las balanzas se conocen desde la Antigüedad, a través de los egipcios.<http://commons.wikimedia.org/wiki/File:Luca_Giordano_013.jpg >

Medir y contar son actividades distintas. Contamos piedras pero no contamos leche, contamos sillas y caballos pero no contamos la longitud de una mesa ni el tamaño de un rancho. De una cantidad de leche *medimos* su volumen, mientras que de una mesa y un rancho *medimos* su longitud y su tamaño, respectivamente.

Para poder medir algo es necesario antes definir las **unidades** en las que se va a expresar lo que se mide. No tiene sentido decir que tenemos 5 de leche, que la mesa mide 2 o que el tamaño de un rancho es 24. Es necesario indicar las unidades que expresan los resultados de nuestra medición. Así, diríamos que tenemos 5 litros de leche, que la mesa mide 2 metros y el rancho 24 hectáreas.

Los números para contar son muchos —de hecho, una infinidad— y aun así, no nos sirven para medir. Lo que medimos no suele constar de un número *exacto* de unidades sino que, frecuentemente, hay que recurrir a fracciones de estas unidades para dar el resultado de la medición. Por ejemplo, la vaca dio "cuatro litros y tres cuartos" de leche, la mesa mide "un metro y ochenta y cinco" centímetros y el rancho tiene "veinticuatro punto tres" hectáreas.

* *Apologia dei Medici* (detalle); fresco en la Galería Palazzo Medici-Riccardi, Florencia (1684-1686).

Los números que utilizamos para contar $1, 2, 3, \ldots$ se llaman **enteros positivos**, mientras que los que usamos para medir se conocen como **fracciones**. Las fracciones son cocientes de dos enteros positivos. Por ejemplo, $\frac{1}{2}, \frac{1}{3}, \frac{5}{7}, \frac{723}{45}$ son fracciones.

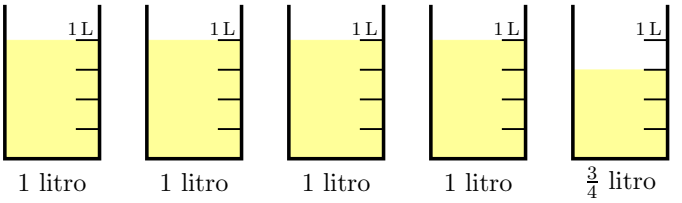


Figura 2.7 Al decir que cuando ordeñamos una vaca obtuvimos 4 litros y tres cuartos de leche, indicamos que la parte adicional —a los primeros 4 litros que obtuvimos— consta de 3 de las cuatro partes del siguiente litro —que hubiera sido el quinto—, mismo que no llegamos a completar quizá porque la vaca alimentó antes a su ternero.

Aun cuando podamos contar algunas cosas, por ejemplo, las fresas o las uvas que compramos en un mercado, preferimos medir su peso y expresarlo no con un número exacto de fresas o uvas, sino como medio kilo de fresas y un kilo de uvas. Observemos que, a veces, no pluralizamos el sustantivo de aquello que compramos por peso, sino que decimos medio kilo de fresa y un kilo de uva, lo cual indica que estamos ignorando, a propósito, el número exacto de objetos y nos concentramos en su peso, que es normalmente una cantidad fraccionaria.

Las unidades que usamos para medir no siempre fueron las mismas. En la Antigüedad, las necesidades del comercio llevaron a los hombres a establecer unidades para medir peso, longitud, área y volumen. En distintas civilizaciones se utilizaron distintas unidades de medición. Cuando el comercio comenzó a crecer cruzando fronteras, se hizo necesario establecer equivalencias entre las unidades de medida de diferentes culturas. Estas equivalencias subsisten hasta nuestros días; por ejemplo, entre el llamado **sistema inglés** y el **sistema métrico decimal**, donde un pie equivale a 30.5 centímetros, una milla a $1\ 609$ metros y una libra a 454 gramos.

En general, los sistemas de pesas y medidas de todas las culturas utilizan fracciones exactas de una unidad para expresar la medida de cualquier cantidad menor que dicha unidad. Así, un pie es la tercera parte de una yarda y una pulgada, la doceava parte de un pie. Estas relaciones, como dijimos, se expresan con fracciones. Por ejemplo, al usar la notación convenida internacionalmente para pies y pulgadas —ft e in, respectivamente— podemos escribir:

$$1\,\text{ft} = \frac{1}{3}\,\text{yarda}$$
$$1\,\text{in} = \frac{1}{15}\,\text{ft}$$

Para facilitar el comercio internacional, hace un par de siglos casi toda la humanidad se puso de acuerdo en hacer uso de un sistema de medición estándar, el sistema métrico decimal, donde las distancias se miden en metros y decenas o centenas de metros, y otras unidades de longitud que siempre son potencias de 10 de un metro —por ejemplo, $1000 = 10^3$ metros son un kilómetro— cuando se trata de distancias muy grandes. Para las pequeñas se usan unidades que representan fracciones de metros y que caben exactamente 10^n —10 "a la n veces", o bien, 10 elevado a la enésima potencia, donde n es cualquier número— veces en un metro, como es el caso de los centímetros —que son $\frac{1}{100} = \frac{1}{10^2}$ m—, los milímetros —$\frac{1}{1000} = \frac{1}{10^3}$ m— o las micras —$\frac{1}{1\ 000\ 000} = \frac{1}{10^6}$ m. Lo mismo se hace con las unidades de peso y con algunas otras, como las de temperatura.

Aun en los países donde impera el sistema métrico decimal —casi todos los del mundo—, las unidades de tiempo y angulares no utilizan —estrictamente— un sistema decimal. La definición original de las unidades de tiempo se basa en lo que tarda la Tierra en girar desde que el Sol está en su punto más alto hasta que ocurre lo mismo al día siguiente,

Figura 2.8 Modelo concreto para el metro —unidad del sistema métrico decimal resguardada en París— en una barra de platino-iridio.

es decir, en un día. En particular, un segundo es la $60 \times 60 \times 24 = 86\ 400$ —"ochenta y seis mil cuatrocientosava"— parte de un día. Si ocurriese un cataclismo —como que un meteorito muy grande chocara con la Tierra— que disminuyera un poco la velocidad de rotación del planeta, la medida del día cambiaría. Sin embargo, ya hay relojes muy precisos basados en las vibraciones moleculares de cristales con los cuales seguiríamos sabiendo lo que es un segundo, aunque ya no coincidiera con la $86\ 400$ parte del día. En efecto, una hora es la veinticuatroava parte de un día, un minuto es la sesentava parte de una hora y un segundo la sesentava parte de un minuto. Estas unidades de medida definidas como una fracción de otras más grandes, se expresan por medio de fracciones propias, es decir, fracciones con numerador igual a uno. Usemos la notación convenida para horas y segundos con afán de mostrar lo anterior:

$$1\,\text{h} = \frac{1}{24}\,\text{día}$$
$$1\,\text{min} = \frac{1}{60}\,\text{h}$$
$$1\,\text{s} = \frac{1}{60}\,\text{min}$$

Para comunicar resultados simples de medición es necesario que sepamos operar con fracciones y decimales. Por ejemplo, si nos dicen que una mesa mide seis pies y cinco pulgadas, la mayoría de nosotros no tenemos una idea clara de lo que esto significa y, si quisiéramos recortar un vidrio para ponerlo sobre la mesa, necesitaríamos convertir el dato al sistema métrico decimal —en el que tenemos nuestra cinta métrica—. ¿Cómo sabemos cuánto mide en metros y centímetros esa mesa? Podemos hacer la conversión de varias maneras, siendo la más conveniente aquella que requiera menos investigación de nuestra parte. Si no sabemos cuántos centímetros mide una pulgada y sí conocemos —conviene basarnos únicamente en— la equivalencia entre pies y centímetros que —sabemos— es de $30.5\,\text{cm}$ por pie. Entonces, primero expresamos la medida de la mesa en pies. Como vimos, una pulgada es $\frac{1}{12}$ de pie, la mesa mide $6 + \frac{5}{12} = \frac{77}{12}$ pies. Para conocer la medida en centímetros, debemos multiplicar por el número de centímetros que hay en un pie —que es 30.5—, por tanto, la mesa mide $\frac{77}{12} \times 30.5\,\text{cm}$.

Hace años, en este punto hubiéramos desarrollado en un papel las operaciones; hoy en día recurrimos a una calculadora y, en cualquier caso, obtenemos que la mesa mide $195.7083333\ldots$ cm; al aplicar el redondeo, concluimos que la mesa mide un metro y 96 centímetros, lo cual se expresa como $1.96\,\text{m}$ o como $196\,\text{cm}$. Si fuera necesaria mayor precisión podríamos decir que la mesa mide un metro, noventa y cinco centímetros y siete milímetros o $195.7\,\text{cm}$.

El simple ejercicio anterior nos muestra cómo, por las necesidades del comercio y la actividad humana en general, fue necesario desarrollar los métodos para operar con fracciones y decimales, es decir, la **aritmética**. La necesidad de medir y comparar medidas llevó al hombre, primero, a desarrollar la aritmética de las fracciones y, posteriormente, a descubrir la conveniencia de utilizar la **notación decimal** para representar los resultados de una medición, así como la de adoptar un sistema básicamente decimal de unidades de pesas y medidas.

De esta forma, el ser humano medianamente instruido debe saber cómo realizar operaciones con fracciones y decimales ya que, sin contar con esta habilidad y aunque posea la mejor calculadora del mundo, le será imposible entender y comunicar resultados básicos de mediciones. No saber aritmética equivale a ser medio analfabeta. Afortunadamente, casi todas las personas aprenden la aritmética de las fracciones —llamadas quebrados hace algu-

nos años— en la escuela primaria que, también, es por fortuna obligatoria en todos los países civilizados. Las operaciones básicas de la aritmética de fracciones y decimales consisten en sumar, restar, multiplicar y dividir —tanto fracciones como decimales—, y en convertir una fracción a decimal y un decimal a una fracción. Las calculadoras pueden ayudarnos a obtener estas conversiones y realizan bastante bien las cuatro operaciones básicas con decimales, pero no operan con fracciones; esto debe saber hacerlo la persona si quiere apoyarse en la calculadora. Por ejemplo, en el problema de convertir la medida de la mesa de pies y pulgadas a centímetros, fue necesario saber expresar las cinco pulgadas como $\frac{5}{12}$ de pie y saber que había que sumar $6 + \frac{5}{12}$ para obtener el número de pies de la mesa como una fracción. Ya planteada la suma hay dos opciones: sumar fracciones y obtener $\frac{77}{12}$, como hicimos en el párrafo anterior, o bien, convertir antes la fracción $\frac{5}{12}$ a decimales y operar con ellos. Ambos caminos son correctos, aunque el primero mantiene la exactitud por más tiempo, de manera que el resultado final, aunque sea un valor aproximado, da lugar a menor incertidumbre que el que obtenemos mediante el segundo camino. De cualquier forma, para efectuar la conversión correctamente, el poseedor de una calculadora tiene que entender lo que va a hacer para plantear las operaciones antes de realizarlas, y esto requiere de un conocimiento en el uso de las fracciones y de la conversión entre fracciones y decimales. En realidad, aunque éstos se consideren temas elementales, tienen suficientes sutilezas para que valga la pena repasarlos. Para conveniencia del lector —que pudiera tener alguna carencia en estos conocimientos básicos—, presentamos aquí un resumen de estas sutilezas en el tema de las fracciones y los decimales.

En primer lugar, varias fracciones distintas pueden representar un mismo número. Por ejemplo, $\frac{1}{2}$ representa lo mismo que $\frac{2}{4}$ y que $\frac{7}{14}$. De hecho, dados una fracción $\frac{m}{n}$ y cualquier entero positivo k, la fracción $\frac{m \times k}{n \times k}$ representa el mismo número que $\frac{m}{n}$. Los números representados como fracciones se llaman **números racionales**. Lo que acabamos de observar muestra que un número racional no es "una" fracción sino muchas, todas las que dan lugar al mismo cociente. Los matemáticos dicen a veces cosas como ésta: los números racionales son las clases de equivalencia de todas las fracciones $\frac{m}{n}$ tales que, para dos elementos $\frac{m_1}{n_1}$ y $\frac{m_2}{n_2}$ de una clase, existen enteros k_1 y k_2 que cumplen:

$$m_1 \times k_1 = m_2 \times k_2$$
$$n_1 \times k_1 = n_2 \times k_2$$

El lector no tiene que leer ni entender estas "precisiones" para saber lo que es un número racional, basta que lo piense como el cociente de una fracción y sepa que si dos fracciones dan lugar al mismo cociente, entonces ambas representan el mismo número racional. Por ejemplo, $\frac{6}{8}$ y $\frac{15}{20}$ son dos fracciones que representan el mismo número racional; también podemos decir que estas fracciones son iguales —pues aunque sean obviamente distintas por sus numeradores y sus denominadores diferentes, los cocientes sí son iguales— y se pueden escribir $\frac{6}{8} = \frac{15}{20}$. Muchas veces es conveniente representar un número racional por su fracción más simple, que es precisamente aquella en la que el numerador y el denominador no tienen divisores comunes; por ejemplo, $\frac{6}{8}$ y $\frac{15}{20}$ pueden representarse por la fracción $\frac{3}{4}$. Una fracción que no tiene divisores comunes —como $\frac{3}{4}$ —, se llama irreducible.

Y... ¿qué es el cociente de dos enteros?, ¿qué es un número racional? Para no recurrir a un lenguaje técnico, es conveniente identificar a los números racionales como las expresiones decimales que se obtienen al dividir el numerador de una fracción por su denominador. En efecto, toda fracción da lugar a una expresión decimal. Para obtenerla, basta dividir el numerador entre el denominador. Algunas fracciones, como $\frac{3}{8}$, tienen una expresión decimal finita o cerrada:

$$
\begin{array}{r|l}
 & 0.375 \\
\hline
8 & 3 \\
 & 30 \\
 & 60 \\
 & 40 \\
 & 0
\end{array}
$$

Por otro lado, toda expresión decimal finita puede escribirse como una fracción. Por ejemplo, $56.3849 = \frac{563\ 849}{10\ 000}$.

Sin embargo, hay fracciones muy simples como $\frac{1}{3}$ o $\frac{1}{10}$ cuyas expresiones decimales son infinitas, aunque ¡atención! Existe una característica que define las expresiones decimales de las fracciones: si no son finitas, entonces son periódicas. Veamos algunos ejemplos de expresiones decimales que se obtienen de fracciones:

$$\frac{1}{3} = 0.33333\ldots$$
$$\frac{1}{9} = 0.11111\ldots$$
$$\frac{3}{7} = 0.428571428571428571\ldots$$

He aquí el cálculo de la última división:

$$
\begin{array}{r|l}
 & 0.428571428571\ldots \\
\hline
7 & 3 \\
 & 30 \\
 & 20 \\
 & 60 \\
 & 40 \\
 & 50 \\
 & 10 \\
 & 30 \\
 & 20 \\
 & 60 \\
 & 40 \\
 & 50 \\
 & 10 \\
 & 3\ldots
\end{array}
$$

Observemos que, a partir de donde el tres aparece como residuo por segunda vez, los dígitos se repiten. La expresión decimal —no finita— de una fracción es periódica cuando se divide un número entre el denominador de la fracción y se obtiene un residuo menor al denominador y, al repetir las divisiones, eventualmente vamos a obtener uno de los residuos que obtuvimos antes —a partir de ese momento, como es lógico, los resultados de la división se repiten. No importa cuán grande sea el denominador, sólo puede haber un número finito de residuos al dividir por él y, por tanto, en algún momento el procedimiento va a repetir el resultado.

Finalmente, vamos a mostrar con un ejemplo cómo puede obtenerse la fracción que corresponde a una expresión decimal periódica. Por ejemplo, si $x = 0.53982539825398253982\ldots$ es el número cuya expansión decimal es periódica, entonces:

$$100\ 000x = 53982.53982539825398253982\ldots$$

Y si hacemos la resta $100\ 000x - x$, obtenemos:

$$
\begin{array}{rl}
 & 100\ 000x = 53982.53982539825398253982\ldots \\
- & \underline{\qquad\quad x = \qquad 0.53982539825398253982\ldots} \\
 & 99\ 999x = 53982
\end{array}
$$

Por ello, al despejar se tiene que:

$$x = \frac{53\ 982}{99\ 999}.$$

Este truco se puede repetir cada vez que la expresión decimal es periódica. Para corroborar lo anterior, hagamos la división:

$$\begin{array}{r|l} & 0.5398253982\ldots \\ \hline 99999 & 53982 \\ & 539820 \\ & 398250 \\ & 982530 \\ & 825390 \\ & 253980 \\ & 539820 \\ & 398250 \\ & 982530 \\ & 825390 \\ & 253980 \\ & 53982\ldots \end{array}$$

Aunque la expresión decimal de un racional sea infinita, la información puede mostrarse con un número finito de cifras al indicar cuál parte se repite de manera interminable con una raya horizontal. Así, se puede escribir:

$$\frac{1}{3} = 0.\overline{3} \quad \text{o} \quad \frac{1}{7} = 0.\overline{142857}.$$

Así, se ve que los números racionales son exactamente aquellos números cuya expresión en decimales resulta periódica. Con los números racionales se puede medir con cualquier exactitud requerida; sin embargo, como se verá en la siguiente sección, hay otros números que se consideran más una necesidad filosófica que práctica.

2.4 NÚMEROS PARA EXPRESAR LO CONTINUO

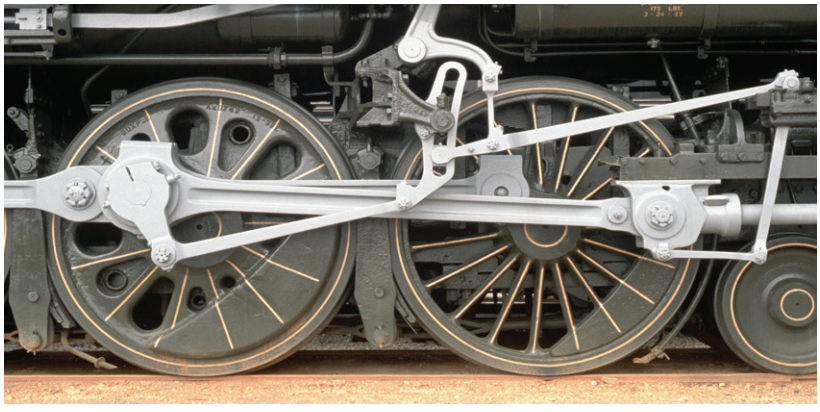


Figura 2.9 En una locomotora, la energía del vapor se traduce continuamente en el movimiento lineal del pistón que, a su vez, genera el movimiento circular en las ruedas y hace avanzar el tren | © Latin Stock México.

Si observamos fenómenos de la naturaleza, es claro que no dan brincos. Lo mismo ocurre en sucesos cotidianos, por ejemplo, si el pistón de un motor se mueve hacia arriba y hacia abajo pasa por todos los estados intermedios. Por lo anterior, conviene buscar un concepto de número que exprese esta continuidad. Estos números se llaman **reales** y se entienden como números decimales con una precisión infinita. En este apartado se verá cómo los matemáticos lograron encontrar dichos números bajo el concepto de continuidad. Vale la pena hacer una advertencia de antemano: estos números son una necesidad filosófica pero no una práctica; las computadoras actuales pueden simular fenómenos continuos de manera

asombrosa, aunque se basan en una precisión limitada. ¿Qué es lo que sucede en una calculadora de bolsillo cuando realizamos la operación $0.4^{0.7}$? Esta y otras preguntas se resolverán más adelante.

Sabemos que no todo número es racional, es decir, una fracción; por ejemplo, $\sqrt{2}$ no es una fracción, como se explicó con anterioridad. Por otro lado, no toda fracción tiene una expresión decimal finita; de hecho, son racionales justamente aquellos números cuya expansión decimal es periódica.

Sin embargo, hay números con expresión decimal infinita como:

$$5.101001000100001000001\ldots$$

donde después de un 1 hay un bloque de ceros, pero cada vez, este bloque tiene un cero de más. Esta expresión no puede ser periódica. Por lo tanto, dado lo que se argumentó con anterioridad, este número —al igual que $\sqrt{2}$ — no es un racional.

Si ponemos todos los números racionales como puntos sobre la **recta numérica**, dejarían huecos y, por eso, no bastan y se requiere rellenar los espacios vacíos entre ellos. Esto puede parecer algo extraño pero los racionales dan una precisión alta, de hecho, arbitrariamente alta. Para tener un ejemplo concreto, veamos cómo se mueve un pistón de un cilindro de una locomotora de vapor.

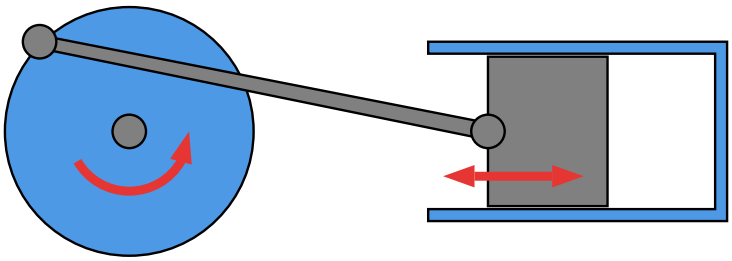


Figura 2.10 Esquema del movimiento de un pistón.

El pistón se encuentra en la parte baja delantera, a la altura de las ruedas, y se mueve horizontalmente por la presión del vapor en el cilindro, como se muestra en la figura 2.10. En cada momento se puede determinar su posición y expresarla con un número —si fijamos bien las unidades de medida y un punto de referencia hacia donde se mide—; por ejemplo, al medir en centímetros la posición del pistón hasta el inicio del cilindro. Es evidente que las mediciones deben arrojar todos los números desde un mínimo hasta un máximo posible. No hay huecos porque el pistón no da brincos. Por ello, los números racionales no bastan; se requieren forzosamente todos los números **reales**: aquellos números con una expansión decimal.

Los números cuya expansión decimal no se vuelve periódica llenan los huecos en la recta numérica. Ésta se debe imaginar como una recta graduada, es decir, con marcas que indican la ubicación de los números.

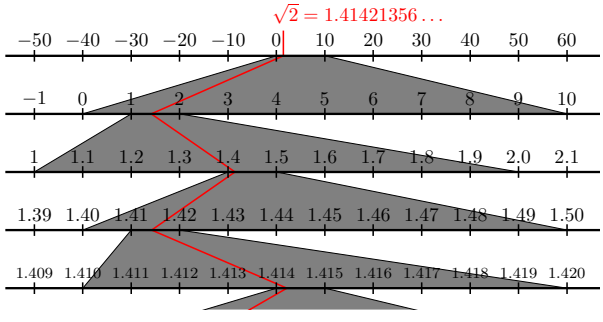


Figura 2.11 Esquema de la recta numérica.

La recta numérica es parecida a una carretera: siempre encontramos las marcas de la distancia hasta un punto de referencia —usualmente el centro de la siguiente ciudad o la anterior—. En cada kilómetro de las carreteras de México, hay un letrero que indica esta distancia. En la recta numérica hay muchos más letreros, uno para cada punto, lo cual significa una infinidad de letreros. Usualmente, se dibuja la recta real de manera horizontal con los números negativos del lado izquierdo y los números positivos del derecho, aunque eso no es realmente importante; también se podría imaginar vertical o curva, o incluso, uniendo la Tierra con la Luna, pues es un concepto abstracto que construimos en nuestra cabeza.

¿Qué tan necesaria es la cola infinita de cifras y cifras para cada número? Realmente, ¿se requiere de tanta precisión? La respuesta honesta es que, para fines prácticos, no se requiere en absoluto; muchas veces bastan algunas cifras decimales. Si dividimos el número 1 entre 3 la calculadora muestra:

$$0.333333333$$

aunque en realidad el número es:

$$0.33333333333333333333333333333333333333333333333333333333\ldots$$

con una cola infinita de tres. La diferencia es tan pequeña que, en la inmensa mayoría de los casos prácticos, no importa. Podemos hacer una comparación: la distancia entre la ciudad de México y Madrid es de $9\,062$ km, aunque esta cantidad tiene un error de 9.062 mm. Es claro por el contexto que los $9\,062$ km también son sólo una aproximación: ¿de dónde a dónde se mide? ¿Del Zócalo a la Plaza Mayor? Pero, ¿de qué punto del Zócalo? ¿Alguna de las esquinas o de la base del mástil? Frente a estas incertidumbres, la diferencia de 9.062 milímetros parece no sólo insignificante, sino errónea para nuestro sentido común.

Todas las calculadoras trabajan con un número finito de dígitos. Para ellas, sólo existen los números racionales y las construcciones abstractas de los números irracionales se aproximan. Por todo lo anterior decimos que los números irracionales son una necesidad filosófica, no una práctica. Los pitagóricos descubrieron que $\sqrt{2}$ no puede ser racional. Este simple hecho basta para demostrar que hay otro tipo de números, además de los racionales.

El siguiente y último argumento para la necesidad de los números reales en matemáticas es de tipo geométrico. Para los griegos, las construcciones con regla y compás eran muy importantes. Sin embargo, si se limitan a puntos en el plano cuyas coordenadas sean racionales, bien puede ser que una construcción dada no tenga intersección, aunque "al ojo" así parece, como lo muestra la figura 2.12.

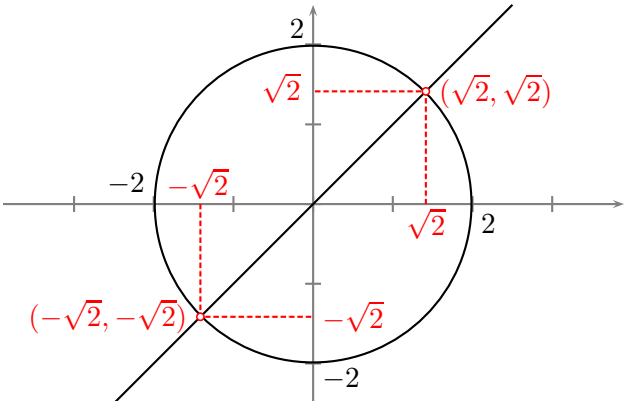


Figura 2.12 Construcción geométrica con regla y compás.

Si se interseca una circunferencia, con centro en el origen y radio 2, con la diagonal entre los dos ejes de coordenadas, entonces, los dos puntos de intersección son $(\sqrt{2}, \sqrt{2})$ y $(-\sqrt{2}, -\sqrt{2})$. Esto quiere decir que los puntos de intersección no tienen coordenadas racionales y, así, ésta es la segunda razón para considerar a los números reales como concepto matemático.

2.5 ¿CÓMO CALCULAR DE MANERA EFICIENTE?

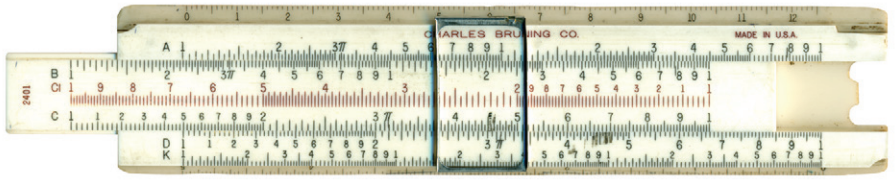


Figura 2.13 La regla de cálculo, tan de moda en la primera mitad del siglo XX, ahora está en desuso por la aparición de las calculadoras de bolsillo. Ambas utilizan la misma técnica para simplificar los cálculos: los logaritmos descubiertos alrededor de 1620.

El principio del siglo XVII es el tiempo de Galilei, de la aceptación del método científico, el tiempo de Tycho Brahe y de Kepler, de los telescopios que permitían una observación más precisa de la bóveda celeste y el tiempo de Blaise Pascal, quien elaboró una de las primeras máquinas que podían sumar y multiplicar. En aquel entonces también aparecen innovaciones importantes al elaborar los relojes de precisión. En resumen, la primera mitad del siglo XVII es una era donde la mecánica fina se vuelve importante para la ciencia y la sociedad.

Sin embargo, la mecánica fina requería también unos cálculos cada vez más precisos. Fue a inicios del siglo XVII cuando Jost Bürgi —matemático y relojero suizo— descubrió los logaritmos al intentar facilitar su trabajo con los cálculos de precisión —que eran bastante tediosos en aquel entonces—, también descubiertos y publicados por John Napier, una década después. Hay muchas historias parecidas en las que se habla de que el tiempo es "maduro" para cierto descubrimiento pues, simultáneamente, varias personas encuentran lo mismo de manera independiente. Hoy en día el mundo reconoce a Napier como el descubridor de los **logaritmos** y, tal vez, ello sea correcto dado que fue él quien divulgó la idea y, así, la puso al servicio de la humanidad.

A continuación veremos en qué consisten estos logaritmos y por qué son tan útiles para facilitar el trabajo con los cálculos. Empezaremos como lo hicieron Bürgi y Napier, al considerar sucesiones algebraicas y geométricas, que se muestran respectivamente a continuación:

$$2, 5, 8, 11, 14, \ldots$$
$$3, 9, 27, 81, \ldots$$

Aquellos que gusten de los retos intelectuales se pueden preguntar cómo sigue cada una de estas sucesiones, es decir, ¿cuál es el siguiente número? Para ello hay que indagar sobre el patrón que rige la sucesión. En nuestro caso no es difícil: en la primera, el aumento de un número a otro es siempre el mismo y cada vez la sucesión crece en tres unidades, mientras que, en la segunda, cada nuevo número se obtiene al multiplicar el anterior por 3.

Veamos la segunda sucesión más de cerca: empieza con 3 y luego este 3 se multiplica repetidamente por 3. Podemos escribir la sucesión de la siguiente manera:

$$3, 3^2, 3^3, 3^4, \ldots$$

En este caso, la simple observación que da lugar al descubrimiento de Bürgi y Napier es el constatar que la sucesión de los exponentes:

$$1, 2, 3, 4, \ldots$$

es una **sucesión algebraica**. El gran problema de los cálculos en aquella época no fue la suma, sino la multiplicación. ¿Cómo multiplicar 3^5 con 3^7? Consideremos que $3^5 = 243$ y $3^7 = 2187$. Ahora, podríamos calcular el producto, pero puede ser provechoso observar que 3^5 es el producto de cinco 3, y 3^7 el producto de siete, por lo que:

$$3^5 \cdot 3^7 = 3^{5+7} = 3^{12}.$$

Hasta aquí parece que no hemos ganado nada, sólo hicimos una adición, sumamos los exponentes. Ahora bien, si pudiéramos representar más números como potencias del 3 podríamos rápidamente hacer más multiplicaciones al efectuar una única suma.

En efecto, lo anterior se hizo pero no con el número base 3 sino con el 10, dado que nuestro sistema es decimal. Aquí se muestran algunos valores:

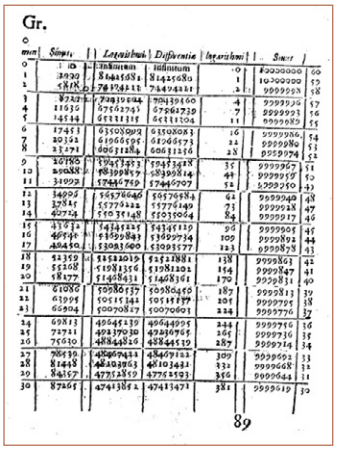
$$\begin{aligned}
1 &= 10^0 \\
2 &= 10^{0.301} \\
3 &= 10^{0.477} \\
4 &= 10^{0.602} \\
5 &= 10^{0.699} \\
6 &= 10^{0.778} \\
7 &= 10^{0.845} \\
8 &= 10^{0.903} \\
9 &= 10^{0.954} \\
10 &= 10^1
\end{aligned}$$

Así, por ejemplo, $2 \cdot 3 = 10^{0.301} \cdot 10^{0.477} = 10^{0.301+0.477} = 10^{0.778} = 6$. ¡Claro! Sabemos multiplicar 2 por 3 más velozmente en nuestra cabeza, pero este simple ejercicio explica el funcionamiento.

En aquel tiempo se empezaron a elaborar libros enteros que contenían tablas de logaritmos, como se muestra en la figura 2.14.

Elevar al cuadrado significa duplicar el exponente: lo vemos si comparamos los exponentes que corresponden a los números 2, 4 y 8, que respectivamente son 0.301, 0.602 y 0.903. De manera similar, obtener la raíz de 5 es ahora fácil: hay que dividir el exponente correspondiente entre 2, lo que es lo mismo que:

$$\sqrt{5} = 10^{0.699 \div 2} = 10^{0.3495}.$$


Gr.

89

Figura 2.14 Fragmento de la primera página de la tabla de logaritmos de Napier.

Con una tabla de logaritmos se puede buscar rápidamente el valor, que es 2.236. Para multiplicar 61.235 por 5.961 se buscan los exponentes correspondientes que son, precisamente, los logaritmos de dichos números en la tabla. Entonces, se encuentra $\log 61.235 = 1.787$ —que equivale a decir que $61.235 = 10^{1.787}$— y $\log 5.961 = 0.775$. Por lo tanto, su producto satisface que:

$$\log(61.235 \cdot 5.961) = 1.787 + 0.775$$

y obtenemos un cálculo sencillo que se puede hacer casi en la cabeza, cuyo resultado es 2.562.

Buscando en las tablas de logaritmos se encuentra que $10^{2.562} = 364.8$, lo cual se acerca bastante bien al resultado correcto, que es $61.235 \cdot 5.961 = 365.021835$. Para tener mayor precisión, se requerían libros cada vez más gruesos con tablas de logaritmos más precisos.

Ya en 1624, Henry Briggs publicó la primera tabla donde reportó los logaritmos de los primeros 20 mil números naturales con una precisión de 14 dígitos decimales. A finales del siglo XVII, se publicaron varios libros con logaritmos en los que se calculaba hasta con 6 dígitos de precisión.

La regla de cálculo que se muestra en la figura 2.13 fue inventada poco después del descubrimiento de los logaritmos. El principio se basa en que ambos lados muestran escalas logarítmicas. Así, con un simple deslizamiento, se podía leer directamente el resultado de una multiplicación hasta, al menos, con un dígito de precisión sin tener que hacer cálculo alguno.

La regla de cálculo desapareció rápidamente con la aparición de las calculadoras de bolsillo, que usan en su arquitectura binaria el principio de los logaritmos y lo combinan con rutinas eficientes para calcular tanto los logaritmos como sus inversos, los exponenciales, para no almacenar tablas grandes de información.

El ejemplo de los logaritmos muestra uno de los grandes logros en el que se aprecia cómo la humanidad usa las matemáticas para simplificar el quehacer de calcular. Al mismo tiempo, muestra que todos estos descubrimientos e inventos están presentes en nuestra cultura de manera oculta, en este caso, en las calculadoras de bolsillo.

2.6 LOS NÚMEROS DE LA COMPUTACIÓN

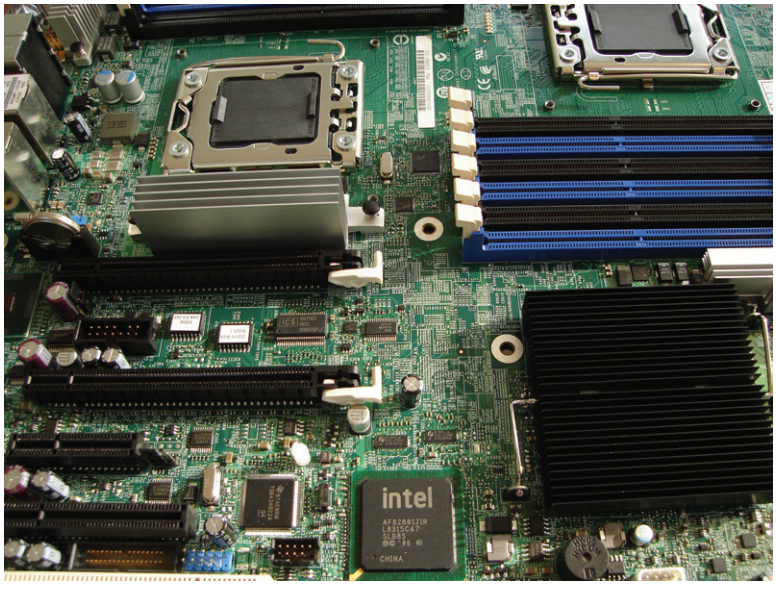


Figura 2.15 Las computadoras han revolucionado nuestra vida de una manera muy determinante. Algunas de sus capacidades son almacenar y procesar grandes cantidades de información. Sorprendentemente, todo lo que sucede dentro de una computadora se realiza con sólo dos estados físicos de algún material que, matemáticamente, se representan como el 0 y el 1. Aquí se muestra la "tarjeta madre" de una computadora moderna.

Las computadoras pueden hacer operaciones numéricas a una altísima velocidad y ésta es una de sus aplicaciones más importantes. Pero, ¿cómo maneja los números una computadora? En esta sección veremos la forma en que las computadoras "almacenan" los números y qué tipo de números pueden utilizar.

Hay distintas representaciones de los números en las computadoras, la mayoría utiliza el **sistema binario** o de base 2, que es el más natural en su medio, pues los sistemas de almacenamiento de información están basados en grandes cantidades de imanes microscópicos, cada uno de los cuales puede estar sólo en dos "estados" o posiciones —digamos, para simplificar, que pueden estar orientados hacia arriba o hacia abajo—. Al guardar y acomodar información en un dispositivo de memoria, se ordenan algunos de esos microimanes hacia arriba y otros hacia abajo. Un sistema tal consta, entonces, de muchos elementos de memoria, cada uno de los cuales, al presentar estos dos estados, puede almacenar lo que se llama un bit de información. Un bit es la información requerida para saber si una cosa es o no es. Repasemos este concepto: supongamos que el "sí" lo representamos con un imán orientado hacia arriba y el "no" con un imán orientado hacia abajo. El "sí" o la orientación hacia arriba, puede a su vez representarse numéricamente con un 1, y el "no" o la orientación hacia abajo, con un 0. Así, por ejemplo, si tenemos ocho de estos microelementos de memoria podemos representar su estado con un conjunto ordenado de ocho dígitos binarios, o sea con ocho números cada uno de los cuales es 0 o 1. Por ejemplo:

$$
\begin{array}{c}
01100101\\
00000001\\
11110000\\
10101010,\\
\text{etcétera,}
\end{array}
$$

son posibles estados de este conjunto ordenado de elementos de memoria o bits. Si el bit tiene un 1, decimos que está encendido, y si tiene un 0, decimos que está apagado. Un conjunto ordenado de ocho bits se llama byte. Los bytes se usan mucho en la computación porque las primeras computadoras se hicieron para operar sobre bytes, que se ha conservado como el elemento mínimo de información que se usa en ellas.

Supongamos ahora que deseamos usar un byte para representar un número. ¿Cómo lo hacemos? ¿Cuántos números distintos podríamos representar con un byte? Lo natural es utilizar el sistema de numeración binario que consiste en asignar al primer bit del byte el valor 1 si está encendido —y 0 si no lo está—; al segundo, el valor 2; al tercero, el $4 = 2^2$; al cuarto, el $8 = 2^3$; al siguiente, el $16 = 2^4$ y así sucesivamente. Recordemos que $2^1 = 2$ y $2^0 = 1$. Por analogía con la notación decimal, consideramos a los bits ordenados de derecha a izquierda. Así, por ejemplo, un byte con la configuración 01100101 representa al número 101, según se muestra a continuación:

$$
\begin{aligned}
01100101 &= 0 \times 2^7 + 1 \times 2^6 + 1 \times 2^5 + 0 \times 2^4 + 0 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0\\
&= 2^6 + 2^5 + 2^2 + 2^0\\
&= 64 + 32 + 4 + 2 = 101
\end{aligned}
$$

Cabe agregar que el 101 es el ciento uno y está en representación decimal y no binaria, es decir, lo interpretamos como:

$$101 = 1 \times 10^2 + 0 \times 10^1 + 1 \times 10^0.$$

Ahora vamos a representar el número 213 en forma binaria como ejercicio. Para ello, hay que comenzar por descomponerlo en una suma de potencias de 2:

Como 213 es mayor que $128 = 2^7$, usaremos 2^7. Nos falta agregar $213 - 128 = 85$, que es mayor que $64 = 2^6$, por lo que también sumaremos 2^6. Nos falta incluir $85 - 64 =$

21, que es menor que $32 = 2^5$ pero mayor que $16 = 2^4$, por lo que no agregaremos 2^5, pero sí 2^4. Finalmente, pongamos lo que nos falta: $21 - 16 = 5 = 2^2 + 2^0$. Por lo tanto, la descomposición del número 213 que buscábamos es $2^7 + 2^6 + 2^4 + 2^2 + 2^0$, que nos lleva a la representación binaria 11010101.

Las computadoras manejan los números enteros en formato binario como el que se acaba de mostrar, excepto que, en general, el primer bit se usa para indicar el signo del número —es decir, si el primer bit está "apagado" el número es positivo, si en cambio está "encendido", el número es negativo. Uno de los formatos más utilizados en las computadoras para los números enteros es de 16 bits, llamado *short* en el lenguaje Java. En un entero de tipo *short*, el primer bit se utiliza para el signo y los 15 restantes para el valor absoluto del número. Así, por ejemplo, la expresión binaria:

10000000 11010101

representa el número -213. Con 16 bits se pueden representar números enteros entre $-(2^{15} - 1) = -32\ 767$ y $2^{15} - 1 = 32\ 767$. En efecto, con 15 bits, el número más grande que puede representarse es aquel en el que todos los bits están encendidos, por lo tanto es $2^{14} + 2^{13} + \ldots + 2^1 + 2^0 = 2^{15} - 1 = 32\ 767$. Para los negativos se puede aprovechar un número más pues el cero está representado entre los no negativos y no hace falta repetirlo, pero no explicaremos la razón de este detalle. La forma más usada para expresar enteros en lenguaje Java es con el tipo *int* que usa 32 bits, con los cuales se pueden representar los enteros entre $-2^{31} = -2\ 147\ 483\ 648$ y $2^{31} - 1 = -2\ 147\ 483\ 647$.

La representación de los números reales es bastante más complicada. En lenguaje Java se emplea la norma IEEE 754—estándar internacionalmente aceptado y usado casi universalmente—. En esta norma hay dos formatos de datos, llamados *float* y *double*, que son los más utilizados. Ambos son "representaciones de punto flotante"; el primero se denomina de precisión simple y usa 32 bits, mientras que el segundo, que es de doble precisión, usa 64 bits. También hay uno de cuádruple precisión que utiliza 128 bits. Aunque el más usado es el de 64 bits, para simplificar aquí la presentación, describiremos solamente el de 32 bits.

De los 32 bits que se usan para representar un número real del tipo *float*, el primero se usa para el **signo**, los ocho siguientes para el **exponente** y los 23 restantes para la **mantisa,** como se muestra a continuación al representar el número -231.125:

$$\underbrace{1}_{\text{signo}}\ \underbrace{100001100}_{\text{exponente}}\ \underbrace{11001110010000000000000}_{\text{mantisa}}$$

La fórmula para obtener el número en cuestión a partir de la representación consiste en, primero, obtener los valores enteros positivos del *signo*, el *exponente* y la *mantisa* a partir de sus representaciones decimales. Así:

$$\begin{aligned} s &= 1 \\ x &= 2^7 + 2^2 + 2^1 = 134 \\ m &= 2^{22} + 2^{21} + 2^{18} + 2^{17} + 2^{16} + 2^{13} = 6758400 \end{aligned}$$

Y luego, al aplicar la fórmula:

$$\text{número} = (1 - 2s) \times \left(1 + \frac{m}{2^{23}}\right) \times 2^{x-127}$$

que, en este caso, nos lleva al resultado:

$$\begin{aligned}\text{número} &= (1-2)\times\left(1+\frac{6758400}{2^{23}}\right)\times 2^{134-127}\\ &= -1.8056640625\times 2^{7}\\ &= -231.125\end{aligned}$$

Estas representaciones de punto flotante están diseñadas no sólo para guardar los números dentro de las computadoras, sino para que éstas puedan realizar las operaciones básicas con gran eficiencia. Los procesadores numéricos de las computadoras son los que se encargan de realizar dichas operaciones. Afortunadamente, las computadoras hacen todo esto con mucha facilidad y rapidez. Lo importante es que tengamos conciencia de que la representación de números reales dentro de las computadoras dista mucho de ser equivalente al concepto matemático de número real; más bien, se limita a utilizar sólo algunos números —todos racionales con expansión binaria finita— y a brindarnos buenas aproximaciones de resultados de los cálculos que les pedimos hacer. Por ejemplo, el número π representado en formato de doble precisión es equivalente al valor decimal:

$$3.141592653589793$$

mientras que sabemos que el verdadero valor de π tiene una representación decimal infinita.

2.7 MEDIR LO INALCANZABLE

Figura 2.16 La imagen muestra un sextante, herramienta para medir la latitud en alta mar usando el Sol. Casi desde que aparecieron sobre la Tierra, los seres humanos han extendido su percepción y medición más allá de lo que directamente alcanzan al usar, por un lado, herramientas sofisticadas como el sextante y, por otro, el razonamiento matemático, como se descubrirá en esta sección | © Latin Stock México.

Medir distancias pequeñas es muy fácil. Basta una cinta métrica, y si con ella no es suficiente, se pueden poner marcas, hacer varias mediciones en serie y luego sumar. También, usando geometría básica se pueden medir distancias enormes o inaccesibles; aquí se hablará de estos métodos para medir distancias. El principio fundamental en que se basan es el de la semejanza de triángulos. Dos triángulos son **semejantes** si sus tres ángulos coinciden. Tienen entonces la misma forma y hay una constante que relaciona los lados de uno con los del otro, o bien, las proporciones entre lados correspondientes son iguales. Así que con pocos datos de un triángulo grande podemos obtener los demás.

Veamos primero un ejemplo de un método que usan los maestros de obras para aproximar distancias. Supongamos que tenemos el brazo extendido hacia enfrente y el pulgar levantado. Al cerrar un ojo y luego el otro, el dedo parece "saltar" o "brincar" de lugar en el fondo, que llamaremos "pared". Lo que pasa es que cada ojo "proyecta" al pulgar en un punto distinto de la pared y, entonces, se forman dos triángulos que comparten el vértice en el dedo. Se arma un triángulo chico con los dos ojos, y el otro, grande, con los dos puntos de la pared donde se proyecta el pulgar. Si estamos frente a la pared, estos dos triángulos son **isósceles** y semejantes —en el vértice del pulgar tienen el mismo ángulo.

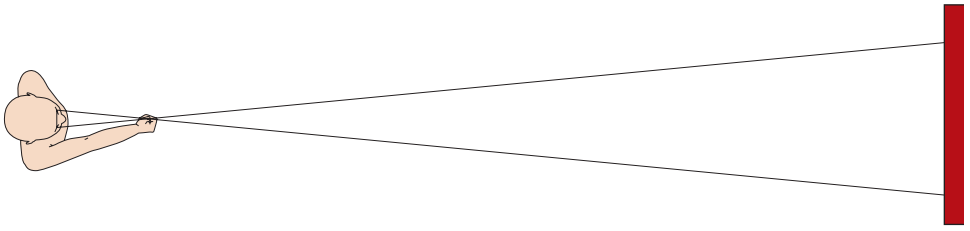


Figura 2.17 Esquema de los triángulos isósceles formados al usar el pulgar como vértice.

Como en el triángulo chico la proporción de la altura a la base —del brazo extendido a la distancia entre los ojos— es, aproximadamente, de 10 a 1, en el triángulo grande se cumple la misma relación. De aquí que, si sabemos que la distancia a la pared es, más o menos, de treinta metros, entonces el "brinco" del dedo en la pared es de aproximadamente tres metros. O bien, si el dedo brincó lo que mide un coche pequeño —alrededor de cuatro metros—, éste debe estar como a una distancia de cuarenta metros.

Así se pueden medir distancias inaccesibles con semejanza de triángulos. Es sorprendente, pero fueron ideas igual de sencillas las que permitieron dar una primera estimación del tamaño de la Tierra y de la Luna. La primera estimación del diámetro de la Tierra la hicieron los griegos. Eratóstenes notó que, en la ciudad de Asuán —en Egipto—, la luz del Sol entraba de lleno a los pozos al mediodía del solsticio de verano —junio 21—, cuando las sombras llegan a su mínima longitud en el hemisferio norte. Esto sucede porque dicha ciudad está casi en el Trópico de Cáncer, que es el paralelo más al norte donde la luz del Sol puede caer o incidir verticalmente —literalmente "a plomo" — y lo hace justo en el solsticio de verano.

El cálculo de Eratóstenes se basó en medir el ángulo con el que inciden los rayos del Sol al mediodía del solsticio de verano en Alejandría, que está al norte de Asuán. Este ángulo resultó ser de $\frac{1}{50}$ de la vuelta completa —2π radianes o $360°$—. Así que, al multiplicar por 50 la distancia entre estas dos ciudades, se obtiene una aproximación de la circunferencia de la Tierra y, por consiguiente, su diámetro al dividir entre una aproximación de π.

Lo impresionante es que, con los métodos para medir ángulos y distancias de aquella época, el error en el cálculo fuera pequeño. No se tiene certeza del cálculo preciso de Eratóstenes pues, en sus escritos, la unidad de medida de longitud que usó fueron los **estadios**, y en la actualidad persiste la discusión histórica de a cuánto equivalen. Eratóstenes consideró la distancia de Asuán a Alejandría de $5\ 000$ estadios. De aquí, la circunferencia de la Tierra resulta de $5\ 000 \times 50 = 250\ 000$ estadios. Con el valor máximo que se tiene de un estadio, que es de $196\,\text{m}$, obtenemos un total de $49\ 000\,\text{km}$, y con el mínimo, $157\,\text{m}$, serían $39\ 250\,\text{km}$. El valor medio de la circunferencia de la Tierra que se estima hoy día es $40\ 000\,\text{km}$, así que Eratóstenes andaba muy cerca.

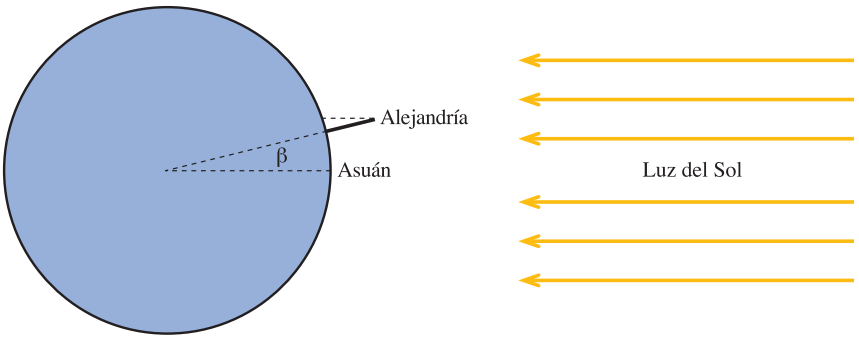


Figura 2.18 Método de Eratóstenes para medir el radio de la Tierra.

Para Eratóstenes, medir la circunferencia de la Tierra era un reto intelectual, "ciencia pura". Casi dos mil años después, cuando Colón planeaba su viaje hacia el oriente navegando en dirección opuesta, el mismo asunto se convirtió en cuestión de vida o muerte —pues la cantidad de víveres que necesitaba para la travesía dependía de la distancia a recorrer—. Por suerte, el cálculo de Colón era erróneo, pues pensaba que la Tierra era más pequeña de lo que en realidad es y se lanzó a la famosa aventura, aunque otro error canceló al primero: se le atravesó un continente insospechado en el camino y los víveres le alcanzaron, aunque él siempre creyó que había llegado a su destino.

Otras mediciones astronómicas que hicieron los griegos fueron la de la distancia de la Tierra a la Luna y el tamaño de la Luna. Tuvieron ayuda de una enorme coincidencia: la Luna y el Sol tienen —en apariencia— casi el mismo tamaño, es decir, el ángulo de nuestro ojo a los dos bordes del Sol o a los dos bordes de la Luna es aproximadamente el mismo. Y este hecho se corrobora en los eclipses solares cuando la Luna se interpone entre el Sol y nosotros.

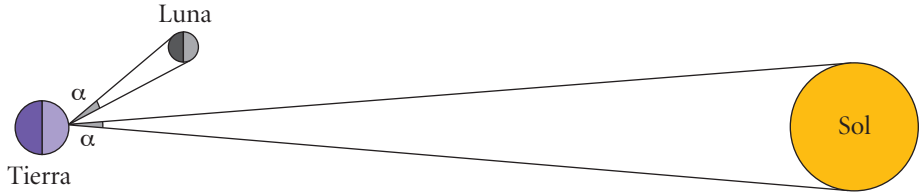


Figura 2.19 El disco del Sol y el disco de la Luna son en apariencia del mismo tamaño, vistos desde la Tierra, como se demuestra en los eclipses solares.

Además, Aristarco de Samos, para medir la distancia a la Luna, se basó en los datos de los otros eclipses: los lunares. En ellos, la Luna entra en el cono de sombra que produce la Tierra.

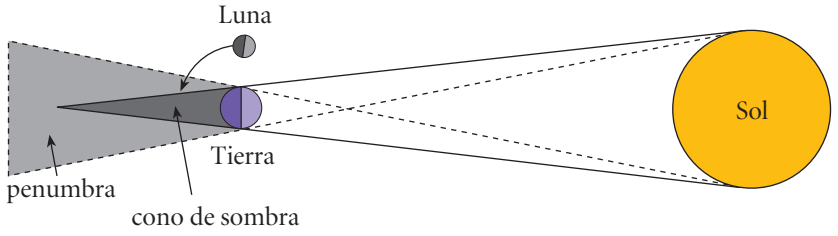


Figura 2.20 Esquema de un eclipse lunar en el que no se mantienen las proporciones.

El cono de sombra se crea porque el Sol se ve como un disco en el cielo y entonces la luz que de él nos llega tiene pequeñas variaciones en el ángulo. El cono de sombra es donde toda su luz queda bloqueada. Se puede observar el cono de sombra que produce un dedo en un día soleado alejándolo del piso hasta una altura de un metro o más. Muy cerca del suelo

se forma una sombra con bordes bien definidos. Pero a medida que lo alejamos, sus bordes se vuelven difusos o desenfocados: son la penumbra, donde si bien parte de la luz que viene del Sol se bloquea, algo de ella pasa. El cono de sombra del dedo, o de una moneda, es "semejante" al cono de sombra de la Tierra, y entonces lo podemos medir.

Los griegos calcularon en forma experimental que el cono de sombra del Sol tiene una proporción aproximada de altura a base de 108 a 1, es decir, la longitud del cono de sombra es aproximadamente 108 veces el diámetro de la Tierra. De aquí, considerando que Sol y Luna tienen el mismo diámetro aparente, se obtiene que si d_{Luna} denota al diámetro de la Luna, entonces:

$$x = 108 \cdot d_{\text{Luna}} \tag{3}$$

es la distancia de la Tierra a la Luna.

En un eclipse lunar, la Luna entra primero a la zona de penumbra y luego al cono de sombra de la Tierra. Por el tiempo que tarda la Luna en cruzar este cono, se puede estimar que la Luna cabe más o menos 2.5 veces en el cono de sombra, es decir, en el lugar donde la Luna cruza al cono, éste mide 2.5 veces el diámetro de la Luna. Con ello ya juntamos las mediciones necesarias para poder determinar el diámetro de la Luna y la distancia que tiene de nosotros; cabe remarcar que todas estas mediciones se obtuvieron a partir de observaciones realizadas desde la Tierra.

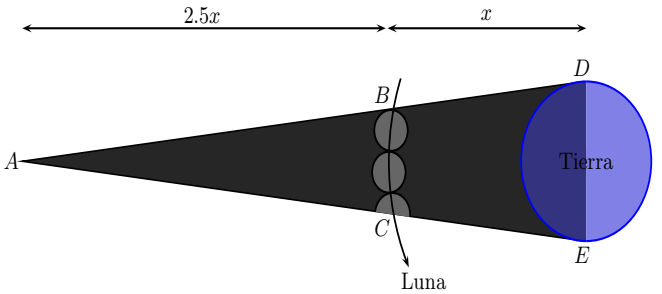


Figura 2.21 Esquema de un eclipse lunar.

En la figura 2.21 tenemos dos triángulos isósceles semejantes: el cono de sombra ΔADE es semejante al triángulo ΔABC. Por otro lado, el triángulo ΔABC tiene base $2.5 \cdot d_{\text{Luna}}$ y, por ello, la altura —horizontal en este caso— del triángulo ΔABC es $2.5x$. Por lo tanto, la altura del triángulo ΔADE es $2.5x + x = 3.5x$. Como sabíamos que esta altura es 108 veces el diámetro de la Tierra, entonces:

$$3.5x = 108 \cdot d_{\text{Tierra}},$$

de donde podemos despejar x, que es la distancia del centro de la Tierra al de la Luna:

$$x = \frac{108}{3.5} \cdot d_{\text{Tierra}}.$$

También obtenemos que el diámetro de la Luna es $d_{\text{Luna}} = \frac{1}{3.5} d_{\text{Tierra}}$, por (3).

Es claro que en los métodos que acabamos de describir hay una considerable posibilidad de error en las mediciones. Los estimados básicos — 108, 2.5 y que los diámetros aparentes de Luna y Sol coinciden—, vienen de Aristarco. Si en vez del diámetro de la Tierra que él usó, empleamos el que se estima en la actualidad —de $d_{\text{Tierra}} = 13\ 000\,\text{km}$—, nuestras fórmulas darían $401\ 142\,\text{km}$ para la distancia de la Tierra a la Luna, y $3\ 714\,\text{km}$ para el diámetro lunar. Los estimados actuales son $384\ 403\,\text{km}$ y $3\ 474\,\text{km}$, respectivamente, que dan errores del 4% y el 6% para el método de Aristarco. Esto demuestra el poder de la geo-

metría euclidiana elemental y la increíble precisión con que los griegos hicieron sus mediciones, así como el enorme poder del razonamiento abstracto.

2.8 LA MEDICIÓN DE LA TIERRA

Figura 2.22 Un billete de diez marcos alemanes —antes de la introducción del euro— muestra a Carl Friedrich Gauss, matemático alemán. Gauss fue tal vez el matemático más importante de todos los tiempos; a él se deben numerosos desarrollos en las matemáticas que hizo para fines prácticos. Por ejemplo, inventó el heliógrafo, un instrumento para aumentar la precisión en las mediciones que dirigió en el país de Hannover.

A principios del siglo XIX no se sabía cuál era la montaña más alta del planeta ni se conocía el tamaño de la Tierra con precisión. La medida conocida como **metro** la establecieron los franceses con la idea de que fuera la $10\,000$-ésima parte de la distancia del Ecuador al Polo Norte. Casi al mismo tiempo, los ingleses iniciaron el *Great Trigonometrical Survey* —gran proyecto de topografía trigonométrica— para medir con precisión toda la India, colonia británica en aquel entonces, pues se pretendía tener una visión más realista del territorio que ocupaba el Imperio británico en Asia y, a la vez, tener una medida más precisa del tamaño de la Tierra. Este proyecto consumió grandes recursos, duró casi todo el siglo XIX y permitió establecer la altura del monte Everest, nombrado así en honor al coronel que estuvo a cargo del proyecto —aunque finalmente ya no vio la montaña pues se quedó ciego—. Fue hasta ese momento que se supo que el Everest, a $8\,850$ metros sobre el nivel del mar, era más alto que los Andes.

En esta sección veremos los principios básicos del proceso para medir la superficie terrestre conocido como "triangulación" —basado en la medición y el cálculo de triángulos, por ejemplo, entre tres picos de montañas. Claro que estos triángulos, en general, no serán ni **rectángulos** —que tienen un ángulo de $90°$— ni isósceles —con dos o tres lados iguales—, sino que son triángulos arbitrarios. Para aclarar la relación que existe entre los lados y los ángulos, tenemos que empezar primero con triángulos rectángulos.

En un triángulo rectángulo, a los dos lados que forman el ángulo recto se les llama **catetos**, y al lado opuesto, **hipotenusa**. Los ángulos de un triángulo rectángulo quedan determinados por cualquiera de sus ángulos no rectos, pues entonces, la medida del otro ángulo corresponde a lo que falta para $90°$ —recordemos que los tres ángulos internos de un triángulo suman $180°$—. Si elegimos uno de los ángulos no rectos y lo llamamos α, podemos diferenciar a los dos catetos. Al lado que forma α con la hipotenusa, se le llama **cateto adyacente** y se le denota CA; al lado opuesto al ángulo α, **cateto opuesto** y se le denota CO y, finalmente, a la hipotenusa con H.

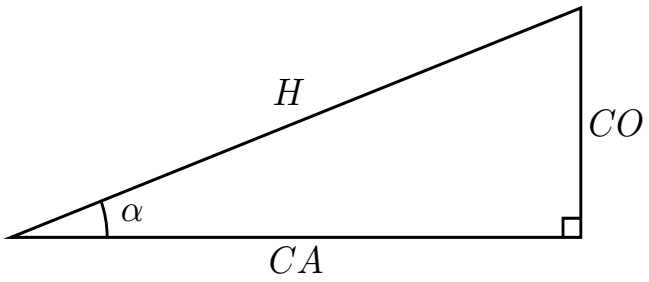


Figura 2.23 Triángulo rectángulo.

Las **funciones trigonométricas** son las proporciones entre los lados de los triángulos rectángulos. Las básicas son el **coseno** y el **seno**, definidas como:

$$\cos\alpha = \frac{CA}{H}, \operatorname{sen}\alpha = \frac{CO}{H}.$$

Observemos que cuando la hipotenusa mide 1 —que es igual a decir que $H = 1$—, el coseno y el seno son, precisamente, lo que miden los catetos. Así, podemos pensar que son las coordenadas cartesianas de un punto en el círculo unitario, es decir, el círculo de radio 1 con centro en el origen. En esta manera de ver al coseno y al seno, los puntos del círculo unitario quedan parametrizados por el ángulo con el eje de las x, como se muestra en la figura 2.24.

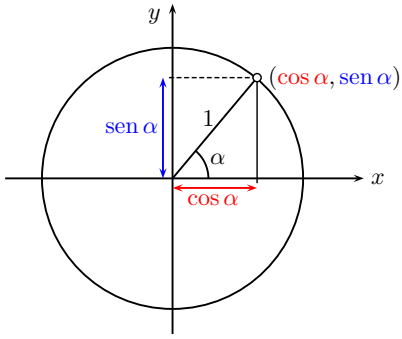


Figura 2.24 Las coordenadas de un punto en el círculo unitario son el coseno y el seno del ángulo correspondiente.

Al pensar en el seno y el coseno como coordenadas es natural extender su definición a cualquier ángulo con los signos correspondientes a los de los cuatro cuadrantes.

Planteemos ahora un problema más complicado: medir la altura de una montaña sin subir en ella. No podemos usar triángulos rectángulos pues no podemos acceder al interior de la montaña, pero sí se pueden hacer mediciones suficientes desde afuera para determinar un triángulo del que se conozcan dos de sus ángulos y un lado. Luego, utilizando la llamada **ley de los senos**, que veremos a continuación, se pueden calcular los otros dos lados.

En un triángulo cualquiera, tracemos una de sus alturas y llamémosla h; a los ángulos opuestos a ella, denotémoslos α y β, mientras que los lados opuestos —en el triángulo original—, serán a y b, respectivamente.

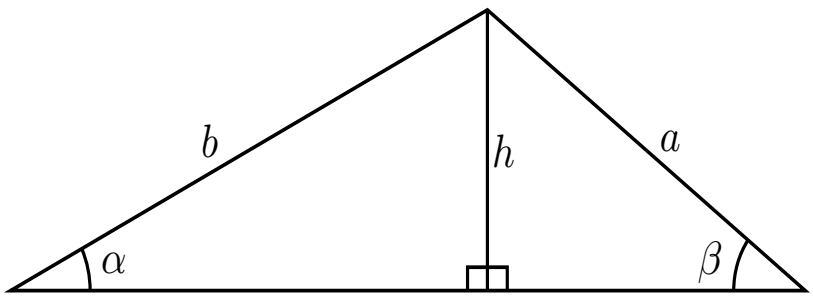


Figura 2.25 Un triángulo con una altura h.

El seno de α se puede expresar como $\operatorname{sen}\alpha = \frac{h}{b}$ de donde podemos despejar h:

$$h = b\operatorname{sen}\alpha.$$

De manera análoga, obtenemos que $h = a\operatorname{sen}\beta$. Hemos descrito a h de dos maneras diferentes, lo que nos da la igualdad: $b\operatorname{sen}\alpha = a\operatorname{sen}\beta$, que también se puede escribir como:

$$\frac{\operatorname{sen}\alpha}{a} = \frac{\operatorname{sen}\beta}{b}.$$

Esta ecuación es el principio básico de la ley de los senos, aunque la manera más común de presentarla es igualando los inversos e incluyendo al tercer ángulo γ y a su lado opuesto, c. Para hacerlo, hay que realizar el mismo razonamiento con alguna de las otras dos alturas. De tal manera, para cualquier triángulo se cumple el que la proporción de los lados con respecto a los senos de sus ángulos opuestos sea la misma. Al expresar la ley de los senos en forma de ecuaciones tendríamos:

$$\frac{a}{\operatorname{sen}\alpha} = \frac{b}{\operatorname{sen}\beta} = \frac{c}{\operatorname{sen}\gamma}.$$

Regresemos ahora al problema de la montaña para aplicar la ley de los senos. Llamemos C a un punto del pico visible. Desde dos puntos A y B en la falda, se pueden medir los ángulos correspondientes α y β del triángulo ΔABC. Supongamos que también conocemos la distancia c entre A y B.

Por la ley de los senos, se obtienen las distancias a y b:

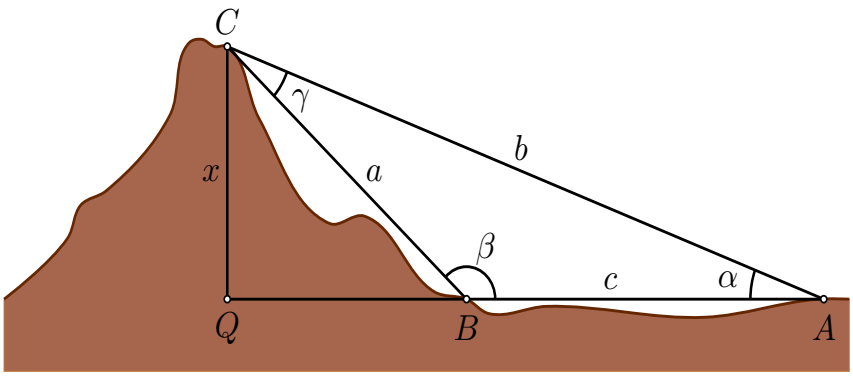


Figura 2.26 Triángulos para medir la altura de una montaña.

$$a = \frac{\operatorname{sen}\alpha}{\operatorname{sen}\gamma}\cdot c \quad \text{y} \quad b = \frac{\operatorname{sen}\beta}{\operatorname{sen}\gamma}\cdot c$$

donde γ es el ángulo en C, que se obtiene a partir de α y β, pues los tres suman $180°$. Conociendo b —y suponiendo que A y B están al mismo nivel—, podemos calcular la altura de C usando la definición del seno:

$$x = b\operatorname{sen}\alpha.$$

Con las mismas ideas que hemos considerado en esta sección, funciona el método de triangulación: primero, se mide una longitud con mucha precisión llamada "base"; después se miden los ángulos hacia un punto muy visible. Al usar la ley de los senos se pueden calcular las distancias a los extremos de la base. Como las herramientas de medición permiten medir el ángulo de elevación por separado del ángulo de giro, también es posible calcular la

altura del punto visible. Después, el equipo se transporta a este punto y divisa la base y otros puntos nuevos. Así, la medición avanza por el terreno midiendo la forma de manera precisa. Gauss condujo así la medición del Hannover, y a finales del siglo XVIII, con este método, se determinó que el Popocatépetl no era la montaña más alta de México, sino el Pico de Orizaba.

2.9 LA PIRÁMIDE TRUNCADA

Figura 2.27 Las pirámides de Giza son unos de los monumentos más impresionantes de las civilizaciones de la Antigüedad. Su forma geométrica reúne tanto estabilidad como belleza | © Latin Stock México.

El ámbito de la producción industrial a principios del siglo XIX demandaba herramientas de medición cada vez más sofisticadas. Entre los hombres que trataron de cumplir esta demanda se encontraba el matemático suizo Jakob Amsler-Laffon, quien fundó su propia fábrica de instrumentos de medición. Uno de sus más apreciados inventos fue el "integrador": un arreglo de varas, ruedas y puntas que medía cualquier área como, por ejemplo, la sección transversal de un riel de hierro mostrada en la figura 2.28. Para medir el área se tenía que colocar el integrador sobre la hoja de dibujo y, luego, pasar una punta por el borde de la figura. Una rueda integrada al aparato registraba continuamente el área, que podía leerse al final sin problemas.

Antes de esta invención era sumamente difícil obtener una buena aproximación para el área de una figura así de compleja. Los instrumentos de Amsler se basaban de manera crucial en desarrollos matemáticos, entre ellos, el cálculo diferencial e integral. La importancia de estas herramientas no debe subestimarse: fueron fundamentales para la industria, en particular, la del acero —una buena aproximación del área de la sección transversal del riel proporciona una estimación certera del costo del material, según su peso por metro.

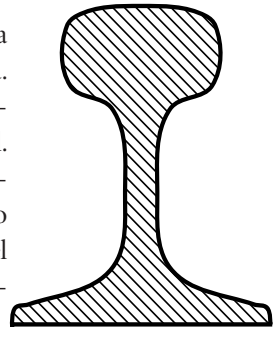


Figura 2.28 Dibujo de una sección transversal de un riel.

Es interesante ver que fue, justamente, la insistencia para calcular áreas y volúmenes de figuras —como la circunferencia— o cuerpos —como la esfera o la pirámide— lo que permitió desarrollar, poco a poco, una teoría más general que finalmente aportó herramientas matemáticas aplicables a situaciones de la vida cotidiana, como el cálculo del área de un riel

que ya hemos revisado. En esta sección y en las tres subsecuentes, se exponen varios de estos cálculos con todo detalle pues, así, se logra comprender mejor las bases del cálculo integral.

Tratar de calcular el área de figuras y el volumen de cuerpos ha llamado la atención a muchas personas desde la Antigüedad. El ejemplo más famoso es, sin duda, el del área y la circunferencia de un círculo. Después de revisarlo en esta sección, sabremos cómo se puede calcular el volumen y la superficie de otros cuerpos geométricos y llegaremos a ver los fundamentos del cálculo integral en la sección 2.12.

En particular, en esta sección se verán diferentes fórmulas que expresan los volúmenes de cuerpos como prismas y pirámides. Estos cuerpos geométricos sencillos están delimitados por **polígonos** o figuras planas que, a la vez, están delimitados por segmentos rectos. Después, se usará esta información para determinar el volumen de una **pirámide truncada** —a la cual se le cortó un pedazo de la punta—. Además, se pondrá un énfasis particular en la lectura de tales fórmulas para explicar cómo las propiedades de éstas se reflejan en propiedades geométricas.

Todos sabemos calcular el volumen de una caja con lados a, b y c: $V = abc$, es decir, el producto de los tres lados que terminan en el vértice de la caja.

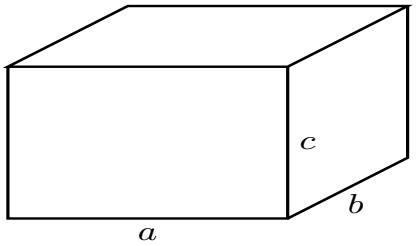


Figura 2.29 Una caja que, en matemáticas, tiene el nombre rimbombante de paralelepípedo rectangular.

La caja tiene un nombre matemático horrible, se llama **paralelepípedo rectangular** y es un caso particular de un **prisma**. Los prismas se obtienen al trasladar un polígono en una dirección, como se muestra en el siguiente dibujo.

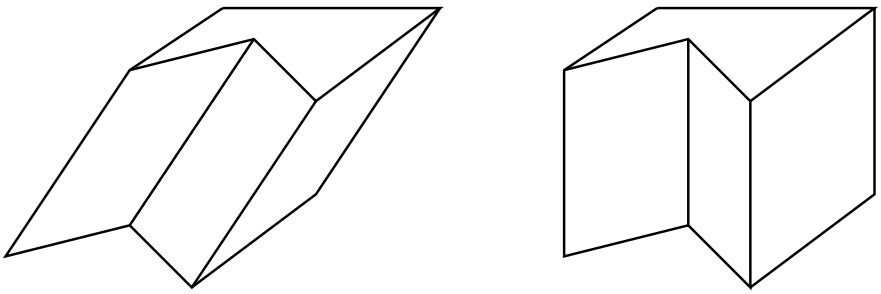


Figura 2.30 Dos prismas con bases congruentes y alturas iguales.

Del lado derecho vemos un prisma vertical. La fórmula del volumen de un prisma es sencilla y se parece mucho a la del rectángulo:

$$\text{volumen de prisma} = \text{base} \times \text{altura}$$

sólo que aquí, "base" significa el área que se traslada y la "altura" es la distancia entre la tapa superior y la inferior. Cuidado: si al generar el prisma la base no se trasladó en dirección vertical, la altura no será la longitud de lo que se trasladó, sino menos.

Se debe observar que la dirección del traslado no importa; siempre y cuando se tenga la misma altura, se tendrán los mismos volúmenes. Esto se debe al **principio de Cavalieri**: dos cuerpos que tienen cortes de áreas iguales con cada plano paralelo a la superficie sobre las que están puestos, poseen volúmenes iguales.

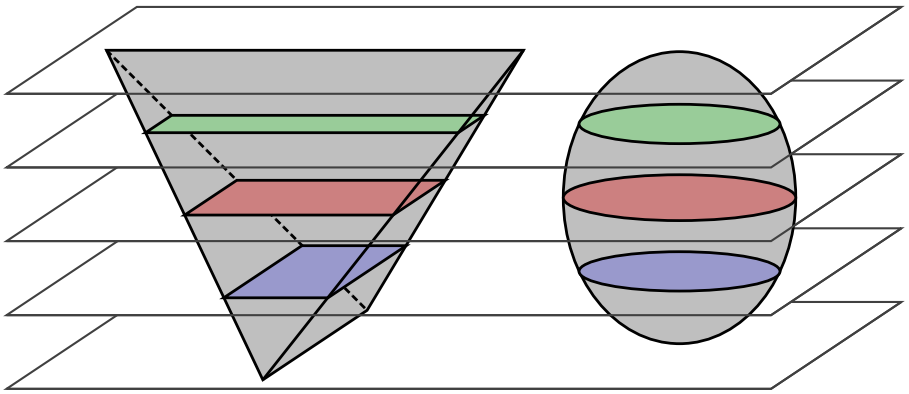


Figura 2.31 Ilustración del principio de Cavalieri: los dos cuerpos tienen el mismo volumen dado que sus áreas de intersección con cada plano paralelo son iguales.

¿Cuál será entonces el volumen de una pirámide? El cuerpo se construye uniendo cada punto de la base con un punto, que es la punta de la pirámide.

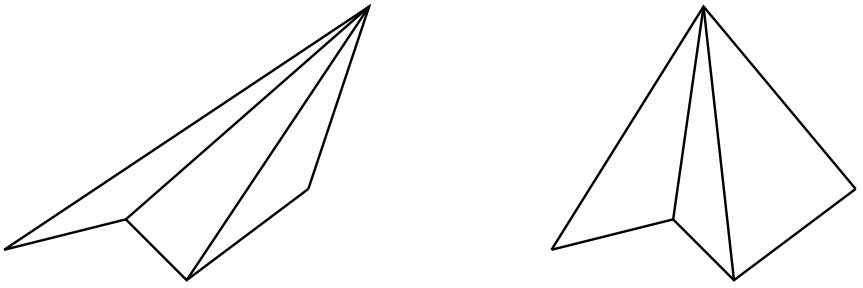


Figura 2.32 Dos pirámides con bases congruentes y alturas iguales.

La pirámide se relaciona con el prisma como lo hace el triángulo con el paralelogramo. Se habla de una **analogía**, es decir, una similitud entre dos relaciones en diferentes ámbitos. Las analogías son un fuerte motor de ideas, no sólo en matemáticas, sino también en el lenguaje, por ejemplo, donde se usan para aclarar una cierta relación.

Por esta analogía, podríamos pensar que el volumen de una pirámide se calcula de manera parecida a como se obtiene el área de un triángulo: base por altura entre dos. Sin embargo, con las analogías hay que tener cuidado pues no siempre todo se traduce uno a uno de un lado al otro, la relación es usualmente más complicada.

Cada prisma —o pirámide— se puede dividir en prismas —o pirámides, respectivamente— con bases triangulares y la misma altura, como se observa en la figura 2.33.

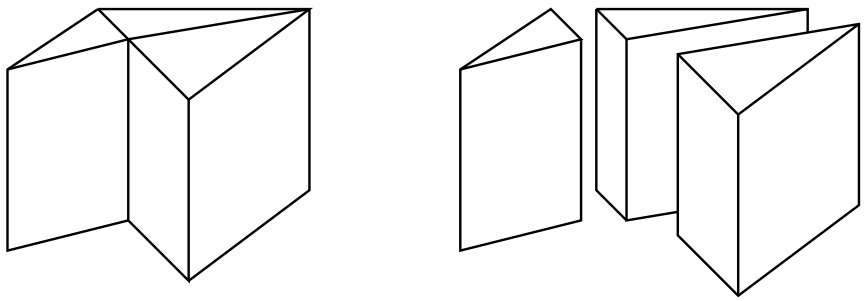


Figura 2.33 Partición de un prisma cualquiera en prismas triangulares.

Por lo anterior, basta entender cómo se calcula el volumen de una pirámide con base triangular. Es fácil colocar dos pirámides con la misma base y altura en un prisma que también tiene la misma base y altura, como lo muestra la siguiente figura.

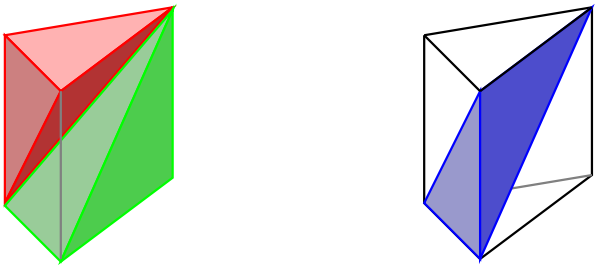


Figura 2.34 Partición de un prisma triangular en tres pirámides triangulares.

Como se observa, dos pirámides no son suficientes para llenar un prisma, así que la primera idea crecida sobre la tierra de la analogía no puede ser correcta. No obstante, el espacio sobrante es, a la vez, una pirámide —la base puede ser, por ejemplo, el costado vertical derecho y la altura la vemos entonces acostada en dirección horizontal.

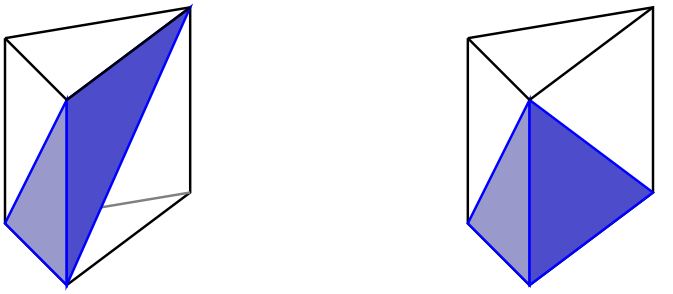


Figura 2.35 Aplicación del principio de Cavalieri.

Por el mismo principio de Cavalieri, podemos mover hacia abajo la punta, que se encuentra arriba en la tapa, ya que no modificamos ni la base ni la altura. Aquí con "la base de la pirámide" nos referimos a la cara azul claro de la figura 2.35 y con "la altura" a la perpendicular —recordemos que ninguna de estas medidas depende de la dirección—. Obtenemos, entonces, que el resto también tiene el mismo volumen que buscamos. En resumen: dentro del prisma caben tres pirámides y cada una tiene el volumen de una pirámide con la misma altura y base que el prisma. Concluimos entonces que:

$$\text{volumen de pirámide} = \frac{\text{base} \times \text{altura}}{3} \quad .$$

Ahora, podemos revisar la analogía de nuevo y descubrimos que sí hay una relación muy asombrosa: el 2 en el denominador se cambió a un 3 al pasar del triángulo a la pirámide. Estos números se explican en términos de la dimensión de las figuras: el triángulo es plano y tiene dimensión dos, mientras que la pirámide tiene volumen y, por ello, dimensión tres.

Recordemos que se dividió la base de cualquier pirámide en triángulos. Ahora queremos ver si podemos deducir la fórmula del volumen de cualquier pirámide. En efecto, si la base original B se dividió en t partes triangulares —$B_1, B_2, \ldots, B_t$— entonces se tiene que:

$$B = B_1 + B_2 + \ldots + B_t,$$

lo que expresa que el área total de la base se obtiene al sumar las áreas de cada uno de los t triángulos. En forma similar:

$$V = V_1 + V_2 + \ldots + V_t.$$

Ahora podemos sustituir cada uno de los volúmenes $V_1, V_2, \ldots, V_t$ por $\frac{B_1 \cdot h}{3}, \frac{B_2 \cdot h}{3}, \ldots, \frac{B_t \cdot h}{3}$. Así, se obtiene que:

$$\begin{aligned} V &= \frac{B_1 \cdot h}{3} + \frac{B_2 \cdot h}{3} + \ldots + \frac{B_t \cdot h}{3} \\ &= \frac{(B_1 + B_2 + \ldots + B_t) \cdot h}{3} \\ &= \frac{B \cdot h}{3} \end{aligned}$$

al usar la factorización en la segunda ecuación. En resumen, la fórmula que obtuvimos para pirámides con base triangular vale, en general, para cualquier pirámide.

Ahora se tienen ya conocimientos suficientes para atacar la fórmula del volumen de una pirámide truncada. Antes de entrar en cálculos y argumentaciones involucradas, observemos que la fórmula que buscamos debe "contener" la del prisma y la pirámide, como "casos particulares".

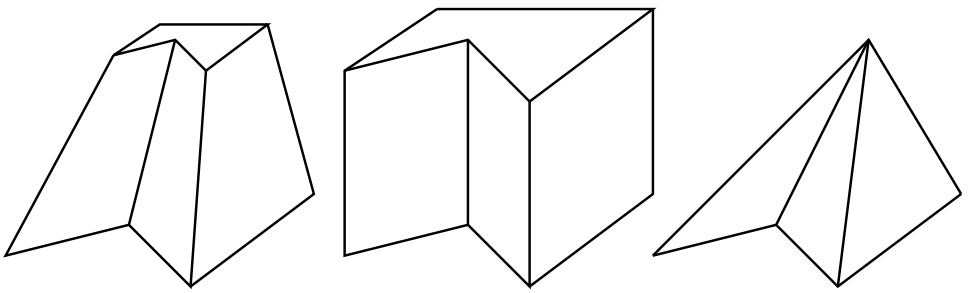


Figura 2.36 El prisma y la pirámide son casos particulares de la pirámide truncada.

Si la base y la tapa son iguales, se trata de un prisma mientras que, si la tapa se contrae hacia un punto, resulta una pirámide. La fórmula que buscamos debe entonces "generalizar" las dos fórmulas anteriores y unirlas en una sola. Además, debe ser **simétrica** en la base y la tapa, es decir, si volteamos la pirámide cortada se intercambian la base y la tapa, pero el volumen no cambia; por ello, dicho intercambio no debe alterar la fórmula.

Para simplificar los cálculos que haremos a continuación, supondremos ahora que la pirámide truncada tiene como base un cuadrado de lado a, como tapa un cuadrado de lado b y la altura será denotada siempre por h.

Si la pirámide trunca se completa, obtenemos una altura H. El volumen es de $\frac{a^2 \cdot H}{3}$, dado que la base tiene área a^2. Lo que se cortó al truncar es, también, una pirámide con base de área b^2 y altura $H - h$. De esta manera, la primera fórmula sería:

$$V = \frac{a^2 \cdot H}{3} - \frac{b^2 \cdot (H - h)}{3} = \frac{a^2 - b^2}{3} H + \frac{b^2}{3} h \qquad (4)$$

La altura H se obtiene por semejanza con:

$$\frac{H}{a} = H - hb$$

de donde, al despejar tenemos que:

$$H = \frac{a}{a - b} h.$$

Si sustituimos H por la expresión $h\frac{a}{a-b}$ en la fórmula (4):

$$V = \frac{a^2 - b^2}{3} \frac{a}{a - b} h + \frac{b^2}{3} h,$$

lo cual puede simplificarse en:

$$V = \frac{a^2 + ab + b^2}{3} h. \qquad (5)$$

Ahora, si $a = b$ resulta que $a^2 + ab + b^2 = 3a^3$ y entonces, $V = a^2 \cdot h$ como lo esperábamos. Si, en cambio $b = 0$, entonces $V = \frac{a^2 \cdot h}{3}$, que también es lo que esperábamos. Se puede observar la simetría entre a y b: si intercambiamos estas dos variables obtenemos $V = \frac{a^2 + ab + b^2}{3} h$, lo mismo que se obtuvo en (5).

Si bien trabajamos con cuadrados para la base y la tapa, ¿cómo habría que generalizar esta última fórmula a una pirámide truncada general? Primero, observemos que $a^2 = B$ es el área de la base y $b^2 = T$ es el área de la tapa. Pero, ¿cómo habría que interpretar el término ab? Como $a = \sqrt{B}$ y $b = \sqrt{T}$, entonces $ab = \sqrt{BT}$ y, a partir de ahí, encontramos que:

$$\text{volumen de una pirámide} = \frac{B + \sqrt{BT} + T}{3} \cdot h,$$

aun cuando la base y la tapa tengan una forma distinta. Todo este razonamiento se hizo a partir del principio de Cavalieri que se analizará en la sección 2.11, con mayor detalle.

2.10 EL NÚMERO π Y LA CUADRATURA DEL CÍRCULO

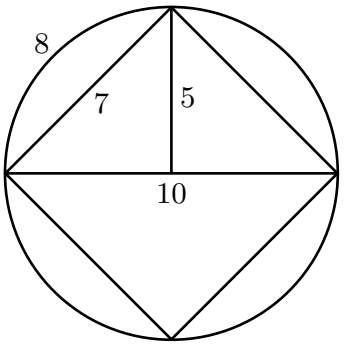


Figura 2.37 Edwin J. Goodwin, un aficionado a las matemáticas, propuso fijar tanto el valor de la raíz cuadrada de 2 como el valor de π —sus aproximaciones eran de 3.2 para π y de 1.4 para $\sqrt{2}$—. La propuesta fue aceptada por unanimidad por The House of Representatives del estado de Indiana, en 1897. Éste es, tal vez, el intento más célebre de establecer una certeza científica por medio de la ley. El Senado de Indiana, advertido de la falsedad, pospuso la decisión indefinidamente.

Civilizaciones tan antiguas como la babilónica, la egipcia, la china y la hindú, reconocieron que había ciertas relaciones entre las dimensiones de algunas figuras geométricas que se mantenían constantes aunque su tamaño variara, es decir, que eran independientes del tamaño de la figura.

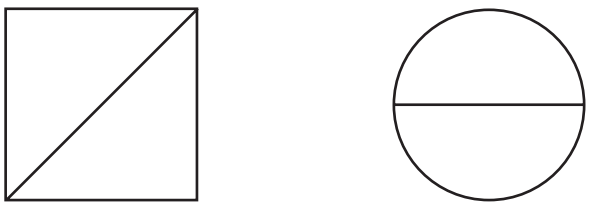


Figura 2.38 Un cuadrado con una de sus diagonales y un círculo con uno de sus diámetros.

Por ejemplo, la diagonal de un cuadrado mide un poco menos que una y media veces su lado, independientemente de si el cuadrado es grande o pequeño. También el perímetro de una circunferencia es poco más que tres veces su diámetro, sin importar si la circunferencia es grande o minúscula como una retina. Se trata de relaciones, razones o proporciones fijas —que no se alteran aunque varíe el tamaño de las figuras—, que sólo dependen de su for-

ma. Estas relaciones permiten calcular, por ejemplo, la diagonal de una plaza cuadrada cuyos lados miden 80 metros. A través de sucesivas mediciones podríamos comprobar que la relación entre la diagonal y el lado de un cuadrado es mayor que $1+\frac{3}{8}$ y menor que $1+\frac{1}{2}$. Por lo tanto, la diagonal de la plaza sería mayor que 110 metros y menor que 120 metros. Pero… ¿cuánto mide exactamente la diagonal de la plaza? Para responderlo, tendríamos que saber cuántas veces es mayor la diagonal de un cuadrado que su lado, es decir, tendríamos que conocer con exactitud la relación entre estas dos dimensiones.

El concepto de número que tenía el hombre de la Antigüedad se limitaba a las fracciones —cocientes de números enteros—, demasiado primitivas para representar cantidades —como los números irracionales— que no pueden ser expresadas mediante fracciones. En particular, fueron los pitagóricos quienes, desde entonces, descubrieron que esta razón o proporción entre la diagonal y el lado de un cuadrado no puede expresarse mediante una fracción, así que no tuvieron otro remedio que hacerlo con algún símbolo y estimar su valor usando fracciones. Hoy en día representamos dicha relación como $\sqrt{2}$ y sabemos que $1.414213 < \sqrt{2} < 1.414214$.

De manera análoga, la relación entre el perímetro de una circunferencia y su diámetro tampoco puede expresarse con fracciones. De hecho, como veremos más adelante, se trata de una relación aún más compleja. Las civilizaciones antiguas intentaron expresarla a partir de fracciones cada vez más exactas, pero eran sólo meras aproximaciones de algo cuya definición, como número, escapaba al lenguaje matemático de la época. No fue sino hasta el esplendor de la civilización helénica cuando se llegó a un concepto y una definición precisas de dicha relación, razón o proporción, que hoy llamamos π e identificamos con los famosos dígitos 3.1416.

Los babilonios usaron durante un tiempo al 3 como una aproximación práctica; más adelante y al mejorarla, adoptaron el valor $3+\frac{1}{8}$, equivalente a 3.125. El papiro Rihnd —que data de 1650 a.C.— incluye una aproximación aún más cercana, equivalente a 3.16049. Estas estimaciones eran el resultado de cálculos de perímetros de figuras formadas por segmentos —más o menos parecidas al círculo— y que representaban valores útiles; sin embargo, no respondían a un concepto matemático bien definido.

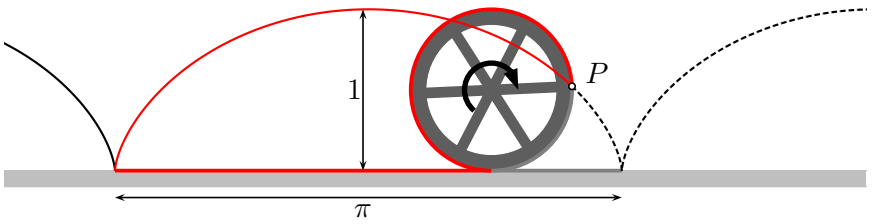


Figura 2.39 π como el recorrido de una rueda de diámetro unitario.

La definición correcta y el primer cálculo de esta relación como un resultado matemático se deben a Arquímedes de Siracusa —siglo III a.C.—. Él demuestra, apoyado en los conocimientos geométricos de la época que, efectivamente, la relación entre el perímetro y el diámetro de la circunferencia es independiente del tamaño de la misma —y se mantiene aunque el tamaño cambie—, y en seguida procede a realizar un cálculo riguroso de la relación demostrando que se encuentra entre $3+\frac{10}{71}$ y $3+\frac{1}{7}$. El uso de la letra π —inicial que denota perímetro en griego— es mucho más reciente y proviene del matemático galés William Jones, quien la usó por primera vez en 1706. Esta notación se popularizó luego a través de los trabajos de Leonhard Euler.

Lo relevante del trabajo de Arquímedes no son los valores concretos que obtuvo, sino el método que inventó para ello y con el que abrió la posibilidad de obtener el valor de π con tanta precisión como se desee, aunque el proceso pueda ser lento por la dificultad de los cálculos requeridos. Es importante tomar en cuenta que, para obtener su resultado, Arquímedes tuvo que realizar varias estimaciones de fracciones y raíces cuadradas —cuya expresión decimal era entonces desconocida— en forma de laboriosas desigualdades entre fracciones. Actualmente, el método de Arquímedes para estimar π puede utilizarse perfectamente en una computadora para obtener estimaciones bastante más precisas que las del célebre siracusano.

El método de Arquímedes para definir y calcular π consiste en comparar el perímetro de la circunferencia con los de dos polígonos regulares, uno inscrito y otro circunscrito a ella, como se muestra en la figura.

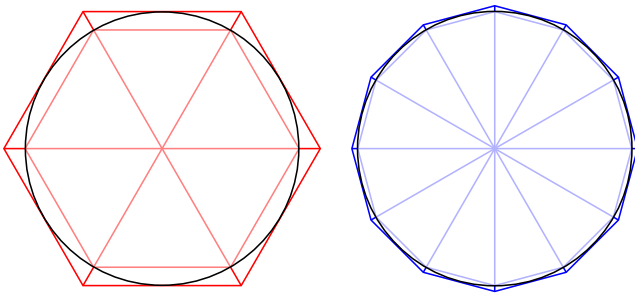


Figura 2.40 Aproximaciones del perímetro de una circunferencia mediante los perímetros de polígonos regulares, de 6 y 12 lados, inscritos y circunscritos.

Al utilizar resultados de la geometría euclidiana, Arquímedes demuestra que el perímetro de la circunferencia debe ser mayor que el del polígono inscrito y menor que el del circunscrito. También demuestra que la diferencia entre ambos puede hacerse tan pequeña como se desee al considerar polígonos con un número muy grande de lados. Estas ideas contienen no sólo el germen del concepto de límite utilizado en el cálculo moderno sino, también, la manera de tratarlo con absoluto rigor lógico.

Analicemos una circunferencia de radio 1 y consideremos π como la mitad de su perímetro. Entonces, el lado del hexágono inscrito es $L = 1$ y, por el teorema de Pitágoras, su apotema —la distancia del centro a cada lado— es $A = \sqrt{1 - \left(\frac{1}{2}\right)^2}$. Por semejanza de triángulos, el lado del hexágono circunscrito mide $\frac{L}{A} = \frac{1}{A}$. De ahí se obtienen las siguientes desigualdades:

$$3 < \pi < 3 \cdot \frac{1}{A}$$

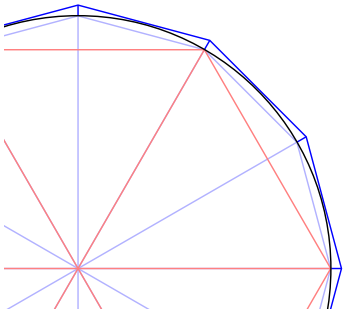


Figura 2.41 Refinamiento de la aproximación pasando de un polígono de 6 lados a uno de 12, es decir, de un hexágono a un dodecágono.

Si ahora se consideran los dodecágonos regulares —inscrito y circunscrito— y se usa nuevamente el teorema de Pitágoras, resulta que el lado del dodecágono inscrito es $L_1 = \sqrt{\left(\frac{1}{2}\right)^2 + (1 - A)^2}$, mientras que su apotema es $A_1 = \sqrt{1 - \left(\frac{L_1}{2}\right)^2}$; por semejanza de trián-

gulos se obtiene que el lado del dodecágono circunscrito es $\frac{L_1}{A_1}$. Al sumar los lados, se obtienen las siguientes desigualdades:

$$3 \cdot 2 \cdot L_1 < \pi < 3 \cdot 2 \cdot \frac{L_1}{A_1}$$

Si se siguen construyendo polígonos regulares de 24, 48, 96, ... lados, y cada vez se divide el ángulo interno entre dos, se tienen las desigualdades:

$$3 \cdot 2^n \cdot L_n < \pi < 3 \cdot 2^n \cdot \frac{L_n}{A_n}$$

donde L_n es el lado del polígono regular con $6 \cdot 2^n$ lados y $A_n = \sqrt{1 - \left(\frac{L_n}{2}\right)^2}$ es su apotema. El valor de L_n se puede obtener recursivamente —que quiere decir "repetir indefinidamente la aplicación de"— mediante la fórmula $L_{n+1} = \sqrt{\left(\frac{L_n}{s}\right)^2 + (1 - A_n)^2}$, consecuencia directa del teorema de Pitágoras. Estas fórmulas ofrecen un algoritmo para calcular aproximaciones de π con cualquier grado de precisión. La siguiente tabla muestra las desigualdades —que se obtienen al usar este procedimiento— expresadas con 6 decimales:

NÚMERO DE LADOS DE LOS POLÍGONOS	ESTIMACIÓN DE π
$6 = 6 \cdot 2^0$	$3.000000 < \pi < 3.464102$
$12 = 6 \cdot 2^1$	$3.105828 < \pi < 3.215391$
$24 = 6 \cdot 2^2$	$3.132628 < \pi < 3.159660$
$48 = 6 \cdot 2^3$	$3.139350 < \pi < 3.146087$
$96 = 6 \cdot 2^4$	$3.141031 < \pi < 3.142715$
$192 = 6 \cdot 2^5$	$3.141452 < \pi < 3.141874$
$384 = 6 \cdot 2^6$	$3.141557 < \pi < 3.141663$
$768 = 6 \cdot 2^7$	$3.141583 < \pi < 3.141611$
$1\,536 = 6 \cdot 2^8$	$3.141590 < \pi < 3.141598$
$3\,072 = 6 \cdot 2^9$	$3.141592 < \pi < 3.141594$

Las desigualdades correspondientes al polígono de 96 lados son ligeramente mejores a las que obtuvo Arquímedes con el mismo polígono: $3 + \frac{10}{71} < \pi < 3 + \frac{1}{7}$ —aproximadamente $3.140845 < \pi < 3.142857$, con precisión de seis decimales. La aproximación que se obtiene con el polígono de 3072 lados corresponde a la que obtuvo en el año 263 d.C. el matemático chino Lui Hui.

Esfuerzos sucesivos en el cálculo han alcanzado más y más precisión. Primero fueron manuales y después se hicieron usando computadoras cada vez más potentes. En la actualidad, el récord en el cálculo de dígitos lo tiene Daisuke Takahashi, con más de dos y medio billones de cifras decimales. Cabe mencionar que estos esfuerzos no tienen ninguna importancia respecto al significado de π, sólo se hacen por el prestigio que da a los programadores y constructores de computadoras el llegar a ellos.

En cambio, es muy interesante saber que el número π aparece en una gran cantidad de resultados matemáticos, aparentemente, sin relación alguna con la circunferencia. Por ejemplo, se sabe que:

$$\frac{\pi}{4} = \frac{1}{1} - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots$$

$$\frac{\pi^2}{6} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots$$

$$\sqrt{\pi} = \int_{-\infty}^{\infty} e^{-x^2} \, \mathrm{d}x$$

Arquímedes no sólo define y calcula π sino que usa su valor para expresar el área de la circunferencia y, también, el de volúmenes y áreas de conos, cilindros y esferas. Se trata de los primeros resultados de cálculos matemáticos precisos de áreas de figuras delimitadas por curvas y de superficies curvas que no pueden aplanarse sin alterar su área. La geometría de Euclides permitía obtener las áreas de regiones delimitadas por rectas, pero al intentar obtener las áreas de regiones con fronteras curvas se experimentaron grandes dificultades. Fue Arquímedes quien señaló el camino adecuado para resolverlas con un método que, a la larga, daría lugar a una de las teorías y herramientas matemáticas más poderosas: el cálculo integral.

Para comprender la importancia de esta parte del trabajo de Arquímedes conviene repasar los problemas de "cuadraturas" que se plantearon los matemáticos griegos. "Cuadrar" una figura era encontrar un cuadrado —un rectángulo o un triángulo— cuya área fuera igual a la de la figura dada. El problema de la cuadratura del círculo se hizo famoso a través de los siglos al encontrar que resultaba imposible hacerlo únicamente con los métodos y elementos de la geometría de Euclides, es decir, sólo con regla y compás. Un caso especial de cuadratura que sí se obtuvo con los métodos euclidianos fue el de las llamadas lunas de Hipócrates de Quios —en el siglo v a.C.—, quien demostró que las figuras obtenidas por defecto de los semicírculos construidos sobre los catetos de un triángulo rectángulo ΔABC, y el semicírculo que tiene como diámetro la hipotenusa y pasa por el ángulo recto, tienen igual área que el triángulo ΔABC.

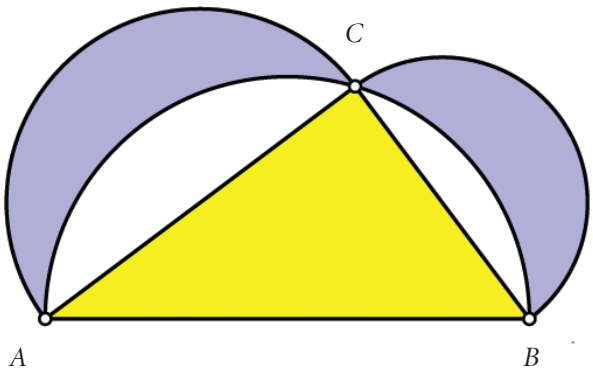


Figura 2.42 El área de las dos lunas de Hipócrates es igual a la del triángulo rectángulo.

El resultado anterior es consecuencia directa del teorema de Pitágoras, aplicado no a los cuadrados construidos sobre los catetos y la hipotenusa, sino a los semicírculos. La suma de las áreas de los semicírculos —que tienen como diámetros a los catetos— es igual al área del semicírculo —cuyo diámetro es la hipotenusa—. La conclusión de Hipócrates se infiere observando que las lunas son el resultado de restar de los semicírculos pequeños las mismas regiones que hay que restar al semicírculo grande para obtener el triángulo.

Esta ingeniosa deducción dio esperanzas a los matemáticos griegos para poder obtener, mediante trucos ingeniosos, la cuadratura de otras figuras delimitadas por curvas, en particular la del círculo. Sin embargo fue imposible, como finalmente logró demostrar Ferdinand Lindemann en 1882. Los métodos euclidianos consisten fundamentalmente de construcciones —hechas con regla y compás— en correspondencia con operaciones algebraicas que incluyen la extracción de raíces cuadradas. Los números algebraicos son los que pueden obtenerse mediante este tipo de operaciones —a partir de los enteros—, mientras que se llama **trascendentes** a los que no se pueden obtener de esta manera. Lindemann demostró que π es un número trascendente y, por lo tanto, la cuadratura del círculo no puede obtenerse usando regla y compás.

Sin embargo, dos mil años antes Arquímedes vio que había otro camino para resolver las cuadraturas: uno mucho más general pero que requería de un concepto que él mismo

no definió completamente, aunque contribuyó a crear, el de **límite**. Veamos en forma resumida cómo obtiene Arquímedes el área de un círculo de radio R.

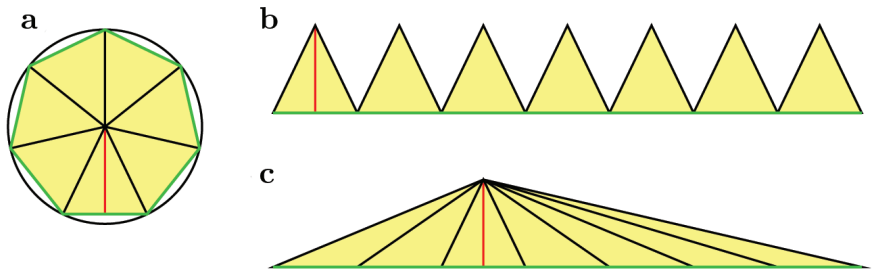


Figura 2.43 El área del polígono inscrito es igual a la del triángulo, cuya base es el perímetro del polígono y cuya altura es el apotema.

Si se inscribe un polígono regular de n lados en un círculo de radio R, su área es la de n triángulos isósceles de base L y altura A, es decir, $n\frac{LA}{2}$, donde L es la longitud del lado del polígono y A el apotema. A medida que n aumenta, el área del polígono tiende a la del círculo y, al mismo tiempo, nL tiende al perímetro $2\pi r$ mientras el apotema A tiende al radio R. Por lo tanto:

$$n\frac{LA}{2} = \frac{nL \cdot A}{2} \quad \text{tiende a} \quad \frac{2\pi R \cdot R}{2} = \pi R^2$$

$$\text{cuando } n \text{ tiende a } \infty,$$

lo cual demuestra que el área del círculo de radio R es πR^2. Este argumento recurre al concepto moderno de "tender al límite", pero Arquímedes hizo su demostración utilizando desigualdades similares a las que usó para el cálculo de π.

En la siguiente sección se estudiará cómo aparece π en el cálculo de las áreas y volúmenes de cilindros, conos y esferas.

2.11 Cilindros, conos y esferas

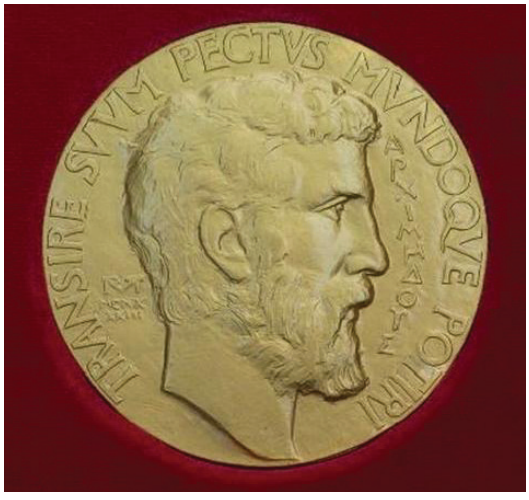


Figura 2.44 Busto de Arquímedes en la medalla Fields, reconocimiento internacional que se otorga cada cuatro años por descubrimientos sobresalientes en matemáticas.

Arquímedes de Siracusa —sin duda el matemático más importante de la Antigüedad— quiso que, como único epitafio en su tumba —encontrada efectivamente con esa inscripción en la isla de Sicilia—, se dibujara una esfera inscrita en un cilindro junto con la relación que había descubierto entre las áreas y los volúmenes de estas figuras. Consciente de ser un gran matemático, Arquímedes se ufanaba de ello y ese resultado suyo era del que más orgulloso se sentía. Y no es para menos. Desde la escuela primaria sabemos las fórmulas para calcular las superficies y volúmenes de cilindros, conos y esferas, pero nunca nos explican cómo se obtienen, sólo nos dan las fabulosas recetas. A continuación repasaremos esas fórmulas, explicando cómo se obtiene cada una de ellas.

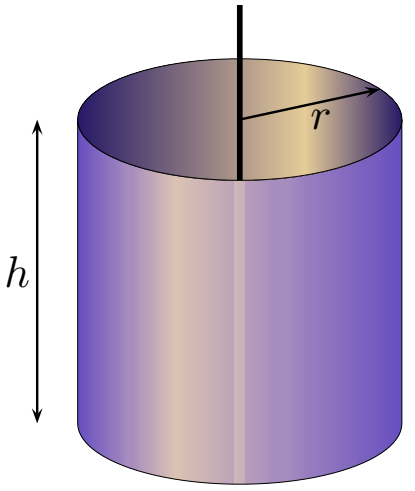


Figura 2.45 La superficie del cilindro con radio r y altura h es 2πrh.

Para el caso del cilindro, la fórmula se justifica con facilidad. Si se corta la figura a lo largo de una de sus alturas y se extiende hasta dejarla plana, se obtiene un rectángulo cuya base es el perímetro $2\pi r$ de la base del cilindro, y cuya altura es la misma del cilindro: h y, por lo tanto, su superficie es $2\pi r \cdot h$.

Un cono circular recto se caracteriza por tener una base circular y el vértice situado sobre la recta perpendicular a la base, que pasa por su centro. Las aristas de estos conos tienen todas la misma longitud.

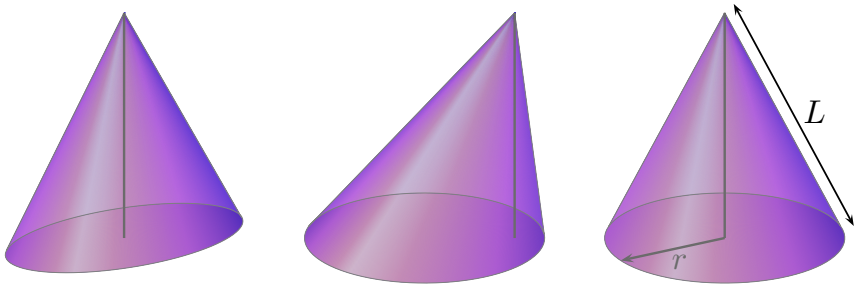


Figura 2.46 De izquierda a derecha, se muestra un cono, un cono circular y un cono circular recto.

Un cono circular recto de lado L puede construirse a partir de un sector circular de radio L y, por lo tanto, su área es igual a la del sector circular, o sea, la mitad del perímetro por el radio.

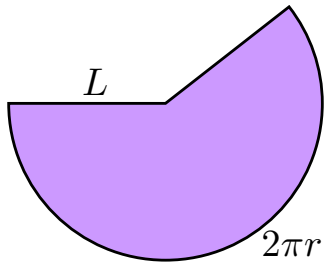


Figura 2.47 Sector circular obtenido al extender un cono.

Como el perímetro es $2\pi r$, donde r es el radio de la base circular del cono y L el radio, entonces el área del cono es $S = \pi r L$.

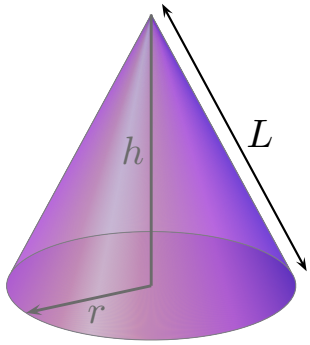


Figura 2.48 Superficie del cono de lado L y con radio de la base igual a r : $S = \pi r L$.

Para estudiar la superficie de una esfera, Arquímedes la aproxima por medio de sectores cónicos. Un sector de cono circular recto es la parte del cono comprendida entre dos planos paralelos a su base, como se muestra en la figura 2.49.

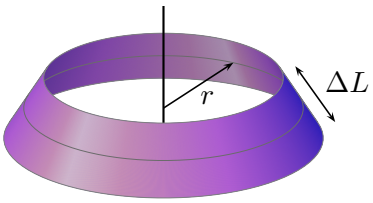


Figura 2.49 Sector de cono circular recto, de radio r y ancho ΔL.

El área ΔS de un sector cónico con radio medio r y ancho ΔL es igual al perímetro de la circunferencia media por el ancho, es decir:

$$\Delta S = 2\pi r \cdot \Delta L$$

Este importante resultado puede obtenerse intuitivamente sumando las áreas de una infinidad de trapecios de ancho **infinitesimal** y altura ΔL. La suma de los anchos medios de los trapecios sería igual al perímetro de la circunferencia media $2\pi r$. También puede obtenerse mediante un cálculo exacto considerando al sector cónico como la diferencia de dos conos: uno cuyo radio de la base es $r + \frac{\Delta r}{2}$ y su lado es $L + \frac{\Delta L}{2}$ y otro cuyo radio de la base es $r - \frac{\Delta r}{2}$ y su lado es $L - \frac{\Delta L}{2}$. Al aplicar la fórmula para las áreas de ambos conos y restarlas, se obtiene:

$$\begin{aligned} \Delta S &= \pi \cdot \left(r + \frac{\Delta r}{2}\right) \cdot \left(L + \frac{\Delta L}{2}\right) - \pi \cdot \left(r - \frac{\Delta r}{2}\right) \cdot \left(L - \frac{\Delta L}{2}\right) \qquad (6) \\ &= \pi \cdot \left(2L\frac{\Delta r}{2} + 2r\frac{\Delta L}{2}\right) = \pi \cdot (L\Delta r + r\Delta L) \end{aligned}$$

Por semejanza de triángulos se tiene que $\frac{\Delta r}{\Delta L} = \frac{r}{L}$ y, por lo tanto, $L\Delta r = r\Delta L$. Al sustituir en (6) tenemos $\Delta S = 2\pi r \cdot \Delta L$, que es lo que se deseaba demostrar.

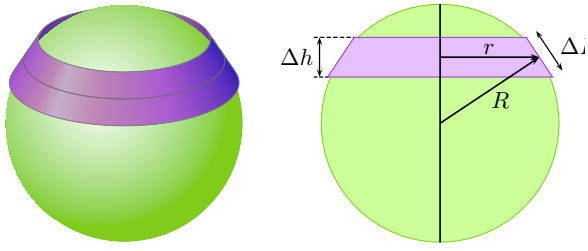


Figura 2.50 Sector de cono circular recto, de radio r, ancho ΔL y altura h, que es tangente a una esfera de radio R.

Si un sector cónico de radio medio r y ancho ΔL es tangente a una esfera de radio R —precisamente a lo largo de la circunferencia media, como muestra la figura 2.50—, entonces el área ΔS del sector cónico es igual a:

$$2\pi r \cdot \Delta h$$

donde Δh es la altura del sector. En efecto, para demostrar esto basta ver que, por semejanza de triángulos,

$$\frac{r}{R} = \frac{\Delta h}{\Delta L} \text{ y, por lo tanto, } r \cdot \Delta L = L \cdot \Delta h \,.$$

Estamos ya en condiciones de calcular el área de una esfera tal como lo hizo Arquímedes. Para hacerlo, ahora nosotros, cubrimos la esfera con sectores cónicos tangentes a ella, como se ilustra en la siguiente figura:

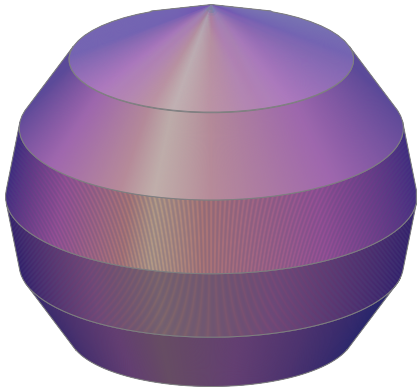


Figura 2.51 Esfera cubierta por sectores circulares cónicos tangentes a ella.

Ya que todos los sectores cónicos son tangentes a la esfera de radio R, sus áreas son $\Delta S_i = 2\pi R \cdot \Delta h_i$, donde Δh_i son las alturas de los diferentes sectores cónicos. La suma de todas estas áreas es igual a:

$$\sum \Delta S_i = 2\pi RH,$$

donde H es la altura total de la cubierta de la esfera. Obsérvese que $2R < H$, y que H puede hacerse tan cercana a $2R$ como se desee, utilizando una cubierta suficientemente fina. Por lo tanto, el área de la esfera satisface la desigualdad:

$$S \leq 4\pi R^2.$$

De manera análoga, al **inscribir** en la esfera los sectores cónicos, se demuestra la desigualdad contraria: $4\pi R^2 \leq S$ y, por consiguiente,

$$S = 4\pi R^2.$$

En otras palabras, el área de la esfera es igual al cuádruple del área de uno de sus círculos máximos. Otra manera de interpretar el resultado es diciendo que el área de la esfera es igual a la del mínimo cilindro que la contiene, sin contar las tapas, o bien, que el área de la esfera es $\frac{2}{3}$ del área del cilindro que la contiene, si se incluyen las tapas.

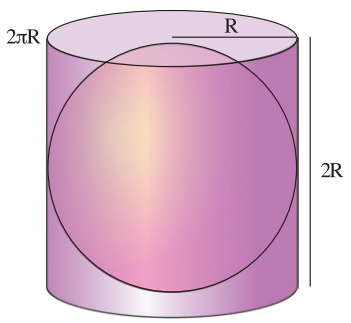


Figura 2.52 Superficies S_E de la esfera y S_C del cilindro: $S_E : S_C = 2 : 3$.

Ésta es la figura y la relación que debían aparecer como epitafio en la tumba de Arquímedes, de acuerdo con sus propios deseos y, también, de acuerdo con el testimonio de Cicerón —que visitó la tumba del gran matemático dos siglos después de su muerte—. Como veremos más adelante, la relación de $\frac{2}{3}$ se da entre los volúmenes de las mismas figuras.

El área de un casquete esférico, que consiste en la parte de una esfera que se encuentra arriba de un plano horizontal —como se muestra en la figura 2.53—, es igual a la de la circunferencia, cuyo radio es la distancia L del polo norte a cualquiera de los puntos de la orilla, es decir, $S = \pi L^2$.

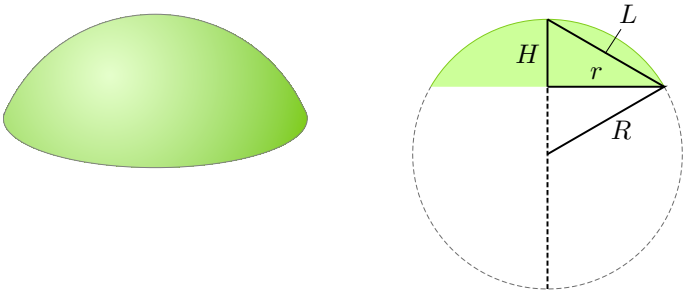


Figura 2.53 Un casquete esférico es la parte de la superficie de una esfera, cortada por un plano que pasa arriba del ecuador.

La prueba de este resultado se hace cubriendo la superficie esférica con sectores cónicos, como se hizo con anterioridad para toda la esfera. Así, se obtiene que $S = 2\pi RH$, donde H es la altura de la superficie esférica. Por el teorema de Pitágoras:

$$L^2 = H^2 + r^2 = H^2 + R^2 - (R - H)^2 = 2RH.$$

Sustituyendo esta igualdad en la fórmula anterior se obtiene el resultado anunciado: $S = \pi L^2$.

Pasemos ahora al cálculo de los volúmenes del cilindro, el cono y la esfera. El caso del cilindro es muy sencillo pues todas sus secciones paralelas a la base son círculos del mismo radio —esto es, tienen la misma área $A = \pi r^2$— y, por lo tanto, el volumen es el producto de esta área por la altura h, es decir:

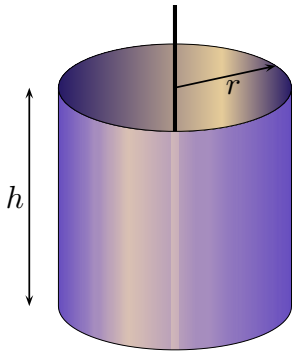


Figura 2.54 Volumen del cilindro de altura h y radio r: $V = \pi r^2 h$.

Para obtener el volumen del cono, se recurre a pirámides inscritas y circunscritas, como muestra la figura 2.55, y se utiliza el resultado conocido de que el volumen de una pirámide es igual a la tercera parte del producto del área de la base por la altura.

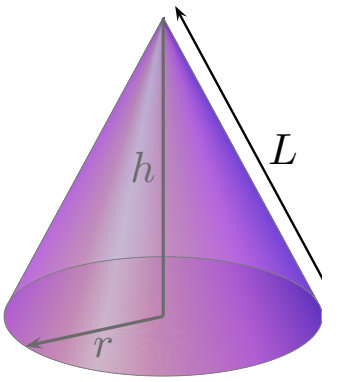


Figura 2.55 Cono con una pirámide inscrita y otra circunscrita.

Esto permite demostrar que lo mismo sucede para un cono, como se muestra en la figura 2.56.

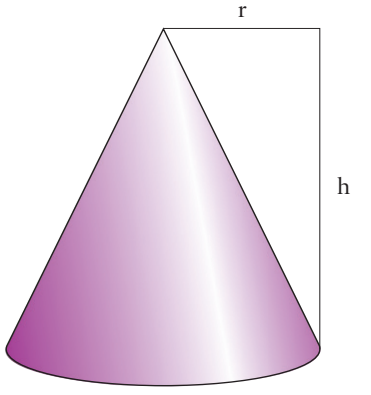


Figura 2.56 Volumen del cono de altura h y radio r: $V = \frac{1}{3}\pi r^2 h$.

Para obtener el volumen de la esfera, Arquímedes utilizó un truco muy ingenioso que consiste en comparar las secciones que se obtienen al cortar por un plano horizontal a un cono, una semiesfera y un cilindro, como se muestra en la figura 2.57.

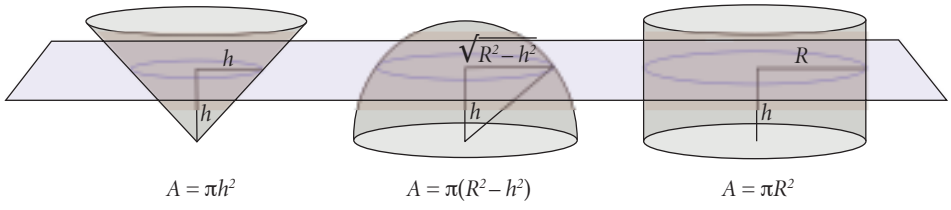


Figura 2.57 Cortes de un cono, un hemisferio y un cilindro por un plano.

La suma de las áreas de las secciones sobre el cono y la semiesfera es igual al área de la sección del cilindro, y esto sucede para todos los cortes. Por consiguiente, se puede deducir que el volumen del cono más el de la semiesfera es igual al del cilindro, es decir, si denotamos por V el volumen de la semiesfera, entonces:

$$\frac{1}{3}\pi R^3 + V = \pi R^3.$$

Por lo tanto,

$$V_{\text{SE}} = \frac{2}{3}\pi R^3$$

y en consecuencia, el volumen de la esfera es el doble:

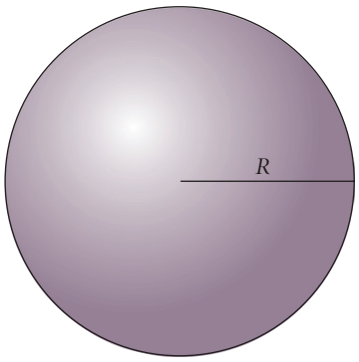


Figura 2.58 Volumen de la esfera de radio R: $V = \frac{4}{3}\pi R^3$.

Comparando el volumen de la esfera con el del cilindro de radio R y altura $2R$ —que vale $2\pi R^3$ —, observamos que:

$$\frac{\frac{4}{3}\pi R^3}{2\pi R^3} = \frac{\frac{4}{3}}{2} = \frac{4}{6} = \frac{2}{3}$$

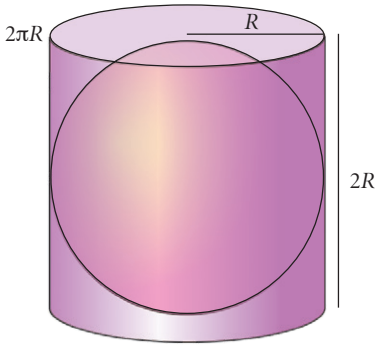


Figura 2.59 Volúmenes V_E de la esfera y V_C del cilindro: $V_E : V_C = 2 : 3$.

y entonces, se cumple la misma relación $2 : 3$ que se obtuvo para las superficies.

2.12 LA CUADRATURA DE LA PARÁBOLA Y EL MÉTODO DE DEMOSTRACIÓN POR INDUCCIÓN

La cuadratura de la parábola, al igual que el cálculo del volumen de la esfera, se debe al genio de Arquímedes. Aquí la obtendremos recurriendo a un procedimiento general para obtener áreas de regiones delimitadas por curvas, esencia del cálculo integral. Este método nos llevará, por un camino diferente al de la sección anterior, hasta obtener los volúmenes del cono y la esfera.

Para encontrar la cuadratura de la parábola necesitaremos la igualdad:

$$1^2 + 2^2 + \ldots + N^2 = \frac{N(N+1)(2N+1)}{6}$$

que es válida para todo entero positivo N. Otra manera de expresar esta igualdad es usando la notación de suma:

$$1^2 + 2^2 + \ldots + N^2 = \sum_{n=1}^{N} n^2$$

Figura 2.60 La integración es un procedimiento para aproximar un área o un volumen mediante sumas de pequeñas partes. Es un proceso matemático tan importante que tiene su propio símbolo: a primera vista se parece a una de las ranuras en un cello, pero proviene de una letra S estilizada que, a la vez, recuerda la suma.

por lo que la fórmula puede también escribirse así:

$$\sum_{n=1}^{N} n^2 = \frac{N(N+1)(2N+1)}{6}.$$

Esta fórmula suele demostrarse por el llamado **método de inducción**, que consta de dos pasos:

1. Se comprueba que la fórmula es válida para $N = 1$.
2. Suponiendo que la fórmula es válida para algún entero positivo N, se prueba que también será válida para el siguiente entero $N + 1$.

El primer paso se cumple porque, para $N = 1$, ambos lados de la igualdad valen 1, en efecto:

$$\sum_{n=1}^{N} n^2 = \sum_{n=1}^{1} n^2 = 1^2 = 1,$$
$$\frac{N(N+1)(2N+1)}{6} = \frac{1(1+1)(2\cdot 1+1)}{6} = \frac{1\cdot 2\cdot 3}{6} = 1$$

Ahora, supongamos que la igualdad se cumple para algún entero positivo N y probemos que entonces también se cumple para $N + 1$.

$$\begin{aligned}
\sum_{n=1}^{N+1} n^2 &= \sum_{n=N}^{1} n^2 + (N+1)^2 && \text{se separa el último término de la suma}\\
&= \frac{N(N+1)(2N+1)}{6} + (N+1)^2 && \text{se usa la hipótesis de inducción}\\
&= \frac{N(N+1)(2N+1) + 6(N+1)^2}{6} && \text{se saca común denominador}\\
&= \frac{(N+1)(N\cdot(2N+1) + 6N + 6)}{6} && \text{se factoriza } (N+1)\\
&= \frac{(N+1)(2N^2 + 7N + 6)}{6} && \text{se desarrolla el segundo factor del numerador}\\
&= \frac{(N+1)(N+2)(2N+3)}{6} && \text{se factoriza el segundo factor del numerador}\\
&= \frac{(N+1)((N+1)+1)(2(N+1)+1)}{6} && \text{se reescribe para exhibir la fórmula}
\end{aligned}$$

Esto completa la demostración por inducción.

Cuadrar la parábola significa encontrar el área encerrada por ella y una recta. Simplificaremos el proceso al pasar directamente a obtener el área debajo de una parábola específica, la que corresponde a la gráfica de $y = x^2$. Denotemos por A al área bajo la parábola entre 0 y a.

Vamos a demostrar que $A = \frac{a^3}{3}$. Para ello, dividimos el intervalo $[0, a]$ en N partes iguales y dibujamos dos gráficas formadas por rectas horizontales —una por debajo de la parábola y otra por arriba—, como se ilustra en la figura 2.61.

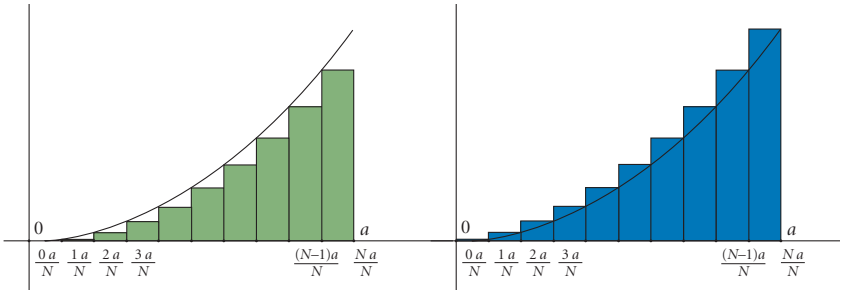


Figura 2.61
Aproximaciones por abajo y por arriba del área bajo la parábola.

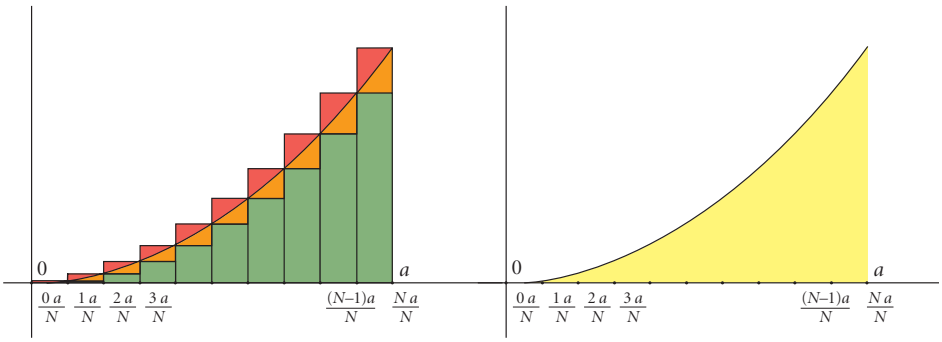


Figura 2.62 Aproximación por ambos lados y el área bajo la parábola.

Entonces, A es mayor que el área de los rectángulos verdes y menor que el área de los rectángulos rojos. La suma de las áreas de los rectángulos verdes es:

$$\left(\frac{a}{N}\right)^2 \cdot \frac{a}{N} + \left(2\frac{a}{N}\right)^2 \cdot \frac{a}{N} + \ldots + \left((N-1)\frac{a}{N}\right)^2 \cdot \frac{a}{N} = \frac{a}{N} \cdot \sum_{n=1}^{N-1} \left(n\frac{a}{N}\right)^2$$

y la de los rectángulos rojos:

$$\left(\frac{a}{N}\right)^2 \cdot \frac{a}{N} + \left(2\frac{a}{N}\right)^2 \cdot \frac{a}{N} + \ldots + \left(N\frac{a}{N}\right)^2 \cdot \frac{a}{N} = \frac{a}{N} \cdot \sum_{n=1}^{N} \left(n\frac{a}{N}\right)^2$$

Por lo tanto, se tienen las siguientes desigualdades para el área A bajo la parábola:

$$\begin{aligned}
\frac{a}{N} \cdot \sum_{n=1}^{N-1} \left(n\frac{a}{N}\right)^2 < A < \frac{a}{N} \cdot \sum_{n=1}^{N} \left(n\frac{a}{N}\right)^2 \\
\frac{a^3}{N^3} \cdot \sum_{n=1}^{N-1} n^2 < A < \frac{a^3}{N^3} \cdot \sum_{n=1}^{N} n^2 \\
\frac{a^3}{N^3} \cdot \left(\sum_{n=1}^{N-1} n^2 - N^2\right) < A < \frac{a^3}{N^3} \cdot \sum_{n=1}^{N} n^2
\end{aligned}$$

Al aplicar la fórmula que demostramos anteriormente por inducción y que nos da la suma de los primeros cuadrados, obtenemos las siguientes desigualdades:

$$\begin{aligned}
\frac{N(N+1)(2N+1) - 6N^2}{6N^3} \cdot a^3 < A < \frac{N(N+1)(2N+1)}{6N^3} \cdot a^3 \\
\left(\frac{2N^3 - 3N^2 + N}{6N^3}\right) \cdot a^3 < A < \left(\frac{2N^3 + 3N^2 + N}{6N^3}\right) \cdot a^3 \\
\left(\frac{1}{3} - \frac{1}{2N} + \frac{1}{N^2}\right) \cdot a^3 < A < \left(\frac{1}{3} + \frac{1}{2N} + \frac{1}{N^2}\right) \cdot a^3
\end{aligned}$$

Finalmente, como estas desigualdades se cumplen para cualquier entero positivo N, tomando "uno" suficientemente grande podemos hacer que, tanto el lado izquierdo como el derecho, estén tan cercanos a $\frac{a^3}{3}$ como queramos. Y entonces:

$$A = \frac{a^3}{3}$$

que es lo que queríamos demostrar.

El método anterior es un proceso infinito de aproximación que nos permite obtener un resultado exacto, y contiene la esencia del cálculo integral que Arquímedes ya usaba en el siglo III a.C. y, que Newton y luego Leibniz —al formalizarlo—, generalizaron casi dos mil años más tarde. En la notación moderna del cálculo integral, inventada por Leibniz, este resultado se expresa así:

$$\lim_{N\to\infty} \frac{a}{N} \cdot \sum_{n=1}^{N-1} \left(n\frac{a}{N}\right)^2 = \int_0^a x^2 \mathrm{d}x$$

y debe leerse como "la integral de cero a a de x^2 con respecto a x". La integral se convierte en el límite de cualquier suma cuando el número de particiones N tiende a infinito —y por lo tanto, el ancho de las particiones tiende a cero.

Intuitivamente, la integral se interpreta como una "suma infinita" de cantidades "infinitamente pequeñas" o bien, "infinitesimales". Aunque esta manera de hablar no es matemáticamente rigurosa, se utiliza con frecuencia como una forma abreviada de indicar el proceso con el que se llega —sí, rigurosamente— al resultado.

El método de integración usado para alcanzar la cuadratura de la parábola nos permite obtener fácilmente, como corolarios, dos resultados ya conocidos: el volumen del cono y el de la esfera.

Para encontrar el volumen del cono cuya altura es H y cuya base tiene radio R, lo partimos en una "infinidad" de discos —cilindros circulares— de grosor "infinitesimal" *dh* y radio $r = \frac{R}{H}h$, donde h varía de 0 a H. Entonces, el volumen del cono es la "suma infinita" de los volúmenes de todos esos discos. En la notación de Leibniz, lo anterior se escribe como:

$$V = \int_0^H \pi r^2 \mathrm{d}h = \pi\frac{R^2}{H^2}\int_0^H h^2 \mathrm{d}h.$$

Salvo por la constante $\pi\frac{R^2}{H^2}$ y un cambio en el nombre de la variable —x por h—, esta es la misma integral que se usó para obtener la cuadratura de la parábola, de donde sabemos que:

$$\int_0^H h^2 \mathrm{d}h = \frac{H^3}{3}.$$

En consecuencia, el volumen del cono es:

$$V = \pi\frac{R^2}{H^2} \cdot \frac{H^3}{3} = \frac{1}{3}\pi R^2 H.$$

En forma análoga, partiendo del resultado de que el área de una superficie esférica de radio r es $4\pi r^2$, el cálculo del volumen de la esfera de radio R es equivalente a la cuadratura de la parábola $y = 4\pi x^2$ entre 0 y R. En efecto, la esfera se puede considerar como la suma de una infinidad de superficies esféricas de grosor infinitesimal $\mathrm{d}r$ y de área —variable— $4\pi r^2$ con r entre 0 y R. Por consiguiente, el volumen de la esfera puede calcularse también como una "suma infinita" de cantidades "infinitesimales":

$$V = \int_0^R 4\pi r^2 \mathrm{d}r.$$

Y de nuevo, salvo por la constante 4π y un cambio en el nombre de la variable —x por r—, esta "suma infinita" de "infinitesimales", que hoy llamamos *integral*, es la misma que se usa para cuadrar la parábola. Por lo tanto, el volumen de la esfera de radio R es:

$$V = \frac{4\pi R^3}{3}.$$

2.13 LAS CÓNICAS Y SU USO

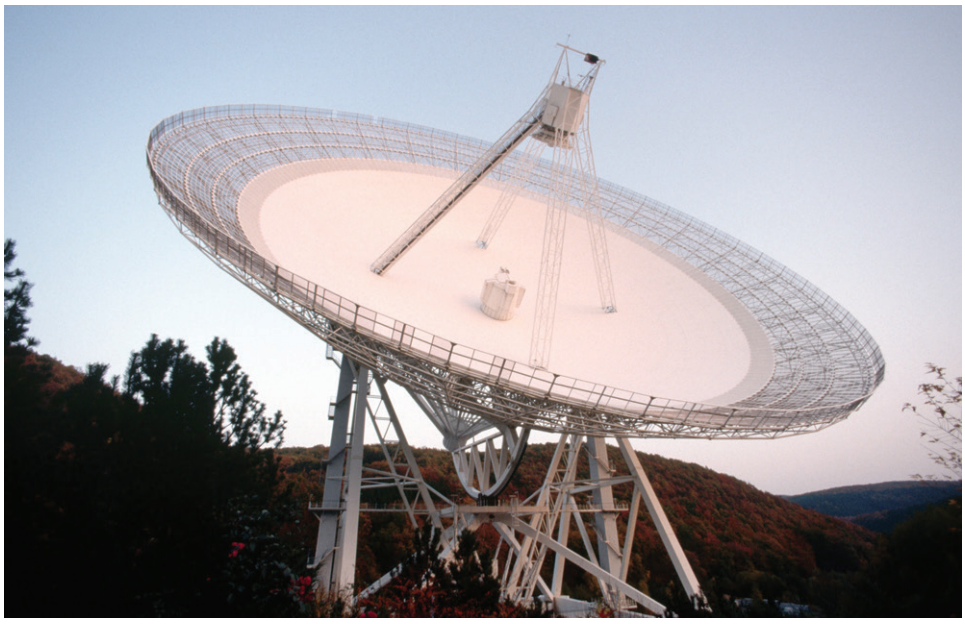


Figura 2.63 La antena de Effelsberg, un radiotelescopio de 100 metros de diámetro, es la más grande en Europa. Estos telescopios se diseñan para registrar señales electromagnéticas con frecuencias de 10 MHz a 100 MHz, lo que corresponde a longitudes de onda de 3 a 30 metros. Por ello, es posible usar mallas de metal para formar la superficie del reflector que concentra las ondas en el receptor. Para que las reflexiones lleguen a un solo punto —donde está el receptor— es necesario que la antena tenga una forma particular: que sea parabólica | © Latin Stock México.

2.13.1 Secciones de un cono

Las cónicas son curvas en el plano que aparecen en diferentes contextos —como en la naturaleza y en la tecnología— y están presentes en muchos objetos de uso diario —por ejemplo, en los faros de los automóviles. Su popularidad se debe a un efecto similar que se describió en la esfera de la sección 1.3.1: cumplen muchas propiedades y éstas las definen de manera definitiva. En esta sección revisaremos varios de esos aspectos.

Se cree que el primer estudio sobre cónicas lo hizo Menaechmus, para resolver el problema de la duplicación del cubo —véase también sección 4.2. Aquí empezaremos un poco distinto y tomaremos un primer acercamiento a las cónicas según el origen de su nombre: las **secciones cónicas**, es decir, las secciones con un cono. Cabe agregar que, en dicho acercamiento, se usa el espacio para definir estas curvas.

Primero, se construye un doble cono: se fija una recta ℓ, el **eje**, un punto P sobre ℓ y un ángulo α, tomando en cuenta que $0° < \alpha < 90°$. Entonces, consideramos todas las rectas que pasan por P e inciden en ℓ con un ángulo α. Podemos pensar que tomamos una de estas rectas, la llamamos k y la rotamos alrededor de ℓ. Lo que obtenemos es un doble cono, como se observa en la figura 2.64.

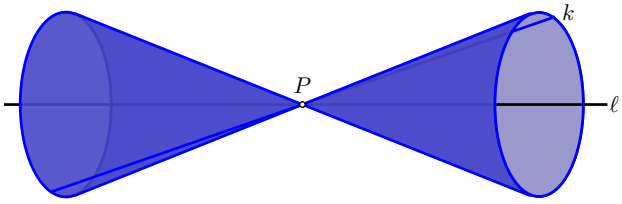


Figura 2.64 Doble cono que se obtiene como resultado al rotar una recta.

Una **sección cónica** es la intersección de un doble cono con un plano y en la figura 2.65 se muestran varios casos típicos. Veamos primero los posibles casos, cuando el plano pasa por P: la sección cónica podrá ser un punto, una recta o un par de rectas que se intersecan. Estos casos se llaman cónicas **degeneradas**. Más interesante es cuando el plano no pasa por P y obtenemos cuatro casos distintos:

1. Si el plano es perpendicular al eje ℓ, entonces la cónica resultante será una **circunferencia**.
2. Si el plano se inclina un ángulo mayor que α pero menor de $90°$, entonces la cónica será una **elipse**.
3. Si el plano se inclina exactamente por el ángulo α, es decir, si algunas de las rectas que definen el doble cono es paralela al plano, entonces la sección resultará ser una **parábola**.
4. Si la inclinación del plano con respecto a ℓ es menor que α, se obtendrá una **hipérbola**. La hipérbola no es conexa, sino que consiste de dos **ramas**.

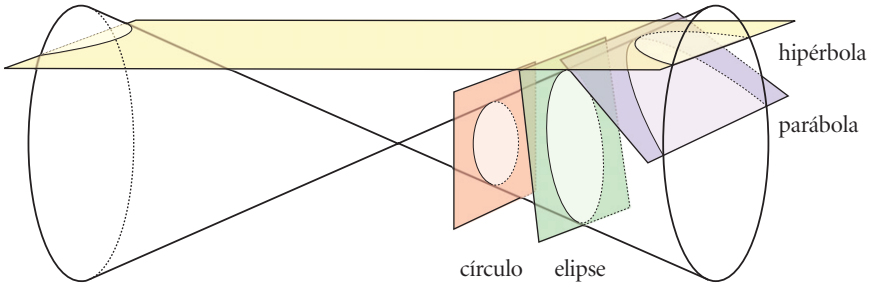


Figura 2.65 Diferentes secciones de un cono.

Podemos ver secciones cónicas en diferentes lugares. Una lámpara de mesa con una sombrilla de apertura circular proyectará —sobre la mesa o una pared cercana— luz en ciertas áreas, mientras otras partes quedarán oscuras. La curva divisoria entre luz y sombra será siempre una cónica. La razón para que ocurra lo anterior es sencilla: la mesa o la pared es plana y la luz se emite en un cono.

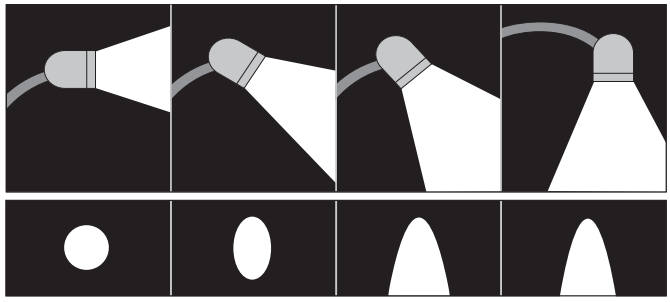


Figura 2.66 La proyección de luz genera diferentes cónicas.

Otro ejemplo son los "spots" empotrados en el techo de muchos cines, que proyectan luz sobre una pared vertical formando parte de una hipérbola.

2.13.2 Las cónicas como lugares geométricos

Fue idea del matemático belga Germinal Pierre Dandelin insertar dos esferas en el cono que tocan el plano de intersección dado. La figura 2.67 muestra un caso donde la cónica es una elipse. Estas esferas tocan al cono en una circunferencia e y al plano que define la cónica en un punto F.

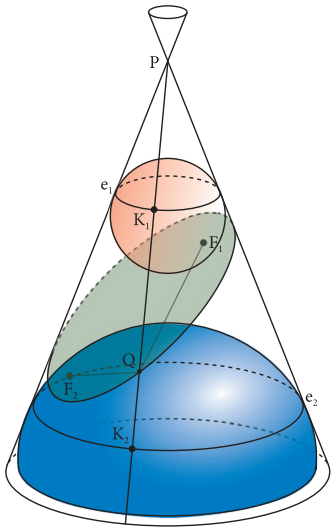


Figura 2.67 Las esferas de Dandelin para el caso de la elipse.

Todas las tangentes de una esfera —desde un punto fijo— tienen la misma longitud. Las dos esferas tocan al cono en dos circunferencias, e_1 y e_2 —donde cada una está formada por puntos que se encuentran a la misma distancia de P—. Por ello, para cada punto Q de la elipse, la distancia K_1, K_2 es la misma — K_1 y K_2 son los puntos de las circunferencias e_1 y e_2 que están sobre la recta PQ —. Como QK_1 y QK_2 son tangentes de la misma esfera, miden lo mismo. En forma similar, QK_2 mide lo mismo que QF_2, por lo tanto:

$$\mathrm{d}\,|Q, F_1| + \mathrm{d}\,|Q, F_2| = \mathrm{d}\,|Q, K_1| + \mathrm{d}\,|Q, K_2| = \mathrm{d}\,|K_1, K_2| = \text{const.} \quad (7)$$

Los dos puntos F_1 y F_2 son los **focos** de la elipse. La propiedad de la elipse como lugar geométrico se enuncia como sigue: la **elipse** es el **lugar geométrico** de todos los puntos Q, tales que $\mathrm{d}\,|Q, F_1| + \mathrm{d}\,|Q, F_2| = h$.

Lo anterior quiere decir que la condición $\mathrm{d}\,|Q, F_1| + \mathrm{d}\,|Q, F_2| = d$ define a todos los puntos de la cónica. Esta propiedad de la elipse se conoce como "la construcción del jardinero", pues se puede trazar una elipse con un hilo amarrado en los cabos a dos postes en el piso —mientras que el hilo esté flojo— al tensar el hilo hacia afuera.

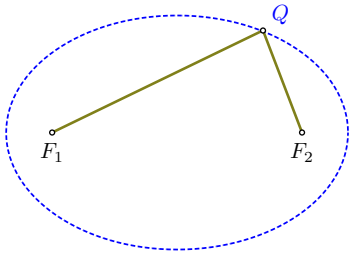


Figura 2.68 La construcción del jardinero de la elipse.

El concepto de **lugar geométrico** es importante en la geometría, dado que es un objeto geométrico definido como un conjunto de puntos que satisfacen alguna propiedad determinada. La siguiente lista muestra algunos lugares geométricos:

1. Dado un punto M y una distancia d, el lugar geométrico de todos los puntos que están a una distancia d de M es una circunferencia con centro M y de radio d.
2. Dados dos puntos A y B, el lugar geométrico de todos los puntos P, tal que $\mathrm{d}\,|P, A| = \mathrm{d}\,|P, B|$ es una recta —la mediatriz— del segmento AB.
3. Dadas dos rectas a y b que se intersecan en P, el conjunto de puntos Q, tal que $\mathrm{d}\,|P, a| = \mathrm{d}\,|P, b|$, son dos rectas perpendiculares —la unión de las dos bisectrices de a y b.

4. Dado un segmento AB, el lugar geométrico de puntos C, tal que $\measuredangle ACB = 90°$, es una circunferencia con diámetro AB.
5. Dadas dos circunferencias que no se intersecan c_1, c_2, el lugar geométrico de todos los puntos P, tal que la tangente de P a c_1 tenga la misma longitud que la tangente de P a c_2, es una recta.

El argumento para la hipérbola es similar, sólo que ahora es una diferencia:

$$|\mathrm{d}Q, F_1 - \mathrm{d}Q, F_2| = |\mathrm{d}Q, K_1 - \mathrm{d}Q, K_2| = |\mathrm{d}K_1, K_2| = \text{const.}$$

Consecuentemente se tiene la siguiente caracterización: la **hipérbola** es el lugar geométrico de todos los puntos Q, tal que $|\mathrm{d}Q, F_1 - \mathrm{d}Q, F_2| = h$.

El valor absoluto de la diferencia se requiere para obtener ambas ramas de la hipérbola. En el caso de la parábola sólo hay una esfera de Dandelin y la descripción geométrica como lugar geométrico se obtendrá más tarde.

A la mitad del siglo XX, se empezó a instalar un sistema de ubicación en alta mar que se llama LORAN —el nombre viene del inglés *long range navigation* — donde se emiten señales desde puntos fijos en la costa. Un barco recibe estas señales en tiempos distintos y puede, a partir de la diferencia de tiempo, calcular la diferencia de distancia. Se sabe entonces que el barco se encuentra sobre una hipérbola. Con las señales de tres emisores se puede calcular una ubicación con buena precisión. El sistema LORAN dejará de operar durante el año 2010, dado que el GPS —del inglés *global positioning system* — otorga una posición mucho más precisa.

2.13.3 La excentricidad

Cada esfera de Dandelin toca al cono en una circunferencia e, que está en un plano A perpendicular al eje ℓ. Si se interseca A con el plano B definido por la cónica, se obtiene una línea d llamada **directriz**. En consecuencia, una elipse y una hipérbola tienen dos directrices, la parábola tiene una y la circunferencia ninguna.

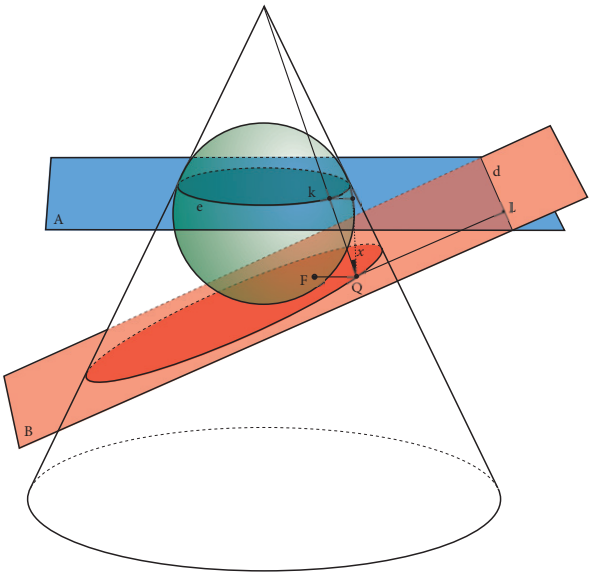


Figura 2.69 Directriz para el caso de la elipse.

Ya vimos que cada punto Q de la cónica satisface $\mathrm{d}|Q,F| = \mathrm{d}|Q,K|$, donde K es el punto de intersección de e con la recta PQ y F es el punto donde la esfera toca al plano de la cónica. Sea x la distancia del punto Q al plano A. Por lo tanto, se tiene que $x = \cos(\alpha)\mathrm{d}|Q,K|$ (aquí recordamos que α es el ángulo entre la recta PQ y el eje ℓ). Sea L el punto de la directriz más cercana al punto Q —en consecuencia, QL y d son perpendiculares—, entonces $x = \cos(\beta)\mathrm{d}|Q,L|$, donde β es el ángulo entre ℓ y B. De ahí obtenemos que:

$$\mathrm{d}|Q,F| = \varepsilon \mathrm{d}|Q,d|, \quad \varepsilon = \frac{\cos(\beta)}{\cos(\alpha)}$$

El valor ε es la **excentricidad** de la cónica. La siguiente tabla muestra la excentricidad para los diferentes casos.

CÓNICA	ÁNGULO ENTRE ℓ Y B	EXCENTRICIDAD
circunferencia	$\beta = 90°$	$\varepsilon = 0$
elipse	$90° > \beta > \alpha$	$0 < \varepsilon < 1$
parábola	$\beta = \alpha$	$\varepsilon = 1$
hipérbola	$\beta < \alpha$	$\varepsilon > 1$

Lo anterior nos proporciona una descripción de la parábola como lugar geométrico: la parábola es el lugar geométrico de todos los puntos Q, tal que $\mathrm{d}|Q,F| = \mathrm{d}|Q,d|$, donde el punto F es el **foco** y la recta d es la directriz de la parábola.

Los planetas orbitan alrededor del Sol siguiendo trayectorias elípticas. Esto fue un hallazgo importante para la cultura y más aún si vemos que las elipses son casi circunferencias —se trata de elipses con excentricidad muy baja—. Por ejemplo, la órbita de la Tierra tiene una excentricidad de $\varepsilon = 0.0167$, mientras la de Marte es de $\varepsilon = 0.0933$.

2.13.4 La forma de una antena

Recordemos que la elipse es el lugar geométrico de los puntos que tienen una constante suma de distancias h a los dos focos. De esta manera, el plano se divide en **curvas de nivel** para diferentes valores de h.

Veamos cómo esta división del plano en curvas de nivel tiene una consecuencia inesperada: cualquier rayo que sale de uno de los focos, es reflejado en la elipse —como si fuera un espejo— y llega justo al otro foco. Esto se conoce como la propiedad de la reflexión de la elipse.

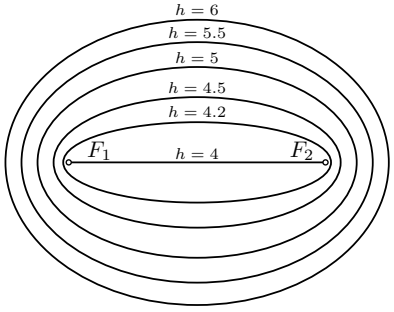


Figura 2.70 Partición del plano en una infinidad de curvas de nivel.

Para entender esta propiedad, consideremos la elipse definida para un valor h. Ahora bien, si Q es un punto de la elipse, entonces definimos como b a la recta que pasa por Q e interseca

a la rectas F_1Q y F_2Q con el mismo ángulo. Es decir, b es el espejo para que el rayo F_1Q se refleje en Q hacia el punto F_2, como se observa en la figura 2.71. Luego definimos al punto L como el reflejo de F_2 en b. Por ello, se tiene que $\mathrm{d}|Q, F_2| = \mathrm{d}|Q, L|$ y entonces:

$$\mathrm{d}|F_1, Q| + \mathrm{d}|Q, L| = \mathrm{d}|F_1, Q| + \mathrm{d}|Q, F_2| = h.$$

Falta demostrar que b es tangente a la elipse, es decir, que cualquier punto R de b —con $R \neq Q$— está afuera de la elipse, o dicho de otra manera, que $\mathrm{d}|R, F_1| + \mathrm{d}|R, F_2| > h$. Por lo anterior, sigue que $\mathrm{d}|R, F_2| = \mathrm{d}|R, L|$ por la desigualdad del triángulo en ΔF_1LR, el lado F_1L tiene menor longitud que la suma de los lados F_1R y RL. Si juntamos ambos argumentos obtenemos que:

$$\mathrm{d}|R, F_1| + \mathrm{d}|R, F_2| = \mathrm{d}|F_1, R| + \mathrm{d}|R, L| > \mathrm{d}|F_1, Q| + \mathrm{d}|Q, L| = h.$$

lo cual muestra que b es tangente a la elipse.

La propiedad de reflexión se usa en la práctica, por ejemplo, en hornos especiales que tienen una forma de elipsoide —un cuerpo que se obtiene al rotar una elipse en el espacio por el eje F_1F_2—, donde en un foco se coloca una fuente de calor y en el otro, el objeto a calentar. Otro caso ocurre en el Desierto de los Leones, donde la bóveda elipsoidal de la "Capilla de los susurros" tiene el siguiente efecto: lo que susurra una pareja en uno de sus focos, se escucha perfectamente bien en el otro.

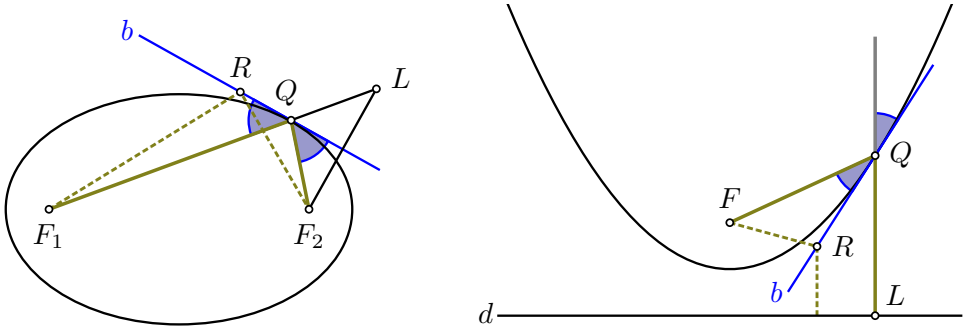


Figura 2.71 La propiedad de reflexión en la elipse y la parábola.

De manera similar se puede ver que en la parábola hay una propiedad de reflexión. Si Q es un punto de la parábola, dibujemos a la recta b que pasa por Q, tal que incluya los mismos ángulos con FQ y QL —donde F es el foco y L el pie de la perpendicular a la directriz d por Q— como se observa en la figura 2.71. Sólo hay que demostrar que b es tangente para terminar de concluir que el reflejo de FQ en la parábola es perpendicular.

El hecho de que b sea tangente se verifica de manera sencilla: para cualquier punto R de b —con $R \neq Q$ —, se tiene que $\mathrm{d}|F, R| = \mathrm{d}|L, R| > \mathrm{d}|L', R|$ donde L' es el pie de R en d. Por lo tanto, R está bajo la parábola.

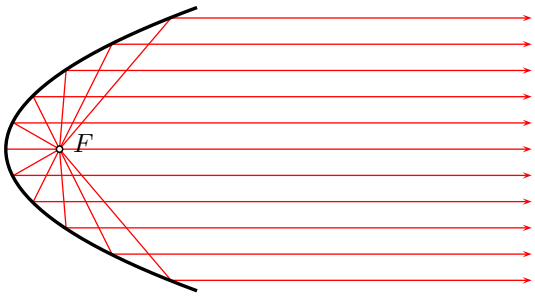


Figura 2.72 Propiedad de reflexión de la parábola.

Esta particularidad de la parábola se aplica en los faros de los automóviles que son paraboloides, es decir, tienen la forma que se obtiene al rotar una parábola en su eje de simetría. El foquito luminoso se coloca en el foco del paraboloide y, de esta manera, se obtiene un haz de luz que se dispersa muy poco y puede alumbrar lejos. El mismo efecto pero a la inversa se usa en las antenas de astronomía y también en las que reciben señales de satélites para la televisión. En ambos casos se reflejan ondas que llegan prácticamente paralelas en un paraboloide que los hace converger en el foco donde se encuentra el receptor. Dicha manera de agrupar los rayos tiene el efecto de multiplicar la señal, es decir, de aumentar su intensidad.

2.13.5 Ecuaciones de segundo grado

La última parte de la descripción de las cónicas es la analítica, es decir, queremos describir las cónicas con ecuaciones. Para ello usamos la descripción de las cónicas como lugares geométricos, mediante un foco y la directriz correspondiente. Por consiguiente, la cónica es el conjunto de puntos Q tal que se satisface la ecuación: $\mathrm{d}|Q,F| = \varepsilon \mathrm{d}|Q,d|$, donde ε es la excentricidad, F es un foco y d la directriz correspondiente.

Podemos elegir nuestro sistema de coordenadas de manera que los cálculos sean lo más sencillos posible. Una manera de hacerlo es al poner el origen en el foco y exigir que la directriz sea paralela al eje de coordenadas x. Entonces, calculamos las dos distancias $\mathrm{d}|Q,F| = \sqrt{x^2+y^2}$ y $\mathrm{d}|Q,d| = y+k$ donde $k = \mathrm{d}|F,d|$.

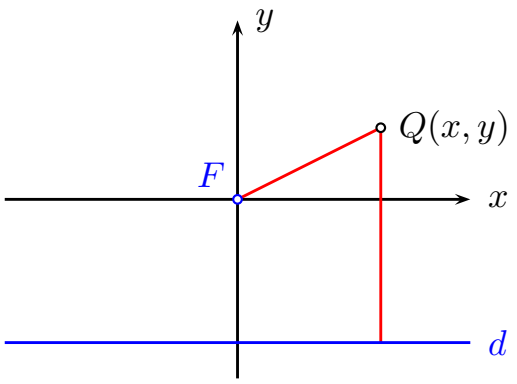


Figura 2.73 Distancias al foco y a la directriz de un punto Q.

Al sustituir estas expresiones en (7) y elevar al cuadrado, se obtiene:

$$x^2 + y^2 = \varepsilon^2 y^2 + 2\varepsilon^2 k y + \varepsilon^2 k^2 .$$

Como ε y k son números fijos para cada cónica dada, se puede escribir esta ecuación como:

$$x^2 + (1-\varepsilon^2)y^2 + (-2\varepsilon^2 k)y + (-\varepsilon^2 k^2) = 0 \qquad (8)$$

Si definimos $a = 1,\ b = 0,\ c = 1-\varepsilon^2,\ d = 0,\ e = -2\varepsilon^2 k,\ f = -\varepsilon^2 k^2$, entonces vemos que la expresión (8) es un caso particular de la **ecuación general de segundo grado en dos variables**:

$$ax^2 + bxy + cy^2 + dx + ey + f = 0.$$

Debe resaltarse que ya demostramos que cada cónica puede escribirse de la forma $a = 1$, $b = 0$, $d = 0$.

No es tan claro que cualquier ecuación de segundo grado con dos variables define una cónica. En efecto, se puede demostrar que siempre define una cónica, pero la demostración es bastante larga y engorrosa.

La idea es cambiar el sistema de coordenadas de tal manera que la ecuación se simplifica lo suficiente como para reconocer que, en efecto, se trata de una cónica. Resaltamos aquí que, en ciertos sistemas de coordenadas, las ecuaciones de las cónicas toman una forma muy especial. Por ejemplo, cualquier ecuación de la forma:

$$y = ax^2 + bx + c,$$

con $a \neq 0$, define una parábola. Podemos cambiar el sistema de coordenadas para escribirla en forma aún más sencilla, como:

$$y = ax^2.$$

Para la elipse, se puede encontrar un sistema de coordenadas de manera que la ecuación toma la forma:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$$

y la hipérbola se representa como:

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1.$$

2.13.6 Cónicas en la física

Galilei argumentó que el movimiento sobre la superficie de la Tierra encuentra en la dirección vertical hacia la Tierra una aceleración constante y, que esto, resulta en una **función cuadrática**:

$$z(t) = at^2 + bt + c \qquad (9)$$

para la altura $z(t)$ y en función del tiempo t. Sin embargo, la dirección horizontal no experimenta aceleración alguna. Se trata de un movimiento uniforme y la posición cambia de manera lineal:

$$h(t) = dt + e\,.$$

Si $d \neq 0$, entonces es posible despejar la variable t que luego se puede sustituir en (9). Lo que resulta es una ecuación de la forma:

$$z = a'h^2 + b'h + c',$$

donde a', b' y c' son coeficientes reales. Como vimos antes, esta expresión corresponde a una parábola. El movimiento sobre la superficie terrestre —de cualquier cuerpo pesado y

extenso— es, en muy buena aproximación, parabólico —no toma en cuenta la fricción del aire, el hecho de que la Tierra sea redonda ni el que "vertical" en un lugar determinado no es paralelo a "vertical" en otro, por lo que no se trata de una ecuación en un sistema de coordenadas ortonormales —ejes verticales con la misma escala.

Con más trabajo se puede ver que la primera ley de Newton tiene como consecuencia el que cualquier cuerpo que orbita un cuerpo pesado —que atrae todos los cuerpos hacia sí— tiene que viajar a lo largo de una trayectoria, que resulta ser una cónica. Como la única cónica que describe una trayectoria cerrada es la elipse —y la circunferencia, como un caso particular—, se obtiene que los planetas giran en elipses alrededor del Sol.

Por último, vale la pena repasar lo que se hizo en esta sección. Las cónicas se definieron como secciones cónicas; después se usaron las esferas de Dandelin para deducir la descripción de las cónicas como lugares geométricos —los griegos ya conocían estas propiedades, aunque Dandelin vivió en el siglo XIX—. En la actualidad se usa el argumento de Dandelin porque es más "elegante", es decir, más corto que los argumentos que dieron los griegos para concluir lo mismo. Al final, esbozamos el tratamiento de las cónicas desde el punto de vista de la geometría analítica, usando ecuaciones. La geometría analítica tiene su origen en el siglo XVII con trabajos de Descartes y de Fermat —quien demostró que cualquier ecuación de segundo grado con dos variables describe una cónica.

Como se puede ver, el discurso dio unos saltos para adelante y para atrás en el tiempo y no siguió el desarrollo histórico. Esto es completamente intencional: un tratamiento histórico sería otra cosa. Pero, ¿por qué habrá introducido Dandelin sus esferas si ya se conocían las propiedades? En matemáticas, la historia nunca acaba, siempre se puede repensar un tema y verlo desde otro ángulo. Los objetos de las matemáticas están tan intensamente interrelacionados que, a veces, resulta que lo que era la causa se transforma en la consecuencia.

2.14 PROBABILIDAD Y ESTADÍSTICA, CALCULANDO EL AZAR

Figura 2.74 Jarrón romano | © Latin Stock México.

En la vida diaria hay acontecimientos que obedecen a causas conocidas y que podemos predecir, mientras otros son resultado de la casualidad. De los primeros se ocupan las ciencias naturales, de los segundos, la probabilidad. Que una moneda lanzada al aire describa una parábola y caiga al suelo es algo totalmente predecible de lo que se encarga la física. Los rebotes consecuentes a la caída y su posición final en el suelo —con una de sus caras hacia arriba— es algo, en cambio, extraordinariamente difícil de anticipar. Por lo tanto, cada vez que lanzamos la moneda el resultado es fortuito, fruto de la casualidad, del azar. Dentro del

margen de lo que sí sabemos es que, aproximadamente la mitad de las veces, la moneda caerá con una cara hacia arriba, y el resto con esa cara hacia abajo. Este conocimiento es justo el que compete al estudio de la **probabilidad**. Estos fenómenos casuales o del azar se llaman **aleatorios**.

Figura 2.75 Tique, diosa de la suerte.

En las civilizaciones antiguas se creía que algún dios se ocupaba de determinar los resultados de los sucesos aleatorios y, con ello, del destino. Se asocia a la diosa griega Tique —en griego, $Tυχη$ significa suerte— con esa actividad, mediante la cual decidía la fortuna de las personas y las ciudades. El nombre romano de Tique es, precisamente, Fortuna. El caos reinante en diversas épocas históricas se atribuía a una intensa labor de la caprichosa diosa, considerada hija de Hermes y Afrodita. Según Libanius, el reconocido retórico, el templo dedicado a Tique en Alejandría era uno de los más impresionantes del mundo helénico. ¡Cómo no edificarlo grandioso y rendirle adoración a alguien que controlara el resultado de todos los fenómenos aleatorios y que, por ende, tendría un poder casi absoluto sobre el destino de cada individuo! Todas las casualidades serían resultado de su antojo. Con seguridad, si alguien así existiera, todos querríamos contarnos entre sus protegidos.

El hombre moderno no cree ya en una deidad que determine los fenómenos aleatorios; sabe que el azar es impredecible e incontrolable, y que todo cuanto puede hacer con él es tratar de comprenderlo, sin buscar causas donde no las hay.

2.14.1 Necesidad de una teoría de la probabilidad

La teoría de la probabilidad es la rama de las matemáticas que se ocupa del azar desde el punto de vista cuantitativo. Sus inicios, que ocurren en el siglo XVII, son increíblemente recientes para ser un estudio que abarca fenómenos tan importantes y frecuentes. Quizá, el tardío surgimiento de la probabilidad como interés dentro de las matemáticas, se deba a que el hombre se resistía a creer que hubiera incidentes verdaderamente impredecibles, o a que es difícil imaginar cómo decir algo sobre situaciones gobernadas por el azar. En su larga historia intelectual, el ser humano detentaba logros tan grandes como llegar a predecir los eclipses, descubrir las leyes matemáticas del Sistema Solar, entender cómo se reproducen los seres vivos... Al aceptar la existencia de fenómenos verdaderamente aleatorios, parecía darse por vencido —acaso, claudicar— en la tarea de conocer la naturaleza.

En realidad, hay dos razones para ocuparse de la probabilidad: una práctica y otra filosófica. La primera es que muchas de las cosas que ocurren a nuestro alrededor no son previsibles; por ejemplo, para predecir el resultado del lanzamiento de una ficha, sería necesario conocer su impulso, la altura alcanzada, su aceleración de caída, la velocidad de giro en el momento del lanzamiento, en fin, poseer los detalles exactos y precisos de todo lo anterior y de la superficie sobre la que va a caer, incluyendo las propiedades elásticas del material del que está hecha; habría que conocer incluso la velocidad del aire en cada momento del

recorrido. Recabar estos datos es poco menos que imposible; además, aunque los tuviéramos, el problema matemático que habría que resolver sería de una extraordinaria complejidad. Nada justificaría tanto esfuerzo. Así que definitivamente claudicamos y aceptamos tratar el lanzamiento de fichas como un fenómeno aleatorio.

La razón filosófica es que la ciencia del siglo xx —concretamente la física cuántica— descubrió que el azar existe, no sólo como producto de una capacidad limitada de cálculo, sino como un hecho natural en sí mismo: es imposible determinar —simultáneamente— y con total exactitud, la posición y la velocidad de una partícula elemental como el electrón, o bien, predecir cuándo un átomo de una sustancia radiactiva va a decaer al emitir energía y convertirse en un átomo diferente. No se trata de incompetencia humana o falta de instrumentos adecuados, sino que ésa es parte esencial de la naturaleza misma de dichas partículas. No es aquí donde analizaremos este asunto, pero conviene señalarlo para que el lector sepa que, en la actualidad, ya no se concibe al mundo físico como determinista, al contrario, se sabe que, inexorable e irremediablemente, el azar interviene en él.

Cualquiera de las razones expuestas con anterioridad es lo suficientemente importante para aceptar a la probabilidad como una rama legítima del conocimiento; aunque el uso práctico motivó su estudio un tiempo antes. De hecho, los fenómenos aleatorios que primero ocuparon la mente de los matemáticos fueron los juegos de azar.

Escrito alrededor de 1560, encontramos el libro *Liber ludo aleae* sobre los juegos de azar de Girolamo Cardano; tenemos después la correspondencia —de 1654 a 1660— entre Pierre de Fermat y Blaise Pascal, que se considera como el inicio del cálculo de probabilidades.

Figura 2.76 Blaise Pascal y Pierre de Fermat | © Latin Stock México.

Aparentemente, el Chevalier de Méré, un jugador empedernido e inteligente, planteó un problema a Pascal que motivó aquella famosa relación epistolar entre él y Fermat. El problema en cuestión trataba sobre cómo sería justo repartir el dinero de las apuestas en un juego de azar si éste se suspendía antes de que hubiera un ganador. La solución requería de un cálculo preciso de las probabilidades que tenía cada participante de ganar el juego en caso de que hubiera continuado. En dicha correspondencia —es importante señalarlo— se establece por vez primera que es posible determinar las probabilidades con absoluta precisión bajo supuestos claros sobre la naturaleza y las reglas del juego en cuestión. Es decir, que hay una única manera lógica de determinar las probabilidades que cada jugador tiene de ganar el juego y éstas pueden emplearse para repartir de una manera justa la apuesta.

2.14.2 El problema con el que se inicia el cálculo de probabilidades

Supongamos que dos personas deciden participar en un juego que consiste en lanzar cinco veces una ficha al aire. Cada una aporta 50 fichas. Si salen más caras o soles que cruces o águilas, el primer jugador se llevará todas las fichas; en caso contrario, se las llevará el segundo jugador. Cuando se han lanzado 3 fichas y se ha visto que 2 son cara y 1 cruz, tiene que suspenderse el juego. ¿Cuántas fichas debe llevarse cada jugador?

Está claro que al parar quien tenía mayor probabilidad de ganar era el primer jugador, y por lo tanto parece justo que él se lleve más fichas que el segundo. Un árbitro imparcial podría decidir que el primer jugador se lleve todas las fichas, ya que en el momento de suspenderse el juego él llevaba ventaja, pero al segundo jugador esta decisión le parece injusta porque aún tenía posibilidades de ganar. El árbitro puede decidir también que, como el juego se suspendió, cada jugador recupere sus 50 fichas; el primer jugador protesta argumentando que esa repartición hubiera sido justa si el juego nunca hubiera comenzado, o se hubiera suspendido cuando se habían tirado dos fichas —una cara y la otra cruz—, pero que en la situación actual se está ignorando la ventaja que evidentemente él llevaba.

A estas alturas parece más justo decidir que cada quien se lleve un número de fichas proporcional a la probabilidad que tenía de ganar en el momento en que se detuvo el juego. Esto requiere esclarecer el concepto de probabilidad y encontrar cómo calcularla numéricamente, y es a lo que Fermat y Pascal se abocaron en su correspondencia. Para abordar la situación sería necesario superar acaloradas discusiones y, finalmente, acordar algunos puntos. Uno es que el resultado parcial de dos caras y una cruz no tiene ningún valor predictivo sobre los dos siguientes lanzamientos —pueden caer tanto cara como cruz—. Lo siguiente es cómo debería calcularse la probabilidad de que el vencedor fuera el primer jugador. Para ello, se calcula el número de casos que lo llevarían a ganar y el resultado se divide entre el número total de casos posibles —esto es válido sólo bajo la hipótesis de que todos los casos posibles sean igualmente probables—. Siguiendo esta línea de argumentación, los dos siguientes lanzamientos darían un total de cuatro posibles resultados:

$$\begin{array}{cc} cara & cara \\ cara & cruz \\ cruz & cara \\ cruz & cruz \end{array}$$

Como se observa, 3 de ellos darían la victoria al primer jugador y sólo 1 al segundo. Por tanto, la probabilidad de que el vencedor sea el primer jugador es de $\frac{3}{4}$, mientras que la del segundo jugador es de $\frac{1}{4}$, lo cual significa que uno debe llevarse 75 fichas y el otro 25. Sin embargo, si las reglas del juego cambian un poco, el problema se complica considerablemente: cabe mencionar que en matemáticas se define la probabilidad de cualquier evento como un número entre 0 y 1 que, si se multiplica por 100, representa una estimación del porcentaje de casos favorables respecto al total.

Supongamos que hay tres jugadores — A, B y C— que ponen la misma cantidad de fichas al jugar a los dados. Deciden otorgarse puntos de acuerdo con la siguiente regla: el jugador A gana un punto cuando el dado cae en 1 o 2, el jugador B lo gana cuando el dado cae en 3 o 4, y el jugador C si cae en 5 o 6. El juego termina cuando algún jugador ha ganado tres puntos y se lleva todas las fichas.

Supongamos que, en un momento dado, A tiene dos puntos y B y C tienen un punto cada uno y deciden suspender el juego. ¿Cómo repartir el lote de fichas? Habría que considerar todas las opciones posibles para cada una de las siguientes tiradas y calcular las pro-

babilidades de que cada jugador resulte ser el vencedor en esa jugada. La siguiente tabla muestra los posibles resultados de los siguientes tres lanzamientos; la letra A, B o C indica para quién fue el punto en esa tirada.

5ª tirada	A									B									C								
6ª tirada	A			B			C			A			B			C			A			B			C		
7ª tirada	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
ganador	A	A	A	A	A	A	A	A	A	A	A	A	B	B	B	A	B	C	A	A	A	A	B	C	C	C	C

Figura 2.77 Con rojo se indica la letra del ganador para cada tirada.

Ahora hay que contar cuántos de estos 27 casos dan la victoria a cada jugador. Para facilitar la cuenta se ha marcado con negritas cuando el jugador, en ese paso, es ya el ganador. Todos los casos debajo de una letra marcada con negritas le darían la victoria al mismo jugador y de hecho, en la práctica, esas jugadas ya no se llevarían a cabo. El resultado es:

$$\begin{aligned} 17 &\text{ dan la victoria a } A \\ 5 &\text{ dan la victoria a } B \\ 5 &\text{ dan la victoria a } C \end{aligned}$$

Por lo tanto, al suspenderse el juego A debe llevarse $\frac{17}{27}$ fichas del lote, mientras que B y C tendrán $\frac{5}{27}$ fichas cada uno. En números: si cada jugador hubiera apostado 9 fichas —el lote completo sería de 27 fichas—, A debería quedarse con 17, B con 5 y C con las otras 5.

Llegar a esta solución llevó a Pascal y Fermat —dos de los matemáticos más reconocidos de la historia— un esfuerzo considerable, pero su relevancia es crucial ya que señaló el camino hacia un método general para calcular las probabilidades que puede explicarse de manera muy sencilla:

> La probabilidad de un evento es el número de casos favorables dividido por el número total de casos, siempre y cuando todos los casos sean igualmente probables.

Éste es el principio básico que sirve como fundamento al cálculo de probabilidades aplicado a los juegos de azar y a muchas otras situaciones en las que es posible construir un modelo basado en casos equiprobables, es decir, que tienen la misma probabilidad. En algunas situaciones concretas, llevar a cabo este cálculo puede ser extremadamente complicado, pero la regla siempre es clara. Quien desee calcular probabilidades debe entender bien este principio.

2.14.3 El modelo matemático general de la probabilidad

Entre el siglo XVII y el XX se realizaron muchos avances en la teoría de la probabilidad, pero no fue sino hasta 1930 cuando el trabajo del famoso matemático ruso Andrei Nikolaievich Kolmogórov la fundó sobre bases sólidas, aprovechando un desarrollo del análisis matemático llamado **teoría de la medida**. La idea intuitiva de este concepto es muy simple. Pensemos en la masa contenida en cada región del espacio. Hay regiones grandes que pueden contener poca masa, y otras pequeñas que pueden contener mucha, como un trozo de plomo. La función matemática que asocia cada región del espacio con la masa que contiene es un ejemplo de una **medida**. Una medida tiene la propiedad de ser **aditiva en conjuntos ajenos**, es decir, si A y B son dos regiones ajenas del espacio, la masa contenida en A y B es la masa contenida en A más la masa contenida en B. En símbolos, lo anterior se escribe así:

$$m(A \cup B) = m(A) + m(B) \quad \text{si} \quad A \cap B = \emptyset$$

donde $A \cup B$ se lee "A **unión** B" y representa el conjunto de todos los puntos del espacio que están en A o B, mientras que $A \cap B$ se lee "A **intersección** B" y representa los puntos que están tanto en A como en B. El símbolo $\emptyset$ es el conjunto vacío, pues no tiene ningún elemento, por lo que la igualdad $A \cap B = \emptyset$ significa que no hay ningún punto que esté simultáneamente en A y B, o bien, que el conjunto de puntos en A y B es vacío.

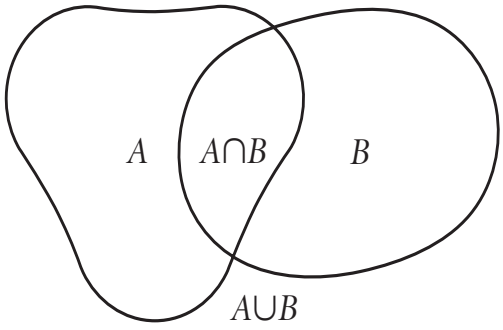


Figura 2.78 Conjuntos que se intersecan.

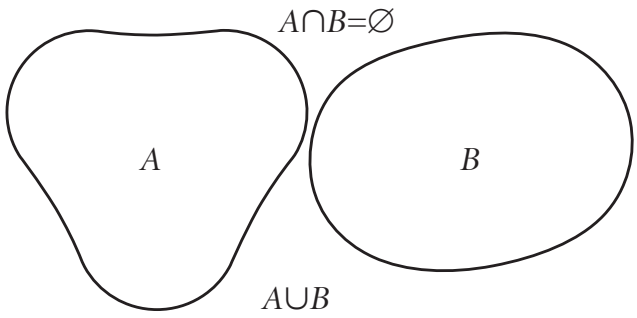


Figura 2.79 Conjuntos ajenos, es decir, con intersección vacío.

La notación de conjuntos presentada brevemente en el párrafo anterior, junto con el concepto de medida, constituyen la herramienta idónea para trabajar la teoría de la probabilidad.

Un **espacio de probabilidad** o **espacio muestral** se define como un conjunto Ω —la letra griega conocida como "omega"— que se interpreta como el de todos los posibles resultados de un experimento; una familia Σ —"sigma" — de subconjuntos de Ω, que se interpreta como los eventos del experimento cuyas probabilidades se conocen o pueden ser calculadas; y una medida p es una función que asigna, a cada evento E de Σ, un número entre 0 y 1. $p(E)$ se interpreta como la probabilidad del evento E, y debe asignar el valor de 0 al conjunto vacío y 1 al total, es decir, $p(\emptyset) = 0$ y $p(\Omega) = 1$.

Por ejemplo, en el caso más simple del lanzamiento de una ficha, Ω consta de dos elementos: *cara* y *cruz*. Σ son todos los subconjuntos de Ω: el total Ω, el vacío $\emptyset$, el que contiene sólo a *cara* y el que contiene sólo a *cruz*. La medida de probabilidad p asigna a estos eventos los valores de 1, 0, $\frac{1}{2}$ y $\frac{1}{2}$, respectivamente.

Para el caso de conjuntos finitos, se especifica un conjunto como la lista de todos sus elementos encerrada entre llaves. Por ejemplo, el evento que ofrece *cara* como el resultado del experimento, se denota por $\{cara\}$, mientras que el evento imposible se denota por el conjunto vacío $\emptyset$, que puede escribirse también como $\{\}$. Al aprovechar estas convenciones, la descripción completa del modelo matemático del experimento aleatorio cuando se lanza una ficha, puede presentarse por la terna (Ω, Σ, p), cuyos elementos quedan totalmente especificados a continuación:

$$\begin{aligned}\Omega &= \{cara, cruz\}\\ \Sigma &= \{\{cara, cruz\},\ \{\},\ \{cara\},\ \{cruz\}\\ p(\{cara, cruz\}) &= 1\\ p(\{\}) &= 0\\ p(\{cara\}) &= \frac{1}{2}\\ p(\{cruz\}) &= \frac{1}{2}\end{aligned}$$

El caso del lanzamiento de un dado queda descrito por la terna (Ω, Σ, p), donde $\Omega = \{1, 2, 3, 4, 5, 6\}$ y los eventos son todos los subconjuntos de Ω, es decir, Σ consta de cada uno de estos subconjuntos de Ω —que son muchos—, como: $\{\}$, $\{3\}$, $\{4, 5, 6\}$, $\{2, 3\}$, $\{1, 2, 3, 4, 6\}$, etc. La probabilidad p asigna a cada evento el número de sus elementos entre seis. Por ejemplo:

$$p(\{2, 3, 5, 6\}) = \frac{4}{6} = \frac{2}{3}.$$

Este modelo es aplicable a situaciones donde los posibles resultados de un experimento no son equiprobables, aunque en estos casos, para estimar los valores de las probabilidades habría que recurrir a métodos estadísticos o a suposiciones cuantitativas que podrían no estar bien justificadas.

Veamos el caso de un dado cargado: puede tener la característica de que los números 2, 3, 4 y 5 y aparezcan con probabilidad $\frac{1}{6}$ pero, en cambio, el 6 aparece con probabilidad $\frac{2}{9}$ y el 1 con probabilidad de $\frac{1}{9}$. Como:

$$\frac{1}{9} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{2}{9} = \frac{4}{6} + \frac{3}{9} = \frac{2}{3} + \frac{1}{3} = 1,$$

entonces definimos p de manera que $p(E)$ se calcula sumando tantos $\frac{1}{6}$ como números entre 2 y 5 contenga, más $\frac{1}{9}$ si contiene al 1 y $\frac{2}{9}$ si contiene al 6. Por ejemplo:

$$p(\{3, 4, 6\}) = \frac{2}{6} + \frac{2}{9} = \frac{6+4}{18} = \frac{10}{18} = \frac{5}{9}.$$

Puede comprobarse que ésta es una buena definición y que podría modelar el experimento aleatorio de lanzar un dado cargado, siempre y cuando las probabilidades de salir del 1 y el 6 fueran, respectivamente, $\frac{1}{9}$ y $\frac{2}{9}$.

Este modelo de espacio probabilístico puede aplicarse a una gran cantidad de situaciones prácticas mucho más complejas que las descritas en los párrafos anteriores, incluyendo casos en los que el espacio muestral Ω de posibles resultados sea infinito.

2.14.4 Probabilidad condicional, eventos independientes y variables aleatorias

Sean A y B dos eventos en cualquier espacio probabilístico. Supongamos que al realizar el experimento nos informan que el evento A ocurrió. ¿Eso nos dice algo acerca del evento B? En general sí. Por ejemplo, consideremos estos dos eventos en el experimento de lanzar un dado:

$A =$ “el resultado es par” y
$B =$ “el resultado es mayor que 3”.

La probabilidad de A es claramente $\frac{1}{2}$ y la de B también es $\frac{1}{2}$. Pero si alguien nos informa que A ocurrió, ¿seguirá B teniendo la misma probabilidad? La respuesta es no. Saber que el resultado es par implica que ya sólo quedan tres posibilidades: 2, 4 y 6 y, como éstas son equiprobables, la probabilidad de que el resultado sea mayor que 3 es ahora $\frac{2}{3}$.

¿Entonces la probabilidad del evento B no estaba bien definida porque antes era $\frac{1}{2}$ y ahora es $\frac{2}{3}$? En realidad lo que sucede es que saber la ocurrencia del evento A nos lleva a una nueva situación probabilística, a un nuevo espacio de probabilidad. El espacio muestral en la nueva situación ya no es $\{1,2,3,4,5,6\}$ sino $\{2,4,6\}$. Como estos cambios de espacio probabilístico se dan mucho en el estudio de la probabilidad, se ha desarrollado un concepto que permite referirse a estas situaciones nuevas manteniendo el lenguaje de la situación original, lo cual resulta cómodo en la práctica. Se trata del concepto de **probabilidad condicional**. El evento de que ocurra B cuando se garantiza la ocurrencia de A, se denomina "B dado A" y se denota por $B|A$. Las probabilidades de estos eventos son: $p(A)=\frac{1}{2}$, $p(B)=\frac{1}{2}$ y $p(B|A)=\frac{2}{3}$. La probabilidad condicional puede definirse así:

$$p(B|A)=\frac{p(A\cap B)}{p(A)}$$

y se puede interpretar como otro espacio de probabilidad donde el espacio muestral es A. Observemos que esta relación se cumple en el ejemplo anterior. En efecto, $p(A\cap B)=\frac{1}{3}$, pues $A\cap B=\{4,6\}$ y, por tanto:

$$\frac{2}{3}=p(B|A)=\frac{p(A\cap B)}{p(A)}=\frac{\frac{1}{3}}{\frac{1}{2}}=\frac{2}{3}.$$

El concepto intuitivo de probabilidad condicional corresponde de manera adecuada a la definición formal.

La probabilidad condicional es uno de los dos conceptos claves de la teoría de la probabilidad. El otro es el de **eventos independientes**, que se describe a continuación.

Consideremos otra vez el experimento de lanzar un dado pero ahora con dos nuevos eventos A y B:

$A=$ "el resultado es impar" y
$B=$ "el resultado es 3 o 4".

Entonces $p(A)=\frac{1}{2}$ y $p(B)=\frac{1}{3}$. Saber que A ocurre nos dice que las únicas posibilidades son 1, 3 y 5, todas igualmente probables; por tanto, $p(B|A)=\frac{1}{3}$. En este caso la probabilidad de B y la de "B dado A" son iguales. En estas circunstancias se dice que B es independiente de A. En términos de probabilidades, decimos que B es independiente de A si $p(B|A)=p(B)$. Usando la definición de probabilidad condicional obtenemos que si B es independiente de A, entonces:

$$\frac{p(A\cap B)}{p(A)}=p(B)$$

o en forma equivalente:

$$p(A\cap B)=p(A)p(B).$$

Esto muestra que el concepto de eventos independientes es simétrico, es decir, B es independiente de A si y sólo si A es independiente de B. Por tal motivo, se habla de eventos independientes sin especificar cuál de ellos es independiente del otro. Como corolario ob-

tenemos que si A y B son eventos independientes, entonces $p(B|A) = p(B)$ y $p(A|B) = p(A)$.

Se dice que dos eventos son **mutuamente excluyentes** si la ocurrencia de cualquiera de ellos implica que la ocurrencia del otro es imposible. Por ejemplo, en el experimento de lanzar un dado y observar el número en la cara de arriba, los siguientes eventos $A =$ "el número obtenido es par" y $B =$ "el número obtenido es 5" son mutuamente excluyentes. Que dos eventos sean mutuamente excluyentes equivale a que los conjuntos de resultados que los representan son ajenos, es decir, dos eventos A y B son mutuamente excluyentes si y sólo si $A \cap B = \emptyset$. Es importante observar que este concepto no es probabilístico sino únicamente lógico y que la probabilidad no interviene en su definición; sin embargo, con frecuencia se comete el error de confundirlo con la independencia de eventos explicada anteriormente.

Una **variable aleatoria** es una función real definida en el espacio muestral de un espacio probabilístico (Ω, Σ, p), en otras palabras, es una función:

$$X : \Omega \longrightarrow \mathbb{R}$$

Las variables aleatorias, a pesar de ser funciones, se suelen denotar con letras $X, Y, \ldots$ mayúsculas en lugar de con una f. Se dice que dos variables aleatorias X_1 y X_2 son independientes si los eventos que se pueden definir con una y otra son siempre independientes, es decir, si eventos como $\{\omega \in \Omega : a_1 \leq X_1(\omega) \leq b_1\}$ y $\{\omega \in \Omega : a_2 \leq X_2(\omega) \leq b_2\}$ son independientes.

Intuitivamente, dos variables aleatorias son independientes si el saber algo de una de ellas no nos da ninguna información acerca de la otra. Por ejemplo, en el experimento de lanzar dos dados, saber el resultado de uno de ellos no nos da información acerca del otro. Las variables aleatorias X_1 y X_2 definidas como el número en la cara superior del primer dado y el mismo número pero del segundo dado, respectivamente, son variables independientes.

Las variables independientes desempeñan un papel fundamental en el **teorema del límite central**, uno de los resultados más importantes y útiles de la probabilidad, que se aplica en la **estadística**.

2.14.5 El lanzamiento de canicas a una pared y la distribución normal

Una muestra sencilla de un experimento aleatorio —donde el espacio muestral sea infinito— es el lanzar rodando por el suelo una canica hacia un punto en la base de una pared, y registrar, para cada lanzamiento, el punto en el que la canica toca la pared. Si los lanzamientos los hace siempre la misma persona, es posible construir un modelo matemático que se aproxime bastante a la realidad. Comencemos haciendo algunas simplificaciones: supongamos que las canicas llegarán con igual probabilidad a la izquierda y a la derecha del blanco —que está ubicado en el punto b—. En segundo lugar, supondremos que es más probable que caigan en un intervalo A cercano a b, que en otro intervalo B de la misma longitud pero alejado de b. Necesitamos una función que sea simétrica respecto al origen, que sea mayor a medida que nos acercamos a b y menor mientras nos alejamos, y que el área bajo su gráfica sea igual a 1. Todas estas propiedades las satisfacen las funciones definidas para cualquier número positivo σ.

A la función $N_{b,\sigma}(x)$ se le llama de distribución normal con **media** b y **varianza** σ^2. También se le conoce como la función gaussiana —en honor a Gauss— o distribución de campana, debido a la forma de su gráfica parecida al perfil de una campana. En la figura 2.80 se muestran tres ejemplos de estas funciones con un mismo valor de $b = 0$ y distintos valores de σ.

$$N_{b,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-b)^2}{2\sigma^2}}$$

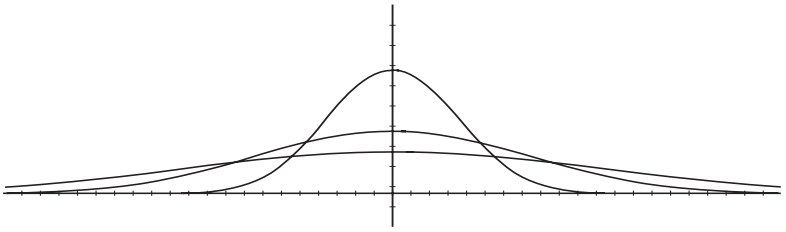


Figura 2.80 Distribuciones normales con varianzas distintas.

La gráfica con varianza pequeña de $\sigma = 0.5$ corresponde a un lanzador de canicas con buen tino, la de varianza $a = 1$ a un jugador medio, y la de varianza $a = 1.5$ a uno bastante errático. Supongamos que el jugador de nuestro experimento tiene buen tino, por lo que su varianza es $\sigma = 0.5$. La probabilidad $p([x_1, x_2])$ del evento que consiste en que la canica caiga en el intervalo $[x_1, x_2]$ está dada por el área bajo la gráfica de la curva normal $N_{0,0.5}$ entre x_1 y x_2 —sombreada en la figura 2.81—, lo cual se representa con la integral definida:

$$p([x_1, x_2]) = \int_{x_1}^{x_2} N_{0,0.5}(x)\mathrm{d}x$$

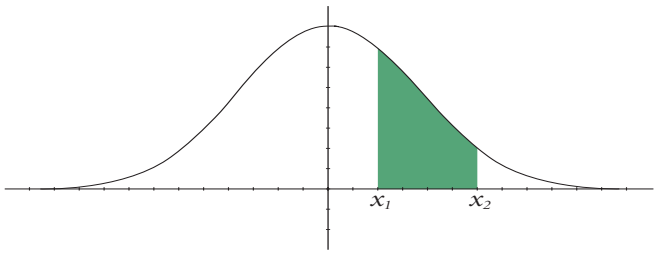


Figura 2.81 Interpretación gráfica —que se representa como el área bajo la curva— de la probabilidad de que la observación caiga entre x_1 y x_2.

y se ilustra como el área bajo la curva en la figura 2.81.

Calcular estas integrales es difícil, antes se hacía usando tablas y hoy en día puede recurrirse a la computadora. La clave es que nuevamente pudimos aplicar el modelo general. En este caso, la medida de probabilidad p se definió mediante una integral definida, es decir, como el área bajo una curva.

Las medidas de probabilidad que se utilizan en las diferentes aplicaciones de la teoría llegan a definirse incluso en espacios de dimensión infinita —como ocurre cuando se modelan los llamados **procesos estocásticos**— en los que cada elemento del espacio muestral puede ser, por ejemplo, una curva.

2.14.6 La ley de los grandes números

Quizá el resultado más general e importante de la teoría de la probabilidad es el que se denomina **ley de los grandes números** y afirma que, a la larga, la frecuencia observada de las ocurrencias de un evento se acerca o converge a la probabilidad del mismo evento. Esta convergencia es lo que da sentido y aplicabilidad a la teoría de la probabilidad. Sin embargo, el acercamiento puede ser muy lento y esa lentitud es causa de no pocos errores en la interpretación de los resultados. Las observaciones realizadas casi siempre parecen indicar que la realidad difiere mucho del modelo probabilístico, que no hay una buena relación entre las

frecuencias observadas —los cocientes entre el número de observaciones favorables entre el total de observaciones— y las probabilidades. En muchas ocasiones, el resultado de un experimento aleatorio parece ser el de uno muy diferente al que estamos aplicando.

Por ejemplo, al realizar 10 veces el experimento de lanzar diez dados hemos obtenido los siguientes resultados:

$$\begin{array}{cccccccccc}
1 & 2 & 3 & 2 & 4 & 4 & 3 & 4 & 4 & 2 \\
6 & 4 & 2 & 1 & 5 & 2 & 3 & 4 & 3 & 1 \\
5 & 2 & 2 & 5 & 6 & 4 & 3 & 5 & 5 & 6 \\
2 & 2 & 2 & 2 & 2 & 2 & 6 & 3 & 1 & 4 \\
6 & 1 & 2 & 1 & 4 & 1 & 5 & 1 & 3 & 2 \\
2 & 5 & 1 & 3 & 2 & 3 & 1 & 5 & 6 & 5 \\
4 & 3 & 3 & 6 & 2 & 4 & 6 & 5 & 2 & 1 \\
3 & 4 & 4 & 2 & 2 & 2 & 2 & 6 & 4 & 5 \\
1 & 2 & 1 & 6 & 3 & 1 & 6 & 6 & 5 & 5 \\
1 & 1 & 2 & 1 & 3 & 2 & 2 & 6 & 5 & 2
\end{array}$$

Figura 2.82 Resultados de cien lanzamientos de un dado.

En la primera línea no aparecen ni el 5 ni el 6, lo cual haría creer que se trata de los resultados de un dado tetraédrico —de cuatro lados— y no cúbico. La cuarta línea parece indicar que el 2 es altamente probable. Las frecuencias de aparición de cada número son:

$$\begin{aligned}
fr(1) &= \frac{17}{100} \\
fr(2) &= \frac{28}{100} \\
fr(3) &= \frac{14}{100} \\
fr(4) &= \frac{14}{100} \\
fr(5) &= \frac{14}{100} \\
fr(6) &= \frac{13}{100}
\end{aligned}$$

y parecen indicar un fuerte sesgo hacia el 2, como si el dado estuviera "cargado" favoreciendo al 2 sobre los otros números. Si hiciéramos otra vez el mismo experimento podríamos observar ciertas peculiaridades, distintas en cada ocasión.

Si repetimos muchas veces el experimento y calculamos las frecuencias para $10\,000$, $1\,000\,000$, $100\,000\,000$ observaciones, veremos que difieren cada vez en menos de $\frac{1}{6}$. No obstante, para llegar a observar claramente esta convergencia sería necesario recurrir a muchísimas observaciones.

En general, dos motivos contribuyen a la impresión de que las muestras resultantes de un experimento aleatorio no son tan aleatorias como se esperaría. Primero, la mente humana está entrenada para detectar irregularidades y, al observar un campo aleatorio, lo regular es precisamente lo que resalta. El segundo es que en muestras relativamente pequeñas es altamente probable que haya alguna desviación importante.

Se sabe de experimentos famosos en que el investigador manipuló ligeramente los datos para que señalaran con mayor claridad el resultado que se deseaba probar. En realidad, para probar una hipótesis en situaciones aleatorias es necesario recurrir a los métodos de la estadística, diseñados para tomar en cuenta la ley de los grandes números y, a pesar de la lentitud de la convergencia, se pueden obtener resultados confiables.

Para aplicar la teoría de la probabilidad a la vida cotidiana se requiere tener muy clara la lentitud con la que las frecuencias convergen a las probabilidades y también que, aunque los eventos poco probables ocurren escasas veces, hay muchos —por lo tanto la probabilidad

de observar alguno de ellos es bastante alta. Para aclarar este último punto, recurramos al clásico ejemplo conocido como la paradoja del cumpleaños.

2.14.7 La paradoja del cumpleaños

¿Qué probabilidad hay de que en un auditorio con 50 personas, dos cumplan años el mismo día? La mayoría creemos que esta probabilidad es muy baja, pero un cálculo preciso nos muestra que, en realidad, es alta y se aproxima a 0.97.

¿Por qué esta discrepancia entre la intuición y el cálculo preciso? Pues porque solemos confundir esta pregunta con otra: ¿cuál es la probabilidad de que **yo** encuentre entre esas 50 personas otra con mi mismo cumpleaños? La respuesta a esta segunda pregunta es, aproximadamente, de $\frac{49}{365} \approx 0.13$. Si en el auditorio hubiera 367 personas, la probabilidad de que dos cumplan años el mismo día sería 1. La probabilidad de que entre 50 personas, dos hayan nacido en la misma fecha se calcula de la siguiente manera. Primero, se obtiene la probabilidad p_n de que n personas tengan diferentes fechas de nacimiento. No entraremos en detalles, pero una reflexión cuidadosa lleva a entender que el resultado debe ser:

$$p_n = \frac{365}{365} \cdot \frac{364}{365} \cdot \ldots \cdot \frac{365-n}{365}.$$

En particular, para $n = 50$, $p_{50} = 0.03$ y por tanto, la probabilidad de que en el auditorio de 50 personas dos tengan el mismo cumpleaños es de $1 - p_{50} \approx 0.97$. La siguiente tabla enlista los valores de las probabilidades para distintas cantidades de personas en el auditorio:

$$\begin{aligned} 1 - p_{10} &\approx 0.14 \\ 1 - p_{20} &\approx 0.44 \\ 1 - p_{30} &\approx 0.73 \\ 1 - p_{40} &\approx 0.90 \\ 1 - p_{50} &\approx 0.97 \\ 1 - p_{60} &\approx 0.99 \end{aligned}$$

En la teoría de la probabilidad es muy frecuente encontrar resultados como el de la paradoja del cumpleaños, que contradicen nuestra intuición ingenua. Este hecho demuestra que la probabilidad es un asunto muy delicado que no debe dejarse a la intuición de personas sin experiencia en sus cálculos.

2.14.8 El teorema del límite central

La distribución normal presentada en el ejemplo anterior aparece con mucha frecuencia en la teoría de la probabilidad y ello se debe a que es la forma natural de muchas distribuciones de probabilidad, precisamente lo que afirma uno de los grandes teoremas de la teoría, el **teorema del límite central**.

Para explicar este teorema, debemos definir lo que se entiende por una variable aleatoria. Toda función definida en el espacio muestral que satisfaga ciertas condiciones mínimas —que aquí no tiene sentido especificar— es una **variable aleatoria**. Por ejemplo, en el experimento de lanzar una ficha al aire, la función que vale 1 cuando el resultado es *cara*, y 0 cuando es *cruz*, es una variable aleatoria. En el experimento de lanzar un dado, la fun-

ción que da el número que aparece en la cara superior del dado, es una variable aleatoria. En el experimento de lanzar canicas hacia un punto de una pared, la función que asigna a cada resultado del experimento la posición x de la canica al chocar con la pared es también una variable aleatoria.

En general, una variable aleatoria es una interpretación numérica —muchas veces parcial e incompleta— del resultado de un experimento, es decir, es una manera numérica pero indirecta de observar los resultados. A las variables aleatorias se les podría llamar **observables**, como de hecho se hace en el ámbito de la mecánica cuántica, pues son formas de "observar" los resultados de un experimento aleatorio.

El enunciado preciso del teorema del límite central es complicado, pero el caso que nos interesa aquí es el que se refiere a la distribución de probabilidad de una variable aleatoria, aplicada a diferentes realizaciones de un mismo experimento. En el ejemplo del lanzamiento de un dado, si usamos la variable aleatoria X —valor del número que aparece en la cara superior del dado—, al repetir muchas veces el experimento vamos a encontrar que toma el valor de:

$$\frac{X_1 + \ldots + X_n - nb}{a\sqrt{n}}$$

y al graficar las frecuencias con las que aparece cada valor de este "promedio", entonces la gráfica de las frecuencias se parece cada vez más a la distribución normal.

Para visualizar este resultado, observemos las gráficas de la figura 2.83 en las que se han marcado con segmentos rojos las frecuencias de los valores obtenidos de los "promedios" de las primeras 5, 20 y 80 observaciones, respectivamente.

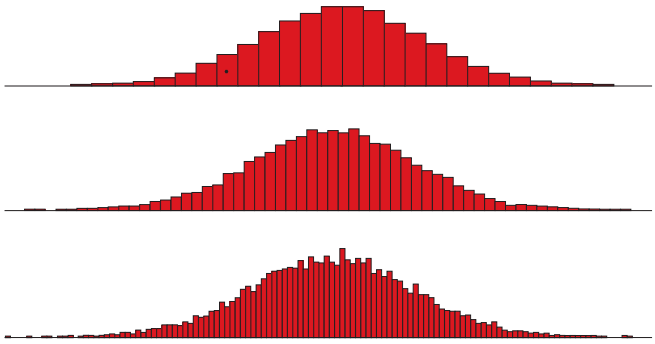


Figura 2.83 Histogramas que corresponden a un número grande de observaciones, con conjuntos de 5, 20 y 80 dados, respectivamente.

Vemos que el parecido de estos **histogramas** con la distribución normal es muy grande. Las observaciones aleatorias tienden a comportarse, en promedio, como una distribución normal. Debido a lo anterior, la distribución normal tiene muchas aplicaciones en la estadística.

Muchas veces se elabora la hipótesis de que una variable aleatoria cualquiera tiene una distribución normal y ésta se determina, sin muchos miramientos, a partir de una muestra de tamaño N que ha dado observaciones $x_1, x_2, \ldots, x_N$, tomando como media μ al promedio de los valores x_i y como varianza σ^2, al promedio de las cantidades $(x_i - \mu)^2$, es decir, se supone que la distribución de la variable aleatoria es:

$$N_{\mu,\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

donde:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x-\mu)^2.$$

En realidad μ y σ^2, calculadas de esta manera, son la media y la varianza de la muestra $x_1, x_2, \ldots, x_N$ y no de la distribución real. Identificar estos promedios con la media y la varianza de la distribución real es un abuso que, aunque ocasionalmente pueda ser útil, es causa de no pocos errores de interpretación.

Para obtener estimaciones adecuadas de las distribuciones de probabilidad de una variable aleatoria a partir de un conjunto de datos, hay que recurrir a los métodos especializados de la estadística.

2.14.9 La estadística

La estadística es el "arte" de obtener información a partir de datos conocidos. Permite medir y cuantificar la confianza de las conclusiones sobre un cierto problema, así como el tamaño del error que se comete. Si bien se puede definir en pocas palabras, en realidad consiste de un cúmulo enorme de conocimientos y recomendaciones encaminados a la necesidad de tomar decisiones con base en información incompleta —cosa que ocurre en casi todas las situaciones de la vida real—, de manera que se pueda hacer con plena conciencia de lo que la información disponible dice y, también, de lo que no dice.

La estadística se divide en dos grandes ramas, la **estadística descriptiva** y la **estadística inferencial**.

La estadística descriptiva consta de todos los métodos recomendados para presentar la información —en general, de forma gráfica— fácilmente comprensible y evitar posibles interpretaciones erróneas. Todos conocemos los diferentes tipos de tablas y gráficas que se muestran en la prensa y en todos los demás sitios donde se presenta información al público —algunos ejemplos en las figuras 2.84 a 2.86.

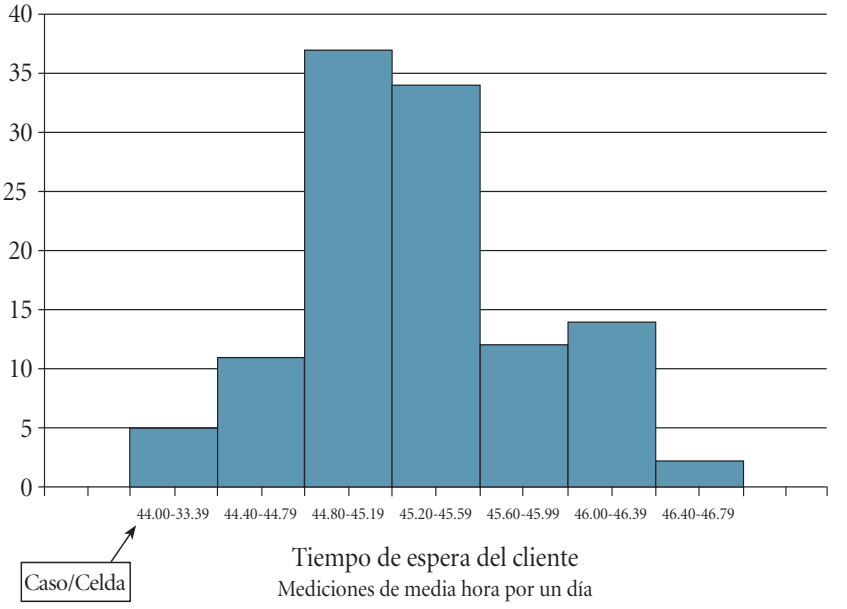


Figura 2.84 Histograma.

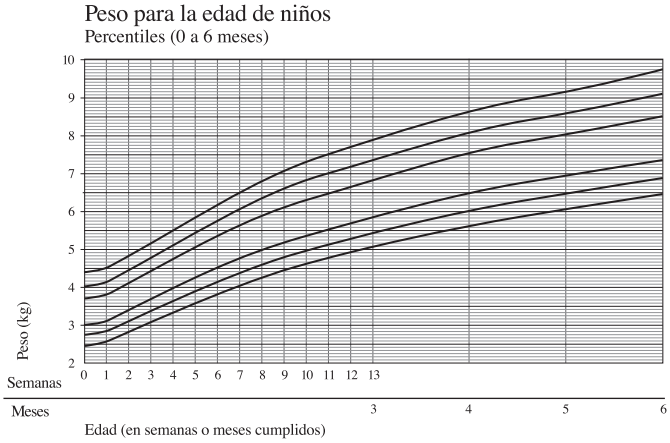


Figura 2.85 Gráfica con la evolución del peso de los bebés.

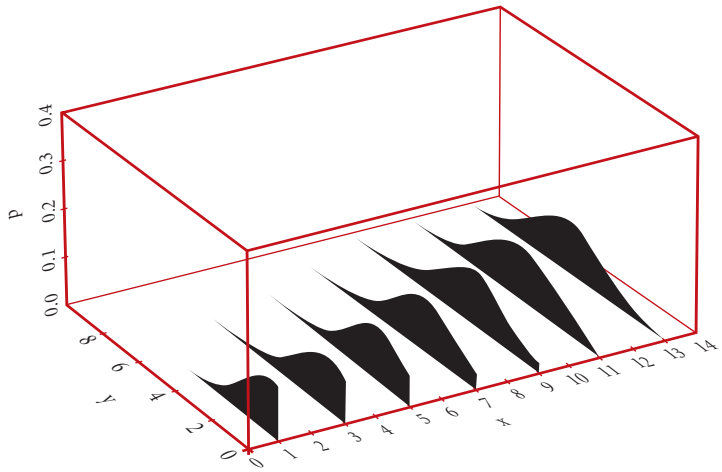


Figura 2.86 Gráfica en tres dimensiones.

Quienes publican gráficas estadísticas no siempre tienen el cuidado de seguir las recomendaciones de los expertos al elaborarlas y eso puede causar pérdidas de calidad en la información que se intenta transmitir. Estos problemas no son de carácter matemático, más bien caen en el ámbito de cómo se transmite la información. En particular, la estadística descriptiva hace muy poco uso —a veces ninguno— de la probabilidad o las matemáticas, a excepción de cálculos más o menos elementales.

En cambio, la estadística inferencial sí tiene un carácter matemático y depende en forma muy importante de modelos probabilísticos. Los métodos de la estadística inferencial son muchos y muy variados, así que aquí no vamos siquiera a intentar describirlos. Únicamente daremos un ejemplo, en el área de la medicina, para exhibir las ideas básicas que se aplican a la obtención de datos y el análisis estadístico de un conjunto de datos, además del tipo de conceptos matemáticos que para ello se emplean. Cabe mencionar que el desarrollo de los diferentes métodos de inferencia estadística ha dado lugar a profundas investigaciones matemáticas y existe, incluso, una rama de investigación —llamada estadística matemática— que se ocupa de este tipo de cuestiones.

Un planteamiento que se hacen frecuentemente los médicos es cómo saber si un nuevo medicamento es mejor que otro —utilizado por algún tiempo y con resultados conocidos—. Para ello, hay que hacer pruebas, de hecho, hay que "diseñar un experimento". El diseño de experimentos es una parte esencial, pues la forma en que tomamos los datos determina el o los modelos estadísticos que se pueden utilizar para su posterior análisis. Al diseñar un experimento es necesario cumplir algunas reglas y tomar varias decisiones importantes que determinarán, a fin de cuentas, la validez y confiabilidad de las conclusiones

a las que se arribe. Cuando se diseñan experimentos con seres humanos —y más cuando se trata de su salud— es imprescindible que los individuos involucrados acepten participar después de entender perfectamente en qué consiste el experimento y los riesgos que, como sujetos de estudio, corren.

Después, es necesario definir el alcance del experimento, lo que se llama la **muestra**. ¿Utilizaremos sólo sujetos enfermos de aquello que el medicamento en cuestión pretende curar? Supongamos que sí —primera decisión—. ¿Cómo se realizará el experimento? Si damos el medicamento a todos los sujetos de la muestra, ¿qué vamos a medir para saber si el nuevo medicamento funciona mejor que el anterior? Digamos que mediremos la temperatura en cada individuo, antes de suministrar el medicamento y un tiempo después de haberlo suministrado, cualquiera que sea el medicamento suministrado. Con esta información podemos calcular el cambio de temperatura en cada individuo.

Aquí surge otro problema: el lograr que la muestra no sea sesgada al obtener, por ejemplo, una muestra de pacientes particularmente fuertes y que responden mejor a cualquier medicina que la media de los pacientes, o bien, una muestra de pacientes enfermos de varios días y que tienden a curarse más rápidamente que los que acaban de contraer la enfermedad, o al revés. Hay métodos recomendados para resolver este problema, los cuales, normalmente, requieren que la elección de los sujetos se haga aleatoriamente —no aceptando, por ejemplo, sólo a los que se autopropongan para el experimento que, quizá, son gente con más ánimo y más fuertes que la media o personas que ya llevan varios días enfermas y están dispuestas a probar cualquier cosa.

Además, ¿cuántas personas involucramos en el experimento? Cuanto más grande sea la muestra, más confiables serán los resultados, pero también será más caro el experimento y se tendrá que someter a un mayor número de personas a un tratamiento del cual aún se ignoran las posibles consecuencias negativas. Por lo anterior, el tamaño de la muestra es de los primeros problemas a resolver. Existen métodos estadísticos específicos para poder estimar el tamaño de muestra mínimo que puede darnos resultados suficientemente confiables de acuerdo con criterios más o menos universalmente aceptados. En particular, éste es uno de los aportes importantes de la estadística a la investigación, pues nos permite obtener inferencias válidas y con cierto nivel de confianza y precisión al disminuir el costo de la investigación.

En un experimento de este tipo se consideran muchos otros detalles, pero vayamos a otra cosa. Supongamos que ya tenemos una muestra no sesgada de N enfermos que han aceptado responsablemente participar en el experimento y que lo llevaremos a cabo al administrar el medicamento a los enfermos, en periodos de tiempo preestablecidos, y al registrar su temperatura un número preestablecido de horas después.

Se pueden identificar tres aspectos fundamentales del diseño experimental: repetición, aleatorización y control. La teoría y los métodos estadísticos nos dicen cómo resolver estos principios en el momento de diseñar un experimento.

El número de repeticiones está determinado por la confiabilidad y precisión que deseamos del experimento, la forma de aleatorizar está definida por las características de nuestros individuos —para obtener una comparación justa entre ambos medicamentos— y, finalmente, los conocimientos biológicos junto con los del diseño experimental nos dicen cuáles factores tenemos que controlar y cómo hacerlo.

Ahora pasemos al análisis de los resultados y a realizar la inferencia deseada: ¿es mejor el nuevo medicamento? Recordemos que debemos llegar a una conclusión a partir de los datos del experimento. Supongamos que, con los resultados experimentales obtenidos, se genera la siguiente gráfica mostrada en la figura 2.87.

¿Podemos saber cuál de los dos medicamentos es mejor a partir de la gráfica? A simple vista, parecería que el nuevo no logró disminuir tanto la temperatura de los enfermos como

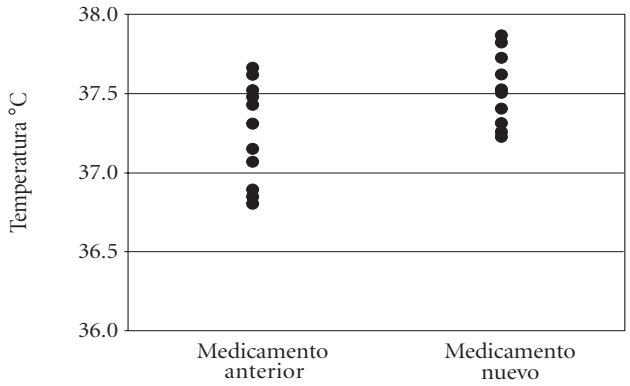


Figura 2.87 Gráfica de temperaturas con el medicamento anterior y el nuevo.

el anterior; pero, si miramos atentamente, veremos que hay una acumulación de varios resultados del nuevo medicamento —alrededor de los 37.25 grados— que parece estar cerca de la media de temperatura para el primer medicamento. Por consiguiente, no resulta obvio cuál de los dos medicamentos es mejor. Para obtener una respuesta clara debemos recurrir a una metodología específica que se denomina **prueba de hipótesis**.

Para hacer dicha prueba se plantea una hipótesis, que llamaremos **hipótesis nula** y la denotaremos como H_0; la prueba consiste en determinar una regla de decisión que nos llevará a "rechazar" o a "no rechazar" dicha hipótesis. Obsérvese que una hipótesis no se acepta, sólo se puede "rechazar" o "no rechazar". La situación que describe esta regla de decisión se muestra en el siguiente cuadro:

INFERENCIA	H_0 EN REALIDAD ES CIERTA	H_0 EN REALIDAD ES FALSA
Rechazamos la hipótesis H_0	Error tipo I con probabilidad α	Inferencia correcta
No rechazamos la hipótesis H_0	Inferencia correcta	Error tipo I con probabilidad β

Las probabilidades α y β indicadas en la tabla tienen significados precisos e importantes. $1-\alpha$ es el **grado de confianza** que nos da la prueba, mientras que $1-\beta$ se llama la **potencia** de la prueba. Para poder explicar esos conceptos en el caso del experimento de los medicamentos y comprender cómo α y β están relacionados con los datos, vamos a introducir la llamada distribución t de Student.

William S. Gosset fue un estadístico que publicó sus trabajos bajo el seudónimo de Student, debido a que la destilería Guinness —en donde trabajaba— había prohibido a sus empleados publicar artículos con cualquier tipo de contenido, debido al secreto industrial. Esta anécdota explica por qué una de las herramientas más utilizadas en la estadística se llama la distribución t de Student y no de Gosset —como sería de esperar, ya que Gosset fue su creador. La distribución de Student aparece en esta situación: supongamos que tenemos una población que tiene una distribución normal con media μ y varianza σ^2; ahora consideremos una muestra de dicha población, de tamaño n, con media m y varianza s. Entonces la distribución de la variable aleatoria que se muestra a continuación:

$$t = \frac{m-\mu}{\frac{s}{\sqrt{n}}}$$

es la llamada distribución t de Student con $n-1$ grados de libertad. Para valores grandes de n, la t de Student se parece mucho a la distribución normal, pero es precisamente en situaciones donde la muestra no es muy grande cuando esta distribución resulta muy útil, como en el ejemplo de los medicamentos.

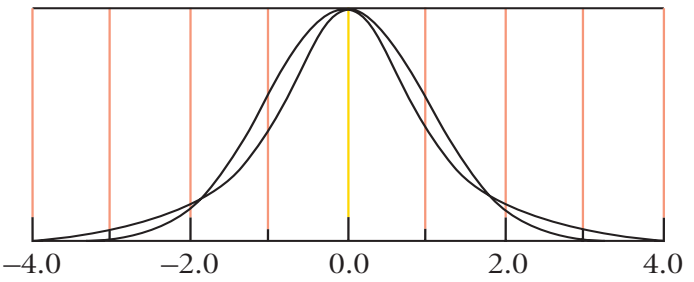


Figura 2.88 La t de Student tiene las colas más pesadas que la normal y mientras más grados de libertad tiene, más se aproxima a la normal.

En este caso, es necesario usar la distribución t con $n_N + n_V - 2$ grados de libertad, donde n_N y n_V son los tamaños de las muestras para el medicamento nuevo y para el anterior, respectivamente, pues no se conocen las varianzas poblacionales y hay necesidad de estimarlas —no olvidemos que, si éstas fueran conocidas, se usaría la distribución normal—. Los grados de libertad de la t corresponden al número de observaciones independientes usadas para estimar la varianza poblacional. El procedimiento a seguir para completar la prueba de hipótesis, consiste en calcular la estadística de prueba:

$$T^* = \frac{(\bar{X}_N - \bar{X}_V) - (\mu_N - \mu_V)}{\sqrt{\left(\frac{1}{n_N} + \frac{1}{n_V}\right) \frac{(n_N - 1)\hat{\sigma}_N^2 + (n_V - 1)\hat{\sigma}_V^2}{n_N + n_V - 2}}}$$

al utilizar las dos muestras de individuos —los que tomaron el medicamento anterior y los que tomaron el nuevo. Si la hipótesis nula H_0 es cierta, esta estadística de prueba tendrá una distribución $t_{n_N + n_V - 2}$, donde intuitivamente esperaríamos que el valor de t^* fuera cercano a cero —ya que esta distribución estaría centrada en el cero y los valores con mayor densidad de probabilidad son cercanos a él—. Si la hipótesis H_0 fuera falsa, la estadística t^* tendría una distribución que estaría a la derecha de la que se describió anteriormente y sería simétrica alrededor de un número mayor que cero. Mientras mayor sea la diferencia entre la efectividad de los medicamentos, esta gráfica estará más alejada del cero, α se iría haciendo más pequeña y $1 - \alpha$ más grande.

En la figura 2.89 se muestran las gráficas de dichas distribuciones t. En el eje x se mide la diferencia de temperaturas entre ambos medicamentos, mientras la línea vertical representa el valor obtenido para la estadística de prueba t^*.

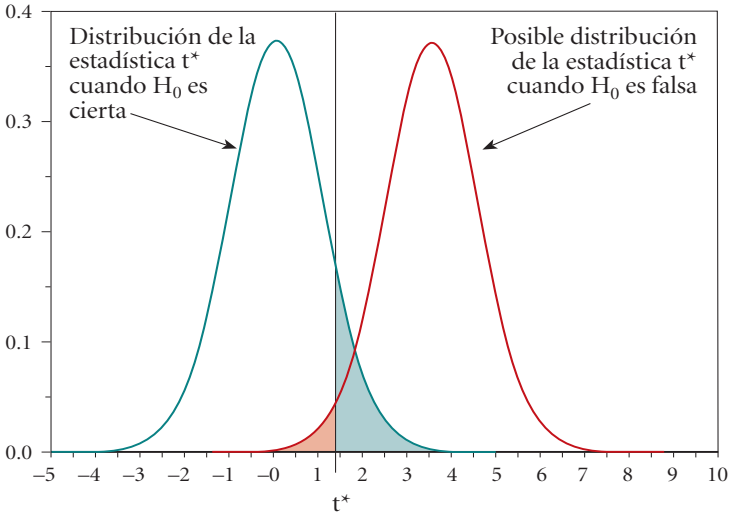


Figura 2.89 A la izquierda la distribución t cuando H_0 es cierta y a la derecha la correspondiente a una posible distribución t cuando H_0 es falsa.

En otros experimentos puede convenir usar otras distribuciones —la normal, la χ^2, etc.— para las pruebas de hipótesis, o bien, para otros procedimientos estadísticos. Lo que es invariable es el hecho de que los resultados estadísticos siempre llevan alguna probabilidad de error. La estadística puede hacer recomendaciones pero, siempre, con cierto grado de confianza, nunca con certeza absoluta, aunque es una poderosa herramienta en la toma de decisiones en muchos campos de la actividad humana, como la industria, el comercio, la política y la ciencia. Por ello, su aplicación no debe tomarse a la ligera y conviene recurrir a los especialistas cuando las decisiones que de ella dependen son muy importantes.

LAS MATEMÁTICAS EN LA NATURALEZA

TEMA 3

Figura 3.1 Ondas en el agua | © Latin Stock México.

3.1 INTRODUCCIÓN

Desde los inicios de la civilización, la comprensión de la naturaleza ha sido una de las preocupaciones fundamentales de los seres humanos. Es innegable que, a medida que el hombre ha ido entendiendo cómo funciona la naturaleza, ha logrado también mejorar su calidad de vida. Toda ciencia —en el fondo— está basada en principios físicos que subyacen en fenómenos tan diversos como la circulación sanguínea en un animal o las reacciones entre las sustancias químicas.

Durante el desarrollo de la ciencia, las matemáticas fueron vinculándose primero en forma tímida y después de una manera cada vez más clara e íntima. Por ejemplo, en la mecánica que es el estudio del movimiento de los cuerpos en función de las fuerzas que actúan sobre ellos. La mecánica es una teoría científica que no puede concebirse, ni siquiera expresarse con un mínimo de claridad, sin las matemáticas. Más adelante, otros campos de la física se desarrollaron siguiendo el mismo patrón de relación íntima con las matemáticas. De hecho, muchas de las ramas de las matemáticas se desarrollaron para poder dar cuenta de algunos fenómenos físicos, por ejemplo, el cálculo vectorial del que se hablará más adelante. El que la física se haya desarrollado en íntima relación con las matemáticas no parece ser un hecho casual o fortuito, más bien, las leyes de la física aparentan que son, realmente, de carácter matemático. Eugene Wigner, Premio Nobel de Física, dice que la ciencia no ha podido explicar por qué el Universo tiene naturaleza matemática y eso es una grave laguna en el conocimiento humano. Pero no hay duda de que para comprender el Universo se necesitan las matemáticas. De hecho, así como la creatividad de un pintor puede quedar limitada por una falta de habilidad en el manejo del pincel, así la del científico puede estar limitada si carece de una sólida formación matemática.

Figura 3.2 Eugene Paul Wigner (1902-1995) fue un físico y matemático húngaro que recibió el Premio Nobel de Física, en 1963, debido a sus contribuciones a la teoría del núcleo atómico y de las partículas elementales, en especial, por el descubrimiento y aplicación de los importantes principios de simetría | © Latin Stock México.

Los aspectos más interesantes de la relación entre las matemáticas y la física requieren de conceptos matemáticos avanzados, la mayoría de los cuales, no están al alcance del lector al cual nos dirigimos. No hemos querido ocultar esta complejidad, sino que hemos tratado de explicar en palabras simples los conceptos más importantes, mostrando al mismo tiempo la simbología matemática que se necesita para expresarlos correctamente (y que a más de un lector podrían asustar). Recomendamos al lector enfrentar este tema con tranquilidad y sentido del humor. Algunas fórmulas le parecerán incomprensibles, pero la intención de los autores no es espantarlo sino mostrarle el tipo de simbología que se usa en las matemáticas avanzadas y ofrecerle una explicación intuitiva de su significado. Para los físicos y matemáticos estas fórmulas con símbolos extraños manifiestan una belleza inaudita por su capacidad de expresar leyes de la naturaleza de una manera compacta y precisa, pero incluso su aspecto visual les resulta estéticamente atractivo. Éste es el caso en particular con las ecuaciones de Maxwell que aparecen como adornos en las playeras y tazas de café de ambientes académicos.

Este tema inicia con la simetría, uno de los aspectos matemáticos más importantes y, a la vez, más fáciles de comprender de la naturaleza por su atractivo visual. Continúa con una descripción de las matemáticas relacionadas con el estudio del espacio y el movimiento, así como los grandes logros de la mecánica que permitieron explicar matemáticamente el movimiento planetario. Después, se muestra la versatilidad de la aplicación de las matemáticas en diversos temas de ciencias, distintos a la física. Finalmente, se plantea cómo la física y las matemáticas continuaron desarrollándose de la mano en temas de teoría electromagnética y física moderna y contemporánea.

3.2 LA SIMETRÍA EN LA NATURALEZA

Figura 3.3 El Nautilus *captura la imaginación. Da la idea de movimiento, de crecimiento, de armonía, de perfección en el diseño, de simetría. Con unos cuantos trazos, nos lleva a pensar en lo infinitamente pequeño y lo infinitamente grande | © Latin Stock México.*

3.2.1 Algunos objetos simétricos

La simetría es una noción fundamental en el arte. Interviene en nuestra idea subjetiva de algo bello, pues surge de la relación que guardan las partes entre ellas y con el todo. Además,

en la naturaleza hay una multitud de ejemplos donde aparece la simetría, no por consideraciones estéticas, sino por eficiencia o economía en el diseño. Más aún, ha resultado ser un elemento indispensable al considerar los modelos más avanzados de la realidad física; es una especie de principio básico en la naturaleza cuya expresión formal es, ineludiblemente, matemática.

La inmensa mayoría de los animales vertebrados tiene simetría bilateral o de espejo. Esto quiere decir que, al partirlos por un plano imaginario, las dos mitades —izquierda y derecha— se corresponden como si estuvieran reflejadas en un espejo, no de manera estricta, sino en el diseño básico o ideal de la estructura ósea y muscular. Y esto es así pues presenta enormes ventajas evolutivas en el desplazamiento, el equilibrio, el control y el desarrollo —con la mitad de los datos y una instrucción "repítase en espejo", se obtiene el todo.

También abundantes en los seres vivos, aunque más escasas, existen simetrías con mayor complejidad. Por ejemplo, en muchas flores no sólo hay un espejo de simetría, sino varios.

Figura 3.4 De izquierda a derecha: flores con simetría de tres, cuatro, cinco y seis rotaciones y reflexiones, respectivamente.

En el esquema simple de una flor de la figura 3.5, aparecen cinco posibles espejos. La simetría de su diseño es la de un pentágono regular: al reflejarlo en cualquiera de esas cinco líneas, regresa a su lugar.

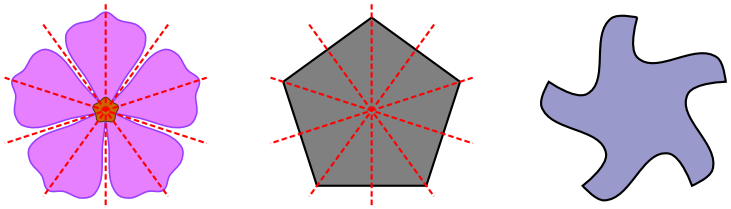


Figura 3.5 Una flor con cinco pétalos tiene la misma simetría que un pentágono regular. Pero una "estrella ninja" de cinco picos sólo comparte con ellas las rotaciones y tiene la simetría de un pentágono con lados orientados cíclicamente.

Cabe aclarar que *reflejar*, en el sentido matemático, es la transformación del plano o del espacio en sí mismo que intercambia los dos lados de una recta o un plano, llamado *espejo*, y mantiene las distancias con respecto a él. En el mundo real, un espejo actúa sobre una mitad del espacio y lo hace aparecer —como visto a través de una ventana— reflejado. Pero una reflexión matemática actúa sobre ambos lados, intercambiándolos.

3.2.2 Composición e inversos de simetrías

Además de las cinco reflexiones, en la simetría pentagonal aparecen cinco rotaciones por ángulos múltiplos de $\frac{2\pi}{5}$ o 72°. Hay figuras planas que sólo tienen estas rotaciones como simetrías, como se muestra a la derecha de la figura 3.4. En la figura 3.6 se observa cómo podemos distinguir estas rotaciones; para ello, se rotularon las esquinas con números, por lo que ahora es posible ver cómo se mueve la esquina número 1 bajo estas rotaciones.

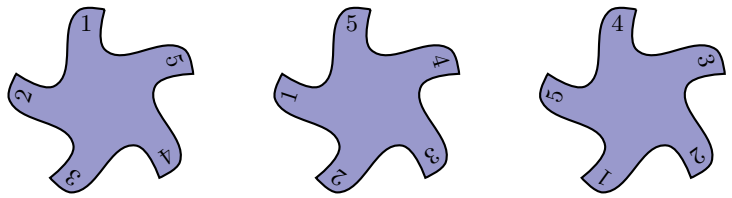


Figura 3.6 Rotaciones de una estrella ninja.

A la izquierda se muestra la posición inicial, en el centro está la posición después de rotar 72° y, a la derecha, después de moverse dos veces según dicho ángulo. Hay algo importante que debe observarse aquí —aunque parezca muy simple— y que es absolutamente crucial: la rotación de la estrella *ninja* a la derecha en la figura 3.6 puede obtenerse como una *rotación* de 144° o *por dos rotaciones consecutivas* de 72°. Esto se puede reformular de una manera más general: dos simetrías, es decir, dos movimientos que dejan invariante a una figura pueden efectuarse uno tras otro para obtener, de nuevo, una simetría.

La figura 3.7 muestra la importancia del orden en que se efectúan dos simetrías. A la izquierda se numeran los pétalos y se muestra una rotación ρ y una reflexión σ —estas letras se leen "rho" y "sigma", respectivamente—. La flor del centro muestra el efecto de primero reflejar (σ) y después rotar (ρ) —esto se expresa en notación de funciones como $\rho \circ \sigma$; y se lee "rho después de sigma"— y en la tercera imagen se muestra el efecto de primero rotar y después reflejar $(\sigma \circ \rho)$. Por ejemplo, si rotamos con ρ, el pétalo 1 queda en el lugar del 2 —podríamos escribir $\rho(1) = 2$, que se lee "rho de 1, o aplicado a 1, es igual a 2"—, y si después reflejamos con σ el pétalo 1 queda finalmente en el lugar donde antes estaba el pétalo 5 —pues $\sigma(2) = 5$ y le da sentido a la notación $(\sigma \circ \rho)(1) = \sigma(\rho(1)) = \sigma(2) = 5$.

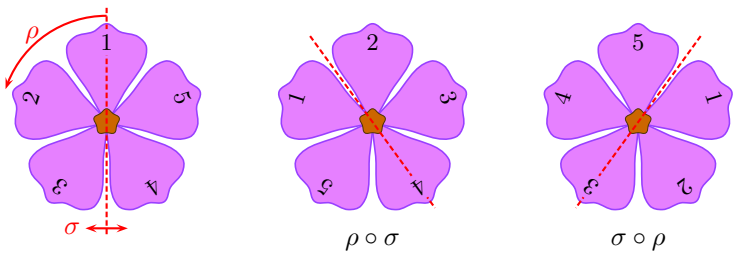


Figura 3.7 Composición de una reflexión (σ) y una rotación (ρ) en sus diferentes órdenes. En el centro, primero σ y luego ρ; en el extremo derecho, primero ρ y después σ.

Se observa en la figura 3.7 —viendo el efecto en los pétalos numerados— que $\rho \circ \sigma$ es lo mismo que reflejar en el eje que pasa por el pétalo marcado con el número 4, mientras $\sigma \circ \rho$ equivale a reflejar en el eje que pasa por el pétalo 3. El proceso de efectuar primero una simetría y luego otra se llama *composición*. Y ahora sucede algo raro, si componemos la reflexión σ —de la figura 3.7— consigo misma, obtenemos un movimiento que no hace nada: cada pétalo regresa a su lugar. Al principio cuesta trabajo considerar al "no hacer nada" como un movimiento, pero es claramente un movimiento del plano que no altera las distancias. A este movimiento se le llama *identidad*, dado que funciona como el número 1 al multiplicar: deja todo como estaba.

También, cada simetría tiene su *inversa*: un movimiento como una reflexión o una rotación siempre tiene una reflexión —o rotación— correspondiente, que deshace el efecto

del movimiento original; para las reflexiones son ellas mismas. En general, dada una simetría, la que deshace su efecto se conoce como su inversa, muy similar al hecho de que $\frac{1}{2}$ es el inverso de 2, pues al multiplicar por $\frac{1}{2}$ se deshace el efecto de multiplicar por 2.

Por estas razones vemos que la identidad es una simetría importante y, sin ella, no podríamos componer dos simetrías cualesquiera ni podríamos considerar la formación de inversos. Las simetrías de una figura forman una estructura que se llama *grupo*. Los grupos tienen un elemento distinguido que es la identidad; además, sus elementos se pueden componer y se permite formar inversos.

3.2.3 El concepto detrás de la simetría

La simetría de muchas galaxias es sencilla: además de la identidad solamente hay una rotación de 180°; muchas orquídeas también sólo tienen una simetría más, que es la reflexión en un plano vertical que atraviesa la planta a la mitad, como se observa en la figura 3.8.

Figura 3.7 Del lado izquierdo, la galaxia NGC 1365 con dos brazos principales; del lado derecho, una orquídea.

En ambos casos, hay dos simetrías: la identidad —que se abrevia con id— y otra que llamamos ρ y que tiene la propiedad de que:

$$\rho \circ \rho = \mathrm{id}\,.$$

Las otras posibles composiciones son:

$$\rho \circ \mathrm{id} = \rho, \quad \mathrm{id} \circ \rho = \rho \quad \text{y} \quad \mathrm{id} \circ \mathrm{id} = \mathrm{id}\,.$$

Estas últimas igualdades siempre se tienen pues id es como el 1 de la multiplicación: deja todo como estaba.

En conclusión, el grupo de simetría de la galaxia con dos brazos es "muy similar" al grupo de simetrías de la orquídea: tienen el mismo número de elementos y, además, si éstos se denotan con id y ρ, entonces se componen de la misma manera. En matemáticas, se dice que dos grupos son *isomorfos* cuando, literalmente, tienen la misma estructura. A menudo, se identifican grupos isomorfos como si fueran iguales, en forma muy similar a como se puede identificar un número de frijoles con un número de puntos de fichas de dominó o con la abstracción: el número. El grupo en abstracto es el que está detrás de la *representación* o *realización* concreta.

3.2.4 Las simetrías del *Nautilus*

El *Nautilus*, cuya imagen inició esta sección, no tiene simetrías en el sentido de que no hay movimientos del plano como reflexiones o rotaciones que lo dejen invariante, excepto —claro está— la identidad. Sin embargo, el *Nautilus* tiene una forma bella y, usualmente, nuestros ojos no mienten: donde hay belleza, hay cierta simetría. En efecto, así es: el *Nautilus* tiene muchas simetrías, pero para verlas hay que relajar la noción de "transformación que deja invariante a la figura", a que signifique no sólo *movimientos rígidos*, es decir, transformaciones del plano que no alteran las distancias.

En el *Nautilus* necesitamos también dilataciones, es decir, transformaciones que agrandan o achican. Para tener una transformación bien definida se requiere establecer el centro de la dilatación y el factor. En la figura 3.9 vemos el ejemplo de una dilatación o, como también se les dice, *homotecia*.

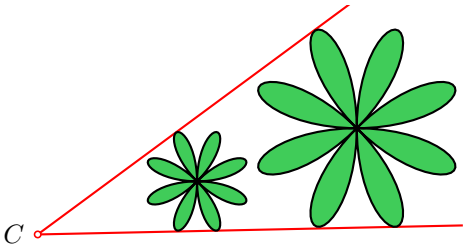


Figura 3.9 El efecto de una dilatación con factor 2.

La espiral del *Nautilus* tiene la propiedad de que es invariante bajo ciertas dilataciones. En la figura 3.10 se muestra una espiral particular: sobre un rayo que sale del centro —marcado con rojo— se numeraron los puntos de la espiral con $\ldots, C_{-1}, C_0, C_1, \ldots$, por lo tanto, es una sucesión cuyo índice son los enteros. En esta espiral cada punto C_n está al doble de distancia al centro que el punto anterior C_{n-1}. Esto no depende de la dirección del rayo rojo y es igualmente cierto para otro rayo como, por ejemplo, el azul marcado en la misma figura. Por ello, una dilatación con factor 2 deja a la espiral invariante. El grupo que se forma con estas dilataciones es isomorfo al grupo de los enteros bajo la adición: al número entero n le hacemos corresponder aquella dilatación que envía el punto C_0 al punto C_n.

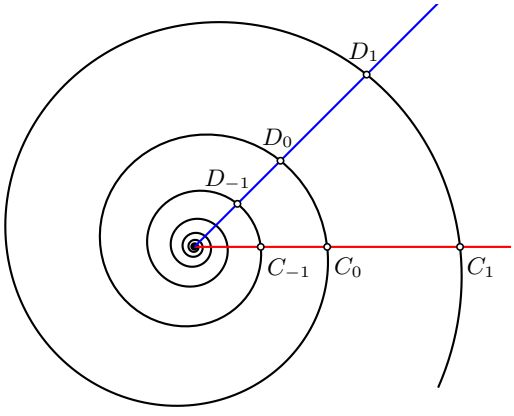


Figura 3.10 Una espiral logarítmica invariante bajo dilataciones con factor de potencias de 2.

Pero éstas no son todas las simetrías de la espiral porque podemos combinar dilataciones con rotaciones. Si primero giramos de manera tal que el rayo rojo coincida con el azul y luego dilatamos por un factor apropiado, obtenemos otra simetría que manda C_0 a D_0. El grupo de simetría de la espiral es isomorfo al grupo de los números reales positivos bajo la multiplicación: a cada número real positivo corresponde una simetría que está compuesta por una dilatación con este factor y una rotación apropiada. Para referencia posterior, usaremos $\mathbb{R}_{>0}$ para denotar a este grupo.

Si observamos de nuevo la imagen del *Nautilus*, destacan las paredes divisorias entre las cámaras internas. Si buscamos simetrías de la espiral que "respeten" también las paredes, entonces solamente ciertos factores de dilatación serían permitidos. Estos factores son de la forma ρ^i, es decir, son múltiplos enteros —con exponentes positivos o negativos— de un solo factor ρ. Este grupo de simetría es como el grupo de los enteros bajo la suma.

3.2.5 Simetría de fórmulas

A primera vista, la simetría es un concepto que aparece meramente de la geometría. Veremos ahora que lo anterior no es sino un punto de vista demasiado limitado. Para ello, consideramos la ecuación:

$$xy - 1 = 0. \tag{1}$$

Esta ecuación es invariante bajo el intercambio de las dos variables x, y, dado que $yx = xy$. En consecuencia, si el par de números (x, y) satisface la ecuación (1), entonces también lo hace el par (y, x). Esta transformación corresponde geométricamente a una reflexión del plano en la diagonal $x = y$, así que esperamos que el conjunto de soluciones sea simétrico respecto a dicha reflexión. Denotemos a este intercambio como σ:

$$\sigma\colon \begin{cases} x \mapsto y, \\ y \mapsto x. \end{cases}$$

El conjunto de soluciones (x, y) describe una hipérbola con los ejes de coordenadas como asíntotas, según se muestra en la figura 3.11. En la misma figura se indica la recta $x = y$ con color rojo. Es —como esperábamos— un eje de simetría de la hipérbola azul. También se dibujó el otro eje de simetría $y = -x$.

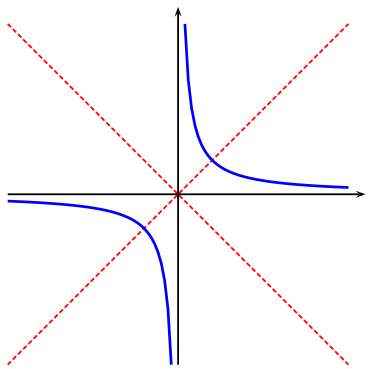


Figura 3.11 La hipérbola descrita por la ecuación $xy - 1 = 0$.

Reflejar en la recta $y = -x$ tiene el efecto de hacer que un punto con coordenadas (x, y) sea enviado al punto $(-y, -x)$, pues intercambia los ejes y les cambia orientación.

En otras palabras, la reflexión geométrica corresponde al intercambio de variables con cambio de signo:

$$\rho\colon \begin{cases} x \mapsto -y, \\ y \mapsto -x. \end{cases}$$

Dado que las simetrías forman un grupo, podemos componer las dos reflexiones y obtenemos una rotación de 180° con centro en el origen del sistema de coordenadas. Esta operación corresponde al cambio de signo τ:

$$\tau = \rho \circ \sigma \colon \begin{cases} x \mapsto -x, \\ y \mapsto -y. \end{cases}$$

Junto con la identidad, hemos encontrado cuatro simetrías: $\mathrm{id}, \sigma, \rho, \tau$; la siguiente tabla muestra cómo se componen estos elementos. Por ejemplo, en el segundo renglón y tercera columna —contando en lo blanco— está τ, que es el resultado de componer ρ con σ —justo los encabezados amarillos correspondientes—. A esta tabla se le llama *tabla de multiplicación* y un grupo queda definido por ella.

$\circ$	id	ρ	σ	τ
id	id	ρ	σ	τ
ρ	ρ	id	τ	σ
σ	σ	τ	id	ρ
τ	τ	σ	ρ	id

Figura 3.12 Tabla de multiplicación que define un grupo.

Al grupo con esta tabla de multiplicación se le conoce como el *grupo de Klein*. Otra realización de este grupo se obtiene al considerar entre las permutaciones de cuatro elementos $1, 2, 3, 4$, sólo aquellas que permuten dos pares: por ejemplo, intercambiar 1 con 2 y al mismo tiempo 3 con 4; esta permutación se denota por $(1\ 2)(3\ 4)$. Además de la identidad sólo hay 3 permutaciones así y éstas están determinadas por cómo se permute el número 1:

$$\rho = (1\ 2)(3\ 4), \quad \sigma = (1\ 3)(2\ 4), \quad \tau = (1\ 4)(2\ 3).$$

Para calcular una composición, como $\rho \circ \sigma$, es útil dibujar un diagrama como el que se muestra en la figura 3.12. Debe recordarse que $\rho \circ \sigma$ denota la permutación que se obtiene al aplicar, primero σ y luego ρ, contrario a la lectura de la expresión de izquierda a derecha. Con este diagrama se verifica sin problema que $\rho \circ \sigma = \tau$ y, de manera similar, se puede comprobar que las cuatro permutaciones id, ρ, σ y τ satisfacen la misma tabla de multiplicación que las simetrías de $xy = 1$.

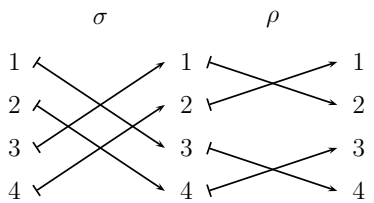


Figura 3.13 Diagrama que muestra cómo se permutan los elementos bajo la composición $\rho \circ \sigma$.

Antes de regresar a la hipérbola, vale la pena observar que el grupo de Klein también se realiza en el espacio como las simetrías rotacionales de la pelota de tenis.

Figura 3.14 Las rotaciones que dejan al diseño de una pelota de tenis en su lugar forman un grupo isomorfo al de Klein.

En la hipérbola, y como vimos con el *Nautilus*, no tenemos que restringirnos a las simetrías formadas por movimientos del plano que conservan la distancia. Para cada número real positivo, la transformación:

$$M_a\colon \begin{cases} x \mapsto ax, \\ y \mapsto \frac{1}{a}y \end{cases}$$

también deja invariante la ecuación de la hipérbola (1). El efecto geométrico de esta transformación es que envía un punto (x, y) al punto $(ax, \frac{1}{a}y)$. En particular, la recta $y = x$ se envía a la recta $y = \frac{1}{a^2}x$. Por ejemplo, para $a = 2$, la transformación M_2 envía la recta $y = mx$ a la recta $y = \frac{m}{4}x$, mientras que todas las hipérbolas de la forma $xy = c$ se mantienen fijas.

El grupo formado por las transformaciones M_a es isomorfo al grupo $\mathbb{R}_{>0}$ ya que las transformaciones se componen según la siguiente ley:

$$M_a \circ M_b = M_{ab}.$$

Éste es el mismo grupo que el de las dilataciones del *Nautilus*. Si permitimos que a también puede ser negativo, es decir, cualquier número real excepto el cero, entonces obtenemos un nuevo grupo y, si consideramos también el intercambio de x y y, obtenemos uno todavía más grande.

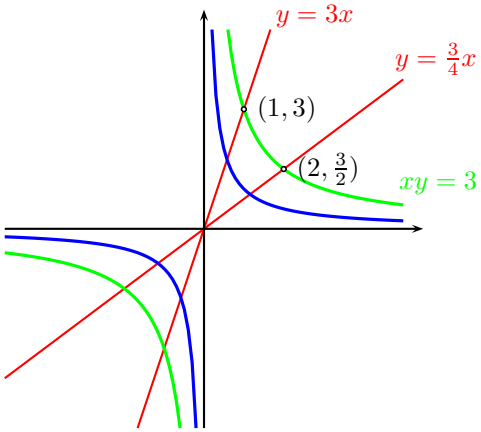


Figura 3.15 Ejemplo de cómo actúa la transformación M_2 en el plano: el punto $(1, 3)$ es enviado al punto $(2, \frac{3}{2})$. Toda la recta $y = 3x$ es enviada a la recta $y = \frac{3}{4}x$, mientras la hipérbola $xy = 3$ permanece —como conjunto— fija.

De esta manera hemos visto que un mismo objeto puede exhibir diferentes grupos de simetría y que el mismo grupo puede aparecer en diferentes objetos.

3.2.6 Simetría conceptual

Existen palabras como *oso*, *rayar* o *somos* que pueden leerse de adelante hacia atrás o al revés y significan lo mismo. Estas palabras se llaman *palíndromos*. También hay frases que son palíndromos:

Anita lava la tina.

Los palíndromos tienen una simetría especial: son invariantes bajo la acción de voltearse. No se trata de una simetría geométrica, sino de cambiar de lugar letra por letra. Si aplicamos dos veces esta simetría obtenemos la identidad. El grupo es isomorfo al grupo de simetría de la galaxia y de la orquídea.

Es usual que recordemos la fórmula para calcular el área de un triángulo como "base por altura sobre dos". Pero, ¿cuál de los tres lados es la base? ¿El que está más abajo? Estrictamente, un triángulo no tiene una base, sino tres. Cualquiera de los tres lados puede fungir como base y "la altura" será, entonces, la altura correspondiente. Por ello, la fórmula mencionada anteriormente no es simétrica en las tres variables, sino que requiere una elección: hay que elegir uno de los tres lados como base.

Si tenemos un triángulo con lados de 6, 9 y 7 centímetros, no podríamos calcular con esta fórmula el área de manera directa. Por otro lado, es claro que el triángulo queda completamente definido al especificar sus tres lados y que debería ser posible calcular su área. Lo anterior se puede resolver con trigonometría, según se hizo en la sección 2.7 o con la *fórmula de Herón*, que dice que:

$$A = \sqrt{s \cdot (s-a) \cdot (s-b) \cdot (s-c)} \qquad (2)$$

es el área del triángulo con lados a, b y c si:

$$s = \frac{a+b+c}{2}$$

es el *semiperímetro*, es decir, la mitad del perímetro del triángulo. La fórmula de Herón es simétrica en los tres lados. Si permutamos los lados obtenemos una expresión que se reduce, al usar la conmutatividad de la multiplicación, a (2). Por otro lado, esta fórmula se puede evaluar simplemente: en nuestro ejemplo el semiperímetro es $s = 11\,\text{cm}$ y por lo tanto:

$$\begin{aligned} A &= \sqrt{11\,\text{cm} \cdot (11\,\text{cm} - 6\,\text{cm}) \cdot (11\,\text{cm} - 9\,\text{cm}) \cdot (11\,\text{cm} - 7\,\text{cm})} \\ &= \sqrt{440}\,\text{cm}^2 \end{aligned}$$

Vale la pena reflexionar sobre cuáles de las fórmulas que usualmente se aprenden en bachillerato son simétricas. Por ejemplo, el teorema de Pitágoras trata de triángulos rectángulos, cuyos lados a, b y c no son simétricos. Uno de ellos es la hipotenusa, los otros dos son los catetos. Si c es la hipotenusa, entonces el teorema de Pitágoras afirma que:

$$c^2 = a^2 + b^2,$$

una fórmula que "es simétrica en a y b", es decir, podemos intercambiar a y b sin alterar la fórmula. Si el triángulo no es rectángulo hay una fórmula similar que toma en cuenta el valor del ángulo opuesto al lado c, que según la convención de Euler se denota por γ. Esta fórmula se llama *ley del coseno* y dice que:

$$c^2 = a^2 + b^2 - 2ab\cos(\gamma).$$

El coseno de 90° es cero, por lo que vemos que la ley de coseno se especializa en el teorema de Pitágoras, o dicho de otra manera, la ley del coseno contiene como un caso particular al teorema de Pitágoras. Nuevamente, la ley del coseno es simétrica en a y b.

Si f y g son dos funciones, entonces la derivada de la suma o del producto es simétrica en f y g:

$$\begin{aligned} (f+g)' &= f' + g' \\ (f \cdot g)' &= f \cdot g' + f' \cdot g \end{aligned}$$

donde denotamos la derivada de una función con una prima, es decir, f' es la derivada de f —aquí no importa la definición precisa de derivada sino las propiedades que debe heredar de las que tienen las funciones. Esta simetría corresponde a que $f + g = g + f$ y $f \cdot g = g \cdot f$, es decir, la conmutatividad de la suma y la multiplicación de funciones. Pero la derivada de la composición de funciones no es simétrica en f y g:

$$(g \circ f)' = (g' \circ f) \cdot f'$$

y ello corresponde al hecho de que la composición de funciones no es conmutativa: si $f(x) = x^2$ y $g(x) = 2x$, entonces $(f \circ g)(x) = (2x)^2$ mientras que $(g \circ f)(x) = 2x^2$, que no es lo mismo.

3.2.7 Simetría en la física

Las leyes de la física toman en cuenta conceptos como el espacio, el tiempo, la materia o la electricidad. Para expresar las cantidades se suelen usar coordenadas. Por ejemplo, la segunda ley de Newton dice que la fuerza que actúa sobre un cuerpo es igual al producto de su masa por la aceleración y se suele escribir como $F = ma$, donde F es la fuerza, m la masa y a la aceleración.

Pero en realidad se trata de una ley en el espacio: la fuerza y la aceleración son *vectores*, es decir tienen dirección y longitud. Si esto lo escribimos en nuestro sistema de coordenadas obtenemos tres ecuaciones:

$$\begin{aligned} F_x &= ma_x \\ F_y &= ma_y \\ F_z &= ma_z \end{aligned}$$

donde F_x, F_y, F_z son los componentes del vector F y similarmente a_x, a_y, a_z del vector a.

En principio, no es claro que estas leyes no cambien si tomamos otro sistema de coordenadas; por ejemplo, si lo rotamos por un ángulo α alrededor del eje de coordenadas z, la transformación es:

$$\begin{aligned} x' &= \cos(\alpha)x - \operatorname{sen}(\alpha)y \\ y' &= \operatorname{sen}(\alpha)x + \cos(\alpha)y \\ z' &= z \end{aligned}$$

donde x', y' y z' significan las coordenadas en el sistema girado. De forma similar se calculan los componentes del vector de la fuerza y la aceleración. Por ejemplo:

$$\begin{aligned} F_{x'} &= \cos(\alpha)F_x - \operatorname{sen}(\alpha)F_y \\ &= \cos(\alpha)ma_x - \operatorname{sen}(\alpha)ma_y \\ &= m(\cos(\alpha)a_x - \operatorname{sen}(\alpha)a_y) \\ &= ma_{x'} \end{aligned}$$

De manera semejante, se obtiene $F_{y'} = ma_{y'}$ y $F_{z'} = ma_{z'}$. Con ello se muestra que la segunda ley de Newton es invariante bajo rotaciones alrededor del eje de coordenadas z. Con argumentos similares se puede mostrar también la invariancia bajo rotaciones alrededor de otros ejes y, por lo tanto, de una combinación de ellos. Éste es un ejemplo sencillo de

una de las motivaciones que llevaron a Albert Einstein a proponer la teoría de la relatividad: que las leyes de la física deben ser las mismas en cualquier parte del Universo y en todo momento; que al expresarlas como ecuaciones, éstas deben ser simétricas respecto al movimiento en el espacio y en todo momento.

La simetría juega un papel importante en la física y, sobre todo en el siglo xx, llegó a ser una consideración sustancial para generar expectativas y hasta teorías enteras. Una simetría de cierta ecuación, la de Dirac, condujo a la predicción de la existencia del *positrón*, la antipartícula del electrón. El concepto de grupo, en particular su abstracción, no sólo abrió camino a las matemáticas en el siglo xix sino también a la física durante el siglo xx. Tiene un papel central en la frontera de la física, en el mundo cuántico de lo muy pequeño y en el relativista de lo muy grande. Se espera, incluso, que guíe la ruta de la buscada "teoría del todo", una unificación de esas dos ramas claves de la física moderna. Para finalizar, hay que resaltar que el concepto de grupo, ese que subyace al de simetría y a las más modernas teorías físicas, surgió a principios del siglo xix al atacar un problema de las mismas matemáticas: el resolver la ecuación de quinto grado.

3.3 ESPACIO, TIEMPO Y MOVIMIENTO

La relación entre la naturaleza y las matemáticas es tan estrecha que muchas veces resulta difícil separarlas. Por ejemplo, ¿qué tanto la geometría es un estudio del espacio físico en el que nos encontramos? Con sus conceptos y teoremas, la geometría proviene del esfuerzo de muchas generaciones para adquirir un modelo teórico que represente fielmente al espacio en que vivimos y sus propiedades. ¿Esto es física o matemáticas? Más bien, ambas.

La concepción moderna de las matemáticas que dominó el pensamiento durante el siglo xx, está basada en la idea de que éstas no tratan con ninguna realidad sino con objetos abstractos, las relaciones entre ellos —que se postulan como axiomas— y las consecuencias lógicas que pueden deducirse de estos últimos. Si aceptamos esta idea, la geometría, como parte de las matemáticas, no podría tratar sobre el espacio. Pero a lo largo de muchos siglos, el hombre intentó modelar y comprender su entorno apoyándose para ello en las matemáticas y concibiéndolas como el lenguaje adecuado para explicar las propiedades y el comportamiento del mundo que le rodea.

¿Existen los puntos? Según la geometría de Euclides, un punto es aquello que no tiene partes. ¿Alguien ha visto alguna vez un punto; no una pequeña mancha en el papel —que por muy pequeña que sea, consta de una zona del papel manchada por la tinta—, sino un punto de verdad que no conste de otras cosas? ¿Existen una recta y un círculo perfectos? Estas preguntas se las hacían los griegos y Platón las respondía diciendo que tales objetos matemáticos "perfectos", como el punto, la recta y el círculo, no son parte del mundo de los sentidos sino que habitan un mundo ideal al que el ser humano sólo tiene acceso mediante el pensamiento. Los objetos reales que señalamos diciendo que son puntos, rectas y circunferencias, no son más que ejemplos materiales imperfectos de aquellos objetos ideales.

El formalismo llegó aún más lejos que la geometría de Euclides en la construcción de una teoría de las matemáticas desligada del mundo material, que se revisa en la sección 4.7, al especificar que las matemáticas tratan con objetos abstractos ajenos al mundo material, es decir, que esos objetos ni siquiera deben intentar representarse por dibujos o algún otro medio, y que las palabras punto, recta y circunferencia bien podrían cambiarse por otras como silla, corcho y sal, sin alterar en absoluto el contenido de la geometría, siempre y cuando los objetos abstractos designados por ellas guarden entre sí las relaciones postuladas, es decir, los axiomas.

Quizá el punto de vista formalista representa el sentir mayoritario de la comunidad académica matemática y, sin embargo, no explica la increíble eficacia de las matemáticas para modelar la realidad, sus componentes y sus leyes, como se verá en la sección 3.5. Tampoco explica la gran utilidad de las matemáticas para representar, plantear y resolver problemas prácticos en el ámbito de la actividad humana en general. El formalismo da cuenta de una parte de la actividad matemática, pero ignora por completo las relaciones de las matemáticas con las ciencias y su utilidad social. En este apartado incursionaremos en el mundo fascinante de las relaciones entre las matemáticas y el mundo físico, entendido como el espacio en el que vivimos, sus propiedades y las de los cuerpos materiales que lo habitan, el movimiento, las leyes que lo rigen, las fuerzas de la naturaleza y los conceptos matemáticos que el hombre ha desarrollado —específicamente— para modelar ese mundo físico.

3.3.1 El espacio

En general, modelamos el mundo en que vivimos como un espacio euclidiano de tres dimensiones. Aunque la teoría de la relatividad apunta en la dirección de que el espacio físico puede no ser euclidiano —podría tener cuatro dimensiones, ser curvo y quizás hasta finito—, nuestro modelo cotidiano del espacio físico es euclidiano, de tres dimensiones e infinito. ¿Qué significa esto? Pues que la posición de cada objeto queda perfectamente especificada mediante tres números x, y, z, a partir de un sistema de coordenadas que podemos fijar arbitrariamente en cualquier sitio, con cualquier orientación y con cualquier unidad de medida.

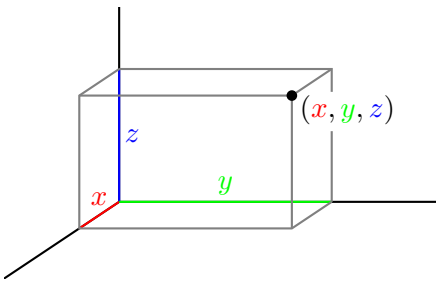


Figura 3.16 Un punto en el espacio se determina por sus coordenadas x, y, z, que son tres números reales.

Por supuesto, los números x, y, z que dan la posición de un punto, dependen de dónde se haya colocado el origen del sistema, cómo se haya orientado y qué unidad se haya elegido. Pero una vez definidas estas variables, la posición de cada punto queda determinada de manera inequívoca mediante una terna ordenada (x, y, z) de números. Que este modelo del espacio sea "correcto" o sólo una buena aproximación a la realidad es una cuestión que corresponde resolver a la física, pero el modelo es claro y se ajusta perfectamente a la realidad cotidiana y a la mayor parte del ámbito científico. Sólo haría falta modificar este modelo utilizando el de la Teoría de la Relatividad, cuando se trata con cuestiones del espacio astronómico donde las distancias, las velocidades y las fuerzas pueden ser mucho mayores que aquellas con las que lidiamos cada día aquí en la Tierra.

La idea de los sistemas de coordenadas proviene de René Descartes, que en su libro *La geometría* propone este modelo para el espacio físico. Descartes no hizo más que plantear el modelo explícitamente pues éste ya existía en la mente humana y se usaba diariamente desde la Antigüedad. Para que el modelo cartesiano quede claro es necesario suponer que contamos con un sistema de numeración que permite asignar números a las coordenadas de cada punto del espacio. En particular, esto requiere que contemos con un modelo de la recta que permita asignar a cada uno de sus puntos un número.

3.3.2 El continuo espacio-tiempo: los números reales

Para modelar el continuo se utiliza el sistema numérico de los **números reales** que, si bien fue utilizado implícitamente por los matemáticos griegos, no se definió con precisión hasta el siglo XIX, muchos años después de que Descartes propusiera su modelo del espacio. Los hombres de ciencia y el público en general ya tenían —desde mucho tiempo atrás— una idea intuitiva del continuo, de que el espacio no puede ser modelado con algo discreto y finito, sino que hace falta considerar que entre dos puntos hay una infinidad de otros puntos y que, por lo tanto, cada punto debe tener un tamaño nulo pues, de otra manera, se podría cubrir un intervalo con un número finito de puntos. Las ideas sobre el continuo provienen de la Antigüedad y siempre causaron controversias intelectuales, como las famosas paradojas de Zenón, según las cuales sería imposible el movimiento. Por ejemplo, una de ellas dice que:

> *Una persona no puede recorrer una cierta distancia, porque primero debe llegar a la mitad de ésta, antes a la mitad de la mitad y, antes aun, debería recorrer la mitad de la mitad de la mitad y así indefinidamente.*

De este modo, en teoría, una persona no puede recorrer una distancia, aunque la experiencia muestra que el movimiento sí es posible.

La paradoja se resuelve eliminando la hipótesis —implícita en el lenguaje— de que para realizar una infinidad de pasos es necesario un tiempo infinito. Todos los pasos pueden llevarse a cabo en un tiempo finito, porque cada uno es menor que el anterior y la suma de todos ellos es una cantidad infinita. Por ejemplo, la suma infinita de un medio de una unidad, más la mitad de un medio de la unidad, más la mitad de la mitad de un medio de la unidad, etc., es igual a una unidad. En símbolos, se representa como:

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots=1$$ y gráficamente como:

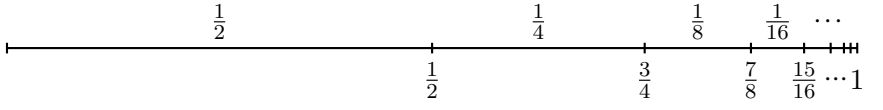


Figura 3.17 Suma de $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots=1$.

Un poco de álgebra nos permite demostrar que:

$$\begin{aligned}\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots+\frac{1}{2^N} &= \frac{1}{2^1}+\frac{1}{2^2}+\ldots+\frac{1}{2^N}\\ &= \frac{1}{2}\frac{1-\frac{1}{2^N}}{1-\frac{1}{2}}\\ &= 1-\frac{1}{2^N}\end{aligned}$$

y, por lo tanto, todas las sumas $\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\ldots+\frac{1}{2^N}$ están acotadas por 1 y tienden a 1 cuando N tiende a infinito.

Como ya se dijo en el tema 2, concebimos a la línea recta como algo continuo, sin huecos, donde en cada intervalo existe una infinidad de puntos y donde, entre cada dos puntos, existen otros. Para representar todos los puntos de una línea recta usamos los números reales. Por lo general, representamos al continuo como una recta horizontal, elegimos un punto en ella que llamamos *origen* e identificamos con el cero y marcamos otro punto a la derecha que identificaremos con el uno. La distancia entre el origen y este segundo punto se toma como la unidad para medir distancias.

Figura 3.18 Unidad para medir distancias.

Dada esta construcción, podemos asignar un número real a cada punto P de la recta. Para ello, contamos el número de unidades a la derecha o a la izquierda que pueden colocarse desde el origen hasta el punto y asignamos este número entero, 2 en el caso ilustrado en la figura, como primera aproximación en la representación que haremos del punto P.

Figura 3.19 Primera aproximación en la representación que haremos del punto P.

A continuación, dividimos el intervalo de longitud uno en el que se encuentra el punto —en este caso, entre 2 y 3— en diez partes iguales y contamos cuántas de esas partes están a la izquierda del punto y, así, definimos el primer decimal. En este caso, obtenemos 2.4 como segunda aproximación a P. Luego dividimos el intervalo de un décimo de unidad en el que se encuentra P en diez partes iguales y contamos cuántas caben antes del punto, con lo cual definimos el segundo decimal. Continuamos este proceso indefinidamente para completar la definición del número real que representa al punto P.

La descripción contenida en el párrafo anterior es la clave del concepto de número real. Dice que el proceso continúa *indefinidamente*, es decir, que nunca termina. De acuerdo con nuestro lenguaje ordinario, esto parecería indicar que nunca terminaremos de definir el número. Sin embargo, con las instrucciones dadas el número ya queda perfectamente definido, pues nos permiten escribir una aproximación decimal del punto tan precisa como deseemos y eso es precisamente lo que entendemos por número real. Preocuparse porque esta definición no ofrece una expresión infinita no nos lleva a ningún lado. En general, un punto de la recta requeriría, para ser representado mediante una expresión decimal, de una cadena infinitamente larga de dígitos. Sin embargo, como por lo general esto es imposible, no tiene caso intentarlo; para considerar que el número está definido, basta describir el proceso mediante el cual se puede encontrar cada uno de sus decimales.

La representación del continuo espacial mediante números nos obliga a la creación del concepto de número real como un proceso infinito de aproximación y no como un resultado que puede exhibirse explícitamente. Se trata de algo incómodo al principio, pero perfectamente correcto desde un punto de vista lógico y, por tanto, algo a lo que la mente humana puede habituarse y manipular sin ambigüedades. La creación de los números reales permite aplicar las matemáticas no sólo a la descripción del espacio físico, sino también a la del tiempo y al estudio del movimiento de los cuerpos en el espacio, lo cual puede considerarse como el origen de la ciencia cuantitativa.

Al igual que el espacio, el tiempo también requiere del concepto del continuo y se beneficia del concepto de número real. En efecto, la observación del movimiento nos indica una continuidad en el flujo del tiempo. Cuando algo se mueve no pasa instantáneamente de una posición a otra, sino que va cambiando continuamente su posición, pasando por todos los puntos intermedios y haciéndolo en instantes intermedios entre el inicio y el final. Es perfectamente concebible un mundo en que el tiempo fuera discreto, donde todo ocurriera como en los fotogramas de una película. Sin embargo, nuestra experiencia indica que si hubiera alguna granularidad en el tiempo, ésta sería tan pequeña que no podríamos distinguirla y, dado que el espacio parece continuo y ya nos acostumbramos a esa idea, resulta natural concebir al tiempo como un continuo y representarlo con números reales.

Para medir el tiempo también se elige un origen, es decir, un evento o acontecimiento que ocurre en un instante preciso, por ejemplo, el nacimiento de Cristo o el momento en que damos el disparo de salida para una competencia deportiva. También se requiere de una unidad, que debe ser el tiempo que pasa entre dos eventos repetibles. Hay una unidad

muy natural para el tiempo que casi todas las civilizaciones han tomado como unidad: el día o el lapso que tarda el Sol desde que asoma por el horizonte, hasta que vuelve a asomar por el mismo lugar al día siguiente. Las otras unidades de tiempo que se utilizan son fracciones enteras del día. La hora es la veinticuatroava parte del día, el minuto la sesentava parte de la hora, etcétera. Conviene observar que estas unidades de tiempo sólo tienen sentido en la Tierra; unos extraterrestres, si existieran, usarían con seguridad unidades diferentes para medir el tiempo y las distancias espaciales, simplemente porque no nos hemos puesto de acuerdo con ellos en una medida común. Las unidades son arbitrarias pero imprescindibles para aritmetizar el espacio y el tiempo. Las unidades de medida que puedan usar otros pueblos o civilizaciones tendrán necesariamente una equivalencia con las nuestras. Quizá algunos extraterrestres usen como unidad del tiempo una que para nosotros equivalga, por ejemplo, a 32.27 horas y quizás algún "pueblo bárbaro" se empeñe en usar una unidad de distancia que equivalga a 1 609 de nuestros metros.

3.3.3 El movimiento

Los objetos pueden moverse, esto quiere decir que en cada momento el cuerpo puede ocupar una posición diferente. Gracias a que podemos describir la posición y el tiempo con números, también podemos describir matemáticamente el movimiento. Consideremos un cuerpo relativamente pequeño, que suele llamarse partícula, que se mueve de una posición a otra en un intervalo de tiempo. Sean t_1 y t_2 los instantes de tiempo al inicio y al final del movimiento, y sean (x_1, y_1, z_1) y (x_2, y_2, z_2) las posiciones inicial y final de la partícula.

Esta situación puede representarse matemáticamente mediante una función del tiempo al espacio, concretamente, una función que asocia a cada instante de tiempo t una posición (x, y, z). Una función así puede representar no sólo el inicio y el final del movimiento, sino todo el recorrido. Llamemos $\vec{r}$ a la función que asigna a cada instante de tiempo t del intervalo $[t_1, t_2]$, la posición (x, y, z) de la partícula en ese instante. Entonces, esta función debe cumplir ciertas condiciones. En primer lugar $\vec{r}(t_1)$ debe ser (x_1, y_1, z_1) y $\vec{r}(t_2)$ debe ser (x_2, y_2, z_2). En segundo lugar, la función $\vec{r}$ debe ser **continua**, es decir, no debe dar saltos sino llevar el punto (x_1, y_1, z_1) al punto (x_2, y_2, z_2) a lo largo de una trayectoria continua, como el arco que se muestra en la figura.

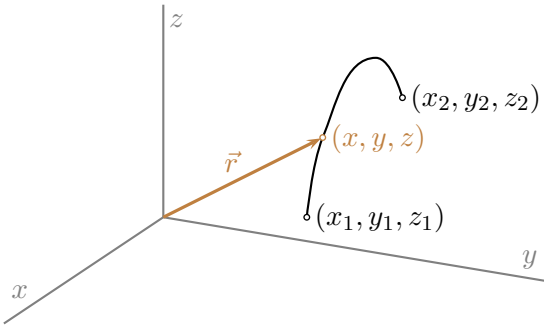


Figura 3.20 Una trayectoria en el espacio se determina por una función de tres coordenadas $\mathbf{r}(t) = (x(t), y(t), z(t))$, *donde la variable t representa al tiempo.*

Una función así puede construirse con tres funciones continuas $x(t), y(t), z(t)$ reales de variable real, definidas en el intervalo $[t_1, t_2]$. De esta manera es posible definir con toda precisión el movimiento de una partícula a través del espacio. La función —vectorial— $\vec{r}$, o bien las funciones escalares $x(t), y(t), z(t)$, determinan completamente el movimiento y se llaman ecuaciones paramétricas, en las que el parámetro es el tiempo.

Por ejemplo, el movimiento de una piedra lanzada desde el origen puede representarse con las siguientes funciones:

$$\begin{aligned} x(t) &= v_x \cdot t \\ y(t) &= v_y \cdot t \\ z(t) &= v_z \cdot t + \frac{g \cdot t^2}{2} \end{aligned}$$

donde t representa el tiempo transcurrido desde que se lanzó la piedra y v_x, v_y y v_z son las componentes de la velocidad inicial en las direcciones de los ejes coordenados. Otro ejemplo, serían las funciones:

$$\begin{aligned} x(t) &= r \cdot \cos(\omega t) \\ y(t) &= r \cdot sen(\omega t) \\ z(t) &= 0 \end{aligned}$$

que representan un movimiento circular uniforme, a lo largo de una circunferencia de radio r que yace en el plano horizontal, de una partícula que gira a razón de ω radianes por unidad de tiempo.

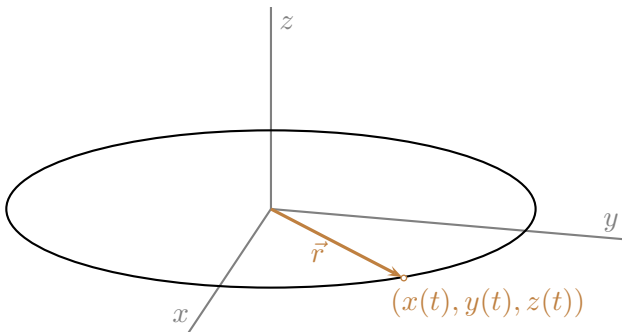


Figura 3.21 Trayectoria de un movimiento circular uniforme en el plano xy.

3.3.4 La velocidad y el concepto de derivada

Si conocemos las ecuaciones paramétricas del movimiento de un cuerpo, ¿podemos determinar su velocidad en cada instante? Esta pregunta, aparentemente sencilla, dio lugar a que Isaac Newton creara el cálculo diferencial. La velocidad de un cuerpo se define como el cociente de la distancia que cubre dividido por el tiempo que tarda en recorrerla. Ésta es una definición adecuada de velocidad para un cuerpo que se mueve con velocidad constante a lo largo de una línea recta. Sin embargo, para un cuerpo que se mueve a lo largo de una curva, el concepto de velocidad es más complicado por dos razones: en primer lugar, necesitamos que la velocidad indique no sólo cuán rápido se mueve sino en qué dirección lo hace; en segundo lugar, la velocidad puede ser distinta en distintos momentos durante el movimiento, por lo tanto, hace falta definir el concepto de velocidad instantánea. El concepto de velocidad requiere de una definición precisa y rigurosa que satisfaga la noción intuitiva que tenemos de ella.

La manera en que se define la **velocidad instantánea** $\mathbf{v}(t)$ en el instante t, es con un límite, específicamente, como el límite de los cocientes del desplazamiento del cuerpo en un intervalo de tiempo $[t, t+h]$, dividido entre el tamaño h de ese intervalo, cuando h tiende a cero. En símbolos, lo anterior se escribe así:

$$\mathbf{v}(t) = \lim_{h \longrightarrow 0} \frac{\vec{r}(t+h) - \vec{r}(t)}{h}$$

Newton llamaba a esto la *fluxión* de $\vec{r}$; hoy día se le llama la *derivada* de $\vec{r}$ con respecto a t.

Dada una función cualquiera f de una variable x, la **derivada** de f en el punto x se denota por $f'(x)$ o por $\frac{df}{dx}$ y se define como:

$$f\,'(x) = \frac{df}{dx} = \lim_{h \longrightarrow 0} \frac{f(x+h) - f(x)}{h},$$

siempre y cuando este límite exista. La definición es muy general y aplica cuando los valores de f son números reales, pero también cuando son vectores en el espacio de tres dimensiones e, incluso, puede extenderse a funciones con valores en espacios mucho más generales —llamados espacios vectoriales topológicos.

El concepto de derivada de una función tiene dos aplicaciones principales, una es la que generó el propio concepto de derivada, es decir, la velocidad instantánea, y la otra es la que está relacionada con la gráfica de una función f, que tiene valores reales. La derivada de la función en un punto x viene a ser *la pendiente de la recta tangente a la gráfica* de la función en el punto $(x, f(x))$, como se indica en la figura 3.22.

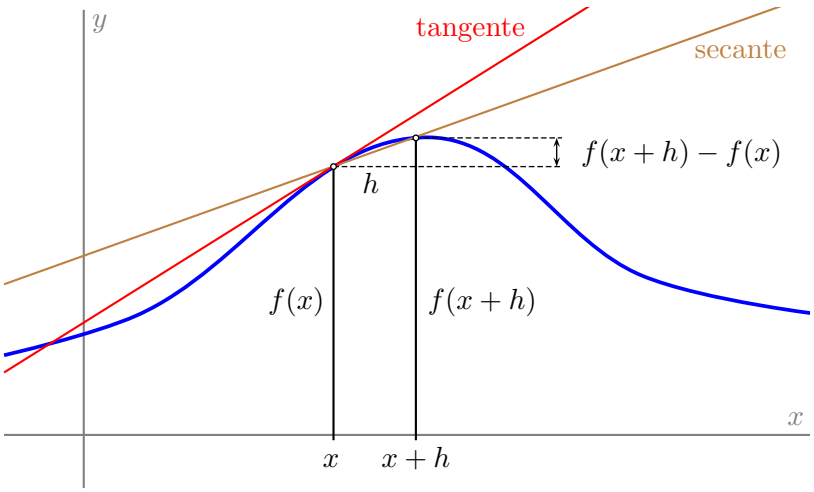


Figura 3.22 La derivada es la pendiente de la recta tangente a la gráfica de f en el punto $(x, f(x))$.

Esto se debe a que los cocientes:

$$\frac{f(x+h) - f(x)}{h}$$

son iguales a la pendiente de la recta —llamada secante—, que pasa por los puntos $(x, f(x))$ y $(x+h, f(x+h))$.

Para obtener las derivadas de las funciones, hay que calcular los límites correspondientes. Calculemos, por ejemplo, la derivada de x^3:

$$\begin{aligned}
\frac{dx^3}{dx} &= \lim_{h \longrightarrow 0} \frac{(x+h)^3 - x^3}{h} \\
&= \lim_{h \longrightarrow 0} \frac{3x^2h + 3xh^2 + h^3}{h} \\
&= \lim_{h \to 0} 3x^2 + 3xh + h^2 \\
&= 3x^2
\end{aligned}$$

Aunque el cálculo de la derivada de una función no suele ser muy complicado si uno tiene claro el concepto de límite, es conveniente memorizar las derivadas de las funciones más frecuentes, como las que aparecen en la siguiente tabla:

$$\begin{aligned}
\frac{d\ c}{dx} &= 0 \\
\frac{d\ x}{dx} &= 1 \\
\frac{d\ x^2}{dx} &= 2x \\
\frac{d\ x^n}{dx} &= nx^{n-10}, \\
\frac{d\ sen(x)}{dx} &= \cos(x) \\
\frac{d\ \cos(x)}{dx} &= -sen(x)
\end{aligned}$$

También es importante saber las reglas que se aplican al derivar sumas y productos de funciones. Sean f y g dos funciones cualesquiera, entonces:

$$\begin{aligned}
\frac{d(cf)}{dx} &= c\frac{df}{dx}, \text{ para cualquier constante } c \\
\frac{d(f+g)}{dx} &= \frac{df}{dx} + \frac{dg}{dx} \\
\frac{d(fg)}{dx} &= f\frac{dg}{dx} + g\frac{df}{dx}
\end{aligned}$$

Con estas pocas fórmulas y reglas se pueden obtener las derivadas de muchas funciones. Los cursos de cálculo suelen dedicar bastante atención a que los alumnos aprendan a obtener derivadas. En ocasiones, se usan otras pocas fórmulas que aparecen en las tablas de derivación de los libros. Sabiendo derivar se pueden resolver muchos problemas como, por ejemplo, saber la velocidad instantánea de una partícula dada su ecuación de movimiento. En particular, si regresamos al ejemplo del movimiento circular uniforme descrito por las ecuaciones paramétricas:

$$\begin{aligned}
x(t) &= r \cdot \cos(\omega t) \\
y(t) &= r \cdot sen(\omega t) \\
z(t) &= 0
\end{aligned}$$

podemos obtener las tres coordenadas del vector velocidad derivando las ecuaciones con respecto a t:

$$\begin{aligned}
v_x(t) &= -r\omega \cdot \cos(\omega t) \\
v_y(t) &= r\omega \cdot sen(\omega t) \\
v_z(t) &= 0
\end{aligned}$$

y, si volvemos a derivar con respecto a t, obtenemos las componentes de lo que se llama el vector **aceleración**:

$$\begin{aligned}
a_x(t) &= -r\omega^2 \cdot sen(\omega t) \\
a_y(t) &= -r\omega^2 \cdot \cos(\omega t) \\
a_z(t) &= 0
\end{aligned}$$

La aceleración y no la velocidad es lo que se relaciona directamente con la fuerza que se aplica a un cuerpo, esto es, el contenido de la segunda ley de Newton. Es interesante observar que el vector aceleración $\vec{a}$ en este caso es proporcional al vector de posición $\vec{r}$, concretamente:

$$\vec{a} = -\omega^2 \vec{r}$$

En otras palabras, la aceleración de un cuerpo en movimiento circular uniforme es hacia el centro de su trayectoria circular y de magnitud igual al radio multiplicado por ω^2, el cuadrado de la velocidad angular. Por lo tanto, si un cuerpo sigue un movimiento circular uniforme, es porque hay una fuerza $\vec{F}$ —llamada *fuerza centrípeta*— que lo impulsa hacia el centro de la trayectoria:

$$\vec{F} = -m \cdot \omega^2 \vec{r}$$

Recordemos al lector que en el tema 1 de este libro hay una deducción de la tercera ley de Kepler basada en este resultado.

De manera similar se pueden obtener muchos resultados útiles sobre el movimiento de los cuerpos utilizando las derivadas.

3.3.5 La integral y el teorema fundamental del cálculo

La segunda ley de Newton nos permite encontrar la trayectoria de una partícula si sabemos las fuerzas que actúan sobre ella. Sin embargo, esto no siempre resulta fácil. Poder resolver este tipo de problemas fue una de las principales motivaciones de Newton para desarrollar el cálculo integral. Si conocemos la fuerza que actúa sobre una partícula, por la segunda ley de Newton, sabemos cuáles son los valores de la aceleración, es decir, de la derivada de la velocidad con respecto al tiempo. Para obtener la velocidad de la partícula necesitamos resolver el problema inverso a la derivación, es decir, necesitamos encontrar la velocidad como una función del tiempo cuya derivada sea la aceleración que conocemos. El proceso inverso de la derivación se llama **integración**. Los matemáticos de la época de Newton y Leibniz, con ellos incluidos, se dieron cuenta de que la integración tenía mucho que ver con el cálculo de áreas y por ese motivo resulta conveniente comenzar el estudio de la integración con el concepto de área bajo la gráfica de una función, que se denomina **integral definida**.

La integral definida de una función entre dos puntos de la recta real a y b se define mediante un límite:

$$\int_a^b f(x)dx = \lim_{N \longrightarrow \infty} \sum_{n=1}^{N} f(x_i)\Delta x \qquad (3)$$

donde N es un número entero positivo, $\Delta x = \frac{b-a}{N}$ y los valores x_i son, por ejemplo, los puntos $a + (i + \frac{1}{2})\Delta x$ que están en medio de los intervalos de ancho Δx y donde también se ha usado el símbolo $\int$ llamado integral. La expresión $\int_a^b f(x)dx$ se lee "la integral entre a y be de efe de equis, de equis". Obsérvese que cuando N tiende a infinito, Δx tiende a cero y viceversa; en símbolos, $N \longrightarrow \infty$ si y sólo si $\Delta x \longrightarrow 0$.

El límite de (3) mide el área bajo la gráfica de $f(x)$ entre los puntos a y b, como se ilustra en la figura 3.23.

De la misma manera que el concepto de la derivada está íntimamente relacionado con su significado geométrico de ser la pendiente de la recta tangente a la gráfica de la función, la integral definida entre dos puntos a y b está íntimamente relacionada con el área delimitada por la gráfica, el eje x y los puntos a y b. De hecho, el área bajo la curva se define mediante la integral definida. Ambos conceptos, el de tangente a una curva y el de área bajo la curva, sólo pueden definirse de manera rigurosa recurriendo a los conceptos de derivada e integral definida, respectivamente.

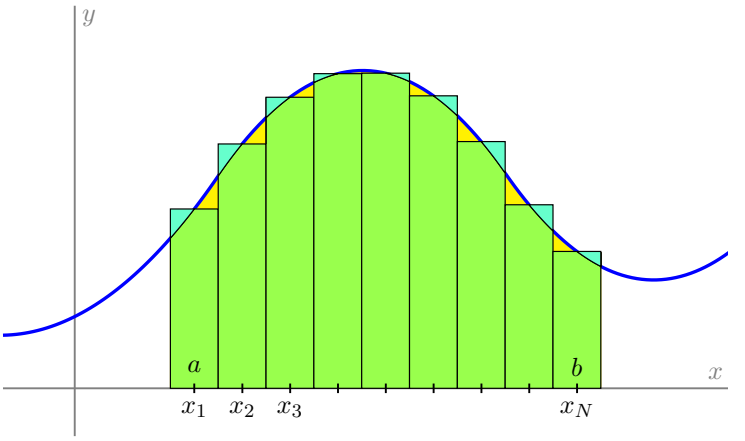


Figura 3.23 Aproximación del área bajo la curva mediante una función escalonada.

Dada una función $f(x)$ puede definirse otra con el área bajo la gráfica de la función entre un punto fijo a y uno variable x. En símbolos, esta función de $F(x)$ se escribe así:

$$F(x) = \int_a^x f(s)ds$$

Las funciones definidas de esa manera tienen mucha importancia en las matemáticas. Por ejemplo, en la teoría de la probabilidad, si $f(x)$ representa la densidad de probabilidad de una variable aleatoria, entonces $F(x)$ es la probabilidad de que la variable tome valores en el intervalo $[a, x]$. En forma geométrica, la función $F(x)$ definida así representa el área bajo la gráfica de f desde a hasta x y, si se concibe como una función del extremo superior de integración x, puede derivarse al igual que cualquier otra función de x, sólo que en este caso el resultado es muy interesante:

$$\frac{dF}{dx} = f(x)$$

es decir, la derivación y la integración son procesos inversos uno del otro. Éste es uno de los resultados más importantes de la historia de las matemáticas pues ha permitido hacer cálculos exactos de muchos procesos límite que, de otra manera, sólo hubieran podido obtenerse en forma aproximada. El nombre de cálculo que se da al estudio de las derivadas, las integrales y su íntima relación, se debe precisamente a que dicho estudio ofrece una metodología muy poderosa para realizar cálculos en campos muy diversos como la física, la geometría y la probabilidad.

El **teorema fundamental del cálculo** exhibe la relación íntima que hay entre las integrales y las derivadas. Muestra que los procesos de integración y derivación son cada uno el inverso del otro. El teorema tiene dos versiones:

1] La derivada de la función $F(x) = \int_a^x f(s)ds$ es igual a $f(x)$. Es fácil entender por qué esto es cierto si aplicamos la definición de derivada a la función $F(x)$:

$$\begin{aligned} F'(x) &= \lim_{h \longrightarrow 0} \frac{\int_a^{x+h} f(s)ds - \int_a^x f(s)ds}{h} \\ &= \lim_{h \longrightarrow 0} \frac{\int_x^{x+h} f(s)ds}{h} = f(x) \end{aligned}$$

en donde hemos interpretado el primer numerador como una diferencia de áreas y lo hemos escrito como una integral definida entre x y $x + h$, y luego, hemos aprovechado el hecho de que cuando h es muy pequeño, $\int_x^{x+h} f(s)ds$ es casi igual a $f(x)h$.

La segunda forma del teorema fundamental dice que:

2] Si la derivada de una función $F(x)$ es otra función $f(x)$, entonces $\int_a^b f(x)dx = F(b) - F(a)$. Para entender intuitivamente por qué ocurre esto, conviene remplazar la integral por una suma ya muy cercana a ella, digamos $\sum_{n=1}^{N} f(x_i)\Delta x$, y sustituir $f(x_i)$ por el cociente:

$$\frac{F(x_i) - F(x_{i-1})}{\Delta x}$$

donde $x_i = a + i\Delta x$. La suma puede escribirse entonces como:

$$\sum_{n=1}^{N} f(x_i)\Delta x = \sum_{n=1}^{N} F(x_i) - F(x_{i-1}) = F(b) - F(a)$$

puesto que los términos $F(x_i)$ se cancelan por aparecer una vez con signo positivo y otra con signo negativo, así que sólo sobreviven los correspondientes a los extremos a y b.

Volvamos a considerar las funciones que se definen usando la integral así: $F(x) = \int_a^x f(x)dx$. Diferentes puntos fijos de inicio, digamos a_1 y a_2, dan lugar a diferentes funciones $F_1(x) = \int_{a_1}^x f(s)ds$ y $F_2(x) = \int_{a_2}^x f(s)ds$ pero tales funciones difieren entre sí sólo por una constante, que es precisamente igual al área bajo la gráfica de la función entre los puntos a_1 y a_2. Las funciones cuya derivada es $f(x)$ se llaman **primitivas** o **antiderivadas** de f. También se dice de ellas que son la **integral indefinida** de f, lo cual se escribe así:

$$\int f(x)dx$$

Las **tablas de integrales** muestran las funciones primitivas, antiderivadas o integrales indefinidas de varias funciones. La tabla siguiente es una pequeña muestra, donde c es una constante arbitraria:

$$\begin{aligned}
\int 1dx &= x + c \\
\int xdx &= \frac{x^2}{2} + c \\
\int x^n dx &= \frac{x^{n+1}}{n+1} + c \\
\int sen(x)dx &= \cos(x) + c \\
\int \cos(x)dx &= -sen(x) + c
\end{aligned}$$

Las tablas de integrales pueden obtenerse de las tablas de derivadas. Cada fórmula de derivación nos da una de integración. Dado que es mucho más fácil calcular las derivadas que las integrales cuando partimos de las definiciones como límites, el teorema fundamental nos resuelve de manera muy general un problema. Nos permite usar las tablas de integración para realizar cálculos que de otra manera serían mucho más complicados. Ésta es la razón por la que el cálculo es una herramienta tan poderosa. Por ejemplo, si intentáramos calcular el área bajo la gráfica de $f(x) = \cos(x)$ entre $-\frac{\pi}{2}$ y $\frac{\pi}{2}$, sin el teorema fundamental tendríamos que calcular un límite muy difícil. Pero gracias al teorema fundamental sólo necesitamos encontrar la antiderivada de $f(x) = \cos(x)$, la cual sabemos que es $sen(x)$, y por lo tanto:

$$\int_{-\pi/2}^{\pi/2} \cos(x)dx = sen(\frac{\pi}{2}) - sen(-\frac{\pi}{2}) = 1 - (-1) = 2$$

Aunque no todos los cálculos que se realizan usando la integral son tan sencillos, el teorema fundamental permite muchas veces llegar a resultados exactos que sin él serían poco menos que imposibles.

3.3.6 Newton y las leyes de Kepler, una nueva concepción del Universo

Como ya vimos previamente, los conceptos del cálculo están íntimamente relacionados con la física. La velocidad no puede definirse con rigor sin la derivada, mientras que el cálculo de la trayectoria de un cuerpo requiere del uso de los poderosos métodos del cálculo.

La primera y más impactante aplicación del cálculo a la física fue la deducción matemática de las leyes del movimiento de los planetas, es decir, de las leyes de Kepler, a partir de la segunda ley de Newton:

$$\vec{F} = m\vec{a}$$

así como de la ley de la gravitación universal, en forma vectorial:

$$F = G\frac{mM}{r^2}$$

Usando los métodos del cálculo se puede demostrar que, si suponemos que la fuerza que actúa sobre un cuerpo celeste es la fuerza gravitatoria del Sol, entonces su trayectoria será una curva cónica —elipse, parábola o hipérbola— que tendrá al Sol en uno de sus focos. Los planetas, como giran alrededor del Sol, no pueden tener órbitas parabólicas o hiperbólicas, por lo tanto sus órbitas son elípticas: segunda ley de Kepler. También, a partir de la segunda ley de Kepler se obtiene que el radio vector del Sol al cuerpo barrerá áreas iguales en tiempos iguales. La tercera ley de Kepler se puede deducir de igual forma con absoluto rigor: los cuadrados de los tiempos de revolución de los planetas alrededor del Sol son proporcionales a los cubos de los semiejes mayores de sus órbitas elípticas.

No vamos a mostrar cómo se usa el cálculo para deducir las leyes de Kepler, tal demostración puede encontrarse en los libros avanzados. Pero es importante que quede claro al lector que se trata de resultados de una extraordinaria importancia, pues llevan al hombre a cambiar su concepción del Universo, le hacen dar el paso definitivo entre la superstición y la ciencia. Y en este paso las matemáticas jugaron un papel central.

El Universo parece comportarse siguiendo unas leyes de naturaleza matemática y todos los detalles del movimiento de los cuerpos pueden deducirse matemáticamente de estas leyes. Éste fue uno de los descubrimientos científicos más importantes de la historia, si no es que el más importante. Se trata de un descubrimiento que pone a la razón y en particular a las matemáticas en el gobierno de la naturaleza, con lo que la existencia de deidades que se encargan de decidir todo lo que ocurre deja de ser lógicamente necesaria. El Universo tiene leyes racionales, leyes que se expresan con fórmulas matemáticas y que determinan todo lo que ocurre. La religión, que se había adueñado por muchos siglos del poder sobre el Universo recibió un duro golpe cuando se descubrió que las maravillas de la naturaleza podían explicarse, en principio, a partir de fórmulas y, por lo tanto, no era necesario recurrir a un sacerdote —supuestamente en contacto privilegiado con los dioses— para cambiar el curso de los acontecimientos, sino que se podía actuar directamente sobre las causas para producir los efectos deseados. Dejan de ser necesarios intermediarios para influir sobre el mundo, ahora ya sólo bastan la ciencia y la tecnología.

A medida que los descubrimientos de Newton se fueron popularizando y alcanzaron a las personas interesadas en la política, se convirtieron en herramientas de liberación. La Ilustración y por tanto la Revolución francesa, madre de todas las revoluciones sociales posteriores, deben mucho a la concepción del Universo que nos legó Newton. El cálculo juega un papel fundamental en esta historia, los propios conceptos que permitieron entender el Universo son entes matemáticos.

La siguiente sección puede dar una idea al lector de por qué la segunda ley de Newton y la ley de la gravitación universal llevan a las leyes de Kepler. No se trata de una deducción rigurosa en la que habría que usar el cálculo, sino de un divertimento geométrico que hace plausible la deducción.

3.4 LAS ÓRBITAS CELESTES

Figura 3.24 Richard Feynman, físico estadunidense que ganó el Premio Nobel por su trabajo sobre la unión de dos áreas de la física: la mecánica cuántica y la electrodinámica. Fue un hombre con extraordinario talento no sólo en física, como se puede constatar en libros autobiográficos | © Latin Stock México.

Como revisamos en la sección 1.3, Newton desarrolló el cálculo diferencial para estudiar el movimiento, el fenómeno de cambios continuos, y explicó de esta manera, con base en principios básicos, las leyes de Kepler. Aquí haremos algo distinto, pero no menos interesante: se discutirá cómo se pueden hacer verosímiles las leyes de Kepler a partir de principios geométricos. El desarrollo se basa en una idea del físico Richard Feynman. Es importante resaltar que se trata de consideraciones verosímiles y no de una demostración exacta; por ello, en casi todos lados tendremos aproximaciones —y usaremos el símbolo $\approx$ para denotarlo— y no igualdades exactas.

Un cuerpo que se mueve en el espacio sin ninguna atracción describe una trayectoria recta y se traslada distancias iguales en tiempos iguales. Es decir, en cada momento avanza la misma distancia respecto a un observador que se piensa inmóvil. Podemos observar que, en este caso, se cumple lo que enuncia la segunda ley de Kepler: el radio vector del Sol a un planeta barre áreas iguales en tiempos iguales.

Para cualquier punto del espacio O, el área del triángulo P_1Q_1O es igual al área del triángulo P_2Q_2O si P_1, P_2 marcan dos posiciones en la trayectoria y Q_1, Q_2 son las posiciones respectivas después de un intervalo Δt de tiempo.

Ahora analizamos un caso más complejo en el que, en algún momento, un asteroide golpea el cuerpo, lo que produce un cambio casi instantáneo en la dirección del movimien-

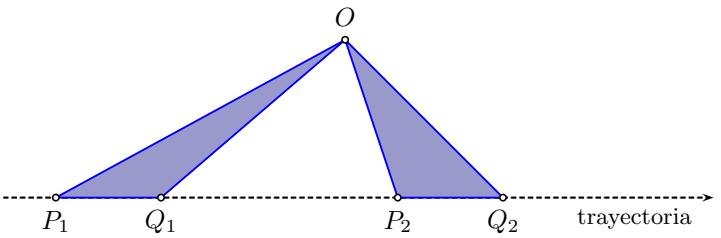


Figura 3.25 Ilustración de la segunda ley de Kepler en el movimiento uniforme.

to, es decir, en la velocidad. Obviamente, el cambio en la velocidad depende de la "intensidad" con la que golpeó el cuerpo y de la dirección del golpe. Pero aquí no nos interesa demasiado analizar el mecanismo del golpe, sino el efecto en la velocidad. Podemos simplificar el fenómeno y pensar que, en este momento, la velocidad se altera por la diferencia $\Delta\vec{v}$, es decir, la velocidad posterior al impacto $\vec{v}_2$ se calcula a partir de la velocidad anterior $\vec{v}_1$:

$$\vec{v}_2 = \vec{v}_1 + \Delta\vec{v}$$

Sabemos que las velocidades se suman formando el paralelogramo correspondiente, según se ve en la figura 3.26. De nuevo, nos preguntamos si es cierto que se cumple la segunda ley de Kepler para cualquier observador, es decir, si para cualquier punto O se tiene que el vector que une O con el planeta barre áreas iguales en tiempos iguales. La respuesta es negativa pues no se cumple para cualquier punto O, pero sí lo hace para puntos que están sobre la línea que marca el vector $\Delta\vec{v}$ desde el punto P_2 del impacto (véase figura 3.26).

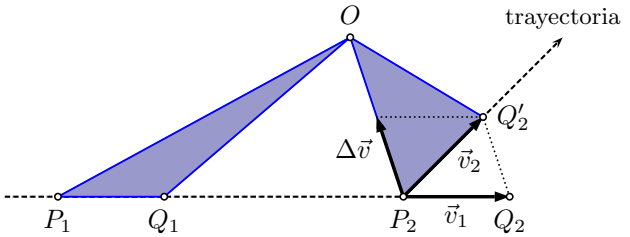


Figura 3.26 En el punto P_2, el asteroide golpea al planeta y hay una desviación en la trayectoria. El cambio de la velocidad $\Delta\vec{v}$ se suma con la velocidad anterior $\vec{v}_1$ para dar como resultado la velocidad posterior $\vec{v}_2 = \vec{v}_1 + \Delta\vec{v}$. Esta suma se forma al usar el paralelogramo con lados $\Delta\vec{v}$ y $\vec{v}_1$.

Para simplificar el dibujo supusimos que el impacto se realizó justo en la posición P_2. En vez de que el cuerpo siga hacia la posición Q_2, éste ahora se encuentra en Q'_2 después del mismo lapso Δt. Los dos triángulos P_2Q_2O y $P_2Q'_2O$ comparten el lado OP_2 y, además, tienen la misma altura —respecto a OP_2— dado que $Q_2Q'_2$ es paralelo a OP_2, porque $\Delta\vec{v}$ está dirigido a O en el momento del impacto.

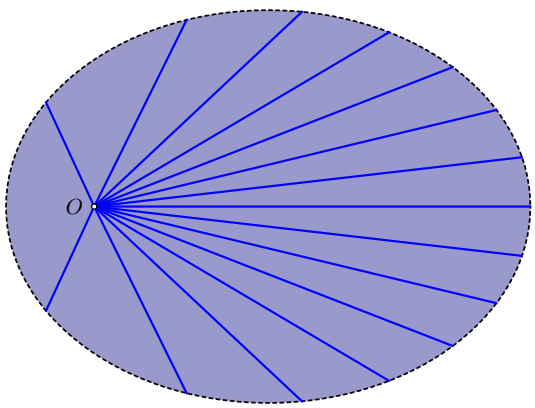


Figura 3.27 La partición de una órbita en intervalos de tiempos iguales, o lo que es lo mismo, en áreas "barridas" iguales.

En las órbitas celestes se supone que la masa del Sol es muy grande y, por lo tanto, éste no se mueve y sólo atrae a los planetas hacia sí mismo. En realidad, también el Sol es atraído hacia los planetas pero esta atracción tiene un efecto pequeño sobre él y, en una primera aproximación, no hay que tomarlo en cuenta. Si nos imaginamos la fuerza gravitacional ejercida por el Sol sobre los planetas como una sucesión de pequeños golpes, cada uno de estos golpes produce un cambio de velocidad dirigido hacia el Sol. Por lo que vimos antes con el asteroide, esto tiene como consecuencia que el radio del Sol a la Tierra barre áreas iguales en tiempos iguales, según se muestra en la figura 3.27. Lo único que se usó hasta este momento en el argumento fue que tenemos una *atracción central*, es decir que la fuerza siempre se dirige al mismo punto O.

En realidad, la fuerza es continua y se ejerce todo el tiempo, pero imaginar que se trata de pequeños impactos nos permite aplicar la geometría y arribar a la segunda ley de Kepler de manera intuitiva y clara sin tener que usar herramientas matemáticas más avanzadas.

Ahora haremos una división diferente de la órbita. La dividimos en partes, llamadas sectores, tal que cada una de ellas incluya el mismo ángulo $\Delta\alpha$ en el punto O, como se muestra en la figura 3.28.

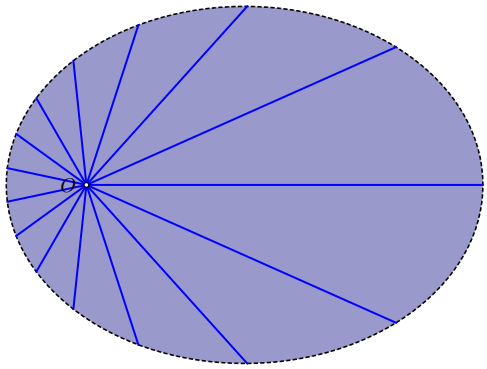


Figura 3.28 La ilustración muestra una partición de una órbita en sectores del mismo ángulo O.

Por lo que vimos con anterioridad, el área es proporcional al tiempo transcurrido en la órbita. Regresamos de nuevo al punto de vista de que el cuerpo recibe impulsos instantáneos en ciertos momentos, que ahora ya no serán después de cierto tiempo sino cada vez que se cubre cierto ángulo. Esto permitirá emplear nuevamente conceptos de la geometría. La figura 3.29 muestra cómo cambian las velocidades $\vec{v}_0, \vec{v}_1, \vec{v}_2 \ldots$

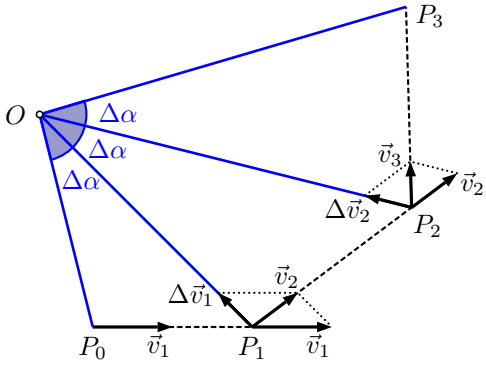


Figura 3.29 Los cambios de velocidades son constantes $|\Delta\vec{v}_1| \approx |\Delta\vec{v}_2|$.

Veremos que los cambios de velocidades $\Delta\vec{v}_1 \approx \vec{v}_2 - \vec{v}_1$, $\Delta\vec{v}_2 \approx \vec{v}_3 - \vec{v}_2, \ldots$ son iguales si estos cambios ocurren cada vez que se gira la órbita por el mismo ángulo. En otras palabras, se tiene $|\Delta\vec{v}_1| \approx |\Delta\vec{v}_2|$, donde las barras verticales indican que se considera solamente la longitud del vector, pero no la dirección. La aproximación tiende a ser una igualdad conforme el ángulo $\Delta\alpha$ se hace pequeño.

Para ello, necesitaremos por primera vez que la fuerza gravitacional que provoca la atracción sea proporcional al inverso del cuadrado de la distancia. Según la ley fundamental de Newton, la fuerza $\vec{F}$ es igual a la masa por la aceleración:

$$\vec{F} = m\vec{a}$$

donde m es la masa del cuerpo y $\vec{a}$ es la aceleración. Esta última es el cambio de velocidad. En nuestro modelo de golpes, la aceleración en el punto P_1 es:

$$\vec{a} \approx \frac{\Delta\vec{v}_1}{\Delta t_1}$$

donde Δt_1 es el tiempo que transcurre para que llegue el cuerpo del punto P_0 al punto P_1. Como la masa es constante y la fuerza proporcional al cuadrado de la distancia $r_1 = OP_1$, se obtiene que:

$$\frac{|\Delta\vec{v}_1|}{\Delta t_1} \propto \frac{1}{r_1^2}.$$

Aquí, el símbolo $\propto$ expresa la proporcionalidad, esto es que:

$$\frac{|\Delta\vec{v}_1|}{\Delta t_1} \approx k\frac{1}{r_1^2}, \quad \frac{|\Delta\vec{v}_2|}{\Delta t_2} \approx k\frac{1}{r_2^2}, \quad \frac{|\Delta\vec{v}_3|}{\Delta t_3} \approx k\frac{1}{r_3^2}, \ldots \quad (4)$$

siempre con la misma constante de proporcionalidad k.

Por otro lado, el área ΔA_1 del triángulo OP_0P_1 es $\Delta A_1 \approx \frac{1}{2r_1^2}\Delta\alpha$, donde se usa que $\operatorname{sen}(\Delta\alpha) \approx \Delta\alpha$ es una buena aproximación para ángulos pequeños. En particular, el área ΔA_1 es proporcional al cuadrado de la distancia r_1, si el ángulo $\Delta\alpha$ se mantiene constante. Pero, por lo que vimos antes, esta área es también proporcional al tiempo Δt_1. En conclusión, si $\Delta\alpha$ es constante, entonces:

$$\frac{\Delta t_1}{r_1^2} \approx \frac{\Delta t_2}{r_2^2} \approx \frac{\Delta t_3}{r_3^2} \approx \ldots$$

y de aquí sigue que a partir de las ecuaciones (4), tenemos que:

$$|\Delta\vec{v}_1| \approx |\Delta\vec{v}_2| \approx |\Delta\vec{v}_3| \approx \ldots$$

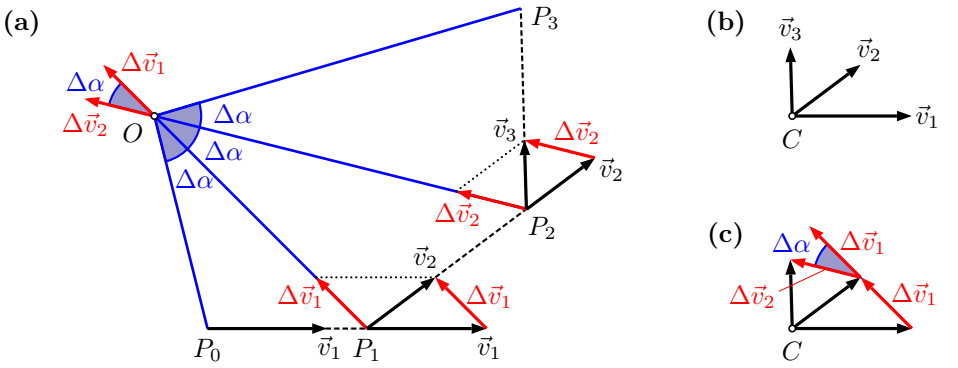


Figura 3.30 Construcción del hodograma.

Ahora podemos dibujar el diagrama de las velocidades que se llama *hodograma*. Éste se obtiene al trasladar todas las velocidades que tiene el cuerpo a lo largo de su trayectoria de manera paralela, tal que todos los vectores empiecen en un mismo punto C, como se observa en la parte (b) de la figura 3.30. Se debe notar que los vectores $\vec{v}_1$ en la parte (a) y (b) son

paralelos. El vector que une la punta de $\vec{v}_1$ con la de $\vec{v}_2$ es $\Delta\vec{v}_1$, dado que $\vec{v}_2 \approx \vec{v}_1 + \Delta\vec{v}_1$, como se observa en la parte baja de la figura 3.30(a). En forma similar, $\Delta\vec{v}_2$ une la punta de $\vec{v}_2$ con la de $\vec{v}_3$.

Finalmente, a partir de todo lo que se ha argumentado se puede llegar a la siguiente conclusión: los vectores $\Delta\vec{v}_1, \Delta\vec{v}_2, \ldots$ son todos igual de largos y, además, difieren en su dirección siempre por el mismo ángulo $\Delta\alpha$, según se puede observar si se trasladan hasta el punto O; véase parte (a) en la figura 3.30. El resultado se muestra en la parte (c) de la misma figura. La conclusión es que la aproximación del hodograma formado por las puntas de los vectores $\vec{v}_1, \vec{v}_2, \vec{v}_3, \ldots$ es un polígono regular, dado que su contorno está formado por los vectores $\Delta\vec{v}_1, \Delta\vec{v}_2, \ldots$ que, como ya dijimos, son igual de largos y siempre incluyen el mismo ángulo.

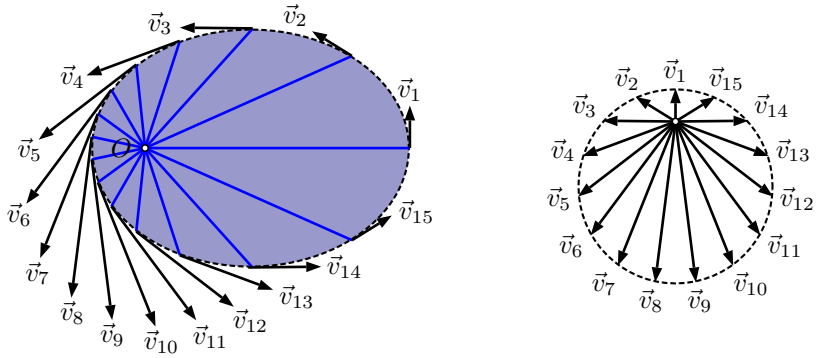


Figura 3.31 El hodograma de una órbita elíptica.

Al hacer más pequeño el ángulo $\Delta\alpha$, el polígono regular se aproximará a una circunferencia. Concluimos que el hodograma es una circunferencia si el cuerpo se mueve en un campo de atracción central con una fuerza que es inversamente proporcional al cuadrado de la distancia. Es importante observar que el punto C —el "origen" del hodograma— no es el centro de la circunferencia. Esto se puede ver en la figura 3.31. Pero todavía no hemos demostrado que las órbitas sean elípticas y la figura 3.31 es, meramente, una ilustración.

A cada punto P de la órbita le corresponde un punto del hodograma P', que se marca por la punta del vector de velocidad $\vec{v}_P$, que es la velocidad que tiene el cuerpo en el punto P. Por ejemplo, si P es el punto más a la derecha de la órbita en la figura 3.31, entonces P' es el punto más arriba en el hodograma.

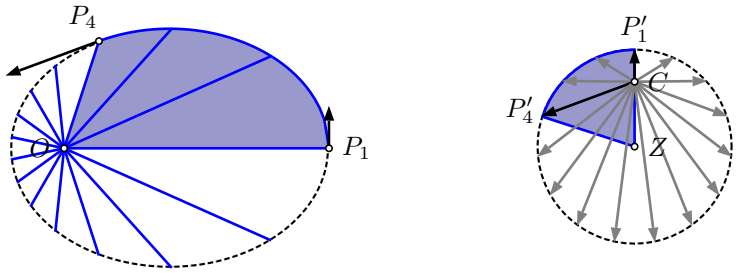


Figura 3.32 La correspondencia de ángulos β.

Hemos visto que si dividimos la órbita en sectores de ángulos iguales, los puntos correspondientes en el hodograma forman un polígono regular que tiene también sectores de ángulo iguales $\Delta\alpha$, si los sectores se miden con el centro Z del hodograma y no con el punto C (véase figura 3.32).

Finalmente, reunimos todos los elementos para poder llegar a la conclusión de que la órbita es realmente una cónica. Para ello, giramos el hodograma 90° y empalmamos el punto Z con el punto central O, como se muestra en la figura 3.33. Esto se hace con el efecto de empalmar los dos ángulos considerados anteriormente en la figura 3.32. Ahora se conside-

ra que el punto P' recorre el hodograma. Se traza la *mediatriz* t de CP' y se interseca con ZP' para obtener un punto P'', según se aprecia en la figura 3.33.

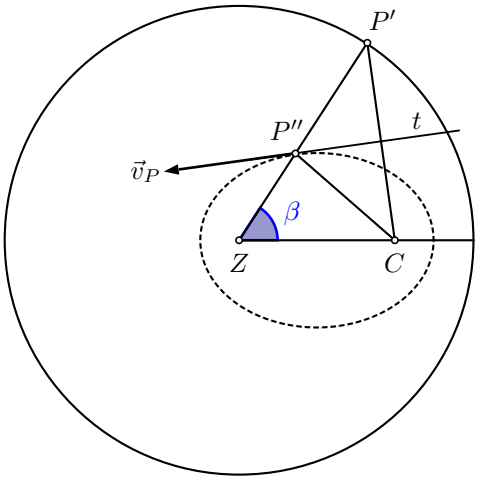


Figura 3.33 La órbita es una cónica.

Como P' recorre una circunferencia con centro Z, la distancia $\mathrm{d}(Z, P')$ no cambia. Pero $\mathrm{d}(Z, P') = \mathrm{d}(Z, P'') + \mathrm{d}(P'', P')$ y, como t es mediatriz de CP', se tiene que $\mathrm{d}(P'', P') = \mathrm{d}(P'', C)$. Por lo tanto:

$$\mathrm{d}(Z, P'') + \mathrm{d}(P'', C) = \mathrm{d}(Z, P')$$

que es constante. En consecuencia, el punto P'' se encuentra en una elipse con los focos C y Z. Además, como se puede verificar en la sección cónicas, la mediatriz t es tangente a la elipse.

El girar 90° el hodograma tiene otro efecto grato: la velocidad $\vec{v}_P$ del cuerpo en el punto P es paralela a t, dado que la velocidad $\vec{v}_P$ se representa en el hodograma original —antes de girar— por el vector CP'. Hemos encontrado una trayectoria para el cuerpo: la elipse. El punto P'' se mueve sobre una elipse y la velocidad $\vec{v}_P$ es, en cada momento, tangente a la trayectoria de P''.

Pero, en principio, estamos haciendo una barbaridad. El punto P'' pertenece a un diagrama de velocidades que es incomparable con un diagrama de posiciones si no se fija una escala común. Si invertimos el argumento, se puede aprovechar y decir que podemos fijar una escala tal que, en un momento, la posición P coincida con la de P''. Entonces, a partir de este momento, P y P'' se moverán al unísono: ambos se encuentran sobre ZP' por la igualdad de los ángulos β en los dos diagramas y ambos tienen, también, el mismo vector de velocidad. Como P'' se mueve por una elipse, lo mismo hace P.

Hemos llegado al final de la conclusión. Se quería demostrar que las trayectorias siempre son cónicas y encontramos lo que ocurre en el caso de las elipses, que se hizo para simplificar el argumento. Lo que se supuso sin argumento alguno es que el punto C *está en el interior* del hodograma. No obstante, habría que considerar realmente tres casos: cuando C está en el interior —esto nos da una elipse—, cuando C está sobre el hodograma —que nos dará una parábola— y finalmente, cuando C está afuera —que produce una hipérbola—. En cualquiera de estos casos se argumenta de manera similar, pero como la parábola y la hipérbola son curvas abiertas, corresponden a cuerpos que pasan al centro de atracción una sola vez y jamás regresan. Consecuentemente, las órbitas de los planetas y los cometas son elipses.

La argumentación anterior no es del todo sencilla, pero es interesante ver que es posible obtener una intuición geométrica del porqué de lo elíptico de las órbitas y no delegar nues-

tra comprensión a un cálculo diferencial que requiere familiaridad con herramientas más avanzadas, como los vectores.

Con ello, concluimos la argumentación geométrica que hace verosímiles las primeras dos leyes de Kepler a partir de dos principios: que hay una atracción central y que ésta es proporcional al inverso del cuadrado de la distancia. Esto es uno de los más grandes logros científicos de la humanidad.

3.5. LAS ECUACIONES QUE MODELAN EL MUNDO

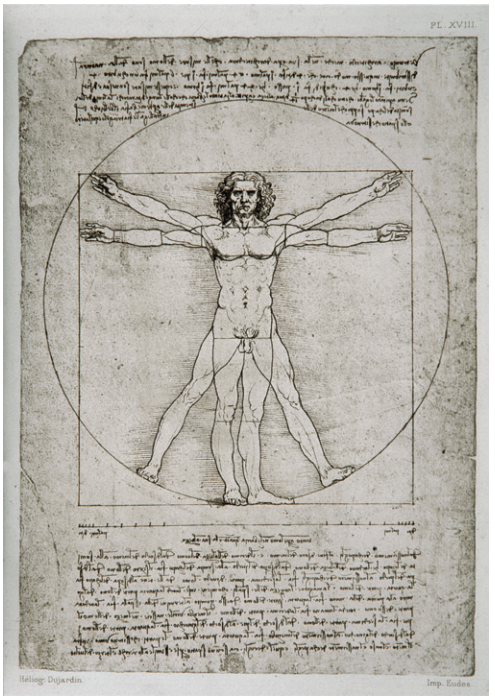


Figura 3.34 El hombre de Vitruvio *es un famoso trabajo de Leonardo da Vinci realizado en uno de sus diarios, alrededor de 1492. El dibujo representa una figura masculina desnuda —inscrita en un círculo y un cuadrado— y en dos posiciones sobreimpresas de pies y manos. Este modelo es un estudio de las proporciones del cuerpo humano que utiliza la razón áurea, descrita en la sección 4.2* | *© Latin Stock México.*

3.5.1 Primeros ejemplos

En el contexto matemático, un modelo es una herramienta que facilita la descripción, representación, explicación, predicción, discusión o evolución particular de un objeto, proceso, sistema o fenómeno. En esta sección veremos diferentes herramientas matemáticas que pueden usarse para la modelación y que hay una que, claramente, es la más potente: la ecuación diferencial. Antes de llegar a ello, veremos primero unos ejemplos que muestran la gran versatilidad de los conceptos.

Un número es impar si no es divisible entre dos. La suma de dos impares es par, como se ve, por ejemplo, en $3 + 5 = 8$. Pero si quisiéramos hacer una demostración de ello, necesitamos tener una expresión que nos modele el número impar en general. Ésta existe y es un término de la forma:

$$2n + 1$$

donde n es cualquier entero. Para diferentes valores de n obtenemos diferentes números impares, es más, obtenemos todos los impares y sólo impares. Con este modelo es fácil comprobar que, en general, la suma de dos impares —digamos $2m + 1$ y $2n + 1$— es par:

$$(2m+1)+(2n+1)=2m+2n+2=2(m+n+1),$$

ya que el lado derecho es un múltiplo de 2.

Veremos ahora cómo una *ecuación* puede modelar una situación interesante. El principio que impulsa los cohetes espaciales se basa en la conservación del impulso: si en el espacio vacío —donde no actúa la gravedad— un astronauta lanza un objeto, se impulsa en la dirección opuesta. El cohete trae combustible y, al quemarlo, expulsa gas a gran velocidad v_e —la velocidad de emisión— al espacio, lo que impulsa al cohete. La *ecuación de Tsiolkovski* (5) representa el aumento en la velocidad Δv que puede alcanzar un cohete al quemar todo su combustible. La ecuación relaciona la masa total inicial m_{inicial} y la masa final m_{final} después de consumir el combustible, con las velocidades Δv y v_e.

$$\Delta v = v_e \cdot \ln\left(\frac{m_{\text{inicial}}}{m_{\text{final}}}\right) \tag{5}$$

En esta ecuación aparece la función ln de logaritmo natural. Por ejemplo, si el 80% de la masa total es combustible, entonces:

$$\frac{m_{\text{inicial}}}{m_{\text{final}}}=\frac{100\%}{20\%}=5$$

y como $\ln(5)$ es aproximadamente 1.6, el cohete alcanzará 1.6 veces la velocidad de emisión. La velocidad de expulsión puede alcanzar hasta los 4 400 metros por segundo en motores que usan hidrógeno y oxígeno; con ello, el cohete alcanzará una velocidad de más de 25 mil kilómetros por hora. Para un cálculo realista hay que tomar en cuenta que el cohete sale de la Tierra y usa gran parte del combustible para alejarse de ella, es decir, para elevarse en contra de la fuerza de gravedad. Para ese cálculo necesitaríamos otra ecuación.

Como vemos, la ecuación puede ser muy útil para calcular ciertas cantidades, pero al mismo tiempo tiene sus limitaciones —en este caso sólo sirve para el espacio vacío, así que su uso real siempre será aproximado—. Lo anterior es la típica característica de un modelo: permite hacer un cálculo que, usualmente, es simplificado y, a partir de él, hacer predicciones. Toda la aeronáutica espacial se basa en modelos que tienen que ser muy buenos, dado que los experimentos reales son muy costosos.

3.5.2 Un modelo para la radiactividad

La radiactividad es un proceso espontáneo que se presenta en la naturaleza cuando en el núcleo de un átomo ocurre un cambio. En el núcleo se encuentran partículas como los protones, con carga positiva, y los neutrones, que no tienen carga y que estabilizan a los protones. El número de protones determina las características al reaccionar con otros átomos. Es por lo anterior por lo que los átomos se clasifican en *elementos* como el oxígeno y el carbono que tienen, respectivamente, 8 y 6 protones. El número de neutrones puede variar. Hay átomos de carbono con 6, 7 u 8 neutrones. Esto da un total de 12, 13 y 14, respectivamente, partículas en el núcleo y, por ello, se habla de carbono 12, carbono 13 y carbono 14. Se dice que son diferentes *isótopos* de carbono.

El carbono 14 es radiactivo, es decir, puede decaer en cualquier momento, mientras el carbono 13 y el 12 son *isótopos estables*. En el carbono 14 puede suceder que uno de los neutrones se desintegre en un protón, que quede atrapado en el núcleo, y un electrón y un antineutrino que se emitan hacia el exterior del núcleo. En este cambio, el núcleo se queda con 7 protones y 7 neutrones, por lo que el átomo ahora es de nitrógeno. La cantidad de energía

de las partículas que se emiten durante el decaimiento se puede medir y es lo que le dio su nombre al fenómeno: actividad en el espectro de radio, es decir, se registró la radiactividad por la radiación que emite.

Nadie puede predecir cuando un átomo de carbono 14 se desintegra pues lo hace de manera espontánea, de repente. Lo único que se puede decir es que en 5730 años hay 50% de probabilidad de que suceda y otro tanto de que no. Más sorprendente aún es que, si no decayó en estos 5730 años, tiene 5730 años para otra vuelta con una oportunidad para decaer o no del $50\% - 50\%$. En otras palabras, los átomos de carbono 14, si no decaen, no envejecen.

Si en un momento aislamos una gran cantidad de átomos de carbono 14 y esperamos 5730 años encontraremos que más o menos la mitad de átomos ha decaído, mientras la otra mitad sigue igual. ¿Qué encontraremos otros 5730 años después? La respuesta correcta no es que ahora la otra mitad también decayó y se convirtió en nitrógeno, sino que de los átomos de carbono que sobrevivieron los primeros 5730 años, ahora la mitad también decayó. En conclusión sólo un cuarto de los átomos originales queda sin decaer y, después del tercer periodo de 5730 años, será sólo un octavo.

En nuestro ambiente hay una concentración baja de carbono 14. ¿Cómo es posible si estos isótopos decaen? ¿No tendrían que haber decaído ya todos, después de los millones de años de existencia en la Tierra? Es cierto que continuamente decaen pero, por otro lado, también se generan nuevos átomos de carbono 14 en la atmósfera alta a causa de la radiación solar. De esta manera, la concentración total se mantiene estable en nuestra atmósfera y también en los organismos pues, cada vez que respiran intercambiando gases, incorporan a su estructura estos átomos. Cuando el ser vivo muere, el intercambio cesa y los átomos de carbono 14 que quedaron integrados en su estructura hasta este momento, decaen en los milenios posteriores.

Lo anterior hace posible calcular en un resto orgánico el tiempo que transcurrió desde la muerte del organismo que lo originó al medir la concentración de carbono 14 que resta en el tejido, un método conocido como *fechamiento con carbono 14*. Para ver cómo funciona dicho método, veremos primero cómo depende de la concentración de carbono 14 según el tiempo —medido en años— que transcurre después de la muerte del organismo. Lo que buscamos es una *función f* que modele dicha concentración. De esta función sabemos que $f(0) = 100\%$ y que $f(5730) = 50\%$. Por lo que vimos antes, tenemos que $f(2 \cdot 5730) = 25\%$. Cada 5730 años, la concentración se divide a la mitad, así que $f(3 \cdot 5730) = 12.5\%$. Si escribimos los porcentajes cómo números tenemos:

$$\begin{aligned} f(0 \cdot 5730) &= 1 \\ f(1 \cdot 5730) &= \tfrac{1}{2} \\ f(2 \cdot 5730) &= \tfrac{1}{4} \\ f(3 \cdot 5730) &= \tfrac{1}{8} \\ f(n \cdot 5730) &= \tfrac{1}{2^n} \end{aligned}$$

En la última línea escribimos una expresión general. Si ahora sustituimos $t = n \cdot 5730$ y observamos que $\frac{1}{2^n} = 2^{-n}$, obtenemos la fórmula:

$$f(t) = 2^{-\frac{1}{5730}t}.$$

Si, por ejemplo, en un hueso se encuentra una concentración de 14.2% = 0.142, entonces debemos resolver la ecuación:

$$0.142 = 2^{-\frac{1}{5730}t}$$

en la variable t. La figura 3.35 muestra la gráfica de esta función.

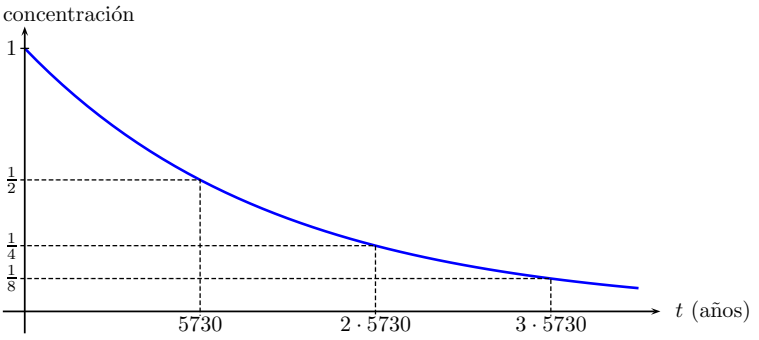


Figura 3.35 La gráfica de la función que modela la concentración de átomos de carbono 14 en el tiempo, después de la muerte del organismo.

Por ello, es conveniente reescribir el lado derecho y usar la base e —el *número de Euler*— en vez de 2, ya que la función $\exp(x) = e^x$ tiene un inverso conocido, que es el *logaritmo natural* $\ln(x)$: si $y = e^x$ entonces $x = \ln(y)$. Véase también el siguiente recuadro.

El número e es definido como el siguiente límite:

$$e = \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$$

La función $\exp(x) = e^x$ tiene la importante propiedad de que es su propia derivada:

$$\frac{\mathrm{d}\exp(x)}{\mathrm{d}x} = \exp(x)$$

y su inverso se llama logaritmo natural y se denota por $\ln(x)$.

Tomamos $y = 2^{\frac{1}{5730}}$ y obtenemos un número:

$$\ln(2^{-\frac{1}{5730}}) = -0.00012097$$

Entonces, si usamos la letra griega "lambda" para denotar a este número —es decir, $\lambda = 0.00012097$—, obtenemos:

$$f(t) = e^{-\lambda t} \tag{6}$$

y la ecuación que debemos resolver es:

$$0.142 = e^{-\lambda t}$$

o, con el logaritmo natural:

$$\ln(0.142) = -\lambda t,$$

es decir:

$$t = \frac{\ln(0.142)}{-\lambda} = \frac{-1.952}{-0.00012097} = 16136$$

Lo cual quiere decir que el organismo del que proviene el hueso murió hace aproximadamente 16 136 años.

La función (6) se llama *función exponencial* y se emplea con frecuencia para modelar fenómenos de la naturaleza. Lo que la función modela en este caso es el comportamiento de la concentración del carbono 14 en el tiempo; pero esta concentración no es otra cosa que la fracción de la cantidad de átomos en el tiempo entre la cantidad inicial.

En la vida real, esta función da brincos, dado que siempre hay un número entero de átomos en cada momento. Nos deberíamos preguntar si realmente es permisible usar una función continua para un fenómeno discreto; más aún, en la realidad la función depende del azar, dado que no se puede predecir cuándo los átomos decaerán. Para entender por qué la función (6) es extremadamente buena para describir el fenómeno real debe tomarse en cuenta que la cantidad de átomos que conforman la concentración es realmente enorme.

De todos los átomos de carbono en nuestro cuerpo sólo uno de un billón —un millón de millones— es un átomo de carbono 14, por lo que en un gramo de carbono hay, aproximadamente, 40 mil millones de átomos de carbono 14. Por ello, es muy improbable que no sea más o menos la mitad que decae en 5 730 años. Si sólo tuviéramos 2 átomos, la probabilidad de que la mitad decaiga es del 50%, mientras hay 25% de probabilidad de que ninguno decaiga y otros 25% de probabilidad de que ambos decaigan. Con 40 mil millones de átomos hay una probabilidad de 99.9953% de que entre 19 999 600 000 y 20 000 400 000 átomos decaigan. Por ello, la función continua aproxima muy bien el fenómeno discreto y probabilístico y el método de fechamiento con carbono 14 se puede emplear con éxito desde su desarrollo, en 1949, a partir de un equipo encabezado por el físico-químico estadunidense Willard Frank Libby.

3.5.3 Modelos para el crecimiento poblacional

En el año 2000 había, por cada mil habitantes de la Tierra, 22 nacimientos y 9 muertos, es decir, un crecimiento de 13 personas por cada mil. ¿Cuánto tiempo transcurrirá hasta que se duplique la población si el crecimiento fuera constante a este ritmo? Es un error pensar que pasarán $\frac{1000}{13} \approx 77$ años. Para entenderlo, simplificamos y digamos que pasa un cierto tiempo T hasta que la población se duplique. ¿En cuánto tiempo se habrá cuadruplicado? El tiempo necesario es sólo el doble de T ya que en cada periodo T se duplicará; por lo tanto, sólo se necesitan dos periodos. El posible error consiste en pensar que en cada año se *suma* la misma cantidad, cuando en realidad en cada año se *multiplica* por el mismo factor.

El factor de crecimiento anual λ se obtiene así: después de un año habrá por cada 1000 unas 13 personas más. Así que:

$$\lambda = \frac{1013}{1000} = 1.013,$$

entonces, $1013 = \lambda \cdot 1000$. Después de dos años la población habrá alcanzado λ^2 veces y, después de tres años, λ^3 veces la población original. En la figura 3.36 se muestra cómo crece λ^n conforme aumenta n.

n	1.014^n
1	1.013
2	1.026
3	1.040
4	1.053
5	1.067
6	1.081
⋮	⋮

n	1.014^n
10	1.138
20	1.295
30	1.473
40	1.676
50	1.908
60	2.171
⋮	⋮

Figura 3.36 Evaluación de la función 1.013^n para diferentes valores de n.

De la tabla se ve que se requieren entre 50 y 60 años para que se duplique la población mundial si sigue creciendo al mismo ritmo que lo hizo durante el año 2000. Este tipo de crecimiento se conoce como *exponencial*. A corto plazo, el modelo es muy exitoso, como se puede observar en la figura 3.37: la estimación no rebasa nunca el 1% de error. El modelo que se usó fue:

$$V(n) = V(0) \cdot \lambda^n \tag{7}$$

donde $\lambda = 1.013$ y $V(n)$ es la población en el año $2000 + n$. Nuevamente, se trata de una función exponencial, pero ahora creciente.

Año	Población mundial	Estimación
2000	6 115 367 000	6 115 367 000
2001	6 194 886 000	6 194 867 000
2002	6 274 302 000	6 275 400 000
2003	6 353 658 000	6 356 980 000
2004	6 432 978 000	6 439 621 000
2005	6 512 276 000	6 523 336 000
2006	6 591 548 000	6 608 139 000
2007	6 670 801 000	6 694 045 000
2008	6 750 062 000	6 781 068 000
2009	6 829 360 000	6 869 222 000
2010	6 908 688 000	6 958 522 000

Figura 3.37 La columna de en medio muestra los datos de las Naciones Unidas, la columna a la derecha exhibe la estimación basada en la hipótesis de que el crecimiento porcentual es constante de año en año, e igual al del año 2000. Los datos fueron redondeados a miles.

El crecimiento exponencial es ilimitado —en el modelo, la población crece y crece, rebasando cualquier límite— y, por lo tanto, inadecuado para predicciones a largo plazo: hay que tomar en cuenta que nuestros recursos son limitados, aun más, el espacio es limitado o, de manera todavía más determinante: la materia en la Tierra es infinita y hay un tiempo para el cual el modelo predice que la población rebasará aun estos límites, lo que es absurdo. Esto se conoce como la *catástrofe de Malthus*, llamada así por el inglés Thomas Robert Malthus, quien publicó seis ediciones de *Ensayo sobre el principio de la población* entre 1798 y 1826.

Por ello, en 1838, el matemático belga Pierre François Verhulst propuso un modelo que tomaba en cuenta lo limitado de los recursos y consideraba que el mundo puede sostener un máximo número de personas K —aunque este máximo sea desconocido—. Para explicar este modelo, reformulamos primero el crecimiento exponencial (7) en términos de ecuaciones diferenciales:

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = r \cdot V(t) \tag{8}$$

donde $r = \ln(\lambda)\frac{1}{\text{año}}$; esta reformulación se justifica en el siguiente recuadro.

El número de personas es una función en el tiempo $V(t)$ y, en un tiempo Δt, la población crecerá ΔV. Por ejemplo, si Δt es un año, entonces $\Delta V = 0.013 \cdot V(0)$, si tomamos el crecimiento durante el año 2000 como indicador. Es decir, el crecimiento es proporcional a la población con un factor de 0.013, para lapsos de un año. Por ello, tenemos:

$$V(t + \Delta t) - V(t) = \Delta V = (\lambda - 1)V(t)$$

donde $\lambda = 1.013$. Si ahora tomamos medio año $\frac{\Delta t}{2}$, entonces tendremos un crecimiento de:

$$V(t + \tfrac{\Delta t}{2}) - V(t) = (\sqrt{\lambda} - 1)V(t).$$

Más general, para la n-ésima parte de un año $\frac{\Delta t}{n}$ se tiene que:

$$V(t + \tfrac{\Delta t}{n}) - V(t) = (\sqrt[n]{\lambda} - 1)V(t).$$

Obviamente, si crecemos el valor de n, este incremento disminuye y se acerca a 1. Pero si consideramos el cociente:

$$\frac{V(t + \frac{\Delta t}{n}) - V(t)}{\frac{\Delta t}{n}} = n(\sqrt[n]{\lambda} - 1)\frac{V(t)}{\Delta t} \tag{9}$$

pasa algo interesante: del lado izquierdo el numerador y el denominador se acercan simultáneamente al valor 0.

Al tomar el límite para $n \to \infty$, obtenemos en (9) la derivada $\frac{\mathrm{d}V(t)}{\mathrm{d}t}$, mientras del lado derecho, el cociente $\frac{N(t)}{\Delta t}$ no depende de n. En tanto, el límite se calcula de la siguiente manera:

$$\lim_{n\to\infty} n(\sqrt[n]{\lambda} - 1) = \ln(\lambda) = 0.0129,$$

como se explica en seguida. Para calcular $\lim_{n\to\infty} n(\sqrt[n]{\lambda} - 1)$, se hace primero un cambio de variable $n = \frac{1}{m}$ y se observa que $\sqrt[n]{\lambda} = \lambda^{\frac{1}{n}} = \lambda^m$:

$$\lim_{n\to\infty} n(\sqrt[n]{\lambda} - 1) = \lim_{m\to 0} \frac{\lambda^m - 1}{m}.$$

Esta última fracción se puede interpretar como la derivada de la función λ^x en el punto $x = 0$:

$$\lim_{m\to 0} \frac{\lambda^m - 1}{m} = \frac{\mathrm{d}\lambda^x}{\mathrm{d}x}\bigg|_{x=0} = \frac{\mathrm{d}e^{\ln(\lambda)\cdot x}}{\mathrm{d}x}\bigg|_{x=0}$$

que, según la regla de derivación de la composición de funciones, se calcula en:

$$\lim_{m\to 0} \frac{\lambda^m - 1}{m} = \ln(\lambda) \cdot e^{\ln(\lambda)\cdot x}\Big|_{x=0} = \ln(\lambda).$$

Si ponemos $r = \ln(\lambda)\frac{1}{\text{año}}$, entonces obtenemos finalmente la siguiente reformulación (9) de la ecuación (8), como se había anunciado.

La modificación de (8) que propone Verhulst es introducir un término correctivo negativo $\frac{r}{K} \cdot V(t)^2$:

$$\frac{\mathrm{d}V(t)}{\mathrm{d}t} = r \cdot V(t) - \frac{r}{K} \cdot V(t)^2. \tag{10}$$

Este término de corrección depende del cuadrado de $V(t)$ y esto tiene una buena razón: el número de encuentros en una población depende aproximadamente del cuadrado del número de integrantes, si se supone que éstos se mueven por el espacio disponible. Con los contactos, se aumenta la propagación de enfermedades, lo que disminuye la población —por ello, el término aparece con signo negativo.

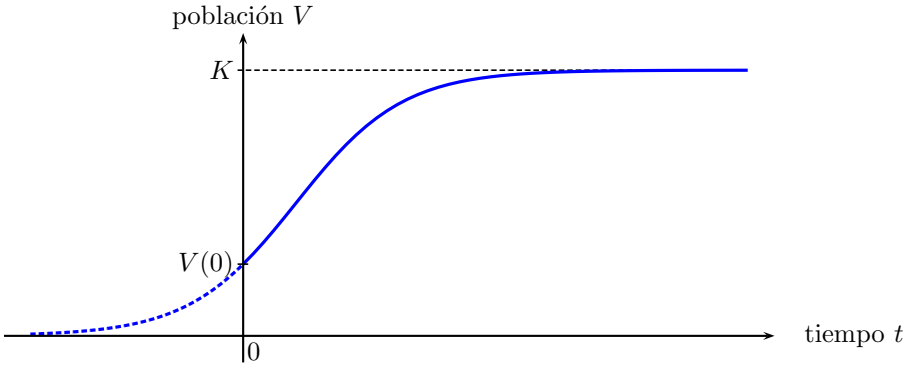


Figura 3.38 El modelo de Verhulst muestra el desarrollo de la población $V(t)$ en el tiempo t de manera cualitativa, es decir, sin especificar valores concretos para r y K, ya que éstos se desconocen para la población humana.

El término de corrección tiene el efecto deseado sobre el modelo: el desarrollo en el tiempo deja que $V(t)$ se acerque a K pero sin sobrepasarlo, como se observa en la figura 3.38 y justo como se deseaba.

3.5.4 Un modelo para la propagación de un virus

Existen muchos modelos matemáticos usados para la prevención de la salud y la estimación de cuán contagiosa es una determinada enfermedad infecciosa. Por ejemplo, se puede modelar cómo un virus —como el de la influenza— se disemina en una comunidad. En este caso se toma en cuenta que los enfermos entran en contacto con personas sanas y las contagian con el agente infeccioso. Veamos el planteamiento del modelo en forma sencilla: sea $I(t)$ el número de personas que se han contagiado con la enfermedad como función del tiempo t y $S(t)$ el número de personas que aún no se han enfermado. Finalmente, también habrá que considerar el número de personas $R(t)$ que ya no se infectarán ni podrán infectar a otros: son aquellas personas que se han recuperado de la enfermedad o que han muerto a causa de ella.

Parece razonable suponer que la rapidez $\frac{\mathrm{d}S(t)}{\mathrm{d}t}$ a la que las personas sanas se enferman es proporcional al número de encuentros o interacciones entre estos dos grupos de personas. También es razonable suponer que el número de interacciones es proporcional al número de enfermos $I(t)$ y al número de personas sanas $S(t)$, es decir, proporcional al producto $I(t) \cdot S(t)$, entonces:

$$\frac{\mathrm{d}S(t)}{\mathrm{d}t} = -k \cdot I(t)S(t), \tag{11}$$

donde k es la constante de proporcionalidad. El signo negativo indica que el número de personas sanas disminuye. Hay dos factores que influyen en el cambio del número de infec-

tados: por un lado, aumenta por los nuevos infectados y, por otro, disminuye por los que se curaron o murieron. En el modelo se supone que el número de personas que deja de ser infeccioso es proporcional al número de infectados. La siguiente ecuación da cuenta de ello:

$$\frac{\mathrm{d}I(t)}{\mathrm{d}t} = k \cdot I(t)S(t) - \ell \cdot I(t), \tag{12}$$

Nuevamente, debe observarse los signos de los dos sumandos. Para finalizar, la tercera ecuación explicará el cambio en el número de personas que dejaron de ser infecciosas.

$$\frac{\mathrm{d}R(t)}{\mathrm{d}t} = \ell \cdot I(t), \tag{13}$$

Se observa que contraer la enfermedad da inmunidad a la persona, es decir, la persona ya no se puede infectar nuevamente. Las tres ecuaciones (11), (12) y (13) forman el modelo que discutimos aquí. Se trata de un modelo continuo para un fenómeno discreto: los números considerados siempre serán enteros ya que no hay fracciones de personas.

Sólo la práctica puede darnos indicaciones sobre los *parámetros* k y ℓ para una población dada y un virus determinado. La correcta estimación de estos parámetros puede conducir a una predicción sobre el porcentaje de población que debe vacunarse para impedir una epidemia.

Las campañas de vacunación masiva son proyectos de la Organización Mundial de la Salud (OMS) y de muchos departamentos de salud pública. Una característica importante de las enfermedades infecciosas es la reproducción básica del factor R0, que permite estimar el número de infecciones secundarias que un individuo infectado producirá durante el tiempo en que está enfermo. Para casi cualquier enfermedad infecciosa, los modelos de ecuaciones diferenciales pueden llevar al cálculo del R0 básico al usar información específica —tal como la tasa de transmisión del patógeno, la duración del periodo infeccioso y la tasa promedio de muerte en la población—. El R0 proporciona un estimado de la situación de la infección, por ejemplo, de cómo se puede diseminar en una población sana: si el R0 es menor a uno, la enfermedad no persistirá en la población, pero si es mayor a uno, entonces la enfermedad posee el potencial para diseminarse en forma de epidemia o de volverse endémica. Un caso interesante para ejemplificar lo anterior es el de la viruela, actualmente erradicada del planeta, a pesar de que tenía un R0 de 3 a 5 y requirió vacunar alrededor del 70 u 80% de la población —el último caso de viruela se reportó en 1977—. La OMS está tratando de erradicar también la polio, que tiene un R0 de 5 a 7, lo cual significa que, aproximadamente, se requiere inmunizar entre 80 y 86% de la población mundial para lograr desaparecer la infección.

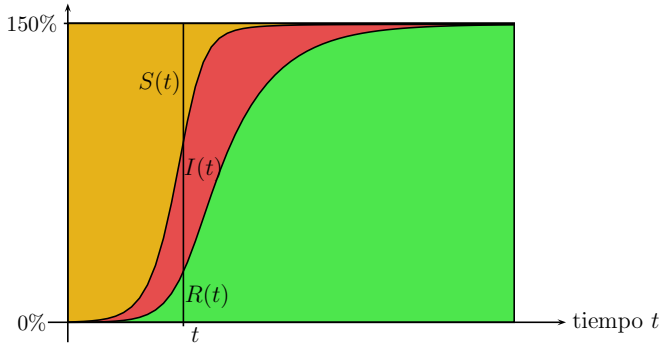


Figura 3.39 El desarrollo cualitativo de las tres cantidades, "sanos" $S(t)$, "infectados" $I(t)$ y "recuperados" $R(t)$, en el modelo de propagación de virus. En cada momento t, la suma de los tres números debe ser el total de la población, es decir, igual a 100 por ciento.

Los modelos permiten estudiar y comprender cómo es que ocurren los fenómenos que no podemos observar fácilmente o manipular directamente. Por ello, son de fundamental importancia para las ciencias. El mayor éxito entre los modelos han sido las ecuaciones diferenciales, de las cuales aquí dimos algunas muestras.

3.6 EL CAMPO Y LOS VECTORES

Figura 3.40 La atracción gravitatoria establece una relación cuantitativa entre dos objetos con masa, como el Sol y la Tierra.

¿Cómo actúa la fuerza gravitatoria? ¿Cómo es que el Sol ejerce una fuerza sobre un planeta que se encuentra a miles de kilómetros de distancia? Resulta aún más sorprendente este hecho si consideramos que esa fuerza ejercida a gran distancia, a juzgar por sus efectos en el movimiento de los astros, parece tomar valores perfectamente definidos con gran precisión a cada momento. No sabemos cómo responder a esas preguntas, pero hay una idea de los físicos que resulta muy útil para tranquilizar nuestras angustias filosóficas. La idea es que el Sol y todo cuerpo con masa en realidad no están localizados en la pequeña zona en la que está comprendida su parte material, su masa, sino que hay algo no material que es también parte del Sol: el campo de fuerza gravitatoria, y esa parte está distribuida en todo el Universo. ¿Qué es este campo? En lugar de decir que es sólo una abstracción matemática, adoptamos la postura más científica que consiste en decir que tiene una realidad física perfectamente clara y bien definida; lo único que no posee, y que estamos acostumbrados a pensar que sí poseen los entes físicos, es masa. Decimos que el campo gravitatorio existe pues se manifiesta físicamente ejerciendo fuerza sobre los objetos con masa.

El estudio de la electricidad impulsó aún más la idea del campo, en este caso, el llamado campo eléctrico. A continuación se describe la notación matemática de vectores y campos vectoriales que se generaron alrededor de la electrostática y resultaron útiles para profundizar en el estudio del electromagnetismo, la mecánica de los fluidos y muchos otros temas de la física.

3.6.1 La electrostática y el concepto de campo vectorial

El campo eléctrico es una función vectorial $\mathbf{E}(x, y, z)$ que se define como la fuerza que actuaría sobre una carga eléctrica unitaria colocada en el punto (x, y, z). Se dice que una función es vectorial si a cada punto de su dominio se le asocia un vector con dirección, magnitud y sentido. El campo vectorial $\vec{E}$ se puede representar mediante tres componentes escalares o numéricas E_x, E_y, E_z, una a lo largo de cada eje de coordenadas. En símbolos, se puede representar al campo eléctrico así:

$$\vec{E} = E_x\ \vec{i} + E_y\ \vec{j} + E_z\ \vec{k}$$

o más explícitamente, así:

$$\vec{E}(x, y, z) = E_x(x, y, z)\ \vec{i} + E_y(x, y, z)\ \vec{j} + E_z(x, y, z)\ \vec{k}$$

donde $\vec{i}, \vec{j}, \vec{k}$ son vectores unitarios —de magnitud 1— en las direcciones de los ejes x, y, z, respectivamente. El lector debe observar que los vectores se escriben en los textos impresos con letra **negrita**. En los manuscritos los vectores se indican poniendo una raya o una flecha encima del nombre del vector.

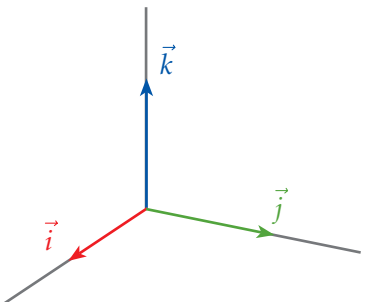


Figura 3.41 Los vectores unitarios $\vec{i}$, $\vec{j}$, $\vec{k}$.

El comportamiento del campo eléctrico está definido por la ley de Coulomb y la ley de superposición. La ley de Coulomb dice que el campo eléctrico producido por una carga eléctrica Q en un punto del espacio es proporcional a la magnitud de la carga e inversamente proporcional al cuadrado de la distancia, y su dirección coincide con la del vector que va de la carga al punto. El sentido del campo es hacia la carga si ésta es negativa y en sentido opuesto si es positiva. Suponiendo que la carga eléctrica Q se encuentra en el origen, que $\vec{r} = x\vec{i} + y\vec{i} + z\vec{k}$ es el vector de posición del punto (x, y, z), que $r = |\vec{r}\,|$ es la distancia del punto (x, y, z) al origen y que $\vec{u}$ es el vector unitario que apunta en la dirección del punto (x, y, z) en sentido opuesto al origen, la ley de Coulomb puede expresarse así:

$$\vec{E} = \frac{Q}{r^2}\vec{u} = \frac{Q}{r^3}\vec{r} = \frac{Q}{r^3}(x\vec{i} + y\vec{j} + z\vec{k})$$

donde la segunda igualdad se justifica porque:

$$\vec{u} = \frac{\vec{r}}{r}.$$

La ley de superposición dice que el campo eléctrico producido por dos o más cargas eléctricas es la suma de los producidos por cada una de ellas. Por ejemplo, el campo eléctrico correspondiente a una carga positiva Q colocada en el punto $(-a, 0)$ y una negativa $-Q$ en el punto $(a, 0)$ tiene esta expresión:

$$E = \frac{Q}{{r_1}^3}\left((x+a)\vec{i} + y\vec{j} + z\vec{k}\right) - \frac{Q}{{r_2}^3}\left((x-a)\vec{i} + y\vec{j} + z\vec{k}\right)$$

donde $r_1 = \sqrt{(x+a)^2 + y^2 + z^2}$ y $r_2 = \sqrt{(x-a)^2 + y^2 + z^2}$.

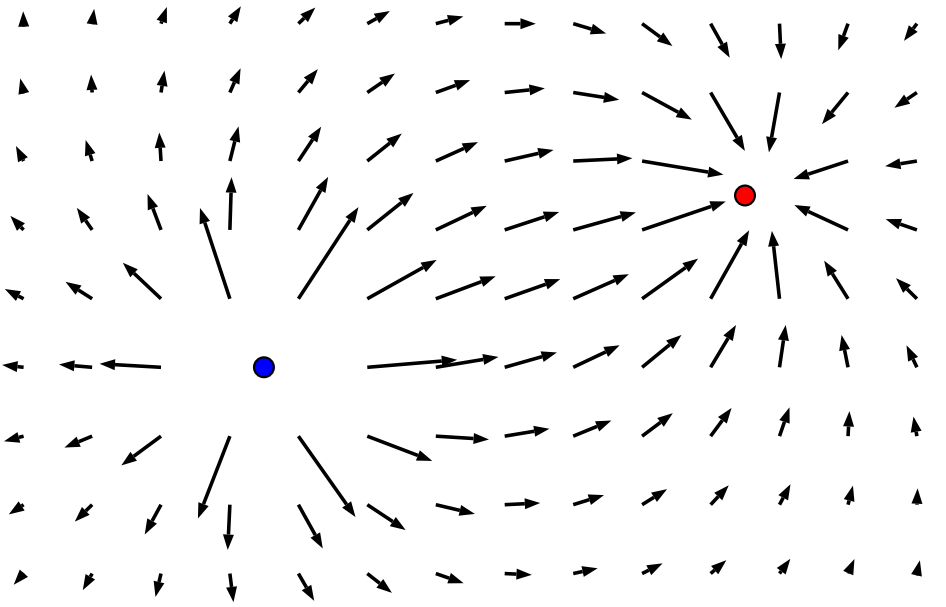


Figura 3.42 Ilustración del campo eléctrico debido a una carga eléctrica positiva (azul) y una negativa (roja).

El estudio de la electrostática es prácticamente imposible sin el uso de estas nuevas herramientas matemáticas que hemos venido introduciendo a lo largo de los párrafos anteriores y que constituyen lo que se llama vectores y campos vectoriales. A continuación, se presenta un resumen de los principales conceptos del álgebra de vectores y el cálculo vectorial. Se trata de una presentación muy escueta cuyo único objetivo es ofrecer al lector una primera idea de la potencia expresiva de estas herramientas matemáticas que, como se verá en la siguiente sección, llevaron a grandes descubrimientos en relación con el electromagnetismo y las ondas electromagnéticas, esenciales para la tecnología de las comunicaciones modernas.

3.6.2 Álgebra vectorial

Dados dos vectores $\vec{a} = a_1\vec{i} + a_2\vec{j} + a_3\vec{k}$ y $\vec{b} = b_1\vec{i} + b_2\vec{j} + b_3\vec{k}$ y un escalar —o sea, un número— c se definen cuatro operaciones algebraicas:

1] Suma de dos vectores: $\vec{a} + \vec{b} = (a_1 + b_1)\vec{i} + (a_2 + b_2)\vec{j} + (a_3 + b_3)\vec{k}$.
2] Multiplicación de un vector por un escalar: $c\vec{a} = ca_1\vec{i} + ca_2\vec{j} + ca_3\vec{k}$.
3] Producto escalar de dos vectores (también llamado producto punto):
$\vec{a} \cdot \vec{b} = a_1b_1 + a_2b_2 + a_3b_3$.
4] Producto vectorial (o producto cruz) de dos vectores:
$\vec{a} \times \vec{b} = (a_2b_3 - a_3b_2)\vec{i} + (a_3b_1 - a_1b_3)\vec{j} + (a_1b_2 - a_2b_1)\vec{k}$.

Las primeras dos operaciones corresponden a lo que se llama álgebra lineal y las dos segundas a lo que propiamente es el álgebra vectorial. Los productos entre vectores son de gran utilidad por sus propiedades geométricas que pueden demostrarse como teoremas a partir de la definición de dichas operaciones, que describimos a continuación.

El producto escalar o producto punto $\vec{a} \cdot \vec{b}$ es un número —un escalar— igual al producto de la magnitud de cualquiera de ellos por la proyección del otro sobre el primero o equivalentemente: $\vec{a} \cdot \vec{b} = |a||b| \cos(\theta)$, donde θ es el ángulo que forman los dos vectores.

El producto vectorial de dos vectores es un vector que apunta en la dirección perpendicular a ambos; el sentido está dado por lo que se llama la regla de la mano derecha, que dice: el producto vectorial $\vec{a} \times \vec{b}$ apunta hacia donde lo hace el pulgar de la mano derecha cuando el dedo índice se alinea con $\vec{a}$ y el dedo medio con $\vec{b}$. Finalmente, la magnitud de $\vec{a} \times \vec{b}$ es igual al producto de las magnitudes de $\vec{a}$ y $\vec{b}$ por el seno del ángulo entre ellos:

$$\left|\vec{a} \times \vec{b}\right| = |a||b|sen(\theta)$$

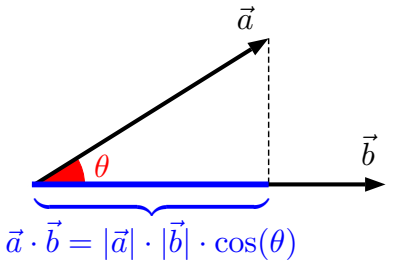


Figura 3.43 Producto escalar.

3.6.3 Cálculo vectorial

Los conceptos de derivada e integral que se desarrollaron para funciones de una variable, se generalizan de distintas maneras a funciones de varias variables y a campos vectoriales. Aquí nos interesan las generalizaciones de esos conceptos a funciones escalares y vectoriales de tres variables, por ser éstas de gran utilidad en el estudio de los fenómenos físicos. El teorema fundamental del cálculo tiene versiones vectoriales, muy interesantes en sí mismas y muy útiles en la física.

Estudiaremos tres conceptos que corresponden vagamente al de derivada: el **gradiente**, la **divergencia** y el **rotacional**. El gradiente se aplica a campos escalares, es decir, a funciones escalares $f(x, y, z)$ definidas en alguna región del espacio de tres dimensiones. Un ejemplo de función escalar es la temperatura $T(x, y, z)$. El concepto de gradiente corresponde al vector que apunta en la dirección y sentido del máximo crecimiento de la función y cuya magnitud es igual a la razón de cambio del valor de la función en esa dirección. Un teorema del cálculo vectorial nos dice que el gradiente $\text{grad}(f)$, de una función $f(x, y, z)$, que también se denota como ∇f, puede calcularse fácilmente usando las derivadas parciales de:

$$f : \text{grad}(f) = \nabla f = \frac{\partial f}{\partial x}\,\vec{i} + \frac{\partial f}{\partial x}\,\vec{j} + \frac{\partial f}{\partial x}\,\vec{k}$$

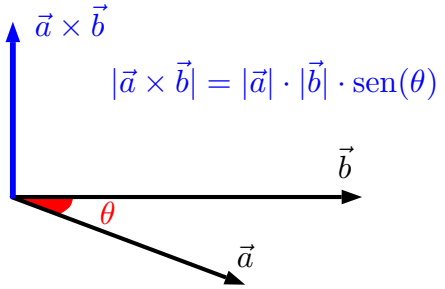


Figura 3.44 Producto vectorial.

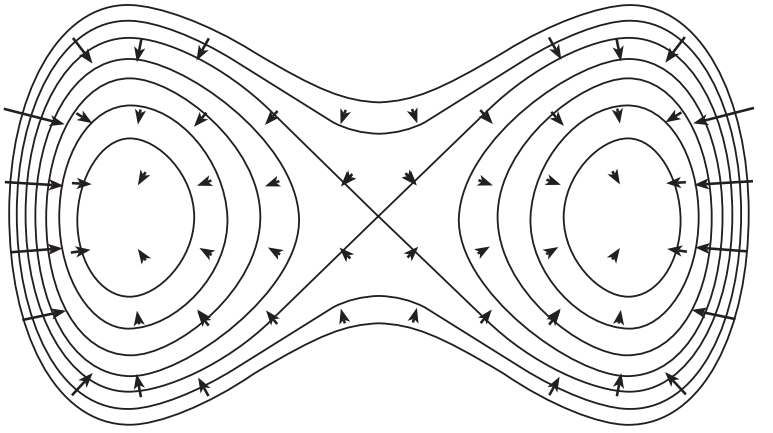


Figura 3. 45 El gradiente de un campo escalar mide la dirección del mayor ascenso: a cada punto le asocia un vector que apunta en la dirección en la cual el campo escalar aumenta en mayor grado. La imagen muestra una función escalar con líneas de nivel: cada línea marca un valor constante de la función. Las flechas indican el gradiente en varios lugares. El gradiente asocia a un campo escalar un campo vectorial.

Para definir los conceptos de divergencia y rotacional es conveniente definir antes dos conceptos de integración para campos vectoriales: la integral de línea a lo largo de una curva y el flujo o integral de superficie. Definiremos estos conceptos utilizando la idea intuitiva de que la integral es una suma infinita de elementos infinitamente pequeños. Lo importante es entender qué representan los elementos que se suman y, por tanto, lo que representa cada una de esas integrales. No nos va a preocupar en este libro la formalización matemáticamente precisa de los conceptos —que por supuesto se hace usando límites—, basta decir que tal formalización puede lograrse y se puede encontrar en casi todos los libros especializados en cálculo —o análisis— vectorial.

La integral de línea del campo vectorial $\vec{F}$ a lo largo de una curva C se denota por:

$$\int_C \vec{F} \cdot \mathrm{d}\vec{r}$$

y representa la suma infinita de elementos infinitamente pequeños $\vec{F} \cdot \mathrm{d}\vec{r}$ a lo largo de la curva. Aquí, $\mathrm{d}\vec{r}$ representa el vector de desplazamiento infinitesimal a lo largo de la curva —en realidad, tangente a ella—. Si pensamos que $\vec{F}$ es una fuerza, cada elemento $\vec{F} \cdot \mathrm{d}\vec{r} = |\vec{F}| \cdot |\mathrm{d}\vec{r}| \cos(\theta)$ representa el trabajo realizado por esa fuerza $\vec{F}$ a lo largo del elemento $\mathrm{d}\vec{r}$ de la misma —recuerda: *trabajo = fuerza* × *distancia*; la multiplicación por $\cos(\theta)$ corresponde a tomar la proyección de $\vec{F}$ a lo largo de la tangente a la curva—. Por tanto, la integral de línea puede interpretarse como el trabajo realizado por $\vec{F}$ sobre una partícula que se mueve a lo largo de la curva C. Cuando C es una curva cerrada, es decir, un ciclo, entonces la integral de línea se escribe así:

$$\oint_C \vec{F} \cdot \mathrm{d}\vec{r}$$

y se llama la circulación de $\vec{F}$ a lo largo de la curva cerrada C.

El flujo de un campo vectorial $\vec{F}$ a través de una superficie S se define como la integral de superficie y se denota como:

$$\iint_S \vec{F} \cdot \mathrm{d}\vec{S},$$

que es la suma de una infinidad de elementos infinitesimales $\vec{F} \cdot \mathrm{d}\vec{S}$ cubriendo toda la superficie S. Los elementos $\vec{F} \cdot \mathrm{d}\vec{S}$ son el producto escalar del vector $\vec{F}$ por el vector $\mathrm{d}\vec{S}$ que se define como perpendicular a la superficie y con magnitud igual a la del área que cubre. Las superficies se consideran orientadas y la orientación de una superficie es, precisamente, el sentido del vector perpendicular a ella. La mejor manera de interpretar el flujo de un

campo vectorial a través de una superficie es pensando que el campo vectorial corresponde a la velocidad de un fluido. Entonces el flujo es la cantidad de fluido que atraviesa la superficie por unidad de tiempo.

El concepto de divergencia $\mathrm{div}(\vec{F})$ o $\vec{\nabla}\cdot\vec{F}$ se aplica a un campo vectorial $\vec{F} = F_x\,\vec{i} + F_y\,\vec{j} + F_z\,\vec{k}$. La divergencia se define como el siguiente límite, si existe:

$$\mathrm{div}(\vec{F}) = \vec{\nabla}\cdot\vec{F} = \lim_{V\to 0}\frac{1}{V}\iint_S \vec{F}\cdot \mathrm{d}\vec{S}$$

donde V representa un volumen encerrado por una superficie S. El límite se toma haciendo el volumen cada vez más pequeño, colapsándolo al punto para el que se está calculando la divergencia. Si $\vec{F}$ representa el campo gravitatorio, resulta que la divergencia es proporcional a densidad de masa ρ:

$$\vec{\nabla}\cdot\vec{F} = \rho$$

En particular, en el vacío no hay masa y, por lo tanto, la divergencia del campo gravitatorio en el vacío es cero. Algo similar ocurre para el campo eléctrico:

$$\vec{\nabla}\cdot\vec{E} = \rho$$

donde ρ representa en este caso la densidad de carga eléctrica. No es difícil demostrar que la divergencia de un campo vectorial puede calcularse con las derivadas parciales de sus componentes, específicamente:

$$\mathrm{div}(\vec{F}) = \vec{\nabla}\cdot\vec{F} = \frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}$$

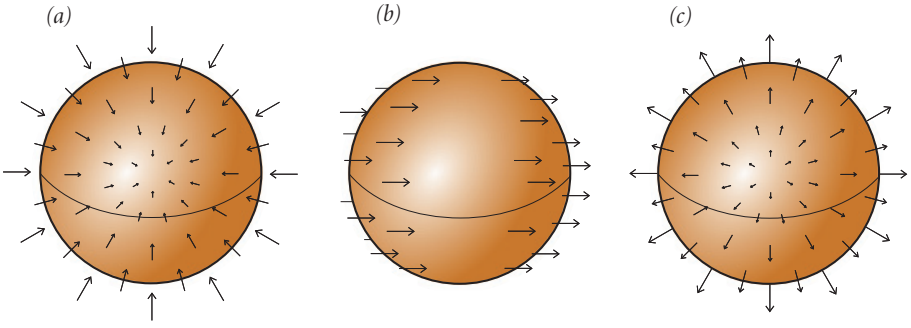


Figura 3. 46 La divergencia de un campo vectorial es una medida para medir flujos que salen de un lugar. En los tres ejemplos, en (a) se tiene una divergencia positiva, en (b) es cero y en (c) es negativa. La medida es local: se puede medir en cada punto al calcular el flujo que sale de una pequeña esfera que rodea este punto y luego se reduce el radio de esta esfera hacia cero en un proceso límite. Así la divergencia convierte un campo vectorial en un campo escalar.

El concepto de rotacional $\mathrm{rot}(\vec{F}) = \nabla\times\vec{F}$ también se aplica a un campo vectorial $\vec{F} = F_x\vec{i} + F_y\vec{j} + F_z\vec{k}$. Se define como otro campo vectorial que tiene la siguiente propiedad para cualquier vector unitario $\vec{n}$:

$$\mathrm{rot}(\vec{F})\cdot\vec{n} = (\vec{\nabla}\times\vec{F})\cdot\vec{n} = \lim_{C\to 0}\frac{1}{L(C)}\oint_C \vec{F}\cdot \mathrm{d}\vec{r}$$

donde C es una curva cerrada sobre un plano perpendicular a $\vec{n}$ que en el límite se encoge o se colapsa a un punto. Se puede demostrar que el rotacional de un campo vectorial puede calcularse con las derivadas parciales de sus componentes, específicamente así:

$$\mathrm{rot}(\vec{F}) = \vec{\nabla}\times\vec{F} = \left(\frac{\partial F_y}{\partial z} - \frac{\partial F_z}{\partial y}\right)\vec{i} + \left(\frac{\partial F_z}{\partial x} - \frac{\partial F_x}{\partial z}\right)\vec{j} + \left(\frac{\partial F_x}{\partial y} - \frac{\partial F_y}{\partial z}\right)\vec{k}$$

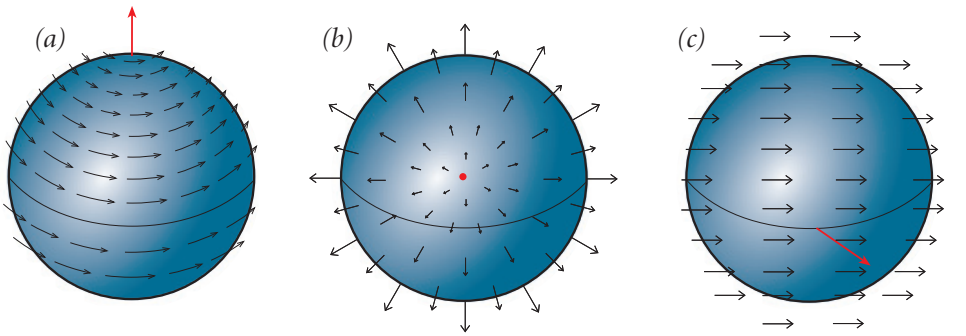


Figura 3.47 La rotacional de un campo vectorial mide el giro alrededor de un punto. En el ejemplo (a) la rotación apunta hacia arriba, en (b) es cero y en (c) apunta hacia adelante. Si se imagina el campo vectorial como la velocidad de un líquido o un gas y se coloca una pelota pequeña con el centro fijo dentro de este líquido, entonces la rotacional apunta en la dirección del eje en el cual gira la pelota. La rotacional asigna a un campo vectorial otro campo vectorial.

Hay tres famosos e importantes teoremas del cálculo vectorial en tres dimensiones que enunciamos a continuación.

a] Teorema de Gauss:

$$\iiint_V \vec{\nabla} \cdot \vec{F} = \iint_S \vec{F} \cdot \mathrm{d}\vec{S}$$

donde V es un volumen encerrado por la superficie S, lo cual se denota por $S = \partial V$.

b] Teorema de Stokes:

$$\iint_S \vec{\nabla} \times \vec{F}\ \mathrm{d}S = \oint_C \vec{F} \cdot \mathrm{d}\vec{r}$$

donde S es una superficie cuya frontera es la curva C, lo cual se denota por $C = \partial S$.

c] Teorema del potencial: Si $\vec{\nabla} \times \vec{F} = 0$ (o equivalentemente, si $\oint_C \vec{F} \cdot \mathrm{d}\vec{r} = 0$ para toda curva cerrada C) entonces existe un campo escalar $f(x, y, z)$, llamado el potencial de $\vec{F}$, tal que:

$$\vec{F} = \vec{\nabla} f = \mathrm{grad}(f)$$

y viceversa, si un campo vectorial es el gradiente de un potencial, entonces su rotacional es cero y las integrales de línea sobre curvas cerradas son cero.

Los dos primeros teoremas son una consecuencia más o menos directa de las definiciones de divergencia y el rotacional, respectivamente. El último es un resultado profundo que en particular nos dice que el campo gravitatorio y el campo electrostático son gradientes de un potencial. La función potencial de estos campos resulta muy útil en los cálculos. El campo magnético, en cambio, no tiene esta propiedad. El campo magnético no es el gradiente de un potencial, su rotacional no es cero y hay curvas a lo largo de las cuales la circulación del campo magnético es distinta de cero.

La siguiente sección explica las interesantes leyes del electromagnetismo, su importancia en la ciencia moderna y cómo las matemáticas jugaron un papel fundamental en el descubrimiento de las ondas electromagnéticas, que son la base de las comunicaciones modernas.

3.6.4 El campo electromagnético

Entender el funcionamiento planetario a través de un modelo matemático que asignó números a ciertas entidades físicas —para luego relacionarlos en fórmulas precisas y claras—, fue un gran logro del intelecto humano. Este hallazgo, la teoría de la gravitación universal de Newton, hizo pensar a los científicos, durante dos siglos, que habían encontrado uno de los secretos más íntimos de la naturaleza. A lo largo del siglo XIX se estudiaron la electricidad y

el magnetismo, y se halló entre ambos fenómenos una correspondencia que también puede describirse perfectamente mediante fórmulas matemáticas que relacionan los campos eléctricos y magnéticos. El descubrimiento de Ørsted —las corrientes eléctricas generan campos magnéticos— y el de Faraday —sobre la creación de corrientes eléctricas con el movimiento de imanes— dieron origen a toda una nueva rama de la física que, a la vez, reforzó la idea de que el Universo se rige por leyes matemáticas.

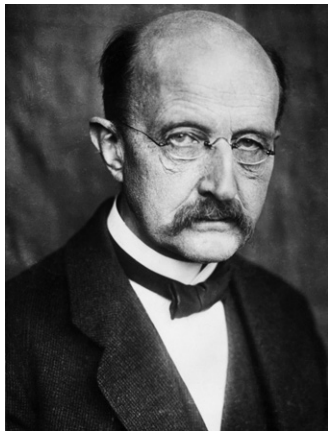


Figura 3.48 Max Karl Ernst Ludwig Planck (1858-1947), galardonado con el Premio Nobel de Física en 1918, fue un físico alemán considerado el fundador de la teoría cuántica | © Latin Stock México.

Uno de los mayores logros científicos que se deben directamente a las matemáticas fue el descubrimiento de las ondas electromagnéticas. Al poner en limpio las leyes del electromagnetismo y tomar en cuenta la tendencia de la naturaleza a exhibir simetrías, Maxwell llegó a un sistema de ecuaciones que describe perfectamente los fenómenos electromagnéticos conocidos hasta entonces. Obtuvo una representación matemática sumamente elegante y compacta del electromagnetismo: las ecuaciones de Maxwell; a la vez, descubrió la inducción mutua de los campos eléctrico y magnéticos, es decir, el hecho de que las variaciones en el tiempo del campo magnético producen uno eléctrico, y, a la vez, las variaciones en el tiempo del campo eléctrico producen otro magnético. Lo anterior se traduce matemáticamente en la ecuación de onda, conocida con anterioridad. Las soluciones de esta ecuación son ondas, es decir, perturbaciones en un medio que viajan por el espacio. Las fórmulas matemáticas del electromagnetismo señalaron que si de verdad este fenómeno podía describirse con esas ecuaciones, entonces deberían existir ondas electromagnéticas. Más aún, las ecuaciones de Maxwell definían con absoluta precisión la velocidad de transmisión de dichas ondas, aproximadamente $300\,000\,\frac{\text{km}}{\text{s}}$. Hoy sabemos que ésta es la velocidad de la luz.

$$\begin{aligned}
\vec{\nabla}\cdot\vec{E} &= \frac{\rho}{\epsilon_0}\\
\vec{\nabla}\cdot\vec{B} &= 0\\
\vec{\nabla}\times\vec{E} &= -\frac{\partial\vec{B}}{\partial t}\\
\vec{\nabla}\times\vec{B} &= \mu_0\vec{J}+\mu_0\epsilon_0\frac{\partial\vec{E}}{\partial t}
\end{aligned}$$

Figura 3.49 Las ecuaciones de Maxwell en forma diferencial.

$$\begin{aligned}
\vec{\nabla}\cdot\vec{E} &= 0\\
\vec{\nabla}\cdot\vec{B} &= 0\\
\vec{\nabla}\times\vec{E} &= -\frac{\partial\vec{B}}{\partial t}\\
\vec{\nabla}\times\vec{B} &= \mu_0\epsilon_0\frac{\partial\vec{E}}{\partial t}
\end{aligned}$$

Figura 3.50 Las ecuaciones de Maxwell en el vacío, que exhiben la simetría de la naturaleza con mayor claridad que las otras ecuaciones.

$$\frac{1}{c^2}\frac{\partial^2\vec{E}}{\partial t^2}=\frac{\partial^2\vec{E}}{\partial x^2}+\frac{\partial^2\vec{E}}{\partial y^2}+\frac{\partial^2\vec{E}}{\partial z^2}$$

$$\frac{1}{c^2}\frac{\partial^2\vec{B}}{\partial t^2}=\frac{\partial^2\vec{B}}{\partial x^2}+\frac{\partial^2\vec{B}}{\partial y^2}+\frac{\partial^2\vec{B}}{\partial z^2}$$

Figura 3.51 Ecuaciones de onda para los campos eléctrico y magnético, que se obtienen manipulando las ecuaciones de Maxwell en el vacío. Sus soluciones son ondas que se propagan con velocidad $c=\frac{1}{\sqrt{\mu_0\epsilon_0}}$.

Motivado por el trabajo de Maxwell, Hertz intentó descubrir si tales ondas de verdad existían y logró producirlas en un punto y registrarlas en otro, es decir, logró transmitir di-

chas perturbaciones de un lugar a otro. Esas ondas se llamaron por un tiempo ondas hertzianas y muy pronto quedó perfectamente claro que correspondían a las que Maxwell había previsto con sus ecuaciones. La primera aplicación de las ondas electromagnéticas fue la radio, y ésta fue el paso inicial de la revolución tecnológica de las comunicaciones, basada en ese gran descubrimiento al que se llegó con unas fórmulas matemáticas.

No es pues sorprendente que parte de la humanidad esté enamorada de la idea de que al mundo lo rigen un conjunto de fórmulas matemáticas, lo cual ha funcionado muy bien en muchos ámbitos de la ciencia, sobre todo en las diversas ramas de la física. En innumerables ocasiones, la creación de un modelo matemático para representar un fenómeno de la naturaleza ha llevado a importantes descubrimientos, más allá de lo estrictamente vinculado con el fenómeno en cuestión. La fórmula de la atracción gravitatoria ayudó no únicamente a comprender cómo se mueven los planetas, sino también a descubrir cuerpos celestes desconocidos hasta entonces.

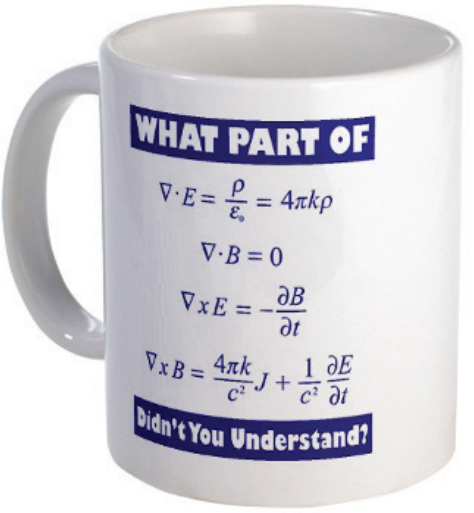


Figura 3.52 Las ecuaciones de Maxwell son tan elegantes que se han vuelto populares entre los estudiantes y las escriben en playeras y tazas de café.

Este resultado matemático —las ecuaciones de Maxwell— para el electromagnetismo llevó no sólo a la aplicación de las ondas electromagnéticas en el mundo de las comunicaciones sino que, además, el haber entendido la interrelación entre la electricidad y el magnetismo guió el avance tecnológico desde finales del siglo XIX. Una de sus consecuencias es el motor eléctrico. Incontables usos de la energía eléctrica son posibles gracias al electromagnetismo. Las computadoras y todo aquello que llamamos "nuevas tecnologías" dependen, en un altísimo porcentaje, de nuestro descubrimiento del electromagnetismo. La enorme cantidad de información que se acumula hoy en discos de computadora y se transmite vía internet es almacenada en medios magnéticos y se extrae y transmite usando procesos electromagnéticos. La medicina moderna ha alcanzado diagnósticos antes insospechados a través de la resonancia magnética que es posible, precisamente, gracias a la manipulación del electromagnetismo. Incluso las nuevas generaciones de trenes de pasajeros llevarán grandes electroimanes y levitarán sobre vías, que son, a su vez, electroimanes. Los ciclotrones, utilizados en la investigación de los secretos de la materia y el origen del Universo, aceleran las partículas haciéndolas colisionar unas con otras, por medio de electroimanes.

La predicción de las ondas electromagnéticas es quizá la aplicación más importante en toda la historia de las matemáticas. Los efectos de este descubrimiento en la forma de vida de la humanidad han sido espectaculares.

3.6.5 La física cuántica

Todas las ciencias utilizan en mayor o menor medida las matemáticas, aunque la física es sin duda la que mayor provecho ha sacado de ellas. Cuanto más alejada se encuentra una ciencia de los fenómenos físicos, más difícil resulta utilizar en ella las matemáticas clásicas y más necesario se hace el recurrir a métodos estadísticos. Podría creerse que la física es determinista y que, sólo al alejarnos de ella, es necesario introducir el elemento de azar que reflejaría cierto grado de ignorancia. Sin embargo, durante la primera mitad del siglo XX se realizaron

descubrimientos extraordinarios en la física, en concreto en la estructura atómica de la materia, que condujeron a la necesidad de crear un modelo matemático, intrínsecamente aleatorio, para explicar el comportamiento del mundo atómico y subatómico.

Como suele ocurrir, un hallazgo de esta relevancia propició la creación de una nueva rama de la física, llamada esta vez mecánica cuántica. Niels Bohr, Heisenberg, Schrödinger y otros físicos llegaron a la conclusión de que si había un modelo matemático para el comportamiento de los electrones alrededor del núcleo en un átomo, éste debería incluir elementos de azar. De hecho, existe algo intrínsicamente aleatorio en la naturaleza. No se trata de simple falta de datos lo que nos lleva a introducir el elemento de azar sino que, en realidad, las partículas subatómicas tienen un comportamiento esencialmente aleatorio. Por ejemplo, al lanzar una partícula sobre una pared con dos pequeños huecos, puede que choque sin pasar por los huecos o, también, puede ser que pase por ellos, pero el que pase o no, no depende de manera determinista de la posición y velocidad inicial de la partícula, como ocurriría en el campo de la mecánica newtoniana que se aplica a los cuerpos macroscópicos. En el caso de las partículas elementales es imposible determinar con precisión absoluta la posición y velocidad inicial de una de ellas, es más: cuanto mejor definida esté su posición, menos definida estará su velocidad y viceversa. Esto se llama el *principio de incertidumbre* de Heisenberg y tiene una formulación matemática muy precisa:

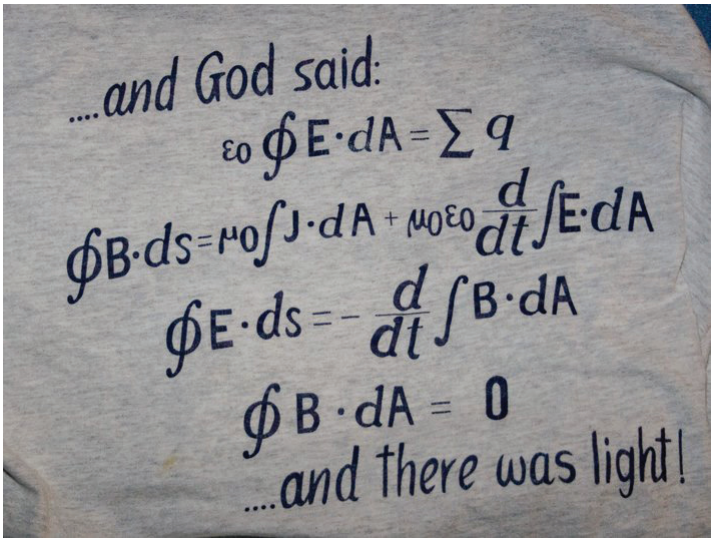


Figura 3.53 Las ecuaciones de Maxwell en forma integral.

$$\Delta x \cdot \Delta p \geqq \frac{h}{4\pi}$$

donde Δx es la incertidumbre en la posición, Δp la incertidumbre en la cantidad de movimiento íntimamente relacionado con la velocidad, y h es una constante, llamada la constante de Planck. El concepto de incertidumbre es equivalente al de desviación estándar que se usa en la estadística. El principio de incertidumbre dice que ninguna de las dos incertidumbres Δx o Δp puede ser cero y que si una de ellas es muy pequeña, la otra deberá ser muy grande. Schrödinger encontró una ecuación que describe la distribución de probabilidad de que un electrón —de un átomo— se encuentre en algún punto del espacio alrededor del núcleo:

$$i\frac{h}{2\pi}\frac{\partial \Psi}{\partial t} = \frac{h^2}{8\pi^2 m}\Delta\Psi + V\Psi$$

La anterior es la ecuación de Schrödinger para la amplitud de onda Ψ.

$|\Psi|^2$ es la densidad de probabilidad de la posición de una partícula.

La ecuación de Schrödinger, además, resulta ser una ecuación de onda, similar a la que describe el campo electromagnético, por lo cual surge la posibilidad de que los electrones de un átomo y las partículas elementales, en general, tengan un comportamiento ondulatorio. De Broglie predice que las partículas podrían comportarse como ondas; al final se demuestra experimentalmente que en verdad sucede así: al cruzar por una rendija, los electrones exhiben un comportamiento típico de las ondas: interferencia.

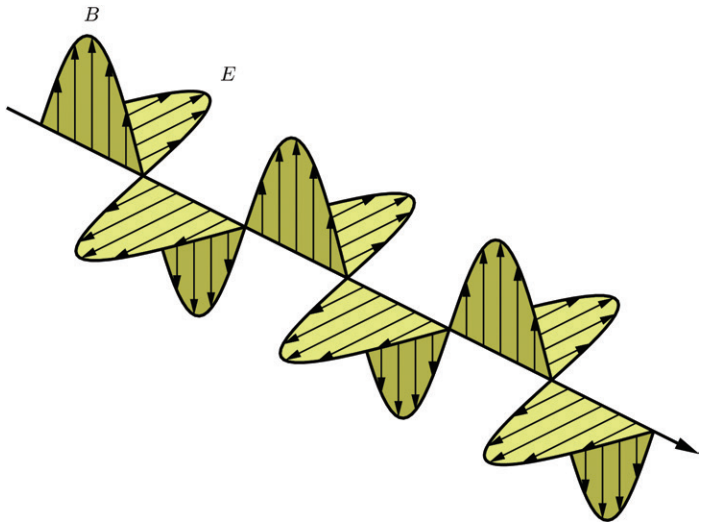


Figura 3.54 Esquema de una onda plana, cuya dirección del avance está indicada por la flecha principal. Las flechas horizontales representan el campo eléctrico y las verticales el magnético.

La distribución de electrones que han pasado a través de una rendija se parece mucho a la de la luz que pasa por una mirilla, similitud por la cual la ciencia estaba convencida de que la luz es un fenómeno ondulatorio. Se pensaba que la luz consistía probablemente en ondas electromagnéticas, pues la velocidad de estas ondas, de acuerdo con las ecuaciones de Maxwell, era exactamente la misma que se había medido para la luz: $300\,000\,\frac{\text{km}}{\text{s}}$.

De nuevo, al crear un modelo matemático para describir un fenómeno se llega a un descubrimiento científico: la materia exhibe un comportamiento ondulatorio. Ya con anterioridad el trabajo de Planck sobre la radiación del cuerpo negro, y el de Einstein sobre el efecto fotoeléctrico, habían sugerido que la luz, considerada en aquel momento como ondas electromagnéticas, se comportaba como si estuviera constituida por pequeñas partículas llamadas "cuantos". Ahora se deducía que lo contrario también era cierto, las partículas de las que estaba constituida la materia se comportaban como ondas. "¿Qué onda?", se preguntaría asombrado un joven mexicano... Por fin, ¿la luz es onda o partícula?, y la materia, ¿es onda o partícula? La conclusión de todas estas interrogantes e investigaciones es increíble: la luz y también la materia son ambas, onda y partícula. Hay una realidad dual. Quizá es difícil comprenderlo —igual que fue difícil entender el concepto del continuo y los límites—, pero esta dualidad pasó a ser un principio fundamental de la física que, apoyado en fórmulas matemáticas, permite describir el comportamiento aleatorio de las partículas-ondas de la materia y la luz.

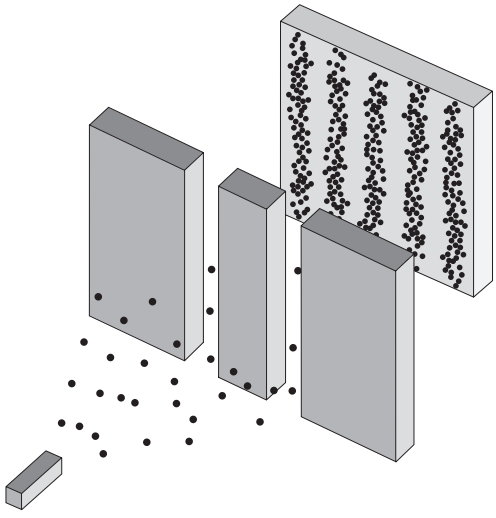


Figura 3.55 Esquema del experimento de la doble rendija para electrones en el que se observa el fenómeno ondulatorio de interferencia.

La física subatómica tiene aún muchos enigmas y cuestionamientos, y lleva años buscando su Santo Grial, es decir, una teoría unificada que, en caso de existir, consistiría en un conjunto de fórmulas matemáticas capaces de describir el comportamiento de la materia-energía, relacionándolas con todas las interacciones de la naturaleza, de las cuales hoy se conocen cuatro: la gravedad, el electromagnetismo, las interacciones nucleares débiles y las interacciones nucleares fuertes.

3.6.6 Albert Einstein y la teoría de la relatividad

El lector pensará que es exageración, pero aún hay otra rama de la física que se creó al definir un modelo matemático. Dicho modelo supuso predicciones sorprendentes que fueron verificadas después mediante observaciones y experimentos. Se trata de la teoría de la relatividad de Einstein.

En 1905, antes de que la física cuántica se desarrollara, uno de los más grandes genios de la humanidad, Albert Einstein —que entonces era un modesto y joven empleado de la oficina de patentes de la ciudad suiza de Zúrich—, publicó sus trabajos sobre tres inquietantes temas de la física de ese momento. Uno de ellos fue una brillante explicación del efecto fotoeléctrico, que consistía en cómo un haz de luz al incidir sobre ciertos materiales — fotoeléctricos— lograba que los electrones saltaran de esos materiales y pasaran a una placa con carga eléctrica positiva frente a ellos, generando de esa manera una corriente eléctrica. El efecto era conocido y no resultaba sorprendente que la luz, consistente en ondas electromagnéticas que transportan energía, interactuara con partículas de carga eléctrica como los electrones. Lo inaudito era el comportamiento cuantitativo del fenómeno. Por ejemplo, en el caso de cierto material con luz roja, los electrones no saltaban aunque ésta fuera de gran intensidad, mientras que con muy poca luz azul sí lo hacían.

Planck, en su teoría sobre la radiación del cuerpo negro, había señalado que la luz podía comportarse como si la luz estuviera constituida por partículas cuya energía era proporcional a su frecuencia: $E = hv$, donde v es la frecuencia y h es la llamada constante de Planck. A esas partículas de luz se les llamó "cuantos" aunque ahora se conocen como fotones. Einstein tuvo la idea de explicar la radiación del cuerpo negro aprovechando la idea de los cuantos de Planck: como la frecuencia de la luz azul es casi el doble que la de la luz roja, los cuantos de luz azul tenían mucha más energía que los de luz roja y por eso podían expulsar a los electrones de su órbita alrededor del átomo, mientras los de luz roja no. Esta sencilla explicación fue una de las ideas generadoras de la mecánica cuántica.

Otro de los trabajos publicados por Einstein, en 1905, contenía un modelo matemático que explicaba, cuantitativamente, el comportamiento del llamado movimiento browniano. Antes de este trabajo las causas del fenómeno estaban claras: se trataba de colisiones entre las moléculas del fluido y otras partículas más grandes, que se hallaban en suspensión. Lo paradójico de esta explicación es que la partícula en suspensión era demasiado grande en relación con las moléculas, para que el efecto individual de éstas al chocar pudiera producir el movimiento observado. El modelo matemático de Einstein resolvía esta aparente paradoja pues demostraba que las trayectorias quebradas descritas por la partícula en suspensión eran el resultado de muchísimos choques de las moléculas. La explicación de Einstein está muy relacionada con la ley débil de los grandes números, que nos dice que las variables aleatorias se acercan a sus valores esperados, pero lo hacen muy lentamente. A pesar de la importancia evidente de los dos trabajos mencionados, el tercero es el que mayor fama dio al sabio alemán y el que tiene un gran peso en la historia de intimidad entre la física y las matemáticas. Se trata del trabajo inaugural de la teoría de la relatividad.

A Einstein le preocupaba un detalle acerca de lo que se sabía sobre la luz y los electrones: por un lado, parecía que la velocidad de la luz era una constante universal; esto podía deducirse de las ecuaciones de Maxwell e incluso había un experimento —el de Michelson y Morley— cuyos resultados eran comprensibles sólo pensando que dos observadores, que se mueven uno con respecto al otro, obtendrían el mismo valor si midieran la velocidad de la luz. Sin embargo, era obvio que estos mismos observadores medirían velocidades diferentes para cualquier objeto en movimiento. Por ejemplo, si dicho objeto se encontrara en reposo para uno, para el otro estaría moviéndose con la misma velocidad que hay entre ambos. En el caso de la luz no pasa esto. La luz, entendida como onda electromagnética, no podría estar en reposo con respecto a uno de los observadores: como para ambos las ecuaciones del campo electromagnético son exactamente las mismas y, de acuerdo con ellas, las ondas electromagnéticas propagándose en el vacío tendrían que moverse a una velocidad específica de $300\,000\frac{\text{km}}{\text{s}}$, resulta que ambos observadores deben medir la misma velocidad para cualquier rayo de luz. Esto parece una paradoja, sin embargo Einstein, al dar más importancia a la lógica que a la intuición —que depende de nuestra limitada experiencia como ínfimos habitantes del Universo—, se dio cuenta de que la verdad era que la velocidad de la luz es efectivamente la misma para todos a quienes ilumine y la midan; en cambio las distancias, las masas y el tiempo, medidos por diferentes observadores, sí cambian.

$$
\begin{aligned}
t' &= \frac{t - vx/c^2}{\sqrt{1 - v^2/c^2}} \\
x' &= \frac{x - vt}{\sqrt{1 - v^2/c^2}} \\
y' &= y \\
z' &= z
\end{aligned}
$$

Figura 3.56
Transformaciones de Lorentz.

Las transformaciones de Lorentz son consecuencia lógica de que la velocidad de la luz es la misma para dos observadores que se mueven a velocidad constante, uno respecto al otro.

Esta conclusión, impecable desde el punto de vista lógico y tan contradictoria con nuestras experiencias vitales, puede representarse perfectamente mediante un modelo matemático y, como ocurrió en muchos otros casos donde se llegó a un modelo matemático de algún fenómeno de la naturaleza, el modelo conocido como la teoría de la relatividad, predice fenómenos que la ciencia acaba por verificar mediante la observación.

Para ejemplificar lo anterior, conviene aclarar que existen dos teorías de Einstein sobre la relatividad. Las conclusiones del estudio de Einstein sobre los cuerpos que se mueven con velocidad constante unos con respecto a otros se conocen como la teoría de la relatividad especial. La teoría de la relatividad general, desarrollada algunos años más tarde por el propio Einstein, incorpora fuerzas, aceleraciones y la atracción gravitatoria. En ella, Einstein encuentra que todo funciona como si la presencia de cuerpos con masa deformara el espacio y estas deformaciones son, precisamente, las causantes de las aceleraciones y curvaturas en las trayectorias de los cuerpos.

La teoría de la relatividad general predice que la órbita de Mercurio —y en realidad, la de todos los planetas— no es en verdad una elipse sino que es casi una elipse, pero con una precisión; es decir, el eje de la elipse va girando poco a poco. Como el fenómeno es más intenso cuanto más cerca está el planeta del Sol, es claramente observable sólo en Mercurio. La teoría de la relatividad también predice que la luz, al pasar cerca de un cuerpo masivo, podría deflectarse, es decir, que la atracción gravitatoria puede actuar sobre un rayo de luz. Esto lleva a considerar que la luz se comporta como si tuviera masa, una nueva sugerencia

de que luz y materia no son cosas diferentes sino manifestaciones distintas de un mismo ente. Pero, sin duda, la consecuencia de la teoría de la relatividad que más ha impactado a la sociedad está representada por la famosa fórmula:

$$E = mc^2$$

que dice que una masa m puede convertirse en una enorme cantidad de energía. La energía que puede llegar a generarse a partir de un objeto material es proporcional a su masa, y la constante de proporcionalidad es el cuadrado de la velocidad de la luz, que es un número muy, muy grande.

Esta fórmula anuncia un hecho angustiante: un solo gramo de materia podría generar suficiente energía para destruir la Tierra. La ciencia y la tecnología han unido esfuerzos para explotar esta importante relación entre masa y energía, produciendo, primero, las bombas atómicas y de hidrógeno, que convierten materia en energía de una manera violenta; después, creando los reactores nucleares que generan energía eléctrica de una manera controlada, a partir de materiales radiactivos y, posteriormente, al controlar la fisión nuclear —que se produce cuando dos átomos de hidrógeno se unen para formar uno de helio— con una pérdida insignificante de masa que libera una gran cantidad de energía. El control de la fisión nuclear representa una de las pocas oportunidades del hombre para obtener una fuente limpia de energía y sustituir el consumo de petróleo y otras formas contaminantes de producción energética.

Figura 3.57 Timbre de correos soviético de 1979, con Albert Einstein y la famosa fórmula $E = mc^2$.

3.6.7 Conclusión

No es de extrañar que los científicos sigan intentando descubrir fórmulas matemáticas que describan aquellos fenómenos de la naturaleza que aún no han podido ser comprendidos y dominados. Están convencidos de que cualquier avance en la representación matemática de la naturaleza volverá a traer consecuencias importantes para la vida y hasta para la supervivencia de la humanidad. En particular, la posibilidad de controlar la fisión nuclear y aprovecharla para generar energía sin producir contaminantes, el poder almacenarla y transmitirla hasta donde se necesite, quizá dependa de que se resuelvan algunos enigmas de la ciencia y de que la solución se exprese en forma de ecuaciones matemáticas.

Albert Einstein dedicó gran parte de su vida a intentar crear una teoría unificada del campo gravitacional y el electromagnético y no lo logró. El descubrimiento de otras dos fuerzas de la naturaleza llamadas nucleares —la interacción débil y la fuerte— complicó más el panorama sobre un posible modelo matemático integral y unificado de las fuerzas de la naturaleza, pero el sueño de Einstein continúa vigente y se cree que es posible realizarlo, a pesar de que las mentes más brillantes del mundo le han dedicado un peculiar y extraordinario esfuerzo durante años sin tener éxito. Desde que el hombre descubrió que su sola capacidad de raciocinio podría ser suficiente para comprender el mundo, se ha propuesto explicarlo todo mediante leyes racionales expresadas con fórmulas matemáticas. La teoría unificada sería el mayor logro imaginable en esa dirección; es, como ya dijimos, el Santo Grial de la ciencia contemporánea y nadie concibe esta teoría sin la participación de las matemáticas.

Las matemáticas se vinculan íntimamente con las leyes de la naturaleza. No sabemos por qué, pero es un hecho evidente dados los muchos ejemplos de esta estrecha relación que se encuentran en la historia de la ciencia. Quizá sea cierto aquello que decía Galileo de que la naturaleza es un libro abierto escrito en caracteres matemáticos. Es probable que el propio Galileo nunca imaginara que su verdad atravesaría los siglos alcanzando un sentido cada vez más profundo.

LAS MATEMÁTICAS DE LAS MATEMÁTICAS

TEMA

4

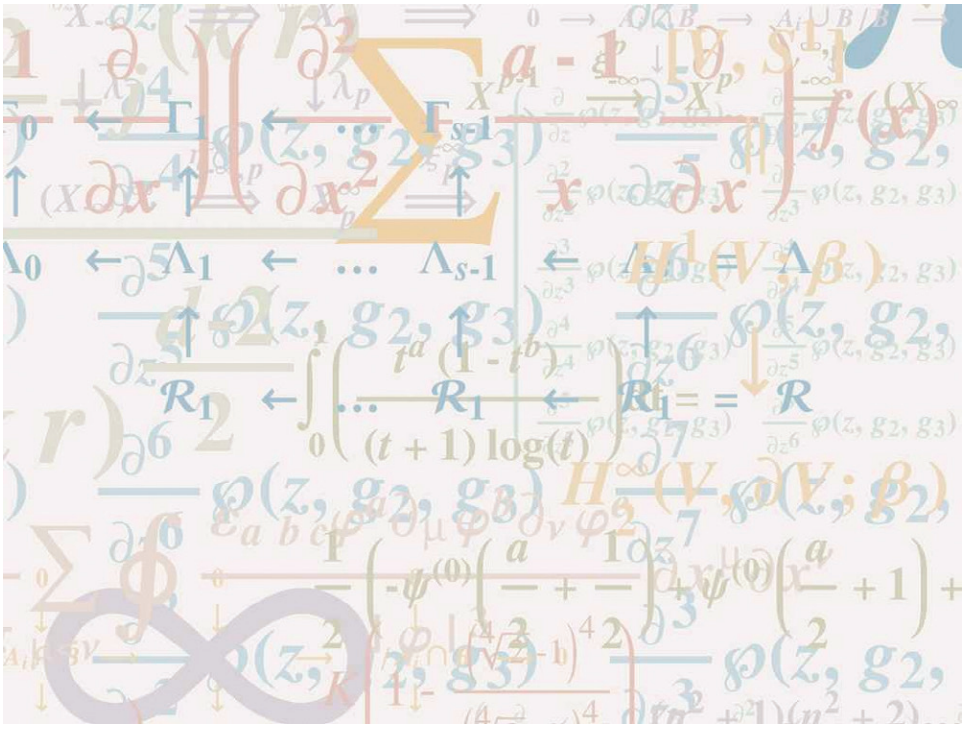


Figura 4.1 Se confunde con frecuencia a las matemáticas con simbología complicada. Pero en realidad la buena simbología es tan sólo una herramienta que ayuda a asimilar y trabajar con ideas profundas, bellas, naturales y con su razón de ser.

4.1 INTRODUCCIÓN

Una de las fuentes de inspiración más fértiles para la creación de las matemáticas son ellas mismas. Muchos de sus desarrollos más espectaculares surgen de mirarse obsesivamente el ombligo, es decir, del trabajo para resolver problemas que ellas mismas plantean. Siendo por naturaleza abstractas, parecería que este ensimismamiento las conduciría a una espiral que se aleja irremediablemente de la realidad. Pero no es así. La historia ha demostrado una y mil veces que matemáticas creadas en la estratósfera de la abstracción se convierten en la herramienta para entender o resolver problemas de otra ciencia o área muy concreta de la actividad humana. El ejemplo más clásico son las cónicas de los griegos que

reaparecen como las órbitas planetarias con Kepler, pero hay innumerables más: el MP3 y los DVD, Google, la relatividad general o la mecánica cuántica, por citar algunos otros ejemplos famosos de este fenómeno en que desarrollos abstractos de las matemáticas se aplican tiempo después.

Los matemáticos creemos, no como fe sino como método de trabajo, que esto sucede así porque hay *naturalidad* en las matemáticas y que éstas se descubren; que están *ahí* como la realidad lo está para las otras ciencias, pues da la sensación de que su *naturalidad* es algo cercano a la naturaleza —y la historia lo indica—. De tal manera que uno de los encantos de hacer matemáticas es ese contrapunto lúdico que se establece entre crear y descubrir. Henry Poincaré, uno de los matemáticos más importantes hace un siglo, decía que "el científico no estudia la naturaleza porque sea útil; lo hace porque se deleita en ello y se deleita en ello porque es bella". Quizá no todos los científicos estén de acuerdo con él hoy día, pero los matemáticos seguro que sí, e incluyen, como Poincaré lo hacía, a su materia de trabajo en el vocablo *naturaleza*.

Las secciones o apartados que se aglutinan en este capítulo responden a motivaciones o describen desarrollos que surgen de las propias matemáticas; de su dinámica interna. Pedimos entonces al lector que se muerda la lengua si lo asaltan las preguntas ¿y esto para qué sirve? o ¿en qué me será útil? Debe ser condescendiente al empezar cada apartado y dar por válida la motivación que se presenta, pues como fenómeno cultural así es como se han desarrollado en gran medida las matemáticas y vale la pena conocerlas como tal; tratar de apreciar por qué han cautivado a mentes tan extraordinarias.

En este capítulo iniciamos con un tópico muy clásico que es la razón áurea, pero poniendo énfasis en su interés matemático más que en el estético. Se sigue con secciones que versan en áreas como la combinatoria, la teoría de conjuntos, el álgebra y la geometría, con enfoques diversos pero siempre partiendo de los cuestionamientos que dieron origen a algunos de sus desarrollos. Después, damos un giro y revisamos aspectos de los fundamentos de las matemáticas así como de sus límites, para intentar transmitir que a partir de cuestionamientos aparentemente filosóficos también surgen matemáticas profundas. Concluimos con una breve exposición de la vitalidad y el crecimiento explosivo de las matemáticas contemporáneas.

4.2 RAZÓN ÁUREA (Y FIBONACCI)

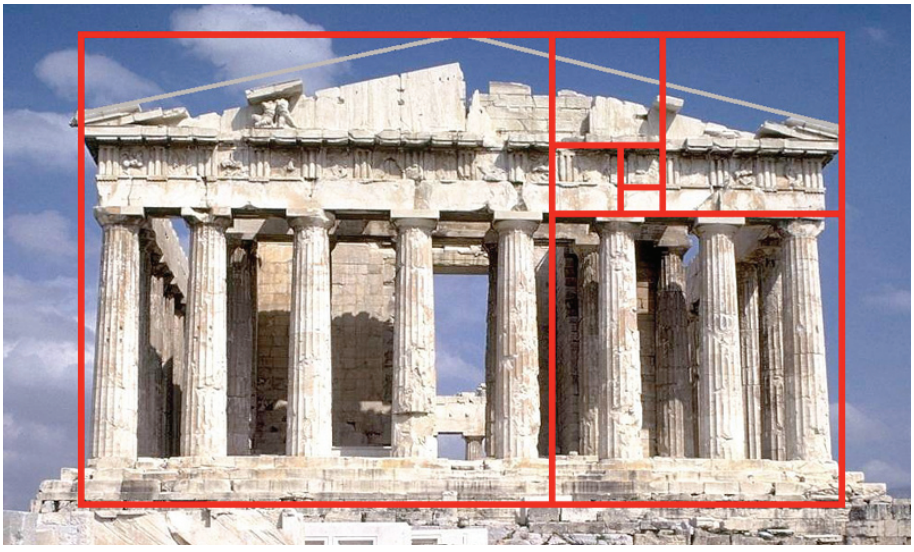


Figura 4.2 El Partenón en la Acrópolis de Atenas, Grecia. Su diseño arquitectónico es una oda a la razón áurea. <http://britton.disted.camosun.bc.ca/goldslide/gold08.jpg>.

No es claro quién descubre la razón áurea y cuándo; pero es un hecho que los griegos la usaron de manera prominente en su arte y arquitectura. El nombre que seguimos usando para ella, razón o proporción "dorada", o a veces también se le llama "divina", viene desde entonces. Uno de los ejemplos emblemáticos es la fachada del Partenón en la Acrópolis, que está en Atenas, Grecia: todo el diseño arquitectónico está basado en la razón áurea, como lo demuestra la imagen de la página anterior. Pero desde mucho antes se le conocía, porque en las grandes pirámides de Egipto también se le utiliza. La gran Pirámide de Giza, muy cerca de El Cairo, tiene en su diseño central un triángulo cuya proporción de la base a la hipotenusa es áurea.

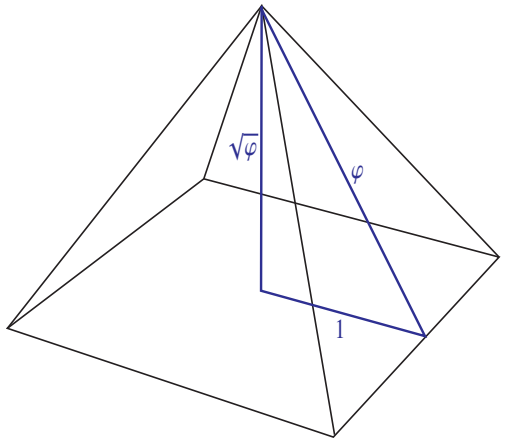


Figura 4.3 Pirámide de Giza y un esquema de su triángulo central | © Latin Stock México.

Pitágoras fue un gran estudioso de la razón áurea. Observó que las proporciones humanas tienen una estrecha relación con ella; observación que Leonardo da Vinci también documentó en el Renacimiento. Pero, ¿qué es la razón áurea?

Un rectángulo es *áureo* cuando al quitarle el cuadrado más grande posible (pegado a una arista corta), el rectángulo que queda tiene la misma proporción que el original.

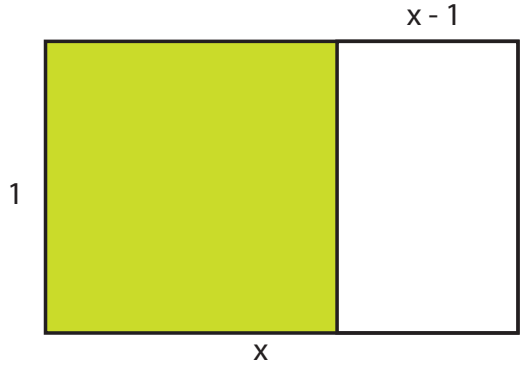


Figura 4.4 Rectángulo áureo.

Veremos que esta propiedad da lugar a una ecuación cuadrática. Como sólo nos interesa la proporción, podemos suponer que el lado chico del rectángulo es 1. Denotemos por x (la incógnita por excelencia) al lado grande. De tal manera que el cuadrado máximo es de lado 1, y el rectángulo que sobra tiene lados $x - 1$(el chico) y 1 (el grande). Si éste tiene la misma proporción que el original, se cumple la ecuación:

$$\frac{1}{x} = \frac{x-1}{1},$$

pues cada lado de la ecuación es la razón de lado chico a lado grande.

Multiplicando por x, se obtiene:

$$1 = x(x-1) = x^2 - x \ ,$$

que equivale a la ecuación cuadrática:

$$x^2 - x - 1 = 0 \ . \tag{1}$$

Se resuelve con la fórmula del "chicharronero" que nos da:

$$x = \frac{1 \pm \sqrt{5}}{2} \ .$$

Hay dos soluciones a la ecuación (1). Cuando usamos el signo $-$, como $\sqrt{5} > 1$, nos da un número negativo. Así que la solución que nos interesa es con el signo $+$, y podemos definir:

$$\varphi = \frac{1+\sqrt{5}}{2} \approx 1.6180339887499... \ . \tag{2}$$

donde ya estamos usando la notación común de llamar φ (es la letra griega que se lee "fi") al número áureo. Su proporción con el 1 es la razón áurea. Hemos demostrado que un rectángulo con lados a y b (con $a < b$) es áureo si:

$$\frac{b}{a} = \varphi \ .$$

Vale la pena hacer notar que la otra solución de la ecuación (1) es $-\varphi^{-1}$.

4.2.1 El pentagrama místico

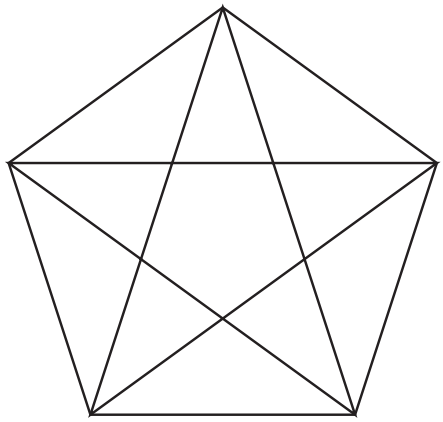


Figura 4.5 Pentágono regular y sus diagonales (el pentagrama).

Uno de los hechos que emocionaba a los griegos sobre la razón áurea, que les daba la seguridad de que su uso era trascendente y no casual, es que apareciera también en el pentágono regular: es la proporción de la diagonal al lado. Que estas dos longitudes, que surgen de un contexto tan diferente, también guarden la proporción divina, produce un estremecimiento o algo parecido al goce estético de estar ante el Partenón o ante un cuadro de Leonardo. Además, como veremos a continuación, en el razonamiento que conduce a este hecho, las piezas caen con tal precisión en el rompecabezas que es inevitable sentir "el toque mágico" de las matemáticas. No extraña, pues, que la escuela pitagórica haya adoptado al pentagrama como su símbolo, y que le hayan conferido poderes místicos.

Consideremos un pentágono regular con sus cinco diagonales (figura 4.5); a veces, se le llama el *pentagrama*. Supongamos que sus lados miden 1 y veremos a continuación que el número áureo φ es efectivamente la longitud de las diagonales.

Observemos primero que todos los ángulos que aparecen en el pentagrama son múltiplos de $\pi/5$ (léase "pi-quintos") o 36 grados (aquí, es mejor medir ángulos con radianes pues los argumentos se basan en que cinco de estos ángulos dan justo media vuelta, es decir, π radianes o 180 grados). Para verlo, considérense los ángulos que se forman en un vértice del pentágono, al trazar segmentos a los otros cuatro vértices y añadir su tangente al círculo que los contiene, como en la figura 4.6 a). Que los cinco ángulos que se forman son iguales, se sigue de un teorema general sobre los ángulos inscritos en un círculo, pues los cinco arcos en que se divide este último son iguales. De este hecho, y de que los ángulos de un triángulo suman π, se pueden obtener todos los ángulos que aparecen en el pentagrama en la figura 4.6 b).

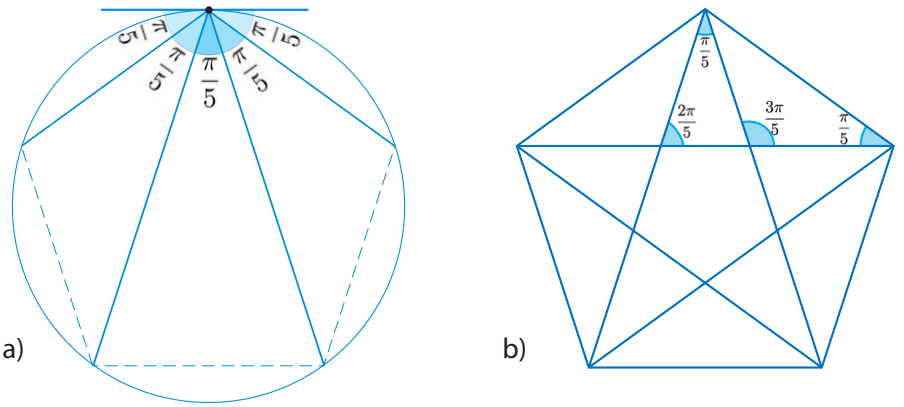


Figura 4.6 a) Los cinco ángulos en un vértice de un pentágono regular. b) Hay tres tipos de ángulos en el pentagrama.

Si se alargan dos lados no consecutivos del pentágono, el AC y el BD en la figura 4.7, se intersecan en un nuevo punto F.

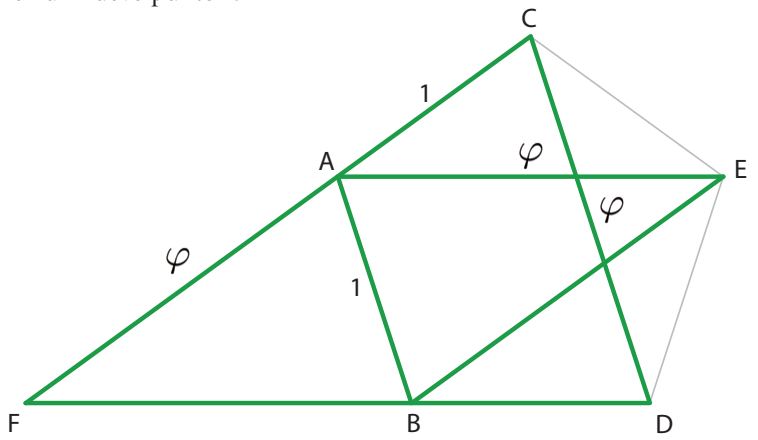


Figura 4.7 El pentágono regular con dos de sus lados alargados y tres diagonales.

Los triángulos isósceles ABF y ABE son iguales, pues su base, que mide 1, coincide y tienen los mismos ángulos (uno de $\pi/5$ y dos de $2\pi/5$, que se lee "dos pi-quintos"). Esto nos dice que los segmentos AF y BF miden φ (como la diagonal AE). Entonces, de la clara semejanza de los triángulos FAB y FCD, se obtiene que:

$$\frac{\varphi}{1} = \frac{\varphi + 1}{\varphi},$$

que es equivalente a la ecuación:

$$\varphi^2 = \varphi + 1 \, . \qquad (3)$$

Ésta es otra forma de escribir la ecuación (1) cuya solución positiva es el número áureo φ. Así que hemos demostrado que la diagonal de un pentágono regular está en proporción áurea con su lado.

La ecuación (3) es conocida como la *ecuación áurea.* Recordando que $\varphi^0 = 1$ y multiplicándola por φ^{n-2} se obtiene:

$$\varphi^n = \varphi^{n-1} + \varphi^{n-2}$$

para cualquier n positiva o negativa. En particular, se obtiene que $\varphi = 1 + \varphi^{-1}$ y que $1 = \varphi^0 = \varphi^{-1} + \varphi^{-2}$, de tal manera que todos los segmentos del pentagrama son potencias de φ, como se aprecia en la figura 4.8.

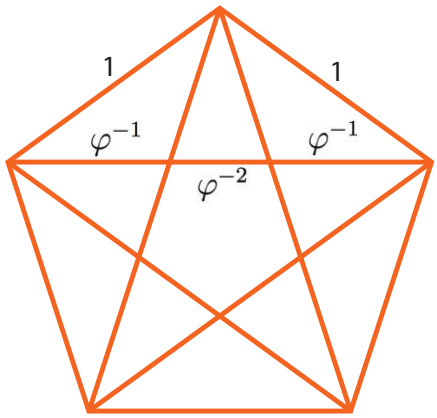


Figura 4.8 Todos los segmentos en el pentagrama son potencias de φ.

Los dos tipos de triángulos isósceles que aparecen en el pentagrama son conocidos como *triángulos áureos.* El *alto,* que usamos en la demostración, tiene dos ángulos de $2\pi/5$ y uno de $\pi/5$; y el *chaparro* tiene dos ángulos de $\pi/5$ y uno de $3\pi/5$. En el pentagrama hay cinco copias (una por arista) del triángulo áureo alto con lados 1 y φ, y diez copias del chaparro (una por vértice tanto del pentágono grande como del chiquito en el centro). Aparecen también, como sus intersecciones, copias de éstos en escalas áureas hacia abajo: diez altos y cinco chaparros con razón de semejanza φ^{-1} a los originales; y, por último, cinco altos en la segunda escala áurea hacia abajo, es decir, con base φ^{-2}, que son los picos de la estrella.

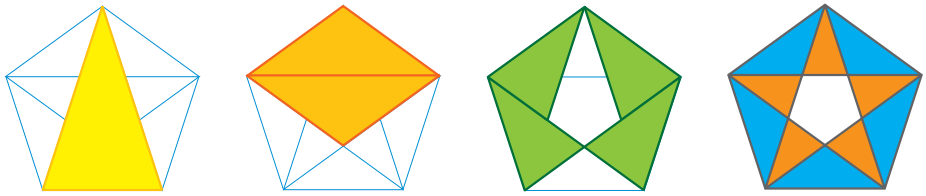


Figura 4.9 Triángulos áureos en el pentagrama.

Aunque el pentagrama y los triángulos áureos han sido estudiados, apreciados y usados por milenios, siguen dando de qué hablar. Recientemente, alrededor de 1975, el físico matemático inglés Roger Penrose descubrió una familia de mosaicos íntimamente relacionados con ellos que ahora se conocen como *mosaicos de Penrose.* Se construyen con dos piezas: los *papalotes* y las *dagas* que, a su vez, se construyen pegando dos triángulos áureos, altos y chaparros, respectivamente. En el entendido de que cada vez que se pegan dos piezas, los extremos de las *curvas de Amman* coinciden (véase figura 4.10), se puede llenar to-

do el plano. Pero siempre da lugar a mosaicos no periódicos, es decir, que no se pueden trasladar para regresar a sí mismos. Ésta es la propiedad que ha atraído la atención de físicos y químicos en los últimos 35 años, pues están relacionados con los llamados cuasicristales. Y, por supuesto, han sido también un deleite para los matemáticos pues tienen propiedades muy interesantes; una de ellas es que la proporción de papalotes a dagas siempre es áurea.

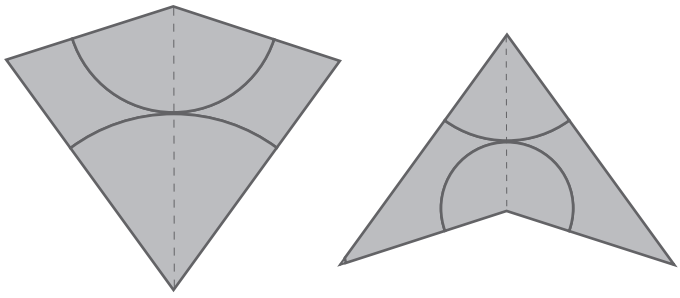


Figura 4.10 El papalote y la daga con los que se arman los mosaicos de Penrose. Los arcos de círculo, cuyos extremos parten las aristas en proporción áurea, deben coincidir al pegar dos piezas por una arista.

Figura 4.11 Mosaico de Penrose, de Juan Sandoval, en la sala de matemáticas de Universum | © Arturo Orta, Universum, DGDC-UNAM.

4.2.2 La sucesión de Fibonacci

Como sucede a menudo con conceptos fundamentales de las matemáticas, la razón áurea surge en varios contextos que, en primera instancia, parecen independientes. Uno de ellos es la famosa sucesión de Fibonacci:

$$1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, \ldots ,$$

donde cada término es la suma de los dos anteriores. Es decir, si llamamos F_n al n-ésimo (léase "enésimo") término de la sucesión, el que le sigue está definido por:

$$F_{n+1} = F_n + F_{n-1} \ . \qquad (4)$$

En dicha ecuación (4) hay que suponer que $n > 1$ y declarar que los dos primeros son: $F_1 = F_2 = 1$.

Fibonacci, cuyo nombre real era Leonardo de Pisa, la introduce en 1202 para estudiar la reproducción de los conejos, en un modelo muy simple que es precursor de lo que ahora se llama biología matemática o biomatemática. Aunque ahora sabemos que muchos siglos antes, los matemáticos de la India conocían esta sucesión y la usaban.

Lo que nos interesa en este momento de la sucesión de Fibonacci es que la razón entre términos consecutivos se aproxima a la razón áurea. Los primeros casos de estos cocien-

tes son: 1/1 = 1, 2/1 = 2, 3/2 = 1.5, 5/3 = 1.666 . . . , 8/5 = 1.6, 13/8 = 1.625, 21/13 = 1.6153 . . . , 34/21 = 1.61904 . . . . Claramente se aproximan a $\varphi = 1.61803$ alternadamente por abajo y por arriba. Pero, al usar la notación moderna de límites, nos interesa ver que:

$$\lim_{n\to\infty} \frac{F_n}{F_{n-1}} = \varphi \ . \tag{5}$$

Esta ecuación se lee "el límite cuando ene tiende a infinito de *efe-ene sobre efe-ene-menos-1* es igual a la razón áurea"; o bien "F_n/F_{n-1} tiende a φ cuando n tiende a infinito". Y quiere decir que entre más grande sea n, F_n/F_{n-1} aproxima mejor a φ.

Para verlo, recordemos nuestra primera definición de razón áurea quitándole un cuadrado a un rectángulo. Podemos usarla al revés: pegándole un cuadrado a un rectángulo áureo se obtiene un rectángulo más grande pero con la misma proporción. Este proceso de pegar cuadrados se puede aplicar a cualquier rectángulo e iterarse; en esa iteración aparecerá la sucesión de Fibonacci.

Si empezamos con un cuadrado unitario, 1×1: al pegarle un cuadrado obtenemos un rectángulo de 1×2; luego le pegamos un cuadrado de 2×2 y obtenemos un rectángulo de 2×3; al pegarle un cuadrado 3×3, obtenemos un rectángulo de 3×5, y así sucesivamente (figura 4.12). Este proceso da una sucesión de rectángulos cuyos lados son términos consecutivos de la sucesión de Fibonacci.

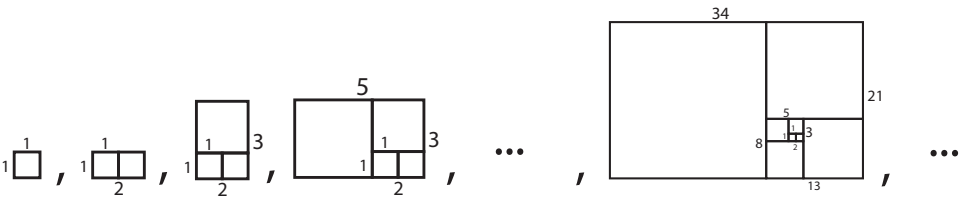


Figura 4.12 Sucesión de rectángulos que se obtiene pegando cuadrados.

En la figura 4.12, nos saltamos tres términos y tuvimos que reducir el octavo (un rectángulo de 21×34) para que cupiera. Pero no importa, porque sólo nos interesan las proporciones, no los tamaños explícitos. Si de nuevo nos saltamos tres términos obtendríamos un rectángulo de 144×233, que al volver a reducir es casi indistinguible del áureo, pues el error, evidente en los primeros términos, se ha hecho muy pequeño. De hecho, $|233/144 - \varphi| < 0.000022$.

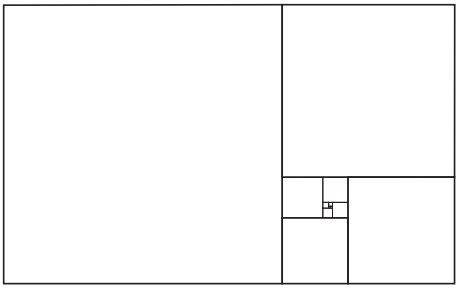


Figura 4.13: Término 12 de la sucesión de rectángulos.

El argumento de que el "error se hace más pequeño" conforme avanzamos en la sucesión de rectángulos, convence intuitivamente. Pero muchos matemáticos, aunque no rebatirían la veracidad de la ecuación (5), pedirían una prueba más contundente; intuirían que se puede tener más control en cómo se da la aproximación para argumentar mejor un hecho que sucede en el infinito. Puede parecer superfluo, pero esta insistencia en demostrar las cosas es lo que da solidez a las matemáticas.

Si generalizamos, al tiempo que recapacitamos en el argumento, obtendremos esa prueba. Por desgracia, puede ser que rebase el nivel de este libro pero, por las ideas que usa, puede resultar interesante y la incluimos.

Un rectángulo queda determinado por sus dos lados, que podemos codificar en una pareja ordenada de números (x, y). El proceso de pegarle un cuadrado es, según lo que hemos hecho, cambiar esta pareja por la pareja $(y, x + y)$. Así que si empezamos con la pareja $(x, y) = (1, 1)$, la que da origen a la sucesión de Fibonacci, y le aplicamos reiterativamente esta regla de cambio nos da las parejas $(1, 2), (2, 3), (3, 5), (5, 8), \ldots$ que son las parejas de términos sucesivos. Pero esto nos da una regla para una *transformación* del plano cartesiano en sí mismo:

$$(x, y) \mapsto (y, x + y) \ . \tag{6}$$

Al entender geométricamente esta transformación, obtendremos la prueba que buscamos.

El rectángulo áureo está representado por la pareja $(1, \varphi)$. Veamos qué le hace la transformación:

$$(1, \varphi) \mapsto (\varphi, 1 + \varphi) = (\varphi, \varphi^2) = \varphi(1, \varphi) \ ,$$

donde hemos usado la propiedad básica de la razón áurea ($\varphi^2 = \varphi + 1$), y denotamos por $t(x, y)$ a la pareja (tx, ty). Si pensamos en términos de vectores: al *vector áureo* $(1, \varphi)$, la transformación simplemente lo alarga en proporción áurea al vector $\varphi(1, \varphi) = (\varphi, 1 + \varphi)$; como era de esperarse. A un vector con esta propiedad (de ser cambiado de escala) se le llama *vector propio* de la transformación. El otro vector propio es su ortogonal:

$$(-\varphi, 1) \mapsto (1, 1 - \varphi) = (1, -\varphi^{-1}) = -\varphi^{-1}(-\varphi, 1) \ .$$

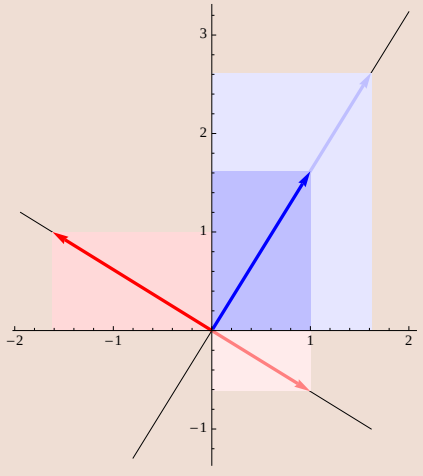


Figura 4.14 Qué le hace la transformación (6) a los vectores áureos $(1, \varphi)$ *y* $(-\varphi, 1)$.

Pero en este caso, el factor de escalamiento, o *valor propio*, $-\varphi^{-1}$, es negativo y de magnitud menor que 1. Éste será el meollo del asunto. Lo negativo, le invierte orientación a $(-\varphi, 1)$, y lo de magnitud menor que 1: ¡lo encoge!

Podemos entender la transformación $((x, y) \mapsto (y, x + y))$ en dos pasos. Primero, refleja en la recta de pendiente φ (lo negativo de $-\varphi^{-1}$), y después cambia de escalas: en la *dirección áurea* (del vector $(1, \varphi)$) hay un alargamiento áureo, de factor φ; y en su dirección ortogonal, un encogimiento áureo inverso, de factor φ^{-1}. De tal manera que si aplicamos la transformación a cualquier punto (x, y), se cambia de lado de la recta de pendiente φ pero se acerca a ella. Si, además, le pedimos que no esté en la recta de pendiente $-\varphi^{-1}$ e iteramos este proceso, se va hacia el infinito en la dirección de $(1, \varphi)$ o de $-(1, \varphi)$ (dependiendo de en qué lado estaba al principio), pues en cada paso se acerca a la recta en una proporción áurea inversa. Esto puede darnos el "control" que se necesita para precisar la aproximación, pero nos saltamos las fórmulas.

Al generalizar, hemos demostrado más de lo que pretendíamos. Casi cualquier sucesión tipo Fibonacci, es decir, dada por (4) pero con F_1 y F_2 arbitrarios, cumple (5). Pues hay que tomar el punto $(x, y) = (F_1, F_2)$ e iterar la transformación (6), para obtener la sucesión correspondiente. Sólo hay dos casos especiales. Cuando $F_1/F_2 = -\varphi^{-1}$, pues empezaríamos con un punto en la recta de pendiente $-\varphi^{-1}$ y entonces las iteraciones se mantienen ahí pero tienden al $(0, 0)$; o bien, cuando $F_1 = F_2 = 0$ y la sucesión es constante 0.

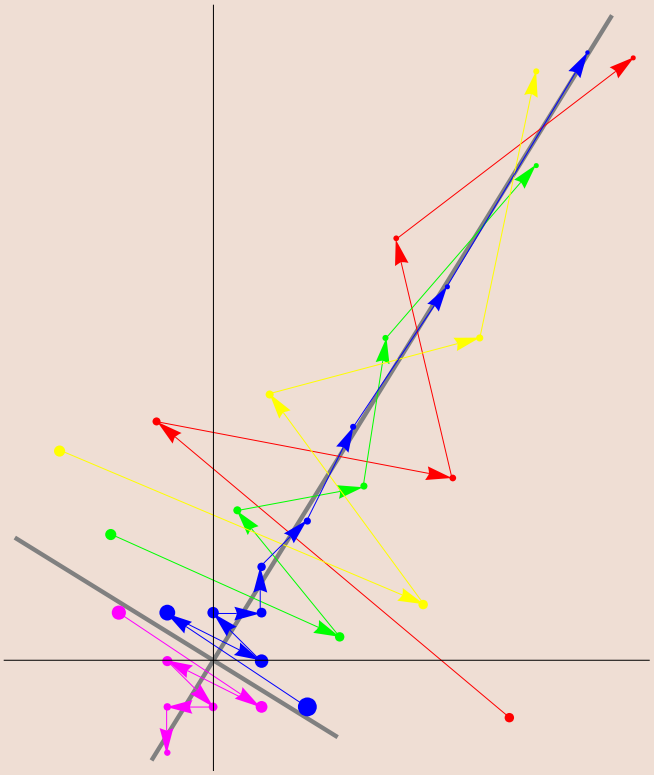


Figura 4.15 Iteraciones de (6) en diferentes puntos. Los azules pasan por la sucesión de Fibonacci.

En particular, hemos indicado cómo demostrar lo que queríamos de la sucesión de Fibonacci (la ecuación (5)). Se explica también por qué nuestros primeros cálculos se alternaban de menor a mayor que φ, y además, que en el proceso de añadir cuadrados a cualquier rectángulo, se obtiene como límite uno de proporciones áureas, pues la pareja (x, y) con que arrancaríamos la iteración tiene entradas positivas.

4.3 DE KÖNIGSBERG A GOOGLE

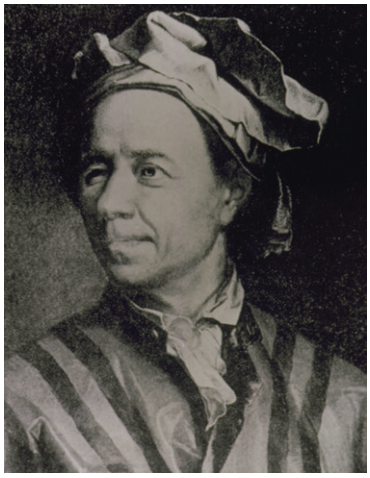


Figura 4.16 El suizo Leonard Euler (1707-1783) es considerado uno de los más grandes matemáticos de todos los tiempos | © Latin Stock México.

4.3.1 Los puentes de Königsberg

La ciudad de Königsberg en la antigua Prusia —que ahora se llama Kaliningrado y está en Rusia— se hizo célebre entre los matemáticos por un problema que resolvió Euler a principios del siglo XVIII. El problema era más bien un divertimento social entre los habitantes de una ciudad que se enorgullecía de sus puentes. Pero dio lugar a un desarrollo teórico que en la actualidad tiene una enorme importancia en muchas áreas de la actividad humana. El problema, juego o desafío consistía en dar un paseo por la ciudad de Königsberg de tal manera que se cruzaran sus siete puentes sin repetir ninguno. El mapa del Königsberg de aquella época está en la figura 4.17.

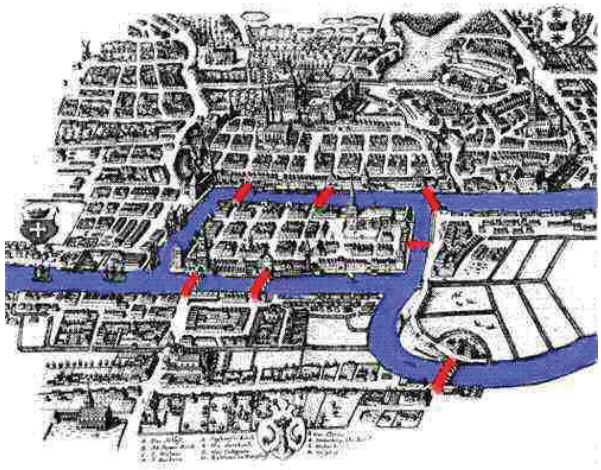


Figura 4.17: Los siete puentes de Königsberg alrededor de 1730.

El problema se hizo famoso porque, no obstante lo sencillo de su planteamiento, nadie daba con una respuesta. La solución de Euler no es diseñar un paseo con las características que se exigen, sino demostrar y hacer evidente que es imposible hacerlo. Para entender su argumento, conviene abstraer la información esencial del problema. Son cuatro los bloques de tierra, que si representamos con pequeños círculos, que llamaremos *vértices,* unidos por aristas que representan a los puentes, se obtiene la *gráfica* de la figura 4.18.

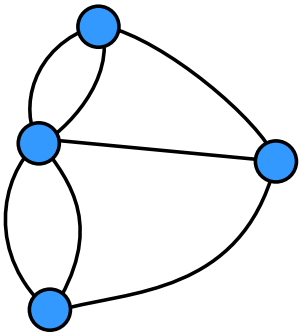


Figura 4.18 Gráfica de los siete puentes de Königsberg.

En lo que se refiere al problema, diseñar un paseo en esta gráfica, en un plano de Königsberg o echárselo a pie, en bici o en carroza por la ciudad, son equivalentes. Concentrémonos entonces en la gráfica, aunque sea válido hacer referencia a los otros dos enfoques. Un *paseo* consiste en una sucesión alternada de vértices y aristas que empieza y termina en vértices, en el entendido de que cada arista de la sucesión está entre los dos vértices que une —al pasear en la ciudad, cada puente comunica dos bloques de tierra—. Es decir, un paseo equivale a dibujar sobre la gráfica con un lápiz sin levantarlo nunca.

Esto remite a un problema común entre los niños de secundaria de la ciudad de México en los años cincuenta y sesenta, conocido como la *firma del diablo*. Consistía en dibujar un cuadrado con sus dos diagonales de un solo trazo: sin levantar el lápiz y sin dibujar sobre lo dibujado. Se insistía, para avivar la curiosidad, en que el diablo sí podía hacerlo.

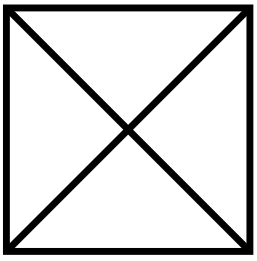


Figura 4.19 La firma del diablo.

Si ponemos un vértice en las cuatro esquinas y uno en el cruce del centro, el problema es equivalente al de los puentes de Königsberg; lo que ha cambiado es la *gráfica* en la cual se debe diseñar un paseo. Hemos *generalizado* el problema de los puentes de Königsberg a una *familia* de problemas: en cualquier gráfica se puede plantear. Y sucede muy frecuentemente que esto de generalizar ayuda a la solución.

La observación básica de Euler es que cada vez que en un paseo en una gráfica pasamos por un vértice, se usan dos de sus aristas —por la que entramos y por la que salimos—. De tal manera que si un vértice —vale la pena ponerle un nombre para referirnos a él en específico, y qué más simple que una letra—, llamémosle v —para pensar en vértice—, no es aquel donde iniciamos o concluimos un paseo que no repite aristas, entonces habremos usado un número par de las aristas que inciden en v. ¿Cuántas? Justo el doble de las veces que pasamos por v. Revisemos la gráfica de Königsberg para concluir que el paseo que se pide en el problema es imposible. En cada vértice inciden un número impar de aristas —hay un vértice con 5 aristas y tres con 3—; ninguno de ellos puede ser intermedio en un paseo que no repite aristas y las usa todas. Y el mismo argumento se aplica a la firma del diablo: hay un vértice con cuatro aristas —el del centro—, pero hay cuatro —las esquinas— con tres aristas. Faltan candidatos a ser vértices intermedios. ¡Ni el mismísimo diablo puede firmarlo!

4.3.2 Paseos eulerianos

Cuando se usa una y otra vez la misma descripción o frase, los matemáticos acostumbramos ponerle un nombre para que el discurso sea más fluido. En honor al primero que los estudió, se les llama *paseos eulerianos* a los paseos en una gráfica donde no se repiten aristas y además pasan por todas ellas, como los que estábamos buscando en Königsberg o en la firma del diablo. Puede haber de dos tipos: *cerrado* cuando el paseo empieza y acaba en el mismo vértice, o *abierto* cuando éstos son distintos. El otro concepto que se repite insistentemente es el del número de aristas que inciden en un vértice: llamémosle su *grado*. Podemos ahora limpiar y resumir el argumento de Euler.

Si una gráfica tiene un paseo euleriano cerrado, entonces todos sus vértices tienen grado par; las aristas donde se empieza y se acaba se pueden aparear para dar ahí el grado par. Pero el argumento da para más: si una gráfica tiene un paseo euleriano abierto, entonces todos sus vértices tienen grado par excepto el del principio y el del final —la arista con que empezamos y con la que acabamos no tienen con quién aparearse.

Euler, siendo un gran matemático, no se quedó satisfecho con su ingeniosa solución al problema de los puentes de Königsberg. Definió en general el concepto de *gráfica* como un conjunto de vértices (o nodos) junto con un conjunto de aristas, cada una de las cuales une (o *incide* con) dos vértices. Y se hizo la pregunta inversa: ¿será cierto que una gráfica que tiene todos los vértices de grado par admite un paseo euleriano cerrado? Y, si tiene únicamente dos vértices de grado impar, ¿tendrá un paseo euleriano abierto? Demostró que así es. Por ejemplo, si en Königsberg se construyera otro puente, cualquiera que sea, ya habría paseos eulerianos abiertos. O bien, cuando el diablo firma, pasa dos veces por una arista exterior pero lo hace con tal precisión y rapidez que no se nota el repaso.

Vale la pena reconstruir el argumento de Euler porque es indicativo del área de las matemáticas que ahora se llama *combinatoria* o *matemática discreta.*

Primero, observemos que el caso de dos vértices de grado impar se puede reducir al otro añadiendo una nueva arista entre los dos vértices especiales.

Dada una gráfica G —de nuevo será importante ponerle un nombre sencillo—, con todos sus vértices de grado par, iremos construyendo un paseo euleriano cerrado poco a poco. Pensemos que la tenemos dibujada con lápiz en un papel con circulitos de vértices y segmentos —rectos o curvos, pero simples— como aristas. Escogemos uno de sus vértices, llamémoslo v_0 —léase "v-cero" — y con un plumón rojo vamos recorriendo a G: armando paso a paso un paseo, con la única precaución de que no se vale pasar por una arista ya pintada de rojo (figura 4.20).

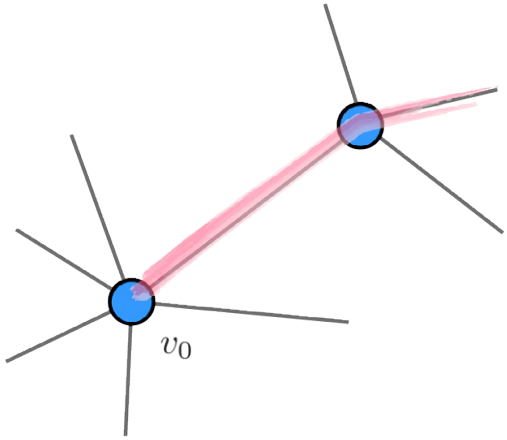


Figura 4.20
Inicio del paseo.

Cuando pasamos por algún vértice se marcan dos de sus aristas, así que de un vértice que no sea v_0 siempre podemos salir porque su grado es par: al llegar a él hay un número impar de aristas rojas y por lo tanto queda alguna en lápiz; por ella salimos y lo volvemos a dejar con "grado rojo" —y "grado en lápiz"— par. Entonces, la única manera en que nos atoremos y no podamos seguir adelante es regresando a v_0 y, además, cuando el resto de sus aristas ya son rojas (fig. 4.21).

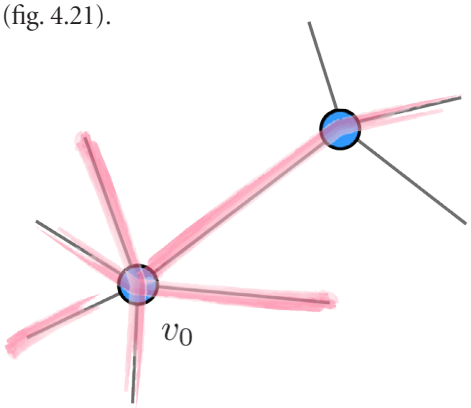


Figura 4.21 Final de la primera etapa.

En ese punto ya no tenemos por dónde salir. Hemos construido un paseo cerrado que empieza y termina en v_0 y usa todas las aristas que inciden en él. Si toda la gráfica G ya es roja, acabamos, pues estamos de suerte y, al primer intento, encontramos el paseo euleriano que buscábamos. Si no es así, quedan en G algunas aristas en lápiz y mejoraremos nuestro intento, pero usándolo como base.

Volvemos a recorrer el primer paseo —justo el mismo recorrido— hasta el momento en que lleguemos a algún vértice v_1 con aristas en lápiz. Hacemos ahí una pausa. Notemos que hay un número par de aristas en lápiz en v_1 —ya que hay un número par de rojas— y que en cada otro vértice también hay un número par de aristas en lápiz —que puede ser 0—. Salimos de v_1 por alguna de estas aristas en lápiz y seguimos y seguimos pintando más aristas de rojo. Por la misma razón que en el caso anterior, este deambular sólo puede terminar regresando a v_1 cuando todas sus aristas ya son rojas. Aquí concluye la pausa y retomamos el camino de nuestro primer paseo que termina en v_0. Este nuevo paseo, basado en v_0, incluye más aristas rojas que el primero.

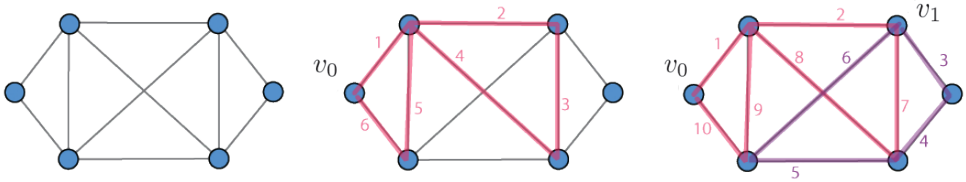


Figura 4.22 En una gráfica con todos los vértices de grado par, construcción de un paseo euleriano en dos etapas. La numeración en las aristas corresponde al orden del paseo.

Si no hemos pasado con el plumón rojo por todas las aristas de G, repetimos el mismo proceso pero ahora con nuestro paseo extendido como base y, así, seguimos una y otra vez. En cada vuelta ampliamos el paseo y pintamos más aristas. Y para demostrar que este proceso termina en algún momento, que el paseo se vuelve euleriano, tenemos que suponer dos cosas que no habíamos hecho explícitas. Primero, que la gráfica G —tanto vértices como aristas— es finita, lo cual sobreentendíamos al pintarla en un papel y obliga a que tiene que llegar el momento en que el paseo ya no pasa por ningún vértice con aristas en lápiz. Y segundo, que G es *conexa,* es decir, que no consiste en dos o más pedazos aislados, porque entonces nuestro proceso sólo pinta de rojo a *la componente conexa* en que vive v_0 nos da un paseo euleriano en ella, pero lo demás, como no tiene manera de comunicarse con v_0 a través de paseos, se queda pintada en lápiz: el plumón rojo nunca llega ahí.

4.3.3 La Red y la red

Éste fue el primer resultado de *teoría de las gráficas,* que ahora es un campo muy activo de las matemáticas. Tanto en sí mismo como por sus aplicaciones a muchas áreas del quehacer humano, pues las gráficas son objetos matemáticos que modelan muy diversas situaciones. Por ejemplo, *la Red:* ese universo informático, también conocido como internet, que está cambiando a pasos agigantados al mundo y la manera de interactuar de la humanidad. Puede pensarse que cada página web es un *nodo* o vértice de una gráfica, gigantesca y cambiante segundo a segundo, pero finita a fin de cuentas.

Y si una página web u incluye un *enlace* —o *link*— a otra página web v, se representa por una arista que va del vértice u al vértice v. Tiene la estructura de una *gráfica dirigida,* en la cual las aristas tienen *dirección,* principio y fin, van *de* un vértice *a* otro. Llamemos *la red,* con minúsculas, a esta gráfica dirigida. Se pierde en ella mucha de la información que hay en cada uno de los nodos, la información concreta de la página web en cuestión —así como en la gráfica de los puentes de Königsberg se pierde la información que hay en el mapa y en éste se pierde la información de la ciudad—. Pero la red modela algo fundamental de la Red. Representa su "conectividad" en abstracto.

En el último lustro del siglo xx, el estudio de la red como gráfica llevó a dos estudiantes de la Universidad de Stanford a diseñar el buscador Google, que se ha convertido en la página web más importante y consultada de la Red. Entre los ingredientes básicos de este buscador está el que, como en cualquier gráfica, sus vértices adquieren un orden natural, llamado *espectral,* que ya se había estudiado. No tenemos herramientas para describirlo, pero con base en ese orden espectral es como se recorren los vértices de la red para hacer una búsqueda en la Red y a este método debe su éxito Google. El punto es señalar otra historia de cómo de un divertimento intelectual surgen conceptos y objetos interesantes para los matemáticos que los estudian *per se,* y siglos después se aplican y cambian nuestro mundo.

4.4 LA CONQUISTA DEL INFINITO

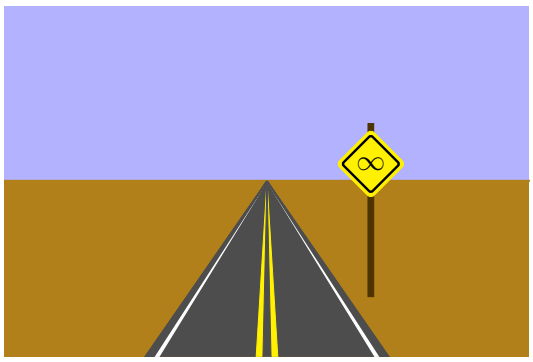


Figura 4.23 Cuando vemos cómo se pierde una carretera en el horizonte, nos acercamos a la forma de imaginar el infinito. Como el infinito fue por mucho tiempo simplemente lo contrario de lo finito, no fue fácil aclarar el concepto y no fue sino hasta finales del siglo XIX *cuando esto se resolvió.*

4.4.1 Cómo medir lo infinito

Aristóteles había escrito que *el todo es más que la suma de sus partes.* Euclides, medio siglo después, lo acortó y dijo que *el todo es mayor que la parte.* Fue Galilei quien por primera vez observó que estas afirmaciones podían fallar si se trataba del infinito. En su libro *Consideraciones y demostraciones matemáticas sobre dos nuevas ciencias* dice, por boca de Simplicio, que "seguramente la infinitud de una línea corta es menor que la infinitud de una línea más larga"; la respuesta de Salviati es que "eso son dificultades que resultan al tratar de compren-

der, con nuestro intelecto finito, algo infinito y se atribuyen al infinito propiedades que se conocen de lo que es finito, pero esto no es válido".

El ejemplo que expone Salviati es interesante. En ello, argumenta que podemos elevar al cuadrado cada número y obtener un *cuadrado perfecto,* es decir, un cuadrado cuya raíz cuadrada es un número natural. De esta manera, se muestra que los números naturales y los números cuadrados *se corresponden uno a uno.* Lo anterior choca con nuestro sentido común, entrenado por el mundo de lo finito, de que una parte —en este caso, los números cuadrados— son sólo una parte de todos los números naturales. A este tipo de problemas se le conoció como *paradojas del infinito* y no se sabía en un inicio cómo lidiar con ellos ya que el infinito fue, por mucho tiempo, lo contrario de lo finito.

El problema de considerar conjuntos, es decir, dos colecciones de objetos, consiste en establecer un criterio claro para compararlos. Si se trata de conjuntos infinitos como, por ejemplo, el conjunto de números naturales, entonces no podemos contarlos: el proceso de contar —decir número tras número señalando un objeto tras otro hasta llegar al número asignado al último objeto, que es el número de objetos en el conjunto— no funciona porque no hay un último objeto.

Por consiguiente, hay dos alternativas. La primera es la contención. Podríamos decir que los números cuadrados perfectos son menos que los números naturales, porque los primeros están contenidos como una parte propia del segundo conjunto. Pero con este criterio no podemos comparar el conjunto de tres manzanas con el de dos peras, ya que peras no son manzanas y el conjunto de peras no es un *subconjunto,* es decir, una parte del conjunto de manzanas. Vemos que el número es una abstracción y, al contar, hacemos una asignación uno a uno de los objetos con los primeros números.

La segunda alternativa consiste en tomar la correspondencia de uno a uno como criterio de comparación. Las consecuencias son sorprendentes. El hombre que trabajó con detalle esta idea fue Georg Cantor, matemático alemán, y lo hizo precisamente con esta manera de comparar que ya Galilei había propuesto. Con ello, concluimos que, en efecto, hay la misma cantidad de números cuadrados perfectos que de números naturales. Si entre dos conjuntos A y B no es posible establecer una correspondencia uno a uno, pero sí es posible establecer una correspondencia del conjunto A con una parte del conjunto B, entonces diremos que A es de menor *cardinalidad* que B. Con este concepto vemos sin problema que el conjunto de dos peras es de menor cardinalidad que el conjunto de tres manzanas.

Los conjuntos se dividen en diferentes cardinalidades formando *clases de equivalencia* de conjuntos que tienen la misma cardinalidad, es decir, conjuntos entre los cuales es posible establecer una correspondencia uno a uno. Los cardinales finitos son aquellos que corresponden a los números naturales. Después, hay un primer cardinal infinito que corresponde al conjunto de los números naturales. Cantor denotó a este cardinal con $\aleph_0$ y se lee "alef cero" —aleph por ser la primera letra del alfabeto hebreo y cero porque es el primer cardinal infinito—. Los conjuntos de cardinalidad $\aleph_0$ se llaman también *numerables,* ya que una correspondencia con los números naturales indica una manera de numerar estos elementos. El elemento que corresponde al número 1 será el primero en la lista, el que corresponde al 2 el segundo y así succesivamente.

Cualquier conjunto infinito contiene una parte que tiene cardinalidad $\aleph_0$, que se puede ver de la siguiente manera. Como el conjunto es infinito, podemos empezar a hacer una lista de elementos sin repetir ninguno. Esta lista siempre es finita y, por lo tanto, siempre se puede extender. De esta manera se obtiene un listado $\ell_1, \ell_2, \ell_3, \ldots$ infinito con cardinalidad $\aleph_0$.

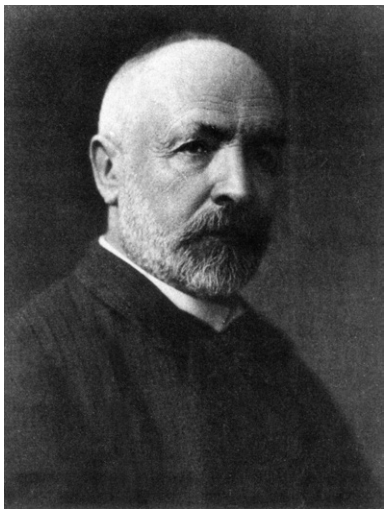


Figura 4.24 Georg Cantor (1845-1918) fue un matemático alemán al que se le atribuye la invención de la teoría de conjuntos, *base de las matemáticas modernas | © Latin Stock México.*

Lo anterior se puede aplicar justo a los números naturales. Si quitamos el primer número, seguimos con un conjunto infinito $\ell_1 = 2, \ell_2 = 3, \ell_3 = 4, \ldots$ de cardinalidad $\aleph_0$. Así, obtenemos una correspondencia uno a uno, como se observa en la figura 4.25, donde se muestra la correspondencia con flechas dobles.

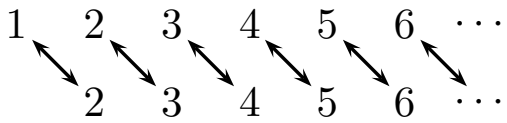


Figura 4.25 La correspondencia uno a uno de los números naturales con una de sus partes.

De hecho, cada conjunto infinito tiene la misma cardinalidad que una de sus partes y sólo los conjuntos infinitos tienen esta propiedad. En efecto, si A es un conjunto infinito, especificamos un subconjunto $L = \{\ell_1, \ell_2, \ldots\}$ y denotamos con $B = A \setminus L$ lo que queda de A al remover los elementos de L. Ahora, sea $A' = A \setminus \{\ell_1\}$, es decir, quitamos únicamente el elemento ℓ_1 y establecemos una correspondencia uno a uno entre A y A', como en la figura 4.25 entre L y $L \setminus \{\ell_1\}$ y la identidad entre B y B.

4.4.2 Diferentes infinitos

Cantor descubrió que hay más que un tipo de infinitos. Demostró que los números enteros —incluyendo los negativos— tienen la misma cardinalidad que los naturales y que también los racionales tienen la misma cardinalidad. Pero los números reales, aquellos que vimos en la sección 2.3 para expresar lo continuo, tienen otra cardinalidad, una que es mayor que $\aleph_0$.

El argumento que dio Cantor se conoce hoy como el *argumento diagonal* y se basa en la contradicción, es decir, supone que los números reales son numerables. Si éste fuera el caso, sería posible hacer una lista completa —aunque infinita— de los números reales. A continuación, se demuestra que hay un número que no está en la lista, lo que constituye una contradicción a la suposición de que la lista era completa pues contenía todos los números reales. Como en la argumentación no hay error alguno, se concluye que la suposición fue falsa y así queda demostrado que los reales no son numerables.

Veamos ahora cómo construir el número real que no está en la lista. Para ello, se usa la misma lista que se suponía completa y se escribe cada número en su expansión decimal.

Después, se construye el número en su expansión decimal al usar los dígitos que se encuentran en la diagonal de la lista. Observemos cómo podría empezar la lista —dado que los dígitos de la diagonal juegan un papel importante, se marcaron con color rojo:

$$\begin{aligned} n_1 &= 4.151784200\ldots \\ n_2 &= 1.332699051\ldots \\ n_3 &= 0.067001004\ldots \\ n_4 &= 4.223991063\ldots \\ n_5 &= 8.010207501\ldots \\ n_6 &= 0.005141699\ldots \\ n_7 &= 8.112086032\ldots \\ n_8 &= 8.888888888\ldots \\ n_9 &= 3.141592653\ldots \end{aligned} \tag{7}$$

De esta manera, obtenemos una sucesión de números:

$$1, 3, 7, 9, 0, 1, 0, 8, 3, \ldots \tag{8}$$

Ahora bien, se construye un nuevo número x de manera que el dígito en posición n sea diferente del dígito en la misma posición de la sucesión (8). Esto se puede hacer, por ejemplo, de forma que cada dígito 1 se convierte en 0 y los otros dígitos en 1. En nuestro caso, obtendríamos:

$$x = 0.011110111\ldots$$

Entonces, encontramos que $x \neq n_1$ ya que difieren en el primer dígito, que $x \neq n_2$ ya que difieren en el segundo y así sucesivamente. La conclusión es que x no es ningún número de la lista y la lista fue incompleta. Hemos llegado a la contradicción que buscábamos.

Cantor compara diferentes conjuntos y es capaz de mostrar que la cardinalidad de los dos conjuntos es la misma:

$$\{(x, y) \mid y = 0, 0 \leq x \leq 1\} \quad \text{y} \quad \{(x, y) \mid 0 \leq y \leq 1, 0 \leq x \leq 1\}\ .$$

Los dos conjuntos describen el lado y la superficie de un cuadrado. Es decir, el conjunto de puntos de un segmento es igual que la cardinalidad de toda el área. Eso fue muy sorprendente. Con el uso de los *conjuntos potencia* se puede demostrar aún más: hay una infinidad de "diferentes infinitos".

El conjunto potencia consiste de todos los subconjuntos de un conjunto dado. Por ejemplo, si $X = \{1, 2, 3\}$, entonces los subconjuntos de X son:

$$\{\}, \quad \{1\}, \quad \{2\}, \quad \{3\}, \quad \{1, 2\}, \quad \{1, 3\}, \quad \{2, 3\}, \quad \{1, 2, 3\}\ .$$

Las "llaves" { } indican la colección del conjunto. Entre ellas, se enlistan los elementos del conjunto, por ejemplo $\{1\}$ que es el conjunto que sólo contiene al elemento 1, mientras { } denota al *conjunto vacío,* es decir, aquel conjunto que no contiene ningún elemento. El conjunto potencia de un conjunto X se denota con (X). Por ejemplo:

$$(\{1,2,3\}) = \{\{\},\{1\},\{2\},\{3\},\{1,2\},\{1,3\},\{2,3\},\{1,2,3\}\}$$

Esto se ve muy complicado en la notación, pero no es tan difícil. Los elementos de (X) son conjuntos a la vez y, por ello, aparecen las llaves dentro de otras llaves.

Ahora se puede demostrar que la cardinalidad de (X) es siempre mayor que la cardinalidad de X. En efecto, se trata nuevamente de un argumento de contradicción. Se supone que X y (X) tienen la misma cardinalidad y que existe una correspondencia uno a uno entre ellos. Escribiremos esta correspondencia como una función:

$$e\colon X \longrightarrow (X).$$

Entonces, para cada elemento $x \in X$, tenemos que $e(x) \in (X)$ es un elemento de (X) y, por lo tanto, un subconjunto de X. Ahora nos podemos preguntar si el elemento x pertenece o no al subconjunto $e(x)$. Dividimos los elementos de X en dos subconjuntos S y N y decimos que $x \in S$ si x sí es un elemento de $e(x)$ y, por el contrario, si x no es un elemento de $e(x)$, entonces $x \in N$.

Como la función e establece una correspondencia uno a uno, hay un elemento n que corresponde al conjunto N, es decir $e(n) = N$. Luego, nos preguntamos si n es un elemento de $e(n)$ o si no lo es. Veamos, si $n \in e(n)$, entonces n es uno de los elementos de S y no de N. Por lo tanto, $n \notin N = e(n)$, que es justo lo contrario de $n \in e(n)$. Contrariamente, si $n \notin e(n) = N$, entonces $n \in S$ y, por definición, tenemos que $n \in e(n)$, ¡otra vez una contradicción! Esto muestra que seguro hay una contradicción si X y (X) tienen la misma cardinalidad.

Es claro que $X \longrightarrow (X), x \mapsto \{x\}$ es una función *inyectiva,* es decir, diferentes elementos se envían a diferentes elementos. Finalmente, esto muestra que la cardinalidad de X es menor que la cardinalidad de (X).

Con los conjuntos potencia se muestra que hay una sucesión infinita de diferentes infinitos, cada vez más grandes.

4.4.3 Números ordinales y números cardinales

Ya vimos que hay muchos infinitos y, con los conjuntos potencia, sabemos que podemos encontrar cardinalidades cada vez más grandes. Se escribe $|X|$ para denotar la cardinalidad del conjunto X. En efecto, se puede mostrar que cualesquiera dos conjuntos X y Y siempre son *comparables,* es decir, que sólo hay tres opciones posibles: $|X| < |Y|$ o $|X| = |Y|$ o $|X| > |Y|$. En otras palabras, si las cardinalidades de X y Y no son iguales, entonces o el conjunto X tiene la misma cardinalidad que un subconjunto propio de Y o, al revés, el conjunto Y tiene la misma cardinalidad de un subconjunto propio de X. Esto no es para

nada obvio y Cantor tuvo que introducir los *ordinales;* primero, demostró que cualesquiera dos conjuntos ordenados, siempre se pueden comparar. Como consecuencia los números cardinales, es decir, las cardinalidades siempre se pueden comparar en forma muy similar a como se comportan los números naturales o reales.

Dado que los cardinales se pueden comparar, Cantor pudo definir el cardinal $\aleph_1$ como el primer número cardinal que es mayor que $\aleph_0$. El número cardinal $\aleph_2$ es el que sigue después de $\aleph_1$, y así sucesivamente. Lo que, en cambio, no quedaba nada claro era la cardinalidad del conjunto de los números reales. Es casi seguro que Cantor pensara que los reales tienen la cardinalidad $\aleph_1$, lo que hoy se conoce como la *hipótesis del continuo.* Pero unos 50 años después de él, se descubrió que esta pregunta es *independiente* del sistema de axiomas de Zermelo-Fraenkel. Como consecuencia, se puede hacer matemáticas con la hipótesis de que los reales tienen cardinalidad $\aleph_1$ o con la hipótesis de que la cardinalidad de los reales es mayor que $\aleph_1$.

Lo anterior debería ser algo inquietante pues afirma que el conjunto de los números reales no se conoce bien, ni siquiera en lo que se refiere a sus subconjuntos. No se puede determinar si hay o no un subconjunto X de números reales que tenga una cardinalidad intermedia entre los naturales y los reales o si tal subconjunto no existe.

4.4.4 La base formal de las matemáticas

Incluso en tiempos de Cantor, un conjunto era simplemente una colección de objetos que tenía la única restricción de que a partir de un conjunto siempre se debía poder decidir si algún objeto pertenecía o no a dicho conjunto. Con esta definición tan amplia se llegó rápidamente a contradicciones, entre las cuales la que formuló Russell es la más famosa.

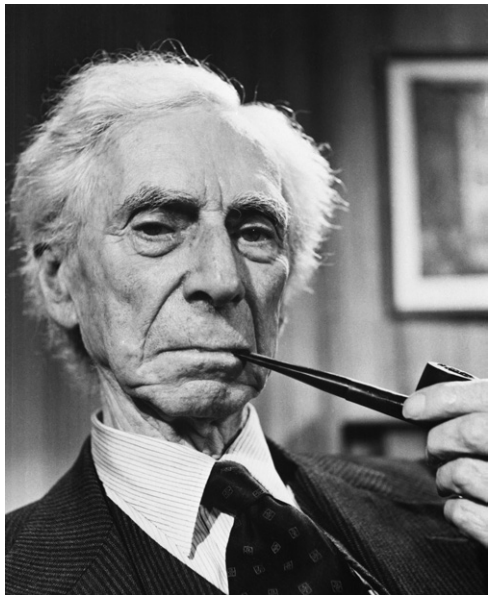


Figura 4.26 Bertrand Russell (1872-1970) fue un matemático, filósofo y escritor británico que también se dedicó al pacifismo y a defender los derechos de las mujeres | © Latin Stock México.

Russell dice que conoce un pueblo donde el único barbero afeita a todos los hombres que no se afeitan por sí mismos. A primera vista, esto suena totalmente plausible. La formulación implica que cada hombre es afeitado: si no lo hace él solo, entonces visita al barbero. Con los hombres del pueblo se forman dos subconjuntos: los que se afeitan solos y los que visitan al barbero. Ahora, ¿a cuál subconjunto pertenece el barbero?

Al hacer las conclusiones lógicas se deduce rápidamente que, de cualquier manera, se llega a una contradicción. Con ello, Russell muestra que los conjuntos no se pueden definir de una manera tan laxa. En efecto, la definición que hoy opera para un conjunto es más restrictiva y evita que se generen este tipo de contradicciones.

Y aunque la discusión sobre cuál debe ser exactamente la base de las matemáticas no se ha concluido, es un hecho que los conjuntos son los que forman la base moderna para las matemáticas. A partir de los conjuntos, se forman todos los demás conceptos desde un punto de vista más formal. Como ejemplo, veremos de qué manera el concepto de *función* se basa en el de conjuntos. Una función tiene un *dominio* A y un *codominio* B y "asigna" a cada elemento de A un elemento de B. Ahora, esto se transforma en el lenguaje abstracto de conjuntos. Para ello, partimos de dos conjuntos A y B y definimos primero el concepto de *producto cartesiano* $A \times B$, que consiste en los pares ordenados (a, b) con $a \in A$ y $b \in B$. Formalmente, también se debe definir *par ordenado* con base en conjuntos; esto lo indicamos al lector interesado en el siguiente recuadro.

Si a y b son dos elementos, entonces $\{a, b\}$ es el conjunto que contiene los dos elementos a y b. Consecuentemente, tenemos $\{a, b\} = \{b, a\}$ ya que en un conjunto no hay orden alguno y sólo se puede saber si un elemento le pertenece o no. En matemáticas, se suele denotar con (a, b) al *conjunto ordenado* que contiene a a como primer elemento y a b como segundo. Esto se puede realizar de la siguiente manera. Se define (a, b) como el conjunto:

$$\{\ \{a\},\ \{b, \{\}\}\ \},$$

es decir, el conjunto que tiene los dos elementos: $\{a\}$ y $\{b, \{\}\}$.

El primero de estos dos conjuntos tiene un solo elemento —que es a—, mientras el segundo conjunto tiene dos elementos: b y el conjunto vacío. De esta manera se pueden recuperar los dos elementos sin confundirse: vemos la cardinalidad de los dos elementos y concluimos, a partir de ello, cuál es el primer elemento y cuál el segundo.

Con el concepto del par ordenado podemos formar ternas $(a, b, c) = (a, (b, c))$, cuartetas o n-adas con n elementos.

Un subconjunto $\Sigma \subset A \times B$ —esto se lee "sigma subconjunto de A cruz B" — se llama *gráfica* si para cada $a \in A$, existe un $b \in B$ —uno y solamente uno—, tal que $(a, b) \in \Sigma$. La condición realiza lo que antes se expresó: "la función asigna a cada elemento $a \in A$, un elemento $b \in B$". Finalmente, una *función* es un triplete (A, B, Σ) donde A, B son conjuntos y $\Sigma \subset A \times B$ es una gráfica.

Los conjuntos no sólo sirvieron para liberar y aclarar el concepto del infinito sino, también, para dar formalidad a todos los conceptos matemáticos en la actualidad. En este sentido, han sido el gran artículo de moda matemática para todo el siglo xx.

4.5 EL DESARROLLO DEL ÁLGEBRA

Figura 4.27 La imagen muestra el libro de álgebra del docente cubano Aurelio Baldor. Centrado en la aritmética y el cálculo con variables es, sin duda, el libro más consultado de álgebra en América Latina. La portada presenta al matemático persa Muhammad ibn Musa Al-Juarismi, que escribió un libro llamado Al-Kitab al-Jabr wa-l-Muqabala *y que significa "Compendio de cálculo por el método de completado y balanceado".*

4.5.1 La aritmética

La *aritmética* se encarga de la operación de números; es la rama más antigua de las matemáticas y, además de la geometría, la que más desarrollo ha generado con el pasar de los siglos.

Al principio, los fines de la aritmética eran sólo prácticos; por ejemplo, enseñaba cómo hacer cálculos correctamente y, además, de manera eficiente.

No debe olvidarse, como ya se mencionó en el tema 2, que la operación con números requiere de considerable destreza si uno no cuenta con herramientas auxilares como un ábaco en tiempos pasados, una regla de cálculo hace 50 años o una calculadora de bolsillo en la actualidad.

La representación de los números en un *sistema posicional* —como el decimal de los árabes— fue fundamental para el desarrollo de *algoritmos,* con los cuales se pudieron efectuar las operaciones básicas.

Fueron precisamente los griegos quienes establecieron las primeras propiedades de los números —los pitagóricos estaban particularmente fascinados con estos números—. Ellos llamaron número n, *perfecto* si la suma de sus divisores propios, es decir, los divisores menores que dicho número, resultaba justamente n. Por ejemplo, 6 es perfecto: sus divisores propios son 1, 2 y 3 que suman 6. Pero 5 no es perfecto, ya que sólo tiene un divisor propio que es el uno. Los números que sólo tienen al 1 como divisor propio se llaman *primos.* Por ejemplo, 5 es primo y 6 no lo es.

Los griegos mostraron que cada número positivo se descompone en factores primos y que esta descomposición es esencialmente única, pues diferentes descomposiciones sólo difieren en el orden. Por ejemplo $24 = 2 \cdot 2 \cdot 2 \cdot 3 = 2 \cdot 3 \cdot 2 \cdot 2$. Al hecho de que se puede y que es único, se le conoce como el *teorema fundamental de la aritmética.*

Además, los griegos demostraron que hay un número infinito de primos. Esto lo demostraron por contradicción, es decir, se supone lo contrario y con ello se llega a una contradicción. En efecto, si solamente hubiera un número finito estos primos se podrían poner en una lista y numerarlos como $p_1, p_2, \ldots, p_N$. Entonces, se formaría el número:

$$x = p_1 \cdot p_2 \cdot \ldots \cdot p_N + 1.$$

Este número no puede ser divisible entre p_1 porque si no, también $x - p_1 p_2 \ldots p_N = 1$ sería divisible entre p_1, lo cual no puede ser ya que $p_1 > 1$. De manera similar, se muestra que x tampoco es divisible por ninguno de los primos $p_2, \ldots, p_N$. En conclusión, x no es producto de primos, lo que contradice el teorema fundamental de la aritmética. Hemos encontrado una contradicción.

Una revisión minuciosa de la argumentación muestra que no se ha cometido ningún error y, por lo tanto, no puede ser cierta la suposición de que hay sólo un número finito de primos.

4.5.2 El largo nacimiento de la notación algebraica moderna

Desde los más antiguos documentos que reportan la enseñanza de las matemáticas, destacan problemas formulados para buscar algún número que cumpla con cierta propiedad. Uno de estos documentos es el papiro Rhind, que se considera una colección de problemas con los cuales se enseñaban las matemáticas en Egipto, 2 000 años antes de nuestra era. El problema 26 del papiro Rhind es:

> Una cantidad y una cuarta de sus partes son 15, ¿cuál es esta cantidad?

Figura 4.28 El papiro de Ahmes, *también conocido como* de Rhind, *es un documento de contenido matemático que data aproximadamente del 1680 a.C. y se le atribuye al escriba Aahmes. Fue encontrado en el siglo XIX; entre las ruinas de Luxor, en Egipto. Actualmente, reside en el Museo Británico de Londres | © Latin Stock México.*

Este problema se podría modelar hoy con la siguiente ecuación lineal $x + \frac{x}{4} = 15$. Sin embargo, ésta es una notación moderna. En India, usaban la palabra *ya* —de *yavat tavat*— para llamar a la incógnita principal, mientras que los nombres de colores se utilizaban para denotar otras variables. Al-Juarismi usaba "shai" para denotar incógnitas, que después se tradujo al latín en "res" o "causa" y que, en italiano, se transformó en "cosa". Los alemanes copiaron el sonido y lo escribieron "coss" y, por ello, durante cierto tiempo a los matemáticos se les conocía como los "cosistas". Hasta aquí se empleaban palabras y sílabas cuyo uso variaba en el tiempo y el espacio, es decir, se utilizaba el lenguaje normal para expresar problemas o ecuaciones. Por ejemplo, Leonardo de Pisa escribió en 1202:

> El cubo y siete cosas menos 5 cuadrados es igual a la raíz de la cosa más 6.

Para no confundirnos, aclaramos que se trata de la siguiente ecuación:

$$x^3 + 7x - 5x^2 = \sqrt{x+6}$$

En un manuscrito de 1485, Nicolas Chuquet usó la notación 5^3 para expresar lo que hoy escribiríamos como $5x^3$.

En 1590, François Viète, un francés, usó consonantes para constantes y vocales para incógnitas, pero escribía, por ejemplo, "a cubum" para lo que hoy denotaríamos como a^3. Descartes, en 1637, combinaba estas ideas aunque con cambios: usaba las últimas letras del alfabeto para denotar incógnitas y las primeras para constantes. Él hubiera escrito:

$$x^3 + 7x - 5x^2 \propto \sqrt{x+6},$$

ya que el símbolo de igualdad que usamos en la actualidad todavía no se había establecido.

Aunque la matemática no se reduce a la notación, ésta sí clarifica y ayuda al pensamiento. Más aún, posibilita conceptos completamente nuevos. Si escribimos $x^2, x^3, x^4, x^5, \ldots$ para expresar las potencias de x, podemos estar más tentados de pensar en expresiones como x^{-2} o $x^{\frac{1}{2}}$, es decir, la notación puede sugerirnos una *generalización,* cosa que no pueden los nombres "censo", "cubus", "censo de censo" y "primo relato" que usaba, por ejemplo, Niccolo Tartaglia y se basa en una tabla amplia de Luca Pacioli de finales del siglo XV.


LA PRIMA PARTE DEL GENERAL TRATTATO DI NV-MERI, ET MISVRE DI NICOLO TARTAGLIA,
NELLAQVALE IN DIECISETTE LIBRI SI DICHIARA TVTTI GLI ATTI OPERATIVI, PRATICHE, ET REGOLE NECESSARIE NON SOLA-mente in tutta l'arte negotiativa, & mercantile, ma anchor in ogni altra arte, scientia, over disciplina, doue intervenghi il calculo.
MALIGNITA'
FORTEZZA
NOVAE NON FVO
CON LI SVOI PRIVILEGII.
In Vinegia per Curtio Troiano de i Nauò.
M D LVI.
LIBRO
Figura ovale

Figura 4.29 El libro Trattato di Numeri et Mesure *de Niccolo Fontana Tartaglia fue escrito en 1556 y es un tratado extenso en el que se encuentra buena parte de la notación usada en la época.*

Además, la notación sugiere nuevas leyes de manera más fácil. Por ejemplo, $x^2 \cdot x^3 = x^5$ se hubiera escrito como "ce.cu–p°.r°" ya que se usaban abreviaturas como "ce" para "censo", "p°.r°" para "primo relato", o la línea para indicar la igualdad. Pero era imposible formular la generalización: $x^a \cdot x^b = x^{a+b}$. De esta manera se ve que la notación sí influye en las ideas, en cómo hacer matemáticas y en cómo una notación adecuada propicia el desarrollo, mientras que una complicada lo inhibe. La notación algebraica se estabiliza más o menos a la mitad del siglo XVII y se puede decir que es una de las grandes aportaciones a las matemáticas y a la ciencia, en general.

4.5.3 Ecuaciones lineales, cuadráticas, cúbicas y de cuarto grado

El método de resolución que se indica en el papiro Rhind fue enseñado hasta principios del siglo xx. En un documento conservado del segundo siglo antes de nuestra era se considera el siguiente problema:

> Para 3 gavillas de una buena cosecha, 2 de una mediana cosecha y una de una mala cosecha se reciben 39 Tou. Para 2 gavillas de una buena cosecha, 3 de una mediana cosecha y una de una mala cosecha se reciben 34 Tou. Para 1 gavilla de una buena cosecha, 2 de una mediana cosecha y 3 de una mala cosecha se reciben 26 Tou. ¿Cuánto se recibe para cada gavilla de una buena, de una mediana y de una mala cosecha?

Hoy diríamos que en este problema se trata de plantear un sistema de ecuaciones lineales. El método que se indica en el documento chino es el que hoy conocemos como *algoritmo de Gauss;* no obstante Gauss vivió dos milenios después de Fang Cheng, el autor del documento. Pero la historia rara vez hace justicia al atribuir los descubrimientos y tiende a dar a quien ya tiene.

Problemas como los que se han presentado sugirieron la invención de variables, una hazaña que tardó varios siglos y tuvo muchos altibajos. Los matemáticos árabes, en particular después de la obra de Al-Juarismi, introdujeron propiamente el *álgebra elemental* como la conocemos hoy día, comprendiendo el concepto de ecuación algebraica y de polinomios, la solución numérica de ecuaciones y la construcción geométrica de soluciones.

Al-Juarismi nació alrededor de 790, vivió y trabajó en Bagdad y murió, aproximadamente, de sesenta años. Su libro contenía una clasificación de ecuaciones cuadráticas y cómo resolverlas de manera geométrica. Los tipos de ecuaciones que trataba Al-Juarismi se leen en la notación moderna como:

1. $ax^2 = bx$
2. $ax^2 = b$
3. $ax = b$
4. $ax^2 + bx = c$
5. $ax^2 + c = bx$
6. $bx + c = ax^2$

Hay que notar que sólo consideraba soluciones y coeficientes positivos. Luego, dividió estas ecuaciones entre el coeficiente a, con el fin de obtener "formas normales". Para cada uno de estos tipos deriva una fórmula de solución usando ejemplos y argumentaciones geométricas. Por ejemplo, la argumentación geométrica para resolver la ecuación $x^2 + q = px$ se basa en la figura 4.30.

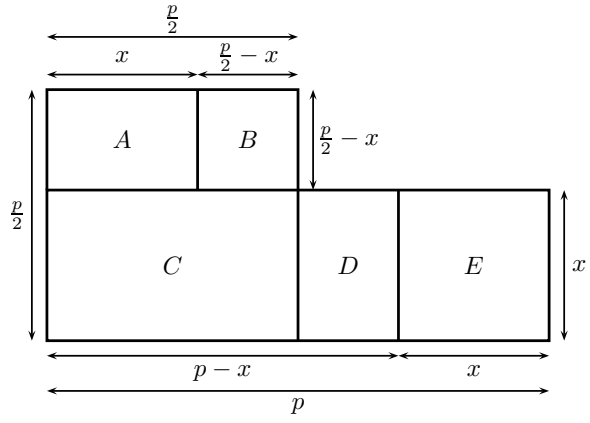


Figura 4.30 La figura geométrica usada en la demostración de la fórmula de solución por Al-Juarismi.

En la figura se ve que el rectángulo C tiene la mitad de ancho que CDE juntos. Por ello, $C = D + E$ si con los mismos símbolos se denotan las áreas. Además $C + D = (p - x)x = q$, según la ecuación que se quiere resolver, que es $x^2 + q = px$. Finalmente, se observa que A y D son congruentes. Entonces, podemos calcular:

$$B = (\tfrac{p}{2})^2 - A - C = (\tfrac{p}{2})^2 - D - C = (\tfrac{p}{2})^2 - q,$$

es decir,

$$(\tfrac{p}{2} - x)^2 = (\tfrac{p}{2})^2 - q$$

lo que ahora se puede resolver fácilmente pues:

$$\tfrac{p}{2} - x = \sqrt{(\tfrac{p}{2})^2 - q}$$
$$x = \tfrac{p}{2} - \sqrt{(\tfrac{p}{2})^2 - q}$$

De esta manera, se comprende mucho mejor el título de la obra que se traduce como "Compendio de cálculo por el método de completado y balanceado", ya que el argumento anterior muestra cómo se balancean los términos para lograr expresar la incógnita en un cuadrado.

Después de Al-Juarismi hubo muchos otros matemáticos árabes, de los cuales destacaron, principalmente, el matemático persa Al-Karaji, quien usó los monomios $x, x^2, x^3 \ldots$ y $\frac{1}{x}, \frac{1}{x^2}, \frac{1}{x^3}, \cdots$ sin representación geométrica, y después Omar Khayyam, quien clasificó y resolvió la ecuación cúbica al usar intersecciones de secciones cónicas.

Los conocimientos de los árabes se difundieron poco a poco en Europa, durante los siglos XII, XIII y XIV; se sabe también que a finales del siglo XV las principales obras de griegos como Euclides, Apolonio, Diofanto, Herón y Arquímedes, entre otros, ya se conocían. Fue precisamente en esta época cuando Tartaglia, Cardano y Ferrer desarrollaron la solución general de la ecuación algebraica de tercer y cuarto grado que detallamos en la sección 4.11.

Los ejemplos mostrados en esta sección ponen de manifiesto que el álgebra básica se desarrolló en diferentes regiones del planeta, de manera independiente.

Lo que sigue es una de las historias más asombrosas y que sólo ocurrió una vez, en Europa.

4.5.4 Nuevos horizontes

El cambio al siglo XVI trajo dos grandes personajes: Galileo Galilei, quien liberó la ciencia de las ataduras de la escolástica y dio a las matemáticas su lugar moderno, y René Descartes, quien inauguró el pensamiento autónomo frente a la fe e inició una unión entre las dos ramas principales de las matemáticas en aquel tiempo, es decir, la geometría y el álgebra. Esta unión se conoce hoy como geometría analítica y consiste en el uso de coordenadas en el plano y el espacio y en la representación de objetos geométricos mediante ecuaciones. La geometría analítica tuvo mucho impacto tanto para la geometría como para el álgebra.

En 1770, aparece el libro *Instrucciones completas para el álgebra* de Leonhard Euler, que reúne la aritmética y la teoría de ecuaciones, y está escrito con tanta claridad que fue reeditado múltiples veces. En él, ya aparecen los números imaginarios, pero no fue sino

hasta unos veinte años después que el concepto de *números complejos,* finalmente, se formalizó.

Más o menos al mismo tiempo, aparece la primera demostración rigurosa del *teorema fundamental del álgebra* que afirma que cualquier ecuación algebraica de grado positivo con coeficientes racionales tiene, al menos, una solución compleja. Es importante resaltar que el teorema sólo afirma la existencia en abstracto sin dar ninguna solución concreta, es decir, el resultado asegura que hay al menos una solución, pero no ayuda a encontrarla. Por ello, la búsqueda de fórmulas para las soluciones cobra aún más importancia.

No obstante que las ecuaciones lineales, cuadráticas, cúbicas y cuárticas podían resolverse con fórmulas, para la ecuación de quinto grado no se conocía ninguna fórmula. Fue un joven sueco, Niels Henrik Abel, quien demostró que dicha empresa es imposible. Vale la pena enfatizar que el teorema fundamental del álgebra afirma que las soluciones existen, mientras que el resultado de Abel implica que dichas soluciones no se pueden encontrar con una fórmula usando sólo las operaciones básicas y raíces, como se mencionó en la sección 4.9.

Figura 4.31 Sello postal fabricado en honor a Leonhard Euler conmemorando los 300 años de su nacimiento.

Sin embargo, la demostración es técnica y hoy se suele usar la teoría de Galois, llamada así por otro joven, el francés Evariste Galois, quien vislumbró una teoría completa para tratar en general la problemática de resolución de ecuaciones algebraicas. Llegamos a un parteaguas en la historia: aunque lo que demuestra la teoría de Galois no es novedoso —Abel ya lo había demostrado unos años antes—, dicha teoría es la que desató realmente el álgebra en su versión moderna.

4.5.5 El álgebra moderna

Galois estudia cómo ciertas permutaciones de las raíces de un polinomio, es decir, las soluciones de una ecuación algebraica, corresponden a ciertos *campos intermedios* —véase sección 4.9—. En lo anterior, hay dos nociones abstractas: el *grupo* y el *campo.* La formalización de estas nociones fue tardía.

Hoy se define al grupo como un conjunto G junto con una *operación binaria:* $\mu : G \times G \longrightarrow G$, es decir, una función del producto cartesiano $G \times G$ al conjunto G con ciertas propiedades. El *producto cartesiano* $G \times G$ es, por definición, el conjunto de pares (a, b) de elementos a, b de G. En otras palabras, la función μ asigna a cada par de elementos un elemento de G.

Un ejemplo de lo anterior es la suma de números enteros que asigna a cada par de números su suma; en este caso se denota $\mu(a, b) = a + b$.

La multiplicción y la exponencial son también otras funciones y se denotan $\mu(a, b) = ab$ y $\mu(a, b) = a^b$, respectivamente.

Para que G sea un grupo debe satisfacerse la *propiedad de asociatividad,* que se escribe de la siguiente manera si $\mu(a, b) = ab$: para todo a, b, c en G se tiene $(ab)c = a(bc)$. La segunda propiedad es en la que se pide que exista un *elemento neutro* e, esto es, un elemento que satisfaga $ea = a = ae$, para todo elemento a de G. La tercera y última propiedad que se pide es que para cada elemento a de G existe un *inverso multiplicativo,* es decir, un elemento b tal que $ab = e = ba$, donde e es el elemento neutro.

Debe observarse que no se requiere que $ab = ba$ para todo a y b de G. Si esta propiedad también se satisface, se dice que el grupo es *conmutativo,* o también se dice que el grupo es *abeliano,* en honor a Abel, quien se dio cuenta de la importancia de la conmutatividad para poder resolver una ecuación con radicales.

Algunos ejemplos de grupos conmutativos son los números enteros con la adición o los números racionales distintos de cero con la multiplicación como operación binaria. Un grupo no conmutativo es el conjunto de todas las permutaciones del conjunto $I = \{1, 2, 3, 4\}$, es decir, todas las funciones biyectivas de I en I. La operación binaria en este caso está dada por la composición de funciones.

La segunda noción es la de *campo.* Un campo es —en el lenguaje moderno— un conjunto con dos operaciones que suelen denotarse como suma y multiplicación, tal que la suma es asociativa, conmutativa, admite elemento neutro 0 e inversos, mientras que la multiplicación es asociativa, conmutativa, admite elemento neutro 1 y cada elemento no cero tiene inverso multiplicativo. Ejemplos de campos así son los números racionales, los reales o los complejos, pero también campos como $(\sqrt{2})$, esto es, los números reales que tienen la forma $a + b\sqrt{2}$ con a y b racionales.

Hoy día hay teorías completas que se llaman *teoría de grupos* y *teoría de campos,* y que se dedican al estudio de cada una de estas dos nociones. Una vez que se han formulado dichas nociones es fácil obtener nuevas estructuras.

Por ejemplo, un *anillo con unidad* se define en forma muy similar a un campo: es un conjunto con dos operaciones binarias denotadas como suma y multiplicación, con la única diferencia de que en la multiplicación no se pide ni conmutatividad ni existencia de inversos. Ejemplos de anillos son los polinomios $[x]$, con coeficientes racionales en la incógnita x, que se pueden sumar y multiplicar. El polinomio constante 1 es la unidad multiplicativa, pero ningún polinomio de grado positivo tiene un inverso multiplicativo.

Aunque la noción de anillo es una generalización de la de campo, la *teoría de anillos* es muy distinta a la teoría de campos, pues se requieren herramientas diferentes para trabajarlas. Una noción importante para la teoría de anillos es la de un *ideal.* Un ideal I de un anillo A es un subconjunto $I \subseteq A$ que contiene el 0, es cerrado bajo la suma, es decir $i_1 + i_2$ es un elemento de I para cada i_1, i_2 de I. Además, el inverso aditivo de cada elemento de I es, de nuevo, un elemento de I y, con respecto a la multiplicación, se tiene que para cada a de A y cada i de I, también los elementos ai, ia pertenecen a I. Para campos, el concepto de ideal no es interesante ya que siempre hay sólo dos ideales: todo el campo y $\{0\}$, el ideal cero.

Para anillos e ideales hay muchos ejemplos naturales. Por ejemplo, los polinomios de grado mayores o iguales que 3 forman un ideal de $[x]$, o más general, para cada polinomio f de $[x]$ los múltiplos de f, es decir, los polinomios gf —donde g es cualquier polinomio— forman un ideal que se denota por (f).

Un ideal se llama *primo* si satisface la siguiente propiedad: si un producto a_1a_2 pertenece a I, entonces, al menos uno de los dos factores pertenece a I. Veamos esto en el ejemplo del anillo Z de los números enteros con la suma y multiplicación usual. El ideal (5) consis-

te en todos los múltiplos de 5 y es un ideal primo: si $a_1 a_2$ es un múltiplo de 5, entonces a_1 o a_2 tienen que ser múltiplos de 5. En cambio, el ideal (6) no es primo ya que el producto 2×3 está en (6), pero ninguno de los factores está en (6). Se ve que el ideal (f) es primo exactamente cuando el número f es primo o es cero.

La formalización de estos conceptos y la abstracción del contexto de las ecuaciones tardó más de un siglo y culminó a principios del siglo xx. En este proceso incidieron muchos matemáticos, como el inglés William Rowan Hamilton y los alemanes Ernst Eduard Kummer, Emmy Noether —una de las mujeres matemáticas más destacadas— y Emil Artin, entre muchos otros.

En 1930, el matemático Van der Waerden publicó un libro llamado *Álgebra moderna,* en donde, por primera vez, se definen todos estos conceptos —y muchos más— en el lenguaje que se sigue usando hasta hoy y que se basa en la teoría de conjuntos, establecida a principios del siglo xx.

4.5.6 Álgebra lineal

Hemos dejado la solución de ecuaciones lineales desde la mención de que los chinos ya conocían un método general que se llama hoy el *método de Gauss,* lo que puede interpretarse correctamente como que en estos 2 000 años no hubo avance sustancial. La situación cambia radicalmente durante la segunda mitad del siglo xix. Matemáticos como Arthur Cayley y Hermann Günther Grassmann lucharon por la noción de *espacio vectorial,* que se hace respecto a un campo k. Hoy, se define un espacio vectorial como un grupo abeliano V junto con una función:

$$k \times V \to V \tag{9}$$

tal que se satisfacen una serie de propiedades, que relacionan la suma en V con la suma y el producto en k. Por ejemplo, se quiere que:

$$(a+b)v = av + bv$$
$$a(v+w) = av + aw$$

para todo a, b de k y todo v, w de V. Estas ecuaciones se ven como leyes de distributividad. Las otras propiedades que deben satisfacerse son:

$$(ab)v = a(bv)$$
$$1_k v = v$$
$$0_k v = 0_V$$

para todo a, b de k y todo v de V. La primera se parece a una asociatividad, la segunda establece que el elemento neutro 1_k del campo k satisface una propiedad similar en V respecto a la multiplicación con escalares (9), mientras que la tercera relaciona los elementros neutros aditivos de k y de V.

Si tomamos el ejemplo de Fang Cheng, entonces el sistema de ecuaciones:

$$3x + 2y + 1z = 39$$
$$2x + 3y + 1z = 34$$
$$1x + 2y + 3z = 26$$

expresa una relación entre las tres variables x, y, z. En el siglo XIX se empezó a pensar muy diferente sobre ello. Al considerar la matriz de coeficientes:

$$M = \begin{bmatrix} 3 & 2 & 1 \\ 2 & 3 & 1 \\ 1 & 2 & 3 \end{bmatrix}$$

se considera una transformación $R^3 \to R^3$, del espacio de tres dimensiones en sí mismo. Aquí, R denota al conjunto de números reales que debemos pensar en su forma geométrica como recta real, mientras R^3 denota $R \times R \times R$, que es el producto cartesiano de tres números reales.

También se formalizó la noción de *dimensión* como el mínimo número de coordenadas que requiere en un espacio vectorial para describir cualquier punto en él, en completa concordancia con nuestro sentido común para la dimensión en la geometría. Es decir, la dimensión es el mínimo número N, tal que existen elementos $v_1, \ldots, v_N$ de V con la propiedad de que cualquier elemento w de V se puede escribir como una *combinación lineal* de $v_1, \ldots, v_N$ con coeficientes en el campo k, esto es, que existen $a_1, \ldots, a_N$ de k, tal que:

$$w = a_1 v_1 + a_2 v_2 + \ldots a_n v_N.$$

Se puede demostrar que estos coeficientes están determinados de manera única. Los números $a_1, \ldots, a_N$ son las *coordenadas* de w en la *base* $v_1, \ldots, v_N$. Por ello, basta fijar el *vector* $(a_1, \ldots, a_N)$ en k^N y lo que se obtiene es una identificación —una *biyección,* en términos matemáticos— de los elementos de V con los elementos de k^N mediante las coordenadas en la base $v_1, \ldots, v_N$.

Con el álgebra lineal se crean fuertes herramientas para tratar todo tipo de fenómenos lineales en el álgebra y aquellos fenómenos geométricos que son lineales, es decir, que sólo tienen que ver con *líneas* y *planos* y su generalización en dimensiones mayores.

4.5.7 Algebraización

Con la teoría de conjuntos, véase sección 4.3, y las estructuras algebraicas a la mano, el siglo XX emprendió una formalización de las matemáticas —o mejor dicho—, una *algebraización.* Amplias áreas de las matemáticas se revisaron y se basaron en el álgebra, y se puede decir que, en algunos casos, la euforia era tal que se pecó de exageración.

Veamos aquí en qué sentido se "algebraizó" la geometría, de nuevo y una vez más, después de la geometría analítica de Descartes. Lo que surgió a partir de entonces se llama hoy *geometría algebraica* y debe bastarnos como ejemplo. Mencionamos que existe también una topología algebraica, que un grupo de matemáticos franceses —que publicaron bajo el seudónimo de Nicolas Bourbaki— algebraizaron el análisis y que la teoría algebraica de los números es de lo más moderno y más difícil.

En la geometría algebraica no sólo se tratan planos y rectas, sino *curvas* y *superficies,* aunque éstas se piensen dentro de algún espacio vectorial; por ejemplo, el C^n, es decir, el espacio de números complejos de dimensión n. Por consiguiente, los objetos que interesan son curvos dentro de un espacio bien conocido. Se exige que la curva o la superficie estén definidas por una o varias ecuaciones algebraicas. Por ejemplo:

$$x_1^2 + x_2^2 + x_3^2 - 1 = 0$$

describe una esfera en el espacio tridimensional, con centro en el origen $(0, 0, 0)$ y radio 1, mientras que el sistema de ecuaciones:

$$\begin{aligned} x_1^2 + x_2^2 + x_3^2 - 1 &= 0 \\ x_3 &= 0 \end{aligned}$$

describe una circunferencia con centro en el origen y radio 1, que está en el plano de coordenadas x_1, x_2.

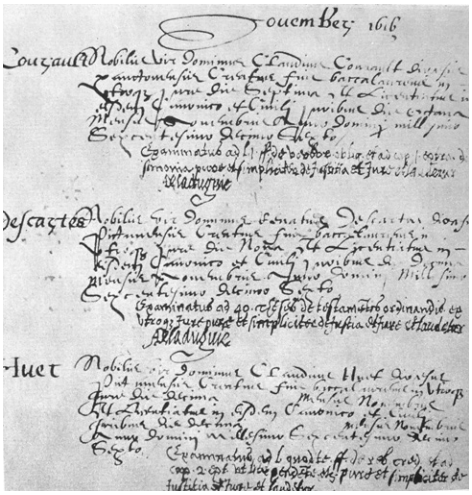


Figura 4.32 Registro de graduación de 1616 de René Descartes (1596-1650), en el Collège Royal Henry-Le-Grand, La Flèche. Descartes fue un filósofo y matemático francés, reconocido por sus grandes contribuciones a las ciencias naturales.

En lo que sigue, sea T un conjunto finito de polinomios en $C[x_1, \ldots, x_N]$, o lo que es lo mismo, T es un conjunto de polinomios en las incógnitas $x_1, \ldots, x_N$ con coeficientes en el campo de los números complejos C. Con $Z(T)$ denotamos al conjunto:

$$Z(T) = \{x \in C^N \mid f(x) = 0, \text{ para cada } f \text{ en } T\},$$

es decir, el conjunto $Z(T)$ consiste en las raíces comunes de todos los polinomios en f. Se escribe $Z(f)$ si $T = \{f\}$.

Inversamente, para cada conjunto algebraico Y de C^N definimos:

$$I(Y) = \{f \in C[x_1, \ldots, x_N] \mid f(y) = 0, \text{ para cada } y \text{ de } Y\}.$$

Se observa que si $T_1 \subseteq T_1$, entonces $Z(T_1) \supseteq Z(T_2)$ y si $Y_1 \subseteq Y_2$ entonces $I(Y_1) \supseteq I(Y_2)$. Esto es que las dos funciones Z, I inviertan la contención de conjuntos, que se parece a la correspondencia de Galois; véase sección 4.9. Además, Y siempre es un subconjunto de $Z(I(Y))$ y se puede mostrar que $Y = Z(I(Y))$ se cumple en general.

Es fácil ver que $I(Y)$ es un ideal de $C[x_1, \ldots, x_N]$ y, por ello, no necesariamente se tiene $T = I(Z(T))$, ya que no cualquier subconjunto T de $C[x_1, \ldots, x_N]$ es un

ideal. Además, $I(Y)$ es un ideal *radical,* eso quiere decir que si f^r pertenece a $I(Y)$, entonces f pertenece a $I(Y)$.

Los subconjuntos de C^N que son de la forma $Z(T)$ para algún T se llaman *conjunto algebraico.* Se dice que un conjunto algebraico es *irreducible* si no puede escribirse como la unión de dos subconjuntos propios que son algebraicos. Por ejemplo, $Z(x_1x_2)$ consiste en los dos ejes de coordenadas $x_1 = 0$ y $x_2 =$ y no es irreducible dado que $Z(T) = Z(x_1) \cup Z(x_2)$. Sin embargo, los ejemplos de la esfera y la circunferencia que dimos anteriormente, sí son irreducibles.

Un *teorema de Hilbert* afirma que hay una correspondencia entre los conjuntos algebraicos de C^N y los ideales radicales de $C[x_1, \ldots, x_N]$. Además, bajo esta correspondencia, los conjuntos algebraicos irreducibles corresponden a los ideales radicales que son primos.

Esto muestra que el estudio de los objetos geométricos que están dados por ecuaciones algebraicas se puede traducir a un lenguaje algebraico. Se critica a la geometría algebraica porque ya no es geométrica, sin embargo, algebraizaciones como ésta han tenido un fuerte impacto en el avance de las matemáticas.

4.5.8 El gran proyecto de clasificación de los grupos simples

Después de su nacimiento, los grupos se hicieron abstractos, es decir, los matemáticos empezaron a estudiarlos sin pensar específicamente en grupos de Galois y en su conexión con la resolución de ecuaciones y permutaciones de raíces. De la misma manera, se empezaron a estudiar los grupos finitos, es decir, los grupos que tienen un número finito de elementos. Como los números enteros mayores que 1 se descomponen en un producto de primos, también los grupos finitos se "descomponen", pero en *grupos simples.* Es decir, así como los primos son la pieza clave para los números, los grupos simples lo son para los grupos.

Lo anterior muestra el gran interés de la comunidad matemática por conocer todos los grupos simples que hay, es decir, de obtener un resultado de *clasificación.* En 1954, en el Congreso Internacional de Matemáticas en Amsterdam, Richard Brauer anunció una idea sobre cómo podría emprenderse esta clasificación, lo cual desató una actividad frenética que tardó más de dos décadas en resolverse e involucró alrededor de cien matemáticos que publicaron sus hallazgos y avances en más de diez mil páginas.

El trabajo concluyó alrededor de 1980, cuando se obtuvo una lista completa de los grupos finitos que son simples. Esta lista consta de tres familias infinitas. Una de ellas es la de los *grupos alternantes:* para cada entero $n \geq 5$, el grupo A_n se forma con las permutaciones que se pueden escribir como composición de un número par de transposiciones; véase sección 4.9. Además, hay 26 grupos *esporádicos,* es decir, grupos que no aparecen en familias, cuyos nombres son muy llamativos: el *Monster* es el grupo esporádico más grande y tiene:

$$808\ 017\ 424\ 794\ 512\ 875\ 886\ 459\ 904\ 961\ 710\ 757\ 005\ 754\ 368\ 000\ 000\ 000$$

elementos. Sin embargo, ya durante la escritura de la demostración surgieron dudas sobre la validez del resultado. La demostración de que éstos son todos los grupos simples es extremadamente difícil y, por ello, Daniel Gorenstein, Richard Lyons y Ron Solomon empren-

dieron una segunda etapa: el proyecto de revisión, que aún no ha concluido aunque ya se publicaron 4 de los 12 tomos planeados para la demostración. Lamentablemente, Gorenstein murió en 1992, situación que frenó el proyecto.

El teorema de clasificación de los grupos simples es un gran logro de la mente humana y completamente único en la historia. Nunca antes los matemáticos se habían reunido para producir una certeza en un estilo casi industrial. El logro es comparable con la gran Pirámide de Giza, que también sólo fue posible por el esfuerzo colectivo. El resultado en sí ha sido muy exitoso: muchas veces que se quiere demostrar una propiedad para grupos finitos, en general se puede mostrar para los simples porque éstos se conocen y, luego, se extiende el resultado a todos los grupos usando el teorema de Jordan-Hölder.

Sin embargo, este ejemplo muestra también la gran dificultad a la cual se enfrentan los matemáticos hoy día: la complejidad de los argumentos puede volverse tan grande que es casi imposible avanzar aún más en esta dirección. Lo anterior es un ejemplo de las limitaciones que existen para el avance científico. Aún más sorprendente es que el grupo llamado *monster* está relacionado con la teoría de cuerdas; esta teoría de la física aún no se ha concluido, pues los cálculos involucrados son extremadamente difíciles. Una vez más se puede apreciar lo que Eugene Wigner llamó "la inexplicable eficacia de las matemáticas": es realmente sorprendente que este lenguaje abstracto sea tan útil en la descripción de las leyes que gobiernan nuestro Universo.

4.6 ¿QUÉ ES LA GEOMETRÍA HOY?

Figura 4.33 Salto de un tigre en pos de una pequeña presa. Aunque no se mida en metros sino en la fuerza muscular que se emplea, hay que calcular distancias para saltar. También se calculan velocidades y con base en ello se deciden trayectorias para escapar de predadores o atrapar presas. De la eficacia de estas decisiones intuitivas depende la supervivencia: en sus cálculos, el tigre también tomó en cuenta el inminente aterrizaje.

Las nociones geométricas elementales residen en algo más primordial que el intelecto humano, pues los animales también las manejan. Con ellos compartimos un espacio —el *espacio físico*—, así como las dificultades, obstáculos y peligros que surgen al moverse en él. De esta manera, la contundencia del espacio que habitamos por medio de las reglas elementales —e implacables— que lo rigen es la fuente de nuestra intuición

geométrica. Sin embargo, no fue tarea fácil precisar esas reglas y eso sí es obra del intelecto humano.

Se puede pensar en la geometría euclidiana como el primer modelo exitoso del espacio físico. Con base en unos cuantos *axiomas* o postulados se construye una teoría en la que se pueden ir demostrando ciertas afirmaciones, llamadas teoremas, por medio de razonamientos lógicos precisos. Con estos axiomas se trata de definir los elementos básicos —pero etéreos— del espacio, que son precisamente los puntos y las líneas, al indicar las relaciones que mantienen. Por ejemplo, con el axioma "por dos puntos pasa una única línea" se expresa la idea de que las líneas son la manera preferente para ir de un punto a otro y que, en ellas, se miden las distancias.

Además de describir con unos cuantos axiomas el espacio que habitamos, Euclides también describe y sienta las bases para estudiar algo mucho más abstracto: el *plano euclidiano,* que tiene sólo dos dimensiones y sus símiles en el mundo real son las superficies lisas como paredes, pisos, papeles y pizarrones. A diferencia de ellos, el plano euclidiano se extiende indefinidamente. Esto es algo que es más fácil de intuir para el espacio: nos sentimos inmersos en él y sabemos por experiencia que nunca veremos sus "límites", pero a los "planos reales" los vemos desde fuera y siempre se acaban. El plano euclidiano es el primer ejemplo de lo que hoy los matemáticos llamamos *espacios.* Sí, en plural, porque hay muchos. Sabemos bien que en el sentido físico no existe el plano euclidiano, que es algo abstracto de lo cual podemos decir muchas cosas y que al estudiarlo obtenemos herramientas para modelar y controlar al espacio físico. Pero también plantea problemas intrínsecos, que aunque parezcan alejados de la realidad, a la vuelta de la historia nos ayudan a entenderla.

4.6.1 El quinto postulado

De los postulados con que Euclides define el plano, el quinto se hizo famoso pues parece no ser tan elemental como los demás y hasta se sospechaba que era teorema, más que axioma. Por más de dos milenios los matemáticos intentaron demostrarlo usando sólo a los otros cuatro, pero no pudieron. Y no pudieron porque no se puede, aunque esto quedó claro y bien establecido sólo hasta el siglo XIX.

Hay varias maneras equivalentes de enunciar al quinto postulado. La más usual es el *axioma de las paralelas:* dada una línea ℓ y un punto P fuera de ella, se puede trazar una única línea que pasa por P y es paralela a ℓ.

Existen dos formas de negar este axioma. La primera es que no existe la paralela y la segunda que por el punto P no pasa sólo una sino muchas paralelas a ℓ. En la primera negación se peca por escasez y en la segunda, por exceso. Ambas suposiciones, al establecerlas como quinto axioma, son válidas y dan lugar a las *geometrías no euclidianas.* La primera negación del quinto postulado equivale a que no existe el paralelismo: cualquier par de rectas se intersecan como sucede en el *plano elíptico,* íntimamente relacionado con la geometría proyectiva. Por otra parte, en el *plano hiperbólico* se cumple que por un punto pasa más de una línea paralela a otra lejana.

Hacia principios del siglo XIX, varios matemáticos, entre los que destacan János Bolyai y Nikolai Ivanovich Lobachevsky, se convencieron de que existía el plano hiperbólico en el sentido matemático —pues aunque fuera algo abstracto, estaba muy bien definido por un sistema de axiomas—. Se podían hacer razonamientos y demostrar teoremas como antes lo hizo Euclides y estos resultados se iban aglomerando con una nueva lógica y elegancia interna. Por ejemplo, se podía demostrar que, en el plano hiperbólico, la suma de los ángulos

internos de un triángulo siempre era menor que π —o 180 grados— pero, además, que lo que le faltaba para π era justamente su área. Entre más chica el área de un triángulo, más se parece a uno euclidiano; pero por el otro extremo, los triángulos hiperbólicos siempre tienen un área menor a π; ¡algo sorprendente y bonito!

Estos últimos resultados se los achacan algunos historiadores a Gauss, pero no quiso publicarlos por temor al descrédito. De hecho, el trabajo de Bolyai y Lobachevsky —que sí publicaron— fue atacado y despreciado por muchos como algo insensato y sin sentido ni fundamentos en nuestra "realidad". Sin embargo, hacia finales del siglo XIX se descubrieron *modelos* del plano hiperbólico dentro de la propia geometría euclidiana y, entonces, no hubo más remedio que aceptar su existencia. El más simple de estos modelos es el de Klein, que consiste en el interior de un disco —sin su frontera— y las líneas son los segmentos o cuerdas de este disco (figura 4.35).

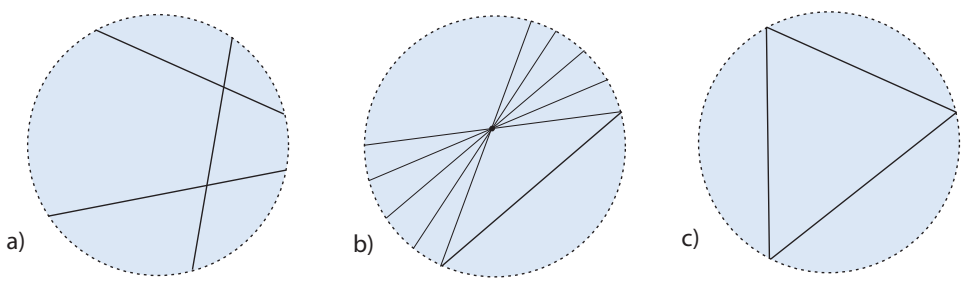


Figura 4.34 a) El modelo de Klein del plano hiperbólico. b) Muchas paralelas a una línea dada pasan por un punto fuera de ella. c) Un triángulo ideal *de área π; se le llama así pues sus vértices no están en el plano hiperbólico.*

En el modelo de Klein del plano hiperbólico, las distancias y los ángulos no corresponden con los euclidianos. Están dados por fórmulas complejas que expresan con precisión cómo dos puntos que nos parecen cercanos pero que están cerca de la frontera, hiperbólicamente están muy lejos. Así que una barra rígida —hiperbólica— se hace chica —euclidianamente— conforme se le acerca al borde, como se verá en los ejemplos de la figura 4.40.

Con la aceptación de estas nuevas geometrías, la matemática se desprende de la realidad inmediata como único objeto de estudio y aclara el papel de la ciencia como proveedor de modelos abstractos para entenderla. De hecho, esta nueva libertad creativa proporciona herramientas para cuestionar nuestra concepción de la realidad, pues Gauss, como buen científico, quiso medir los ángulos de triángulos muy grandes para ver si realmente sumaban π, o lo que es lo mismo, si a gran escala éramos euclidianos o hiperbólicos. Sin embargo, el control en los errores de medición no le permitieron decidir y esa duda sigue en pie.

4.6.2 El toro plano y el plano elíptico

Para describir un modelo del plano elíptico, nos conviene considerar primero un espacio en el que muchos han jugado. Entre los primeros juegos de computadora, hay uno en el que una navecita en la pantalla se mueve cambiando su dirección con las flechas y acelerando —por propulsión a chorro— con la barra espaciadora o alguna otra tecla. Con un impulso, la nave viaja inercialmente en línea recta pero, para que no choque con los bordes, el programa la hace aparecer del lado opuesto y viajando en la misma dirección. Así que si vamos a salir por arriba, aparecemos por abajo y si salimos por la derecha reaparecemos por la izquierda. El *espacio* donde se mueve la nave, y las rocas que hay que destruir a balazos, se llama el *toro plano* y se muestra en la figura 4.35.

No es difícil ver que el toro plano es un espacio de dos dimensiones donde cada punto

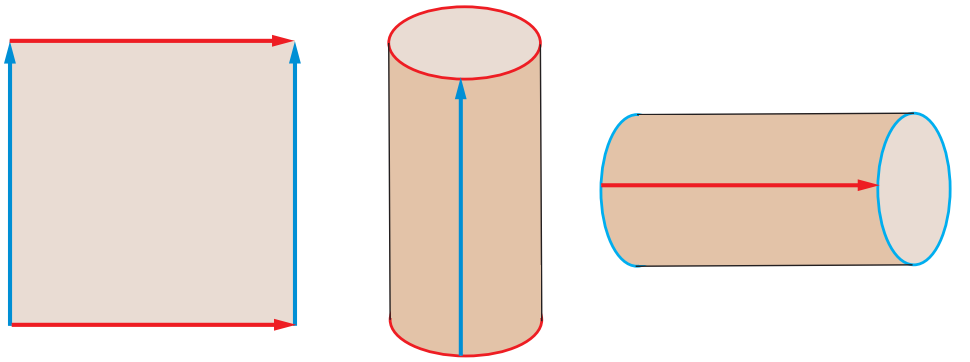


Figura 4.35 Para obtener el toro plano se identifican las aristas opuestas de un cuadrado, de manera que su borde o frontera aparente ya no es tal. En el espacio físico y con el cuadrado de papel, se pueden identificar dos de las aristas para formar un cilindro, pero no las cuatro sin arrugar el papel o cambiarle su geometría.

tiene una pequeña vecindad, igual que si estuviera en el plano euclidiano; no importa dónde ande la navecita, "siente" que está en un plano. Este espacio también presenta ciertas nociones geométricas como distancias, ángulos y trayectorias inerciales o líneas, pero no cumple todos los axiomas de una geometría. Por ejemplo, no es cierto que haya una única línea entre dos puntos, como se muestra en la figura 4.36. Además, no se le puede sumergir en el espacio euclidiano de tres dimensiones sin deformar algunas de sus propiedades geométricas, de lo que hablaremos más adelante. Por el momento, nos interesa para usar la posible familiaridad personal con él como símil de nuestra siguiente definición.

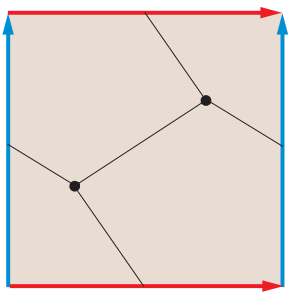


Figura 4.36 En el toro plano, hay varias posibilidades para unir dos puntos por un segmento.

El *plano elíptico* se obtiene a partir de media esfera, al identificar los puntos opuestos en su borde o frontera. Para fijar ideas, pensemos que la media esfera es el hemisferio norte, es decir, que el borde es el ecuador y está horizontal. Si una navecita viaja en el plano elíptico, al llegar al borde va de bajada, pero reaparece por la posición diametralmente opuesta y de subida. También podemos hacer que la nave viaje inercialmente a lo largo del ecuador y, en este caso, sólo la veríamos a la mitad, mientras que la otra mitad estaría justo del lado opuesto y viajando —en la misma dirección y con la misma velocidad—; después de "media" vuelta, la navecita estaría en el mismo lugar, pero sus lados se habrían intercambiado.

Las líneas del plano elíptico son los círculos máximos de la esfera o, mejor dicho, sus mitades; son las trayectorias inerciales que seguiría la nave. Parecen acabar en puntos opuestos del ecuador, pero éstos son en realidad el mismo punto. Ahora sí, tenemos una geometría "a la Euclides" con todas las de la ley, donde por cada par de puntos pasa una única línea y, al seguirla, se obtiene la trayectoria más corta entre ellos. A diferencia de los planos euclidiano e hiperbólico, las líneas en el plano elíptico no son infinitas sino circulares, en el sentido de que al viajar en ellas se regresa eventualmente al mismo lugar. Y se cumple que no hay paralelismo: cualquier par de rectas se interseca en un único punto.

Parecería que los puntos del ecuador —y que éste como línea— son diferentes a los demás. Pero no es así: para verlo, imaginemos que la media esfera es parte de una esfera com-

pleta de la que sólo vemos el hemisferio superior porque está incrustada en el plano opaco del ecuador —algo similar a las bolitas que, a veces, se usan como ratón—. Si la esfera gira libre y lentamente, los puntos que van desapareciendo debajo del plano son remplazados —y ahora representados— por sus correspondientes puntos antípodas que emergen por abajo y en el lado opuesto del ecuador, conforme a la regla de pegado de este espacio. Entonces, vemos al ecuador como un medio círculo cualquiera y se sigue que el plano elíptico es *homogéneo,* pues podemos moverlo rígidamente para llevar un punto hasta cualquier otro al girar la esfera; lo anterior quiere decir que todos los puntos son iguales y tienen pequeñas vecindades equivalentes a las de la esfera o, en otras palabras, que el espacio es localmente esférico y su geometría se "hereda" de la esfera.

Al girar un poco la esfera, observamos que sale un "gajo" o *sector angular,* cuya área es proporcional al ángulo de giro. Si suponemos que el radio de la esfera es 1, entonces su área total es 4π y, por lo tanto, el área del plano elíptico es la mitad o 2π. Así, el área de un sector angular de ángulo α es 2α.

Consideremos un triángulo en el plano elíptico con ángulos internos α, β y γ; y sea A su área. Si pintamos los tres sectores angulares con pintura, se pinta todo; pero el triángulo queda pintado con tres manos, mientras que el resto sólo con una. Al contabilizar la pintura total que se usó, se obtiene:

$$2\alpha + 2\beta + 2\gamma = 2\pi + 2A \ ,$$

es decir, que:

$$\alpha + \beta + \gamma = \pi + A \ .$$

Como el área A es positiva, ello implica que la suma de los ángulos de un triángulo en el plano elíptico es mayor que π, además de que el exceso sobre π, es justo su área. Lo mismo vale para triángulos en la esfera.

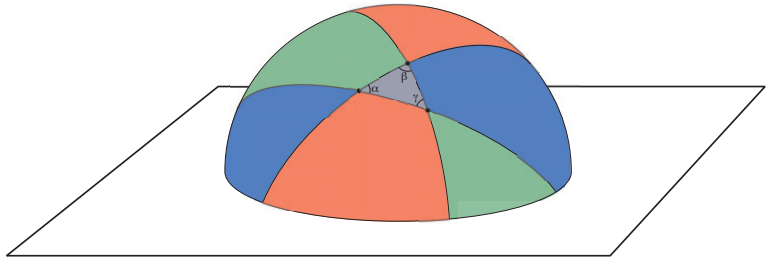


Figura 4.37 Triángulo en el plano elíptico o en la esfera.

4.6.3 El plano proyectivo

En el Renacimiento, los artistas plásticos acabaron de establecer los principios técnicos de la *perspectiva*. La idea básica es que un cuadro representa mejor la realidad si se le concibe como una ventana —a través de la cual se ve el paisaje— en la que se plasma lo que el pintor ve. Geométricamente, lo anterior equivale a *proyectar* al plano del lienzo lo visible —representado con su color— desde un punto fijo: el ojo del pintor que será posteriormente el del obsevador. Entonces resulta que las escalas de las cosas se hacen más chicas conforme están más lejos, y que líneas rectas que en la realidad son paralelas deben dibujarse como líneas que concurren a un punto llamado su *punto de fuga.*

Las líneas reales se proyectan en líneas al lienzo, pues son la intersección del lienzo con el plano que une la línea real con el ojo. Resulta que los puntos de fuga de todos los haces paralelos en un plano forman una línea: su línea al infinito; por ejemplo, la *línea al infinito* del piso es el *horizonte* en el lienzo, al cual se fugan todas las líneas horizontales. Supongamos que queremos dibujar en perspectiva un edificio que se encuentra enfrente y a nuestra izquierda. El plano paralelo a su fachada que pasa por nuestro ojo se interseca con el lienzo en su línea al infinito: cualquier cosa en el plano de la fachada se verá a la izquierda de esta línea, como se aprecia en la figura 4.38 b.

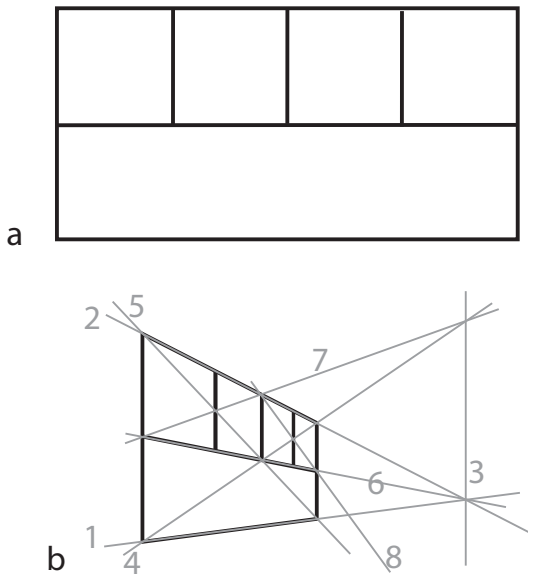


Figura 4.38 a) Una fachada sencilla en un rectángulo de 2×4*. b) Si se dibuja a ojo el contorno del rectángulo —en este caso, con las verticales aún verticales—, la fachada se reconstruye con trazos de líneas auxiliares en el orden de la numeración. Si se alargan las líneas 5 y 8, concurren en un punto de la línea 3 —que es la vertical en la intersección de 1 y 2— y corresponde a la línea al infinito en el plano (a), donde 5 y 8 son paralelas.*

Al proyectar planos a planos desde un punto, resulta natural y teóricamente indispensable considerar que sus líneas al infinito constan de un punto de fuga por cada haz paralelo. Al plano euclidiano con un nuevo punto para cada dirección, en el que concurren las paralelas correspondientes y forman una nueva línea —al infinito— se le llama *plano proyectivo;* tiene un grupo de transformaciones asociado, las *transformaciones proyectivas,* que son las que preservan líneas. Un ejemplo es el de la figura 4.38 a, que manda la fachada hacia su perspectiva en 4.38 b. En este caso, basados en que las líneas van en líneas, sólo se hicieron los trazos necesarios para reconstruir la fachada, pero la transformación está definida en todo el plano, es decir, se puede decidir cuál debe ser la imagen de cualquier punto. En términos de proyecciones y volviendo a pensar en el paradigma del pintor, en la figura 4.39 la fachada es parte de un plano (a) y el lienzo es parte de otro plano (b); al pintor sólo le interesa lo que está enfrente pero, matemáticamente, se considera todo: lo que está atrás del pintor se proyecta del otro lado de la línea al infinito, de (a) vista en (b).

El plano elíptico y el plano proyectivo están en correspondencia natural. Para entender lo anterior, se considera al plano tangente a la esfera en su polo norte y, después, se proyecta desde el centro de la esfera. Las líneas se corresponden, pues en ambos casos son intersecciones de planos por el centro de proyección. Los puntos en el ecuador de la esfera van hacia el punto correspondiente en la dirección de la línea al infinito del plano proyectivo.

De hecho, a estos dos planos se les piensa como el mismo espacio y lo que los distingue es el grupo de transformaciones que se consideran: en el plano elíptico se juega sólo con las transformaciones rígidas y, entonces, se puede hablar de distancias, ángulos y áreas; en el proyectivo se usan todas aquellas transformaciones proyectivas o *colineaciones*, que de hecho son muchas.

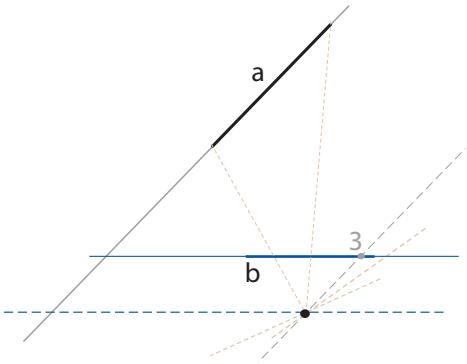


Figura 4.39 Vista superior de la transformación proyectiva de la figura 4.38 considerada como proyección de un edificio real (a) en el lienzo (b). La línea al infinito de (a) se proyecta en la línea 3 en (b), que aquí se ve como un punto.

Aunque aquí hemos hablado de la geometría proyectiva en forma sintética, también se puede trabajar analíticamente —es decir, con coordenadas— y ésta es hoy día una de las herramientas básicas para el despliegue de la realidad virtual en las computadoras.

4.6.4 Los espacios multidimensionales

Se considera a Girard Desargues (1591-1661) el padre de la geometría proyectiva, pues sentó sus bases y demostró los primeros teoremas. Sin embargo, éstos no fueron retomados por otros matemáticos durante los dos siglos siguientes, quizá porque su contemporáneo, René Descartes (1596-1650) atrajo los reflectores de la historia al "coordinatizar" el plano euclidiano. Descartes estableció la correspondencia entre parejas ordenadas de números reales y puntos en el plano euclidiano; con ello, nace la *geometría analítica* y la posibilidad de usar nuevos métodos en la geometría. Más aún, también se abre la puerta para trabajar en otras dimensiones.

Si se describe el plano con parejas de números y el espacio con ternas, al considerar todas las cuartetas de números reales, tendremos el espacio de dimensión cuatro. Estos espacios se denotan R^2, R^3 y R^4 respectivamente. No hay que pararse en cuatro: para cualquier número n se puede definir el *espacio euclidiano de n dimensiones* como el conjunto de n -adas ordenadas de números reales:

$$R^n = \{(x_1, x_2, \ldots, x_n) \mid x_i \in R\} \ .$$

Aunque no se pueda ver o se dude de su existencia en el sentido físico, matemáticamente está ahí, para ser explorado. Por ejemplo, se pueden definir líneas, segmentos y distancias de manera que se extiende naturalmente a los conceptos euclidianos para dimensiones 2 y 3.

A mediados del siglo XIX, Bernhard Riemann amplía, aún más, la noción de "espacio". Ciertos subconjuntos de R^N son especiales, pues localmente se parecen a o se pueden modelar como R^n, para alguna $n < N$; se les llama ahora variedades de dimensión n. Por ejemplo, las superficies lisas —$n = 2$— en el espacio —$N = 3$—, son *variedades de dimensión* 2; o bien, los puntos que equidistan de uno dado en R^N son las esferas de dimensión $N - 1$. Y, por vivir en R^N, heredan una geometría explícita.

Se puede medir la distancia entre dos puntos como la longitud más corta de las trayectorias que los unen dentro de la variedad; y a estas trayectorias más cortas o eficientes para viajar se les llama *geodésicas*. Por ejemplo, si tomamos una esfera en R^3 o en R^N, sus geodésicas son los círculos máximos o intersecciones de planos que pasan por su centro. Riemann observa que ahí hay muchísimos espacios donde se puede hacer geometría, y establece las bases y herramientas para hacerla: lo que ahora llamamos *geometría riemanniana.*

Cuando a principios del siglo xx, Albert Einstein le demuestra al mundo que el Universo debe concebirse como algo de dimensión cuatro y, además, no euclidiano sino con *curvatura* producida por la masa, tenía ya la herramienta matemática para hablar de ello.

Otro personaje importante en la geometría del siglo xix fue Sophus Lie. Estudió los grupos de transformaciones de los espacios euclidianos multidimensionales y observó que son variedades; son grupos continuos llamados *grupos de Lie*. El ejemplo más sencillo es el círculo, pensado como las rotaciones del plano alrededor del origen y, a su vez, como variedad de dimensión 1. Los grupos de Lie han resultado ser una herramienta indispensable para la física del siglo xx, sobre todo para la física cuántica.

Hacia el final del siglo xix, cuando la geometría se convulsionó y expandió en múltiples direcciones, Felix Klein trató de resumir la definición de geometría diciendo que "es el estudio de los invariantes de un espacio bajo un grupo escogido de sus transformaciones". Pongamos un ejemplo de esta idea, retomando el modelo del plano hiperbólico del propio Klein. El espacio dado es el interior del disco —como se muestra en la figura 4.34—; si lo pensamos dentro del plano proyectivo —que contiene al euclidiano—, podemos considerar todas las transformaciones proyectivas que lo dejan en su lugar: éste es el grupo de *transformaciones hiperbólicas*. Entonces, resulta que cualquier punto del interior se puede mover a cualquier otro; que el espacio es homogéneo y que los "invariantes", como distancia y ángulo, se pueden construir a partir del grupo.

Para ver un ejemplo de una transformación hiperbólica, además de las obvias, las rotaciones, supongamos que la fachada de la figura 4.38.a) está inscrita en el círculo del modelo de Klein y queremos trasladarla —como figura hiperbólica— en su línea media horizontal. Basados en que las líneas tangentes al círculo deben ir a líneas tangentes al mismo, pues éste se queda, como conjunto, en su lugar —ésa es la definición de transformación hiperbólica—; una vez fijado el punto donde va el centro de la fachada, con unos cuantos trazos más se obtiene su traslación hiperbólica, como se observa en la figura 4.40.

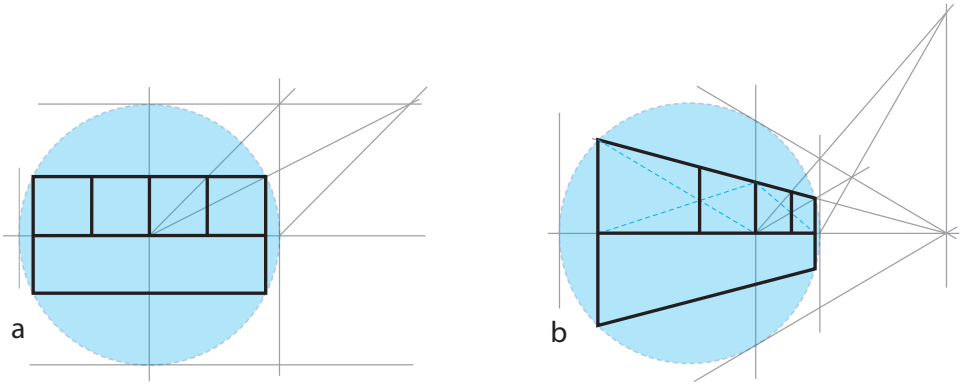


Figura 4.40 Una traslación en la línea hiperbólica que corresponde al diámetro horizontal del disco, deja a las dos tangentes verticales al círculo en su lugar, y por lo tanto manda líneas verticales en líneas verticales.

4.6.5 Topología

En el cambio de siglo del xix al xx, ya se tenía claro que había una geometría aún más libre, donde los espacios se pudieran deformar sin romper su "continuidad" y se sentaron las bases formales para hacerlo: la *topología*. La intuición básica en esta área de las matemáticas es dejar de lado la noción rígida de distancia, pero mantener la de vecindad de los puntos, cambiar la cercanía estricta y rígida por una más laxa y flexible.

Veamos el ejemplo del toro plano. Lo teníamos como un cilindro de papel en R^3 —figura 4.35—, donde había que identificar sus dos extremos circulares —esto es, unirlos punto por punto—; no nos atrevimos a hacer tal cosa, pues el papel o la pantalla de la compu-

tadora regían su geometría, importante en ese momento. Pero si lo imaginamos de un material elástico —idealmente flexible— como espacio topológico, podemos doblarlo y deformarlo, poco a poco, hasta identificar las dos bocas y obtener la superficie de una dona o una llanta, en la cual todas las identificaciones prescritas ya están hechas, según muestra la figura 4.41.

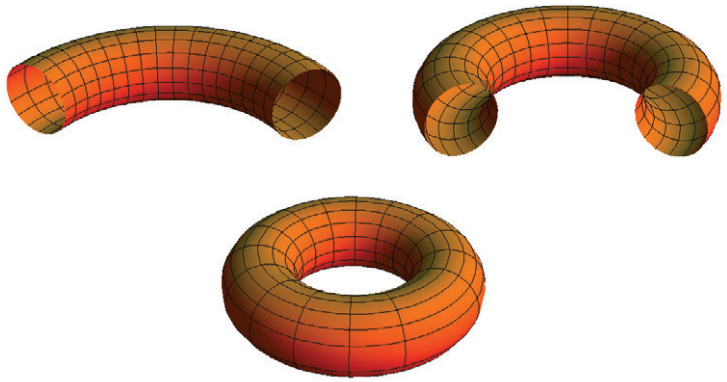


Figura 4.41 Al identificar las dos bocas de un cilindro, se obtiene el toro.

Se le llama el *toro,* sin apellidos. El toro plano es un toro topológico con una estructura extra: la variedad riemanniana que le da rigidez. La topología es algo más elemental o general, pero llegar a ella es, técnicamente, más complicado. Por eso, históricamente se dio después. Otro ejemplo indicativo es un cuadrado y un disco en el plano, donde es fácil concebir una deformación gradual de uno en el otro: aunque geométricamente sean distintos, topológicamente son equivalentes.

A la topología, que estudia los espacios en esta generalidad plástica, se le puede considerar como "la geometría del siglo xx", aunque tal nombre es injusto con la gran cantidad de áreas especializadas en las que se ha desarrollado la geometría.

Uno de los problemas que rigió el desarrollo de la topología en el siglo xx fue planteado por Henry Poincaré en 1905 y se le conoce como la *conjetura de Poincaré.* Ésta tiene que ver con una caracterización elegante de la esfera de dimensión 3, entre todas las variedades de dimensión 3 —o 3-variedades—. En la década de los ochenta, William Thurston demostró que la conjetura de Poincaré sería consecuencia de otra conjetura aún más osada: la de *geometrización,* que dice, a grandes rasgos, que cualquier 3-variedad topológica se puede partir en pedazos, cada uno modelado sobre una de 8 posibles "geometrías" tridimensionales —un ejemplo en dimensión 2 es el toro plano, que está modelado sobre el plano euclidiano.

En 2002, Grigori Perelman demostró la conjetura de geometrización. Todavía son pocos los matemáticos que entienden dicha demostración a cabalidad, pese a que la idea base sea comprensible. Una variedad riemanniana tiene una manera natural de homogeneizarse o de "acomodarse" y si la dejamos fluir —en su *flujo de Ricci*— tenderá a adquirir una estructura geométrica como la que previó Thurston. Por ejemplo, si empezamos con el toro de la figura 4.42, con la estructura riemanniana o geometría que se hereda de vivir en R^3, los círculos horizontales del centro "querrían" crecer y los de afuera decrecer para parecerse entre ellos, pero los verticales se quedarían rígidos porque ya son iguales. Hacer todo esto a la vez en R^3 es imposible; sin embargo, en abstracto o considerando una dimensión más, en R^4, este toro tendería naturalmente a convertirse plácidamente en el toro plano.

En fin, los matemáticos ya tienen un catálogo de posibles formas que puede tener un universo tridimensional. Toca a los cosmólogos, como alguna vez lo intentó Gauss, decidir cuál se aproxima más al "real".

4.7 ¿CÓMO FUNDAMENTAR LAS MATEMÁTICAS?

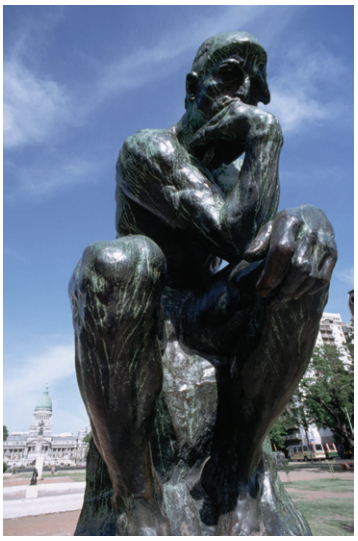


Figura 4.42 El pensador, *realizada en bronce en 1880 por Auguste Rodin, es una de las esculturas más famosas. A pesar de que en su origen Rodin buscaba representar a Dante frente a las puertas del infierno, se considera que simboliza al hombre —en sobria meditación— mientras se debate ante un poderoso dilema interno | © Latin Stock México.*

"Reparte 10 sacos de cebada entre 10 hombres de modo que la diferencia entre la parte de cada hombre y la de sus vecinos sea 1/8 de saco." Este problema es uno más que se plantea en el papiro de Rhind.

Alrededor del año 1700 a.C., Ahmes —un escribano egipcio— elaboró un texto en el que se recopilaban ochenta problemas matemáticos, todos ellos relacionados con situaciones concretas de la vida cotidiana, como la medición de la Tierra o la repartición de bienes. Este papiro fue encontrado en 1858 en Luxor, Egipto, por un anticuario escocés que se llamaba Henry Rhind y en 1863, luego de ser entregado al Museo Británico, se le llamó el *papiro de Rhind.*

Hace miles de años, en las civilizaciones babilonia y egipcia, las matemáticas estaban orientadas, fundamentalmente, a resolver problemas prácticos. No fue sino hasta varios siglos más tarde cuando los matemáticos griegos dieron a las matemáticas un carácter completamente distinto. Para ellos, no bastaba que el conocimiento matemático resolviera problemas concretos que se comprobaran a través de la experiencia, sino que buscaban la generalización de los resultados y su deducción, a partir de otros. Por consiguiente, organizaron las matemáticas como un sistema deductivo donde el único método válido para obtener resultados era el método axiomático.

El ejemplo más importante en las matemáticas griegas es la obra *Los elementos* de Euclides, quien vivió en Alejandría hacia el año 300 a.C. En este libro —compuesto por trece capítulos— Euclides recopiló y organizó en forma deductiva gran parte de la geometría y aritmética conocida en su época.

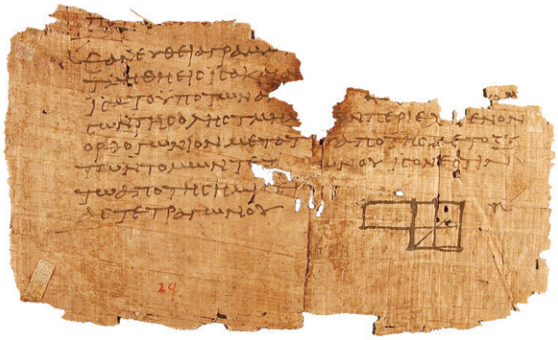


Figura 4.43 Un fragmento de Los elementos, *de Euclides, hallado en Oxirrinco, Egipto. Data del año 100 a. C., aproximadamente.*

Los griegos hicieron de las matemáticas una ciencia teórica y deductiva en la que, para justificar una afirmación, no es suficiente probarla al comprobar que se cumple en casos particulares —aunque sean muchísimos—, sino que es necesario establecerla como un teorema.

Un teorema es una afirmación que se demuestra por medio de la argumentación lógica y de forma deductiva a partir de otros teoremas que, a la vez, ya han sido demostrados. Como las demostraciones no pueden ser indefinidas, se parte de ciertos principios tan claros y evidentes que se pueden asumir sin necesidad de demostrarlos: las definiciones, los postulados y los axiomas. El método axiomático consiste, a grandes rasgos, en lo anterior y con ello los griegos dieron un gran paso hacia la fundamentación de las matemáticas.

Un filósofo griego que no podemos dejar de mencionar aquí es Aristóteles, que nació en el 384 a.C. en Macedonia y que propuso el razonamiento deductivo a partir de silogismos, como el que se muestra a continuación:

- Todos los hombres son mortales,
- Sócrates es hombre,

por lo tanto,

- Sócrates es mortal.

El nombre de Aristóteles estará siempre ligado al de su maestro, Platón, quien a la vez era discípulo de Sócrates y estudioso y admirador de Pitágoras. Es decir, los nombres de los filósofos, matemáticos y físicos griegos están entrelazados en la búsqueda de una estructura que le diera rigor a las formas de pensamiento matemático.

Debemos a Euclides el método axiomático, lo que constituye un gran paso en la historia de las matemáticas, pues introduce la manera moderna o actual de hacerlas.

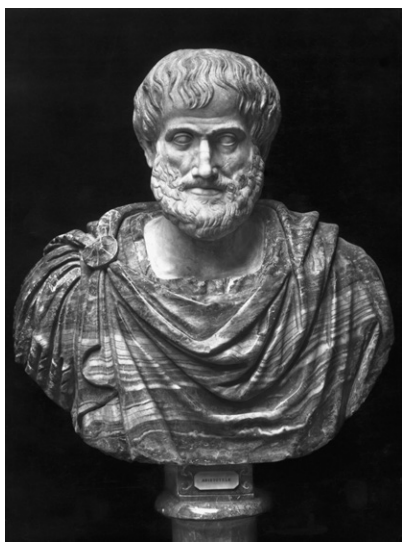


Figura 4.44 Aristóteles | © Latin Stock México.

En un principio, el método axiomático era intuitivo, es decir, los axiomas y los postulados se establecían como evidentes —y por lo tanto no era necesario demostrarlos—, porque así se veían en la realidad, y las deducciones que se hacían a partir de ellos eran, también, bastante intuitivas.

Poco a poco, y con el paso de los siglos, los sistemas axiomáticos se fueron haciendo más abstractos y, entonces, lo importante no era ya que los axiomas y los teoremas se adecuaran a la realidad, sino que formaran *un cuerpo de afirmaciones consistente:* libre de contradicciones. En otras palabras, que mediante los axiomas no pudieran demostrarse a la vez una afirmación y su negación, por ejemplo: "a es igual a b" y, al mismo tiempo, que también se demostrara que "a no es igual a b".

¿Cómo saber que un sistema de axiomas no lleva a contradicciones, es decir, cómo saber que, a partir de los axiomas y de las reglas lógicas que usamos para hacer deducciones, no se pueden deducir un teorema y su negación? Aunque a primera vista pueda parecer increíble, esta pregunta ha ocupado a varios filósofos y matemáticos a lo largo de los siglos.

La sistematización del pensamiento matemático fue dando lugar a preguntas muy importantes —complejas y entrelazadas entre sí—, que hoy día siguen vigentes, acerca de los fundamentos de las matemáticas:

- ¿Las matemáticas se crean o se descubren?
- ¿Cómo se define el concepto de verdad en matemáticas
- ¿Cómo fundamentar correctamente un edificio intelectual tan sofisticado como las matemáticas?

Estos cuestionamientos dieron lugar a un gran debate durante la segunda mitad del siglo XIX y primera del XX.

Para empezar esta historia, explicaremos las dos posturas que existen sobre la esencia de los objetos matemáticos: aquella que sostiene que se construyen a través de procedimientos intuitivos y aquella que, por el contrario, afirma que los objetos matemáticos se conciben o se entienden como parte de una totalidad existente al margen de quien los estudia, es decir, el matemático. Esta segunda postura se conoce como platonismo matemático y ha sido muy relevante en el desarrollo de las matemáticas modernas durante el siglo XX. El afirmar que los objetos matemáticos gozan de una existencia real —análoga en algún sentido a la existencia de los objetos físicos— permite concebir un "universo matemático" independiente de las técnicas o los procedimientos que se usen para estudiarlo.

Son muchos los intentos a lo largo de la historia por mecanizar el razonamiento matemático y probar que estaba libre de contradicciones, es decir, por exhibir fundamentos sólidos sobre los cuales podía erigirse.

El trabajo más importante en este campo comenzó a finales del siglo XIX, cuando los objetos matemáticos empezaron a definirse en términos de conjuntos. Por consiguiente, se intentó definir los números, las funciones y los distintos espacios en términos de conjuntos, con el fin de tener un sistema de reglas para deducir los teoremas a partir de conjuntos pequeños de axiomas y lograr, así, un sistema matemático riguroso.

La teoría de conjuntos se desarrolló con el trabajo de Cantor —un matemático alemán que vivió de 1845 a 1918—, aunque varios matemáticos ya habían trabajado en la orientación "conjuntista" de las matemáticas. La teoría de conjuntos tiene sus raíces en el análisis real, en la teoría de series trigonométricas y, en particular, en la representación de funciones discontinuas a través de series de Fourier y en la caracterización de los conjuntos de puntos de discontinuidad.

Cuando Cantor descubrió, en 1873, que el conjunto de los números reales no es numerable, es decir, que no puede ponerse en correspondencia uno a uno con el conjunto de los números naturales, la concepción que hasta entonces se tenía del infinito cambió radicalmente. Surgieron muchos conceptos nuevos a partir del trabajo de Cantor, como el de conjunto bien ordenado, número ordinal, número cardinal y número transfinito. Cada vez era más fácil caracterizar a los objetos matemáticos en términos de conjuntos.

Pero con la teoría de conjuntos llegaron también las paradojas. A mediados del siglo XIX, ya era bastante grande el grupo de matemáticos y filósofos muy importantes que sostenían una discusión sobre los fundamentos de las matemáticas. El objetivo era eliminar las paradojas y discutir qué era o no aceptable en las matemáticas. Por ejemplo, el *axioma de elección* o la *hipótesis del continuo* tenían que ser discutidos.

Las paradojas nacían de asumir que cualquier colección de objetos —tangibles o pensables— que satisficiera una propiedad era un conjunto. Y así había una correspondencia entre *conjuntos* y *propiedades,* pero Bertrand Russell —filósofo y matemático galés— mostró

que al menos había una propiedad, $X \notin X$, que no determina ningún conjunto si no se quiere caer en una tremenda contradicción.

Una manera de replantear este problema es con la paradoja que ya habíamos revisado anteriormente: "el barbero de un pueblo afeita a todas las personas del pueblo que no se afeitan a sí mismas y sólo a ésas", ¿debe el barbero afeitarse a sí mismo?

Sea *A* el conjunto de todas las personas que no se afeitan a sí mismas, entonces podemos hacer la siguiente argumentación: el barbero pertenece a *A* si y sólo si no se afeita a sí mismo, si y sólo si lo afeita el barbero, si y sólo si se afeita a sí mismo, si y sólo si no pertenece al conjunto *A*. Entonces el barbero es elemento del conjunto *A* si y sólo si no es elemento del conjunto *A*.

Pasa lo mismo al definir el conjunto de todos los conjuntos que no se pertenecen a sí mismos y preguntarse si dicho conjunto pertenece a sí mismo. Con esta paradoja, Russell planteó la imposibilidad de asociar un conjunto a cada propiedad.

Para resolver la paradoja, Russell desarrolló la *teoría de tipos*, en la que la idea básica es establecer tipos o clases que, a la vez, contienen tipos o clases de una jerarquía inferior.

La necesidad de resolver las paradojas tuvo como consecuencia determinar que el camino más sólido para desarrollar las matemáticas era establecer una **axiomática** para la teoría de conjuntos, similar a la que se tenía para la geometría.

Hubo varias propuestas: la de Zermelo-Fraenkel y la de von Neumann-Bernays-Gödel. Por otra parte, el matemático alemán David Hilbert (1862-1943), que en 1899 había publicado una *axiomática* para la geometría euclidiana —más completa que la que hiciera el propio Euclides—, desarrolló los conceptos necesarios para el estudio de las propiedades formales de las teorías axiomáticas en lo que él llamó *teoría de prueba.*

La idea de Hilbert era genial pues se trataba de conseguir una formalización completa de las matemáticas, en particular de la teoría de conjuntos o de la aritmética, de manera que al unir la lógica con una interpretación formal de una teoría matemática concreta, se pudiera formalizar cualquier afirmación de dicha teoría. Esto quiere decir que cualquier afirmación podría formalizarse como una sucesión finita de fórmulas abstractas cuyos símbolos, aislados y sin interpretación, carecerían de significado por sí mismos, evitando así la aparición de paradojas.

Hasta antes del siglo XIX, la intuición era la forma de saber si un sistema axiomático pudiera llevar a contradicciones o no. Euclides formuló sus axiomas para la geometría tomando como modelo la realidad física. Pero, conforme los sistemas axiomáticos se fueron haciendo más abstractos, la intuición y la experiencia dejaron de ser la forma adecuada de verificar si dichos sistemas eran consistentes o no.

Así que una de las cosas más importantes al trabajar con un sistema axiomático es determinar si es consistente o no. Y esta pregunta se hace y se fue haciendo extensiva a todas las matemáticas, a lo largo de la historia: ¿están las matemáticas construidas sobre bases y fundamentos sólidos? ¿Cómo saber que las matemáticas no contienen ninguna contradicción en su seno? Se trataba, a fin de cuentas, de garantizar que cualquier afirmación verdadera pudiera ser demostrada dentro del sistema.

Un personaje muy importante en la sistematización rigurosa de las matemáticas es Giuseppe Peano, matemático y filósofo italiano que vivió de 1858 a 1932. Peano hizo un análisis muy serio del proceso demostrativo de las matemáticas: estableció la formulación axiomática de la aritmética a través de un conjunto de cinco axiomas que hoy se conocen como los axiomas de Peano y que definen a los números naturales, en términos de la teoría de conjuntos. Los axiomas de Peano son:

1. El número 0 es natural.
2. Si n es un número natural, entonces $s(n)$ —el sucesor de n— también es un número natural.
3. El número 0 no es sucesor de ningún número natural.
4. Si para dos números naturales m y n sucede que $(m) = (n)$, es decir, sus sucesores son iguales, entonces $m = n$.
5. Si dada una propiedad p, sucede que:
   a] el número 0 cumple la propiedad p,
   b] cada vez que un natural n cumple la propiedad p, ocurre que $s(n)$ también cumple la propiedad p,
   entonces todos los números naturales cumplen la propiedad p.

Este último axioma se conoce como el principio de inducción.

Durante la segunda mitad del siglo XIX y la primera del XX, el "edificio matemático" fue adquiriendo cada vez más rigor. Fueron muchísimos los matemáticos y filósofos que participaron en la discusión sobre los fundamentos de las matemáticas, con posturas a veces antagónicas y otras veces complementarias.

Y entre estas posturas pueden distinguirse tres que son cruciales y en las que las personas de las que hemos venido hablando pueden ubicarse: el logicismo, el intuicionismo y el formalismo.

Figura 4.45 Alfred N. Whitehead (1861-1947) | *© Latin Stock México.*

El logicismo fue iniciado por Gottlob Frege, un filósofo y matemático alemán que vivió de 1848 a 1925, pero sus máximos exponentes fueron el galés Bertrand Russell (1872-1970) y el inglés Alfred N. Whitehead (1861-1947). El objetivo principal de los logicistas era mostrar que las matemáticas clásicas eran una parte de la lógica, así que la pregunta ¿están las matemáticas libres de contradicciones? se replanteaba como ¿está la lógica libre de contradicciones? Russell y Whitehead escribieron, entre muchas otras obras, *Principia Mathematica,* y en ella intentaron mostrar que todas las matemáticas conocidas hasta ese momento se podían derivar de un sistema axiomático que ellos proponían en sus *Principia.* Sin embargo, nunca pudieron probar que la lógica estaba libre de contradicciones.

El intuicionismo comenzó alrededor de 1908 con un matemático holandés llamado Luitzen Egbertus Jan Brouwer (1881-1966). Mientras que los logicistas jamás habían dudado de la solidez de las matemáticas, los intuicionistas cuestionaron —desde el principio— las matemáticas clásicas y propusieron reconstruirlas por completo, partiendo de cero.

El formalismo existía ya en el siglo XIX, pero, sin duda, su exponente más importante es David Hilbert. Los formalistas propusieron que era necesario formalizar todos los sistemas axiomáticos.

Es interesante ver que, tanto los logicistas como los formalistas, intentaron formalizar todas las ramas de las matemáticas por diferentes razones. Los primeros querían usar la formalización para mostrar que ese campo pertenecía a la lógica; los segundos, para probar matemáticamente que en esa rama no había contradicciones.

Russell y Whitehead construyeron un sistema lógico intentando que de él pudiese derivar toda la matemática: ése fue el sistema que expusieron en sus *Principia.* ¿Era posible deducir todas las verdades matemáticas a partir de ese sistema? ¿Cómo podían tener la certeza de que ese sistema era consistente, es decir, que no podían deducirse contradicciones a partir de él? Este último cuestionamiento era tan importante que Hilbert lo propuso a los matemáticos como una de las preguntas abiertas más importantes para ser contestada en el siglo xx.

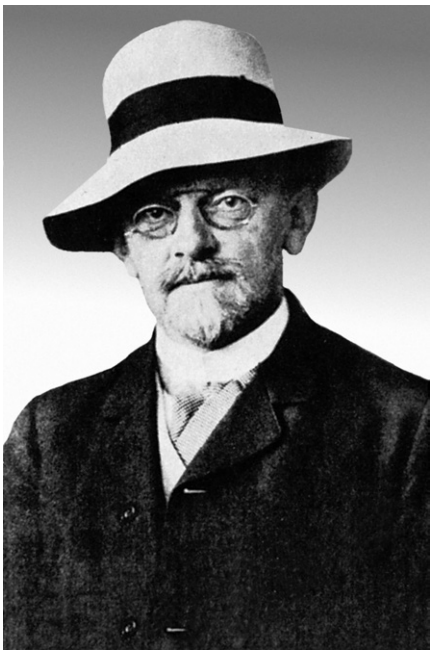


Figura 4.46 David Hilbert (1862-1943) | © Archivo Digital.

El intuicionismo asumió como suyas las críticas que emergieron frente al carácter abstracto de las matemáticas. Con Brouwer se estructuró una visión sobre la naturaleza de las matemáticas, presente también entre los matemáticos decimonónicos como Kronecker y Baire. Los intuicionistas se colocaban en un terreno opuesto, de alguna manera, al formalismo y al logicismo, pues para ellos era indispensable recurrir a la intuición.

Mientras que los logicistas elevaban la lógica a una categoría muy alta, para los intuicionistas se trataba tan sólo de un instrumento absolutamente accesorio.

Los intuicionistas no trataban de probar la consistencia de la matemática sino de hacer "matemática verdadera", apegada a la intuición introspectiva.

Si bien los intuicionistas argumentaban —permanentemente— que no estaban preocupados por los fundamentos de las matemáticas, siempre trabajaron por hacer estable el edificio matemático en el que se movían. Las ideas en la filosofía de las matemáticas y, en particular, en torno a los fundamentos, nunca han sido blancas o negras; todas fueron entretejiéndose.

La historia no tuvo un final feliz. En 1931, Kurt Gödel mostró al mundo en su artículo "Sobre proposiciones formalmente indecidibles en *Principia Mathematica* y sistemas relacionados" que no es posible fundamentar las matemáticas a través de un sistema formal que sea consistente y completo; dicho de otra manera, ningún sistema formal que contenga a la aritmética puede probar su propia consistencia, como se verá en la sección 4.11.

Figura 4.47 Leopold Kronecker (1823-1891) | © Latin Stock México.

Gödel evidenció una limitación fundamental del método axiomático al probar que, para cualquier conjunto consistente en axiomas que contenga a los axiomas de la aritmética, existen afirmaciones verdaderas que no pueden deducirse a partir de dicho conjunto.

En matemáticas, hay muchos enunciados que parecen evidentes para los que no se ha encontrado un contraejemplo y, sin embargo, no han podido ser demostrados. Un ejemplo es la *conjetura de Golbach,* que afirma que todo nú-

Figura 4.48 Kurt Gödel (1906-1978) | © IFPA, Interfoto, Archivo Digital.

mero par mayor que dos es la suma de dos números primos.

El debate sobre los fundamentos sobrepasó la necesidad de elegir alguna de las posturas enfrentadas. Legó, tanto a matemáticos como a filósofos, el conocimiento de hasta dónde podía llegarse en la formalización de teorías matemáticas, a través de la lógica, y puso de manifiesto la diferencia entre la postura *constructivista* y la concepción formalista —que considera a los objetos matemáticos como existentes—, que no estaba tan clara antes del siglo XIX. Esta definición ha sido una de las contribuciones más importantes del debate sobre los fundamentos.

Si bien el logicismo y el formalismo no lograron fundamentar totalmente las matemáticas, el desarrollo en el campo de la filosofía de las matemáticas ha tenido repercusiones muy importantes en otros ámbitos. El trabajo realizado por estos filósofos y matemáticos ha contribuido en otros campos como la lingüística, el ánalisis del razonamiento deductivo, la informática y la computación.

4.8 ¿QUÉ SE PUEDE MEDIR?

Siempre que medimos algo hay que indicar la *unidad* en la que se mide. Decir que el objeto "pesa 4" carece de sentido cuando no se aclara si se trata de kilogramos, gramos, toneladas, libras u onzas. Las medidas utilizadas han cambiado durante la historia y buena parte de las que usamos en la actualidad se establecieron en Francia, después de la Revolución y a principios del siglo XIX. El metro, por ejemplo, se estableció como la diezmilésima parte de la distancia del Ecuador al Polo Norte. Lo importante es que cualquier medida fija puede servir como unidad.

Si tenemos un segmento de 1.2 cm de longitud y otro de $\frac{10}{3}$ cm, podemos usar un tercero como unidad, de $\frac{1}{30}$ cm de longitud, y medir los dos primeros: el segmento de 1.2 cm de largo es 36 veces la unidad, mientras que el segmento de $\frac{10}{3}$ cm es 100 veces la unidad. Por lo tanto, encontramos una unidad de medida para expresar la longitud de ambos segmentos como un múltiplo *entero* de esta unidad. Estas ideas se atribuyen a los pitagóricos, la escuela que fundó Pitágoras en el siglo V a. C. Dos segmentos cualesquiera que se pueden medir con una misma unidad los llamaron *conmensurables.* Los pitagóricos partieron tácitamente de la suposición de que cada dos segmentos son conmensurables.

Para los pitagóricos, los números enteros eran centrales pues, con ellos, se expresaba la naturaleza, por ejemplo, en las relaciones de las notas con la música. Grande fue su asombro al descubrir que hay segmentos que no se pueden medir con una sola unidad; tal es el caso del lado de un cuadrado y su diagonal, que no son conmensurables. La leyenda cuenta que Hippasus, el pitagórico que hizo este descubrimiento, fue echado por la borda del barco cuando comunicó su hallazgo.

En la sección 1.4 expusimos una demostración de este hecho: $\sqrt{2}$ no es racional y, de ahí, sigue la inconmensurabilidad del lado y la diagonal: si hubiera una unidad u para ex-

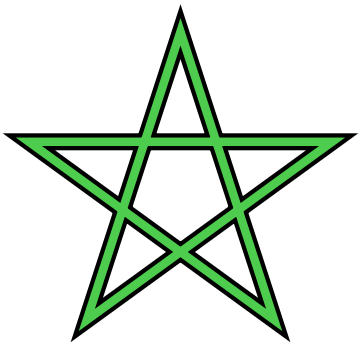
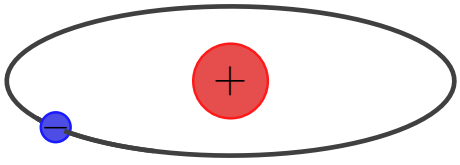


Figura 4.49 El pentagrama *es una estrella de cinco puntas que se forma con cinco segmentos rectos. Fue un símbolo que representaba la perfección para los pitagóricos. Entre las diferentes longitudes aparece la razón áurea, a la que hoy día se asocian significados muy diversos según la cultura.*

presar la longitud del lado de un cuadrado $\ell = mu$ —donde m es un entero— y también la diagonal $\sqrt{2}\ell = nu$ — con n, un entero—, entonces $\sqrt{2} = \frac{n}{m}$ es racional, lo cual es una contradicción pues sabemos que $\sqrt{2}$ no es racional.

La existencia de segmentos *inconmensurables* es, tal vez, la primera ocurrencia en las matemáticas de una demostración de *algo que no existe:* no existe y nunca existirá una unidad para medir el lado y la diagonal de un cuadrado con números enteros. En este tema se tratarán varios hallazgos matemáticos y cada uno de ellos afirmará que algo no existe de manera determinante. A la vez, cada uno tiene que ver con una pregunta importante y una teoría completa que hubo que desarrollar para poder llegar a la conclusión de la no-existencia.

4.9 ¿QUÉ SE PUEDE RESOLVER?



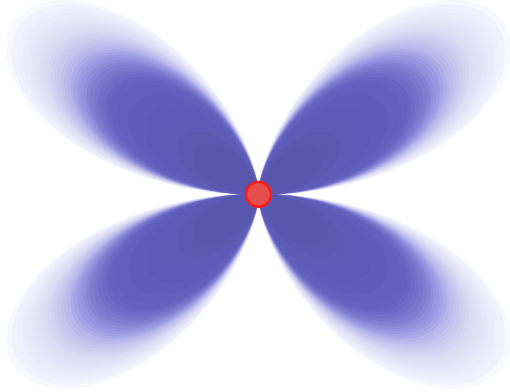


Figura 4.50 A partir de la tercera década del siglo XX, se cambió la manera de pensar sobre los átomos. Los electrones ya no son partículas que orbitan alrededor de un núcleo, sino que se encuentran dispersados como nubes de probabilidad alrededor del núcleo. Como pueden estar en diferentes estados, influyen en la forma de la nube probabilística. La imagen de abajo muestra la probabilidad de encontrar un electrón en un átomo de hidrógeno en el estado ($l = 2, m = 1$).

4.9.1 Limitación a ecuaciones algebraicas

Ya los hindúes se daban a la tarea de proponer y resolver ecuaciones desde el siglo VIII a.C. En una ecuación aparecen *variables* y *constantes*, que se enlazan con operaciones para formar los dos *términos* en ambos lados de la ecuación. Por ejemplo, en la ecuación:

$$x + 3 = 6, \tag{10}$$

x es una variable, los números 3 y 6 son constantes y hay una sola operación involucrada, que es la suma. Nuestras constantes serán números enteros o racionales o, a veces, reales.

Resolver la ecuación significa encontrar los valores para las variables, de manera que la ecuación sea válida. Si sustituimos $x = 2$ en la ecuación 10, ésta se transforma en:

$$5 = 6$$

que, evidentemente, es falso. En cambio, si sustituimos $x = 3$, obtenemos la ecuación:

$$6 = 6$$

que es correcta. Por ello, $x = 3$ es solución de la ecuación 10 y $x = 2$, no lo es.

También hay ecuaciones que involucran *funciones*. Por ejemplo:

$$\cos(\pi x) = x^2 \tag{11}$$

es una ecuación que involucra dos funciones: coseno y elevar al cuadrado. Sin embargo, estas dos funciones no juegan el mismo papel, pues podemos entender al cuadrado de x como el producto de x con x misma, lo que equivale a decir que $x^2 = x \cdot x$. En cambio, la función coseno no se puede reducir a una expresión sencilla que sólo involucre las operaciones básicas de sumar, restar, multiplicar y dividir.

Las *ecuaciones algebraicas* son aquellas que sólo involucran variables, constantes y operaciones básicas. Veamos algunos ejemplos sencillos.

Si en (10) cambiamos los valores de las constantes, cambiamos la ecuación pero no su estructura. De manera más general, podemos considerar ecuaciones que tienen la forma:

$$ax + b = 0, \tag{12}$$

donde a y b son constantes dadas. Si en una ecuación se indican constantes con una variable, se habla de *parámetros*. La ecuación (12) se llama *ecuación lineal* y podemos resolverla siempre que a no sea cero, en cuyo caso la solución es:

$$x = -\frac{b}{a}.$$

Un poco más complicada es la *ecuación cuadrática*:

$$ax^2 + bx + c = 0, \tag{13}$$

donde x es la variable mientras que a, b y c son los parámetros, es decir, constantes. Esta ecuación puede tener hasta dos soluciones:

$$x = \frac{-b + \sqrt{b^2 - 4ac}}{2a}, \quad x = \frac{-b - \sqrt{b^2 - 4ac}}{2a}. \tag{14}$$

Estas fórmulas se obtienen con un truco para completar cuadrados y que se explicará con detalle más adelante.

El número de soluciones depende del valor del *discriminante* $D = b^2 - 4ac$, es decir, del valor de lo que está dentro de la raíz cuadrada. Si $D = 0$, entonces sólo hay una solución: $x = -\frac{b}{2a}$. Cuando $D > 0$, las dos soluciones (14) son distintas.

Un caso interesante es cuando $D < 0$ pues, a partir de la fórmula, se tendría que calcular la raíz cuadrada de un número negativo. Pero no existe número real λ tal que $\lambda = \sqrt{-2}$, ya que al elevar al cuadrado se obtendría $\lambda^2 = -2$. ¡Es imposible tener el lado izquierdo positivo y el derecho negativo! Por ello, podríamos decir que la ecuación (13) no tiene solución. También podríamos tratar de extender la noción de los números e incluir las raíces de números negativos. Aunque sea algo bastante raro, sorprendentemente, funciona muy bien. La humanidad tardó siglos en aceptar estos nuevos números que, en la actualidad, son muy comunes. Los números complejos son un concepto difícil, así que para entenderlos iremos despacio.

4.9.2 El cálculo con "el número" $\mathbf{i} = \sqrt{-1}$

Todo parte del supuesto de que existe un número $\mathbf{i}$, que llamaremos *unidad imaginaria* —por razones históricas—, tal que $\mathbf{i}^2 = -1$. Lo que buscamos es extender los números reales a un nuevo conjunto <más grande, llamado el de los *números complejos*, de manera muy similar a como se extienden los números naturales a los enteros, al añadir los negativos. Así, debemos poder formar nuevos números —por lo pronto, los marcamos con negritas para distinguirlos— como:

$$\mathbf{x} = 2 + \mathbf{i}, \quad \mathbf{y} = 3 - 2\mathbf{i}.$$

Con ellos, debemos poder calcular productos:

$$\mathbf{x} \cdot \mathbf{y} = (2 + \mathbf{i})(3 - 2\mathbf{i}).$$

Pero entonces, necesitamos que los nuevos números satisfagan las mismas leyes que los reales, respecto a la adición y el producto. En este caso, podríamos seguir:

$$\begin{aligned}
\mathbf{x} \cdot \mathbf{y} &= (2 + \mathbf{i})(3 - 2\mathbf{i}) \\
&= 6 + 2(-2\mathbf{i}) + 3\mathbf{i} + \mathbf{i}(-2\mathbf{i}) && \text{por distributivo} \\
&= 6 + 2(-2\mathbf{i}) + 3\mathbf{i} + (-2\mathbf{i})\mathbf{i} && \text{por conmutativo de la multiplicación} \\
&= 6 + (-4)\mathbf{i} + 3\mathbf{i} + (-2)\mathbf{i}^2 && \text{por asociativo de la multiplicación} \\
&= 6 + (-4 + 3)\mathbf{i} + (-2)(-1) && \text{por distributividad y } \mathbf{i}^2 = -1 \\
&= 8 - \mathbf{i} && \text{por conmutatividad de la adición}
\end{aligned}$$

Como se observa en el ejemplo anterior, obtuvimos de nuevo la forma $a + b\mathbf{i}$, donde a y b son dos números reales. Esto, en efecto, siempre es así y podemos hacer lo mismo que en el ejemplo para ver que:

$$(a + b\mathbf{i}) \cdot (c + d\mathbf{i}) = (ac - bd) + (ad + bc)\mathbf{i} \qquad (15)$$

La suma es más fácil:

$$\begin{aligned}
\mathbf{x} + \mathbf{y} &= (2 + \mathbf{i}) + (3 - 2\mathbf{i}) \\
&= (2 + 3) + (\mathbf{i} - 2\mathbf{i}) && \text{por conmutatividad y asociatividad de la adición} \\
&= 5 - \mathbf{i}
\end{aligned}$$

En general, se tiene que:

$$(a + b\mathbf{i}) + (c + d\mathbf{i}) = (a + c) + (b + d)\mathbf{i}.$$

También podemos restar:

$$(a + b\mathbf{i}) - (c + d\mathbf{i}) = (a - c) + (b - d)\mathbf{i},$$

y para ver que podemos dividir, tenemos que hacer un par de maniobras adicionales. Dividir entre 2 es lo mismo que multiplicar por $\frac{1}{2}$, esto es, por el multiplicativo inverso de 2. En general, esperamos que:

$$\mathbf{x} \div \mathbf{y} = \mathbf{x} \cdot \frac{1}{\mathbf{y}},$$

y como sabemos multiplicar gracias a la fórmula (15), es suficiente calcular el inverso multiplicativo de cualquier número complejo con forma $\mathbf{y} = c + d\mathbf{i}$. Para ello, observamos que:

$$(c + d\mathbf{i}) \cdot (c - d\mathbf{i}) = (c^2 - d(-d)) + 0\mathbf{i} = c^2 + d^2,$$

es un número real que no es cero si $\mathbf{y} \neq 0 + 0\mathbf{i} = 0$. Por ello, si $\mathbf{y} \neq 0$ tenemos que:

$$(c + d\mathbf{i}) \cdot \left(\frac{c}{c^2 + d^2} - \frac{d}{c^2 + d^2}\mathbf{i} \right) = \left(c \cdot \frac{c}{c^2 + d^2} - d(-\frac{d}{c^2 + d^2}) \right) + 0\mathbf{i} = 1.$$

Por consiguiente, se obtiene la siguiente fórmula para la formación de inversos multiplicativos:

$$\frac{1}{c + d\mathbf{i}} = \frac{c}{c^2 + d^2} - \frac{d}{c^2 + d^2}\mathbf{i}.$$

Para resumir, debemos considerar que todo parte de dos supuestos:

- Existe un número i, que llamaremos *unidad imaginaria*, tal que $i^2 = -1$.
- El número i y todos los nuevos números satisfacen las mismas leyes que involucran la adición y la multiplicación que los números reales.

Entonces, los números complejos son de la forma $a + b\mathbf{i}$, con a y b reales y que cumplen las siguientes leyes:

$$(a + b\mathbf{i}) + (c + d\mathbf{i}) = (a + c) + (b + d)\mathbf{i} \quad \text{adición} \quad (16)$$

$$(a + b\mathbf{i}) \cdot (c + d\mathbf{i}) = (ac - bd) + (ad + bc)\mathbf{i} \quad \text{multiplicación} \quad (17)$$

$$\mathbf{0} = 0 + 0\mathbf{i} \quad \text{el cero (neutro aditivo)} \quad (18)$$

$$\mathbf{1} = 1 + 0\mathbf{i} \quad \text{el uno (neutro multiplicativo)} \quad (19)$$

$$-(a + b\mathbf{i}) = -a - b\mathbf{i} \quad \text{inverso aditivo} \quad (20)$$

$$\frac{1}{c + d\mathbf{i}} = \frac{c}{c^2 + d^2} - \frac{d}{c^2 + d^2}\mathbf{i} \quad \text{inverso multiplicativo} \quad (21)$$

4.9.3 REPRESENTACIÓN GEOMÉTRICA DE LOS NÚMEROS COMPLEJOS

Todo lo anterior está muy bien, sólo que es muy abstracto. ¿Cómo imaginarse un número complejo? Para los números reales tenemos la recta real como objeto geométrico, que nos da una cierta posibilidad de imaginación e intuición. Para los números complejos hay también un modelo geométrico llamado el *plano complejo*, idea de Jean-Robert Argand —un suizo aficionado a las matemáticas—, donde cada número complejo $\mathbf{x} = a + b\mathbf{i}$ está dado por dos números reales a y b, que podemos imaginar como coordenadas.

Figura *4.51 Jean-Robert Argand (1768-1822)* | *© Latin Stock México.*

El eje que se dibujó horizontalmente es la recta real, mientras que el eje vertical o imaginario contiene los múltiplos de $\mathbf{i}$. Ahora, debemos aclarar cómo se representan la adición y la multiplicación geométricamente.

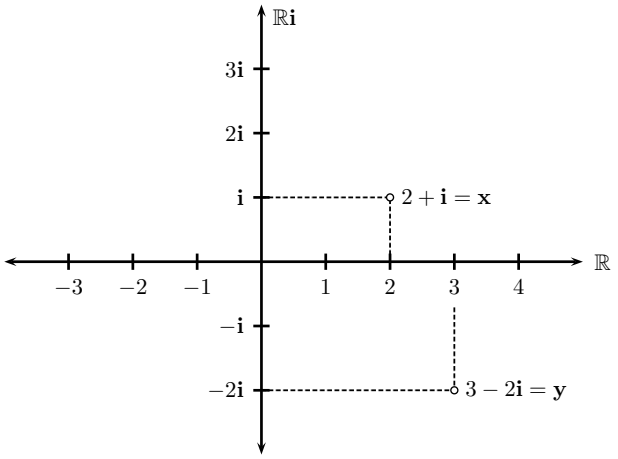


Figura 4.52 El plano complejo.

De la fórmula (16), observamos que la adición se obtiene como la cuarta esquina de un paralelogramo, con sus otras tres esquinas $\mathbf{0}$, $\mathbf{x}$ y $\mathbf{y}$. Formar geométricamente el inverso aditivo no es otra cosa que la reflexión en el punto $\mathbf{0}$. Para entender el significado de la multiplicación, tenemos primero que trabajar algebraicamente. Dado que cada número complejo $\mathbf{x}$ se encuentra en el plano, se puede representar en *coordenadas polares* al indicar, por un lado, la distancia al origen r, que también se llama el *valor absoluto* de $\mathbf{x}$ y, por el otro, el ángulo $\measuredangle\mathbf{1}\text{-}\mathbf{0}\text{-}\mathbf{x}$, es decir, el ángulo que forman los puntos $\mathbf{1}$, $\mathbf{0}$ y $\mathbf{x}$.

El número complejo con coordenadas polares r y φ es $\mathbf{x} = r(\cos(\varphi) + \operatorname{sen}(\varphi)\mathbf{i})$. Las conversiones entre $\mathbf{x} = a + b\mathbf{i}$ y la forma polar se calculan con las siguientes fórmulas:

$$a = r\cos(\varphi), \qquad r = \sqrt{a^2 + b^2},$$
$$b = r\,\mathrm{sen}(\varphi), \qquad \varphi = \tan^{-1}\left(\frac{b}{a}\right).$$

Entonces podemos reescribir la fórmula de multiplicación (17) de la siguiente manera:

$$\begin{aligned} r(\cos(\varphi) + \mathrm{sen}(\varphi)\mathbf{i}) \cdot s(\cos(\psi) + \mathrm{sen}(\psi)\mathbf{i}) =& rs\left(\cos(\varphi)\cos(\psi) - \mathrm{sen}(\varphi)\,\mathrm{sen}(\psi)\right) + \\ &+ rs\left(\cos(\varphi)\,\mathrm{sen}(\psi) + \psi(\varphi)\cos(\psi)\right)\mathbf{i} \\ =& rs\left(\cos(\varphi + \psi) + \mathrm{sen}(\varphi + \psi)\mathbf{i}\right) \end{aligned}$$

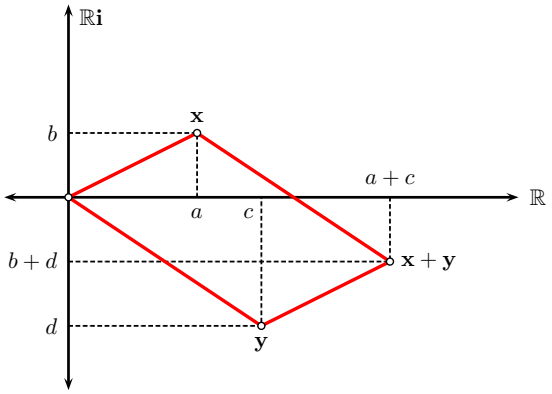


Figura 4.53 La suma de dos números complejos.

Esto significa que al multiplicar dos números complejos se multiplican sus valores absolutos y se suman sus ángulos.

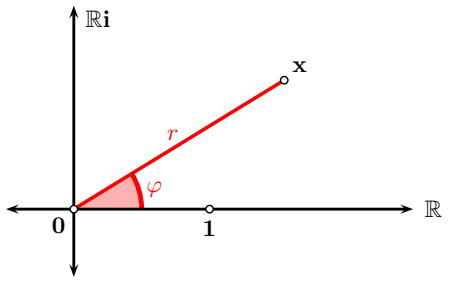


Figura 4.54 Coordenadas polares para números complejos.

Por último, regresamos al punto de partida. Al inicio del discurso, nos dejamos llevar hacia los números complejos con la finalidad de obtener raíces de números reales negativos. Ahora, queremos ver si podemos sacar la raíz cuadrada de cualquier número complejo o si es necesario extender nuestra noción de números una y otra vez.

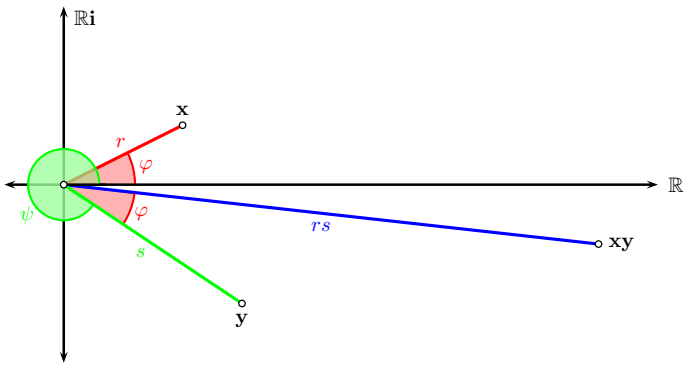


Figura *4.55 El producto de dos números complejos.*

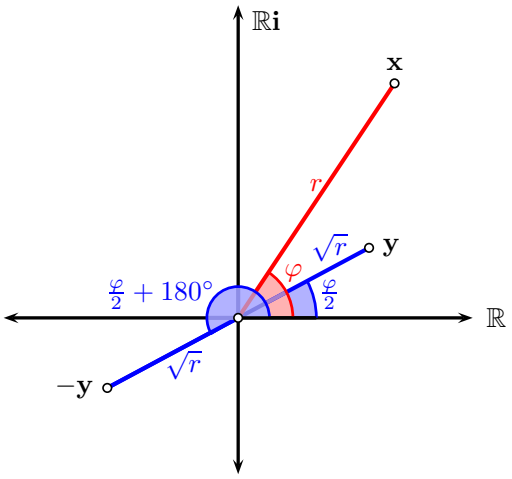


Figura *4.56 Las dos raíces de un número complejo.*

Si queremos sacar la raíz cuadrada de un número complejo $\mathbf{x}$, debemos encontrar otro número complejo $\mathbf{y}$, tal que $\mathbf{y} \cdot \mathbf{y} = \mathbf{x}$. Con la intuición geométrica, esto ya no es difícil: usamos coordenadas polares para $\mathbf{x}$ —digamos que tiene valor absoluto r y ángulo φ—. Entonces, $\mathbf{y}$ debe tener valor absoluto $\sqrt{r}$ y ángulo $\frac{\varphi}{2}$. Observamos que siempre —salvo cuando $\mathbf{x} = 0$— hay dos soluciones: una es el inverso negativo de la otra.

Al recapitular, tenemos que los números complejos son expresiones de la forma $a + b\mathbf{i}$, donde a y b son números reales que pueden sumarse y multiplicarse, restarse y dividirse. Por lo anterior, los números complejos forman un conjunto que los matemáticos llaman un *campo* y, además, tienen una representación geométrica: el plano complejo.

Todavía queda la duda de si estos números complejos son algo más que una invención medio rara para poder sacar la raíz cuadrada de cualquier número real. ¿Qué relevancia tienen? ¿Se aplican en algún lugar o son un mero juguete matemático? Hasta el principio del siglo xx, los números complejos no tenían gran interés fuera de las matemáticas. Sin embargo, en los años veinte del siglo xx se forjó la *mecánica cuántica*, una teoría que logra describir con gran precisión los fenómenos de las interacciones de partículas atómicas. Esta teoría usa los números complejos de una manera esencial: sin ellos, no se podría formular adecuadamente pues explica, en forma contundente, el sistema de los elementos químicos que antes se había formulado con base en observaciones.

4.9.4 La ecuación cuadrática con coeficientes complejos

Antes nos desviamos considerablemente de nuestro objetivo principal: estudiar cómo podría responderse la pregunta ¿qué ecuación se puede resolver? Primero, la restringimos a las ecuaciones algebraicas. Y, hasta ahora, sólo llegamos a la conclusión de que con los números complejos podemos resolver cualquier *ecuación cuadrática* (13) con parámetros a, b y c *reales*. No obstante, vimos que podemos sacar la raíz cuadrada de cualquier número complejo, lo cual nos podría dar la idea de cuestionar si sería posible resolver (13), aun cuando los parámetros $\mathbf{a}, \mathbf{b}, \mathbf{c}$ sean *números complejos*, como en:

$$\mathbf{a}x^2 + \mathbf{b}x + \mathbf{c} = 0.$$

En efecto, lo anterior es cierto y se usa exactamente la misma fórmula (14) que antes. La razón se debe a que esta fórmula se deriva de manera *algebraica* sin hacer ninguna suposi-

ción, excepto que $\mathbf{a} \neq 0$. Consideremos esta deducción y pensemos que $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ son números complejos. Primero, se divide la ecuación entre $\mathbf{a}$:

$$x^2 + \frac{\mathbf{b}}{\mathbf{a}}x + \frac{\mathbf{c}}{\mathbf{a}} = 0.$$

Luego, se suma el término $\frac{\mathbf{b}}{4\mathbf{a}^2}$ en ambos lados y se observa que $x^2 + \frac{\mathbf{b}}{\mathbf{a}}x + \frac{\mathbf{b}}{4\mathbf{a}^2}$ es un cuadrado perfecto:

$$\begin{aligned} x^2 + \frac{\mathbf{b}}{\mathbf{a}}x + \frac{\mathbf{b}}{4\mathbf{a}^2} + \frac{\mathbf{c}}{\mathbf{a}} &= \frac{\mathbf{b}}{4\mathbf{a}^2} \\ \left(x + \frac{\mathbf{b}}{2\mathbf{a}}\right)^2 + \frac{\mathbf{c}}{\mathbf{a}} &= \frac{\mathbf{b}}{4\mathbf{a}^2} \\ \left(x + \frac{\mathbf{b}}{2\mathbf{a}}\right)^2 = \frac{\mathbf{b}}{4\mathbf{a}^2} - \frac{\mathbf{c}}{\mathbf{a}} &= \frac{\mathbf{b} - 4\mathbf{ac}}{4\mathbf{a}^2} \end{aligned}$$

En la última fila, se restó el término $\frac{\mathbf{c}}{\mathbf{a}}$ en ambos lados. Ahora se puede sacar la raíz. Recordemos aquí que ello siempre se puede, dado que $\frac{\mathbf{b}}{4\mathbf{a}} - \frac{\mathbf{c}}{\mathbf{a}}$ es un número complejo y, por lo tanto, siempre hay dos soluciones, salvo cuando $\frac{\mathbf{b}}{4\mathbf{a}^2} - \frac{\mathbf{c}}{\mathbf{a}} = 0$.

$$\begin{aligned} x + \frac{\mathbf{b}}{2\mathbf{a}} &= \pm\sqrt{\frac{\mathbf{b} - 4\mathbf{ac}}{4\mathbf{a}^2}} = \frac{\sqrt{\mathbf{b} - 4\mathbf{ac}}}{2\mathbf{a}} \\ x &= \pm\frac{\sqrt{\mathbf{b} - 4\mathbf{ac}}}{2\mathbf{a}} - \frac{\mathbf{b}}{2\mathbf{a}} = \frac{-\mathbf{b} \pm \sqrt{\mathbf{b} - 4\mathbf{ac}}}{2\mathbf{a}} \end{aligned}$$

De nuevo obtuvimos la fórmula de solución (14) y de nada importó que los coeficientes $\mathbf{a}$, $\mathbf{b}$ y $\mathbf{c}$ sean complejos.

4.9.5 Las ecuaciones de tercer y cuarto grados

No sólo la ecuación cuadrática, sino también la ecuación algebraica de *tercer grado* que se representa como:

$$ax^3 + bx^2 + cx + d = 0, \tag{22}$$

admite una fórmula de solución. Lo mismo sucede con la ecuación algebraica de *cuarto grado*:

$$ax^4 + bx^3 + cx^2 + dx + e = 0.$$

Aquí ya no resaltamos que los coeficientes pueden ser números racionales, reales o complejos, pues el tratamiento que daremos no depende de ello.

La resolución de la ecuación de tercer grado se le atribuye a Niccolò Tartaglia en 1530. La historia cuenta que sólo después de considerables insistencias por parte de Gerolamo Cardano, Tartaglia le confió su secreto bajo la promesa de nunca publicarlo. Unos diez años después, Lodovico Ferrari —un estudiante de Cardano— encontró la solución para la ecuación de cuarto grado. En 1545, el propio Cardano publicó ambos resultados en un libro que hoy se conoce como las *fórmulas de Cardano*.

A continuación, daremos una breve idea de la solución de la ecuación de tercer grado. Primero, se hacen dos maniobras para reducir el número de parámetros involucrados, al dividir la ecuación (22) entre a:

$$x^3 + b'x^2 + c'x + d = 0,$$

donde $b' = \frac{b}{a}, c' = \frac{c}{a}, d' = \frac{d}{a}$. Después, se sustituye:

$$x = y - \frac{b'}{3}, \tag{23}$$

y se obtiene que:

$$\left(y - \frac{b'}{3}\right)^3 + b'\left(y - \frac{b'}{3}\right)^2 + c'\left(y - \frac{b'}{3}\right) + d' = 0,$$

que al expandir, se convierte en:

$$y^3 - 3y^2\frac{b'}{3} + 3y\left(\frac{b'}{3}\right)^2 - \left(\frac{b'}{3}\right)^3 + b'y^2 - 2b'y\frac{b'}{3} + b'\left(\frac{b'}{3}\right)^2 + c'y - c'\frac{b'}{3} + d' = 0$$

$$y^3 + \underbrace{\left(c' - \frac{b'^2}{3}\right)}_{=p} y + \underbrace{\left(\frac{2b'^3}{27} - \frac{b'c'}{3} + d'\right)}_{=q} = 0$$

Con las definiciones de p y q como hemos indicado, reducimos el número de parámetros a sólo dos. Ahora, sólo tenemos que resolver la ecuación:

$$y^3 + py + q = 0. \tag{24}$$

El truco consiste en suponer que tenemos una solución y escribirla como una diferencia $y = r - s$. Al sustituir, se obtiene que:

$$\begin{aligned} r^3 - 3r^2s + 3rs^2 - s^3 + pr - ps + q = 0 \qquad (25)\\ (r^3 - s^3 + q) + (r - s)(p - 3rs) = 0 \end{aligned}$$

Ahora, tratemos de encontrar los números r y s, tales que tenemos simultáneamente que ambos sumandos en (25) son cero:

$$r^3 - s^3 + q = 0, \tag{26}$$
$$(r - s)(p - 3rs) = 0. \tag{27}$$

Revisemos primero la segunda ecuación. Como tenemos un producto que es cero, uno de los dos factores también debe ser cero. Si $r - s = 0$, entonces $y = 0$. Pero si $y = 0$ es una solución, entonces $q = 0$ y la ecuación (24) se puede escribir como $(y^2 + p)y = 0$, donde las otras dos soluciones son $\pm\sqrt{-p}$. Por lo tanto, en caso de que $y = 0$ sea solución, no tenemos que seguir.

El caso $y \neq 0$ es más interesante, pues obtenemos de (26) que:

$$3rs = p,$$

por lo que $s = \frac{p}{3r}$. Al sustituir s por $\frac{p}{3r}$ en (27), se obtiene:

$$r^3 - \frac{p^3}{27r^3} + q = 0.$$

Si multiplicamos por r^3 y sustituimos $R = r^3$, obtenemos la ecuación:

$$R^2 + qR - \frac{p^3}{27} = 0$$

que es cuadrática para la variable R y cuyas soluciones son:

$$R = \frac{-q \pm \sqrt{q^2 + \frac{4p^3}{27}}}{2} = -\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}.$$

Además podemos fijar un signo, como veremos pronto. Digamos que $R = -\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}$. Entonces:

$$r = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}, \quad s = \sqrt[3]{\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}, \tag{28}$$

donde la expresión para s se obtiene de (27): simplemente al sustituir las dos expresiones para r y s del lado izquierdo de (27) y después se simplifica para obtener el lado derecho. Así, hemos encontrado una solución de (24):

$$y = \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \sqrt[3]{\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}$$

Si quisiéramos encontrar la solución para la ecuación de tercer grado (22), tendríamos que sustituir y en 23 para obtener x.

Observemos que en 28 tenemos tres soluciones de $r^3 = R$: si tenemos una solución r, también αr y $\alpha^2 r$ son soluciones, donde $\alpha = \cos(120°) + \operatorname{sen}(120°)$.

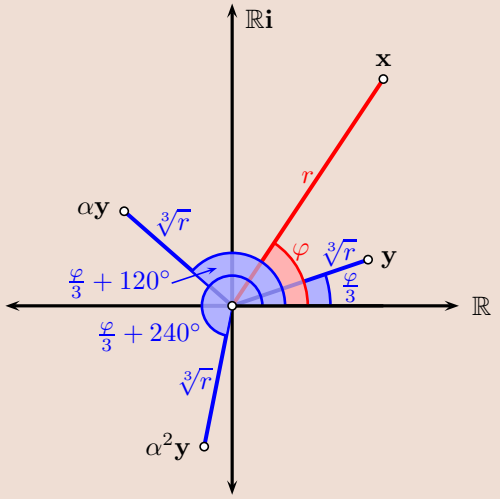


Figura 4.57 Soluciones de $r^3 = R$.

Por lo tanto, las otras dos soluciones son:

$$y_2 = \alpha \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \alpha^2 \sqrt[3]{\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}},$$

$$y_3 = \alpha^2 \sqrt[3]{-\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}} - \alpha \sqrt[3]{\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}}}.$$

Las fórmulas para la ecuación de cuarto grado son aún más complicadas y ni siquiera las indicaremos.

Queremos ahora ocuparnos de las ecuaciones de grado mayor que cuatro. Posterior a Cardano, los matemáticos intentaron reiteradamente resolver ecuaciones de grado cinco y encontrar una fórmula de solución, pero siempre fracasaron. En 1799, Paolo Ruffini dio ideas sobre cómo demostrar que tal fórmula no puede existir. En 1824, Niels Hendrik Abel, un matemático sueco, llenó los "huecos" de la demostración de Ruffini.

4.9.6 El teorema fundamental del álgebra

Figura 4.58 Niels Hendrik Abel (1802-1829) | © Latin Stock México.

Hay un punto delicado aquí: la demostración de Abel no afirma que *no hay solución*, sino que *no hay una fórmula de solución*. Ya desde el siglo XVII se pensaba que la ecuación algebraica de grado n tiene n soluciones que pueden repetirse. Para hacer más preciso lo anterior: hay n números complejos $y_1, y_2, \ldots, y_n$, tal que:

$$a_n x^n + a_{n-1} x^{n-1} + \ldots + a_2 x^2 a_1 x + a_0 = a_n \cdot (x - y_1) \cdot (x - y_2) \cdot \ldots \cdot (x - y_n).$$

Así, si quisiéramos resolver la ecuación:

$$a_n x^n + a_{n-1} x^{n-1} + \ldots + a_2 x^2 a_1 x + a_0 = 0,$$

tendríamos que resolver:

$$a_n \cdot (x - y_1) \cdot (x - y_2) \cdot \ldots \cdot (x - y_n) = 0,$$

lo que resulta fácil, dado que un producto de factores es cero sólo si uno de sus factores lo es. Por ello —y como $a_n \neq 0$—, se tiene que $x - y_i = 0$ para algún i, así que $x = y_i$ para algún i. Lo anterior muestra que sí tiene sentido decir que una solución es "doble" o "múltiple". El resultado formal se llama el teorema fundamental del álgebra y dice que: *Cada ecuación algebraica*:

$$a_n x^n + a_{n-1} x^{n-1} + \ldots + a_2 x^2 a_1 x + a_0 = 0 \qquad (29)$$

con coeficientes $a_0, a_1, \ldots, a_n$ *números complejos tiene* n *soluciones complejas, si se cuentan posibles repeticiones.*

Por ejemplo, la ecuación $x^2 - 2x + 1 = 0$ tiene las dos soluciones: 1 y 1 ya que $x^2 - 2x + 1 = (x - 1)(x - 1)$.

La primera demostración correcta es de Jean-Robert Argand, en 1806. Gauss hizo varias demostraciones de este resultado a lo largo de su vida. El teorema fundamental del álgebra muestra que siempre hay soluciones complejas, pero no es constructivo, es decir, no da ningún método para encontrarlas. Es un punto delicado: el resultado afirma la existencia de soluciones sin exhibir ninguna de ellas. Sólo dice que existen, nada más. Por ello, se percibe la gran importancia de fórmulas para calcular las soluciones de manera explícita. Sin embargo, el teorema de Abel-Ruffini dice que éstas no se pueden encontrar con una fórmula si $n \geq 5$. Lo anterior quiere decir que las soluciones existen y pueden encontrarse de manera *aproximativa*, es decir, hasta cualquier precisión de dígitos, pero si $n \geq 5$, *no es posible* sustituir los coeficientes $a_0, a_1, \ldots, a_n$ de la ecuación (29) en alguna fórmula para obtener de golpe las soluciones.

4.9.7 Polinomios, raíces y simetrías

La demostración de Abel-Ruffini tiene una desventaja: es técnica y dificulta el entender las razones por las que ecuaciones de grado 1, 2, 3 y 4 se resuelven con una fórmula, pero las de grado 5 —y mayores— no se pueden resolver de esta manera. En la actualidad, se usa una teoría basada en las ideas de Evariste Galois —un matemático francés contemporáneo de Abel— para demostrarlo. Sin embargo, antes de llegar a la *teoría de Galois*, tendremos que preparar unas nociones sobre polinomios.

Los términos de la forma:

$$f(x) = a_n x^n + a_{n-1} x^{n-1} + \ldots + a_2 x^2 a_1 x + a_0 \qquad (30)$$

se llaman *polinomios* en la variable x con coeficientes $a_0, a_1, \ldots, a_n$. Si $a_n \neq 0$, entonces n es el *grado* de $f(x)$.

En lo que sigue, será importante considerar polinomios *sobre diferentes campos*: si todos los coeficientes pertenecen al campo K de los números racionales, diremos que $f(x)$ es un *polinomio racional*. Los polinomios racionales forman un conjunto $[x]$ donde es posible sumar y multiplicar también hay elementos neutros aditivos y multiplicativos e inversos aditivos, pero no existen inversos multiplicativos. Se dice que $[x]$ es un *anillo*. De manera similar si K es un campo, denotamos con $K[x]$ al anillo de polinomios con coeficientes en K. Un caso importante son los números complejos: $K = \mathbb{C}$.

Si en la expresión 30 se sustituye la variable x por un número s, entonces se obtiene un número que se denota por $f(s)$. Un número s que satisface $f(s) = 0$ se llama *raíz* del polinomio $f(x)$. De hecho la notación $f(x)$ sugiere que se trata de una función y eso es cier-

to: se trata de la función polinomial que se obtiene al sustituir diferentes valores para x. Si s es raíz de $f(x)$, entonces $f(x)$ se puede factorizar:

$$f(x) = g(x) \cdot (x - s),$$

es decir, existe un polinomio $g(x)$ con grado menor que $f(x)$ y factor lineal de $x - s$, tal que su producto es $f(x)$. Pero los coeficientes del polinomio $g(x)$ serán expresiones que pueden involucrar a la raíz s, como se explica en el siguiente ejemplo.

Por ejemplo, si $f(x) = x^3 - x + 1$ y s es alguna de sus raíces, entonces se puede hacer la división de polinomios $(x^3 - x - 1) \div (x - s)$, que se explica a continuación.

Para hacer la operación de $f(x) = x^3 - x - 1$ entre $x - s$, se divide el monomio de $f(x)$ de mayor grado entre el monomio de mayor grado de $x - s$; en este caso hay que dividir x^3 entre x y se obtiene x^2. Luego, se forma la resta $f_1(x)$:

$$\begin{aligned} f_1(x) &= (x^3 - x - 1) - x^2 \cdot (x - s) \\ &= x^3 - x - 1 - x^3 + sx^2 \\ &= sx^2 - x - 1 \end{aligned}$$

Después se procede en forma similar: el monomio de $f_1(x)$ de mayor grado es sx^2, por lo que al dividirlo entre x, tenemos sx. Esta vez, se resta $(sx) \cdot (x - s) = sx^2 - s^2x$ de $f_1(x)$ y se obtiene:

$$\begin{aligned} f_2(x) &= (sx^2 - x - 1) - sx \cdot (x - s) \\ &= sx^2 - x - 1 - sx^2 + s^2x \\ &= (s^2 - 1)x - 1, \end{aligned}$$

cuyo monomio de grado mayor es $(s^2 - 1)x$. Nuevamente, al dividirlo entre x se obtiene $(s^2 - 1)$ y al calcular la resta:

$$\begin{aligned} f_3(x) &= ((s^2 - 1)x - 1) - (s^2 - 1) \cdot (x - s) \\ &= (s^2 - 1)x - 1 - (s^2 - 1)x + (s^2 - 1)s \\ &= s^3 - s - 1. \end{aligned}$$

Ahora hay que observar que $f_3(x) = s^3 - s - 1 = f(s) = 0$, lo que muestra que la división se hizo sin resta. Se obtuvo la división:

$$(x^3 - x - 1) \div (x - s) = x^2 + sx + (s^2 - 1)$$

El teorema fundamental del álgebra tiene una interpretación alterna: cada polinomio $f(x) \in \mathbb{C}[x]$ se puede escribir como producto:

$$a_n \cdot (x - s_1) \cdot (x - s_2) \cdot \ldots \cdot (x - s_n), \tag{31}$$

donde $s_1, \ldots, s_n$ son las soluciones de la ecuación (29). En otras palabras, cada polinomio con coeficientes complejos se descompone en *factores lineales*. Por ejemplo:

$$x^2 + 1 = (x - \mathbf{i}) \cdot (x + \mathbf{i}), \tag{32}$$

dado que $\pm\mathbf{i}$ son las dos soluciones de la ecuación $x^2 + 1 = 0$. El polinomio $x^2 + 1 \in \mathbb{Q}[x]$ no se descompone en factores lineales pues $\pm\mathbf{i}$ no son racionales, pero $x^2 + 1 \in \mathbb{C}[x]$ sí se descompone como (32). La descomposición en (31) es única en el siguiente sentido: el factor a_n es único y los elementos $s_1, \ldots, s_n$ son únicos salvo permutación, es decir, dos representaciones (31) —si son iguales— sólo pueden distinguirse en el orden en que se escriben los factores $(x - s_1), \ldots, (x - s_n)$.

Hemos llegado a un punto crucial: ¿cuál es el número $\mathbf{i}$ y cuál es $-\mathbf{i}$? Recordemos que antes simplemente postulamos su existencia y vimos que podíamos hacer cálculos con números de la forma $a + b\mathbf{i}$. El número $\mathbf{i}$ es una solución de la ecuación $x^2 + 1 = 0$, mientras que la otra es $-\mathbf{i}$. Pero no hay ninguna manera de saber cuál es cuál, sólo sabemos que una solución es la inversa aditiva de la otra. Si otra persona escoge $\mathbf{j} = -\mathbf{i}$ como solución, entonces obtiene una copia igual de números complejos. Lo anterior se puede decir todavía de otra manera: hay una *simetría* en los números complejos, donde la función:

$$a + b\mathbf{i} \mapsto \overline{a + b\mathbf{i}} = a - b\mathbf{i}$$

se llama *conjugación*. La conjugación es un *automorfismo*, es decir, es biyectiva y conserva las operaciones de adición, multiplicación y la formación de inversos:

$$\begin{aligned}
\overline{\mathbf{x} + \mathbf{y}} &= \overline{\mathbf{x}} + \overline{\mathbf{y}}, \\
\overline{\mathbf{x} \cdot \mathbf{y}} &= \overline{\mathbf{x}} \cdot \overline{\mathbf{y}}, \\
\overline{-\mathbf{x}} &= -\overline{\mathbf{x}} \\
\overline{1/\mathbf{x}} &= 1/\overline{\mathbf{x}}.
\end{aligned}$$

Los números fijos bajo la conjugación son los números reales, dado que $a + b\mathbf{i} = \overline{a + b\mathbf{i}}$ implica que $b = 0$.

Sin embargo, en la realidad no necesitamos los números reales para estudiar la ecuación $x^2 + 1 = 0$. Los coeficientes son números enteros. El campo *más pequeño* (respecto a la contención de conjuntos) que contiene a los coeficientes 0, 1 es $\mathbb{Q}$, el campo de los números racionales. Por ejemplo, el campo más pequeño que contiene los coeficientes y las soluciones $\pm\mathbf{i}$ es el campo K, cuyos elementos son de la forma $a + b\mathbf{i}$, donde $a, b \in \mathbb{Q}$ Se dice que K es el *campo de descomposición* del polinomio $x^2 + 1$ sobre $\mathbb{Q}$, dado que $x^2 + 1$ se descompone en K factores lineales (32) y K es el campo más chico con esta propiedad.

En general, fijamos un polinomio $f(x)$ como en (30), con coeficientes $a_0, a_1, \ldots, a_n \in \mathbb{Q}$, y definimos K como el *campo de descomposición* de $f(x)$, es decir, el campo más pequeño que contiene $\mathbb{Q}$ y todas las soluciones de la ecuación $f(x) = 0$. Por ejemplo, si $f(x) = x^3 - 2$ es el polinomio a descomponer, entonces $\sqrt[3]{2}$ es una raíz y $f(x)$ se descompone como:

$$x^3 - 2 = (x - \sqrt[3]{2})(x^2 + \sqrt[3]{2}x + \sqrt[3]{2}^2), \tag{33}$$

en $K[x]$, donde K es el campo:

$$K = \mathbb{Q}(\sqrt[3]{2}) = \{b_0 + b_1\sqrt[3]{2} + b_2\sqrt[3]{4} \mid b_0, b_1, b_2 \in \mathbb{Q}\}.$$

Con ayuda de la fórmula general para resolver ecuaciones cuadráticas y después de al-

gunas transformaciones algebraicas, se obtiene que el segundo factor de (33) se puede escribir como:

$$x^2 + \sqrt[3]{2}x + \sqrt[3]{2}^2 = (x - \rho\sqrt[3]{2})(x - \rho^2\sqrt[3]{2}),$$

donde $\rho = \frac{-1+\sqrt{3}}{2}$ es una *tercera raíz primitiva de la unidad*, es decir, una solución de la ecuación $x^3 = 1$ que no es solución de una ecuación $x^p = 1$ para $p < 3$. Las tres raíces de $f(x) = x^3 - x - 1$ son entonces $z_1 = \sqrt[3]{2}$, $z_2 = \rho\sqrt[3]{2}$ y $z_3 = \rho^2\sqrt[3]{2}$.

En seguida, denotamos con $\mathrm{Gal}(K : \mathbb{Q})$ al conjunto de todos los automorfismos de K que fijan los elementos de $\mathbb{Q}$ punto por punto, es decir $\varphi(z) = z$ para cada $z \in \mathbb{Q}$ y cada $\varphi \in \mathrm{Gal}(K : \mathbb{Q})$. Por construcción, $\mathrm{Gal}(K : \mathbb{Q})$ es un *grupo*, dado que podemos componer dos automorfismos que fijan $\mathbb{Q}$ para obtener otra vez un automorfismo que fija $\mathbb{Q}$; la identidad de K siempre es un elemento de $\mathrm{Gal}(K : \mathbb{Q})$ y un inverso de un elemento de $\mathrm{Gal}(K : \mathbb{Q})$ es, nuevamente, un elemento de $\mathrm{Gal}(K : \mathbb{Q})$.

El grupo $\mathrm{Gal}(K : \mathbb{Q})$ se llama *grupo de Galois* de K sobre $\mathbb{Q}$. En general, un grupo es un conjunto con una operación binaria —como una multiplicación, adición, composición o cualquier otra operación binaria— que tiene elemento neutro e inverso.

Cada elemento φ de $\mathrm{Gal}(K : \mathbb{Q})$ fija los coeficientes del polinomio $f(x)$, es decir, $\varphi(a_i) = a_i$, dado que $a_i \in \mathbb{Q}$. Por otro lado, el automorfismo φ induce una función entre los polinomios en $K[x]$:

$$b_m x^m + \ldots b_1 x + b_0 \mapsto \varphi(b_m)x^m + \ldots + \varphi(b_1)x + \varphi(b_0).$$

Esta función también conserva suma, multiplicación, elementos neutros e inversos aditivos de $K[x]$, dado que φ es un automorfismo que fija los elementos de K. Esto tiene una fuerte implicación si aplicamos φ a la ecuación:

$$f(x) = a_n x^n + \ldots + a_1 x + a_0 = a_n \cdot (x - s_1) \cdots \ldots (x - s_n),$$

entonces, el lado izquierdo no cambia. El lado derecho, $a_n \cdot (x - \varphi(s_1)) \cdot \ldots \cdot (x - \varphi(s_n))$ es una factorización de $f(x)$. Como ésta es única, concluimos que φ induce una *permutación* de las raíces $s_1, \ldots, s_n$.

Por otro lado, si $\varphi, \psi \in \mathrm{Gal}(K : \mathbb{Q})$ inducen la misma permutación en las raíces $s_1, \ldots, s_n$, entonces $\varphi = \psi$ —esto sigue de que K es el campo más pequeño que contiene $\mathbb{Q}$ y $s_1, \ldots, s_n$, un argumento que aquí no detallamos. Como sólo hay un número finito de permutaciones de $s_1, \ldots, s_n$ concluimos que $\mathrm{Gal}(K : \mathbb{Q})$ es un grupo *finito*. Por ejemplo, si $f(x) = x^2 + 1$, entonces $\mathrm{Gal}(K : \mathbb{Q})$ contiene dos elementos: la conjugación y la identidad.

Si $f(x) = x^4 + 1$, entonces los cuatro números:

$$z_1 = \frac{\sqrt{2}}{2}(1 + \mathbf{i}), \quad z_2 = \frac{\sqrt{2}}{2}(1 - \mathbf{i}), \quad z_3 = -z_1, \quad z_4 = -z_2,$$

son raíces de $f(x)$, como se comprueba al sustituir y, dado que $f(x)$ tiene grado 4, no puede tener más raíces. El campo de descomposición K contiene, por lo tanto, estos cuatro números pero también al cuadrado de z_1:

$$z_1^2 = \frac{2}{4}(1 + 2\mathbf{i} + \mathbf{i}^2) = \frac{1}{2}(1 + 2\mathbf{i} - 1) = \mathbf{i}$$

y además, la suma:

$$z_1 + z_2 = \sqrt{2}.$$

Por lo tanto, K contiene números de la forma $a_0 + a_1\sqrt{2} + a_2\mathbf{i} + a_3\sqrt{2}\mathbf{i}$, donde los coeficientes a_0, a_1, a_2 y a_3 son números racionales. No es difícil ver que, en efecto, K consiste exactamente en estos números: hay que ver que el conjunto formado por dichos números es cerrado bajo la multiplicación e inversos multiplicativos.

Veamos ahora cómo los elementos del grupo $\mathrm{Gal}(K : \mathbb{Q})$ permutan las raíces: si por ejemplo $\varphi(z_1) = z_3$, entonces $\varphi(z_3) = \varphi(-z_1) = \varphi(-1)\varphi(z_1) = -z_3 = z_1$, ya que $\varphi(-1) = -1$. Además, $\varphi(\mathbf{i}) = \varphi(z_1^2) = \varphi(z_1)^2 = (z_3)^2 = (-z_1)^2 = \mathbf{i}$. Entonces, $\varphi(z_1) = \frac{1}{2}\varphi(\sqrt{2}) + \frac{1}{2}\varphi(\sqrt{2})\mathbf{i}$, que debe ser igual a $-z_1 = -\frac{1}{2}\sqrt{2} - \frac{1}{2}\sqrt{2}\mathbf{i}$. Lo anterior muestra que $\varphi(\sqrt{2}) = -\sqrt{2}$. Con ello se concluye que $\varphi(z_3) = -z_3 = z_4$ y, por lo tanto, φ intercambia z_1 con z_3 y z_2 con z_4. De manera similar se puede mostrar que el automorfismo ψ que satisface $\psi(z_1) = z_2$ intercambia z_1 con z_2 y z_3 con z_4, y que el automorfismo η que satisface $\eta(z_1) = z_4$ intercambia z_1 con z_4 y z_2 con z_3. Con la identidad —que es el automorfismo que deja todo fijo— se obtiene un grupo de 4 automorfismos.

Se observa que $\varphi(\mathbf{i}) = \mathbf{i}$, $\psi(\sqrt{2}) = \sqrt{2}$ y $\eta(\sqrt{2}\mathbf{i}) = \sqrt{2}\mathbf{i}$. El conjunto de todos los números y que satisfacen $\varphi(y) = y$, forma un campo L que contiene $\mathbb{Q}$ y que es contenido en K. Se llama el *campo de puntos fijos* de φ. Más general, a los campos que contienen $\mathbb{Q}$ y que están contenidos en K se les llama *campos intermedios* y juegan un papel importante en el resto de esta sección. El campo de puntos fijos de φ es justamente el campo $\mathbb{Q}(\mathbf{i})$, que se llama el campo de los números de Gauss, en honor a Carl Friedrich Gauss. El campo de los puntos fijos de ψ es $\mathbb{Q}(\sqrt{2})$ y el de η es $\mathbb{Q}(\sqrt{2}\mathbf{i})$.

4.9.8 La teoría de Galois

La teoría de Galois parte de un polinomio *irreducible* $f(x) \in \mathbb{Q}[x]$, es decir, de un polinomio que no se puede factorizar como $g(x) \cdot h(x)$, donde $g(x)$ y $h(x)$ son polinomios *no constantes* en $\mathbb{Q}[x]$. Se denota con K el campo de descomposición de $f(x)$, es decir, el campo K más pequeño que contiene $\mathbb{Q}$.

La teoría de Galois establece una relación entre dos conjuntos. Por un lado, están los *subgrupos* de $\mathrm{Gal}(K : \mathbb{Q})$, es decir, los subconjuntos de $\mathrm{Gal}(K : \mathbb{Q})$ que son cerrados bajo la multiplicación y la formación de inversos —y que siempre contienen la identidad como elemento—. Por otro, están los campos L *intermedios*, es decir, los campos L que satisfacen $\mathbb{Q} \subset L \subset K$. La correspondencia se da por dos funciones r y s:

$$\{H \mid H \text{ es subgrupo de } \mathrm{Gal}(K : \mathbb{Q})\} \underset{s}{\overset{r}{\rightleftarrows}} \{L \mid L \text{ es campo intermedio } \mathbb{Q} \subset L \subset K\}$$

Es decir, r asigna a cada subgrupo de $\mathrm{Gal}(K : \mathbb{Q})$ un campo intermedio e, inversamente, la función s asigna a cada campo intermedio uno de estos subgrupos. La función r se define por:

$$r(H) = \{z \in K \mid \forall \varphi \in H, \varphi(z) = z\},$$

es decir, $r(H)$ es el conjunto de números $z \in K$ que permanecen fijos bajo cualquier φ de H. Por otro lado, se define que:

$$s(L) = \{\varphi \in \mathrm{Gal}(K : \mathbb{Q}) \mid \forall z \in L, \varphi(z) = z\},$$

Figura 4.59 Evariste Galois (1811-1832) | © Latin Stock México.

es decir, $s(L)$ es el conjunto de automorfismos $\varphi \in \mathrm{Gal}(K:\mathbb{Q})$ tal que cualquier $z \in L$ permanece fijo bajo φ. No es difícil ver que r y s son funciones inversas una de la otra.

Las contenciones se invierten bajo r y s: si $H \subset H'$ son subgrupos de $\mathrm{Gal}(K:\mathbb{Q})$, entonces $r(H) \supset r(H')$ y viceversa, si $L \subset L'$ son dos campos intermedios, entonces $s(L) \supset s(L')$. Al subgrupo $H = \{\mathrm{id}\}$, que contiene solamente la identidad y que es el subgrupo más pequeño, corresponde el campo intermedio más grande, que es K mismo, mientras que al subgrupo más grande, que es $\mathrm{Gal}(K:\mathbb{Q})$, corresponde el campo intermedio más pequeño, que es $\mathbb{Q}$.

Así, en el caso del polinomio $f(x) = x^4 + 1$ se ve que los cálculos anteriores ejemplifican la teoría de Galois, como se ve en la figura 4.58.

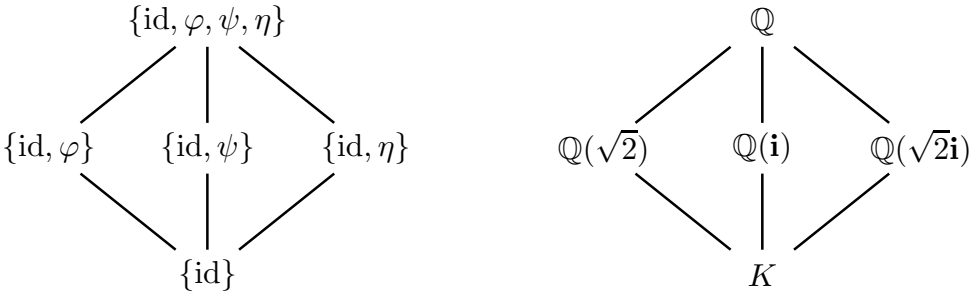


Figura 4.60 Del lado izquierdo, la correspondencia entre subgrupos del grupo de Galois, y del derecho los campos intermedios.

Se puede mostrar que el polinomio $f(x) = x^5 + x + 1$ es irreducible con cinco raíces distintas $s_1, \ldots, s_5$ y el grupo de Galois $\mathrm{Gal}(K:\mathbb{Q})$ es todo el grupo de permutaciones de las cinco raíces $s_1, \ldots, s_5$. En general, para cada n es posible encontrar un polinomio irreducible $f(x) \in \mathbb{Q}[x]$ de grado n tal que el grupo $\mathrm{Gal}(K:\mathbb{Q})$ es todo el grupo de permutaciones de las n raíces distintas.

Por ello, es importante estudiar los subgrupos del grupo de permutaciones de n elementos distintos $s_1, \ldots, s_n$. Dicho estudio es el inicio de la teoría de grupos que se estudian por sí mismos desde entonces, sin conexión con el origen de la resolución de ecuaciones. El grupo de permutaciones tiene una estructura que no depende de cuáles sean los n elementos a permutar. Por consiguiente, las podemos elegir como los números $1, 2, \ldots, n$ para facilitar la notación. El grupo resultante se denota por S_n y se llama *grupo simétrico*. A continuación, resumimos algunos de los hechos de la teoría de grupos que serán de relevancia en lo que sigue.

Una *transposición* es una permutación que deja fijos todos los elementos, salvo dos que intercambia. Si i y j son los dos elementos intercambiados, la transposición se denota por $(i\,j)$. Cada permutación se puede escribir como composición de transposiciones. El nú-

mero de transposiciones usadas en una composición no es invariante, por ejemplo, se tiene que:

$$(1\,2)(2\,3)(1\,2) = (1\,3).$$

Sin embargo, la paridad del número de transposiciones usadas sí es una invariante: una composición de 3 transposiciones no se puede escribir como composición de un número par de transposiciones. Por ello, el conjunto de composiciones A_n de un número par de transposiciones forma un subgrupo, llamado el *grupo alternante*.

El subgrupo A_n es un *subgrupo normal* de S_n, es decir, para cada $\alpha \in \mathrm{A}_n$ y cada $\sigma \in \mathrm{S}_n$ se tiene $\sigma^{-1}\alpha\sigma \in \mathrm{A}_n$. Si $n = 4$, entonces A_4 tiene un subgrupo normal, a la vez, que es el *subgrupo de Klein*:

$$K = \{\mathrm{Id},\ (1\,2)(3\,4),\ (1\,3)(2\,4),\ (1\,4)(2\,3)\}.$$

Éste es justamente el grupo de Galois $\mathrm{Gal}(K : \mathbb{Q})$ si K es el campo de descomposición del polinomio $f(x) = x^4 + 1$, el ejemplo que hemos seguido más de cerca hasta ahora.

Aquí se debe tener cuidado: K es *subgrupo no normal* de S_4 pero sí es normal en A_4. El grupo K tiene, a la vez, tres subgrupos normales: $Z_1 = \{\mathrm{Id},\ (1\,2)(3\,4)\}$, $Z_2 = \{\mathrm{Id},\ (1\,3)(2\,4)\}$ y $Z_3 = \{\mathrm{Id},\ (1\,4)(2\,3)\}$. Si H es subgrupo normal de H', se denota por $H \lhd H'$. Con ello tenemos la siguiente sucesión de subgrupos normales:

$$\{\mathrm{Id}\} \lhd Z_1 \lhd K \lhd \mathrm{A}_4 \lhd S_4.$$

Ésta es una *sucesión de composición* de S_4 donde ya no es posible insertar más subgrupos en medio sin tener igualdad. Si escribimos la cardinalidad en lugar de los grupos, obtenemos la siguiente sucesión de divisores $1 \lhd 2 \lhd 4 \lhd 12 \lhd 24$. Los factores entre dos números sucesivos son primos.

En la teoría de grupos es posible formar cocientes de subgrupos si éstos son normales y, además, la cardinalidad de estos grupos cocientes es justo el factor de los números consecutivos correspondientes. Más aún, cualquier grupo con un número primo de elementos es *conmutativo*. Los grupos conmutativos se llaman también *abelianos*, en honor a Abel, quien enfatizó la importancia de este hecho para la resolución de las ecuaciones.

Las siguientes cadenas muestran sucesiones de composición de S_2 y S_3:

$$\{\mathrm{Id}\} \lhd \mathrm{A}_2 \lhd \mathrm{S}_2, \quad \{\mathrm{Id}\} \lhd \mathrm{A}_3 \lhd \mathrm{S}_3\,.$$

Las cardinalidades son $1 \lhd 2 \lhd 4$ y $1 \lhd 3 \lhd 6$, respectivamente. En consecuencia, $\mathrm{S}_2\mathrm{S}_3$ y S_4 tienen sucesiones de composición con factores abelianos. También S_5 y, más general, S_n para $n \geq 5$ tiene una sucesión de composición:

$$\{\mathrm{Id}\} \lhd \mathrm{A}_n \lhd \mathrm{S}_n,$$

pero A_n *no es conmutativo*. Los únicos subgrupos normales de A_n son $\{\mathrm{Id}\}$ y A_n. Es fácil ver que, A_n para $n \geq 5$, no es conmutativo. La importancia de esta diferencia reside en lo que se explica a continuación.

La teoría de Galois establece una muy buena correspondencia entre los subgrupos de $\mathrm{Gal}(K : \mathbb{Q})$ y los campos intermedios $\mathbb{Q} \subset L \subset K$: el subgrupo $H \subset H'$ es normal en

H' si y solamente si $L = r(H)$ es una *extensión normal* de $L' = r(H)$, es decir, cada $g(x) \in L'[x]$ que tiene una raíz en L, tiene todas las raíces en L. Además, en este caso podemos considerar el cociente H'/H. Este cociente es *cíclico* —es decir, es isomorfo al grupo dado por $\{1, \ldots, n\}$ bajo la adición módulo n— si y solamente si L es una extensión *cíclica* de L', es decir existe un elemento $a \in L'$ tal que L es el campo de descomposición del polinomio $x^n - a \in L'[x]$. En otras palabras, todos los elementos de L se pueden escribir como combinaciones de $1, \sqrt[n]{a}, \sqrt[n]{a}^2, \ldots, \sqrt[n]{a}^{n-1}$ con coeficientes en L.[1]

Si se junta toda esta información, se obtiene que el hecho de que $\mathrm{Gal}(K : \mathbb{Q})$ tiene una sucesión de composición con factores abelianos corresponde a que las soluciones de la ecuación $f(x) = 0$ se pueden escribir con números racionales, las cuatro operaciones básicas y raíces. El grupo S_n no es soluble, es decir, no tiene una sucesión de composición con factores abelianos. Consecuentemente, un polinomio $f(x)$ cuyo grupo de Galois $\mathrm{Gal}(K : \mathbb{Q})$ es isomorfo a S_n, no será *soluble por radicales*, lo cual significa que sus raíces no se podrán escribir usando los coeficientes, las operaciones básicas y raíces cuadradas. Como hemos visto, hay siempre polinomios con esa propiedad. Así se establece una visión más profunda de la imposibilidad de poder dar una fórmula de solución para las ecuaciones algebraicas de grado mayor que 4.

Es una coincidencia triste que, tanto Abel como Galois, murieran a temprana edad. Niels Henrik Abel nació en Noruega en 1802, pero era muy pobre y sólo gracias a una beca pudo viajar a Europa para presentar sus brillantes ideas en los centros de investigación matemática. Viajó a Berlín y luego a París, pero no obtuvo una recepción favorable ni logró despertar el interés de los líderes matemáticos de aquel tiempo, como Gauss, Legendre o Cauchy. En París contrajo tuberculosis y murió a la edad de 27 años, dos días antes de que llegara la carta de Berlín que le informaba que había conseguido un puesto de profesor.

Figura 4.61 Sello conmemorativo por los 100 años de la muerte de Evariste Galois (1811-1832), matemático francés que murió con tan sólo 20 años en un duelo. Años más tarde, otro matemático alemán —Felix Klein— diría: "En Francia apareció hacia 1800 una nueva estrella de inimaginable brillo en el firmamento de las matemáticas: Evariste Galois".

Galois nació en 1811, en Francia, y murió a los 20 años en un duelo en el que había retado a un oficial, por motivos desconocidos. Vivió en tiempos turbios y difíciles. Lo que se sabe de su corta vida habla de que era una persona rebelde. Fue expulsado de la escuela, estuvo en prisión y lideró protestas. La noche anterior al duelo, escribió una última carta donde reportó y resumió frenéticamente sus ideas matemáticas, a sabiendas de que la muerte lo esperaba al principio del día siguiente. A Galois debemos la idea de considerar varias permutaciones de las raíces a la vez y, con ello, toda la teoría de grupos.

[1] Esto da la idea correcta, pero habría que precisarla: sólo es correcta si se cambia el campo $\mathbb{Q}$ por un campo k que contiene $\mathbb{Q}$ y todas las soluciones de $x^n - 1 = 0$. Cabe mencionar que lo anterior no afecta la cuestión de ser resoluble por radicales, es decir, de poder escribir una raíz de un polinomio con las operaciones básicas y raíces cuadradas.

Hemos recorrido un camino largo y sinuoso. Primero, necesitábamos delimitar la pregunta y llegamos a la noción de ecuaciones algebraicas. Expandimos la noción de los números reales a los números complejos para poder resolver cualquier ecuación cuadrática. Después, consideramos la ecuación de tercer grado y vimos que, por el teorema fundamental del álgebra, la ecuación algebraica de cualquier grado positivo siempre tiene soluciones. El problema consiste en encontrar una fórmula para las soluciones a partir de los coeficientes que aparecen en la ecuación. Finalmente, descubrimos que tal fórmula no existe por el teorema de Abel-Ruffini y que la teoría de Galois aporta una explicación profunda de este hecho. Esta teoría requiere de considerable abstracción pero es, indudablemente, una de las joyas del álgebra. Así concluimos el resumen sobre una de las cuatro preguntas fundamentales que consideraremos en este capítulo.

4.10 ¿QUÉ SE PUEDE CONSTRUIR?

Figura 4.62 El compás permite medir distancias y trazar circunferencias. En esta ilustración del Codex Vindobonensis 2554*, una Biblia ilustrada, el compás es la herramienta usada por Dios en el acto de creación. Con ello se expresa la importancia del compás como herramienta geométrica y de la geometría como disciplina del pensamiento.*

4.10.1 Delimitación de la pregunta

Los científicos griegos estaban fascinados por la posibilidad de realizar construcciones con las herramientas más básicas: la regla y el compás. La regla sirve simplemente para trazar pedazos de rectas, mientras el compás sirve para trazar circunferencias. ¿Por qué en muchos de sus estudios se habrán restringido a estas dos herramientas? ¿Qué tienen en particular la recta y la circunferencia?

La circunferencia con centro C y radio r es el conjunto de puntos que están a distancia r del punto C. Las circunferencias miden así lo que está cerca o lejos de un punto y en ellas se encuentra codificada la distancia. En la recta es diferente pues los griegos sólo consideraban pedazos finitos de rectas: *segmentos*. Un segmento marca el camino más corto entre sus dos extremos. ¿Cómo habrá entonces que interpretar la prolongación de un segmento sobre una de sus dos extremidades? La recta marca cómo seguir "la ruta marcada" sin desviarse hacia la derecha o la izquierda.

Sobre superficies curvas como una esfera o una dona ya no podemos hablar de circunferencias o rectas, pero sí podemos hablar del conjunto de puntos que están a una distancia

dada de un punto C o del camino más corto entre dos puntos. En el toro nos damos cuenta rápidamente de que la noción del camino "más corto" no siempre se puede prolongar sobre los extremos. Por ello, mejor se habla de las *geodésicas*, que son las curvas que tienen la propiedad de que para puntos cercanos sobre ellas sí marcan los caminos más cortos.

4.10.2 Los problemas clásicos

La circunferencia y la recta son entonces conceptos muy fundamentales en la geometría plana. Alrededor de 300 a.C. se escribió *Los elementos*, un libro de gran influencia, en donde se reunió gran parte de lo que se sabía dentro de un marco común, que es la construcción con regla y compás. La fascinación griega por los problemas geométricos data de muchos años atrás. Algunos de ellos se hicieron tan famosos que ahora se llaman "clásicos" y son los siguientes:

- *Duplicación del cubo: hallar con regla y compás la longitud del lado de un cubo que tiene el doble del volumen de un cubo cuyo lado esté dado.*
- *Trisección del ángulo: encontrar con regla y compás el tercio de un ángulo dado.*
- *Construcción de polígonos regulares: construir con regla y compás la longitud de un lado de un polígono regular con n lados inscrito en una circunferencia dada.*
- *Cuadrar un círculo: hallar con regla y compás la longitud del lado de un cuadrado que tenga la misma área que una circunferencia cuyo radio sea conocido.*

La expresión "hallar con regla y compás" merece una aclaración: lo que se busca es una lista de instrucciones para construir lo indicado al usar solamente la regla —con la cual se puede trazar una recta por dos puntos ya conocidos— y el compás —con el cual se puede construir la circunferencia con centro dado que pasa por un punto ya conocido. Nuevos puntos sólo se obtienen como intersección de dos objetos —rectas o circunferencias— ya construidas. Para hallar la circunferencia inscrita en un triángulo dado, no se pueden dibujar una serie de circunferencias probando con diferentes centros y radios para aproximarse, poco a poco, a la solución; lo que debe indicarse es cómo se obtienen el centro y el radio para cualquier triángulo dado, no sólo uno en particular.

En *Los elementos* se procede paso a paso al explicar cómo resolver tareas cada vez más complicadas, pero no se resuelve ninguno de los problemas clásicos. Cada uno de estos problemas tiene una historia larga e interesante de muchos intentos fallidos. En la primera mitad del siglo XIX, se demuestra que los primeros tres de estos problemas no tienen solución, y en la segunda mitad del mismo siglo se concluye que tampoco la cuadratura del círculo tiene solución. Para ello, intervinieron matemáticos como Carl Friedrich Gauss, Evariste Galois, Pierre Wantzel y Ferdinand von Lindemann.

El adjetivo "que no tiene solución" tiene un carácter definitivo: no hay y no habrá nunca. No es que los matemáticos no han buscado lo suficientemente bien —lo hicieron por más de dos mil años y, a veces, hasta con una dedicación feroz—, sino que no puede haber solución si uno se restringe a la regla y al compás como las únicas herramientas de construcción. Desde la época de Euclides se conocen varias soluciones a los primeros tres problemas si se elimina la restricción de hacerlo sólo con regla y compás; por ejemplo, con una herramienta que traza parábolas. Con la herramienta que se muestra en la figura 4.68 se puede trisecar un ángulo con facilidad.

La herramienta consiste en varas y articulaciones en forma similar al compás. Para trisecar un ángulo $\measuredangle EFG$ se coloca el punto C en F y, luego, se ajustan las dos manecillas A

y B sobre la recta FE y FG, respectivamente. Las puntas P y Q indican entonces la ubicación de los rayos que trisecan el ángulo dado.

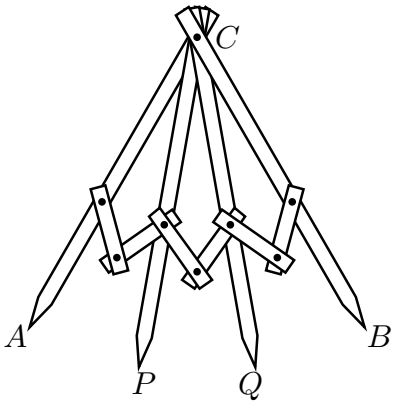


Figura 4.63 Herramienta de construcción para trisecar ángulos.

La dificultad de los problemas clásicos reside en la restricción de las herramientas a, únicamente, regla y compás. En lo que sigue, daremos una idea sobre cómo es posible llegar a la conclusión de que no existe solución a los problemas clásicos. Para ello, usaremos los *números complejos* de la sección 4.2. Cada construcción comienza con unos datos iniciales, como una circunferencia o un triángulo dado. Empezaremos aquí con el mínimo absoluto: a partir de dos puntos dados.

4.10.3 El plano complejo como modo algebraico

El punto clave es usar, como modelo para el plano, el *plano complejo*, véase sección 4.2. Podemos pensar que los dos puntos dados son los números 0 y 1. ¿Qué otros números del plano complejo se pueden construir a partir de estos dos? A continuación, veremos que el conjunto de puntos $\mathscr{C}$ que se puede construir forma —visto como números complejos— un *campo*, es decir, los números en $\mathscr{C}$ se pueden sumar, restar, multiplicar y dividir. Antes de ello, observemos que se pueden construir perpendiculares con regla y compás si damos un punto P y una recta ℓ, pues se debe construir la perpendicular k a ℓ que pasa por P.

Como la figura 4.64 sugiere, hay que distinguir el caso en que P pertenece a ℓ del caso en que no pertenece. La figura indica los cuatro pasos en ambos casos.

También podemos construir *paralelas*: la paralela a la recta ℓ que pasa por P se obtiene al construir la perpendicular a k que pasa por P, donde k es la perpendicular a ℓ que pasa por P. Por la interpretación geométrica de la suma de números complejos (véase sección 4.9.2), es posible construir el número $\mathbf{x} + \mathbf{y}$ si ya se tienen construidos $\mathbf{x}$, $\mathbf{y}$ con anteriori-

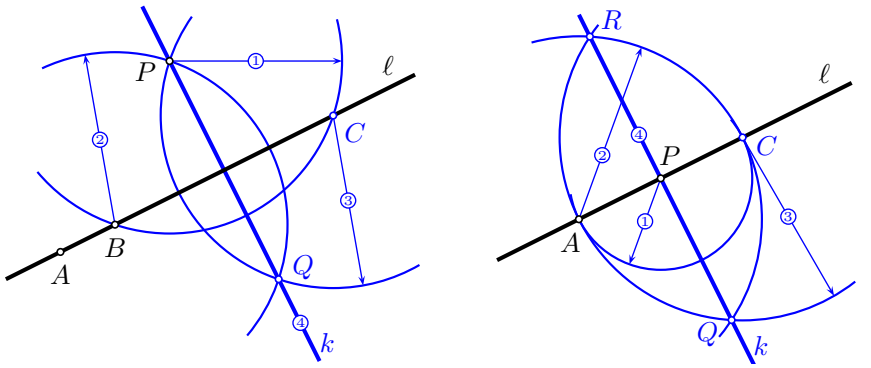
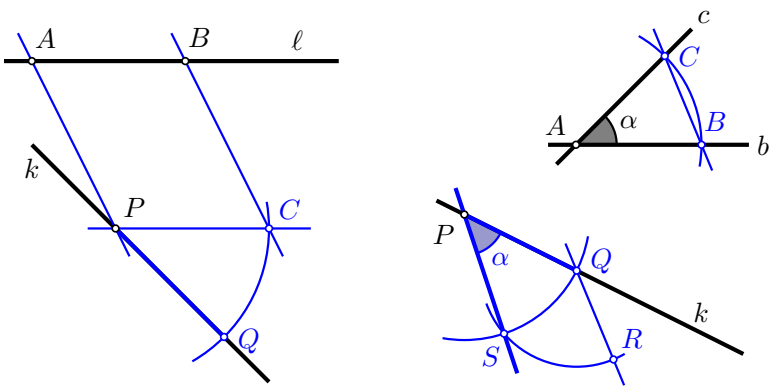


Figura 4.64 Indicación de cómo construir perpendiculares. A la izquierda para puntos P fuera y a la derecha para puntos P sobre la recta ℓ.

dad: ahora sólo necesitamos intersecar la paralela a $0\mathbf{y}$ por $\mathbf{x}$ con la paralela a $0\mathbf{x}$ por $\mathbf{y}$. En otras palabras, si $\mathbf{x}, \mathbf{y}$ pertenecen a $\mathscr{C}$, entonces también $\mathbf{x}+\mathbf{y}$ lo hará.

Por otro lado, podemos construir $-\mathbf{x}$ si ya está construido $\mathbf{x}$, dado que $-\mathbf{x}$ es simplemente el reflejo de $\mathbf{x}$ en el punto 0. Consecuentemente, si $\mathbf{x}, \mathbf{y}$ pertencen a $\mathscr{C}$, entonces también $-\mathbf{y}$ y, por lo tanto, también $\mathbf{x}-\mathbf{y}=\mathbf{x}+(-\mathbf{y})$.

Antes de revisar la multiplicación, analicemos el compás como herramienta. Un compás puede levantarse de la hoja de papel y, sin cambiar su apertura, trazar una circunferencia en otro lugar con el mismo radio. Parece que el compás es una herramienta más poderosa de lo que pensábamos —que, simplemente, construir rectas y circunferencias a partir de puntos ya construidos—. Pero no es así: podemos usar solamente rectas y circunferencias para implementar toda la funcionalidad del compás, como lo muestra la figura 4.65 a la izquierda.



4.65 Indicaciones de cómo transportar distancias y ángulos de un lugar a otro sin el uso del compás.

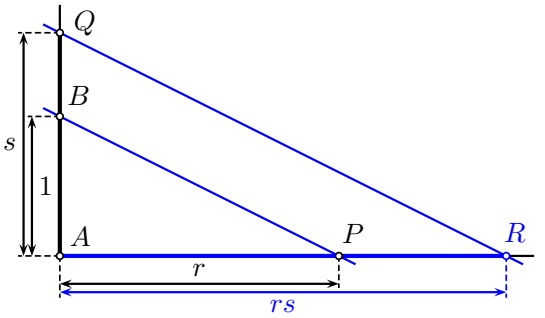
En esta figura se *transportó* la distancia entre los puntos A y B a la recta k, partiendo del punto dado P. Como resultado, se obtiene el punto Q sobre k, que está a la misma distancia que B de A. Del lado derecho se muestra cómo se pueden transportar ángulos. Si se quiere transportar el ángulo α entre dos rectas b y c, al punto P a partir de la recta k, entonces, primero se construye el punto Q —al transportar la distancia AC a P—, luego el punto R —al transportar la distancia CB a una paralela a BC por Q— y, finalmente, se interseca la circunferencia con centro P que pasa por Q con la circunferencia con centro Q que pasa por R, para obtener el punto S y $\measuredangle QPS = \measuredangle CAB = \alpha$.

Claro es que los números enteros son construibles si se transporta, por ejemplo, el segmento s con extremos 0 y 1 a la derecha de 1, para encontrar el punto 2. Luego se transporta s a la derecha de 2, para obtener 3.

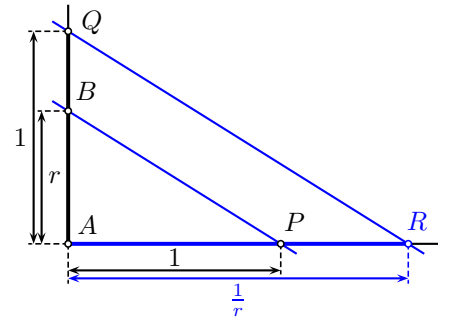
Veamos ahora la multiplicación. Según la interpretación geométrica que se observa en la figura 4.66, debemos sumar los ángulos y multiplicar los valores absolutos. Ambas cosas son bastante fáciles de lograr con regla y compás: los ángulos se suman con el método para transportar explicado arriba y el producto de los valores absolutos se obtiene por semejanza, como en la figura 4.67.

A continuación, veamos la última de las cuatro operaciones básicas: la división. Como en la resta, es suficiente mostrar que podemos construir con regla y compás el inverso multiplicativo de un número construible $\mathbf{z}$ dado. Con tal fin usamos la representación geométrica: debemos construir el negativo del ángulo de $\mathbf{z}$ y el inverso multiplicativo $\frac{1}{r}$, de su valor absoluto r. Esto se hace de nuevo por semejanza, como lo indica la figura 4.67.

Como consecuencia, podemos construir todos los números racionales, es decir, $\mathbb{Q} \subset \mathscr{C}$, si $\mathbb{Q}$ denota al conjunto de los números racionales.

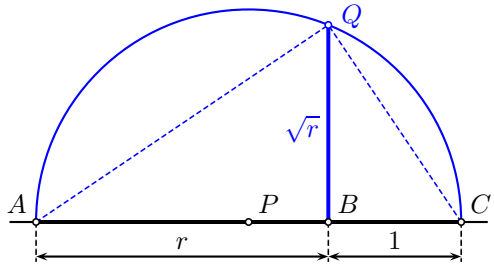


4.66 Construcción para encontrar el producto rs de dos números reales r y s positivos.



4.67 Construcción para encontrar el inverso multiplicativo $\frac{1}{r}$ de un números real r positivo.

Por último, observemos que si un número complejo $\mathbf{z}$ es construible, entonces también lo serán sus dos raíces cuadradas $\pm\sqrt{\mathbf{z}}$; este hecho se suele abreviar diciendo que los números construibles son "cerrados bajo la operación de sacar la raíz cuadrada". Otra vez recurrimos a la interpretación geométrica de sacar raíces, según lo hicimos en la figura 4.67. Debemos dividir el ángulo de $\mathbf{z}$ en dos —que es fácil con regla y compás— y obtener la raíz cuadrada del valor absoluto de $\mathbf{z}$, como se indica en la figura 4.68.



4.68 Construcción para encontrar la raíz cuadrada $\sqrt{r}$ de un número real r positivo.

En efecto, por semejanza de los triángulos ΔABQ y ΔQBC, sigue que la altura $x = BQ$ debe satisfacer $x : r = 1 : x$, es decir, $x^2 = r$ y por ello $x = \sqrt{r}$. Con ello, hemos llegado al primer resultado importante de esta sección:

Teorema 1: Los números construibles $\mathscr{C}$ forman un *campo* que es cerrado bajo la operación de sacar raíces cuadradas.

4.10.4 Descripción alterna del campo de los números construibles

Del resultado anterior se deduce que cualquier número complejo que se puede escribir como una expresión que involucre solamente números racionales, las cuatro operaciones básicas y raíces cuadradas, es construible. Denotamos por $\mathscr{N}$ al conjunto de los números que se pueden escribir así. Por ejemplo:

$$\frac{8-2\sqrt{3}}{\sqrt{-2-\sqrt{\frac{4}{7}}}}-\sqrt{\frac{3}{17}-\frac{1+4\sqrt{\frac{171}{3}}}{\sqrt{2}-\sqrt{-2}}}$$

es uno de los números en $\mathscr{N}$. Hemos demostrado que cada número en $\mathscr{N}$ pertenece a $\mathscr{C}$, es decir, $\mathscr{N} \subset \mathscr{C}$.

Ahora veamos que cada número construible es, en efecto, un número en $\mathscr{N}$ o lo, que es lo mismo, que $\mathscr{C} \subset \mathscr{N}$ y, por ello, $\mathscr{C} = \mathscr{N}$. Demostraremos:

$$\text{si } \mathbf{z} = a + b\mathbf{i} \in \mathscr{C}, \text{ entonces } a, b \in \mathscr{N}. \tag{34}$$

De ahí seguirá que $\mathbf{z} = a + b\sqrt{-1}$ pertenece a $\mathscr{N}$, es decir $\mathscr{C} \subset \mathscr{N}$.

Para ello, primero observamos que si un número complejo $\mathbf{x} = a + b\mathbf{i}$ es construible, entonces también lo son la parte real a y la parte imaginaria b, y viceversa: si a y b son dos *números reales* construibles, entonces también $a + b\mathbf{i}$ es construible. La parte real y la parte imaginaria de los números complejos sirven como las dos coordenadas en el plano. Lo anterior nos permite elaborar argumentos con técnicas de geometría analítica.

Los números que son construibles se organizan en *niveles* $\mathscr{C}_N$. En el primer nivel $\mathscr{C}_1$ están solamente los dos números iniciales 0 y 1. En el segundo nivel $\mathscr{C}_2$ se encuentran los números del primer nivel más aquellos otros que se puedan obtener a partir de intersecciones de rectas y circunferencias definidas por números del nivel anterior. Por ejemplo, $\frac{1}{2} + \frac{\sqrt{3}}{2}\mathbf{i}$ es un número del nivel 1 —es una de las dos intersecciones de la circunferencia con centro 0 que pasa por el punto 1, con la circunferencia con centro 1 que pasa por el punto 0—. Inductivamente, los puntos del nivel N son los puntos del nivel $N-1$ más aquellos puntos que se pueden obtener como intersección de rectas o circunferencias definidas a partir de puntos del nivel anterior.

Como cada punto construible se obtiene por una sucesión finita de pasos de construcción, tenemos que para cada $\mathbf{z} \in \mathscr{C}$ existe un N tal que $\mathbf{z} \in \mathscr{C}_N$. Ahora hemos preparado las bases para una *demostración por inducción*: demostramos (34) por inducción sobre N, es decir demostramos que:

$$\text{si } \mathbf{z} = a + b\mathbf{i} \in \mathscr{C}_N, \text{ entonces } a, b \in \mathscr{N} \tag{35}$$

para los diferentes N, uno tras otro. El inicio de la inducción es cuando $N = 1$. Como $\mathscr{C}_1 = \{0, 1\}$, tenemos que $\mathscr{C}_1 \subset \mathscr{N}$ y con ello (35) para $N = 1$.

Ahora, si suponemos que ya demostramos (35) para $N = L - 1$ y argumentamos que (35) también satisface para $N = L$.

Sea $\mathbf{z} = x + y\mathbf{i} \in \mathscr{C}_L$. Entonces, existen dos objetos ℓ y k —rectas o circunferencias—, dadas por puntos en $\mathscr{C}_{L-1}$: el objeto ℓ está dado por $\mathbf{P} = a + b\mathbf{i}$ y $\mathbf{Q} = c + d\mathbf{i}$ mientras que el objeto k está definido por los puntos $\mathbf{P}' = a' + b'\mathbf{i}$ y $\mathbf{Q}' = c' + d\mathbf{i}$. Si ℓ es la recta que pasa por los puntos $\mathbf{P}$ y $\mathbf{Q}$, entonces ℓ se representa por la ecuación:

$$(a - c)y + bc = (b - d)x + ad, \tag{36}$$

y si ℓ es la circunferencia con centro $\mathbf{P}$ que pasa por $\mathbf{Q}$, entonces ℓ se representa por la ecuación:

$$(x - a)^2 + (y - b)^2 = (c - a)^2 + (d - b)^2. \tag{37}$$

Ecuaciones similares definen al objeto k. Ahora tenemos que distinguir tres casos.

Caso 1: cuando ℓ y k *son rectas*. Aquí, las ecuaciones que representan ℓ y k definen un sistema de dos ecuaciones con dos incógnitas x y y. La solución está dada por expresiones que involucran los números $a, b, c, d, a', b', c', d'$ y las cuatro operaciones básicas, en concreto:

$$x = \frac{(a'-c')(bc-ad)-(a-c)(b'c'-a'd')}{(a'-c')(b-d)-(a-c)(b'-d')},$$
$$y = \frac{(b'-d')(ad-bc)-(b-d)(a'-d'-b'c')}{(b'-d')(a-c)-(b-d)(a'-c')}.$$

Como $a, b, c, d, a', b', c', d'$ están en $\mathscr{N}$ por hipótesis de inducción —porque $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{P}'$, $\mathbf{Q}' \in \mathscr{C}_{L-1}$—, tenemos que $x, y \in \mathscr{N}$ por definición de los números que pertenecen a $\mathscr{N}$.

Caso 2: cuando ℓ *es recta* y k *es una circunferencia, similar al caso en que* k *es recta y* ℓ *circunferencia*. Si $b \neq d$, se puede despejar x de la ecuación (36) y sustituirlo en la ecuación que representa a k. Con ello se obtiene una ecuación cuadrática en y de la forma:

$$Ay^2 + By + C = 0 \tag{38}$$

donde:

$$A = (a-c)^2 + (b-d)^2$$
$$B = 2(a-c)\,[bc - ad - a'(b-d)] + 2(b-d)^2 b'$$
$$C = [bc - ad - a'(b-d)]^2 + (b-d)^2[b'^2 - (c'-a')^2 - (d'-b')^2]$$

Nuevamente, $a, b, c, d, a', b', c', d' \in \mathscr{N}$ por hipótesis de inducción y, con ello, $A, B, C \in \mathscr{N}$ por definición de la forma de los números que pertenecen a $\mathscr{N}$. Por lo tanto, también la solución de (38):

$$y = \frac{-B \pm \sqrt{B^2 - 4AC}}{2A} \tag{39}$$

pertenece a $\mathscr{N}$. La sustitución de (39) da una ecuación cuadrática en x con coeficientes en $\mathscr{N}$. Consecuentemente, x pertenece a $\mathscr{N}$.

Si $b = d$, entonces se puede despejar y de la ecuación (36) y se obtiene que:

$$y = \frac{ad - bc}{a - c} = b,$$

que está en $\mathscr{N}$ por hipótesis de inducción. La sustitución en la ecuación que representa k da:

$$(x-a')^2 - (b-b')^2 = (c'-a')^2 + (d'-b')^2,$$

que es cuadrática en x con coeficientes en $\mathscr{N}$. Por lo tanto, las soluciones están en $\mathscr{N}$.

Caso 3: cuando ℓ y k *son circunferencias*. En vez de trabajar con el sistema de las dos ecuaciones que representan ℓ y k, se trabaja con el sistema de ℓ y m, donde m es la diferencia entre ℓ y k:

$$m: \quad (x-a)^2-(x-a')^2+(y-b)^2(y-b')^2=(c-a)^2-(c'-a')^2+(d-b)^2-(d'-b')^2,$$
$$m: \quad (2a-2a')x+(2b-2b')y+(a^2-a'^2+b^2-b'^2)=(c-a)^2-(c'-a')^2+(d-b)^2-(d'-b')^2$$

La segunda línea muestra la ecuación después de expandir y simplificar el lado izquierdo. Se ve que es lineal en las variables x, y, por lo que se puede proceder como en el caso 2. La ecuación m representa la recta que pasa por los dos puntos de intersección de ℓ y k, o la tangente común si ℓ y k son tangentes en Z.

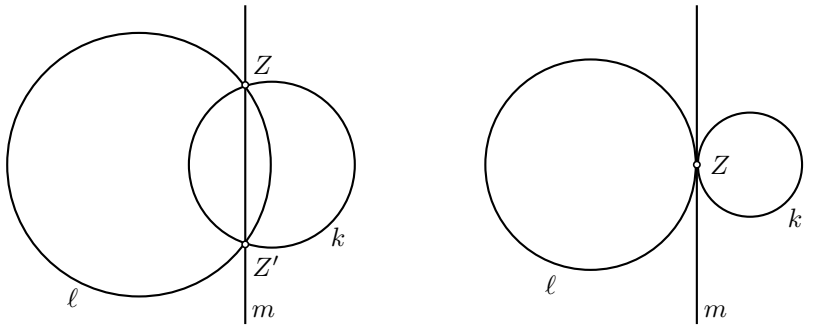


Figura 4.69 Definición de la recta m de dos circunferencias ℓ y k que se intersecan.

Con todo lo anterior, demostramos el segundo resultado importante:

Teorema 2: los números construibles son precisamente aquellos que se pueden escribir a partir de números racionales usando un número finito de las cuatro operaciones básicas y raíces cuadradas.

4.10.5 Sobre la imposibilidad de resolver los problemas clásicos

Con los dos resultados anteriores demostrados, tenemos las bases para indicar los razonamientos para deducir que los cuatro problemas clásicos no se pueden resolver.

En la descripción dada del teorema 2 vemos que cualquier número construible es un *número algebraico*, es decir, un número que es solución de alguna *ecuación algebraica* con coeficientes racionales, como se vio en la sección 4.9.

En 1882, Ferdinand von Lindemann demostró que el número π es *trascendental*, o lo que es lo mismo, que no es solución de ninguna ecuación algebraica con coeficientes racionales. Si se quiere cuadrar el círculo, lo que se busca es encontrar la longitud del lado de un cuadrado que tiene la misma área que la de un círculo dado, por ejemplo, el círculo con radio 1, que posee un área de π, como se vio en la sección 2.9. Por lo tanto, se busca construir la longitud $\sqrt{\pi}$. Si pudiéramos construir esta longitud, entonces también podríamos hacerlo con su cuadrado, es decir π, lo que es imposible por el teorema de Lindemann. Así se muestra que no es posible cuadrar el círculo.

La imposibilidad de cuadrar el círculo es la última de las cuatro demostraciones, pero la indicamos primero dado que, conceptualmente, es la más clara. Veamos ahora la duplicación del cubo.

Si tenemos un cubo dado con aristas de longitud ℓ, entonces el volumen es de ℓ^3. El cubo que tiene el doble de volumen presenta aristas de longitud $\sqrt[3]{2}\ell$. Para hallar esta longitud con regla y compás, debemos construir $\sqrt[3]{2}$ dado que podemos multiplicar. Sin embargo, según el teorema 2 sólo podemos construir raíces cuadradas, no cúbicas. Pero esto no es su-

ficiente: ¿por qué no es posible que alguna expresión que sólo involucra raíces cuadradas sea igual a $\sqrt[3]{2}$? Para resolver lo anterior, necesitamos algunas herramientas de la teoría de *campos*.

En la sección 4.9 consideramos campos intermedios como $\mathbb{Q} \subset K \subset \mathbb{C}$, es decir, campos K contenidos en los números complejos $\mathbb{C}$ y que también contienen los números racionales $\mathbb{Q}$. Un campo así surgió como *campo de descomposición* de algún polinomio $f(x)$ y siempre tiene una *dimensión sobre* $\mathbb{Q}$. La dimensión de campos generaliza la idea del "número de coordenadas que se usan para describir una ubicación": en la recta basta una coordenada, en el plano se requieren dos y en el espacio tres. Para usar coordenadas se requiere de un origen y de ejes.

Por ejemplo, en los números complejos tenemos dos ejes: el real y el imaginario. Cualquier número complejo $a + b\mathbf{i}$ se expresa con dos coordenadas reales a y b. En otras palabras, cualquier número complejo es combinación con coeficientes reales de los números complejos 1 e $\mathbf{i}$.

Si consideramos ahora una situación de un campo de descomposición K sobre $\mathbb{Q}$, entonces fijamos $k_1, \ldots, k_d \in K$ y consideramos *combinaciones lineales*, como en:

$$q_1k_1 + q_2k_2 + \ldots q_nk_n, \tag{40}$$

con coeficientes $q_1, \ldots, q_d \in \mathbb{Q}$. La *dimensión* de K sobre $\mathbb{Q}$ es, entonces, el mínimo número d tal que existen elementos $k_1, \ldots, k_d \in K$ llamados *base*, con la propiedad de que cualquier elemento de K es combinación lineal (40) para algunos coeficientes $q_1, \ldots, q_d$. Esta dimensión se denota por $\dim_{\mathbb{Q}} K$. Se demuestra que, cualesquiera dos bases, siempre tienen la misma cardinalidad. Además, si $\mathbb{Q} \subset L \subset K$ entonces:

$$\dim_{\mathbb{Q}} K = \dim_{\mathbb{Q}} L \cdot \dim_L K \tag{41}$$

Una consecuencia del Teorema 4.10 es que la dimensión sobre $\mathbb{Q}$ de cualquier campo K que consiste en números construibles es una potencia de 2, dado que en cada paso, al aumentar una raíz cuadrada, se aumenta por un factor de 2. El campo de descomposición L de $f(x) = x^3 - 2$, que contiene la raíz $\sqrt[3]{2}$, tiene dimensión 3 sobre $\mathbb{Q}$, por lo tanto, una base es:

$$1, \quad \sqrt[3]{2}, \quad \sqrt[3]{2}^2 = \sqrt[3]{4}, \tag{42}$$

y, por lo tanto, cualquier base tiene cardinalidad 3. Si suponemos que $\sqrt[3]{2}$ es un número construible, entonces este número pertenece a un campo K que tiene dimensión 2^t sobre $\mathbb{Q}$ para algún t. Pero si K contiene $\sqrt[3]{2}$, entonces también la contiene su cuadrado y, por lo tanto, la base (42). Por consiguiente, tenemos ahora las siguientes inclusiones de campos $\mathbb{Q} \subset L \subset K$, pero $\dim_{\mathbb{Q}} L = 3$ y $\dim_{\mathbb{Q}} K = 2^t$, lo que contradice la igualdad (41).

Todo lo anterior muestra que $\sqrt[3]{2}$ no es construible y, por lo tanto, no es posible construir sólo con regla y compás el lado de un cubo que tenga el doble de volumen de un cubo dado.

Un argumento muy similar muestra que los otros dos problemas clásicos tampoco tienen solución con regla y compás: si queremos trisecar un ángulo α conocido debemos construir $\cos(\frac{\alpha}{3})$ o $\operatorname{sen}(\frac{\alpha}{3})$. Para ello, requerimos de una identidad trigonométrica que se deriva de la fórmula de Euler:

$$e^{\varphi\mathbf{i}} = \cos(\varphi) + \operatorname{sen}(\varphi)\mathbf{i} \tag{43}$$

donde e es el *número de Euler* —otro número trascendente— cuyo valor aproximado es 2.7182818284. Al elevar (43) al cubo, se obtiene:

$$\cos(3\varphi)+\operatorname{sen}(3\varphi)\mathbf{i} = (\cos^3(\varphi)-3\cos(\varphi)\operatorname{sen}^2(\varphi))+(3\cos^2(\varphi)\operatorname{sen}(\varphi)-\operatorname{sen}^3(\varphi))\mathbf{i}.$$

Si comparamos las coordenadas, tenemos, después de sustituir $\operatorname{sen}^2(\varphi) = 1 - \cos^2(\varphi)$ y $\cos^2(\varphi) = 1 - \operatorname{sen}^2(\varphi)$, respectivamente, las siguientes identidades para el triple de ángulo:

$$\cos(3\varphi) = 4\cos^3(\varphi) - 3\cos(\varphi) \tag{44}$$
$$\operatorname{sen}(3\varphi) = -4\operatorname{sen}^3(\varphi) + 3\operatorname{sen}(\varphi) \tag{45}$$

Podemos usar estas fórmulas para calcular $\cos(\frac{\alpha}{3})$, si sustituimos $\varphi = \frac{\alpha}{3}$ en (44):

$$\cos(\alpha) = 4\cos^3(\tfrac{\alpha}{3}) - 3\cos(\tfrac{\alpha}{3}) \tag{46}$$

Dado que el ángulo α ya está construido, podemos hallar $c = \cos(\alpha) = \cos(3\varphi)$ con regla y compás, y debemos resolver la ecuación (46) en la incógnita $x = \cos(\frac{\alpha}{3})$, es decir, tenemos que resolver la ecuación cúbica:

$$x^3 - \frac{3}{4}x - \frac{c}{4} = 0$$

En la sección 4.2 calculamos la solución y vimos que había que calcular raíces cúbicas. Si no se trata de ángulos muy particulares, $\cos(\frac{\alpha}{3})$ no será construible. Por ejemplo, $\cos(20°)$ no es construible.

Por ello, el problema de la trisección de ángulos no se puede resolver con regla y compás. Por la misma razón, tampoco se pueden construir polígonos con cualquier número de lados pues, por ejemplo, para construir un polígono con 9 lados, tendríamos que construir primero el $\cos(20°)$, que es imposible como acabamos de ver.

Así concluimos el recuento de una larga historia de búsqueda por la solución de algo que nunca fue posible. Los problemas clásicos no tienen solución mientras se mantenga la restricción de que las herramientas de construcción sean regla y compás, únicamente. Para comprender la imposibilidad, es necesario salir del ambiente de las construcciones: es necesario formular qué quiere decir el construir con regla y compás en un lenguaje que permite una vista distinta y que, en nuestro caso, fue el lenguaje algebraico con el uso de las herramientas de la teoría de campos. Ver que algo es imposible requiere de cierta distancia y reflexión y, en esta situación particular, una considerable cantidad de abstracción. Por ello, no debe sorprender que muchos aficionados han seguido y siguen intentando resolver alguno de los problemas y presentan construcciones ingeniosas para cuadrar el círculo, todas ellas falsas como ahora sabemos.

4.11 ¿QUÉ SE PUEDE DEMOSTRAR?

Q.E.D.

Figura 4.70 Las tres letras Q.E.D. son la abreviatura del latín quod erat demostrandum, *que se traduce como "lo que se quería demostrar". Con ellas se indica el fin de una demostración. Hoy pueden sustituirse también por un cuadrado "□", cuyo uso lo inició Paul Halmos. Las demostraciones juegan un papel central en las matemáticas. Con ellas se establece certidumbre y, a la vez, sirven para comunicar las matemáticas.*

4.11.1 El sistema axiomático

En esta sección queremos indagar un poco sobre las siguientes preguntas: ¿en qué consiste una demostración?, ¿cómo se demuestra algo?, ¿cómo se establece la veracidad de una afirmación? y, finalmente, ¿se puede demostrar todo lo que es verdadero?

Lo que hoy se llama "el Teorema de Pitágoras" se conoció más de 1 000 años antes de Pitágoras, que vivió en el siglo v a.C. A diferencia de las culturas anteriores, los griegos no se dieron por satisfechos con la validez en ejemplos aislados: percibieron la generalidad y la necesidad de una explicación, argumentación o *demostración* general.

Al principio, sólo fueron resultados sueltos como el Teorema de Tales, los teoremas de triángulos isósceles y el Teorema de Pitágoras. Pero, alrededor de 300 a.C., Euclides escribió un libro realmente extraordinario conocido como *Los elementos*, donde reunió en 13 tomos gran parte de lo que se sabía en aquel entonces de matemáticas. El logro de Euclides no fue la tarea de simplemente reunirlo y juntarlo, sino de idear una forma de organización sin precedente que hoy se llama *sistema axiomático*.

La idea de un sistema axiomático consiste en dar al principio los *axiomas*, es decir, aquellas afirmaciones que se aceptan como verdaderas sin cuestionar, y después derivar, me-

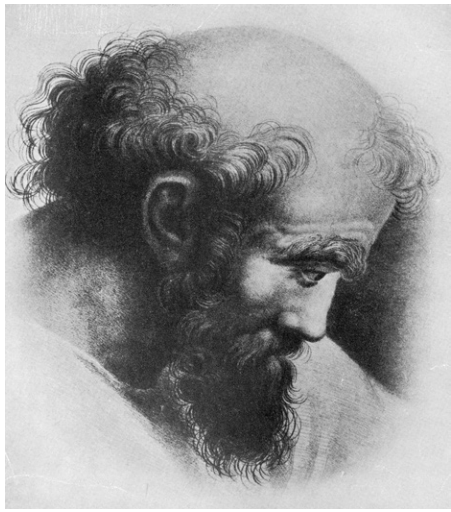


Figura 4.71 Pitágoras |
© Latin Stock México.

diante deducción, todas las demás afirmaciones. Los axiomas deben ser elegidos con cuidado. En particular, se debe evitar que alguno de los axiomas se pueda deducir a partir de los otros. Más importante aún, los axiomas deben ser *consistentes*, es decir, no debe ser posible deducirlos a partir de una afirmación y, a la vez, también de su negación.

Los axiomas formulados en *Los elementos* se dividen en dos grupos. Primero, vienen cinco "Postulados":

1. Por dos puntos diferentes sólo se puede trazar una línea recta.
2. Todo segmento rectilíneo se puede prolongar indefinidamente.
3. Con un centro y un radio dados sólo se puede trazar una circunferencia.
4. Todos los ángulos rectos son iguales.
5. Si una recta corta en forma transversal a otras dos formando ángulos internos del mismo lado de la transversal cuya suma es menor que 180 grados, entonces las dos rectas prolongadas lo suficiente se intersecarán de ese mismo lado.

Después, siguen cinco "Nociones comunes". La primera de ellas dice que "Cosas iguales a una tercera son iguales entre sí", mientras que la quinta dice: "El todo es mayor que la parte".

Se ve que los axiomas que se enuncian como "nociones comunes" mencionan afirmaciones que, posiblemente, nunca reflexionaremos tan detenidamente mientras que las afirmaciones que se enuncian como postulados tienen un carácter más específico, dado que tratan de la geometría en particular.

Estos axiomas son precedidos por las 23 "definiciones" que tratan de aclarar qué se entiende por "punto" o "recta". Por ejemplo, la primera definición dice que "un punto es lo que no tiene partes", y la segunda, "una línea es una longitud sin anchura". Desde un punto de visto moderno, estas definiciones sobran. Son los axiomas los que deben aclarar la relación que tienen los objetos entre sí. Es decir, los objetos como "punto" y "recta" se definen por sus propiedades, enunciadas en los mismos axiomas.

Veamos ahora la primera *proposición* que dice que se puede construir un triángulo equilátero sobre cualquier segmento dado. En seguida se da la demostración:

Sea AB el segmento dado. [P.1]
Así pues, se ha de dibujar sobre la recta AB un triángulo equilátero.
Dibujar el círculo BCD, con centro A y radio AB. [P.3]
Dibujar también el círculo ACE, con centro B y radio AB. [P.3]
A partir del punto C, que es intersección de los dos círculos, dibujar el segmento CA. [P. 1]
A partir del punto C, que es intersección de los dos círculos, dibujar el segmento CB. [P. 1]
Dado que el punto A es el centro del círculo BCD, AC es igual a AB. [D.15]
Dado que el punto B es, a la vez, el centro del círculo ACE, BC es igual a AB. [D. 15]
Como BC es igual a BA y AC es igual a BA, entonces BC es igual a AC. Ahora bien, las cosas iguales a una misma cosa son también iguales entre sí, por lo tanto, CA es también igual a CB y entonces CA, AB y BC son iguales entre sí. [N.C.1]
Por lo tanto, el triángulo ABC es equilátero y ha sido construido sobre el segmento dado AB.

En cada paso se indica la definición, el axioma o la noción común que se usa. Esto refleja la honradez al argumentar: se trata de no suponer nada, de explicar paso a paso la razón de cada argumento para transitar de las premisas —el segmento inicial dado— a la conclusión —el triángulo equilátero sobre este segmento inicial.

Con *Los elementos* se estableció la forma de trabajo de las matemáticas como disciplina, donde se deducen las afirmaciones en forma similar a un gran edificio que se construye poco a poco. Es por ello por lo que *Los elementos* fue el gran modelo a seguir. Por ejemplo, en 1537 se imprimió el libro *La Nova Scientia* de Niccolò Tartaglia sobre la balística, redactado en el mismo estilo. Esto nos muestra la gran aceptación que tuvo el método de Euclides. Sin embargo, Tartaglia determina como *Suposición II* —del segundo libro— el axioma de que un cuerpo que se mueve fuera de la perpendicular describe una trayectoria primero rectilínea y después curva —como arco de circunferencia—, aunque este axioma no corresponde a la realidad, según lo que Galilei demostraría —tiempo después—- acerca de las trayectorias parabólicas.

En consecuencia, queda claro que la elección de los axiomas es crucial. Pero no sólo eso: también a la hora de argumentar puede ser —no obstante lo mucho que se trató de evitar— que se usen algunas intuiciones que no se estipularon previamente. Por ejemplo, en la demostración del primer postulado de Euclides se asume la existencia de la intersección de dos circunferencias, resultando un punto de intersección C. ¿Cómo se sabe que tal punto existe? Actualmente, se sabe que aquí hay un problema en la argumentación y que se requiere del *teorema de curvas de Jordan* para tener la certeza de que tal C exista.

Esto muestra que los estándares de rigor en las demostraciones matemáticas han cambiado con el tiempo. No obstante los puntos débiles que se encontraron en *Los elementos,* este libro marcó un camino a seguir de honestidad en la argumentación. No fue sino hasta el principio del siglo xx cuando se dio otro avance sustancial en el rigor matemático.

4.11.2 La teoría de conjuntos como base para las matemáticas

Figura 4.72 Galileo Galilei (1564-1642) | © Latin Stock México.

Galilei se dio cuenta de que la quinta noción común, "El todo es mayor que la parte", no era siempre correcta cuando se trataba de una infinidad de cosas. En 1638, se imprimió su libro *Discorsi*, en el cual explica que para cada número natural hay un cuadrado perfecto —el

cuadrado de un entero como $4, 9$ o 100— y viceversa, para cada cuadrado perfecto obtenemos un número natural al sacar la raíz. Galilei concluye que hay, por consiguiente, tantos números naturales como cuadrados perfectos, no obstante que los cuadrados perfectos son sólo una parte de todos los números naturales.

Con ello, Galilei usó una noción para comparar el "tamaño" de dos conjuntos, lo que hoy se llama *correspondencia uno a uno* o *biyección* y que se introduce, por primera vez, en 1903. Galilei no persigue más esta idea y lo que encontró ha quedado simplemente como una de las paradojas del infinito. Fue Georg Cantor quien enfrentó estas paradojas de frente, al estudiar los *conjuntos* formalmente y, para ello, usó los *números trascendentes* (véase también sección 4.9). Los conjuntos no tuvieron una fácil aceptación al principio; en particular, Leopold Kronecker se oponía fervientemente.

Al principio, un conjunto era simplemente una colección de cosas. Lo único que se pedía era que, de cada "objeto", siempre se podía decidir si pertenecía o no a dicho conjunto. Relativamente pronto se dieron cuenta de que la noción, así de general, producía paradojas a su vez, como la que hoy se llama *paradoja de Russell*. Para este caso se considera E como el conjunto de todos los conjuntos que no se contienen como elemento y que, al usar el símbolo de pertenencia "$\in$", se podría escribir:

$$E = \{X \mid X \notin X\}.$$

La paradoja surge entonces al preguntar si $E \in E$ o $E \notin E$. En el primer caso, E es un conjunto X con la propiedad $X \notin X$, una contradicción con $E \in E$. En el segundo caso, E satisface la condición $X \notin X$ y, por lo tanto, $E \in E$, lo cual es otra vez una contradicción.

Fue necesario un trabajo adicional de Ernst Zermelo y Abraham Fraenkel, matemáticos alemanes en las primeras dos décadas del siglo xx, para salvar la idea de los conjuntos y redefinirlos en forma completamente nueva. Además, Hilbert —quien era "formalista" y apreciaba la precisión y el rigor que surgió de la teoría de conjuntos a la matemática— acuñó la frase: "Nadie debe poder expulsarnos del paraíso que nos creó Cantor".

Durante la primera mitad del siglo xx tuvo lugar una gran formalización en la matemática que, además, se extendió prácticamente por toda la disciplina. Con este espíritu debe verse también la obra *Principia Mathematica*, de Alfred North Whitehead y Bertrand Russell. En este libro, los autores querían establecer, de una vez por todas, las bases axiomáticas para todas las matemáticas. Los autores publicaron tres tomos cubriendo "sólo" teoría de conjuntos, números cardinales, números ordinales y números reales; para ello, necesitaron aproximadamente 2 000 páginas. El cuarto tomo sobre geometría ya no lo escribieron, por agotamiento.

Hoy día, todas las matemáticas se basan —de manera formal— en la teoría de conjuntos, el sistema de axiomas *ZFC*, que quiere decir el sistema de axiomas de Zermelo-Fraenkel, incluyendo el axioma de elección.

4.11.3 El programa de Hilbert

En 1920, David Hilbert propuso un ambicioso programa para subsanar los problemas de los cimientos de las matemáticas, que hoy se conoce, simplemente, como el *programa de Hilbert*.

Los puntos principales del programa de Hilbert eran:

1. Una *formalización* de todas las matemáticas, es decir, formular todas las expresiones matemáticas en un lenguaje formal y manipularlas de acuerdo con reglas fijas y bien definidas.
2. Demostrar que el sistema es *consistente,* es decir, que no es posible encontrar contradicciones dentro de él.
3. Demostrar que el sistema es *completo,* lo cual quiere decir que cada afirmación matemática que sea correcta puede ser demostrada dentro del mismo sistema.
4. Demostrar que existe un *algoritmo* para decidir la veracidad o falsedad de cualquier afirmación matemática.

Como veremos en esta sección, el programa de Hilbert fracasó: era demasiado ambicioso. No es que fracasara porque no hubiera suficientes esfuerzos vertidos en él, sino porque había obstáculos definitivos para lograrlo. Lo que Hilbert quería no se puede hacer, pero para comprenderlo más a fondo, es necesario entender mejor el primer punto: la formalización en sí.

Las matemáticas se escriben con símbolos que, en ocasiones, son muy raros. Por ejemplo, se suele usar el símbolo $\vee$ para la conjunción lógica "o". La cadena de símbolos $p \vee q$ expresa entonces "p o q", donde p y q son dos afirmaciones cualesquiera. En forma similar, se usa $\wedge$ para expresar la conjunción "y", el símbolo $\neg$ para expresar "no". Por ejemplo:

$$(p \wedge q) \vee \neg q \tag{47}$$

quiere decir "se tiene p y q o se tiene no q". Esto tiene el mismo significado que "si p entonces q", que se suele escribir como $p \Rightarrow q$. No se puede saber si $p \Rightarrow q$ es verdadero o falso, sin conocer la validez de p y q. Si p y q son verdaderos, entonces $p \wedge q$ también es verdadero, mientras $\neg q$ es falso. En consecuencia, 47 es verdadero.

La expresión $p \wedge q \Rightarrow p$ siempre es verdadera, sin importar la validez de p y q. Por lo tanto, si sabemos que $p \wedge q$ es verdadero, podemos sustituir esta expresión por p y obtenemos una nueva cadena que es verdadera. Similarmente, podemos sustituir $p \wedge q$ por $q \wedge p$, es decir, $p \wedge q \Rightarrow q \wedge p$ es siempre verdadera.

Un *sistema formal* consiste en varias partes:

- Un conjunto finito de símbolos, llamado *alfabeto*.
- Un conjunto —usualmente infinito— de *fórmulas bien definidas*. Cada una de estas fórmulas es una cadena de símbolos del alfabeto. Se espera que exista un procedimiento para decidir si una cadena de símbolos está bien definida o no.
- Un conjunto finito de *axiomas*. Cada axioma debe ser una fórmula bien definida.
- Unas *reglas de inferencia* que explican las maneras permitidas para obtener fórmulas bien definidas a partir de otras.

Un *sistema lógico* asigna valores de "verdadero" o "falso" a cada fórmula bien definida.

Hagamos un ejemplo sencillo. El alfabeto es $A = \{1, +, =\}$. Una cadena de símbolos está bien formada si contiene exactamente un símbolo $=$ y, además, cada símbolo $+$ y cada símbolo $=$ está encerrado entre dos símbolos 1. Por ejemplo:

$$111 + 11 + 1 = 111111 \tag{48}$$

es una fórmula bien definida. Hay un único axioma que es $1 = 1$. Las reglas de inferencia son las siguientes, las variables p, q, r que aparecen son cadenas en los símbolos $1, +$ tal que cada $+$ está encerrado entre dos 1:

1. Si $pq = r$, entonces $p + q = r$.
2. Si $p = q$, entonces $p1 = q1$.
3. Si $p = q$, entonces $q = p$.

Si queremos podemos interpretar este sistema formal como el de la adición de números enteros positivos representados en el sistema unitario, es decir, simplemente indicando el número por la cantidad de símbolos 1. Habría que interpretar la ecuación (48) como $3 + 2 + 1 = 6$. Con estas reglas de inferencia es posible deducir cualquier expresión matemáticamente correcta entre dos sumas o números enteros positivos. Las cadenas bien formadas que no se pueden deducir son *falsas* si las interpretamos en este sentido y las que sí se pueden deducir son *verdaderas*.

Del axioma $1 = 1$ derivamos $111111 = 111111$, al usar varias veces la segunda regla de inferencia. Con la primera regla de inferencia, $p = 111$ y $q = 111$, obtenemos $111 + 111 = 111111$ y, si la aplicamos nuevamente con $p = 111 + 11$ y $q = 1$, obtenemos 48. Formalmente, una *demostración* es una sucesión de fórmulas bien formadas, donde se indica en cada paso si es axioma o cómo se obtiene de una fórmula bien definida anterior. Nuestro ejemplo es sencillo, la demostración es:

$$1 = 1 \xrightarrow{I.2} 11 = 11 \xrightarrow{I.2} 111 = 111 \xrightarrow{I.2} 1111 = 1111 \xrightarrow{I.2} 11111 = 11111 \xrightarrow{I.2} 111111 = 111111 \xrightarrow{I.1} 111 + 111 = 111111 \xrightarrow{I.1} 111 + 11 + 1 = 111111$$

4.11.4 El teorema de Gödel

Las nociones de sistemas formales nos dan herramientas para hablar de lo que se puede demostrar dentro de un sistema de axiomas dado. Un sistema así se llama *completo* si, para cada fórmula bien formada f, se puede demostrar f o $\neg f$ usando sólo las reglas de inferencia. Es posible *a priori* o al menos pensable que existen sistemas donde hay fórmulas bien formadas f de las cuales ni f ni tampoco $\neg f$ se pueden derivar usando nada más que las reglas de inferencia. En este caso f es *independiente* de los axiomas dados y es posible aumentar el conjunto de axiomas por f o por $\neg f$.

Hay varios casos conocidos de axiomas que son independientes de otros como, por ejemplo, el quinto postulado de Euclides mencionado anteriormente, que es independiente de los otros cuatro postulados. Si se sustituye por su negación, se obtienen las *geometrías no-euclidianas*. También la *hipótesis del continuo* es independiente de los axiomas ZFC; esta hipótesis afirma que cada subconjunto de $\mathbb{R}$, el conjunto de los números reales, tiene la misma cardinalidad que $\mathbb{R}$ o que $\mathbb{Q}$, pero no hay nada intermedio.

Lo que pedía Hilbert en su programa era dar un conjunto de axiomas que establecieran las bases para todas las matemáticas, como se hacía en el libro *Principia Mathematica* para la aritmética y, luego, demostrar que este sistema de axiomas era completo. Pero eso no es posible, según lo comprobó Kurt Gödel, un matemático que nació en 1906 en Austria y murió en 1978 en Estados Unidos.

El resultado que demostró se llama hoy el *teorema de Gödel* y afirma que un sistema axiomático consistente y "suficientemente poderoso para describir los números enteros" es necesariamente incompleto. Consecuentemente, en cualquier sistema formal que incluya la aritmética como se hace en *Principia Mathematica*, siempre habrá fórmulas bien definidas que no se puedan derivar ni tampoco su contrario. En otras palabras, siempre habrá afirmaciones en las que no podremos decidir si son verdaderas o falsas.

La idea central de Gödel consiste en codificar cualquier fórmula bien formada como un número, lo cual se puede hacer de varias maneras. Lo importante es que se puede recuperar sin problema la fórmula a partir del número. La siguiente tabla muestra posibles códigos para los símbolos de un alfabeto sencillo.

Símbolo	Código	Significado
0	1	cero
S	2	función sucesora, ($SS0$ significa el número 2)
$=$	3	relación de igualdad
$<$	4	relación de orden
$+$	5	operación de adición
$\times$	6	operación de multiplicación
$($	7	paréntesis abre
$)$	8	paréntesis cierra
$\vee$	9	operador lógico "o"
$\wedge$	10	operador lógico "y"
$\neg$	11	operador lógico "no"
x	12	una variable
$'$	13	apóstrofe (se usa para crear más variables: $x', x'', \ldots$)
$\forall$	14	cuantificador "para cada"
$\exists$	15	cuantificador "existe"

Figura 4.73 Posibles códigos para los símbolos de un alfabeto sencillo.

La codificación de una fórmula con k símbolos se hace con los primeros k números primos $2, 3, 5 \ldots, p_k$ —aquí usamos p_i para denotar al i-ésimo primo. El primo p_i se eleva a la potencia c_i, si c_i es el código del i-ésimo símbolo. Por ejemplo, la fórmula:

$$\forall x \forall x' (x + x' = x' + x),$$

tiene 16 símbolos y se codifica en el siguiente número:

$$2^{14} \cdot 3^{12} \cdot 5^{14} \cdot 7^{12} \cdot 11^{13} \cdot 13^{7} \cdot 17^{12} \cdot 19^{5} \cdot 23^{12} \cdot 29^{13} \cdot 31^{3} \cdot 37^{12} \cdot 41^{13} \cdot 43^{5} \cdot 47^{12} \cdot 53^{8},$$

Para recuperar la fórmula a partir del número, hay que hacer la descomposición en factores primos y contar cuántas veces aparece cada uno de ellos. Esto puede ser muy tedioso de hacer en ejemplos concretos; el número de arriba es:

99 729 218 036 602 955 429 934 818 571 840 953 217 751 753 892 397 683 751 686 555 355 842 609 377 305 551 005 904 156 314 767 308 829 381 973 185 267 401 685 826 988 220 470 759 865 745 426 481 243 147 842 988 402 164 767 156 563 454 278 100 000 000 000 000.

¡Un número con 194 cifras! Pero aquí no nos interesa cuán práctico es sino, simplemente, que es posible.

También las demostraciones dentro del sistema formal se pueden expresar como número. Recordemos que una demostración es una cadena de fórmulas bien formadas, como:

$$F_1, F_2, \ldots, F_t \tag{49}$$

donde las primeras fórmulas son axiomas y cada fórmula siguiente es una consecuencia de las anteriores, al aplicar alguna de las reglas de deducción. La última, es decir F_t, es la fórmula que se demostró. Usando los códigos de Gödel para cada una de las fórmulas, obtenemos una cadena de números:

$$m_1, m_2, \ldots, m_t.$$

El código asociado a la demostración (49) se forma al elevar los primeros primos a las potencias $m_1, \ldots, m_t$:

$$3^{m_1} \cdot 5^{m_2} \cdot \ldots \cdot p_{t+1}^{m_t},$$

es decir, se forma muy parecido al código de las fórmulas bien formadas con el único cambio de que se empieza con el primo $p_2 = 3$ y no con $p_1 = 2$. Este pequeño cambio tiene la ventaja de poder distinguir, inequívocamente, entre los códigos de fórmulas y los códigos de demostraciones: los primeros siempre serán pares y los segundos impares.

Por ello, ahora podemos considerar la siguiente función:

$$\mathrm{DEM} : \mathbb{N} \times \mathbb{N} \longrightarrow \{0,1\}, \mathrm{DEM}(d,m) = \begin{cases} 1, & \text{si } d \text{ es el código de una demostración} \\ & \text{para la afirmación } F \text{ cuyo código es } m, \\ 0, & \text{si no.} \end{cases}$$

Lo importante de esta función es que es posible evaluarla en la práctica: si el número m no corresponde al código de fórmulas bien formadas, el valor de $\mathrm{DEM}(d,m)$ es cero, y lo mismo vale si d no corresponde a una sucesión de fórmulas bien formadas. En caso contrario, d corresponde a $F_1, \ldots, F_t$ y m a una fórmula F. Si $F \neq F_t$, entonces nuevamente $\mathrm{DEM}(d,m) = 0$. Y también es posible verificar si cada F_i es axioma o consecuencia de $F_1, \ldots, F_{i-1}$.

Lo sorprendente es que encontramos una función de dos números naturales que nos indica si el primer número corresponde a una demostración de lo que corresponde al segundo número. Además, esta función se puede expresar en nuestro sistema formal, es decir, la expresión:

$$\forall x(\mathrm{DEM}(x,m) = 0) \tag{50}$$

se puede escribir explícitamente con los símbolos dados en la tabla de arriba. Lo escribimos para mejor legibilidad aunque muy comprimido: tan sólo el número m se tendría que escribir como:

$$\underbrace{SS\cdots S}_{m \text{ veces}}0 \tag{51}$$

y DEM se expandirá a una expresión muy grande. Con ello, hemos expresado dentro de nuestro sistema el hecho de que una fórmula bien formada F con código m

no es demostrable, pues no existe código de Gödel que represente una sucesión de fórmulas bien definidas.

Sin embargo, en (50) entró en juego el número específico m. Por ello, se considera mejor la fórmula bien definida:

$$\forall x(\mathrm{DEM}(x, x') = 0) \tag{52}$$

que tiene una variable libre, a saber x'. Esta variable la podemos sustituir por diferentes valores para obtener diferentes fórmulas bien definidas.

Es crucial que la sustitución de una variable libre, como x', en una fórmula pueda hacerse de manera mecánica con los códigos de Gödel. Es decir, la aplicación que asigna al código de (52) el código de (50) es, a la vez, algo que se puede codificar dentro del sistema formal. Será un procedimiento complicado: habrá que ubicar todos los primos p_i dentro del código de (52) que ocurren con una multiplicidad de 12 —el código de x—, tal que la multiplicidad de p_{i+1} sea 13 —el código del apóstrofe— y la de p_{i+2} no sea 13 y, luego, recorrer los primos para poder insertar en lugar de los dos símbolos x', los $m + 1$ símbolos $SS\cdots S0$.

Representaremos con $\mathrm{SUB}(k, 2^{12} \cdot 3^{13}, \ell)$ el código de Gödel de la fórmula que se obtiene al sustituir la variable libre x' —que tiene el código $2^{12} \cdot 3^{13}$— por el valor ℓ en la fórmula cuyo código es k. Repetimos: $\mathrm{SUB}(k, 2^{12} \cdot 3^{13}, \ell)$ es un número y se puede expresar dentro del sistema formal. Lo que aquí importa es que la formulación de SUB dentro del sistema formal no depende de los valores exactos de k, ℓ. Así será posible usar variables en vez de k y ℓ.

Ahora, juntemos todo y consideremos la siguiente fórmula con variable libre x':

$$\forall x(\mathrm{DEM}(x, \mathrm{SUB}(x', 2^{12} \cdot 3^{13}, x')) = 0), \tag{53}$$

como fórmula bien formada, lo cual significa que $\mathrm{SUB}(x', 2^{12} \cdot 3^{13}, x')$ no tiene demostración dentro del sistema formal. Como fórmula bien formada, 53 tiene un código de Gödel n, un número natural —que, sin duda, será gigantesco—. Sustituimos ahora la variable x' por el código n:

$$\forall x(\mathrm{DEM}(x, \mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)) = 0) \tag{54}$$

Lo que obtenemos es una fórmula bien formada que ya no tiene variable libre. ¿Qué significa? Ya lo sabemos: que $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$ no tiene demostración dentro del sistema formal. Pero $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$ es la fórmula que se obtiene al sustituir en (53) —que es la fórmula cuyo código es n— la variable x' por n. Es decir, $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$ no es otra cosa que (54), que dice que $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$ no es demostrable.

El círculo se cerró: (54) es una fórmula bien formada que afirma que ella misma no es demostrable. Falta ver que ni (54) ni su negación se pueden deducir a partir de los axiomas. En efecto, si existiera una demostración $F_1, \ldots, F_t$ con código d de 4.54, entonces:

$$\mathrm{DEM}(d, \mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)) = 1$$

dado que $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$ es (54). Esto nos da una demostración de la fórmula:

$$\exists x(\mathrm{DEM}(x, \mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)) = 1). \qquad (55)$$

Pero como (55) es justo la negación de (54), demostramos una afirmación y su contradicción, cosa que es imposible si el sistema formal es consistente, lo que supusimos desde el principio. Así, llegamos a la conclusión de que no puede haber demostración de la fórmula (54) dentro del sistema y de que tampoco puede haber demostración de la negación de $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$, pues si hubiera una demostración con código d, entonces:

$$\mathrm{DEM}(d, \neg\, \mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)) = 1$$

y, por la consistencia, encontraríamos que no puede haber demostración de $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$, es decir, dedujimos dentro del sistema la fórmula (54), que no es otra cosa que $\mathrm{SUB}(n, 2^{12} \cdot 3^{13}, n)$ mismo. Nuevamente, tenemos una contradicción con la consistencia.

Así es como Gödel llega a la conclusión de que cualquier sistema que sea lo suficientemente poderoso para poder operar con números naturales, siempre contendrá afirmaciones que no son demostrables ni tampoco su negación.

Con ello, concluimos este viaje por la lógica. En resumen, cualquier formalización que incluye la aritmética es necesariamente incompleta. Por lo tanto, cualquier formalización de *todas* las matemáticas, necesariamente, es incompleta. En este sentido, los puntos 2 y 3 del programa de Hilbert no son compatibles: si el sistema es consistente, es necesariamente incompleto.

Lo anterior puede parecer desalentador, ya que con ello se derrumbó el ambicioso programa de Hilbert. Por otro lado, se puede ver como una invitación para hacer matemáticas, ya que éstas no se pueden realizar completamente dentro de sistemas formales. La demostración misma del Teorema de Gödel es una joya de razonamiento que no se da en un sistema formal, sino que se razona sobre ellos. Visto de esta manera, podríamos interpretar este resultado como una confirmación de la importancia para la creatividad del intelecto humano dentro de las matemáticas.

4.12 Y... ¿SI TODO QUEDARA DESCUBIERTO?

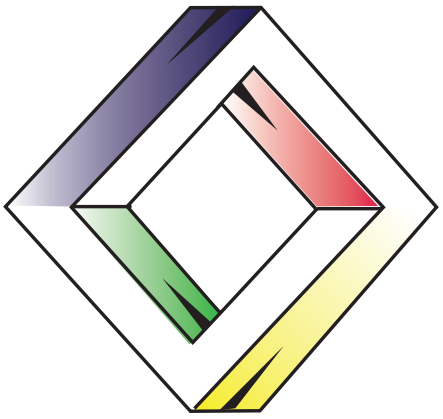


Figura 4.74 La Sociedad Matemática Mexicana eligió como logotipo una cinta torcida en forma de cuadrado, una figura que aparenta un cuerpo tridimensional pero que es imposibe de realizar. ¿Es una metáfora para lo que representan las matemáticas? ¿Será que aun lo que parece imposible se puede lograr en las matemáticas? La Sociedad Matemática Mexicana se compone actualmente de varios cientos de matemáticos y matemáticas, investigadores y profesores.

Es común escuchar expresiones de sorpresa y hasta de incredulidad respecto al quehacer de un matemático en nuestros días: "pero ¿qué van a investigar?, ¿la tabla del 18?, ¿los números?" Es cierto, los números han estado ahí desde siempre y es la asociación más frecuente a la palabra "matemáticas". Pero, aunque hay muchas preguntas respecto a los números que nadie ha podido contestar aún, las matemáticas no se reducen sólo a ellos.

En la historia ha habido tres etapas en las cuales las matemáticas prosperaron particularmente bien: la primera en la antigua Grecia, la segunda en el siglo XVIII y la tercera es ahora. Nunca hubo una comunidad tan grande de científicos que se dedicaran a las matemáticas como ahora, ni tampoco existía la actual producción frenética de resultados.

La revista *Mathematical Reviews* se imprime cada mes y cada uno de sus tomos tiene alrededor de mil páginas. En ellas se encuentran las reseñas, es decir, los resúmenes elaborados por matemáticos de los artículos que se publicaron hace poco. Cada mes se reseñan alrededor de cinco mil artículos y libros, lo que equivale, aproximadamente, a 50 mil páginas de matemáticas nuevas cada mes. Este volumen de conocimiento es tan grande que es absolutamente imposible leerlo todo y, por ello, las matemáticas se han dividido en áreas y en ramas.

La ramificación de las matemáticas no es estática, sino que se encuentra en permanente desarrollo. Tan sólo el catálogo actual de los grandes temas requiere de diez páginas para imprimirse. El cúmulo de información generada por estos científicos es tan grande que no resulta fácil, para los propios matemáticos, mantenerse actualizados.

A lo anterior, debe añadirse que los artículos de investigación publicados por los matemáticos —como reporte de sus investigaciones— no son de fácil lectura. Leer uno de estos artículos puede costar días y, a veces, hasta meses de trabajo para comprender cabalmente lo que se reportó. La frontera del conocimiento avanza lentamente y, al mismo tiempo, se esparce y se ramifica cada vez más.

Por ello, para mantenerse informados, los matemáticos suelen reunirse con frecuencia en conferencias de especialistas que se dedican a alguna de las ramas específicas de las matemáticas. Por otro lado, también se asocian en sociedades como la Sociedad Matemática Mexicana, para mantenerse en contacto permanente.

¿Y no tendrá esta actividad frenética, de repente, su culminación cuando todo esté descubierto? Hay que detenerse un momento para reflexionar acerca de qué significaría que, llegado un momento, ya no hubiera nada que estudiar. Esto sólo podría suceder si todos los problemas del mundo quedaran resueltos, si toda la naturaleza fuera entendida y se contestaran todas las preguntas matemáticas. Pero, como caja de Pandora, cada pregunta que se contesta genera una, tres o más nuevas preguntas. Un fenómeno exponencial que, a fin de cuentas, está muy lejano de poner un fin a las actividades matemáticas por falta de preguntas. Quizá podría llegar por falta de interés. Pero las sociedades exitosas han encontrado que sus logros dependen cada vez más del desarrollo de las ciencias, y las economías emergentes lo son por estar promoviéndolas. El interés social en las matemáticas también crece. ¿Y el individual? Que decayera sería como pensar que dejara de haber músicos, poetas o artistas pues, al igual que los matemáticos, expresan fibras muy profundas del espíritu humano. Queda claro que no se le ve el fin.

Se podría pensar que las preguntas estudiadas por los matemáticos son cada vez más sofisticadas, cada vez más retiradas de lo que se pueda comprender y, en cierta medida, eso es cierto si no fuera por algunos desarrollos que, a finales del siglo XX, resolvieron unas preguntas muy viejas. Entre ellas, la conjetura de Fermat que sólo concierne a los números y que ya hemos mencionado: la ecuación $a^n + b^n = c^n$ sólo tiene soluciones de números enteros positivos si $n \leq 2$. Esta conjetura debida a Fermat, que vivió en el siglo XVII no se demostró sino hasta 1995 por Andrew Wiles.

Actualmente hay siete problemas que se llaman "los problemas del milenio". Para cada uno de ellos, el Instituto Clay ofrece un millón de dólares a aquel que logre resolverlo. Esto quiere decir que la solución se tiene que publicar en una revista internacional y tiene que sostenerse sin problema durante los siguientes dos años, después de su publicación. Durante este tiempo, se supone, los especialistas revisarán la prueba para detectar posibles errores o huecos en la argumentación. Por ejemplo, la primera prueba de Andrew Wiles presentó un hueco que requirió dos años más de trabajo.

En conclusión, las matemáticas se encuentran en una de las eras más prósperas y donde no hay señales de que todo lo que se puede resolver, pronto esté resuelto. Las matemáticas son un campo de investigación activo que da servicio a todas las ciencias al proveer un lenguaje abstracto pero tremendamente útil, al mismo tiempo que se desarrollan por intereses generados por las propias matemáticas y sirven, de manera casi inexplicable, como lenguaje para describir la naturaleza. Por todo lo anterior, vale la pena acercarse a ellas, tratar de adquirir conocimientos profundos en ellas al familiarizarse con sus métodos y su sensibilidad para enfrentar diversos problemas.

BIBLIOGRAFÍA

EL PORQUÉ DE LAS MATEMÁTICAS

Aczel, Amir D., *El último teorema de Fermat. El secreto de un antiguo problema matemático*, México, Fondo de Cultura Económica, 2005.

Amster, Pablo, *Fragmentos de un discurso matemático*, Buenos Aires, Fondo de Cultura Económica, 2008.

———, *Las matemáticas como una de las bellas artes*, Buenos Aires, Siglo XXI, 2005.

Beckmann, Peter, *Historia de* π, México, Libraria/cnca, 2006.

Bell, Eric Temple, *Historia de las matemáticas*, México, Fondo de Cultura Económica, 2010.

Collette, Jean-Paul, *Historia de las matemáticas*, vols. I y II, México, Siglo XXI, 1985.

Courant, Richard y Herbert Robbins, *¿Qué son las matemáticas? Conceptos y métodos fundamentales*, México, Fondo de Cultura Económica, 2006.

Galilei, Galileo, *Discorsi e Dimostrazioni Matematiche, intorno a due nuove scienze attenenti alla meccanica & i movimenti locali* (Diálogos sobre dos nuevas ciencias).

———, *La gaceta sideral y Johannes Kepler: Conversación con el mensajero sideral*, Madrid, Alianza, 2007.

Gardner, Martin, *Mathematical Puzzles of Sam Loyd*, seleccionados y editados por Martin Gardner.

Kasner, Edward y James Newmann, *Matemáticas e imaginación*, prólogo (y una reseña) de Jorge Luis Borges, México, Libraria/cnca, 2007.

Kline, Morris, *Matemáticas para los estudiantes de humanidades*, México, Fondo de Cultura Económica, 2a. ed., 2009.

Koestler, Arthur, *Los sonámbulos. Origen y desarrollo de la cosmogonía*, México, qed/Consejo Nacional para la Cultura y las Artes, 2007.

Maor, Eli, *e: historia de un número*, México, Libraria/cnca, 2006.

Paenza, Adrián, *Matemática... ¿estás ahí?: sobre números, personajes, juegos, lógica y reflexiones sobre la matemática: episodio 2*, Buenos Aires, Siglo XXI, 2001.

Peña, José Antonio de la, *Álgebra en todas partes*, México, Fondo de Cultura Económica, 2010.

Prieto, Carlos, *Aventuras de un duende en el mundo de las matemáticas*, México, Fondo de Cultura Económica, 2000.

Singh, Simon, *El enigma de Fermat*, Barcelona, Planeta, 1998.

MATEMÁTICAS DE LA ACTIVIDAD HUMANA

Abreu, J. L. y H. Fetter, *Sistemas numéricos*, vol. II, México, Limusa, 1980.

Beckmann, Peter, *A History of pi*, Nueva York, St. Martin's Press, 1971.

Berlanga, Ricardo, Carlos Bosch, Juan José Rivaud, *Las matemáticas, perejil de todas las salsas*, México, Fondo de Cultura Económica, 4a. ed., 2009.

Bosch, Carlos, *El billar no es de vagos. Ciencia, juego y diversión*, México, Fondo de Cultura Económica, 2009.

Courant, Richard y Herbert Robbins, *¿Qué son las matemáticas? Conceptos y métodos fundamentales*, México, Fondo de Cultura Económica, 2006.

Feller, William, *Introducción a la teoría de probabilidades y sus aplicaciones*, México, Limusa/Wiley, 1973.

González Urbaneja, Pedro Miguel, *Arquímedes y los orígenes del cálculo integral*, Madrid, Nivola.

Infante Gil, Said, Guillermo P. Zárate de Lara, *Métodos estadísticos, Un enfoque interdisciplinario*, México, Trillas, 1986.

Kline, Morris, *Matemáticas para los estudiantes de humanidades*, México, Fondo de Cultura Económica, 2a. ed., 2009.

Newman, James R., *Sigma. El mundo de las matemáticas*, Barcelona, Grijalbo, 1997.

Peña, José Antonio de la, *Álgebra en todas partes*, México, Fondo de Cultura Económica, 2010,

Vera, Francisco, *Científicos griegos*, Madrid, Aguilar, 1970.

LAS MATEMÁTICAS EN LA NATURALEZA

Abreu, J. L., J. A. Canavati, J. Ize y A. Minzoni, *Cálculo diferencial e integral,* vol. I: *Introducción a los conceptos del cálculo*, México, Limusa, 1980.

Atalay, Bülen, *Las matemáticas y la Mona Lisa*, España, Almuzara, 2008.

Berlanga, Ricardo, Carlos Bosch, Juan José Rivaud, *Las matemáticas, perejil de todas las salsas*, México, Fondo de Cultura Económica, 4a. ed., 2009.

Conway, John H., Heidi Burgiel y Chaim Goodman-Strauss, *The Symmetries of Things*, Wellesley, Massachusetts, A. K. Peters, 2008.

du Satoy, Marcus, *Symmetry. A Journey into the Patterns of Nature*, Nueva York, Harper Perennial, 2008.

Galilei, Galileo, *Discorsi e Dimostrazioni Matematiche, intorno a due nuove scienze attenenti alla meccanica & i movimenti locali* (Diálogos sobre dos nuevas ciencias).

García Álvarez, Miguel Ángel, *Introducción a la teoría de la probabilidad I. Primer curso*, México, Fondo de Cultura Económica, 2008.

González Urbaneja, Pedro Miguel, *Fermat y los orígenes del cálculo diferencial*, Madrid, Nivola.

Goodstein, David L., Judith R. Goodstein, *Feynman's Lost Lecture: The Motion of Planets around the Sun*, Londres, Vintage, 1997

Hawking, Stephen, *La gran ilusión. Las grandes obras de Albert Einstein*, Barcelona, Crítica, 2008.

Koestler, Arthur, *Los sonámbulos. Origen y desarrollo de la cosmogonía*, México, qed y Consejo Nacional para la Cultura y las Artes, 2007.

Maor, Eli, *e: historia de un número*, México, Libraria/cnca, 2006.

Shey, H. M., *Div, Grad, Curl and all That: An Informal Text on Vector Calculus*, W. W. Norton, 3a. ed., 1996.

Swokowski Earl, *Cálculo con geometría analítica,* México, Iberoamérica, 1989.

LAS MATEMÁTICAS DE LAS MATEMÁTICAS

Aschbacher, Michael, "The Status of the Classification of the Finite Simple Groups", *Notices of the American Mathematical Society,* vol. 51, núm. 7, 2004, pp. 736-740.

Beckmann, Peter, *Historia de i*, México, Libraria/cnca, 2006.

Bombal, Fernando, Nicolas Bourbaki, en *Historia de la matemática en el siglo xx,* Real Academia de la Ciencia de Madrid, 1988, pp. 313-323. En línea: <http://ochoa.mat.ucm.es/_bombal/Personal/Historia/BOURBAKI.pdf>.

Bracho, Javier, *¿En qué espacio vivimos?*, México, Fondo de Cultura Económica, 3a. ed., 2010.

———, *Introducción analítica a las geometrías*, México, Fondo de Cultura Económica, 2009.

Hofstadter, Douglas, *Gödel, Escher, Bach: un eterno y grácil bucle*, Tusquets, 2007.

Illanes Mejía, Alejandro, *La caprichosa forma de Globión*, México, Fondo de Cultura Económica, 1999.

Infeld, Leopold, *El elegido de los dioses. La historia de Evariste Galois*, México, Siglo XXI, 1974.

Jones, Arthur, Sidney A. Morris, Kenneth R. Pearson, *Abstract Algebra and Famous Impossibilities*, Nueva York, Berlín, Heidelberg, Springer, 1991.

Kasner, Edward y James Newmann, *Matemáticas e imaginación*, prólogo (y una reseña) de Jorge Luis Borges, México, Libraria/cnca, 2007.

Klein, Felix, *Elementary Mathematics from an Advanced Standpoint: Geometry*, Nueva York, Dover.

———, *Matemática elemental desde un punto de vista superior. Aritmética - Álgebra - Análisis*, Madrid, Nivola.

Lavine, Shaughan, *Comprendiendo el infinito*, México, Fondo de Cultura Económica, 2005.

Montejano Peimbert, Luis, *La cara oculta de las esferas*, México, Fondo de Cultura Económica, 3a. ed., 2003.

Nagel, Ernest, James R. Newman, *El teorema de Gödel,* Madrid, Tecnos, 1999.

Nahin, Paul J., *Esto no es real. La historia de i*, México, Libraria/cnca, 2006.

Ongay, Fausto, *Mathema: el arte del conocimiento*, México, Fondo de Cultura Económica, 2000.

Pesic, Peter, *Abel's Proof: An Essay on the Sources and Meaning of Mathematical Unsolvability*, mit Press, 2003, pp. viii y 213.

Solomon, Ron, "On Finite Simple Groups and Their Classification", *Notices of the American Mathematical Society*, vol. 42, núm. 2, 1995, pp. 231-239.

GENERAL

Atalay, Bülen, *Las matemáticas y la Mona Lisa*, España, Almuzara, 2008.

Collette, Jean-Paul, *Historia de las matemáticas*, vols. i y ii, México, Siglo XXI, 1985.

Courant, Richard y Herbert Robbins, *¿Qué son las matemáticas? Conceptos y métodos fundamentales*, México, Fondo de Cultura Económica, 2006.

Hofstadter, Douglas, *Gödel, Escher, Bach: un eterno y grácil bucle*, Barcelona, Tusquets, 2007.

Klein, Felix, *Elementary Mathematics from an Advanced Standpoint: Geometry*, Nueva York, Dover.

———, *Matemática elemental desde un punto de vista superior. Aritmética -Álgebra -Análisis*, Madrid, Nivola.

Kline, Morris, *Mathematical Thought from Ancient to Modern Times*, Nueva York, Oxford University Press, 1972.

Newman, James R., *Sigma. El mundo de las matemáticas*, Barcelona, Grijalbo, 1997.

Polya, George, *¿Cómo plantear y resolver problemas?*, México, Trillas, 1965.

———, *How to Solve it : A New Aspect of Mathematical Method*, Princeton, NJ, Princeton University Press, 1957.

———, *Matemáticas y razonamiento plausible*, traducción de José Luis Abellán, Madrid, Tecnos 1966.

APÉNDICE

MATEMÁTICAS

MICHEL SERRES
Los orígenes de la geometría

ISAAC NEWTON
Principios matemáticos de la filosofía natural

LOS ORÍGENES DE LA GEOMETRÍA

MICHEL SERRES

[Publicado en Michel Serres, *Los orígenes de la geometría*, México, Siglo XXI Editores, 1996, pp. 157-174, 255-273]

PRIMERO EN LA HISTORIA: TALES

De la pirámide al tetraedro: origen óptico

Diógenes Laercio: "Jerónimo dice que Tales midió las pirámides a partir de su sombra, después de saber la hora en que nuestra propia sombra iguala a nuestra estatura." *Vidas, doctrinas y sentencias de filósofos ilustres*; Tales, I, 27.

Plutarco: "…a él le gustó tu manera de medir la pirámide… Colocando solamente tu baastón en el límite de la sombra arrojada por la pirámide y haciendo que el rayo tangente del sol engendrara dos triángulos, tú has demostrado que la relación de la primera sombra con la segunda era también la de la pirámide con el bastón. Pero también te han acusado de no amar a los reyes…" *Sept. Sap. Conv*. II, 147 A.

Estos textos ponen en escena el teorema de Tales, cuyo esquema compara un primer triángulo formado por la altura de una pirámide, su sombra proyectada en la arena y el rayo de sol rasante, con un segundo, constituido, a su vez, por un cuerpo cualquiera, accesible en su altura, por la proyección también de su sombra, y por un rayo luminoso semejante: ambos rectángulos, de ángulos iguales, son homotéticos.

Jerónimo informa de un caso particular de triángulos isósceles y Plutarco del caso general. Eso depende del momento del día: no es posible observar el primero más que en un instante único.

Mediante los dos gráficos del ilustre teorema, ¿describen esas fuentes cierta aplicación o, por el contrario, el origen: eso que nosotros llamamos el milagro griego, la aparición de una forma y de un razonamiento abstractos sobre el fondo de una práctica o de una percepción previas, alineamiento óptico y medida de los cuerpos?

¿Cómo leer esos relatos, auténticos o míticos? ¿Lo sabremos alguna vez? Veamos algunas leyendas.

Ardid de origen

Tenemos, pues, la pirámide y su sombra proyectada: ésta es accesible, puesto que puedo medir directamente esa mitad negra del monumento funerario; pero es inaccesible la altura de la tumba o la del Sol.

Augusto Comte: "Debemos considerar como suficientemente comprobada la imposibilidad de determinar, midiéndolos directamente, la mayoría de los tamaños que deseamos conocer. Es este hecho general el que exige la formación de la ciencia matemática… Pues, renunciando, en casi todos los casos, a la medida inmediata de los tamaños, el espíritu humano tuvo que buscar cómo determinarlos indirectamente, y así fue como se vio conducido a la creación de las matemáticas." La geometría resulta de un ardid, de un sesgo, en el que la ruta indirecta permite acceder a aquello que no consigue una práctica inmediata. Ésta consiste, aquí, en construir una reducción de la pirámide: no importa cuál sea el objeto vertical, nuestro cuerpo por ejemplo. De hecho, Tales descubre el

módulo o modelo reducido. Para acceder a la inaccesible pirámide, inventa la escala.

De ahí, nuevamente, Augusto Comte: "Así fue, por ejemplo, como Aristarco de Samos calculó la distancia relativa del Sol y de la Luna respecto de la Tierra, tomando medidas sobre un triángulo construido lo más exactamente posible, de forma que fuera semejante al triángulo rectángulo formado por los tres astros en el instante en que la Luna se encuentra en cuadratura, momento en que, por consiguiente, bastaba para definir el triángulo, observar el ángulo con la Tierra." Como Tales, Aristarco fabrica el modelo reducido de tal situación astronómica. Medir lo inaccesible consiste en reproducirlo o imitarlo en lo accesible.

Observemos el caso de las naves en el mar: comentando la vigesimosexta proposición del primer Libro de *Los elementos* de Euclides, Proclo escribe: "En sus *Historias geométricas*, Eudemo hace remontar ese teorema a Tales; pues dice que este último tuvo necesariamente que servirse de la manera en que, según relata, determinaba la distancia de los barcos en el mar." En su *Geometría griega* (p. 90) Tannery reconstruye la técnica de medición inspirándose en la célebre *fluminis variatio* del agrimensor romano Marcus Junius Nipsus.

Se trata, en todos los casos, de trasponer a lo próximo, miniaturizándola, una situación de posiciones inabordables.

¿Qué es la aplicación?

Accesible, inaccesible, ¿qué quiere decir eso? Próximo, alejado; tangible, intocable. Directa o inmediata, la medición exige operaciones de superposición, en el sentido en que una métrica recurre a una ciencia aplicada; pero sobre todo al sentido del tacto.

Esa unidad o esa regla se aplica sobre la cosa a medir: colocada sobre ella, la toca tanto como sea preciso; inmediata o directa, la medida es posible o imposible en tanto que esa superposición lo sea o no. Así, lo inaccesible se convierte en ese intocable hacia el que no puedo transportar la regla, o aquello sobre lo que la unidad no puede superponerse. Pasando de la práctica a la teoría, la astucia imagina un sustituto de esas longitudes a las que mi cuerpo no puede acceder: la pirámide, el Sol, el navío en el horizonte, el otro lado del río.

La matemática descendería de los circuitos de esas astucias.

Tocar o ver: ¿el origen está en nuestros sentidos?

Eso conduce a subestimar el alcance de las actividades prácticas o a encerrarlas en nuestras manos. Puesto que, en fin, esos circuitos consisten en pasar del tacto a la vista, de la medición mediante superposición a la mirada. Aquí, teorizar vale por ver, que es lo que dice la lengua griega. La vista es un tacto sin contacto. Descartes, que sabía lo que es una medida, describía la mirada del ciego en la punta, alejada, pero táctil, de su bastón. La mirada accede algunas veces a ese inaccesible. De ahí la agrimensura a ojo del Sol y de la Luna, del barco y de la pirámide.

Tales descubre las virtudes precisas de la mirada y organiza sabiamente una escena de luz, la representación óptica. A falta de poder transportar una regla, transporta líneas de visión o deja que la luz las proyecte sin él. El pragmático Comte piensa con sus dos manos, sin comprender a Tales el contemplativo, cuyos ojos no hacen más que dejar que las cosas se alineen por sí mismas. Nada tan exacto como una alineación de marcas marinas.

Que yo sepa, incluso para los objetos accesibles, la vista por sí sola me asegura que la regla se superponga sobre ellos. Medir o alinear: el ojo sólo testimonia ese recubrimiento. El de Tales conduce lo visible a lo tangible.

Medir es relacionar. Sí, pero la relación supone un transporte: el de la regla, el del punto de vista, el de las cosas recubiertas por un alineamiento. En lo accesible, el transporte siempre es posible; para lo inaccesible, sólo la vista se encarga del desplazamiento: de ahí el ángulo de visión, de ahí la sombra que llamamos proyectada.

¿Quién relaciona, quién transporta? Ni usted, ni yo, ni nadie, con sus manos. Esperemos a que la luz conduzca la sombra hacia nuestros pies.

El fin último de este libro encontrará ese transporte.

Espacio y tiempo: primer origen astronómico

Desde Diógenes o desde Plutarco, los esquemas presentan cosas que cambian y otras que permanecen. Inmóvil desde hace diez siglos bajo el cielo de Egipto, he ahí la pirámide, invariable; variables, por el contrario, el movimiento aparente del Sol, la longitud y la posición de la sombra. La experiencia corriente dicta que aquéllas dependan a la vez del astro y del monumento.

De ahí la figura del *gnomon*, eje o poste en pie, cuyo rastro dice la hora. La medida de las variaciones de la sombra ritman el curso del Sol. He aquí el cuadrante solar, objetivo civil o astronómico en el que las medidas

espaciales indican el tiempo. De ahí, en Diógenes y en Plutarco, los restos del viejo problema del momento: alcanzar el instante de igualdad entre la sombra y la altura, u observar las dos sombras en un mismo momento de la jornada; dejar que el sol escriba sobre la arena su curso diurno. De ahí la cita de Aristarco: mejor que un reloj, he aquí un observatorio astronómico. Hablaremos de ello dentro de poco.

Invirtiendo todo ese proceso, Tales se plantea y luego resuelve el problema inverso del *gnomon*. En vez de dejar que la pirámide hable del Sol, o sea que el invariante declare la escala de lo variable, pide al Sol que hable de la pirámide, es decir a lo cambiante que diga constantemente algo de aquello que permanece. Astucia más profunda que la de Comte: el invariante no discierne ya las desviaciones regulares de lo variable, sino que a la inversa, en lo variable, Tales discierne el invariante estable y descubre lo desconocido.

Mejor aún, mediante el *gnomon*, quien medía el espacio medía el tiempo. Invirtiendo de nuevo los términos, Tales detiene el tiempo para medir el espacio, fija el curso del Sol en el instante singular de los triángulos isósceles, homogeneiza el día para el caso general.

¿Es preciso verdaderamente congelar el tiempo para concebir la geometría? Bergson quería también que la inteligencia geómetra, toda y siempre espacial, se divorciase de la duración.

Origen óptico

Lo esencial, decimos nosotros, está en el transporte. Puesto que si solamente la medida puede desembocar en resultados exactos, sólo la relación o la referencia del esquema gigante con el modelo reducido accede al rigor.

Las génesis precedentes se relacionan con transportes: reducción o paso del tacto a la vista y viceversa, inversión de la función gnomónica, intercambio de lo estable y lo variante, sustitución del espacio por el tiempo.

Estable por el movimiento aparente del Sol, al menos en su segunda versión, el esquema de Tales diseña un diagrama óptico. Pero la vista y su espectáculo suponen: un sitio o punto de vista, una fuente de luz, el objeto, por último, umbroso o claro. De aquí surgen nuevas preguntas.

¿Dónde situar el punto de vista? No importa dónde. En la fuente de luz o al ras de la tierra. Puesto que el alineamiento de marcas hace posible la aplicación, la relación y la medida, podemos ver alineados o el Sol y lo alto de la tumba, o la cima de la pirámide y el punto extremo de la sombra proyectada. El sitio puede desplazarse.

¿Dónde encontrar el objeto? Hay que hacerlo, también a él, transportable: mediante la sombra proyectada o transportada; o por el modelo que lo imita.

¿De dónde viene la fuente de luz? En el caso del *gnomon* ella variará, y transporta al objeto bajo la forma de sombra. Ella va, y a esto lo vamos a llamar el milagro, a residir en el objeto.

Orígenes múltiples

Balance temporal: nueva proliferación de génesis finos. ¿Cómo llegó la geometría a los griegos? La fabricación de un modelo reducido y el transporte de lo alejado a lo próximo marca un origen pragmático; la representación visual de aquello que no se puede tocar, muestra otra, más sensorial; la inversión de la cuestión del *gnomon* indica un origen civil, geográfico, a partir de la astronomía; pero también conceptual o estética, puesto que borra el tiempo para medir el espacio; incluso epistemológico, cuando cambia los papeles de lo variable y de lo invariable. En las fuentes de la geometría confluyen, pues, varias génesis.

Remontaremos, dentro de poco, nuevos afluentes.

Señal del teorema: el origen mnemotécnico

Otro avatar del transporte, interceptamos primero, de paso, el mensaje. Pues los dos fragmentos citados parecen explicar menos una constitución que poner en escena una forma que ya estaba ahí: el teorema de Tales.

La primera leyenda, con varias génesis, descifra la matemática extrayendo el esquema implícito del relato anecdótico, a propósito del cual el comentario expone cierto color local destinado a mostrar que el sabio griego lo aprendió todo de los sacerdotes de Egipto. La relación de la forma circunstancial con el esquema da en efecto menos a pensar la invención del segundo en la acción relatada por la primera que el recubrimiento de éste por aquélla.

Suponiendo, pues, que yo desee acordarme del teorema de Tales, la historia de la pirámide puede servirme de ayuda mnemotécnica. En una cultura de tradición oral, el relato ocupa el lugar del esquema, escena vale por intuición, y el espacio viene en ayuda de la memoria. El diagrama del teorema se transmite luego por escrito; pero, de la boca a la oreja, la dramatización mejora el vehículo del

saber. Más vale reconocer, entonces, en el relato, no tanto una leyenda originaria como la forma misma de la transmisión; él comunica un elemento de ciencia más que testimoniar su emergencia.

Aquí la matemática proporciona la clave de la historia, y no la historia la de la matemática. El esquema declara el objetivo del relato y no el relato el origen del esquema. Saber, entonces, y, en especial, saber el teorema de Tales, consiste en acordarse del cuento egipcio y enseñarlo, en relatar el pseudo-mito de origen. Así presentado, el más ignorante no tiene ninguna dificultad para recordarlo, inolvidable.

El espacio de los transportes: circunstancia idéntica al esquema

¿Qué es lo que se transmite o se transporta?

El teorema de Tales se reduce, ya lo sabemos, a una presentación del concepto profundo de similitud en el espacio formal de los transportes.

Profundizar el esquema hasta sus consecuencias más abstractas permite reencontrar la variedad vivida, circunstancial y coloreada del relato.

Si, por lo tanto, el teorema se remite al grupo de las similitudes, inscribible sobre o en ese espacio en el que los transportes no deforman las formas, entonces, venido de viaje a las pirámides, Tales no ve más que objetos de la misma forma y de dimensiones diversas.

La percepción de las tres tumbas se desarrolla en el espacio de las similitudes, como si éste se constituyera en esos lugares electivamente: cada uno es otro y el mismo, como todos los triángulos del teorema de Tales.

¿Se ha inventado otra cosa que lo que ya estaba ahí?

Estrictamente fieles al concepto, la historia o el relato, evidentemente y visualmente se asemejan a la idea mimética de lo semejante y la imitan.

Otro transporte: elevación hacia el volumen

La sombra negra del edificio mortuorio se estira sobre la planicie del desierto, en el plano. Si no miramos más que esta proyección plana, nos quedamos en la métrica de dos dimensiones, la de los agrimensores o la de los harpedonaptas, medida agrícola o arquitectural, alegada por las *Historias* de Herodoto, las técnicas usuales del agrimensor, que escribe y dibuja; permanecemos en la representación tal como nos la da la escritura sobre una tablilla o un papiro, ambos planos.

El teorema de Tales no escribe sino que muestra, en el espacio, que el plano se sumerge en la oscuridad, que toda representación plana, discurso o esquema escritos, no accede jamás más que a una sombra negra: el escriba no accede a las luces del nuevo conocimiento.

Origen del espacio de los desplazamientos

Para comprender los acontecimientos del plano, los gráficos y la escritura, hay que elevarse hacia otra y nueva representación en el espacio de tres dimensiones: toda esta historia de Tales se desarrolla, en efecto, delante o en cuerpos voluminosos de los que nunca se puede obtener una completa representación, porque sus diversos planos proyectados, dibujados o escritos, no muestran más que perfiles parciales, difíciles de descifrar.

¿Quién ve una pirámide en esos rasgos perfilados de triángulos y de bases poligonales? ¿Quién adivina los caminos aéreos de las superficies y de las líneas, complejas y enmarañadas según la profundidad, en el gráfico simplista que recibe de ellas un corte? A primera vista, prejuzgamos enigmático todo lo que se inscribe planamente en el plano.

Así pues, para saber y comprender, para ver, hay que poder desplazarse según la nueva dimensión, siguiendo el sentido que, justamente, en el curso de la proyección, acompañan los rayos del sol.

El espacio se convierte en un conjunto de desplazamientos posibles.

Liberación respecto de la escritura

El astro ilumina el espacio, pero se acuesta a ras del plano, dejando la escritura en la noche. Tales libera las matemáticas del escrito, asimila, aquí, a las inscripciones fúnebres en la sombra de la tumba del faraón, rey lapidado, bajo las piedras.

De ahí proviene el milagro maravilloso: los elementos de la geometría no pueden seguir siendo los de la lengua hablada ni los signos de la escritura, sino que vienen de afuera, de otro espacio, tan diferente del espacio usual, plano, de la representación escrita, como el exterior mundial de la caverna se distingue del muro plano que miran, en la sombra del fuego maligno, los prisioneros platónicos de la representación óptica artificial. En efecto, sólo la tercera dimensión permite resolver los problemas imposibles de tratar solamente en el plano.

Así como, en el *Menón*, ni Sócrates ni Platón dicen que, para resolver la cuestión de los lados, insoluble siguiendo su medida lineal, hay que huir en una dirección diagonal, abriendo la segunda dimensión, así hay que abrir la tercera para hacer resolubles diez cuestiones planarias. En esta historia originaria el Sol no sólo es portador de la luz, sino que constituye el fondo y la condición de un espacio voluminoso y transparente. Tales inventa esta estereometría que Platón pretende que constituye la única verdadera geometría, aun cuando vemos que, en realidad, ella la funda.

Vuelo

En el libro de sus *Historias* consagrado a Egipto, Herodoto relata (II, 147-148) que a la muerte de un sacerdote de Hefestos que había reinado solo, aparecieron doce reyes y se repartieron el país y el poder en otros tantos lotes, imponiéndose una ley de nunca destruirse entre ellos ni dominar unos sobre otros; permanecieron, pues, amigos. Sin que podamos comprender claramente por qué, decidieron construir, por primera vez, el famoso laberinto, como monumento común, sin duda en lugar de una pirámide, que era la forma de la tumba del poder cuando residía en manos de uno solo. ¿Nos perdemos para siempre en los corredores interminables y las encrucijadas recomenzadas de un poder compartido?

A ese análogo inesperado de la forma piramidal se agrega una leyenda semejante precisamente a la de Tales. Convertido en símbolo de una dificultad tan grande que quien entra en ella se extravía, el laberinto, cretense ahora, fue construido, se dice, por Dédalo, inventor inteligente, por añadidura, de una célebre técnica de vuelo, fatal para su hijo Ícaro que quiso, también él, salir de la noche, en donde uno se pierde, hacia el Sol, a la luz del cual uno se reencuentra. Para salir del dédalo, no hay, pues, sino dos soluciones: o el hilo de Ariadna o el vuelo vertical. La primera pone en escena el pensamiento algorítmico que aparecerá más adelante en este libro; la segunda, el invento de la geometría: reencontramos el punto alto anteriormente dibujado.

¿Cómo decir mejor que los dibujos planarios plantean problemas insolubles de los que se sale retomando la tradición, antigua en el creciente fértil, de las operaciones reversibles y paso a paso, o bien tomando prestada la tercera dimensión?

La cuestión del origen se resume en esas imágenes.

Origen arquitectural

Más aún que de percepción o de conducta corporal, los dos fragmentos citados hablan de técnicas y de arquitectura, pues la similitud descubre un secreto de construcción: a la mirada como al espíritu, las tres pirámides vecinas hacen brotar, en efecto, el espectáculo de la homotecia. Como el poste o el cuerpo de pie, Kefrén y Micerinos reproducen, ya, modelos reducidos de Keops. Para edificarlas semejantes, era preciso tener a Tales y su teorema. Física y técnicamente, una filosofía de la *mimesis* recomienza, como si se reencontraran, del lado de la *physis* y de la *praxis* los usos del *nomos*.

La talla y la disposición de las piedras suponen, pues, el teorema: prácticas ciegas a semejante saber, o aplicación de un concepto claramente explícito: he aquí una verdadera cuestión.

En las técnicas, ¿el origen de las ciencias?

¿Cuál es el estatuto de un saber implícito en una técnica? ¿Se reduce ésta a una práctica que envuelve una teoría? Toda la cuestión —aquí la del origen— se resume en una interrogación sobre la modalidad de este envolvimiento. Las matemáticas emergen a veces de ciertas técnicas: ¿explicitando un saber implícito?

El que encontremos con frecuencia un secreto en las tradiciones artesanales, a menudo significa que permanece un secreto para todos, incluyendo para el maestro o para el inventor.

Porque un saber claro se oculta en las manos y en la relación obrera con las piedras y con los ladrillos, y puede permanecer encerrado, guardado con doble llave, como en la sombra de la pirámide.

La sombra del secreto

Contemplemos ese teatro primordial del saber, la puesta en escena o el relato de los orígenes: el secreto del constructor y del tallador de piedras, negro para ellos, para Tales y para nosotros, se oculta en la sombra: bajo la sombra proyectada de las pirámides, inmensas cajas negras, Tales se sitúa en lo implícito de un saber, que el Sol, detrás, explicita.

Todo el problema de la relación entre el esquema y la historia, entre el saber implícito y la práctica obrera, se plantea en términos de Sol y de oscuridad, como dramatizado al estilo platónico: el astro deslumbrante del cono-

cimiento y de lo mismo luce, mientras que se extinguen la opinión, los oficios empíricos, los objetos del mundo, en esta sombra.

El pliego abierto o cerrado

Nuestras primeras lecturas acaban de desvelar el saber implícito que oculta un objeto fabricado.

En general, captar la naturaleza de una teoría mezclada con una actividad obrera es fácil como vía normal de la ciencia: fácil, factible, es decir, a veces difícil, sin duda, pero no imposible; complicado o complejo, ciertamente, pero resoluble al fin. Lo difícil, incluso inextricable, es decir indefinidamente explicitable con los residuos esenciales siempre recomenzados, lo que permanece embrollado para siempre, es describir el pliego de esta implicación que, en el otro sentido, se vuelve más claro bajo el nombre de aplicación. El origen del saber a partir de una práctica queda del lado de la sombra, mientras que el origen de una práctica a partir del saber viene del lado de la luz.

La sombra muestra los pliegues en donde yace oculta la ciencia. La actividad técnica de origen pone el saber en la sombra, y nosotros mismos permanecemos ciegos, tanto como agentes, como en la medida en que tratamos de colocar la teoría en la luz.

El nuevo logos

La pirámide proyecta una sombra y cada uno ve la suya, transportada, bajo el Sol de Egipto. ¿Qué hacer o saber sin medir la relación entre las dos sombras, la del objeto altivo y la del sujeto activo, sino estimar la relación entre el secreto que duerme en las piedras y el que cierra los ojos del practicante? ¿La relación entre los dos secretos dice, designa, describe el secreto de la relación del hombre con su objeto trabajado?

En esta leyenda primordial, la geometría de Tales expresa, pues, la relación entre dos cegueras, entre la práctica negra y su sujeto arrojado en esa ceguera. Ella dice y mide el problema, pero no lo resuelve; dramatiza su concepto, pero no lo explica; diseña admirablemente la cuestión sin responderla, refiere la relación de dos cifras, la del albañil y la del edificio, sin descifrar ninguna de ellas; y quizá nunca es posible más que hacer eso, si se permanece en el *logos*.

Puro, aquél pierde sus contenidos y no dice nada fuera de esa relación; ya no designa una palabra llena de sentido, ni un verbo, fuerte por sus acciones, ni la luz advenida por la palabra, sino que agrega entre ellas dos instancias que pueden traernos sin cuidado conocer. Inaudito, desconocido, ese nuevo *logos* corre el riesgo inmenso de la ausencia de todo sentido. Avanza en el conocimiento tendiendo un puente entre dos ignorancias: resplandor nuevo surgido de una oscuridad doble.

Replicación

La relación entre dos sombras, he ahí el problema en su designación, la apelación pura del modo de envolvimiento de un saber por una técnica. La medida, astucia de aplicación o, como dice Auguste Comte, vía indirecta, repite la implicación pero no la explica. De una técnica, Tales extrae otra, de una práctica obtiene una práctica; ciertamente, la arquitectura y la medición encierran ambas el mismo saber, la homotecia y el famoso teorema; sin embargo, la aplicación se repite en otra aplicación, como si de un pliegue se desplegase otro pliegue. La homología de la repetición acaba por decir de nuevo la homotecia, pero cada vez en la ganga de lo aplicado. La teoría expresada por las sombras permanece en la sombra. Ella no ha nacido, en su pureza, ese día: como dice Platón, como lo repetirán los siglos por venir, la geometría no se reduce en absoluto a esta métrica, simple propedéutica que inaugura un largo camino de ciencia.

Más milagro originario: las técnicas se engendran y se perpetúan en la repetición, la medida ve el teorema diferentemente que la arquitectura, eso es todo. Y nosotros permanecemos en la gran sombra del secreto. Pues, de nuevo, no sabríamos pensar el origen de la técnica sin el origen del hombre mismo, *faber* de su emergencia o, mejor, emergente porque *faber*. En el origen, la técnica permite la perpetuación y la repetición del hombre y de la técnica.

Así Tales repite su propio origen, como el nuestro: su métrica de pre-geómetra replica o, lo más sencillamente del mundo, designa de otro modo la modalidad de nuestra relación técnica con los objetos; la homología del fabricante y de lo fabricado ocupa su lugar en la cadena abierta de esas afirmaciones y de sus designaciones, pero no da la clave de la cifra, ni sacan a relucir la articulación secreta del saber y de la práctica en donde se encontraría lo esencial de un posible origen.

Relación entre dos sombras, dos secretos, dos formas o dos rastros; relación, decir vacío que transmite esa relación; esta geometría arcaica mide el problema, toma sus dimensiones, lo plantea, lo pesa, lo hace ver, lo relaciona, pero no lo resuelve.

¿El *logos* de las sombras sigue siendo aún una sombra del *logos*?

Desaparición del sujeto, objeto proyectado

Desde su aurora, la matemática de Tales enuncia sin embargo el descentramiento del sujeto del pensamiento claro por relación al cuerpo que proyecta su sombra transportada: situado más allá del monumento, el sujeto Sol deja el cuerpo del astrónomo del lado de los objetos del mundo o entierra el sujeto conocedor tan profundamente como el cuerpo del faraón, en la sombra. ¡Qué inversión copernicana, ya, en esta representación de los dos triángulos bajo las antorchas del solsticio! El sujeto-Sol escribe sobre la arena formas cambiantes como perfiles, que describen un ciclo de representación. Cada momento de este ciclo, detenido, fijado en la arena plana, está sin embargo provisto de una invariante: la relación estable con la misma sombra, en el mismo momento, de otro objeto, yo por ejemplo. La medida perspectiva habla de una invariante para las variaciones de la representación. Las sombras transportadas cambian, pero tienen entre ellas una relación que, por su parte, no cambia, y que abre lo desconocido, el secreto de la pirámide, la inaccesible altura. Voluble, la representación designa una estabilidad propia del objeto, su medida.

De donde se desprende que, aquí colocado, yo no pueda conocer claramente del volumen más que lo que escriben o describen las sombras proyectadas, las informaciones transportadas sobre la arena por un rayo de Sol tras la intercepción por las aristas y la cima del prisma opaco.

¿Cómo llamar a esta geometría? ¿Una perspectiva, una arquitectura, una física, una óptica?

Representación

El teatro de la medida muestra la descodificación de un secreto, el desciframiento de una escritura, la lectura de un dibujo. La arena en donde el Sol deja su rastro se convierte en la pantalla, la pared de proyección en el fondo de la caverna. He aquí la escena de la representación celebrada milenariamente por el saber occidental, la forma históricamente estable de la contemplación desde lo alto de estas pirámides.

La historia de Tales instaura, tal vez, ese momento de la representación, indefinidamente reasumido por la filosofía, pero sobre todo por las geometrías, de las coordenadas de Descartes al punto de vista arguesiano, del boceto descriptivo a Monge y Gergonne… Primera palabra de una perspectiva, de una proyectiva, de una óptica arquitectural de los volúmenes, de una matemática intuitiva, enteramente sumergida en el *organon* global de esta misma representación.

Pero de la sombra se olvidó, desde Tales hasta nuestros días, que era portada, transportada por algún soporte, que ella transportaba una información. De este primer análisis espectral se hacía la lectura sin descubrir la condición. La gran pregunta, ¿qué mensajero transporta —y cómo— tal y cual mensaje?, se hallaba recubierta durante siglos por la escenografía resplandeciente de la diferencia sombra-luz.

Historia de lo aplicado

Sí, la historia de Tales se parece a la de Platón: el Sol mismo, el objeto otro y empírico, la sombra (proyectada) de la sombra (misma), la similitud mimética, el plano umbroso de la representación; o a la historia de Desargues: talla de las piedras, geometría de los perfiles, teoría de las sombras… Aunque Descartes intervenga, luego Monge y tantos más, trabajan y trabajan aún del lado de la aplicación al mismo tiempo que del de la representación; perpetuando la habilidad de los ingenieros, hacen sobrevivir, pues, el arcaísmo de las pre-matemáticas y bloquean el nacimiento de la mencionada ciencia en su pureza. Pero ésta emerge cuando muere esta habilidad: casi no queda ninguna. Husserl escribe *Origen de la geometría* a la hora sonante de su desaparición, como si algún ciclo histórico por fin se cerrase. El relato conservado de Tales describe aún una métrica, pero no refiere el nacimiento de las matemáticas.

Para prueba Platón, que pide otra cosa para que se realice el milagro: la realidad esencial de las idealidades. Pregunta: ¿cómo puede nacer la pirámide misma como forma ideal?

Regresemos, para responder a eso, al análisis espectral.

Las entrañas negras de los volúmenes

Platón hace retroceder hasta el fondo de la caverna la escena de Tales: el volumen escribe la sombra proyectada sobre una pared plana y clara, la luz describe sobre el sólido su sombra propia. El saber se limita a las dos sombras, he ahí la sombra del saber.

Pero he aquí una tercera, de la cual las otras dos traducen la imagen o la proyección, secreto profundo enterrado en las entrañas del volumen.

Sin duda, el verdadero saber de las cosas del mundo yace en la sombra esencial de los sólidos, en su compacidad opaca y negra, encerrada para siempre tras las múltiples puertas de sus bordes, atacadas solamente por la práctica y por la teoría. La talla puede hacer estallar la piedra y la geometría dividir o duplicar el cubo; he ahí que los sólidos, no agotables mediante el análisis de sus caras, conservan siempre, al abrigo, un núcleo de sombra a la sombra de sus bordes: hay que recomenzar.

De ahí regresamos a la talla de las piedras y a la pirámide. Volumen de volúmenes, poliedro compuesto por bloques recortados, he ahí el edificio. Ahora bien, de semejante sólido, ¿cómo tomar conocimiento sino por proyección planaria? ¿Y cómo tomarlo en las manos sino atacando sus caras?

La geometría de Tales dice eso, y lo dice al mismo tiempo que la técnica arquitectural y la práctica del albañil. Se trata, en los tres casos, de tratar un sólido por la reunión de todas las informaciones recogidas sobre los diversos planos que pueden hablar de ello: secretos de las sombras propiamente dichas y de las sombras proyectadas. Un volumen se expresa por sus proyecciones, que suponen un punto de vista y un dibujo sobre una superficie lisa, ella misma sin sombra propia y sin repliegue oculto. Pero, leyendo y destacando esos rasgos del volumen, Tales no descifra ningún secreto sino el de la impotencia para penetrar en los arcanos de lo sólido, en donde lo cerrado se abriga indefinidamente tras lo abierto necesario para toda información, en donde el saber está para siempre sepultado, de donde la historia infinita de los progresos analíticos brota como de un manantial.

Su historia relata entonces una sucesión inusual de este enfrentamiento con los objetos sólidos, el ataque de los volúmenes compactos, tomados como temas teóricos objetivos indefinidos. La cosa existe como tal, desconocida y correlacionada, secreto involucionado en pliegues y repliegues por esencia inaccesibles, puesto que la explicación despliega y por lo tanto deja, detrás de la cara de lo abierto, lo cerrado plegado sobre sí mismo.

O reconozco el objeto de las dos sombras, la propia y la proyectada; o admito un tercer núcleo de sombras en su seno: entonces teoría y práctica desarrollan infinitamente ese secreto en una historia siempre abierta, la de la ciencia, que admite que todas las cosas implican siempre lo explicable.

Fiat lux

Entonces, la historia comenzada, según se dice, en el delta del Nilo se enrosca por un golpe súbito de una audacia increíble: la negación radical de esas sombras interiores.

El Sol de Tales y de Ra, cuyos rayos, rectamente interceptados, recortan una impecable definición de los triángulos negros, se relaciona con el débil fuego de los prisioneros de la representación en la caverna platoniana, tan llena de humo que todos lloran enceguecidos. Solamente resplandecen los lados o líneas puras, en razón de esos rayos y formados por ellos, así como las puntas o cimas, puntos luminosos, pequeños diamantes sin dimensión en donde los trazos irradiantes convergen. Otra vez los bordes.

Fuera, el nuevo Sol emite una luz trascendente que traspasa las cosas y transmite una visión pasa-muralla. He aquí que se realiza el milagro maravilloso: la transparencia de los volúmenes, apelación metafórica del realismo de las idealidades.

De la caverna al exterior, cambia la escenografía por una iconografía: la sombra de los sólidos actuaba en el plano de la representación y los definía mediante límites y cortes; la luz, ahora, los atraviesa y expulsa la sombra interior. En lugar de la indefinida triangulación de la geometría, he aquí la estereometría de las formas vacías mediante la *epifanía* de la diafanidad.

He aquí el espacio de la geometría pura, atravesada por la intuición del vacío transparente. Entonces y solamente entonces nace la pirámide, el tetraedro puro, primero de los cinco cuerpos platonianos.

Milagro, he aquí el Sol en la pirámide: el lugar, la fuente, el objeto, se reúnen en un mismo punto.

La caja negra de la tumba, del lado del *nomos* o de los usos religiosos y de las leyes civiles relativas a la muerte, se convierte en una caja blanca del lado de la *physis*, bajo la claridad del Sol.

Tercera sombra, segunda epifanía: las cajas blancas

Así como la luz resbalaba sobre el foco diamantino, la recta radial y el plano, tan brillante que la *epifanía* dibujaba, desde los pitagóricos hasta Euclides, la superficie percibida como centelleante, asimismo bajo el nuevo Sol los sólidos ya no contienen ni sombra ni secreto, la misma claridad los atraviesa, los pasa sin intercepción: constituyen a partir de ahora un mundo real cognoscible de par-

te a parte. Se comprende la importancia constantemente concedida por Platón y su escuela a la estereometría de los volúmenes.

La titubeante percolación de las explicaciones infinitas, de las explicitaciones agotadoras en pliegues cerrados sin recurso, se clausura con ese golpe de fuerza, con un resplandor que desgarra los velos de sombra y cuya luminosidad súbita excluye toda tiniebla. Ya sin ningún espectro ni análisis, las tres sombras, la propia, la proyectada, la enterrada, son arrebatadas juntamente por el Sol del Bien. A partir de ese auténtico milagro, la aparición de estas formas llamadas trascendentes, de estas cajas blancas y vacías; en fin, sin obstáculos ya *por la definitiva expulsión* de las sombras, brota de ese recipiente-fuente el nuevo discurso interminable del gran relato de la Geometría.

Y, como para rizar el rizo, con todo rigor y para la coherencia de la historia global, el *Timeo* va a constituir el mundo por medio de los cinco cuerpos platonianos, transparentes y blancos: geometrizada, la Tierra global se convierte, íntegramente, en una caja blanca. La geometría, por fin, ha sido bien nombrada.

Todos los pequeños cursos o hilos, perseguidos hasta aquí, desembocan en la fuente señalada por ese recipiente blanco, vacío y traslúcido, más allá del umbral de la percolación.

Pero el primer cuerpo, el más simple, el tetraedro justamente, designa el fuego. El helenismo hace nacer la pirámide pura bajo la hoguera del Sol, y de ese tetraedro hace renacer el fuego. ¿El del físico Heráclito?

Doble milagro que cumplen las escrituras, la leyenda egipcia y la iniciación de la intuición, colocando la fuente de luz en el seno mismo del poliedro. Cuando la pirámide es ella misma fuego —¿acaso su denominación ha influido en su leyenda? —, el Sol la atraviesa.

Mística más física igual a geometría

Todo ese relato de origen, desde Tales hasta *La República*, se sumerge en una visión o dramatiza un rito de fuego. El nuevo geómetra no percibe ya sombras bajo la hoguera conjugada de la forma pura y del foco solar: gemelidad original de la estereometría matemática, de la física elemental y del entusiasmo religioso, atmósfera enceguecedora de las primeras filosofías de la intuición.

El núcleo del saber es envuelto incesantemente por el mito, cuyo relato no deja de conjugarse con el teatro de la representación: teoría, visión, luz, fuego.

Nueva génesis de múltiples ramas, en donde las técnicas naturales y la historia de las religiones, como dos afluentes, se mezclan; y con la astronomía y la óptica, la métrica y la arquitectura, la talla de piedras y la devoción solar, para librar a los objetos de sus obstáculos negros.

Nomos y *physis* se anudan en y por el proceso de exclusión, común a los dos gestos.

Blanco y negro

Atravesando el sólido, la luz anuncia y produce una historia, la de la primera geometría.

Pero el futuro del cuadrado, de la diagonal, se jugará tanto sobre la arena en donde lo describimos y a través del lenguaje que los codifica como en el cielo blanco de las formas.

Una filosofía de la representación baña todavía el realismo de las idealidades transparentes. Ciertamente, la iconografía sustituye a las diversas escenografías, pero sigue siendo una trans-representación desde el punto de vista divino. Para superar el de Tales, el teatro sin sombra no cierra todavía la escena. Tan pura y abstracta como se la conciba, la idea no se desmarca del ídolo. El realismo inevitable sigue siendo un idealismo.

La forma geométrica dice claramente esta dificultad: prejuzgada sin sombra ni secreto, no recela de nada que exceda a la definición que sea posible pensar, existente como idealidad, transparente a la vista como al pensamiento, conocida teóricamente de parte a parte, vista y destilada sin residuo; deslumbrada por su existencia, la intuición la atraviesa. Existiendo en sí misma, sin embargo, esta forma recela de los obstáculos que superan lo bastante al pensamiento como para obligarlo a inclinarse.

A medida que muera esta geometría tan pura, cuando nada pueda ya fundarse sobre la intuición, cuando el teatro de la representación, poco a poco, se cierre, el secreto, la sombra, la implicación van a explotar de nuevo, entre esas formas abstractas, ante los ojos de los matemáticos asombrados, explosiones continuadas a todo lo largo de la historia.

Todavía más sombras

La recta, el plano, el volumen, sus intervalos y sus regiones, caóticas, densas, compactas… irán muy pronto a hormiguear de nuevo repliegues y escondrijos negros.

Ni tan simples ni tan puras, las formas puras y abstractas, modeladas sobre las ideas simples, ya no son conocidas, vistas y destiladas sin residuos, sino que se convierten

en temas teóricos objetivos infinitamente replicados, de enormes virtualidades de noemas, como las piedras y los objetos del mundo, nuestras construcciones de bloques y nuestros objetos trabajados. La forma oculta, bajo su forma de núcleos transfinitos de los que se puede temer que la historia no baste ya para agotarlos, instancias fuertemente inaccesibles como tareas que nos superan.

El realismo de las idealidades se sobrecarga y retoma una compacidad que había disuelto el sol platónico. Llenas de sombra, las idealidades puras o abstractas se vuelven negras como las pirámides y, como ellas, hacen sombra. Nueva manera de reescuchar la vieja leyenda egipcia y métrica de Tales, de lo que nació antiguamente un discurso indefinido surgido de cajas blancas transparentes.

De esas estelas de sombras que no cesan, brotarán, a todo lo largo de la historia, nuevas luces y otros discursos interminables.

Nuestro origen

Así, la historia de las ciencias matemáticas resuelve sin agotarla la cuestión del origen. No acaba de responderla ni de librarse de ella. El relato de inauguración evoca ese discurso interminable que mantenemos sin reposo desde nuestra propia aurora, equipotencial al conjunto de las matemáticas.

De hecho, ¿qué es un discurso interminable? Aquél que se refiere a un objeto presente pero celado, a un objeto que viene pero que se ausenta inaccesiblemente.

Revelado pero velado al patriarca Abraham en Ur, en Caldea, hace cuatro mil años, Dios, adorados sean su nombre y su hijo, venido y por venir; el amor, venido desde los trovadores de la lengua de oc, en la Francia meridional, hace setecientos años, bendito sea en su paraíso perdido; y la geometría, bañando con su luz nueva la inteligencia de Tales, en Mileto, en Jonia, hace veinticinco siglos, y la nuestra, algunas veces…

…Dios, el amor y la geometría, porque ellos se retiran infinitamente de su presencia, extraen de nosotros un torrente irrepresible de gestos y de discursos, bellos, de los cuales, continuamente, nosotros nacemos, mejores.

MEDIDA DE LA TIERRA: HERODOTO

¿Qué es la geometría, una vez más y para concluir? La medida de la Tierra. Se trata menos, aquí, de su nacimiento que de su etimología: del origen de su nombre. ¿Qué tierra nombramos?

Al regreso de su viaje a Egipto, Herodoto da una buena respuesta a la pregunta.

Orígenes naturalistas

Nuestros predecesores leyeron su relato y nos han transmitido una leyenda, que es la siguiente: venido el tiempo regular, las crecidas del Nilo anegaban los límites de los campos cultivables en el valle aluvial que el río fecundaba; por eso, en el momento del estiaje, ciertos funcionarios reales, los harpedonaptas, de otro modo llamados agrimensores o geómetras, medían de nuevo las tierras confundidas por el fango y el limo para redistribuirlas o reatribuir las partes. La vida recomenzaba. Cada uno regresaba para dedicarse a sus labores.

Primera interpretación de la historia o del relato de Herodoto, versión fisicalista, en la que la tierra significa, simplemente y solamente, el área arable de la que el arado extrae, a fuerza de labor y de sementera, el arroz y el trigo, parcela local, agraria o cultivable, del campesino: el *pagus* precede a la página y al plano.

Las generaciones positivistas que nos precedieron no soñaban los orígenes sino a partir de la naturaleza o bien de la física; las religiones de los primeros dioses, decían, derivaban de terrores ancestrales inspirados por el mundo, los fuegos de los volcanes o de las tormentas, y las crecidas o las inundaciones; pensaban, pues, el origen de la geometría como el surgimiento de una ciencia natural.

O de la naturaleza misma: laborando el valle con sus aguas, el diluvio lleva la tierra al desorden, al caos del origen, al tiempo cero, exactamente a la naturaleza, en el sentido que esa palabra toma si queremos decir que las cosas se aprestan a nacer; la medida correcta la reordena y la hace renacer a la cultura, al menos en sentido agrícola. ¿Cómo no se la encontraría, de manera tautológica, como al principio, si ella expresa por sí misma el nacimiento?

En otro contexto, el *Génesis* escribe que, de las aguas primeras, Dios separó la tierra y la limitó. Al comienzo de los tiempos, de igual manera, vemos el barullo de la inundación seguido por la partición: las condiciones de la definición, de la medida y de la emergencia aparecen juntas a partir del caos: "a partir de", que significa comienzo, quiere decir también repartición, que es lo que yo quiero demostrar.

Pero aquí la tierra deja el campo local y las actividades agrarias para designar ya uno de los cuatro elementos del globo, llamado terrestre, en su totalidad.

La decisión sobre los bordes y fronteras parecía, en efecto, original; sin ella, no hay ningún oasis separado del desierto ni, perforando el bosque, ningún calvero en donde los campesinos se consagren a las labores agrícolas, ningún espacio sagrado o profano, aislados el uno del otro por el gesto de los sacerdotes, ninguna definición, encerrando un dominio, así pues, ni lenguaje preciso sobre el que entenderse ni lógica; ninguna geometría, en fin, al menos en el sentido de la métrica. Henos aquí de nuevo en las meditaciones de Anaximandro.

De golpe, no hemos descubierto más que el origen de las condiciones generales de una medida. Cuando supiéramos limitar los cuadros de cultivo y contemplar las riberas que separan los continentes del mar, no habríamos hecho ningún progreso en geometría, esta ciencia de un espacio tan distinto del de las sementeras.

Esta versión, cuya integridad y suficiencia resumen la segunda parte de este libro, ¿es suficiente, de hecho? ¿Logra expresar el origen de la ciencia abstracta de los griegos?

No, dos veces, a las dos preguntas.

Desde los orígenes culturalistas

¿De dónde viene el discurso de mi generación que, durante aproximadamente medio siglo, ocupó el campo de las ciencias humanas, y cuya demostración completó la versión fisicalista, venida de las ciencias duras, por la pregunta *ad hominem*: quién toma originalmente la decisión del reparto de tierras, de la división, de la creación de un bordo? El faraón, el rey, Sesostris o sus funcionarios, además de los sacerdotes egipcios de quienes Herodoto saca su relato.

La asignación de límites hace cesar, en efecto, los contenciosos entre vecinos; he aquí el derecho de propiedad, el de cercar exactamente un terreno y atribuirlo, he aquí el derecho civil y privado. Además, la misma delimitación por mojones permite al catastro real poner a cada uno en su lugar y fijar el monto del impuesto y de las tasas diversas: ¡he aquí el derecho público y fiscal! Los derechos abundan, pues, en esta leyenda de origen, en donde ellos solos toman la decisión y recortan los campos, cualquiera que sea la persona física, enviada por el faraón, el harpedonapta o geómetra misteriosos que de hecho los restituyen. ¿Quién decide? El legislador o quienquiera que sea el que dice o sigue la jurisprudencia y la hace aplicar.

Este, pues, realiza primero el gesto originario del que nace la geometría, que, en cuanto a ella, va a producir más adelante un acuerdo nuevo entre quienes demuestran, como si la justeza resultase todavía mejor que la justicia, y sobre un mismo terreno; pero esta última, en este punto, ha precedido a aquélla, identificándola consigo misma. Antes del sabio consenso sobre la precisión de la división o la necesidad de la demostración, un contrato jurídico se impone y para empezar pone de acuerdo a todas las personas afectadas.

Ahora bien, una vez más, así como el diluvio borró los límites y linderos de los campos cultivables, así desaparecieron, al mismo tiempo, las propiedades: de vuelta al terreno ahora caótico, los harpedonaptas lo redistribuyen y así hacen renacer el derecho borrado. Éste reaparece al mismo tiempo que la geometría; o, más bien, ambos nacen con la noción de límite, de borde y de definición, con el pensamiento analítico.

De la primera parte de esta obra vuelve ahora Anaximandro y el indefinido, precediendo a la definición de la forma precisa que implica las propiedades, para la geometría, del cuadrado o del rombo y, para el derecho, del propietario: sobre la misma palabra y sobre la misma operación, el pensamiento analítico se arraiga, y surgen de él dos ramas, el derecho y la ciencia.

El harpedonapta o agrimensor tira, sostiene, ata el cordel: su título misterioso se descompone en dos palabras en las que el sustantivo expresa el lazo y el verbo que él le fija. El comienzo es esta cuerda. La que, en un templo, por ejemplo, delimita lo profano y lo sagrado; la que evocan los términos contrato u obligación.

El primer sacerdote que, con ese extremo en su mano, habiendo cercado un terreno, halló a sus vecinos satisfechos con los bordes de su cercado común, fue el verdadero fundador del pensamiento analítico y, a partir de él, del derecho y de la geometría. Y fue así, por la firmeza del contrato, concluido para largo tiempo, por la exactitud y el rigor del dibujo, y por la correspondencia entre la precisión de éste y la estabilidad del aquél; pacto tanto mejor cuando sus términos se afinan, cuando los valores se precisan, cuando las partes se encuentran exactamente separadas. Estos requisitos caracterizan tanto el contrato definido por el jurista como aquél del que nace la ciencia.

La geometría, a la manera griega, refluye hacia la *Maat* egipcia que significa la verdad, el derecho, la moral, la medida y la parte, el orden surgido de la mezcla desordenada, un cierto equilibrio de justeza y de justicia, en fin, la rectitud lisa de un plano. Si un cronista egipcio cualquiera hubiese escrito esta historia, y no Herodoto, el comentario hubiera concluido con el nacimiento del derecho, como si los griegos hubieran orientado hacia la ciencia

un proceso de emergencia del orden que los egipcios orientaron hacia las formas del procedimiento.

El derecho precede a la ciencia y acaso la engedra; o mejor: un origen común, abstracto y sagrado, las une.

Antes de este origen, no se puede imaginar más que el diluvio, la gran crecida primera y recursiva de las aguas en la que el caos indefinido mezcla a los hombres sin estado ni sociedad civil, las cosas del mundo, las causas, las formas, las relaciones de atribución, y confunde a los sujetos. Regresamos justamente al principio de este libro.

Naturaleza y cultura, todas las ciencias confundidas

En la primera versión, la de la *physis*, la tierra sigue siendo la que el cultivador labora y siembra; en la otra, que se saca de Anaximandro, del *nomos* y de las enseñanzas más recientes, ella se convierte en el catastro, el plano, dibujado sobre papiro y destinado a la administración fiscal para permitirle calcular el monto del impuesto.

Hay aquí dos tierras, la negra de limo y la blanca o gris en el libro jeroglífico; la dura y la suave, material o logical; agrícola o estatal, nutricia o jurídica; física o formal; fisiológica o legislativa; inerte y viva, por una parte, colectiva y social por la otra; objeto que forma parte, primeramente, del mundo tal cual y, en ese sentido, sometido a las leyes físicas y naturales, y transformado por técnicas en las que los sólidos concretos obedecen a esas mismas leyes; objeto, en segundo término, de las leyes surgidas de diversos derechos, público, fiscal, administrativo; objeto, pues, único y doble, referido al mundo y al Estado, a las cosas y a los hombres, a dos tipos de leyes. Y, muy pronto tendremos que aprenderlo, a dos órdenes de ciencias. La tierra para los pies, la de los poderosos.

Todavía cultivadores, nuestros padres se referían a la primera, para pensar; viviendo solamente en megalópolis, convertida no hace mucho en exclusivamente política, la generación actual no piensa más que en la segunda, y en el poder.

Así pues, en el origen, una de esas dos tierras habría recubierto a la otra si los harpenodaptas hubieran diseñado el plano del catastro a la escala uno por uno, bella imagen de la utopía y de su semejanza imposible con lo verdadero. El Nilo sube: todos los campos de las dos riberas son cubiertos por un lago de superficie lisa, tan sedosa, que se diría, ya, la uniformidad graneada del papiro. Cuando el remo deja su rastro o su estela sobre el plano del agua, nada queda del frágil rastro de su escritura. Un plano pasa sobre los campos, uniforme, sin límite ni memoria.

Esto es justamente lo que estamos buscando: un recubrimiento; pero no podemos proyectar sobre el agua las propiedades de la tierra o sobre la ceguera de la crecida las distinciones de la decrecida. Ellas permanecen distintas, un poco como la sociedad de los hombres puede, a veces, dejar las presiones de la realidad física o como el anacoreta que ama los arroyos solitarios puede olvidar el ruido y la furia de los grupos.

¿Cómo las separamos, cómo podemos comprenderlas?

La versión derivada de las ciencias humanas, ¿expresa fielmente el relato de Herodoto? Tan escasamente como la de las ciencias de la naturaleza. ¿Es, pues, tan difícil leer nuestras leyendas? ¿Pero hemos leído verdaderamente el texto original? No.

Helo aquí.

El texto original u originario

En el capítulo 109 del segundo libro de las *Historias*, consagrado a la musa *Euterpe*, la protectora de las fiestas, podemos leer esto: *Sesostris, decían los sacerdotes, repartió la tierra entre todos los egipcios, atribuyendo a cada uno un lote igual a los demás, cuadrado; según esta repetición, estableció sus ingresos, prescribiendo que se pagara una cantidad anual. Si sucedía que la crecida despojaba a alguno de una parte de su lote, éste iba a verle y le señalaba lo sucedido; él enviaba a sus gentes para examinar y medir en cuánto había quedado disminuido el terreno, a fin de que posteriormente se fijara una disminución proporcional en el pago del impuesto fijado. Esto fue lo que dio lugar, a mi juicio, a la invención de la geometría, que los griegos llevaron a su país. Puesto que, en cuanto al uso del* polos*, del* gnomon *y de la división del día en doce partes, fue de los babilonios de quienes los griegos lo aprendieron.*

¿Sabemos que en aquellos tiempos la definición de los días separaba la puesta y la salida del Sol, de suerte que, según el invierno o el verano, cortos o largos, las horas variaban, puesto que se repartían, sobre el cuadrante solar, con ángulos cambiantes? Siempre doce, a pesar de todo, como una cuenta fija de cantidades todos los días variables. Lo habíamos olvidado, nosotros para quienes la jornada se compone de doce horas legales y estables, sin relación con el astro del día. ¿Por qué Herodoto acerca la métrica de los campos con la de las horas, la medida espacial de la tierra con la del cielo y del tiempo?

La sola aparición del verbo medir, y por lo tanto la de su operación, en ese capítulo de *Euterpe*, no se refiere a la repartición primera del valle agrícola en lotes, ni a la repartición del tiempo, del lado de Babilonia, sino a una suerte de catástrofe, fuera de la crecida del Nilo, del que todo Egipto es el don y que no interviene jamás en este relato, del que todas las interpretaciones conocidas la extraen, sin embargo, abusivamente. ¡Todas las teorías de la Tierra han necesitado muchos siglos, ellas también, para salir adelante del Diluvio!

En régimen regular sucede muy frecuentemente que, a lo largo de una curva o por un derrubio del río, crecido, no importa cuál río produce un efecto contrario a los aterramientos, o sea el hundimiento parcial o total de un campo aluvial. Todo el texto no hace sino hablar y no hablar de esta diferencia. A un campo mermado le falta tierra.

Perjudicado, el agricultor se desplaza y va a quejarse a las autoridades del accidente del que acaba de ser víctima; entonces, el rey despacha al lugar a su harpedonapta para medir en cuánto ha disminuido su terreno; diferencia, ciertamente, que se reduce a una sustracción, ya que el aterramiento o adición constituye o forma la tierra arable incluso a lo largo del valle. De regreso a su oficina, el funcionario calcula la disminución proporcional del impuesto fijado: *kata logan*.

Logos entre *physis* y *nomos*

He aquí la invención del *logos* o de la proporción entre la diferencia medida sobre el terreno y la que el funcionario calcula para la deducción fiscal: ésta es pues la escala que, sin paradoja, ocupa su lugar en la leyenda: cómo hay que leer el origen.

En la otra leyenda de origen, Tales mide la relación entre las longitudes de sombras y, de golpe, inventa la homotecia, es decir la escala; de igual manera, aquí, aparece y no aparece más que esta invención: el *logos* es esta escala misma que pone en relación la tierra agraria del fellah originario y el plano real del harpedonapta. Ella emerge en el momento mismo en que hacía falta.

Herodoto no habla de geometría para la medición de una parcela de trigo ni para el cálculo, en el catastro, del monto del impuesto, sino para la relación entre una disminución, constatada sobre el terreno por el campesino, y una cuenta, calculada por el alcabalero, sobre el plano catastral.

Él describe, pues, un invariante por variaciones, y compara esta estabilidad, entre el accidente físico y el pago fiscal con la de la cuenta civil babilonia de las horas por la variedad de su longitud real.

La misma relación se establece tanto sobre la tierra como en el cielo. El espacio de la geometría no reproduce en absoluto la primera ni imita al segundo, sino que adopta una especie de camino misterioso, de escala de Jacob, que pone en comunicación la naturaleza y la cultura, la tierra, negra, del campesino y la gris de la administración, la cosa y su representación, el campo y el plano, lo duro y lo blando, material y lógica, las ciencias físicas y las ciencias humanas, la generación que me precede y la mía, la primera interpretación y la segunda, la primera parte del libro y la que le sigue, Tales y Diógenes, Anaximandro físico y él mismo hambriento de justicia, *nomos* y *physis*.

Reales o supuestos, esos transportes, cuyo conjunto condiciona la medida, conducen del *pagus* a la página, del jardín al tribunal, del campo a la plaza pública, del trabajo a la discusión, del campo a la ciudad, de la víctima al rey o a su teniente e, inversamente, del pretorio al terreno o del contencioso contradictorio a la parcela de trigo, en fin, de la ley a la tierra y regreso.

Abren así, crecen y hacen fáciles todos los caminos concebibles cuyos bucles circundan este libro que acaba de describir su coronamiento. La facilitación de las rutas hace franquear el umbral de la percolación.

El plano de los reyes o del Estado imita a la gleba, bajo los pies, como una utopía se asemeja a otra que es, sin embargo, su opuesta: el fisicalismo se mueve sobre un ala de una quimera cuya otra ala lleva el exclusivismo de las ciencias humanas. La estabilidad descubierta aquí, lo real de la geometría, puentea esas dos utopías. El *logos* constituye la clave de ese puente.

¿El espacio abstracto nos permitiría habitar esta tierra en la que el área agraria se inserta en las leyes del Estado como la política se agita bajo el cielo físico?

Objeción

Hay aquí, sin embargo, una objeción que, invencible, nos reconduciría a las interpretaciones precedentes: es preciso que esa relación haya existido antes, para que el rey ordene, con anterioridad al accidente, repartir el valle en lotes o partes cuadradas, primeras, y por ello más originarias que ese cálculo de las diferencias.

Respuesta: Herodoto, justamente, no explica cómo hizo Sesostris la primera repartición, porque quiso, conscientemente, proclamar su origen a partir de la variación. Vagas, aproximativas, las primeras reparticiones no captaron la

medida precisa, exacta, rigurosa sino después o según esas faltas, mediante la relación y el *logos* entre las dos diferencias que, a su vez, pudieron precisar la posición y la forma de los lotes: la proporción precede a la porción; la preposición misma o prefijo lo dice. Antes de la primera, la segunda no accede ni a la justeza ni a la justicia.

La parte y el lote importan menos que la relación, y ésta no nace sino después de que aquéllos se hayan trastocado, por los derrubios del río que corre o por el correr de las horas, variables. El cielo y la tierra importan menos que sus variaciones, sus diferencias, sus fallas, menos de lo que nos es sustraído de la duración fluyente o de la fina tierra móvil, y que los esfuerzos humanos para compensarlo.

La superficie satinada del valle, la inundación plana, los cambios de los astros o del clima importan menos que el espacio en el que se desplazan los campesinos corriendo hacia la administración, los harpedonaptas viniendo a medir la esquina derrumbada, nuestros griegos trayendo de Babilonia el *polos* o el *gnomon*, y de Egipto la relación entre el hundimiento de la ribera y la deducción fiscal. El *logos* o la relación inventan este espacio de transporte que todo el mundo atraviesa permaneciendo invariable: he ahí el espacio puro de la geometría, sin obstáculo notable, donde todo fluye fácilmente, el río como las horas, la historia y los grandes relatos.

Logos

La porción importa menos que la proporción, la relación o transporte en donde el sustantivo –porte, de nuevo, importa menos que sus prefijos o que las preposiciones pro–, re–, ad– y trans–, es decir las relaciones, que pueden perfectamente permanecer estables gracias a la inestabilidad variable de la naturaleza y de los usos, de las cosas, de las causas, de las sustancias, de los sustantivos o de los verbos puestos en juego. Antes de que el harpedonapta o el campesino lo piensen, los dos compensan daños y pérdidas, desplazándose.

El *logos* no dice el ser sino la relación. La abstracción no se produce en absoluto a partir de la tierra, de la superficie unida de las aguas o de la pureza del cielo de los que el plano geométrico imitaría, posteriormente, las dimensiones, la plenitud, la transparencia y la luz, sino que nace a lo largo de transportes, siguiendo las relaciones que puentean y compensan sus variaciones. No se produce tampoco a partir del catastro, del plan escrito y dibujado sobre el papiro real del que la geometría imitaría, posteriormente, la exactitud y la justa precisión, sino que nace de transportes entre el campo y el libro jeroglífico fiscal.

La abstracción no tiene lugar en y por la posición fija o móvil de la tierra o del cielo, en y por la posición móvil o fija de la *Maat* escrita o dicha, sino que sigue la preposición, en general, antes de que se plantee lo que quiera que sea. El origen de la geometría se lee, a libro abierto, sobre el prefijo o la preposición que precede a la palabra misma preposición, sobre esa presencia misma que precede al acto de plantear, antes de la tesis o del ser ahí.

En particular, a lo largo de las relaciones, ausentes o inadvertidas, entre las preocupaciones que nos dan la tierra sobre la cual asentarnos nuestros pies o lo real duro que nos usa las manos o incluso ese río que nos quita el pan de la boca y las inquietudes que nos procuran los otros hombres, el poder, los impuestos, el trabajo, la servidumbre.

Sí, lo esencial sucede a lo largo de relaciones que las ciencias duras olvidan mantener con las ciencias humanas o de las que estas últimas no mantienen con las primeras; a tal punto olvidadas que las reencontramos en el origen, enterradas bajo el aluvión de las tierras, en el misterio de los jeroglíficos y el rechazo a leer, en la leyenda, la diferencia de esas dos ciencias y las relaciones que la compensan; bajo la inextricable, exquisita y transparente red de las preposiciones y los alivios de la declinación.

Sobre la medida, de nuevo

¿Qué es la geometría, de nuevo? Una cierta medida de la tierra.

En francés, como en las demás lenguas grecolatinas, la palabra medida designa exactitud, precisión y justeza en la relación que las cosas mismas mantienen con una regla dada, pero, al mismo tiempo, una moderación, completamente humana, que se espera ver aparecer en las disposiciones oficiales o jurídicas, y, puede ser, ante todo, una especie de medio, mitad, eje o centro por donde se perciben, por una parte, las ventajas y los desacuerdos, y por la otra, como un balance de una justicia. ¿La medida de la tierra originaria traduce ese primer temperamento arbitral? Justamente el texto de Herodoto refiere una rebaja derivada de una retención: el faraón acepta perder el impuesto cuyo equivalente el Nilo ha quitado al campesino; el harpedonapta arbitra midiendo, a fin de que el cultivador y el funcionario fiscal, juntamente, se pongan de acuerdo, moderados.

¿Quién dirá cuál es el consenso que ocurre primera-

mente, el concordato del pago establecido o el acuerdo concerniente a la proporción y cuál de ellos induce al otro? Se diría que la medida, en efecto, la medianería sigue la travesía de la naturaleza a la sociedad: las matemáticas parecen nacer en el medio mismo del paso del Noroeste.

Metis, madre de medida

Nosotros ya no nos acordamos del tiempo en que la gran separación de lo inteligible y de lo sensible no reinaba, imponente y todopoderosa, sobre nuestros espíritus. El divino Platón la extrajo de las idealidades matemáticas: sin ellas, antes de ellas, no había cielo abstracto, poblado de modelos cuyo rigor y belleza las cosas concretas, aquí abajo, imitarán mal y vagamente. Para poder pensar la era que precede a la geometría, hace falta al menos acordarse de actitudes que no rompen en absoluto esta cesura. Dicho de otra manera, ¿cómo pensamos cuando no pensamos obligatoriamente tales dos mundos?

¡Puesto que todo, siempre, se mezcla, tenemos que arreglárnoslas! Hacer pues, sin abstracción. Como los tallistas que proveyeron las piedras de las pirámides o las claves de bóveda de las grandes catedrales medievales, nosotros no siempre supimos, que yo sepa, de geometría en el espacio. Así que inventamos mil trucos, cien dispositivos, astucias, giros y finezas para salir del paso. Sin formalidades ideales, se puede ya pecar, cazar, poner trampas, montar un refugio, prensar las uvas, moler la harina, largar las velas, tratar de seducir a la vecina: ¡he ahí que tenemos ya más de tres cuartas partes de la vida!

Para expresar esta inteligencia vital de base, los griegos usaban, precisamente, *una palabra de la misma familia que medida: metis*, astucia fina que se desliza entre riesgos imposibles, impuestos por la fuerza de las cosas y el poder de los hombres y que pasa, desembarazadamente, entre dos escollos, el Caribdis natural de las turbulencias del Nilo y la Escila cultural de la sociedad, faraón y alcabalero. Sí, la astucia que permite salir del atolladero permite al débil, a veces, ganar la partida al fuerte, mandar —por ejemplo a la naturaleza— simulando obedecer, como suele hacerse ante los poderosos: Bacon seguirá, ciertamente, pero también precede a Tales y Platón.

Tercero excluido anterior a todo dualismo, abstracto y concreto juntamente, aunque la *metis* jamás haya escuchado hablar ni de lo inteligible ni de su imagen, sumergida en el *ápeiron*, sin exclusiva y mediana, *inventa la medida* de donde viene la geometría, que nos hace ver y cortar, mediante la exclusión, dos mundos: cesura segunda que nos hizo racionalistas.

La teogonía correspondiente

Al fin amo del Olimpo, después de haber matado a su padre, Zeus vivió en el terror de que un hijo, mañana, hiciera lo mismo con él; cuando Metis, su mujer, cae encinta por obra suya, él la devora; Atenea nacerá, se dice, de su cráneo, abierto con un hacha por Hefestos.

¿Conocemos alguna mejor manera de olvidar que incorporarse? Ya no recordamos nuestras andanzas ni los trucos múltiples urdidos en el mundo unido que precedió al nacimiento de Atenea, diosa de la razón. Antes de ser racionales, éramos inteligentes.

¿Se acuerda de ello nuestro cuerpo?

La cuestión del cambio

La única aparición del verbo medir, en el capítulo de Herodoto, interviene, pues, a propósito de cambios: el Nilo sube y baja, la tierra aumenta o disminuye, las horas varían, todo fluye y oscila, ¿cómo pensar semejantes variaciones?

Argumentos inatacables en apoyo de la invariancia, Parménides y Zenón velan sobre ese bello problema planteado en la aurora del helenismo por Heráclito, como si un diálogo inmenso hubiera opuesto dos voces, de una y otra parte del mar: la una del este, sobre el litoral de Jonia, en Éfeso, rutilante de fluctuaciones, y la otra, repitiendo la eternidad, de Elea, pequeño puerto de la Italia del sur, al oeste. ¿Podemos pensar lo variable, es decir, lo fluyente, mientras el ser es y el no-ser no es? El Nilo fluye, la tierra se inunda, la aurora cambia su ángulo, la moneda del impuesto deja ir su liquidez…

A esta primera pregunta planteada por la física, tres respuestas aparecen desde el alba griega.

Una teoría cualitativa mezcla los cuatro elementos para extraer lo cálido, lo frío, lo húmedo y lo seco: haciéndola suya, Aristóteles bloqueará al Occidente hasta Galileo.

La hipótesis combinatoria extrae toda evolución de las mezclas diversas de átomos permanentes: volvemos a encontrar la intuición genial de Demócrito, de Epicuro y de los abderitanos.

La tercera, cuantitativa, evalúa, en las mezclas, títulos o proporciones.

Estables en sus combinaciones, las tres respuestas divergen para los elementos.

En el texto de Herodoto, la única aparición del término *logos*, relación, sigue a la de medida de la tierra, para la disminución proporcional del impuesto fijado.

Mediante la mencionada proporción, los cambios del agua, de la tierra, del catastro, del impuesto, de la moneda debida, retornan de nuevo a lo fijo y a lo estable; invariancia en las variaciones, que armoniza Heráclito y Parménides. Es decir, la analogía o el *logos* en general resuelven la pregunta, ella misma general, del reposo y del movimiento, de la fluctuación y de la estabilidad.

Pero la responden transversalmente.

Analogía mejor que proporción

La proporción: he aquí la gran invención griega, que pasa, se desliza de una región a otra: *aritmética*, cuando dos o más fracciones se igualan; *geométrica*, por el teorema de Tales; casi *algebraica*, en tanto que las series de relaciones sirven a los matemáticos griegos, desde los orígenes hasta las fechas más tardías, como lengua universal para la demostración; *musical*, por los intervalos cifrados de la gama, sobre la que discutiremos todavía largo tiempo, sin duda, para decidir quién comenzó, desde el origen, si el pitagórico que evaluó la armonía sobre las cuerdas vibrantes o si aquél que contó fracciones previas para aplicarlas sobre ellas; *astronómica*, por la misma armonía de las esferas —¿quién cantará alabanzas suficientes para celebrar el talento de Eudosio, cuya hipopeda, dibujada tan temprano y casi a partir de nada, salvó la errancia aparentemente ondulada de los planetas?—, pero también por las relaciones contadas sobre las sombras del *gnomon* en pie; *cosmogónica*, por las mezclas determinadas de los elementos del Universo; *física*, por las proporciones definidas por todas partes en relación con los estados físicos primeros de la materia, tierra, agua, aire y fuego, según Empédocles; *química* incluso, por las mismas proporciones reguladas de todas las cosas en todas las cosas, según Anaxágoras de Clazomene; y *médica*, en el *corpus* hipocrático…

Nosotros no sabremos nunca nada del progreso, noción demasiado global y vaga; pero podemos reducir la cuestión a intervalos pequeños: ¿cuántos resultados fueron obtenidos en cuánto tiempo por cuántas personas? Aquí, en menos de cien años, un puñado de hombres estableció toda la ciencia, casi unitariamente; extraordinaria explosión vertical que no se reproducirá sino pocas veces en la historia occidental.

Administrando, a lo largo de su historia, según rendimientos decrecientes, ese capital tan pronto acumulado, la mayoría de los sabios, perezosos, temen, a partir de entonces, el regreso de esos truenos, en los que un aumento superabundante sobreviene en sorprendentes cabezas sobrevivientes, en breves y soberbios momentos. Si hablamos de milagro, ciertamente está aquí.

Pero también, en lengua latina, los términos proporción y fracción nos exponen a no comprender lo que nace aquí. Los griegos no conocían la relación simple a sobre b; solamente la *analogía*, *a* sobre *b* igual a *b* sobre *c*, les interesa, ya que gracias a ella establecen uno o varios términos proporcionales, medios.

He ahí, justamente, el *logos*, la media o mediana proporcional, que va de una relación a otra y, por sustitución, corre, aún, de ésa a una tercera, y así sucesivamente.

Transporte local, término a término, como desde la gleba al plano; transporte global, de ciencia a ciencia.

Sí, he aquí el gran invento griego: la ana-logía, el *logos* que transita, pasa de abajo arriba y de arriba abajo, *kata logon*, la palabra que se cuela y pasa, deambula y se cambia, y que sin embargo no pasa, puesto que todo se evalúa y se mide gracias a su transporte, mensaje fijo de Hermes volante. No, no se trata en absoluto de dividir alguna cosa en partes, es decir compartir o extraer lo que cada uno, generoso o leonino, sabe hacer desde que el mundo yace bajo la luz del Sol y la ferocidad de la guerra, sino de construir, paso a paso, una cadena, o sea, hallar aquello que, estable, transita a lo largo de su encadenamiento.

Logos ana o *kata*… sentido o signo, forma o llamado, descubrimiento, quién sabe, la palabra importa menos aquí que su desplazamiento, que el espacio de su movimiento, que su deslizamiento, que las preposiciones que la acompañan y muestran, indexan, rigen, demuestran su paso, marcan la sintaxis o el ordenamiento, el cómo de su encadenamiento, de su flujo controlado: esas largas cadenas de razón, tan simples y fáciles…

He aquí inventado el primer lenguaje de ciencia, sí, ahí está, la auténtica invención, el descubrimiento o el desbloqueo del elemento fluyente, estable y deslizante, de ese discurso interminable cuyo curso se pone a fluir infinitamente desde el momento en que atraviesa, así, el umbral de la percolación. La logos-relación engendra el logos-discurso, por encadenamiento del logos-palabra. Así comienza la génesis del gran relato de ciencia.

La lección de las dos leyendas

De la naturaleza a las costumbres, del trueque al cambio regido por el dinero, de lo político a lo religioso, de Aristóteles, que trata de la justicia distributiva, al *Libro X* de Euclides, la analogía se desliza: de las ciencias duras a las blandas e inversamente.

Si separamos los dos tipos de leyendas, la del mundo y la de los hombres, leemos la utopía de los sociólogos o de los políticos cuya visión y su vida desdeñan el mundo exterior de las cosas tal cual son y la utopía simétrica de los sabios que se dicen realistas, pero cuyos ojos y su acción menosprecian las relaciones entre los hombres. En la primera isla o sobre una de las alas de la quimera, todo no es más que político, social o humano; en la segunda ala o isla, los objetos bastan. Pero las cosas se vengan de los hombres, que, entre ellos, se vengan de las cosas, que, sin decir nada…

Reunión, intersección

Suponiendo que supiéramos reunir, en un sentido concertado, las dos utopías, nosotros contemplamos o producimos, ¡oh sorpresa!, lo concreto en su plenitud, a saber, las sociedades, ciudades y campos, el mundo de las montañas y las planicies, los marinos en el mar, los artesanos provistos de sus útiles, la totalidad densa de lo real y no solamente de las palabras: paisaje cuyo raro esplendor hace latir el corazón del filósofo e inspira al religioso.

Pero si esta reunión constituye un milagro en la teoría, cuando vivimos cotidianamente sumergidos en ella, no sabemos verdaderamente si su intersección, siempre en sentido concertado, existe y, si es así, si ella está llena o vacía.

Si la suponemos llena, entonces, se puebla de seres existentes de estatutos objetivo y colectivo a la vez, de aquellos que en otra ocasión llamé casi-objetos, trazadores objetales de relaciones intersubjetivas en el grupo.

Si la suponemos vacía, entonces cualquier dimensión que se le suponga, desplazándose sobre ella o en ella, debe llegar, en algún momento, al borde del colectivo o a las orillas de lo objetivo, como si ella representase un papel en el proceso de comienzo.

El conjunto de las leyendas que relatan su instauración hablan constantemente de esas dos vías principales: los harpenodaptas egipcios reparan los daños ocasionados por la crecida del Nilo, he ahí el mundo, y aseguran el plano del catastro por el cálculo del monto del impuesto, he ahí la política de los hombres; al pie de las tres pirámides, Tales observa los rayos del Sol y estructura nuestra visión de su teorema, he ahí el mundo y he aquí los hombres, se burla de los poderes del faraón, tallados en las piedras de la tumba, así como Diógenes exigía a Alejandro que se quitara de su Sol; esta doble y dudosa balanza, ¿expresa el equilibrio de los cuerpos pesados o la justicia distributiva? Tal texto inicial de Euclides, ¿habla, igualmente, de la estabilidad de las cosas pesadas o del acuerdo entre participantes? ¿Hace tal diálogo oír un ruido objetivo o el desacuerdo colectivo?

El problema de los múltiples orígenes de las formas matemáticas, el desciframiento de las leyendas que los relatan, se reúnen en el espacio abierto o cerrado por esta conjunción o esta disposición de separación o de coordinación que diseñan y describen, ambas, la intersección, vacía o llena, de esas dos utopías. Ahí está el lugar de la fuente y de su surgimiento.

Objetivo-colectivo

¿Cómo llamar a la intersección vacía entre dos utopías? Lo abstracto. ¿Cómo llamar al casi-objeto arrojado en esta abstracción? El objeto matemático.

¿Por qué? Porque primeramente realiza, sobre él, el acuerdo completo de la comunidad, cualquiera que ésta sea, contrato único jamás asumido por los hombres; porque, en segundo lugar, se aplica a placer sobre los objetos del mundo tal cual, es decir, libres de toda intervención colectiva. Modelo objetivo perfecto, como jamás se encuentra otro igual; trazador excelente de una red sin ruido, como jamás se vio uno semejante. En suma, objeto paradójico, excepcional en cada punto, pero sobre todo por esa asociación.

No sabemos si existe una intersección entre lo objetivo y lo colectivo, pero si es que existe, llena o vacía, ella se llena con esos objetos ausentes que llamamos las idealidades matemáticas. La ciencia que nosotros calificamos así, no es, en efecto, ya lo sabemos, una ciencia social, ni, tampoco, como igualmente sabemos, una ciencia de las cosas del mundo; ni una política, ni una sociología, ni una física, ni una biología… puede ser incluso que no sepamos cómo definirla; por eso es que a veces la referimos a un cielo ideal en los extremos de lo real o a un conocimiento trascendental, íntimo de lo íntimo, en los límites de los dos espacios de la utopía, doble cuerpo de un unicornio estéril.

Por más que la matemática no sea ni la una ni la otra, es sin embargo tanto una como la otra, puesto que se aplica igual de bien a las cosas del mundo que nadie puede conocer sin ella y puesto que realza tan bien el acuerdo universal entre los hombres que no conocemos ningún otro ejemplo de acuerdo tan perfecto ni de universalidad tan completa y saturada.

El acuerdo colectivo se funda en su necesidad objeti-

va, al mismo tiempo que esta necesidad se funda sobre ese acuerdo: semejante simultaneidad no se encuentra más que ahí.

Lugar-fuente

El espacio geométrico o la cuenta aritmética o el proceso algorítmico paso a paso… nacen los tres de esta intersección que durante largo tiempo se creyó vacía y nula, y de la que yo creo y sé que es, sin embargo, el verdadero mundo, paradisíaco, real, rico y completo, la densa realidad, de la que los dos componentes, natural y cultural, no son más que dos sustracciones utópicas, flacas y pobres.

Eso explica superabundantemente por qué las matemáticas dan a quienes las aman, las practican, las utilizan o, mejor aún, a quienes las inventan, la certidumbre inmediata y experimental, sí, vívida, de la presencia inevitable de un cuerno de la abundacia de donde se saca siempre todo de nada. Ciertamente, todo se encuentra ahí, ¡pero no tenemos ojos para verlo!

He aquí, en el centro de los dos cuerpos, el punto ciego, ¡he aquí el fondo del cuerno de la abundancia!

¿Acaso alguna vez hemos observado verdaderamente hasta qué punto esta ciencia, tan comúnmente compartida por todas las demás, permanece única, rara y paradójica, hasta los últimos límites? Fuera del mundo y en el mundo, inmanente y trascendente, sin presencia humana y sin embargo universal en las relaciones colectivas.

Espacios sin objeto en donde todo objeto, cualquiera que sea, se sitúa o se mide; espacios sin ojo en donde toda escena óptica se aclara y se ordena; espacios vacíos de hombre en donde las relaciones sociales elementales se canonizan y aplacan, como en un contrato excelente, en donde, por ejemplo, beneficio o pérdida igual venta menos compra, en donde las reparticiones se organizan, en donde se equilibran los intercambios, en donde se calculan los impuestos y los tributos, en donde la ecuación garantiza la equidad; se diría que el propio Hermes, dios de la suerte, pasa por ahí. Sin ningún objeto, con todo objeto; sin ninguna relación, desprendida o abstracta, definiendo y comprendiendo toda relación; ciencia, pues, no objetiva y totalmente objetiva; toda relacional y no relacional.

La matemática es, por lo tanto, a tal punto objetiva que sólo ella es verdaderamente colectiva; a tal punto colectiva que sólo ella es verdaderamente objetiva; a tal punto inútil que sólo ella es verdaderamente útil; a tal punto exterior que sólo ella es verdaderamente interior; a tal punto interior que sólo ella es verdaderamente exterior; a tal punto dentro del ser que sobresale en el conocimiento; a tal punto en el conocimiento que sobresale en el ser; a tal punto abstracta que sólo ella es verdaderamente concreta, tan concreta incluso que, a veces, creímos que su espacio era la forma del sentido externo…

…a tal punto concreta, por último, que sólo ella es verdaderamente abstracta: el nacimiento de su abstracción, como yo lo demuestro, se desprende de la suma integral de lo real más concreto que ella atraviesa.

Eminentemente objeto, ella absorbe todos los objetos; sujeto colectivo, eminentemente, ella piensa completamente sola, de suerte que nosotros nos hemos convertido en sus guardianes y sus sirvientes. Desde su nacimiento, voluntariamente o no, vivimos y pensamos en ella y por ella.

¿Qué es pues, para terminar, un objeto matemático? Un casi-objeto límite y excelente.

La tierra y la Tierra

Henos aquí al cabo de los viajes de Tales o de Demócrito, del campesino egipcio y del harpedonapta real…

¿Quién, de hecho, se transporta? Hermes, traducción griega de Tot, dios de Egipto, éste es su doble nombre; pasando y volando ambos, conectan los lugares separados y organizan así los espacios lisos. La homogeneidad del espacio viene de la suma de esos transportes.

El *gnomon* que sale de la tierra la vincula al cielo y a la luz. Así el espacio puro de la geometría suma primero el cielo y la tierra, físicos, pero también el *templum* y el *pagus*, el ágora y el pretorio, y, de golpe une la tierra al Estado, y el mercado a las listas de tarifas llenadas por los escribas y por los ministros.

Una Tierra unitaria aparece entonces, astronómica, natural, real, habitada, cultivada, regulada por las leyes de los dioses y de los reyes, tierra espesa, agraria, pragmática, geográfica, religiosa, política, jurídica y conocedora al mismo tiempo, en donde la geometría nivela el área, y en donde la abstracción y la pureza suman o hacen la sinopsis y la síntesis de esta realidad común y plena.

La conexión súbita entre estas especificaciones, los transportes fulminantes que las asociaron crearon una caja blanca, generalizando las líneas semejantes al segmento de Zenón o los cuerpos regulares como el tetraedro; en suma, un recipiente tal que la fuente apareció y nosotros superamos el umbral de la percolación.

Desde el blanqueamiento de este espacio por el paso y las conexiones de Hermes, todo fluye. Incluso nuestra historia.

Nuestro hábitat

Este espacio blanco, la Grecia lo habitó e hizo que nosotros no cesemos de habitarlo después de ella, como nuestro propio territorio. La geometría integra todos nuestros hábitat prácticos o ideales, como la luz blanca suma todos los colores, en la transparencia o la traslucidez.

Recuerdo haberme aproximado, un día, humildemente, al tetraedro puro y transparente de la Geometría en el espacio para aguardar, durante casi medio siglo, que un sol nuevo se levantase, detrás de ese prisma; y proyectase, sobre la arena, ante mí, niño ignorante atravesando el desierto, deslumbrado como Tales o Diógenes en sus tiempos, el abanico completo de numerosos lazos diversos, como los matices que componen una invisibilidad, distribuidos como en el lenguaje se distinguen el templo, el área agraria, el campamento, la ciudad, la utopía, la escena del teatro, el pretorio, la página… elementos de nuestro antiguo hábitat, y para esperar, durante más de cuatro decenios, comprender quién habita allí y cómo reside, en la violencia o la paz, por o contra la exclusión. Yo no comprendí nuestro antiguo hogar sino en el momento en que capté que el poliedro puro y traslúcido era el *gnomon*, cuyo espacio *comprendía*.

Sí, su abstracción es una suma y no una sustracción. Sin esta síntesis blanca del espacio de todos los pasajes, habría que recurrir a un perpetuo milagro para no comprender por qué las matemáticas, en general y, especialmente, este espacio de la geometría se aplican, universalmente, a los hombres y a las cosas del mundo, sin excepción.

Nuestra tierra de luz, casa medida, integra el conjunto de esos hábitat.

Nosotros habitamos desde entonces este espacio como una casa o, mejor aún, como nuestra tierra: *el metro es la Tierra*, he aquí el sentido profundo del vocablo *geometría*. Nosotros no tenemos la menor idea ni percepción de una tierra sin ella, antes de ella o privada de su extensión cuya transparencia homogénea nos baña y atraviesa nuestro cuerpo, acostado o erguido, extendiendo su envergadura, con su triple flecha ancha, larga y alta, tan universal que el Universo se sume en ella enteramente. ¡Tanto la aculturación griega nos informó y naturalizó así el mundo, que el poco perspicaz Kant tomó el espacio así purificado por la forma de nuestro sentido externo!

Sí, las cosas del mundo y nuestros cuerpos se volvieron entonces euclidianos y se anclaron a tal punto en esta tierra paradójica, extraña por isótropa y traslúcida, que resultó muy difícil, incluso hoy día, mostrar a los filósofos que nuestros sentidos, a veces, se hunden en espacios totalmente distintos, topológicos o proyectivos, caóticos o fractales, tan fuerte sigue siendo su creencia de que el espacio surgido de la geometría antigua sigue siendo nuestra única tierra, mientras que la Tierra, arcaica y nueva, se construye globalmente en otra parte, sin su mirada ciega.

Esperanza

En el curso del siglo xx despegamos poco a poco del espacio de la tierra que habitamos desde hace tres milenios, de suerte que desaparece, a nuestros ojos, poco a poco, el de la luz solar, de la agricultura, de lo sagrado, de la guerra, de los estados, de la página de escritura, que la geometría expresaba, juntamente, en su pureza sumadora.

Circulan, en masa desde entonces, cuerpos, mensajes, informaciones, conocimientos, la luz en su velocidad más que en su claridad: un espacio nuevo de transportes nuevos se instala, sobre una Tierra global, menos pura que mezclada, menos lisa u homogénea que desnuda, arlequinada, atigrada, acebrada, en redes múltiples y conexas.

Puede ser que abandonemos las conexiones simples que Hermes anudaba con su caduceo, para volver a ganar los transportes de las legiones abigarradas de miríadas de arcángeles, por la ubicuidad de los mensajes. La antigua ciencia hablaba de tablas y de causas, la nueva circula en las computadoras y en los escenarios de lo posible.

A través de esas redes percolantes nuevas, una nueva ciencia y otro hábitat, una ciudad nueva, un nuevo Universo se preparan, y por las mismas razones que las que este libro evoca del antiguo saber y de la antigua casa, cuyo nacimiento ocupó las ciudades y las islas del *Logos* escrito por Tales, Heráclito, Eudosio, Herodoto o San Juan Evangelista…

Aguardo a la aurora, mañana, el pasaje, en esta red, del umbral de percolación.

Un flujo circulará: nuevos discursos interminables, otros grandes relatos.

Aubière, mayo de 1958
Kyoto, noviembre de 1992

PRINCIPIOS MATEMÁTICOS DE LA FILOSOFÍA NATURAL

ISAAC NEWTON

[Traducción de Siglo XXI]

DEFINICIONES

Definición I

La cantidad de materia es la medida de la misma que se desprende de su densidad y su masa, conjuntamente.

De esta forma, el aire de doble densidad, en un espacio doble, es el cuádruple en cantidad; en un espacio triple, el séxtuple en cantidad. Lo mismo debe entenderse de la nieve, y de los polvos finos que se condensan por compresión o licuefacción, y de todos los cuerpos que se condensan por cualesquiera otras causas. No considero aquí un medio, en caso de haberlo, que penetre libremente en los intersticios entre las partes de los cuerpos. A esta cantidad es a lo que me refiero en lo sucesivo con el nombre de cuerpo o masa. Y lo mismo se conoce como el peso de cada cuerpo, porque es proporcional al peso, como he descubierto experimentando con péndulos elaborados con gran precisión, como se demostrará después.

Definición II

La cantidad de movimiento es la medida del mismo que se desprende de la velocidad y cantidad de la materia, conjuntamente.

El movimiento del todo es la suma de los movimientos de todas las partes; y por lo tanto, en un cuerpo del doble de cantidad, a igual velocidad, el movimiento es doble; con el doble de velocidad, es cuádruple.

Definición III

La vis insita*, o fuerza intrínseca de la materia, es un poder de resistencia, por el cual todo cuerpo continúa en su estado actual, sea éste de reposo o de movimiento uniforme hacia adelante en línea recta.*

Esta fuerza siempre es proporcional al cuerpo cuya fuerza es, y no difiere de la inactividad de la masa más que por nuestra manera de concebirla. Un cuerpo, por la naturaleza inerte de la materia, no cambia sin dificultad de su estado de reposo o de movimiento. En ese sentido, esta *vis insita* podría ser denominada, con un nombre mucho más significativo, inercia (*vis inertia*), o fuerza de inactividad. Pero un cuerpo sólo ejerce esta fuerza cuando otra fuerza impresa sobre él procura modificar su condición; y el ejercicio de esta fuerza puede considerarse a un tiempo como resistencia y como impulso; es resistencia en la medida en que el cuerpo, para mantener su estado actual, se opone a la fuerza impresa; es impulso en la medida en que el cuerpo, al no ceder fácilmente a la fuerza impresa por otro, procura modificar el estado de ese otro. Habitualmente se la describe como la resistencia de los cuerpos en reposo, y el impulso de los que están en movimiento; pero movimiento y reposo, tal como se los concibe comúnmente, sólo se distinguen en términos relativos; tampoco están siempre verdaderamente en reposo los cuerpos que se toman como tales.

Definición IV

Una fuerza impresa es una acción ejercida sobre un cuerpo a fin de cambiar su estado, ya sea de reposo o de movimiento uniforme en línea recta.

Esta fuerza consiste sólo en la acción, y no perdura en el cuerpo una vez que la misma concluye. Porque un cuerpo mantiene todos los nuevos estados que va adquiriendo, tan sólo por su inercia. Pero las fuerzas impresas son de orígenes diferentes, como por percusión, por presión, por fuerza centrípeta.

Definición V

Una fuerza centrípeta es aquella por la cual los cuerpos son atraídos o impelidos, o tienden de cualquier manera hacia un punto central.

De este tipo son la gravedad, por la cual los cuerpos tienden al centro de la Tierra; el magnetismo, por el cual el hierro tiende a la piedra magnética, y esa fuerza, cualquiera que sea, por la cual los planetas son desviados continuamente de los movimientos rectilíneos que seguirían de otra manera, y hechos girar en órbitas curvilíneas. Una piedra que se hace girar en una honda procura retroceder de la mano que le da vueltas; y al hacerlo distiende la honda con tanta mayor fuerza cuanto mayor es la velocidad con la cual se la hace girar, y tan pronto como se la suelta se aleja volando. Esa fuerza que se opone a este propósito, y por la cual la honda continuamente vuelve a atraer la piedra hacia la mano, y la mantiene en su órbita, porque es dirigida hacia la mano como si fuese el centro de la órbita, la denomino fuerza centrípeta. Y lo mismo debe entenderse de todos los cuerpos que giran en cualquier órbita. Todos pretenden retroceder de los centros de sus órbitas; y de no ser por la oposición de una fuerza contraria que se los impide, y que los detiene en sus órbitas, que por lo tanto denomino centrípeta, saldrían volando en línea recta con movimiento uniforme. De no ser por la fuerza de gravedad, un proyectil no se desviaría hacia la Tierra sino que proseguiría alejándose de ella en línea recta, con movimiento uniforme, si se suprimiese la resistencia del aire. Por la gravedad es continuamente apartado de su curso rectilíneo y hecho desviarse hacia la Tierra, en mayor o menor medida, de acuerdo con la fuerza de su gravedad y la velocidad de su movimiento. Cuanto menor es su gravedad, o la cantidad de su materia, o mayor la velocidad con la que se lo proyecta, menos se desviará de un curso rectilíneo y más lejos llegará. Si una bola de plomo proyectada desde la cima de una montaña por la fuerza de la pólvora, con una velocidad dada y en una dirección paralela al horizonte, es llevada en una línea curva a la distancia de dos millas antes de caer al suelo, la misma, si se eliminase la resistencia del aire, con una velocidad doble o décuple, volaría dos o diez veces más lejos. Y al aumentar la velocidad podemos incrementar a placer la distancia a la cual es posible proyectarla y disminuir la curvatura de la línea que puede describir, hasta que finalmente caiga a una distancia de 10, 30 o 90°, o incluso pueda darle toda la vuelta a la Tierra antes de caer; o finalmente no caer nunca a Tierra, sino adentrarse en los espacios celestiales y proseguir en ese movimiento *in infinitum*. Y de la misma manera que un proyectil, por la fuerza de gravedad, puede hacerse girar en una órbita, y darle la vuelta a toda la Tierra, también la Luna, ya sea por la fuerza de gravedad, si es que está dotada de gravedad, o por cualquier otra, que se impele hacia la Tierra, puede ser desviada continuamente del camino rectilíneo que seguiría por su fuerza intrínseca, y hacérsela girar en la órbita que describe ahora. Y sin alguna fuerza de este tipo la Luna no podría ser mantenida en su órbita. Si esta fuerza fuese demasiado pequeña no alcanzaría para apartar a la Luna de un curso rectilíneo; si fuese demasiado grande, la desviaría en exceso y atraería a la Luna desde su órbita hacia la Tierra. Es necesario que la fuerza sea de la cantidad precisa, y es labor de los matemáticos encontrar la fuerza que pueda servir exactamente para mantener a un cuerpo en una órbita dada a una velocidad dada; y, *viceversa*, determinar el camino curvilíneo al cual es posible hacer que se desvíe un cuerpo proyectado desde un lugar dado, a una velocidad dada, de su senda rectilínea natural, por medio de una fuerza dada.

Puede considerarse la cantidad de cualquier fuerza centrípeta como de uno de estos tres tipos: absoluta, acelerativa y motriz.

Definición VI

La cantidad absoluta de una fuerza centrípeta es la medida de la misma, proporcional a la eficacia de la causa que la propaga desde su centro por el espacio que la rodea.

Así, la fuerza magnética es mayor en una piedra magnética y menor en otra, de acuerdo con sus tamaños y la fuerza de su intensidad.

Definición VII

La cantidad acelerativa de una fuerza centrípeta es la medida de la misma, proporcional a la velocidad que genera en un tiempo dado.

Así, la fuerza de la misma piedra magnética es mayor a menor distancia y menor a mayor distancia; también la fuerza de gravedad es mayor en los valles, menor en la cumbre de las montañas extremadamente altas, y menor aún (como se demostrará adelante) a distancias mayores del cuerpo de la Tierra; pero a distancias iguales es la misma por doquier; porque (eliminando o tomando en consideración la resistencia del aire) acelera por igual todos los cuerpos que caen, ya sean pesados o ligeros, grandes o pequeños.

Definición VIII

La cantidad motriz de una fuerza centrípeta es la medida de la misma, proporcional al movimiento que genera en un tiempo dado.

Así, el peso es mayor en un cuerpo más grande, menor en uno más pequeño; y, en el mismo cuerpo, es mayor cerca de la Tierra y menor a distancias más remotas. Esta clase de cantidad es la centripeticidad como propensión de todo el cuerpo hacia el centro o, podría decir, su peso; y siempre se conoce por la cantidad de una fuerza igual o contraria precisamente suficiente como para impedir el descenso del cuerpo.

En pro de la brevedad, a estas cantidades de fuerzas podemos aplicarles los nombres de motriz, acelerativa y absoluta; y, para distinguirlas, considerarlas con respecto a los cuerpos que tienden hacia el centro, a los lugares de sus cuerpos y a la fuerza del centro hacia el cual tienden; es decir, me refiero a la fuerza motriz de un cuerpo como un propósito y propensión del todo hacia un centro, que surge de las propensiones de las diversas partes tomadas en su conjunto; a la fuerza acelerativa al lugar del cuerpo, como cierto poder que se difunde desde el centro hacia todos los lugares a su alrededor para mover los cuerpos que están en ellos; y a la fuerza absoluta hacia el centro como dotada de alguna causa sin la cual esas fuerzas motrices no se propagarían por los espacios que las rodean. Ya sea que esa causa sea un cuerpo central (como el imán que está en el centro de la fuerza magnética, como la Tierra en el centro de la fuerza gravitacional) o cualquier otra cosa que todavía no aparece. Porque aquí sólo deseo dar una noción matemática de esas fuerzas, sin tomar en consideración sus causas y sedes físicas.

Por consiguiente, la fuerza acelerativa guardará con la motriz la misma relación que la celeridad con el movimiento. Porque la cantidad de movimiento emana de la celeridad multiplicada por la cantidad de la materia; y la fuerza motriz emana de la fuerza acelerativa multiplicada por la misma cantidad de materia. Porque la suma de las acciones de la fuerza acelerativa sobre las diversas partículas del cuerpo es la fuerza motriz del todo. A eso se debe que cerca de la superficie de la Tierra, donde la gravedad acelerativa, o fuerza productora de gravedad, es la misma en todos los cuerpos, la gravedad motriz o el peso es como el cuerpo; pero si ascendiésemos a regiones más elevadas, donde la gravedad acelerativa fuese menor, el peso disminuiría de igual manera, y siempre sería igual al producto del cuerpo por la gravedad acelerativa. De manera que en aquellas regiones en las que la gravedad acelerativa disminuye a la mitad, el peso de un cuerpo dos o tres veces menor será de cuatro o seis veces menos.

Asimismo denomino atracciones e impulsos, en el mismo sentido, a los acelerativos y motrices, y utilizo las palabras atracción, impulso o propensión de cualquier índole hacia un centro, de manera indistinta e indiferente entre sí; considerando estas fuerzas no física sino matemáticamente, por lo cual el lector no debe imaginar que con esos términos he decidido de manera alguna definir la clase o forma de una acción, las causas por razón física de la misma, ni que atribuyo fuerzas, en un sentido verdadero y físico, a ciertos centros (que sólo son puntos matemáticos) cuando en cualquier momento hablo de centros que atraen o que están dotados de poderes de atracción.

Escolio

Hasta ahora he asentado las definiciones de los términos menos conocidos, y explicado en qué sentido han de ser entendidos en el discurso que sigue. No defino tiempo, espacio, lugar y movimiento, puesto que son bien conocidos para todos. Sólo debo observar que las personas comunes conciben esas cantidades de acuerdo con la relación que guardan con los objetos sensibles. Y de ello se desprenden ciertos prejuicios, para eliminar los cuales será conveniente distinguirlos en absolutos y relativos, verdaderos y aparentes, matemáticos y comunes.

I. El tiempo absoluto, verdadero y matemático, en sí mismo y por su propia naturaleza, fluye de manera uniforme sin relación con nada externo, y por otro nombre se denomina duración; el tiempo relativo, aparente y común es cierta medida de duración sensible y externa (sea

precisa o desigual) por medio del movimiento, que se usa habitualmente en lugar del tiempo verdadero, como una hora, un día, un mes, un año.

II. El espacio absoluto, por su propia naturaleza, sin relación con nada externo, permanece siempre similar e inmóvil. El espacio relativo es alguna dimensión o medida movible de los espacios absolutos, que nuestros sentidos determinan por su posición en relación con los cuerpos, y que usualmente se toma por un espacio inmóvil; tal es la dimensión de un espacio subterráneo, aéreo o celeste, determinado por su posición con respecto a la Tierra; el espacio absoluto y el relativo son iguales en figura y en magnitud, pero no siempre siguen siendo numéricamente iguales. Porque si la Tierra, por ejemplo, se mueve, un espacio de nuestro aire, que relativamente y en relación con la Tierra permanece siempre igual, será en un momento parte del espacio absoluto por el cual pasa el aire; en otro momento será otra parte del mismo y así, entendido de manera absoluta, cambiará continuamente.

III. El lugar es una parte del espacio que ocupa un cuerpo, y es acorde con el espacio, ya sea absoluto o relativo. Quiero decir, como parte del espacio, no en la situación ni en la superficie externa del cuerpo. Porque los lugares de los sólidos iguales son siempre iguales, pero muchas veces sus superficies, debido a sus figuras disímiles, son desiguales. En sentido estricto las posiciones no tienen cantidad, y no son tanto los lugares mismos cuanto las propiedades de ellos. El movimiento del todo es el mismo que la suma de los movimientos de las partes; es decir, la traslación del todo de su lugar es lo mismo que la suma de las traslaciones de las partes de sus respectivos lugares, y por consiguiente el lugar del todo es lo mismo que la suma de los lugares de sus partes, y por esa razón es interno, y está en todo el cuerpo.

IV. El movimiento absoluto es la traslación de un cuerpo de un lugar absoluto a otro, y el movimiento relativo es la traslación de un lugar relativo a otro. Así, en un barco de vela, el lugar relativo de un cuerpo es esa parte del barco que posee el cuerpo; o esa parte de la cavidad que el cuerpo llena, y que por lo tanto se mueve junto con el barco; y el reposo relativo es la permanencia del cuerpo en la misma parte del barco o de su cavidad. Pero el reposo real, absoluto, es la permanencia del cuerpo en la misma parte de ese espacio inamovible en el cual se mueve en el barco mismo, su cavidad, y todo lo que contiene. Por consiguiente, si la Tierra realmente está en reposo, el cuerpo, que descansa relativamente en el barco, se moverá real y absolutamente con la misma velocidad que tiene el barco sobre la Tierra. Pero si ésta también se mueve surgirá el movimiento verdadero y absoluto del cuerpo, parcialmente del movimiento verdadero de la Tierra en un espacio inmóvil, y parcialmente del movimiento relativo del barco sobre la Tierra; y si el cuerpo se mueve también relativamente en el barco, su movimiento verdadero se deberá en parte al movimiento verdadero de la Tierra en espacio inamovible, y en parte a los movimientos relativos, tanto del barco en la Tierra como del cuerpo en el barco; y de estos movimientos relativos surgirá el movimiento relativo del cuerpo en la Tierra. Como si esa parte del mundo en la que está el barco se moviese verdaderamente hacia el este a una velocidad de 10 010 partes, mientras que el barco mismo, con un nuevo vendaval y todas las velas desplegadas se traslada hacia el oeste, con una velocidad expresada por 10 de esas partes; pero un marinero camina por el barco hacia el este con 1 parte de dicha velocidad; el marinero, entonces, se moverá verdaderamente en el espacio inmóvil hacia el este, con una velocidad de 10 001 partes, y relativamente sobre la Tierra hacia el oeste, con una velocidad de 9 de esas partes.

En astronomía se distingue el tiempo absoluto del relativo por la igualdad con la corrección del tiempo aparente. Porque los días naturales son verdaderamente desiguales aunque comúnmente se consideran iguales, y se usan como medida del tiempo; los astrónomos corrigen esta desigualdad para poder medir los movimientos celestes con un tiempo más preciso. Puede ser que no exista un movimiento uniforme gracias al cual el tiempo pueda medirse con precisión. Puede que todos los movimientos estén acelerados y retrasados, pero el flujo del tiempo absoluto no es pasible de cambio alguno. La duración de la perseverancia de la existencia de las cosas sigue siendo la misma, sin importar si los movimientos son rápidos o lentos, o si no los hay, y por lo tanto esta duración debería diferenciarse de lo que no son más que mediciones sensibles de la misma, y de la cual podemos distinguirla por medio de la ecuación astronómica. La necesidad de esta ecuación para determinar los tiempos de un fenómeno se manifiesta por igual mediante los experimentos del reloj de péndulo que por los eclipses de los satélites de Júpiter.

Así como el orden de las partes del tiempo es inmutable, lo es también el orden de las partes. Supongamos que esas partes se moviesen de sus lugares; pues se moverían, si se me permite la expresión, fuera de sí mismas. Porque los tiempos y los espacios son, por así decirlo, tanto los espacios de sí mismos como de todas las demás cosas. Todas las cosas se ubican en el tiempo en orden de

sucesión, y en el espacio en orden de situación. Hay lugares a partir de su esencia o naturaleza; y resulta absurdo que los lugares primarios de las cosas pudiesen ser movibles. Éstos, por consiguiente, son los lugares absolutos, y las traslaciones a partir de ellos son los únicos movimientos absolutos.

Pero como las partes del espacio no pueden verse ni distinguirse una de otra por medio de nuestros sentidos, usamos mediciones sensibles de las mismas. Porque a partir de las posiciones y distancias de las cosas en relación con cualquier cuerpo considerado inmóvil definimos todos los lugares, y después, con respecto a tales lugares calculamos todos los movimientos, considerando que los cuerpos se transfieren de algunos de sus lugares a otros. Así, en lugar de lugares y movimientos absolutos, usamos los relativos, sin inconveniente alguno en los asuntos cotidianos; pero en las disquisiciones filosóficas debemos de hacer abstracción de nuestros sentidos y considerar a las cosas en sí mismas, distintas de lo que sólo son mediciones sensibles de ellas. Porque puede ocurrir que no haya un cuerpo que esté realmente en reposo al cual pudiesen referirse los lugares y movimientos de otros cuerpos. Pero podemos distinguir el reposo y el movimiento, lo absoluto y lo relativo, uno de otro, por sus propiedades, causas y efectos. Una propiedad del reposo es que los cuerpos que están verdaderamente en reposo lo están en su relación mutua. Y por consiguiente es posible que en las regiones remotas de las estrellas fijas, o quizá mucho más allá de ellas, pueda existir algún cuerpo que esté absolutamente en reposo; pero es imposible saber por la posición de los cuerpos entre sí en nuestras regiones si algunos de los mismos mantienen la misma posición en relación con ese cuerpo remoto, de lo que se desprende que el reposo absoluto no puede determinarse de la posición de los cuerpos en nuestras regiones.

Es una propiedad del movimiento que las partes que conservan posiciones dadas a sus todos participan de los movimientos de esos todos. Porque todas las partes de los cuerpos que giran procuran retroceder del eje del movimiento; y el ímpetu de los cuerpos que se mueven hacia adelante se desprende del ímpetu conjunto de todas las partes. Por lo tanto, si se mueven los cuerpos que nos rodean, aquellos que están relativamente en reposo dentro de los mismos participarán de ese movimiento. Debido a esta razón el movimiento verdadero y absoluto de un cuerpo no puede determinarse por la traslación de éste en relación con aquellos que sólo parecen estar en reposo, porque los cuerpos externos no sólo deben parecer estar reposo sino estarlo realmente. Porque de lo contrario todos los cuerpos incluidos, aparte de su traslación de las proximidades de los que los rodean, participan asimismo de sus movimientos verdaderos; y aunque no se realizase esa traslación, tampoco estarían realmente en reposo, sino que sólo parecerían estarlo. Porque los cuerpos circundantes conservan con los circundados la misma relación que guarda la parte exterior de un todo con su interior, o la cascarilla con el grano, pero si la cascarilla se mueve, el grano se moverá también, por formar parte del todo, sin que haya traslación de lo cercano a la cascarilla.

Una propiedad casi igual que la anterior es ésta: que si un lugar se mueve, cualquier cosa que se haya colocado en él se mueve también, y por consiguiente un cuerpo que se mueve de un lugar en movimiento participa también del movimiento de su lugar. En razón de lo cual todos los movimientos, de los lugares en movimiento, no son otra cosa que partes de movimientos enteros y absolutos; y todo movimiento entero está compuesto del movimiento del cuerpo de su lugar inicial, y el movimiento de este lugar de su lugar, y así sucesivamente, hasta que llegamos a algún lugar inmóvil, como en el ya mencionado ejemplo del marinero. De manera que los movimientos enteros y absolutos no pueden determinarse más que mediante lugares inmóviles, y por esa razón refería antes esos movimientos absolutos a lugares inmóviles, pero los relativos a los lugares móviles. Pero no hay otros lugares inmóviles que aquellos que, de infinito en infinito, conserven las mismas posiciones dadas entre ellos, y por esta razón deben permanecer siempre inmóviles, y por lo tanto constituyen un espacio inmóvil.

Las causas por las cuales se distinguen entre sí los movimientos verdaderos y los relativos son las fuerzas impresas a los cuerpos para generar movimiento. El movimiento verdadero no se genera ni se altera más que por alguna fuerza impresa sobre el cuerpo movido; pero el movimiento relativo puede generarse o alterarse sin que se imprima fuerza alguna sobre el cuerpo. Porque basta con sólo imprimir cierta fuerza a otros cuerpos con los que se compara el primero para que, al ceder, pueda cambiar esa relación en la que consistía el reposo o el movimiento relativo de ese otro cuerpo. Una vez más, el movimiento verdadero experimenta siempre algún cambio por cualquier fuerza que se imprima sobre el cuerpo en movimiento, pero el movimiento relativo no necesariamente experimenta cambio alguno debido a tales fuerzas. Porque si las mismas fuerzas se imprimen asimismo a esos otros cuerpos con los cuales se hace la comparación, para que pueda preservarse la posición relativa, se mantendrá entonces esa condición en la que

consiste el movimiento relativo. Y por consiguiente, cualquier movimiento relativo puede cambiarse cuando el movimiento verdadero permanecía inalterado, y el relativo puede preservarse cuando el verdadero experimenta algún cambio. Así, el movimiento verdadero no consiste de manera alguna en tales relaciones.

Los efectos que distinguen el movimiento absoluto del relativo son las fuerzas por las que retroceden del eje del movimiento circular. Porque no existen fuerzas tales en un movimiento circular puramente relativo, pero en un movimiento circular verdadero y absoluto son mayores o menores de acuerdo con la cantidad del movimiento. Si un recipiente colgado de una cuerda larga se hace girar tantas veces que la cuerda se enrolla mucho, luego se llena de agua y se mantiene en reposo junto con el agua, tras lo cual, por la acción súbita de otra fuerza, se lo hace girar en sentido contrario, y mientras la cuerda se desenrolla, el recipiente permanece cierto tiempo en ese movimiento; pero después de eso, al comunicarle gradualmente su movimiento al agua, hará que empiece a girar sensiblemente, y a alejarse poco a poco del centro, y ascender por los lados del recipiente, adoptando la forma de una figura cóncava (cosa que he experimentado), y cuanto más rápido se vuelve el movimiento más se eleva el agua hasta que por fin, ejecutando sus revoluciones en los mismos tiempos que mostró el verdadero movimiento circular del agua que aumentaba continuamente, llega a estar relativamente en reposo dentro de él. Este ascenso del agua muestra su intento de alejarse del eje de su movimiento; y el movimiento circular verdadero y absoluto del agua, que aquí es directamente contrario al relativo, se torna conocido y puede medirse gracias a ese intento. Al principio, cuando el movimiento relativo del agua en el recipiente era más grande, no producía ningún esfuerzo por alejarse del eje. El agua no mostraba tendencia alguna hacia la circunferencia ni ascenso ninguno hacia los lados del recipiente, sino que mantenía una superficie plana, por lo cual su verdadero movimiento circular no comenzaba aún. Pero después, una vez que hubo decrecido el movimiento relativo del agua, el ascenso de la misma hacia los lados del recipiente probó su esfuerzo de alejarse del eje, y este esfuerzo demostró el verdadero movimiento circular del agua que aumentaba continuamente hasta que adquirió su máxima cantidad, cuando el agua estaba en reposo relativo dentro del recipiente. Y de esta manera, este esfuerzo no depende de traslación alguna del agua con respecto a los cuerpos circundantes, ni el verdadero movimiento circular puede definirse por tal traslación. Existe sólo un verdadero movimiento circular en cualquier cuerpo en rotación, y corresponde a un único poder de esforzarse para retroceder de su eje de movimiento, como efecto propio y adecuado; pero los movimientos relativos, en un mismo cuerpo, son innumerables, de acuerdo con las diversas relaciones que mantiene con los cuerpos externos, e, igual que otras relaciones, carecen por entero de todo efecto real, aparte de que tal vez pueden participar en ese único movimiento verdadero. Ya eso se debe que haya quienes allí, en su sistema, suponen que nuestros cielos, al girar alrededor de la esfera de las estrellas fijas, transportan consigo a los planetas; las diversas partes de esos cielos, y los planetas, que de hecho están en reposo relativo en los cielos, sin embargo se mueven realmente. Porque cambian de posición uno en relación con otro (lo que nunca ocurre con los cuerpos que están en verdadero reposo), y al ser transportados junto con los cielos participan de sus movimientos, y como partes de conjuntos rotatorios se esfuerzan por alejarse del eje de sus movimientos.

Por consiguiente, las cantidades relativas no son las cantidades mismas cuyos nombres tienen, sino las mediciones sensibles de aquellas (ya sea precisas o imprecisas) que suelen usarse en lugar de las cantidades medidas en sí mismas. Y si el significado de las palabras ha de estar determinado por su uso, entonces con los nombres de tiempo, espacio, lugar y movimiento, sus medidas [sensibles] pueden ser correctamente comprendidas. Y si se hace referencia a las cantidades medidas mismas, la expresión será poco usual y puramente matemática. Al respecto, violan la precisión del lenguaje, que debe mantenerse exacto, quienes interpretan estas palabras por las cantidades medidas. Y no profanan menos la pureza de las verdades matemáticas y filosóficas quienes confunden las cantidades reales con sus relaciones y mediciones sensibles.

Es de hecho cosa de gran dificultad descubrir, y distinguir eficazmente, entre los movimientos verdaderos de cuerpos particulares y los aparentes; porque las partes de ese espacio inmóvil en las que se llevan a cabo esos movimientos de ninguna manera están expuestas a la observación de nuestros sentidos. Mas no es algo desesperado, porque tenemos argumentos que nos digan, en parte gracias a los movimientos aparentes, que son las diferencias de los movimientos verdaderos, y en parte a las fuerzas, que son las causas y efectos de los movimientos verdaderos. Por ejemplo, si dos esferas que se mantienen a determinada distancia una de la otra por medio de una cuerda que las conecta se hiciesen girar alrededor de su centro de gravedad común, a partir de la tensión de la cuerda podríamos descubrir el esfuerzo de las esferas por ale-

jarse del eje de su movimiento, y a partir de allí seríamos capaces de calcular la cantidad de sus movimientos circulares. Y entonces, si se imprimiesen fuerzas iguales simultáneamente a las caras alternas de las esferas para aumentar o disminuir sus movimientos circulares, a partir del aumento o el decremento de la tensión de la cuerda podríamos inferir el incremento o decremento de sus movimientos; y a partir de ello se descubriría sobre qué caras debían ejercerse esas fuerzas, a fin de que los movimientos de las esferas pudiesen aumentar lo más posible; o sea que podríamos descubrir sus caras más alejadas o aquellas que, en el movimiento circular, le siguen. Pero al conocerse las caras que le siguen, y en consecuencia las opuestas, que la preceden, podríamos saber asimismo la determinación de sus movimientos. Y así seríamos capaces de encontrar tanto la cantidad como la determinación de este movimiento circular incluso en un vacío inmenso, en el que no haya nada externo o sensible con lo cual pudiésemos comparar las esferas. Pero ahora, si en ese espacio se ubicaran algunos cuerpos remotos que mantuviesen siempre una posición dada entre sí, como lo hacen las estrellas fijas en nuestras regiones, de hecho no lograríamos determinar a partir de la traslación relativa de las esferas entre esos cuerpos si el movimiento pertenecía a aquéllas o a éstos. Pero si observásemos la cuerda, y descubriésemos que su tensión era esa misma tensión que requerían los movimientos de las esferas, podríamos llegar a la conclusión de que el movimiento estaba en las mismas, y que los cuerpos estaban en reposo; y por fin, finalmente, a partir de la traslación de las esferas entre los cuerpos encontraríamos la determinación de sus movimientos.

LOS AUTORES

RAÚL ARTURO ESPEJEL MORALES

Estudió la licenciatura en la Facultad de Ciencias de la UNAM y la maestría en ciencia e ingeniería de la computación en el Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas de la UNAM, actualmente es candidato a doctor. Es profesor asociado de tiempo completo en la Facultad de Ciencias. Cuenta con el nivel B del Programa de Estímulos a la Productividad Académica. Ha impartido diferentes cursos en el Departamento de Física de la Facultad de Ciencias, UNAM. Ha publicado artículos en revistas de investigación de circulación internacional y cuenta con varios libros de enseñanza media superior. Su labor académica se basa en la física computacional, visualización 3D y simulaciones. Ha participado en la organización de congresos nacionales de la Sociedad Mexicana de Física y es director técnico de la *Revista Mexicana de Física*.

MARÍA LUISA MARQUINA FÁBREGA

Estudió la licenciatura en la Facultad de Ciencias de la UNAM. Es profesora titular de tiempo completo en la Facultad de Ciencias. Cuenta con el nivel C del Programa de Estímulos a la Productividad Académica. Pertenece al Sistema Nacional de Investigadores nivel I. Ha impartido diferentes cursos en el Departamento de Física de la Facultad de Ciencias. Ha publicado más de 55 artículos en revistas de investigación de circulación internacional y cuenta con varios libros de enseñanza media superior. Se dedica a las espectroscopías Mössbauer, Raman e Infrarroja en superconductores, materiales amorfos e inorgánicos, así como al estudio de diferentes materiales. Es miembro de la Comisión de Mejoramiento de la Docencia del Departamento de Física y recibió la medalla Sor Juana Inés de la Cruz por la UNAM. Ha participado en la organización de 16 congresos nacionales de la Sociedad Mexicana de Física (SMF) y ha dirigido su *Boletín*. Es editora asociada del *Catálogo Iberoamericano de Programas y Recursos Humanos en Física*. Actualmente es tesorera de la SMF y decana del Consejo Técnico de la Facultad de Ciencias.

MARCO ANTONIO MARTÍNEZ NEGRETE

Estudió física en la Facultad de Ciencias de la UNAM, en donde obtuvo en 1973 el grado de doctor en ciencias. Ha realizado trabajos acerca de la formulación axiomática-epistemológica de algunas subdisciplinas de la física: termodinámica, mecánica clásica, electrostática, relatividad especial y probabilidad. Ha escrito libros de texto sobre termodinámica, en los niveles de bachillerato y licenciatura, como parte de propuestas didácticas basadas en enfoques epistemológicos, constructivistas,

de aprendizaje significativo y ciencia, técnica y sociedad. Su labor de divulgación se ha centrado en fundamentos de la física, desarme nuclear, fuentes renovables de energía y alternativas a los energéticos agotables, habiendo impartido más de 500 conferencias. Ha sido secretario académico de la Facultad de Ciencias, coordinador de la comisión que modificó el actual plan de estudios de la carrera de física. Obtuvo mención honorífica en el concurso de cuento de ciencia ficción, como parte de los festejos del Año Internacional de la Física, en 2005. Ha sido distinguido con la Cátedra Especial Carlos Graef Fernández en 2007 y 2008.

JOSÉ LUIS MORÁN LÓPEZ

En 1972 obtuvo el título de físico otorgado por la Escuela de Física de la Universidad Autónoma de San Luis Potosí. Continuó sus estudios en el Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional donde obtuvo la maestría en física teórica en 1974. Posteriormente viajó a Berlín donde realizó los estudios de doctorado obteniendo el grado de *Doctor Rerum Naturalium* con la nota de *Magna Cum Laude* en 1977. Fue profesor del Departamento de Física del Centro de Investigación y de Estudios Avanzados del IPN (1980-1986), del Instituto de Física de la Universidad Autónoma de San Luis Potosí (1987-2000). Fue el director general fundador del Instituto Potosino de Investigación Científica y Tecnológica (IPICYT), de 2000-2005. Miembro del Departamento de Materiales Avanzados para la Tecnología Moderna (2000-2008). Actualmente es profesor titular de la Facultad de Ciencias de la UNAM. Es miembro del Sistema Nacional de Investigadores, nivel III, con la distinción de Investigador Nacional de Excelencia. Ha publicado más de 200 artículos científicos en revistas y libros de difusión internacional, once artículos de divulgación en revistas de circulación nacional y múltiples contribuciones en periódicos. Es editor de nueve libros especializados. En 1984 se hizo acreedor a la beca de investigación "John Simon Guggenheim", otorgada en Nueva York. Fue investigador asociado del Centro Internacional de Física Teórica de Trieste. La Academia de la Investigación Científica (México) le otorgó el Premio en el área de Ciencias Exactas en 1985. Se hizo merecedor, en 1990, al premio internacional C. V. Raman del International Centre for Theoretical Physics. En 1996 se hizo acreedor del Premio Nacional de Ciencias y Artes en el área de Ciencias Físico-Matemáticas y Naturales, máxima distinción a un científico en nuestro país por parte del Gobierno Federal. Recibió la beca para académicos J. Tinsley Oden de la Universidad de Texas, en Austin, en 2007.

MIGUEL NÚÑEZ CABRERA

Estudió la licenciatura en física en la Facultad de Ciencias de la UNAM y realizó estudios de posgrado en el programa de maestría en enseñanza superior en la Facultad de Filosofía y Letras de la misma universidad. Cuenta con 30 años de experiencia en labores de docencia, investigación y divulgación en enseñanza de la física y las matemáticas, en los niveles medio superior y superior, así como impartido cursos a profesores de enseñanza media y media superior. Profesor de tiempo completo (en retiro) en la Facultad de Ciencias de la UNAM y de asignatura en el plantel núm. 9 de la Escuela Nacional Preparatoria. Es autor de varios libros en relación con la enseñanza y divulgación de la física y de varios artículos de divulgación publicados en revistas de esta especialidad. Ha dirigido tesis a nivel licenciatura, y ha participado en múltiples ocasiones como jurado calificador en concursos estudiantiles de experimentos de física. Ha impartido conferencias en torno a la física y su enseñanza, presentado ponencias sobre enseñanza y divulgación de la física en congresos nacionales, y prestado asesoría a instituciones educativas, en relación con la enseñanza de la física.

AGRADECIMIENTOS

Los autores agradecemos a Paris Sánchez C. por su apoyo en la elaboración de la tipografía de las ecuaciones, así como al físico Juan Tonda Mazón; asimismo, al Departamento de Física de la Facultad de Ciencias de la UNAM y a la Sociedad Mexicana de Física por las facilidades otorgadas.

p. 288 | *© Latin Stock México.*

INTRODUCCIÓN

En el Universo, por un lado se encuentra lo infinitamente gigantesco y lejano, como las galaxias y los cuásares, así como los eventos que ocurrieron hace miles de millones de años. En el otro extremo se encuentra lo infinitamente pequeño, como los átomos, los quarks y los leptones, así como los fenómenos subatómicos que ocurren en fracciones infinitesimales de segundo.

Entre estos extremos se encuentra el ser humano, cuya vida es muy corta si la comparamos con la edad del Universo, y gigantesca con respecto a la duración de algunos fenómenos atómicos. Sin embargo, lo más grandioso del ser humano es su ingenio y su capacidad para cuestionarse y tratar de comprender todo lo que le rodea. Es capaz de diseñar y construir sofisticados instrumentos que sirven para estudiar tanto lo muy grande como lo muy pequeño. El Universo se expande segundo a segundo, e incluye todo lo que existe, como las estrellas, las nebulosas, los cuásares, los hoyos negros y hasta el conjunto de todas las galaxias. Entre ellas la que habitamos, nuestra galaxia, llamada la Vía Láctea. A ella pertenecen el Sol y los planetas Mercurio, Venus, Marte, Júpiter, Saturno, Urano, Neptuno y, desde luego, la Tierra. En este planeta nos tocó vivir y es donde ocurren la mayoría de los fenómenos físicos que conocemos: los terremotos, los rayos, la lluvia, el viento, el calor, el frío, la nieve, las mareas y también los espectaculares impactos de meteoritos, los maremotos y tsunamis, y las erupciones volcánicas como las del Popocatépetl. Estamos rodeados de fenómenos físicos; algunos los podemos percibir con nuestros sentidos, mientras que para estudiar otros necesitamos instrumentos especializados como telescopios, aceleradores de partículas cargadas, láseres, microscopios de varios tipos y muchos más.

En el primer capítulo se explican las hipótesis y evidencias sobre la estructura general del Universo observable y las características de la gran explosión que, se supone, le dio origen. En relación con los fenómenos que ocurren a escala humana, su explicación se aborda en los capítulos 2 al 6. En el último capítulo, se expone una visión del mundo microscópico, al que pertenecen las partículas elementales.

Así como no es fácil ver lo que ocurre en el Universo, tampoco es fácil ver lo que ocurre dentro de nuestro cuerpo. Para hacer un viaje hacia el interior de nosotros mismos tendríamos que hacernos microscópicos. Primero, para atravesar nuestra piel, tenemos que hacernos cien mil veces más pequeños; entonces podríamos ver las células que la constituyen. Si nos achicamos a la centésima parte del tamaño anterior podemos ver las mo-

léculas que forman estas células, y si nos hacemos aún cien veces más pequeños, podremos ver los átomos de estas moléculas. Los átomos, de los que estamos constituidos los seres vivos y las cosas que nos rodean, se mueven continuamente y están formados por partículas llamadas electrones y núcleos. Además, intercambian energía con otros objetos microscópicos en múltiplos enteros de una cantidad bien definida llamada *quanto*, cuyo estudio dio origen a la mecánica cuántica.

Los núcleos atómicos están compuestos de partículas aún más pequeñas, llamadas protones y neutrones. Sin embargo, los científicos que estudian las partículas elementales han propuesto y demostrado que existen otros componentes más pequeños llamados *quarks*. Éstos, junto con los *leptones*, que son una clase de partículas a las que pertenece el electrón, constituyen, hasta ahora, lo indivisible.

Lo anterior constituye una breve descripción del Universo, desde lo inmenso, aquello que abarca distancias de miles de millones de años luz, hasta lo más pequeño: las células y el mundo de las partículas elementales. La cosmología es una especialidad de la física, que ha unido a los dos extremos de la escala. Los científicos que la estudian tratan de comprender el origen y la evolución del Universo con base en observables astronómicas, y como todas las áreas de la física, se apoya en los conocimientos que se han obtenido mediante muchos experimentos y observaciones realizados a través de la historia. El ser humano es así constructor y testigo único de tan maravillosa ciencia.

DESDE LA GRAN EXPLOSIÓN

El Cosmos es:
todo lo que es,
todo lo que fue,
todo lo que será.

TEMA

1

La Tierra vista desde la Luna | © NASA | *Véase video en* CD: *"Introducción a la cosmología".*

1.1 LA GRAN TEORÍA

El solo hecho de pensar en el Universo, sus galaxias y sus estrellas, agudiza los sentidos. Pensar y hablar del Cosmos es ingresar a un inmenso valle de misterios en donde la ciencia y la fantasía se confunden.

Durante los últimos 100 años hemos visto el nacimiento y desarrollo de la teoría más importante surgida de la ciencia moderna: la *expansión cósmica*.

Muchas leyendas se han creado acerca de lo que vemos en el cielo noche tras noche. Así, han surgido las constelaciones; imágenes que cada cultura ha colocado en el cielo para

expresar sus emociones, deseos y necesidades. El nombre de la Vía Láctea surge de la leyenda donde la diosa griega Hera deja una marca de leche en el cielo al descubrir a Hércules prendido de su seno, hijo bastardo de su esposo Zeus.

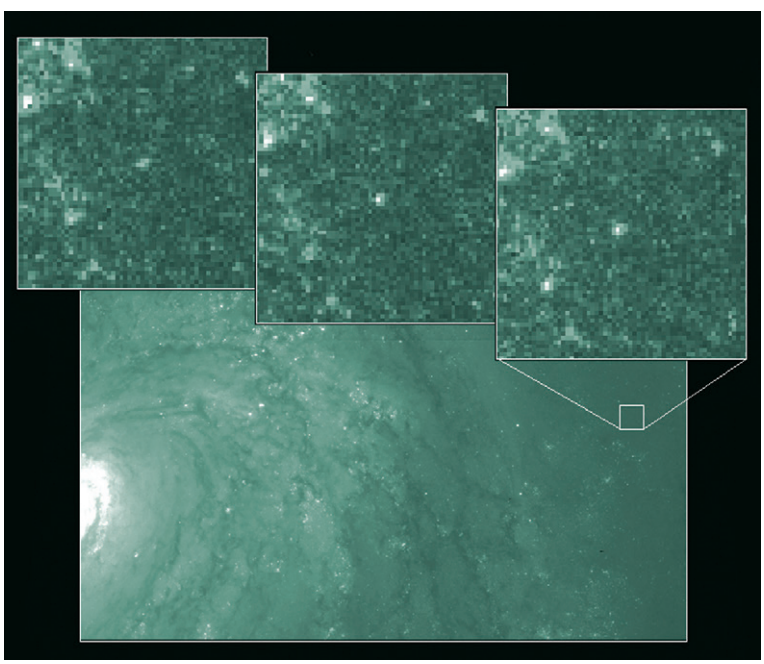
Los cosmólogos suponen que todo el Universo, alguna vez, estuvo concentrado en un solo punto, una especie de huevo primordial, que estalló. Hoy, de aquella *gran explosión*, sólo vemos su proceso expansivo.

Mayor alarde de imaginación es imposible. En la actualidad se concibe al Cosmos, con sus estrellas, nebulosas y galaxias, como el gran protagonista de un evento, el más grande, el más espectacular. Sin embargo, si nos detenemos un momento a reflexionar, tal vez debiéramos preguntarnos: ¿cómo podemos llegar a determinar la forma en que evoluciona el Universo?, ¿cómo evidenciamos que su movimiento es expansivo?

Sólo cuando se pueda asegurar que conocemos la evidencia y la forma en que se interpreta, podremos asegurar que la Gran Explosión o *Big Bang* no es una especulación, sino un hecho.

1.2 UN BRILLO VARIABLE

A principios del siglo xx, bajo la dirección de Edward C. Pickering (1846-1919), el observatorio del Harvard College se ocupaba de detalladas observaciones estelares. Para realizar el inmenso trabajo de ver miles de fotografías, comparar brillos, hacer gráficas y cálculos, Pickering contrató a mujeres astrónomas dispuestas a trabajar y, en su opinión, temperamentalmente más aptas para dicha labor. A pesar de sus grandes contribuciones prácticamente no fueron reconocidas por la comunidad astronómica de su tiempo. Tal vez por eso resalta más la obra de una de ellas: Henrietta Swan Leavitt (1868-1921), quien habría de establecer un método revolucionario para la medición de grandes distancias. Hija de un ministro congregacional, sorda, de modales reservados y notablemente brillante, su trabajo se centró en las placas fotográficas obtenidas durante varios años por un telescopio de 60 cm que el observatorio de Harvard tenía en las montañas de Perú, las cuales mostraban un enjambre estelar muy conocido por los observadores del hemisferio sur: la Pequeña Nube de Magallanes.



Variable Cefeida en M-100.

Leavitt se encontró con numerosos casos de estrellas cuyo brillo variaba periódicamente. En virtud de que la primera de estas curiosas luciérnagas fue la estrella Delta, en la constelación de Cefeo, se les conoce como estrellas variables *cefeidas*. Las cefeidas se expanden y contraen con regularidad; como consecuencia, brillan intensamente, se apagan y vuelven a brillar, repitiendo el ciclo. El lapso de tiempo en que se repite este fenómeno, es decir, el periodo de una variable cefeida, puede ser tan corto como un día o tan largo como varios meses. Sea cual sea el ritmo, por lo general, la periodicidad tiene una sorprendente precisión de hasta uno o dos minutos(figura 1).

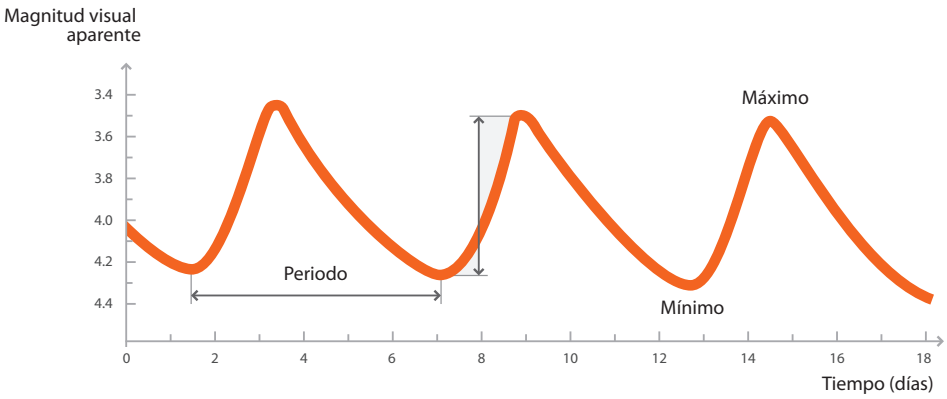


Figura 1. Magnitud visual aparente contra *tiempo.*

Para 1908 Leavitt había compilado una lista de más de un millar de estas estrellas en la Pequeña Nube de Magallanes.[1] Dieciséis de ellas aparecían en suficientes fotografías como para poder determinar sus periodos. De su detallado estudio empezó a surgir una curiosa característica: *mientras más largo es su periodo, mayor es su brillo máximo.* Esto se muestra con claridad en la figura 2.

En 1912, cuando ya había ampliado su estudio a 25 cefeidas, publicó un artículo en el que mostraba que el brillo y el periodo están, en efecto, relacionados matemáticamente.

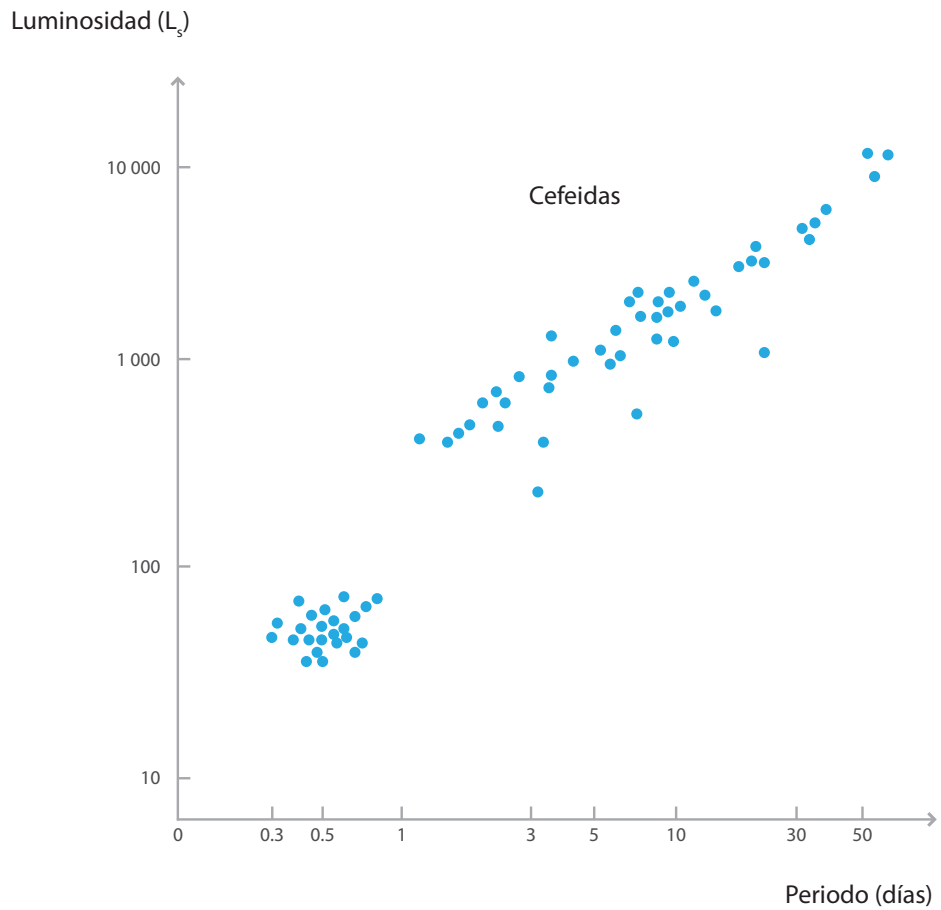


Figura 2. Luminosidad contra *periodo.*

[1] Henrietta S. Leavitt, "1777 variables en las nubes de Magallanes", *Annals of Harvard College Observatory*, pp. 87-108.

Puesto que estas estrellas variables se hallan, probablemente, a casi la misma distancia de la Tierra, sus periodos están aparentemente asociados a su emisión real de luz.[2]

El resultado fue impactante; Henrietta Leavitt había encontrado, como se explica a continuación, la forma de conocer el brillo intrínseco de una estrella y de establecer la distancia a la que se encuentra.

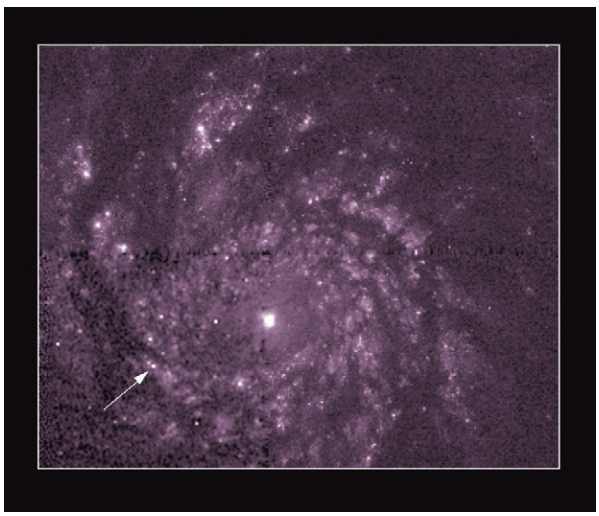
1.3 GRANDES DISTANCIAS

Para conocer la distancia a la que se encuentra una fuente luminosa, se necesita conocer su *brillo intrínseco.* Por ejemplo: para conocer la distancia a la que se encuentra una lámpara que se sabe que es de 100 W (brillo intrínseco), basta medir el brillo aparente y considerar que el brillo decae como el cuadrado del inverso de la distancia. Pero Henrietta Leavitt había encontrado una relación de proporcionalidad directa entre el periodo de oscilación de las cefeidas y su brillo intrínseco.

Entonces, si se conoce el brillo intrínseco de una cefeida cercana y su distancia, puede inferirse la distancia a la que se encuentra una cefeida más lejana del mismo periodo.

En 1917 existía el debate sobre la estructura de las nebulosas espirales y no se lograban resultados convincentes. Sin embargo, los fotógrafos de estas nebulosas empezaron a notar la existencia de objetos cercanos a ellas que incrementaban notablemente su brillo. Se pensaba que ello se debía a que muchas estrellas sufrían una explosión. Dicho comportamiento se traduce en que estrellas con poco brillo, lo aumentan en forma súbita, dando la impresión de ser una estrella nueva. Esto dio origen a su nombre latino: *nova*.

Supernova cerca del núcleo de la galaxia M-51.



Parecería razonable que, accidentalmente, alguna nova pudiera estar ubicada en la línea de observación de una nebulosa espiral; pero, el hecho de que hubiera muchas originó preguntas como: ¿es posible que estas novas formen parte de la espiral?, ¿acaso aquellas nebulosas espirales que son invisibles aun para los más poderosos telescopios, son un conglomerado de estrellas?

Este razonamiento resultó relevante. Si las nebulosas espirales y sus novas son del mismo tipo que la Vía Láctea, el brillo promedio de las novas, decenas de miles de veces más débiles, indica que están cien veces más lejos. Algunos astrónomos empezaron a calcular las distancias basándose en esta hipótesis.

Los resultados indicaban que las nebulosas espirales estaban mucho más alejadas que la frontera de la Vía Láctea, que también es espiral.

Para investigadores como Harlow Shapley (1885-1972) estas observaciones eran pruebas inequívocas de que las nebulosas espirales son galaxias independientes, *Universo Isla*. Sin embargo, cuando formuló su teoría de *La Gran Galaxia*, cambió de opinión. La Vía Láctea, tal como la imaginaba, era demasiado grande comparada con todas las otras nebulosas espirales.

[2] Henrietta S. Leavitt y Edward C. Pickering, "Periodos de 25 estrellas variables en la pequeña Nube de Magallanes", *Harvard College Observatory Circular*, pp. 1-3.

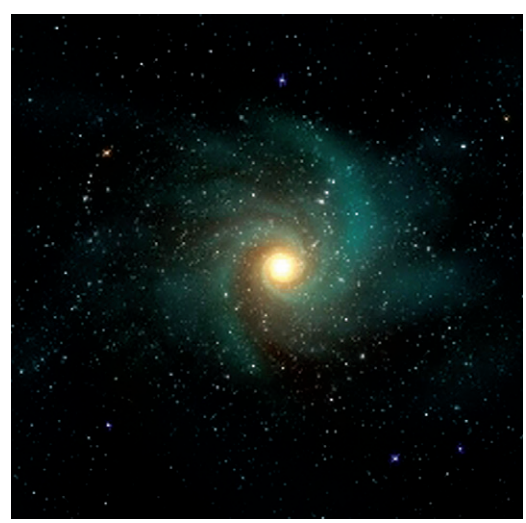
Parte de la Vía Láctea.

1.4 GRANDES VELOCIDADES

Desde 1912, en el observatorio Lowell de Flagstaff, Arizona, un espectroscopista hábil y paciente llamado Vesto Slipher (1875-1969), quien había logrado acumular suficiente luz de una galaxia espiral de la constelación de Andrómeda, midió su desplazamiento Doppler y, usando los cálculos de Fizeau, determinó sus velocidades. El resultado fue impactante: la nebulosa de Andrómeda se está acercando a la Tierra a una velocidad de 300 km/s.[3] Ésta era la velocidad más grande jamás medida para algún objeto celeste.

Para 1914 Slipher había determinado la velocidad, a lo largo de la línea de observación de quince nebulosas espirales. Sorprendentemente, trece de ellas se alejan, algunas a casi 800 km/s; más del doble que la velocidad de aproximación de Andrómeda. Según Heber Doust Curtis (1872-1942), descubridor de las novas, y otros astrónomos, ése era un poderoso argumento a favor de la teoría del Universo Isla. Las velocidades de las espirales son demasiado grandes para estar ligadas gravitacionalmente a la Vía Láctea.

Nebulosa espiral.



1.5 ¡ES VARIABLE!

Cuando Edwin Hubble (1889-1953) estaba terminando el doctorado en la Universidad de Chicago en 1917, George E. Hale estaba preparándose para contratar al personal que habría de operar el telescopio de 2.5 m que se concluiría en Monte Wilson, en las montañas de San Gabriel, California, así que lo invitó a incorporarse al proyecto. Sin embargo, en ese año Estados

[3] V. M. Slipher, "La velocidad radial de la nebulosa de Andrómeda", *Lowell Observatory Bulletin 58*, pp. 56-57.

Unidos decidió participar en la primera guerra mundial, por lo que declinó la oferta por alistarse en el ejército, enviándole un telegrama que decía: *Lamento no poder aceptar su invitación. Marcho a la guerra.*

Cuando volvió, en 1919, la invitación seguía en pie y aceptó. Empezó clasificando nebulosas, entre ellas las nebulosas espirales. Hubble creía en la teoría del Universo Isla y esperaba que el enorme telescopio de 2.5 m de Monte Wilson le ayudara a validarla. Pero ni siquiera este poderoso instrumento podía capturar inequívocamente estrellas individuales en las fotografías de las espirales. En algún momento creyó ver estrellas, pero sus colegas se mostraron escépticos ya que el ver estrellas no resolvía nada si no se podía determinar la distancia a las espirales.

Nebulosa M-51.

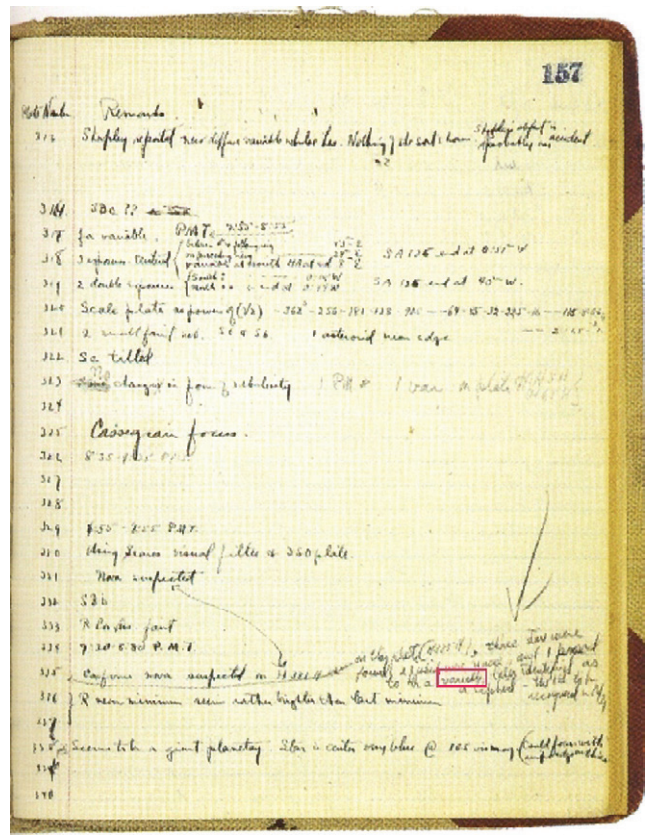
Sin embargo, las fotografías de las espirales eran suficientemente nítidas como para mostrar los puntos de luz que Curtis había identificado como novas y Hubble centró su atención en ellas. Un día de 1923, cuando trabajaba con varias fotografías de la, hasta entonces considerada, *nebulosa Andrómeda*, reexaminó un punto de luz al que había marcado con *N* para identificarla como nova. Revisó placas anteriores y encontró que el cambio de brillo era periódico. Eufórico, tachó la *N* y escribió *¡variable!* La nova se comportaba exactamente como una variable cefeida.

Al fin Hubble tenía una forma de determinar la distancia a la nebulosa de Andrómeda. La cefeida en la espiral era muy débil, mucho más que las que Shapley había descubierto en los cúmulos globulares. Usando la calibración de Shapley, Hubble obtuvo que la nebulosa de Andrómeda (M-31 en los catálogos) estaba a 900 mil años luz de distancia, mucho más allá de la Vía Láctea. La conclusión entonces fue contundente: *Andrómeda es una galaxia independiente, completamente desarrollada*.

Nebulosa espiral | © NASA.

Hubble no se apresuró a enviar su artículo porque la relación del periodo con la luminosidad todavía era muy controvertida como indicador de distancias. En uno de sus últimos trabajos escribió:

> Con el incremento de las distancias nuestro conocimiento se desvanece, y se desvanece rápidamente hasta que en el último e impreciso horizonte buscamos, entre fantasmales errores de observaciones, puntos de referencia que apenas son más sustanciales. La búsqueda continuará. El deseo de conocimiento es más antiguo que la historia. Nunca resulta satisfecho, y nunca podrá ser suprimido.

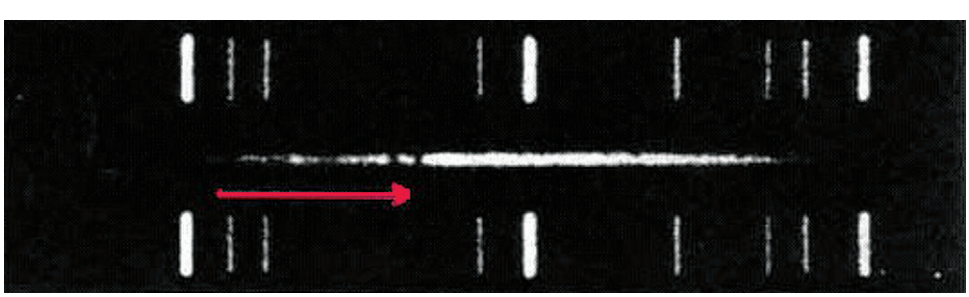


Nota original de Hubble.

1.6 RECESIÓN GALÁCTICA

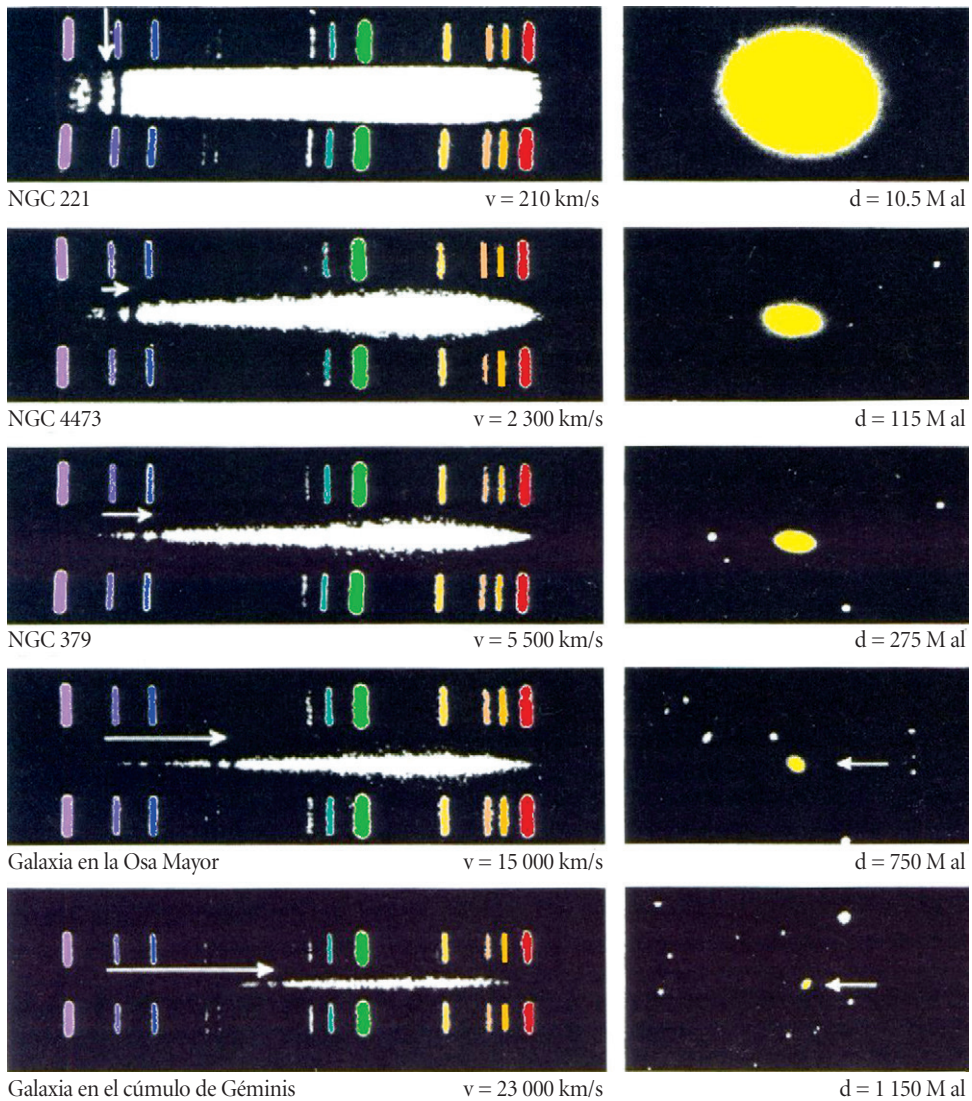
El astrónomo estadunidense Milton Lasell Humason (1891-1972), siguiendo los pasos de Slipher y Hubble, retomó el trabajo sobre las velocidades radiales de las galaxias. Con sumo cuidado comenzó a tomar fotografías que requerían días enteros de exposición para registrar los espectros de galaxias cada vez más tenues. Entre las galaxias más débiles descubrió velocidades mucho mayores que las estudiadas anteriormente. En 1928 Humason midió la velocidad radial de la galaxia NGC 7619, obteniendo 3 800 km/s.[4] Hacia 1936 estaba midiendo velocidades de 40 mil km/s, más de un octavo de la velocidad de la luz. Y siempre se alejaban de nuestro planeta.[5]

Las velocidades que se reportaban eran tan grandes que los astrónomos empezaron a poner en tela de juicio la interpretación Doppler del corrimiento al rojo de la luz observada. Un corrimiento hacia el rojo, ¿implica necesariamente que la fuente se está alejando o existe alguna explicación alternativa que evite tener que aceptar velocidades tan grandes? Tal vez la luz de las galaxias lejanas se ve enrojecida, en su larga travesía, por el polvo fino del espacio intergaláctico.



Corrimiento hacia el rojo.

[4] M. L. Humason, "¿Está el Universo expandiéndose?", *Astronomical Society of the Pacific Leaflets*, p. 161.
[5] M. L. Humason, "No. 531, La velocidad radial aparente de 100 nebulosas extragalácticas", *Contributions from the Mount Wilson Observatory,* pp. 1-13.



Corrimiento hacia el rojo de diferentes galaxias.

Hubo quien propuso que, por su interacción con el polvo, la luz pierde energía. Dicha pérdida se traduce en un alargamiento de la longitud de onda. El espectro se habrá enrojecido. Así, el corrimiento al rojo no se debería a gigantescas velocidades de recesión de las galaxias, sino a que se recibe su luz alterada.

Sin embargo, en algunas galaxias, la desviación medida puede ser hacia el azul; ¿esto significa que la luz gana energía en su viaje?

En resumen, la interpretación más aceptada del corrimiento al rojo es que las galaxias se alejan de la Tierra a velocidades gigantescas y muy pocas se acercan. A este fenómeno de alejamiento se le llama *recesión galáctica*.

Hubble siguió trabajando paralelamente a Humason y logró estimar la distancia a las galaxias; en 1929 usó los resultados de Slipher y Humason encontrando que la velocidad de recesión de las galaxias aumentaba proporcionalmente a la distancia que nos separaba de ellas.[6] Una galaxia que se encuentra dos veces más alejada de nosotros, retrocede con el

[6] Edwin P. Hubble, "Una relación entre la distancia y la velocidad radial entre nebulosas extragalácticas", *Proceedings of the National Academy of Sciences of the United States of America*, pp. 168-173.

doble de la velocidad, y si la distancia es el triple, así será la velocidad de alejamiento.

A este comportamiento se le conoce como *ley de Hubble*. Graficando la velocidad de recesión *v* contra la distancia *r* obtenemos una recta (figura 3).

La pendiente la llamaremos *H* y es conocida como la *constante de Hubble*. Esto es:

$$H = \frac{\Delta v}{\Delta r},$$

donde $\Delta v = v - v_0$ y $\Delta r = r - r_0$. Ya que la gráfica pasa por el origen podemos tomar los puntos iniciales v_0 y r_0 como cero. Así, la forma más común de esta ley es:

$$v = Hr,$$

$$\text{con } H = 2.3 \times 10^{-18}\frac{1}{\text{s}}.$$

Lo más desconcertante de la ley de Hubble se resumía en la pregunta: *¿por qué todas las galaxias se alejan de la nuestra como si ésta fuera el centro de su movimiento de expansión?*

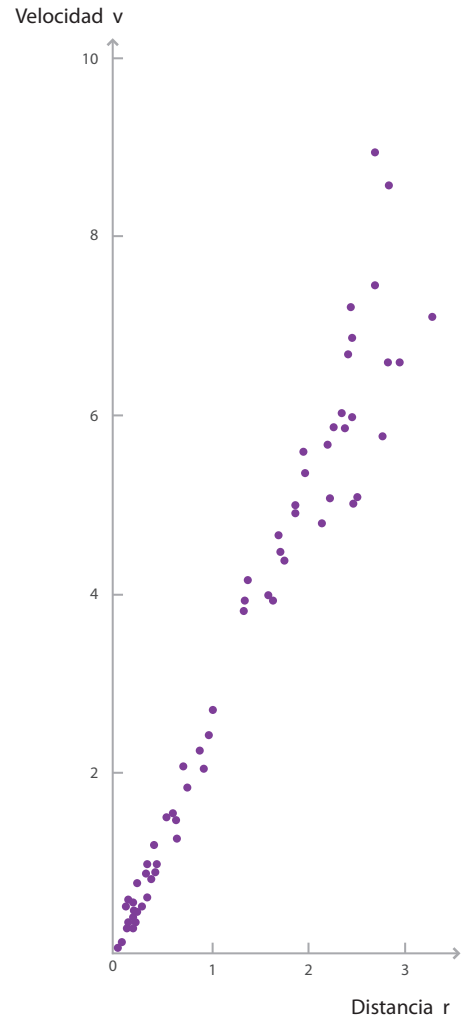


Figura 3. Velocidad de recesión contra distancia de algunas galaxias.

1.7 SE VE LO MISMO

Supongamos un conjunto de galaxias en una malla cuadrada de lado *r*: ¿qué vería un observador situado en cualquiera de las galaxias?

Por la ley de Hubble $v = Hr$, $2v = H2r$, … entonces la galaxia de referencia, la 4 en la figura 5 (p. 16), tendrá $v = 0$; la primera de la derecha lleva *v*, la segunda *2v*, la tercera *3v*, y así. Simétricamente, hacia el lado izquierdo la primera tiene *–v*, la segunda *–2v*, y así. El signo negativo es porque tienen diferente sentido.

Tomemos otra galaxia cualquiera como segundo punto de observación, digamos la 6. Para este segundo observador, su galaxia lleva una velocidad $v_2 = 0$, mientras que para el primero era $2v = v_1$. De esta forma se cumple que

$$v_2 = v_1 - 2v.$$

Haciendo esta operación para cada galaxia, resulta que el segundo observador ve el mismo comportamiento en las velocidades que el primero, es decir, se cumple la ley de Hubble, pero ahora como si él fuera el centro. Podemos hacer lo mismo con cualquier otra galaxia y el resultado será idéntico (figura 4).

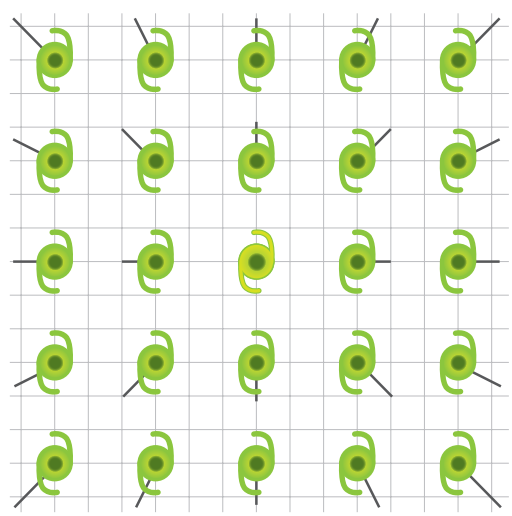


Figura 4. Las galaxias alejándose.

Entonces, desde cualquier punto del Universo se ve que las demás galaxias se alejan. La linealidad de la ley de Hubble implica que todo lugar del Universo es equivalente (figura 5).

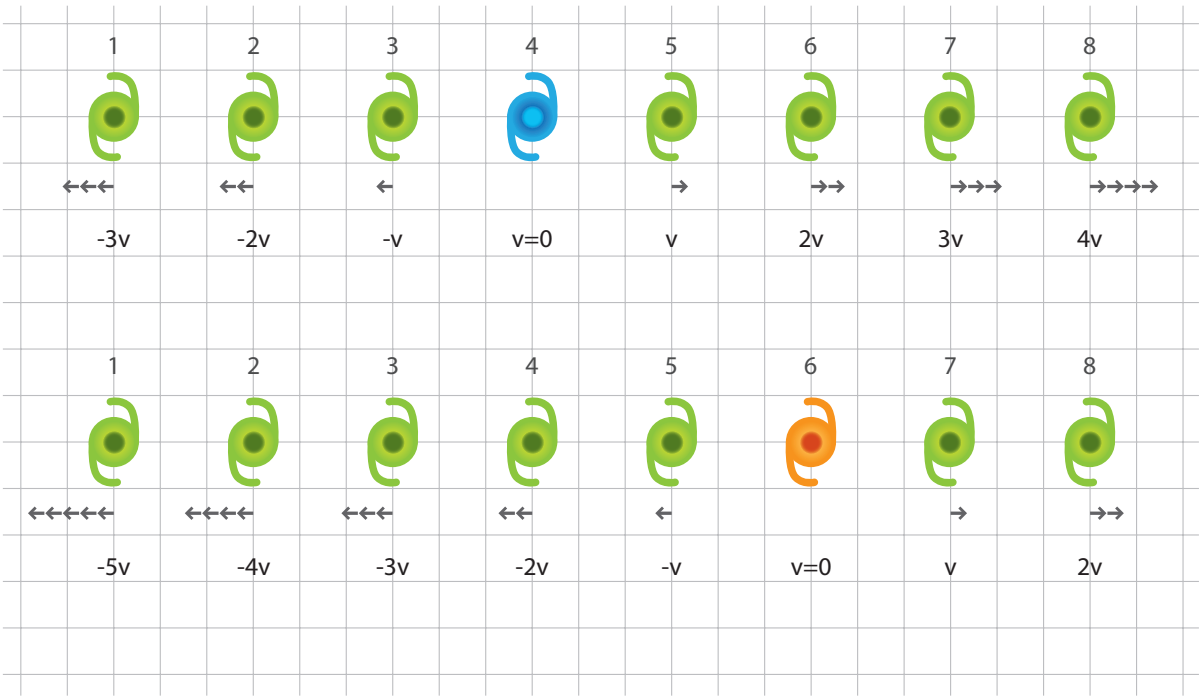


Figura 5. Galaxias alineadas.

1.8 EL HUEVO CÓSMICO

En 1927 el astrónomo belga Georges Édouard Lemaître (1894-1966) sugirió que, en el tiempo cero, toda la materia y energía del Universo se hallaban, efectivamente, comprimidas en una gigantesca masa cuyo diámetro no rebasaba unos cuantos años luz, a la que Lemaître llamó el *huevo cósmico*.

La velocidad V, tal vez constante, con la que se mueve la galaxia debe ser $V = \Delta d/\Delta t$, siendo Δt el tiempo transcurrido desde el estallido. Despejando, tenemos $\Delta t = \Delta d/V$. En nuestro caso, $\Delta d = r$; y, por la ley de Hubble, $V = Hr$. Entonces

$$\Delta t = \frac{r}{Hr}.$$

Así que la edad del Universo es, justamente, el inverso de la constante de Hubble.

$$\Delta t = \frac{1}{H}.$$

El huevo cósmico era inestable y estalló en la más fantástica y espectacular explosión que es posible imaginar. Los fragmentos producidos por dicha explosión fueron violentamente despedidos en todas direcciones, convirtiéndose en las galaxias actualmente en recesión, según dramatizaba Lemaître.

El modelo del huevo cósmico, su explosión y sus remanentes, parecía incluir los razonamientos teóricos de algunos cosmólogos como Einstein, Friedman y Sitter, así como las observaciones de Slipher, Hubble y Humason. Esto fascinó al astrónomo ruso-estadunidense

George Gamow (1904-1968), quien, con su enorme talento divulgador, impulsó esta teoría.

Durante una conferencia para la estación radiofónica BBC de Londres, el astrónomo inglés Fred Hoyle (1915-2001) defensor de la teoría del Universo en estado estacionario, estableció que el Universo es estático, siempre igual, sin principio ni fin. Al referirse a las teorías expansivas, en forma burlona mencionó que se proponía una teoría "fantástica", que se atrevía a afirmar que el Universo nació de una Gran Explosión.

Con el tiempo, el tono despreciativo desapareció, igual que la idea del Universo estacionario, para convertir a la Gran Explosión en la teoría más popular del nacimiento del cosmos. Pero las dudas sobre su realidad permanecieron y crecieron. ¿Realmente hubo una Gran Explosión?

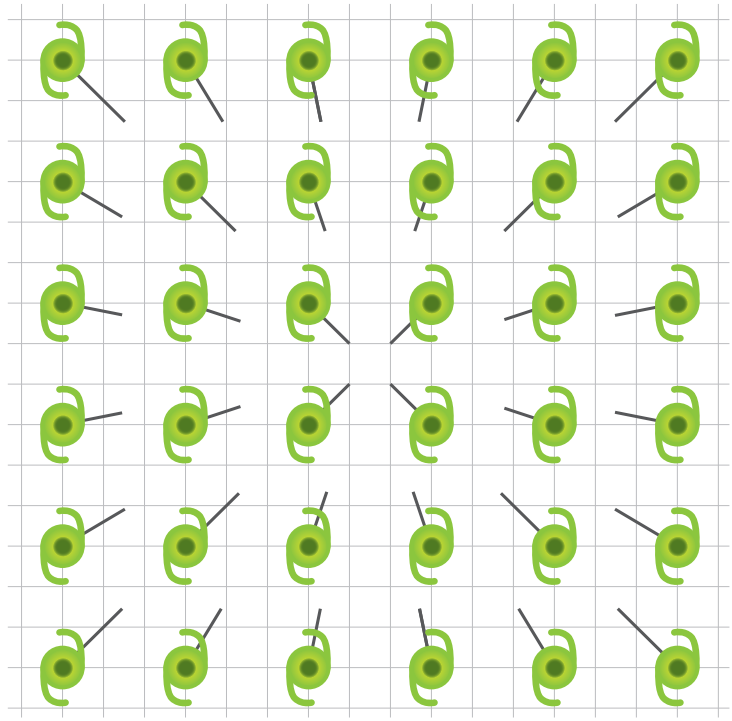
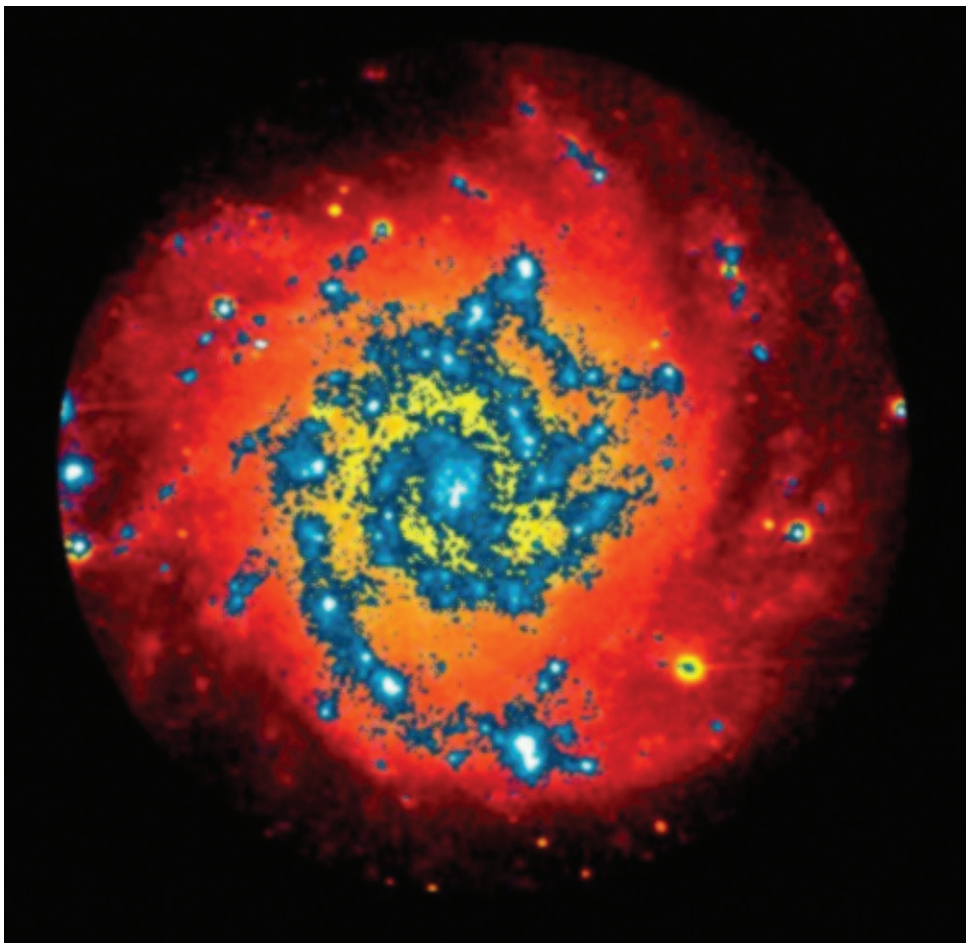


Figura 6. Galaxias contrayéndose.



Representación imaginaria de la Gran Explosión.

1.9 LA RADIACIÓN FÓSIL

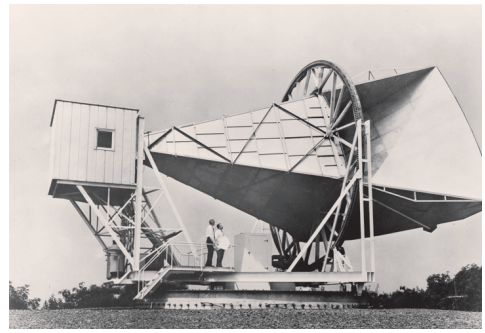
En 1960 los Laboratorios Bell construyeron una antena gigante en Holmdel, Nueva Jersey, como parte de un sistema pionero para la transmisión vía satélite llamado *Eco*. Este sistema colectaba y amplificaba las señales débiles de radio que rebotaban en grandes globos metálicos colocados en la alta atmósfera, enviando así señales a través de grandes distancias. Sin embargo, un par de años después se lanzó el satélite Telstar, haciendo obsoleto este sistema.

Mientras tanto, el alemán Arno Penzias (1933), astrónomo especializado en ondas de radio, que se unió a los Laboratorios Bell en 1958, y Robert Wilson (1936), se interesaron en el sistema *Eco*. Penzias había hecho su posdoctorado usando la técnica del efecto *máser* (amplificación de microondas por emisión estimulada de radiación) para amplificar y medir señales de radio provenientes del espacio intergaláctico. Wilson también había usado el máser para amplificar las señales de radio débiles provenientes de la Vía Láctea.

Penzias y Wilson pensaron que la antena de Holmdel podría ser empleada como un gran telescopio de radio y deseaban usarla para continuar sus observaciones, pero debían esperar a que terminara su uso comercial. El lanzamiento del Telstar en 1962 dio a ambos

investigadores lo que deseaban: la liberación de la antena de Holmdel y su dedicación a la investigación básica.

Cuando comenzaron a usarla como un telescopio para radio (hoy *radiotelescopio*), detectaron que había un ruido de fondo (como la estática en un radio). Esta molestia era una señal uniforme en la frecuencia de las microondas que parecía provenir de todas las direcciones. Llegaron a pensar que el ruido era generado por el propio telescopio.



Antena de Holmdel | © <http://commons.wikimedia.org/wiki/File:Horn_Antenna-in_Holmdel,_New_Jersey.jpeg>.

Verificaron todas las posibles fuentes de aquel ruido. Colocaron la antena en una dirección perpendicular a la ciudad de Nueva York. Y ¡no!, no era la interferencia urbana. Tampoco era la radiación proveniente de nuestra galaxia.

El ruido permaneció igual durante un año; no podía venir del Sistema Solar ni de la prueba nuclear subterránea realizada en Nevada en el año de 1962, porque a un año habría mostrado una disminución. Finalmente, los radioastrónomos decidieron medir las características de la radiación de fondo, encontrando que se le podía asociar una temperatura de alrededor de tres grados Kelvin. Enseguida, empezaron a buscar explicaciones teóricas.

Radiotelescopio | © Latin Stock México.

Al mismo tiempo, Robert Dicke (1916-1997) había elaborado una teoría sobre la Gran Explosión, sugiriendo que el residuo de la explosión tomaba la forma de una radiación de fondo de baja temperatura. Dicke buscaba evidencia para esta teoría cuando Penzias y

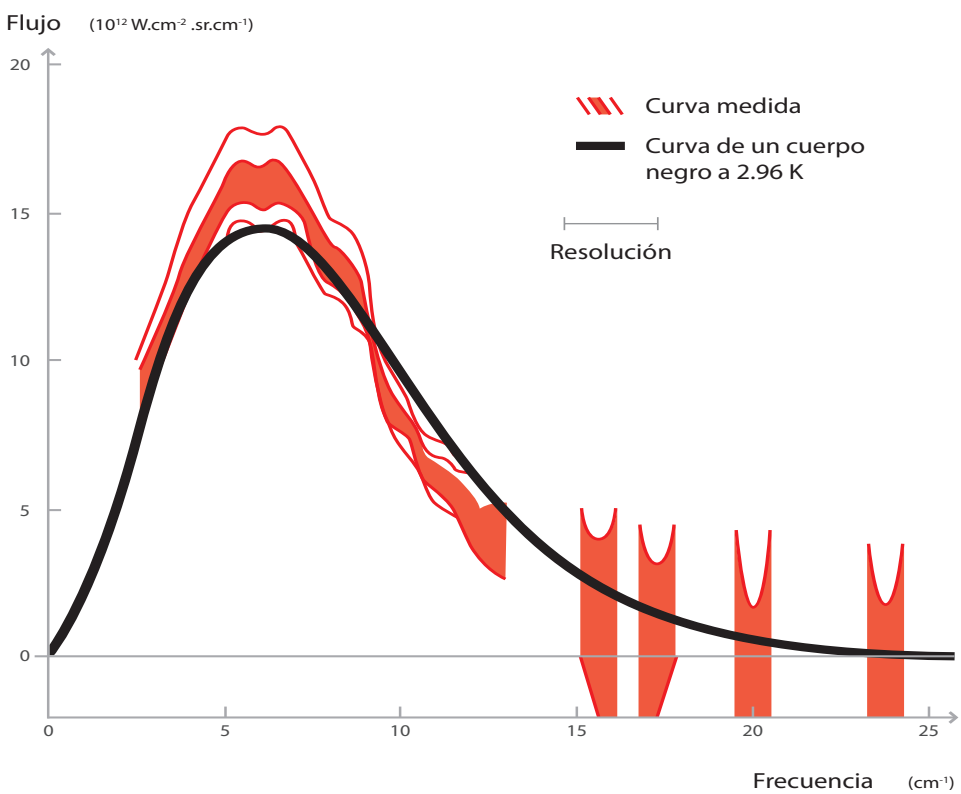
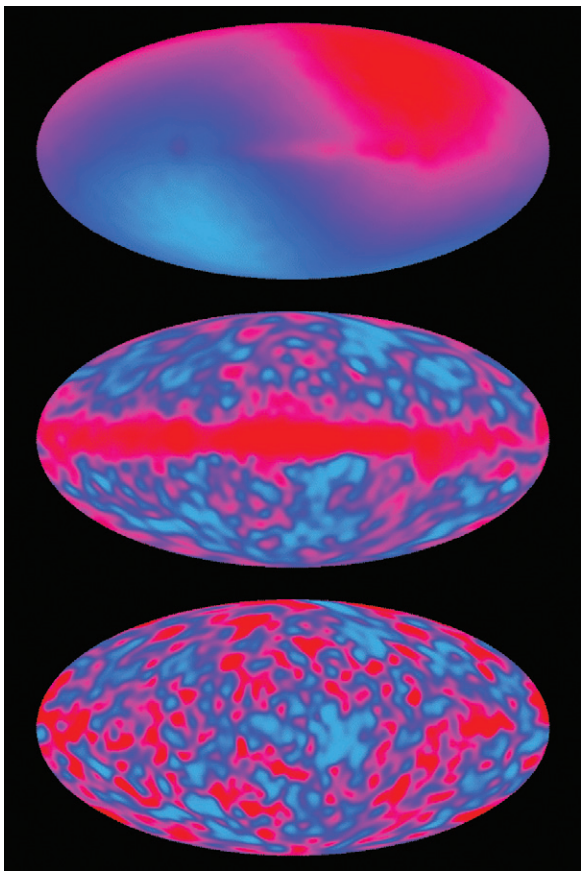
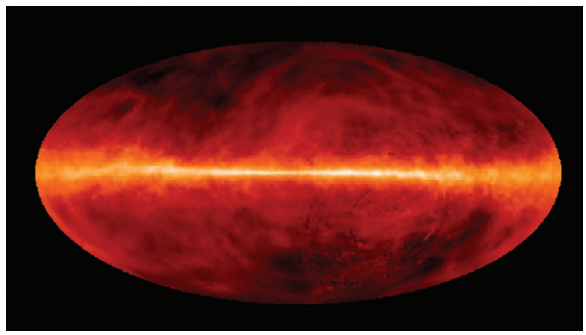


Figura 7. Flujo como función de la frecuencia.





Wilson se comunicaron a su laboratorio; él compartió sus ideas con los radioastrónomos y, al conocer sus observaciones, comentaría a sus colaboradores: *hemos actuado como paleontólogos.*

Curiosamente, cuando Robert Wilson realizó sus estudios, se creía en la teoría del estado estacionario y, por lo tanto, se sentía incómodo con la explicación de que el ruido de radio que detectaba,surgiera de la Gran Explosión. Cuando, conjuntamente Penzias, Wilson y Dicke, publicaron sus trabajos, los investigadores de los Laboratorios Bell insistieron en que fueran *sólo los hechos*: simplemente informar de las observaciones registradas. Así que las contribuciones fueron escritas por separado: dos cartas publicadas juntas, una seguida de la otra. La de Dicke y sus colaboradores contenía la reflexión teórica: *si el Universo tuvo un origen singular, pudo haber sido extremadamente caliente en sus estadios tempranos.*

¿Podría el Universo haberse llenado con una radiación de fondo a partir de este posible momento de alta energía? La temperatura de la radiación podría variar inversamente con el parámetro de expansión, es decir, el radio del Universo.[7]

En su artículo exponían escuetamente los detalles técnicos de su trabajo e informaban de su resultado: *Calculamos que la temperatura remanente en la antena es de 3.5±1.0 K (Kelvin) medida con 4 080 MHz (megaciclos por segundo)*.[8] Hoy en día se ha precisado este dato y se ha llegado a una temperatura de un poco menos de 3 K.

Es bastante irónico que muchos investigadores, tanto teóricos como experimentales, se habían encontrado antes con este fenómeno, pero nunca lo consideraron. En parte porque, como Steven Weinberg escribió, "en el decenio de 1950, se pensaba ampliamente que el estudio del Universo temprano no era del tipo de cosas a las que un científico respetable debía dedicar su tiempo". Todo cambió con el trabajo de Penzias, Wilson y Dicke.

La medida de la *radiación cósmica de fondo* (el ruido del telescopio de Holmdel, se dice ahora), combinada con el hallazgo anterior de Edwin Hubble sobre la recesión galáctica, le dio un gran impulso a la teoría de la *Gran Explosión*.

Arno Penzias y Robert Wilson recibieron el Premio Nobel de física en 1978 por estos estudios.

[7] R. H. Dicke *et al*., *Astrophysical Journal*, pp. 414-419.
[8] A. Penzias, R. Wilson, *Astrophysical Journal*, pp. 419-420.

1.10 UNA TERCERA EVIDENCIA

…Ya van a dar las 5 de la tarde. Lo sé porque escucho a alguien que apaciblemente se acerca, por los andadores de la Facultad de Ciencias, silbando un conocido son. Es Manuel Peimbert que pasa, frente a nuestro cubículo, a ver a sus alumnos.

El decenio de 1970 se caracterizó por la aparición de cálculos sobre la evolución química que debió sufrir el Universo, si se hubiera dado un evento tan espectacular como la Gran Explosión. La síntesis nuclear propiciada por tan intenso proceso debía manifestarse con la aparición de elementos más complejos que el hidrógeno. Entre otros, el astrónomo de la Universidad de Cornell, Robert V. Wagoner,[9] había obtenido varias posibles combinaciones de elementos químicos, las cuales dependían de las condiciones que predominaron durante los primeros minutos de la Gran Explosión.

A mediados de este decenio empezaron a llamar la atención los resultados de las observaciones, sobre las proporciones de elementos químicos, del astrónomo nacido en la ciudad de México, Manuel Peimbert Sierra (1941). Él mostró que las proporciones entre los elementos químicos,

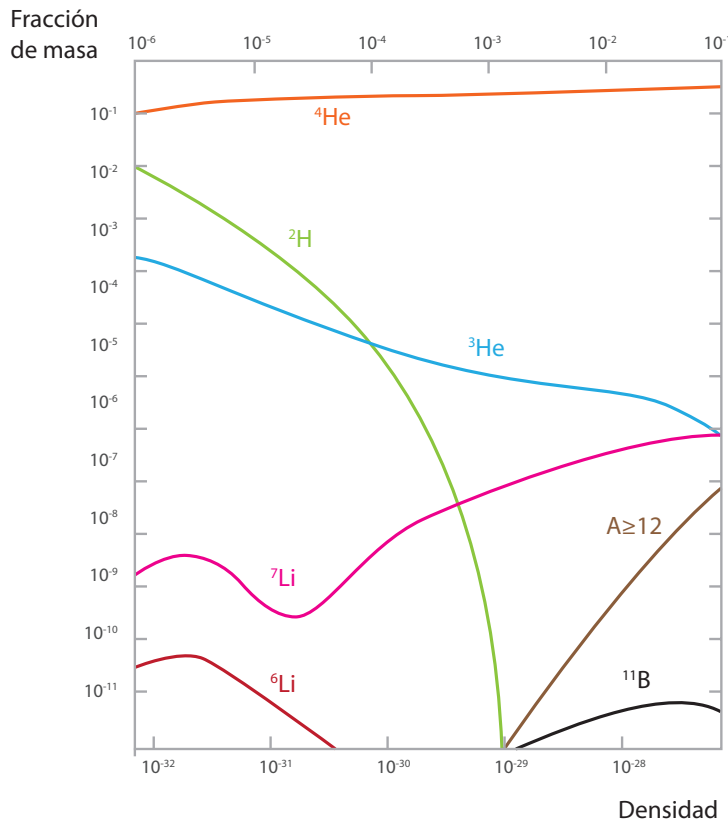


Figura 8. Fracción de masa como función de la densidad.

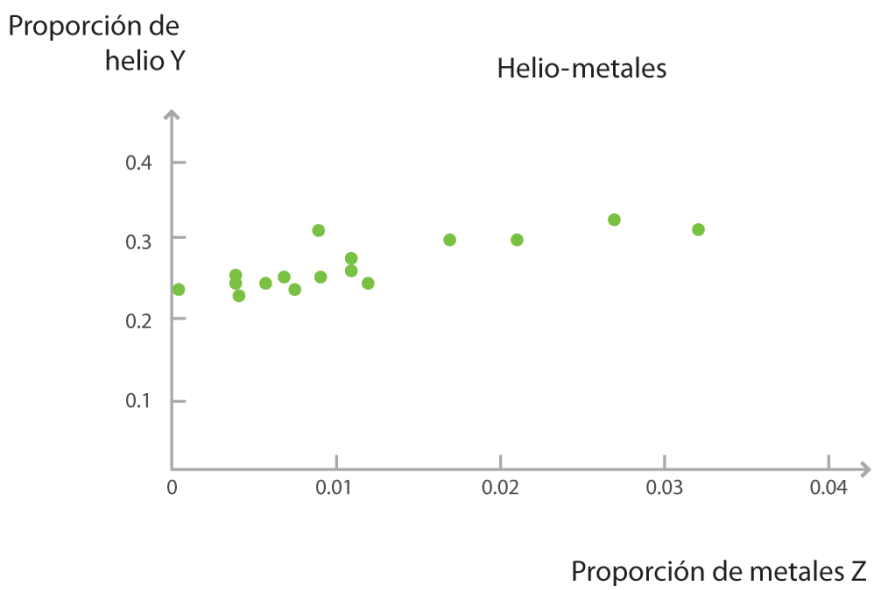


Figura 9. Proporción de helio como función de la proporción de metales.

[9] R. V. Wagoner, *Astrophysical Journal*, pp. 343-360.

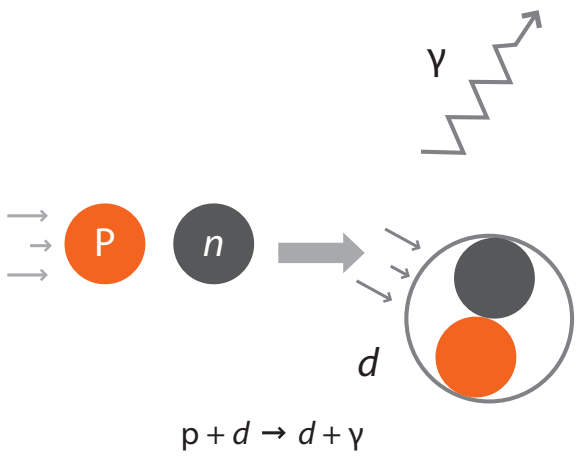


Figura 10. Reacción que da origen a los deuterones.

medidas en objetos celestes muy distantes, son similares (véase figura 9, p. 309), como si provinieran de un mismo proceso que actuó por todas partes.

$$p + n \rightarrow d + \gamma \text{ (figura 10).}$$

Cuando la temperatura cósmica era del orden de diez mil millones de grados Kelvin, las reacciones nucleares debían producir núcleos de deuterio *d*, a partir de neutrones *n* y protones *p*. Pero la temperatura era suficientemente alta para destruirlos enseguida, así que no era posible formar elementos más pesados. Al disminuir la temperatura del Universo, el deuterio se volvió estable y fue posible, a partir de reacciones nucleares de deuterio con protones, formar partículas de ^3He (se lee como helio 3); es decir, un isótopo de helio con dos protones y un neutrón. En las fórmulas, la letra griega γ denota la radiación que se emite junto con la producción de los núcleos.

$$p + d \rightarrow {}^3He + \gamma.$$

Por último, del ^3He y un núcleo de deuterio es posible producir un núcleo de ^4He, que está formado por dos protones y dos neutrones.

$${}^3He + d \rightarrow {}^4He + p.$$

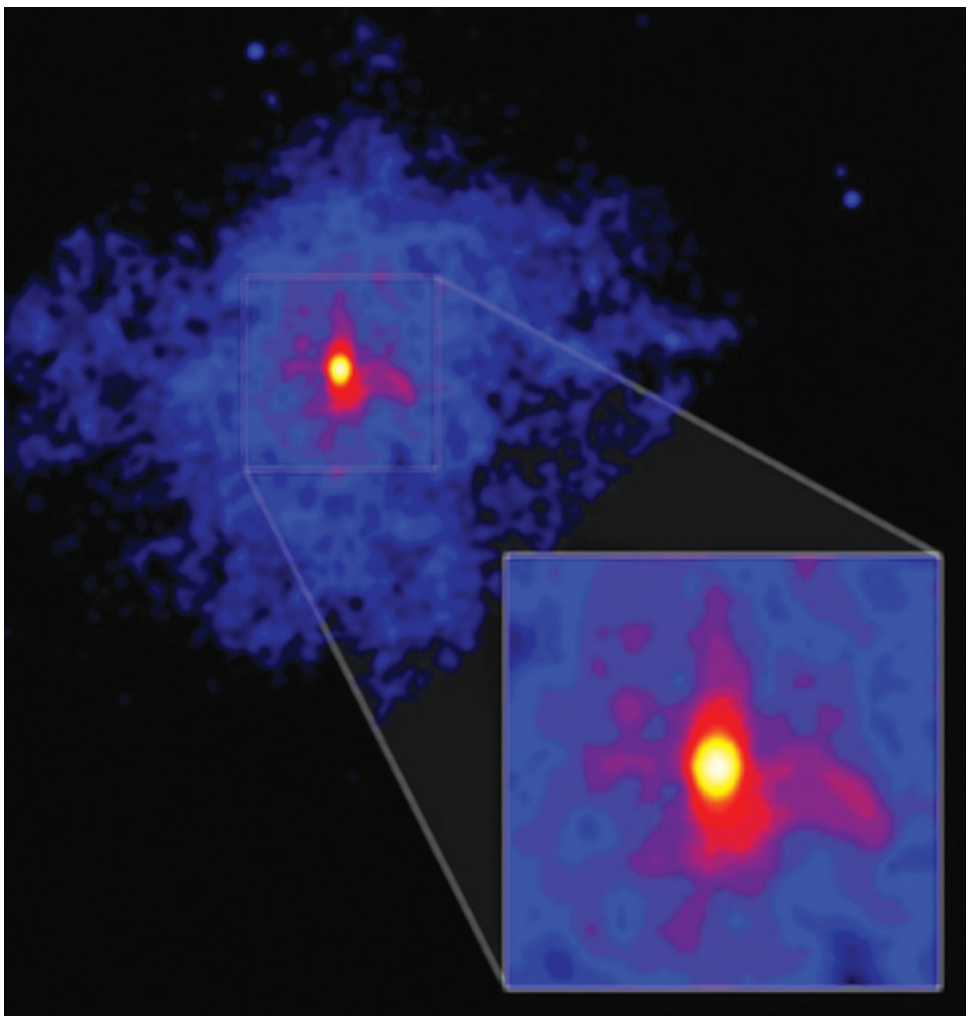
En la actualidad, la mayoría de los átomos de helio del Universo observable son ^4He. Después de formarse este isótopo, la temperatura y la densidad cósmicas disminuyeron lo suficiente como para que ya no fuera posible constituir más elementos pesados. Después de los primeros cuatro minutos, probablemente la temperatura disminuyó hasta unos 800 millones de grados Kelvin, por lo que las reacciones nucleares se detuvieron. Desde ese momento, la composición química del Universo se mantuvo constante, constituida fundamentalmente por hidrógeno, helio y pequeñas cantidades de deuterio y litio. En teoría, de acuerdo con el proceso de la Gran Explosión, la composición química no volvería a modificarse sino hasta la formación de las galaxias y las estrellas, dos mil millones de años después.

Manuel Peimbert y sus colaboradores, estudiando el espectro de emisión de las nebulosas, determinaron la composición química del medio interestelar. En nuestra galaxia y en otras en las que ha sido posible determinar con precisión su composición química, se ha encontrado que los seis elementos más abundantes son hidrógeno, helio, carbono, nitrógeno, oxígeno y neón. Las abundancias relativas de estos elementos en el medio interestelar se pueden obtener a partir de analizar el espectro de emisión de la radiación que emiten las remanentes de supernovas, las nebulosas planetarias y las regiones H II, llamadas así porque casi todo el gas está formado por hidrógeno ionizado. Podemos establecer la aportación de las estrellas al medio interestelar y restarla de la que medimos; así obtendremos la abundancia de los elementos que existían antes de formarse las galaxias.

Se pueden comparar las predicciones teóricas del modelo de la Gran Explosión, con las abundancias por unidad de masa de hidrógeno, deuterio, helio tres, helio cuatro y litio siete. Según lo indicaban Peimbert y sus colaboradores:

> Se ha encontrado que todas las galaxias, con una buena determinación de abundancias, se formaron con aproximadamente 25% de helio y 75% de hidrógeno por unidad de masa, coincidiendo con los resultados de la teoría.[10]

A este 25% de helio se le conoce como *helio pregaláctico* o *primordial*.



Ampliación de una galaxia.

Si se hubiera tratado de una explosión lenta, la temperatura permanecería alta por más tiempo, produciendo mayor abundancia de elementos pesados, particularmente helio. Una explosión rápida tendría una abundancia mucho menor de estos elementos. De esta forma, la abundancia nos indica el tipo de explosión y, más aún, nos permite calcular la edad del Universo. Quizá la parte más impactante del trabajo del científico mexicano es que la abundancia del helio primordial por unidad de masa obtenida observacionalmente coincide con la predicha por la teoría estándar de la Gran Explosión y reproduce el valor de la constante de Hubble con este procedimiento insólito, confirmando los resultados surgidos de la recesión galáctica (figura 11).

[10] Manuel Peimbert, "Evolución química del Universo", *Temas selectos de astrofísica*, pp. 307-331.

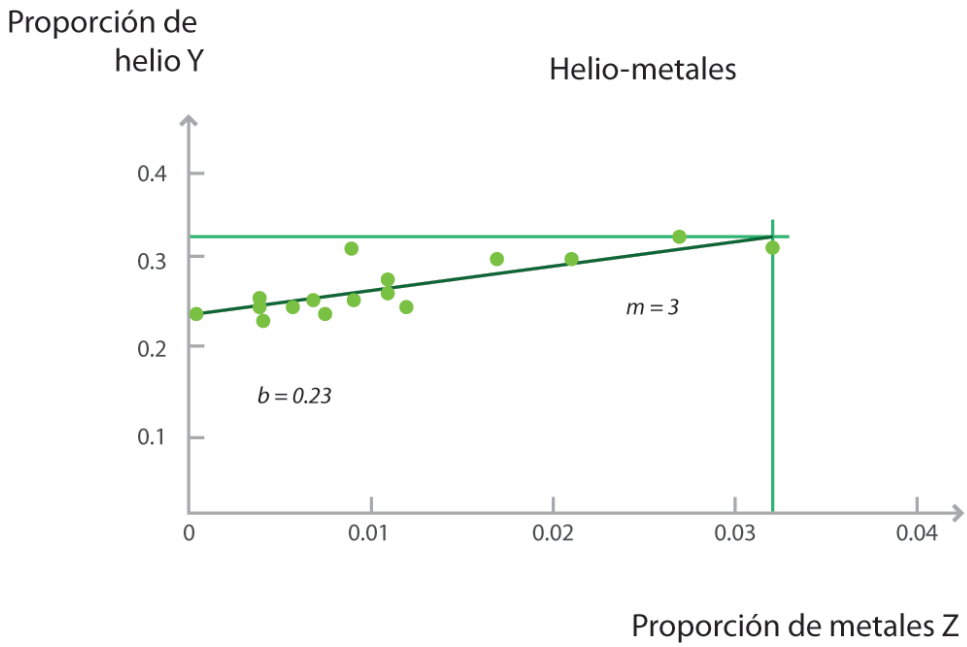


Figura 11. Proporción de helio contra la proporción de metales. A partir de la gráfica se puede saber que la proporción de helio es alrededor de 25 por ciento.

Este trabajo, paciente y meticuloso, de casi cuatro decenios se ha convertido en uno de los pilares en que descansan los modelos que asumen la existencia de la Gran Explosión para explicar la estructura actual del Universo observable. De esta manera, Manuel Peimbert encontró, después de la recesión galáctica y la radiación fósil, una tercera evidencia a favor de la Gran Explosión.

> Recuerdo una ocasión en que estábamos discutiendo un problema. Mi amigo Sergio estaba parado frente al pizarrón. Manuel se asomó para saludarnos, vio la guitarra que estaba sobre el archivero y, sin comentar nada, entró al cubículo, tomó la guitarra, se sentó sobre el escritorio con las piernas cruzadas, y empezó a cantar el son veracruzano "La bruja". Ahí estaba, entre nosotros, como cualquier estudiante, con su suéter sobre la espalda, quien era ya en aquel momento un investigador de renombre mundial en el ámbito de la astronomía.
>
> *Jorge Daniel Marroquín de la Rosa*

MECÁNICA

TEMA

2

© *Latin Stock México.*

INTRODUCCIÓN

La mecánica estudia el movimiento de los cuerpos, es decir, sus posiciones relativas en el espacio al transcurrir el tiempo. La mecánica se divide en tres partes: cinemática, dinámica y estática.

La mecánica se ha estudiado desde hace varios siglos por muchos científicos, destacando dos de ellos por la trascendencia de sus aportaciones: el italiano Galileo Galilei y el inglés Isaac Newton.

Galileo introduce el método experimental en el estudio del movimiento de los cuerpos; asimismo, introduce el lenguaje matemático y muestra que es el apropiado para el

trabajo científico. Con estas propuestas estudió el movimiento de caída de los cuerpos y llegó a plantear la posibilidad del movimiento rectilíneo uniforme ante la ausencia de fuerzas sobre un cuerpo inicialmente en movimiento.

Galileo apoyó la propuesta de Nicolás Copérnico de un sistema heliocéntrico para el sistema formado por los pocos planetas conocidos hasta entonces. Como es sabido, divulgar esas observaciones le costó ser enjuiciado por la Santa Inquisición pero, afortunadamente, no le costó la vida. Fue a partir de Galileo que los científicos de todo el mundo y las personas de los sectores instruidos de las diversas sociedades, empezaron a aceptar que la Tierra, nuestro planeta, no es el centro del Universo. Esto representó un cambio trascendente en la percepción que del Universo tenían las personas y, en particular, un cambio en la percepción de nuestra posición y tamaño en él.

Uno de los principales aportes de Galileo fue la construcción de un telescopio con el que observó manchas en el Sol, cráteres y montañas en la Luna, así como el movimiento de algunos satélites naturales de Júpiter.

Newton, con sus tres leyes sobre el movimiento y su ley de la gravitación universal, explicó el movimiento de los cuerpos, tanto sobre la superficie terrestre como en el espacio. Sus aportaciones, además de explicar la interacción entre el Sol, planetas, satélites y cometas, han sido la base para el cálculo de la puesta en órbita de diversos satélites artificiales ya sea para las telecomunicaciones o para el uso meteorológico, para realizar exitosamente viajes tripulados a la Luna y, recientemente, enviar sondas a Marte.

A lo largo de este capítulo se analizará la importancia de las interacciones para el estudio del movimiento, haciéndose énfasis en los principales aspectos de las leyes de Newton y de la gravitación universal para explicar el movimiento de los cuerpos, los tipos de movimiento y la relación con la energía como aspecto importante para el estudio de los fenómenos físicos.

2.1 LA IDEA DE MOVIMIENTO

¿Qué es el movimiento y cómo lo percibimos?

El concepto de movimiento implica el cambio de posición o desplazamiento de un objeto y en él están implícitas las nociones de espacio y tiempo. Sin duda, el estudio del movimiento ha sido importante en todas las épocas y culturas.

El interés por estudiar el movimiento de los cuerpos está asociado a la supervivencia del hombre mismo. El hallazgo arqueológico de armas y herramientas de piedra, encontrados en Zhoukoudian, China, en 1921, por el geólogo sueco Gunnar Anderson, a las que se les estima una antigüedad de 500 000 años, permite imaginar a los cazadores prehistóricos armados con lanzas y rodeando al mamut para cazarlo.

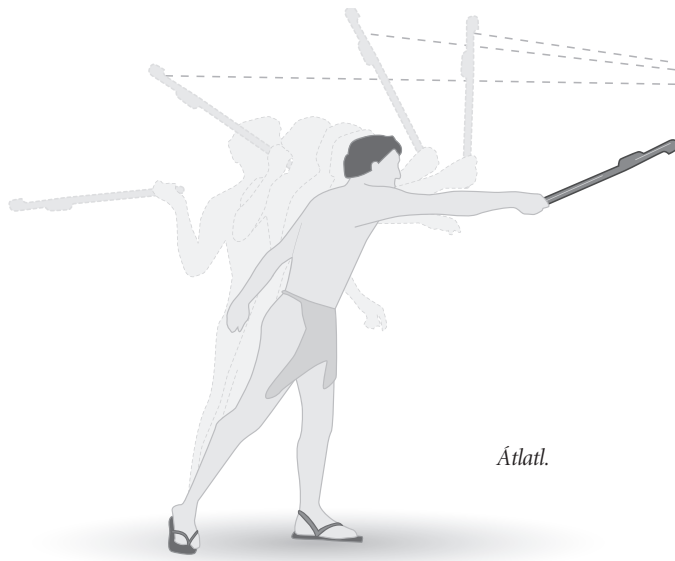
También sería posible pensar en las celebraciones después de una buena caza, tal vez con bailes y concursos entre los mejores lanzadores, quienes, luciendo su pericia aprendida y perfeccionada a lo largo de su vida, tenían clara la importancia de sus habilidades para la supervivencia del grupo.

Seguramente el hombre primitivo observó el movimiento de un sinnúmero de animales tan diferentes como el mamut y las aves, y diseñó diferentes instrumentos y técnicas específicas para cazarlos. En algunas culturas los pájaros eran cazados con cerbatanas, dado que la presa es muy rápida y pequeña; si el cazador no disparaba con puntería sólo la hería y podía escapar; entonces idearon ponerle veneno de acción rápida, como el curare, a los dardos. Con el paso del tiempo, mejoraron sus utensilios y fabricaron un instrumento más sofisticado llamado *átlatl*, que era un lanza-dardos con mayor potencia y alcance,

La caza de mamut | © Latin Stock México.

permitiéndoles cazar presas más rápidas y relativamente grandes. De este modo, el éxito en la cacería dependía del movimiento de los proyectiles y la habilidad del cazador.

Así, el éxito en la cacería dependía del movimiento de tres objetos: los animales, la habilidad del cazador y los proyectiles.



Átlatl.

2.1.1 El movimiento de los astros

Existe otro tipo de movimiento, el de los astros en el cielo. En la Antigüedad todos los astros estaban relacionados con un origen divino. Dado que el hombre era incapaz de influir sobre este movimiento, sólo se restringió a observar y registrar. El principal astro de observación fue desde luego el Sol, aunque los movimientos de la Luna y Venus también desempeñaron un papel importante.

Producto del conocimiento adquirido a través de muchos años de observación, se determinaron las estaciones del año, lo que fue de enorme utilidad para el desarrollo

de la agricultura, pues fue posible determinar con bastante exactitud las etapas del proceso agrícola. Pero, ¿qué significa "el movimiento del Sol" y en relación con qué?

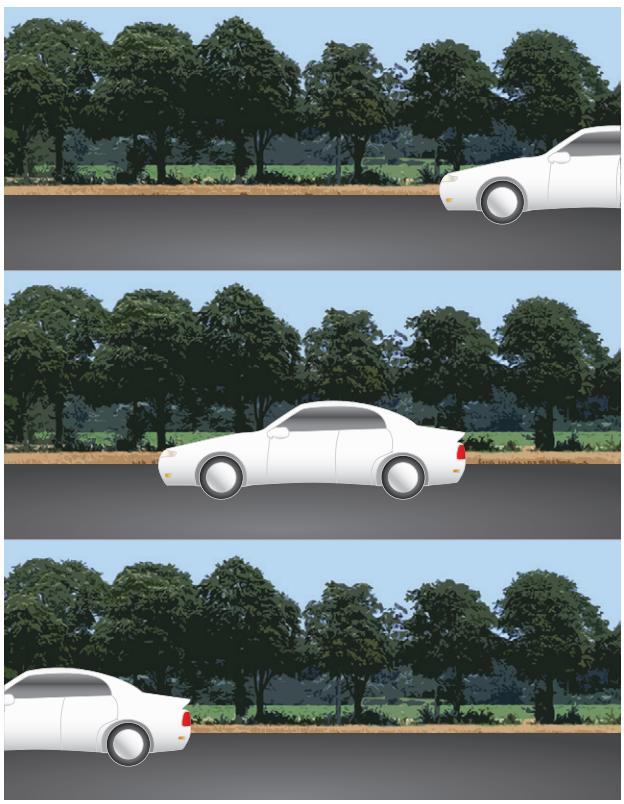
Para empezar, ¿cuántos movimientos del Sol se pueden distinguir? Un movimiento obvio es el que empieza con el amanecer y termina en el atardecer y el otro movimiento está asociado con las estaciones del año.

El primer movimiento es el que tiene la Tierra alrededor de su propio eje y produce el día y la noche. El segundo es el movimiento de la Tierra alrededor del Sol y da origen a las estaciones del año. Llevó mucho tiempo aceptar que el movimiento de la Tierra y los planetas se describía más fácilmente si se asumía que ellos giraban alrededor del Sol. La aseveración de que es la Tierra la que se mueve y no el Sol, le costó la vida a Giordano Bruno (1548-1600) y casi sucede lo mismo con Galileo Galilei (1564-1642). [Véase animación en CD, "El Sol, la Tierra y la Luna".]

2.1.2 Percepción sensorial del movimiento

¿Cómo sabemos que algo se mueve? ¿Cuáles sentidos son los que están implicados más directamente en la percepción del movimiento?

Si se viaja en un autobús y nos asomamos por la ventana, se ve que el paisaje va cambiando. Después de un buen rato, el paisaje será diferente. Alguien parado junto al camino podrá decir que es el autobús lo que se mueve. Sin embargo, el pasajero, cómodamente sentado, oyendo música con sus audífonos o leyendo, podría decir que lo que se mueve es el paisaje. Otro ejemplo de la percepción del movimiento sería el observar la constelación de Orión durante un par de meses, siempre a la misma hora. Conforme pase el tiempo, la posición en el cielo de esta constelación iría cambiando.



Cuadros de automóvil en movimiento.

En cada caso, se dice que algo se mueve porque vemos que ha cambiado su posición, medida como una distancia o un ángulo respecto a un objeto que se toma como referencia, que bien puede ser un punto, el pico de una montaña, una marca sobre el piso, el horizonte, etcétera.

Pero, si un objeto se mueve en la oscuridad no lo veríamos pues para que podamos observar el movimiento de los objetos, éstos deben tener luz propia (ser una fuente luminosa) o deben ser iluminados.

Un cuerpo con luz propia que todos los días vemos moverse es el Sol, y un cuerpo iluminado, que también miramos en el firmamento, es la Luna. Sólo podemos ver los cuerpos moverse si tienen luz propia o son iluminados por una fuente luminosa natural o por una fuente luminosa artificial.

Si pensamos en otros ejemplos diferentes, la tinta y el café soluble en agua o el petróleo en el mar se mueven azarosamente en el líquido, impulsados por el incesante aunque invisible movimiento microscópico de las moléculas del agua; tal es el caso del perfume. El aroma del perfume se percibe porque sus partículas, aunque no las veamos, se mueven entre las del aire.

Fuentes luminosas: linterna, Sol, luciérnagas | *© Latin Stock México.*

Al proceso mediante el cual las partículas de una sustancia se mueven en algún medio como agua o aire, dispersándose, se le llama *difusión* y fue brillantemente explicado por Albert Einstein en 1905, asumiendo que las moléculas que constituyen el medio golpean sin cesar a las partículas agregadas, obligándolas a moverse azarosamente.

El movimiento no sólo se puede ver y oler, sino también escuchar, como por ejemplo el ruido del motor de un auto, el golpe de un objeto al caer, el trueno de una descarga eléctrica en la atmósfera, etc. En los fenómenos donde se manifiesta el sonido, debe haber algo que lo produce y algo que lo capta.

Dos ejemplos que nos dejan ver con claridad el origen del sonido son: el repique de campanas y el rasgar las cuerdas de una guitarra. Tanto al rasgar la cuerda de una guitarra, como al golpear con el badajo el cuerpo principal de una campana, se genera un movimiento vibratorio que produce el sonido que percibimos. Entonces, puede decirse que: *cuerpos vibrando producen perturbaciones en el medio que los rodea*, lo que en los casos anteriores es el aire.

¿Cómo es que se perciben las vibraciones generadas como sonido? La vibración perturba al medio que rodea a la fuente vibratoria; este medio puede ser un gas como el aire, un líquido o un sólido. La perturbación se propaga en el medio y llega a los oídos, golpea al tímpano, lo hace vibrar y nuestro cerebro lo interpreta como sonido. Si no hubiera algún medio, el sonido no se podría propagar.

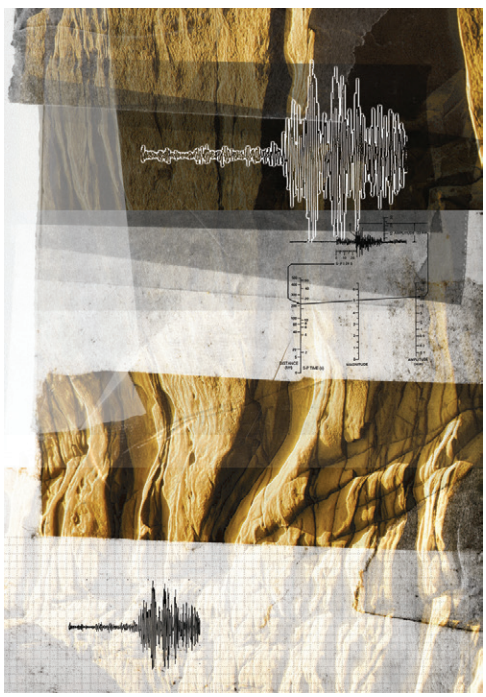
El físico y químico inglés Robert Boyle (1627-1691) diseñó un experimento que hoy se puede hacer con una bomba de vacío, la que no difiere mucho de una aspiradora casera. Se introduce en un frasco transparente de vidrio, una fuente de sonido, que puede ser un radio o un despertador, después se cierra herméticamente el frasco y se le extrae el aire, mientras la fuente de sonido continúa trabajando. El sonido disminuye conforme se extrae el aire, que al regresar al frasco permitirá oír nuevamente el sonido.

Con este experimento se comprueba que *el sonido es el resultado de una perturbación* que necesita de un medio para desplazarse, el aire en este caso. Entonces, *la perturbación en el aire que percibimos como sonido es un movimiento que podemos detectar.*

El oído y los instrumentos que se han desarrollado para ampliar sus facultades (micrófonos y audífonos, junto con los amplificadores) son importantes detectores de un tipo de movimiento, que por lo general es difícil percibir de otra manera.

El movimiento también es posible percibirlo por medio del sentido del tacto. Se siente el movimiento del aire, el agua y la tierra. El viento es aire en movimiento que se percibe en el rostro y el resto del cuerpo como el vaivén del agua en una alberca; o como en el caso de un *tsunami*, que es una pared de agua de 10 metros de altura y varios kilómetros a lo largo de la costa que puede arrasar grandes poblaciones, como lo hizo en 2004. Éste dejó casi 300 mil víctimas en Indonesia, Tailandia, Bangladesh, India, Sri Lanka, las Maldivas e incluso Somalia, al este de África.

Tsunami, 2004 | © <*http://es.wikipedia.org/wiki/Archivo:2004-tsunami.jpg*>.



Esquema ondulatorio de un terremoto y efectos de esa ondulación | © *Latin Stock México.*

Otros movimientos que sentimos y nos impactan son los de tierra: pueden ser de origen local como cuando un cuerpo pesado golpea contra el piso o un tráiler pasa cerca; pero también más devastadores, como los terremotos, que desde hace mucho tiempo son una de las grandes razones por las que ha sido preciso estudiar los movimientos de la naturaleza.

El movimiento se ve, se escucha, se siente y hasta se huele. Se sabe del movimiento por medio de los sentidos y, cuando éstos son insuficientes para proporcionar información satisfactoria, se hace uso de aparatos cada vez más complejos, de mayor alcance, rapidez y precisión. En particular, los movimientos muy rápidos y los demasiado lentos han sido estudiados apoyándose en dispositivos técnicos que fueron diseñados y desarrollados con base en los avances del conocimiento científico.

También es posible percibir el paso del tiempo a través de la sucesión del día a la noche; de hecho, desde tiempos inmemoriales la humanidad ha sabido que el cambio de longitud de la sombra de un objeto indica la hora del día y que la sombra se acorta hacia el mediodía y se alarga hacia el atardecer. Así, el primer reloj de Sol consistía simplemente en una estaca clavada en el suelo.

El más antiguo reloj de Sol conocido, encontrado en Egipto, estaba dividido en doce partes. Pero este reloj tenía un obvio inconveniente: no funciona en días nublados, ni en la noche. Entonces, aparecieron los relojes de agua, que consistían simplemente de un recipiente con un orificio pequeño, por el cual salía el agua. Guillaume Amontonas (1663-1705) construyó uno de estos relojes. Otro reloj similar es el de arena, que funciona de forma parecida al del agua.

Reloj de Sol, reloj de arena, reloj de péndulo y cronómetro | © Latin Stock México.

Otra opción que tuvieron los científicos, antes de la aparición de los relojes que conocemos, fue el conteo de sus propias pulsaciones mientras ocurría algún evento. También podían contar las oscilaciones de un péndulo, lo cual dio origen a relojes más modernos, en los cuales éste hace girar engranes, que a su vez mueven las manecillas.

2.1.3 Marcos de referencia del movimiento

Como ya se ha mencionado, para describir el movimiento es necesario tener una referencia. Es muy difícil saber lo que ve, oye o siente otra persona y notar que el mundo que percibimos

puede ser diferente, es decir, que el mundo que vemos depende desde dónde lo vemos, esto es, depende del punto de referencia. ¿Nos alejamos o nos acercamos? ¿Estamos arriba o abajo?

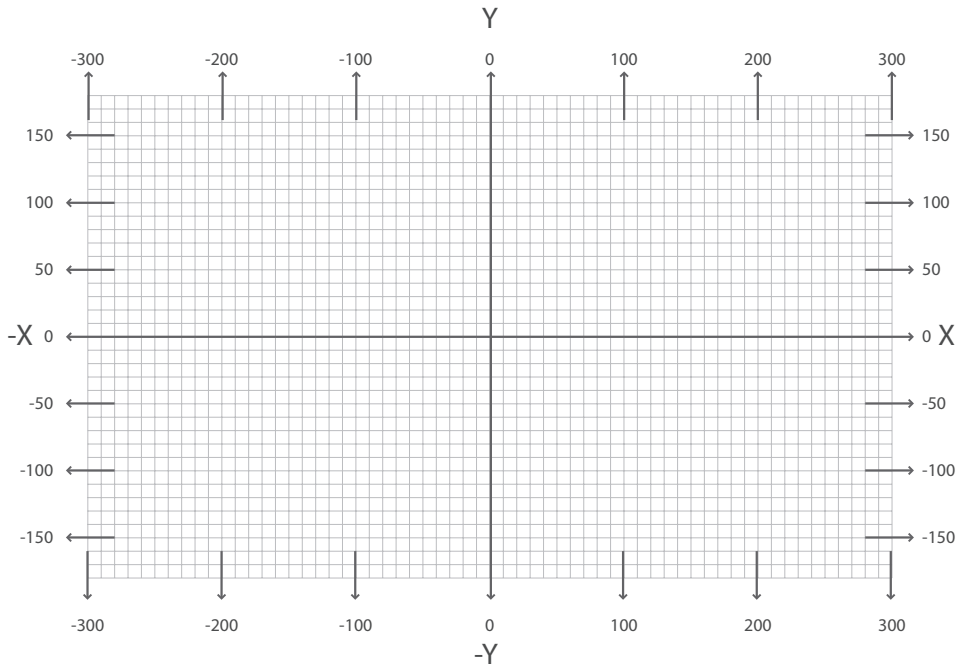


Figura 1. Plano cartesiano.

Para responder a estas preguntas debemos indicar respecto a qué o desde dónde observamos. Si no se especifica el punto de referencia en cada caso la respuesta puede ser diferente, y aun así correcta. Incluso podemos llegar a descalificar a la otra observación, porque simplemente tiene otra referencia.

Sin embargo, hablando de movimiento, el que un cuerpo se mueva quiere decir que pasa de un lugar a otro; esto es, que primero ocupaba una posición y luego otra; sin perder de vista que las posiciones y los cambios de posición son respecto a una referencia.

En física, como se verá más adelante, al estudiar el movimiento usamos el concepto de marco de referencia, que es un esquema geométrico compuesto por dos rectas perpendiculares, sobre las cuales se pueden ubicar distancias; el origen de esas distancias es el punto donde se cruzan las dos rectas (figura 1). Este esquema fue ideado por el científico francés René Descartes (1596-1650).

Para ilustrar lo anterior se puede pensar en un partido de futbol en donde el movimiento de los jugadores y el balón durante el juego se indica tomando como referencia las características de la cancha y sus divisiones: área chica, área grande, medio campo, círculo central; línea de meta, banda lateral, esquinas y manchón de penalti (figura 2).

Con estos conceptos se pueden relatar las jugadas de forma más precisa. "El balón de tiro de esquina fue tomado *de palomita* por el centro delantero en el área chica, a la altura del manchón de penalti y salió un cabezazo al ángulo superior derecho del poste contrario que dejó parado al portero. ¡Fue un golazo!" Como lo narraría un buen cronista.

Si se quiere describir la jugada anterior sin emplear las palabras "esquina", "área chica", "manchón de penalti", "ángulo superior derecho", "poste contrario", no se puede. Cada una de ellas son referencias que nos dan una idea del movimiento del balón; sin ellas no se podría saber su localización, ni por dónde se movió. No se puede dar una descripción de

Figura 2. Campo de futbol.

lo que llamamos *trayectoria.* Además, en la descripción de la jugada, también se usaron los términos de "*cabezazo*" y "*dejar parado al portero*". Ambos conceptos tienen que ver con la *rapidez* de la acción y, con ellos, se complementa la descripción del movimiento del balón, el cual se compone de dos conceptos: trayectoria y velocidad.

Además, para dar una descripción más exacta de la jugada, se requiere saber ¿dónde estaban precisamente los demás jugadores y cómo se movían?, ¿en qué dirección y sentido?, ¿lenta o rápidamente?, así como una descripción precisa del movimiento del balón. Es decir, se requiere de una descripción geométrica del movimiento.

Entonces, para la descripción geométrica del movimiento se requiere de un *marco de referencia*, dentro del cual pueda definirse con precisión una *trayectoria*, que es la línea que marca la *posición* de un objeto en el espacio durante su recorrido.

En el ejemplo del campo de futbol hay lugares que pueden especificarse fácilmente, como el centro del círculo central; mientras que los manchones de penalti necesitan adicionalmente que se especifique si es el de uno u otro lado de la cancha; las esquinas de las áreas chica y grande también requieren de referencias adicionales; las cuatro esquinas del campo demandan que se diga el lado de la cancha a que pertenecen y, además, si es la superior o la inferior. Así que si el balón o el jugador se hallan sobre cualquiera de los lugares anteriores, su posición queda bien determinada.

Pero, ¿qué sucede en cualquier otro lugar?

Si se recurre a un marco de referencia cartesiano, un punto cualquiera sobre el terreno queda determinado por un par de números; a éstos les llamamos *coordenadas*. A la línea horizontal le llamamos *eje x* y *eje y* a la línea en el plano perpendicular a ella. Con esto, la localización de un punto cualquiera se denota por un par de números dentro de un paréntesis *(x, y)*. De este modo, ya se puede determinar la posición del balón y de los jugadores sobre la superficie, pues para ello simplemente se dan las coordenadas *(x, y)* de cada uno de los objetos (figura 3, p. 322). [Véase animación en CD: "Jugada de futbol 2D".]

Sin embargo, falta la altura que lleva el balón. Esto se puede representar asignando un tercer número que indique la altura sobre el plano, representado por la letra *z*. Es claro que el movimiento del balón ocurre en un espacio de *tres* dimensiones y, por lo mismo que antes, se requieren *tres* números para fijar su posición. El espacio de tres dimensiones es en

el que nos movemos. De esta forma el balón y los jugadores pueden localizarse simplemente dando las coordenadas de cada uno, según los tres números *(x, y, z)*. [Véase animación en cd: "Jugada de futbol 3D".]

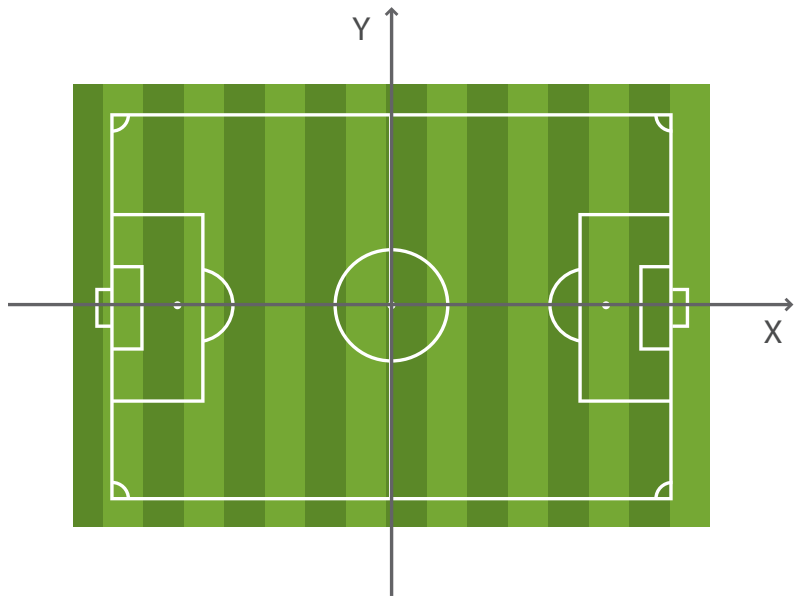


Figura 3. Cancha de futbol con ejes cartesianos.

Un último aspecto es el que tiene que ver con la rapidez con que se mueve el balón. Para esto se debe introducir el tiempo en la descripción del movimiento y definir los conceptos de *rapidez* y *sentido*.

Para introducir el tiempo en la descripción geométrica del movimiento tendríamos que asignar a cada punto un tiempo, que representamos por la letra *t;* ahora los *cuatro* números *(x, y, z, t)* marcan la posición de un evento en un espacio de *cuatro* dimensiones. Este espacio es el *espacio de eventos* de cuatro dimensiones que describe completamente el movimiento del objeto.

2.1.4 Sistema Internacional de Unidades

Desde las civilizaciones más antiguas (sumerios, egipcios, persas o mayas), el ser humano ha establecido unidades de medida para diversas magnitudes de manejo cotidiano como longitudes, tiempos, pesos, superficies y volúmenes. Esas unidades de medida han sido muy diversas, variando en cada lugar y época.

Con el avance de las comunicaciones y contactos entre los diferentes pueblos del mundo, que permitieron un mayor intercambio de mercancías, tecnologías y conocimiento científico, se evidenció la necesidad de acordar un sistema universal de unidades de medida. El actual Sistema Internacional de Unidades (SI) fue adoptado de forma oficial en 1960, por la Conferencia General de la Oficina Internacional de Pesas y Medidas.

Las medidas de tiempo y longitud mencionadas bastan para movernos con soltura en la vida cotidiana, ya sea para llegar a tiempo a una cita, comprar hilo, tela, etc., o describir el movimiento del balón en el espacio de eventos. Una trayectoria en el espacio de eventos de cuatro dimensiones describe completamente el movimiento del objeto, nada más que es difícil de visualizar.

2.1.5 Rapidez y velocidad

Vectores

Cuando se dice que un automóvil viaja a "ciento veinte kilómetros por hora", se está considerando una *cantidad escalar,* que en este caso es la rapidez. Para expresar una cantidad escalar, basta con dar su tamaño y las unidades en que se mide, por ejemplo, *25°C, 45 s, 30 m, 3 kg,* etcétera.

Ahora bien, para expresar la velocidad es necesario especificar su orientación, es decir, hacia dónde se dirige; por ejemplo, 120 km/h hacia el sur. A las cantidades que tienen magnitud y orientación se les llama *cantidades vectoriales*. La velocidad es una cantidad vectorial, así como el desplazamiento y la fuerza.

El *desplazamiento* de un balón de la posición inicial x_i a la posición final x_f se denota por $\Delta x = x_f - x_i$ (esta ecuación se lee "delta equis *es igual a equis 'f' menos equis 'i'*", donde Δ es la cuarta letra mayúscula del alfabeto griego y generalmente se usa para denotar el cambio en una variable, en este caso, la posición x).

El desplazamiento, al igual que cualquier cantidad vectorial, tiene dos características: tamaño o *magnitud* (también conocida como el valor absoluto, pues no se toma en cuenta el signo y se escribe como $|x_f - x_i|$); y orientación (que es positiva si $x_f > x_i$, es decir, que la posición final es mayor que la posición inicial y negativa si $x_i > x_f$). De esta forma, el desplazamiento de un cuerpo posee un determinado tamaño y dirección.

Geométricamente un vector se representa con una flecha. Si escribimos una cantidad vectorial, le agregamos una flecha arriba de la letra: al desplazamiento lo escribimos como $\Delta\vec{x}$ y a la velocidad por $\Delta\vec{v}$. En el caso del vector desplazamiento, el origen de la flecha se ubica en la posición inicial x_i y su punta en la posición final x_f.

Vector.

Para calcular la rapidez con la que se mueve un objeto se toma el cociente de la magnitud del desplazamiento entre el lapso de tiempo $\Delta t = t_f - t_i$:

$$v = \frac{|x_f - x_i|}{t_f - t_i} = \frac{|\Delta x|}{\Delta t},$$

donde v (sin flecha) denota la rapidez o magnitud de la velocidad.

Observamos que para un intervalo Δt fijo, la rapidez varía directamente proporcional con la distancia recorrida Δx. Si, por ejemplo, un carro lleva el doble de la rapidez, recorre el doble de distancia en el mismo intervalo de tiempo Δt. Si ahora fijamos la distancia recorrida Δx, entonces, la rapidez varía inversamente proporcional con el intervalo de tiempo Δt. De esta forma, si el carro lleva el doble de la rapidez, recorre la misma distancia Δx en la mitad del intervalo de tiempo Δt.

2.2 MOVIMIENTO RECTILÍNEO UNIFORME

Si calculamos $\Delta x / \Delta t$ para cualquier intervalo de tiempo y obtenemos el mismo cociente, esto indica que el carro se movió con la misma rapidez en todo el intervalo; a este tipo de movimiento se le llama *movimiento rectilíneo uniforme.* Si para cualquier intervalo de tiempo que se seleccione, su correspondiente intervalo de desplazamiento Δx vale siempre cero, pues x_i y x_f valen lo mismo, la rapidez vale cero en todo ese intervalo de tiempo. A este estado de movimiento se le llama *reposo* (figura 4).

En la gráfica del desplazamiento contra el tiempo, la rapidez está directamente relacionada con la inclinación o pendiente de la recta *"x en función de t"*. A mayor inclinación, mayor rapidez.

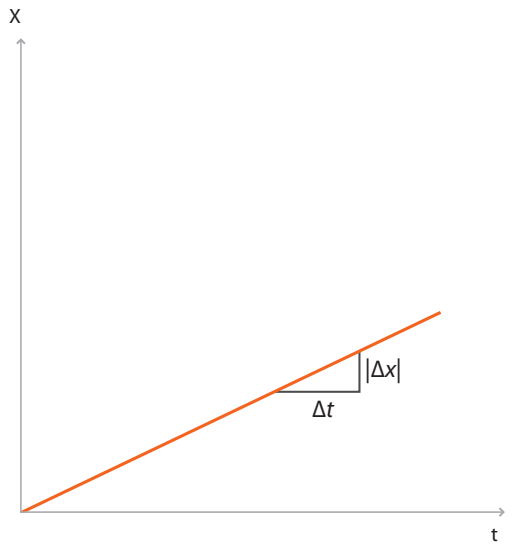


Figura 4. Desplazamiento contra tiempo, x contra t.

2.3 MOVIMIENTO OSCILATORIO. MOVIMIENTO ONDULATORIO TRANSVERSAL Y LONGITUDINAL

Una onda es un movimiento oscilatorio que se propaga en un *medio*. En las ondas, como las que se producen en un estanque donde se ha lanzado una piedra, el movimiento oscilatorio del agua es perpendicular o transversal a la dirección en que la onda se desplaza; a éstas se les denomina *ondas transversales*. Esto mismo sucede en una onda transversal producida en un "resorte de gusano".

En las llamadas *ondas longitudinales*, que también pueden transmitirse en un "resorte de gusano", el movimiento de un paquete de espiras es hacia adelante y hacia atrás, en la misma dirección en que longitudinalmente ocurre la propagación de la onda.

En los casos de ondas transversales y longitudinales, las partículas del medio (del agua o del resorte), que oscilando producen la onda, no se trasladan con ella. Lo que la onda transporta es una *señal*, que es precisamente el movimiento oscilatorio local del medio (figura 5).

La *ola* en un estadio de futbol nos permite ver lo dicho con más sencillez y claridad. Como dato curioso, la ola fue inventada por aficionados mexicanos, durante el campeonato mundial de futbol en 1986 en nuestro país. A una señal, un grupo de aficionados cercanos entre ellos, situados sobre una hilera de asientos, se levanta del asiento levantando al mismo tiempo los brazos y, en seguida, se sientan bajándolos.

Con un movimiento oscilatorio arriba-abajo, los vecinos de uno de los lados, por ejemplo del derecho, inician con un ligero retraso su propio movimiento arriba-abajo consi-

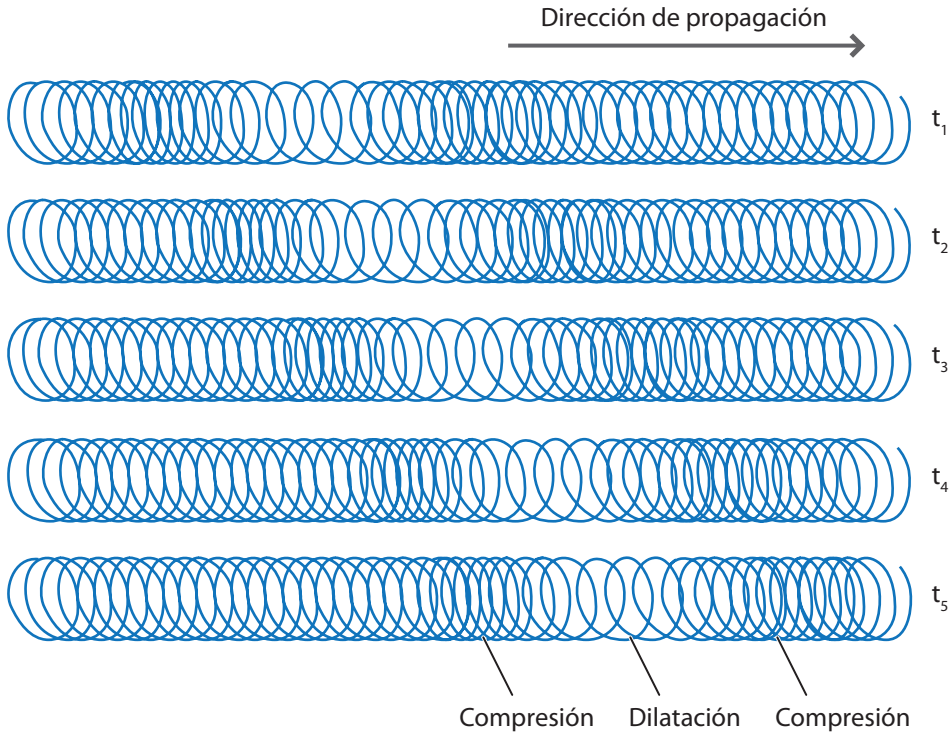


Figura 5. Ondas longitudinales en un "resorte de gusano".

guiendo, con esto, estar completamente de pie una fracción de segundo después de que el vecino lo hizo. Y así para cada hilera, lo que se ve en el estadio es la propagación de la señal inicial; se ha producido *la ola*.

En algunos aspectos, un *tsunami* y un terremoto son olas como las del estadio. Pero, mientras que la ola en el estadio de futbol no ocasiona daños, un tsunami o un terremoto son capaces de provocar una catástrofe.

La ola | © Latin Stock México. [Véase video en CD: "La ola en el estadio".]

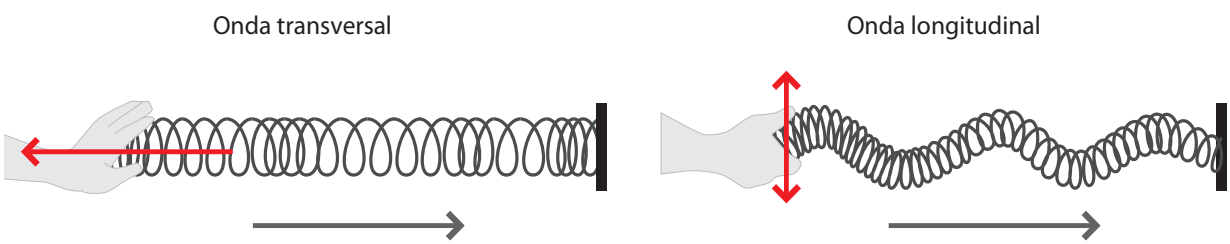
En el caso de *la ola*, la señal o perturbación que se propaga es el movimiento oscilatorio arriba-abajo de las personas, pero ellas no avanzan con la onda, pues se quedan en su lugar. Podemos decir que, en este caso, la ola es la propagación de una perturbación sin transporte de materia.

Ahora bien, la importancia del movimiento ondulatorio radica precisamente en que éste es capaz de transmitir una señal, que a su vez es capaz de producir un efecto físico en algún receptor. Es el caso del oído, que al captar las compresiones del aire, producidas por las ondas, hace que las percibamos como *sonido*. El sonido nos trae señales que contienen *información*, como las palabras o la música. En los tsunamis y terremotos, desgraciadamente, el objeto receptor de la señal es el ambiente y las poblaciones.

Otro tipo de ondas, como las electromagnéticas, transportan señales que son la base de las telecomunicaciones y son indispensables para nuestra civilización. Actualmente no se puede pensar en un mundo sin televisión, teléfonos celulares, radio o Internet. Sin las ondas electromagnéticas tampoco habríamos podido generar nuestra concepción del Universo, pues la información de su estructura nos llega mediante este tipo de ondas. Lo mismo podemos decir en relación con el conocimiento actual del mundo atómico y subatómico.

Una descripción más detallada de las características del movimiento ondulatorio tanto para ondas longitudinales como transversales, tomando como ejemplo al "resorte de gusano", es la siguiente:

Figura 6. Ondas en el "resorte de gusano". [Véase video en CD: Ondas transversales y longitudinales".]

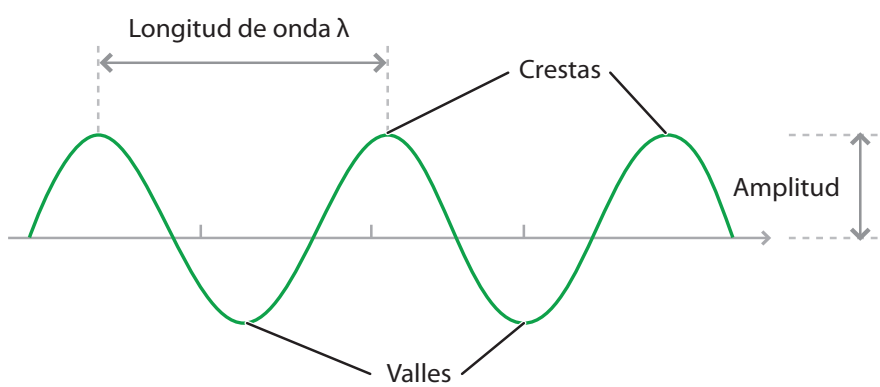


Ondas longitudinales son aquéllas en que las partículas del medio oscilan en la dirección en que avanza la perturbación.

Ondas transversales son aquéllas en que las partículas del medio oscilan en dirección perpendicular a aquella en que se propaga la perturbación (figura 6).

Desde luego que las *partículas del medio* sólo oscilan en torno a un punto, pero *no viajan con las ondas*, es decir, las ondas no transportan materia; entonces, ¿qué es lo que transportan las ondas? Transportan una señal en forma de perturbación del medio de propagación.

Figura 7. Parámetros de una onda.



En la propagación de ondas transversales, en el anterior diagrama se pueden identificar algunos parámetros básicos:

- Los puntos máximos y mínimos se denominan crestas y valles, respectivamente.
- El tamaño de una cresta o valle representa la amplitud de la onda A.
- La distancia entre dos crestas o dos valles consecutivos se denomina longitud de onda, se representa con la letra λ del alfabeto griego (se lee lambda) y se mide en metros, como cualquier otra longitud.
- El tiempo en que se repite una onda completa se conoce como periodo, se representa con la letra T y se mide en segundos.
- El número de ondas completas que pasan por un punto en una unidad de tiempo es la frecuencia, que se representa con la letra *f*. Si una onda completa pasa en un segundo tiene una frecuencia de un Hertz (y se denota 1 Hz). De modo que si en un segundo pasan 100 ondas, tendremos una frecuencia de 100 Hz. Una onda de radio tiene una frecuencia que puede llegar a más de 100 mega Hz, es decir, se repite más de cien millones de veces en un segundo, y hay frecuencias mucho más grandes.

El periodo y la frecuencia están íntimamente relacionados. Por ejemplo:

- Si el periodo es de un segundo, la frecuencia será de 1 Hz.
- Si la frecuencia es de 2 Hz, ocurren dos ondas en un segundo; cada onda se completa en 1/2 segundo y éste es su periodo.
- Si el periodo es de 1/10 s, la frecuencia será de 10 Hz y así sucesivamente.

De tal forma que la frecuencia y el periodo son recíprocos, es decir, el producto de la frecuencia por el periodo es siempre 1.

$$fT = 1,$$

por lo tanto

$$f = \frac{1}{T}.$$

Lo cual quiere decir que cuando la frecuencia crece, el periodo decrece y viceversa. Ahora, la propagación de las ondas tiene también una velocidad, como en el caso del movimiento rectilíneo. Para calcular la velocidad de propagación de las ondas en el resorte de gusano, consideramos *n* ondas de la misma longitud λ en un intervalo de tiempo $\Delta t = t_2 - t_2 = nT$. Las "*n*" ondas ocupan una longitud total $\Delta x = x_2 - x_2 = n\lambda$.

Entonces, la velocidad de las ondas será:

$$v = \frac{\Delta x}{\Delta t} = \frac{n\lambda}{nT} = \frac{\lambda}{T}.$$

Ya que $f = \frac{1}{T}$, entonces

$$v = \frac{\lambda}{T} = \lambda f.$$

Se ha mencionado que las ondas se pueden propagar en materiales líquidos, gaseosos y sólidos. Si imaginamos a un sólido como formado de átomos y moléculas fuertemente unidas por resortes duros, y que los átomos y moléculas de un líquido están débilmente unidos

por resortes "blandos", entonces se puede predecir correctamente que las ondas viajan más rápido en los sólidos que en los líquidos.

¿Qué ocurre con el sonido? Imaginemos a un peatón que pasa frente al patio de una escuela y oye los gritos y risas de los estudiantes, a pesar de la gruesa barda que la rodea. A este efecto del sonido doblando esquinas y saltando bardas se le llama *difracción* y es tan común que muy poca gente le presta atención. La difracción por obstáculos se puede presentar cuando las dimensiones de éstos son del tamaño de la longitud de onda del sonido.

Pero si la longitud de onda es muy pequeña comparada con los obstáculos, la onda no los rodea, no se difracta; se produce una "sombra de sonido" muy bien definida, de manera que para escuchar el sonido habría que situarse sobre una línea recta sin obstáculos entre el que escucha y el lugar de donde proviene el sonido.

Cuando las ondas sonoras inciden sobre una superficie, una parte de su intensidad es absorbida y el resto se refleja. Las ondas sonoras se absorben mejor en las superficies blandas; por ejemplo, las cubiertas con materiales como fieltro, corcho, algodón o alfombras, que se utilizan para recubrir las paredes de teatros y cines.

Las perturbaciones en el aire, producidas por las vibraciones de un objeto, por ejemplo, las producidas por la superficie de un tambor, se propagan longitudinalmente como en uno de los casos del "resorte de gusano". En la figura 8 se puede observar que las espiras comprimidas del resorte corresponden a una capa de alta densidad de aire y las espiras extendidas del resorte a una capa de baja densidad, representadas por zonas oscuras y claras, respectivamente.

Figura 8. Ondas saliendo de la superficie de un tambor. Más oscuro, mayor densidad; menos oscuro, menor densidad. [Véase simulación en CD: "Compresiones y expansiones formadas por un tambor".]

La "dureza" o la "blandura" de un resorte se caracteriza diciendo que tiene baja o alta *elasticidad*, dependiendo de si, con el mismo esfuerzo, se estira poco o mucho. Robert Hooke (1635-1703), físico inglés contemporáneo de Newton, hablaba, en 1680, del "resorte del aire" para indicar que su compresión se parecía a la compresión de un resorte. Como la elasticidad del aire es menor que la de un sólido, por ejemplo un metal, una onda se propagará en el primero con menor rapidez.

2.4 EFECTO DOPPLER

Hay movimientos oscilatorios o vibraciones que no necesariamente son detectadas por nuestro oído. La propiedad de la vibración que determina si la escuchamos o no es la frecuencia. El oído de un adulto puede captar vibraciones con frecuencias entre 15 y 15 000 Hz, mientras que un niño puede escuchar vibraciones hasta de 20 000 Hz.

Las vibraciones con frecuencias menores de 15 Hz y mayores 20 000 Hz no las oímos, y las llamamos infrasónicas y ultrasónicas, respectivamente. Algunos animales, como los perros y los murciélagos, pueden captar vibraciones ultrasónicas. Al transmitirse por di-

ferentes medios, las ondas ultrasónicas tienen aplicaciones importantes en la medicina, la industria y el control del tráfico vehicular. El efecto en el que se basan estas aplicaciones es el cambio de frecuencia de la onda al ser reflejada por un objeto en movimiento, lo que se conoce como el efecto Doppler.

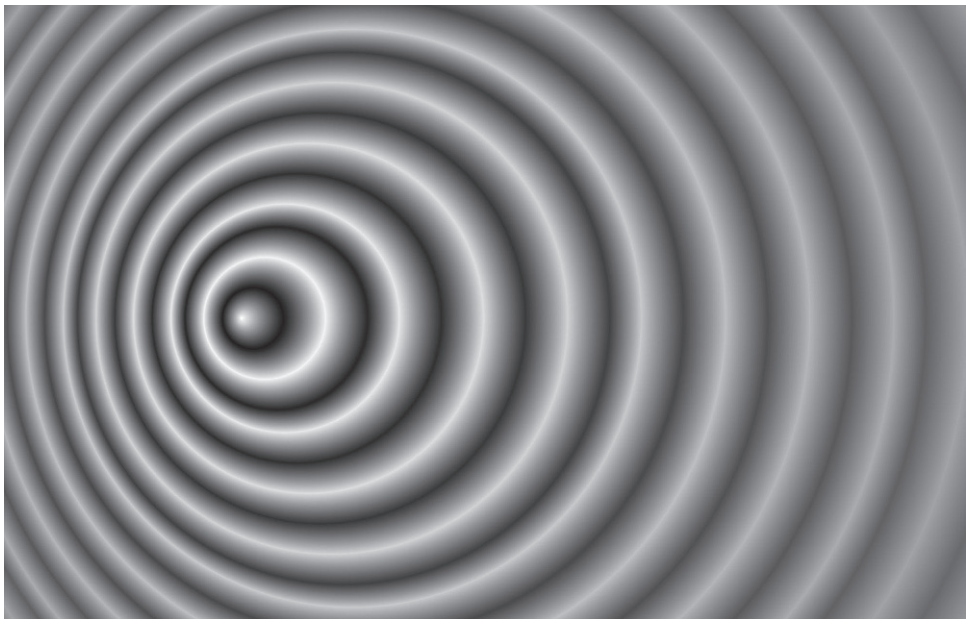
Al aumentar la frecuencia, la longitud de la onda disminuye debido a que

$$\lambda = \frac{v_s}{f},$$

desaparece la difracción y la onda empieza a ser reflejada nítidamente por objetos cada vez más pequeños. Es por eso que un murciélago emitiendo ondas con frecuencias de hasta 150 000 Hz puede captar el eco proveniente de insectos menores de 1 cm de longitud.

El estudio de las vibraciones en el aire es importante por varias razones, pero en particular porque mediante ellas nos comunicamos: las palabras son producidas por la vibración de nuestras cuerdas vocales.

2.4.1 Efecto Doppler en el agua



Efecto Doppler. [Véase simulacro en CD: "Efecto Doppler".]

Si se deja caer un objeto en una capa de agua quieta, se observa que se forman ondas circulares concéntricas que se alejan del punto de perturbación. Si se genera la onda metiendo en el agua una vara o la punta de un lápiz, y en lugar de sacar el lápiz se desplaza lateralmente dentro del agua, se observa que las ondas delante del lápiz se amontonan y, atrás de él, tienden a separarse. A este cambio de frecuencia de la onda debido al movimiento de la fuente se le llama *efecto Doppler*, y ocurre también con el sonido y con la luz. En el sonido lo notamos porque cambia su tono: con la luz cambia su color.

Si se aumenta la velocidad con la que se saca el lápiz, delante de él aumentará la compactación de las ondas, y atrás se separarán más. El hecho de que este cambio de frecuencia dependa de la velocidad de la fuente convierte al efecto Doppler en un valioso fenómeno que permite medir velocidades de fuentes, objetos, estrellas y galaxias, la sangre en nuestras venas, etcétera.

2.5 DE ARISTÓTELES A GALILEO: UNA APORTACIÓN IMPORTANTE PARA LA CIENCIA

Si analizamos el movimiento de una hoja de papel delgado, como un filtro de café que cae libremente, es decir, sin que nadie lo empuje, observamos que éste cae a velocidad constante, atraído por la Tierra. ¿Qué ocurre con la caída si al filtro se le cambia de forma, por ejemplo, formando una bola?, ¿y si se le da forma de lanza, en qué caso cae primero?, ¿la rapidez de caída de estos objetos depende de la forma que se les dio?, ¿el aire interviene en la explicación de esta diferencia? Si dejamos caer dos objetos al mismo tiempo, nos daremos cuenta de que el más compacto cae más rápido.

Ahora, imaginemos un tubo de vidrio o plástico al que se le ha extraído el aire y dentro de él se dejan caer simultáneamente: una pluma de ave y una canica. Se observará que caen al mismo tiempo. De estas experiencias, se pueden sacar dos conclusiones:

a] Lo objetos en el aire tardan tiempos diferentes en caer, dependiendo de su forma.
b] Cuando no hay aire todos los objetos caen en el mismo tiempo, sin importar su forma y composición.

El caso del movimiento vertical del filtro de café se muestra en la figura 9 (x, t).

Para una mejor descripción del movimiento, se representa la rapidez y el tiempo, en donde el eje vertical es la rapidez (v) y el horizontal es el tiempo (t), como se ve en la figura 10. Aquí, al movimiento del filtro de café le corresponde una línea horizontal, lo que quiere decir que la rapidez no cambia con el paso del tiempo, es constante.

Pensemos ahora en un movimiento de caída más complejo, en el cual del tiempo t_0 al tiempo t_1 el objeto se mueve con una rapidez v_1. Del tiempo t_1 al tiempo t_2 el objeto se mueve con una rapidez v_2 mayor que v_1; del tiempo t_2 al tiempo t_3 el objeto se mueve con una rapidez v_3 mayor que v_2 y así sucesivamente, como se muestra en el diagrama (figura 11).

En este caso, la rapidez es constante en cada uno de los intervalos de tiempo, pero a su vez crece formando una escalera. Si estos intervalos son cada vez más pequeños, entonces los peldaños reducen su ancho hasta convertirse eventualmente en puntos. Si los unimos obtenemos una recta con cierta inclinación. Véase figura 12.

¿Cómo sería la correspondiente gráfica (x, t) para este movimiento? Recordemos que cuando estudiamos el movimiento con velocidad constante, vimos que su gráfica (x, t) era una recta cuya inclinación es proporcional a la rapidez. Como la rapidez se incrementa

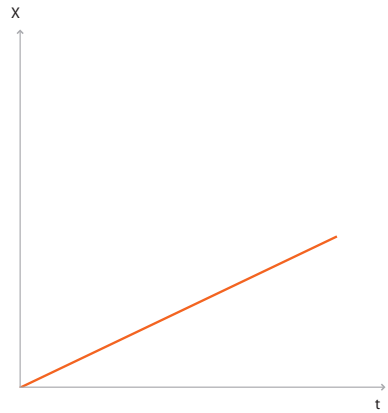


Figura 9. Representación en (x, t) *de un movimiento rectilíneo a rapidez constante.*

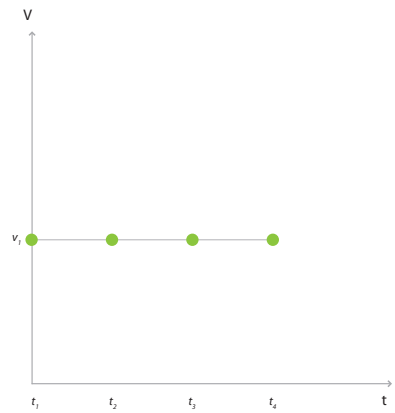


Figura 10. Representación en (v, t)*, de un movimiento rectilíneo a rapidez constante.*

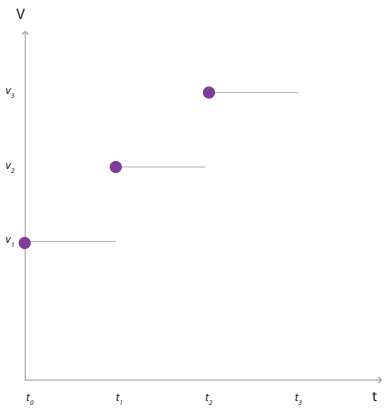


Figura 11. (v *contra* t).

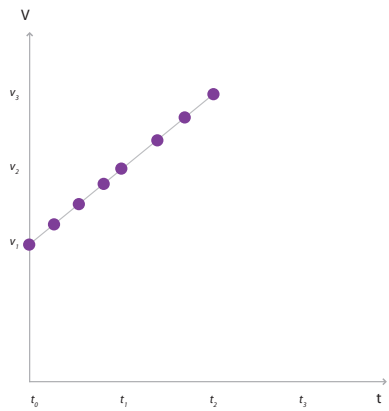


Figura 12. (v *contra* t).

en cada intervalo de tiempo, la inclinación de la recta en cada intervalo será cada vez mayor.

Si hacemos estos intervalos muy pequeños, las rectas inclinadas se convierten en puntos que, al unirlos, nos dan una curva continua creciente (figura 13).

Decimos que un objeto se acelera cuando aumenta su rapidez y que se desacelera en caso contrario.

En la historia de la ciencia se registran diferentes respuestas al estudio del movimiento de los objetos sobre la superficie de la Tierra, empezando por el problema de la caída de los cuerpos en presencia del aire. La caída en el vacío es imaginada hasta el siglo XVII, en Europa, sobre todo a partir del invento de la bomba de vacío en 1650 por Otto von Guericke (1602-1686).

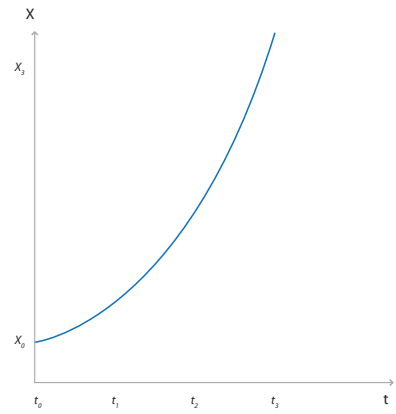


Figura 13. (x *contra* t).

Pero no sabemos lo que al respecto pensaban los científicos en distintas culturas de la Antigüedad, como los mayas o los babilonios, pues no hay registro de su pensamiento. Ni siquiera sabemos con certeza si era un problema que les importaba, aunque la evidencia indirecta, como el diseño del átlatl, presente en muchas culturas de la Antigüedad, demuestra que el movimiento de lanzas sí era un problema importante al que, además, le dieron una brillante solución.

Otro ejemplo es el movimiento de piedras lanzadas a mano o por hondas; de hecho, el invento de la honda es un ejemplo de la existencia de la solución a una necesidad.

2.5.1 Modelo aristotélico

Es una fortuna contar con el registro escrito de algunos pensadores griegos, como Aristóteles (384-322 a.n.e.), quien en su libro titulado *Física* trató el problema del cambio y el movimiento.

Aristóteles sólo se basaba en la observación de hechos, formulando suposiciones que constituían una representación o modelo de la naturaleza. A partir de las suposiciones se explican los hechos observados y hasta es posible hacer predicciones. Si estas predicciones concuerdan con lo observado, hablamos de una corroboración del modelo. Si el objeto no se comporta como predice el modelo, entonces tenemos una discrepancia entre el modelo teórico y el comportamiento observado del objeto. En este caso habrá que des-

echar el modelo, modificar alguna de sus suposiciones o comprobar si las observaciones han sido realizadas correctamente.

Para explicar el modelo de Aristóteles acerca del movimiento de los objetos, en particular para la caída libre sobre la superficie de la Tierra, observemos el movimiento de los objetos que nos rodean: unos objetos se mueven porque otros los empujan o porque los jalan, como cuando se lanza una pelota o una piedra. Estas acciones se pueden ejecutar porque los seres vivos tenemos la capacidad de imprimir movimiento a otros objetos.

En algunas culturas antiguas se pensaba que los fenómenos como el viento o la lluvia eran ocasionados por agentes invisibles o dioses. Éstos tenían que existir, pues todo movimiento fuera de piedras, humo, aire o agua, debería ser provocado por alguien que los lanzaba o jalaba. El viento era aire empujado, la lluvia agua lanzada desde arriba, los truenos fuego lanzado hacia abajo desde el cielo, etc. Estas creencias estaban encaminadas a la explicación del fenómeno o de los hechos del movimiento de objetos, basadas en la suposición de que un agente externo los provocaba.

El modelo desarrollado por Aristóteles no supone la acción de ningún agente externo para explicar el movimiento natural de los objetos, más bien indica que todos los movimientos son debidos a agentes o causas naturales.

Aristóteles se interesó en describir y explicar el movimiento de los objetos comunes del mundo circundante, tal y como lo percibimos con nuestros sentidos. Su atención se centraba en el movimiento de objetos de su entorno, como el lanzamiento de piedras, lanzas y flechas, el vapor de agua, el viento, una llama, la lluvia, los astros, etc. Trató de explicar estos movimientos mediante una sola suposición: todos los objetos tienden a moverse hacia su lugar natural. De acuerdo con él, hay distintos lugares naturales, según sea la composición de los objetos, además de que todos los objetos están formados de cuatro elementos básicos o fundamentales: tierra, agua, aire y fuego.

La tierra tiene como lugar natural el centro de la Tierra. Por eso nuestro planeta asume la forma que tiene, ya que si sus partes se van aglomerando tendiendo hacia su centro, no puede tener otra forma que la esférica. Si los objetos terrosos, como una piedra, se sueltan de la mano, inmediatamente tenderán a caer hacia su lugar natural, el suelo. ¿Qué sucede cuando a una piedra se la lanza hacia arriba? ¿Por qué se mueve en sentido contrario al sitio donde está su lugar natural?

Aristóteles supuso que si la piedra se lanza hacia arriba, estaría dotada de un impulso proveniente del aire detrás de ella. El impulso es ocasionado por la mano y desaparece en el momento en que el movimiento del aire detrás de la piedra se desvanece, para entonces irremediablemente caer hacia su sitio natural.

Los objetos terrosos, siendo más pesados que el agua, la van expulsando hacia arriba, obligándola a acumularse por encima de ellos. Por eso los lagos, mares y ríos están naturalmente situados sobre la superficie de la Tierra.

El aire, siendo más ligero que la tierra y que el agua, tiende a ir hacia arriba, terminando por acumularse por encima. El fuego, más ligero que los tres elementos anteriores, tiende a elevarse hacia la parte superior de la esfera del aire, es decir, se situará en el extremo superior de la atmósfera. Ésa es la razón de que las lengüetas de una llama apuntan hacia arriba.

El modelo de Aristóteles consta de objetos compuestos de combinaciones de los cuatro elementos y supone que ellos se mueven hacia su lugar natural. En un mundo organizado así, primero tenemos la esfera de la Tierra; a ella la rodea en gran parte la capa esférica del agua; por encima de ella está la capa esférica del aire y, finalmente, en la parte superior se sitúa la capa del fuego.

Para el caso del movimiento de una piedra que se lanza manualmente hacia arriba, Aristóteles decía que el impulso dado se transmitía al aire y que éste transportaba la pie-

dra a lo largo de su trayectoria. Sin embargo, conforme el impulso era transmitido punto a punto en el aire, éste se debilitaba, de manera que el movimiento natural de la piedra se iba haciendo cada vez más dominante. El movimiento hacia arriba disminuía, transformándose en un movimiento hacia abajo, hasta que la piedra quedaba, de nuevo, finalmente en reposo sobre el suelo. Ni la fuerza del brazo ni la de una catapulta podrían, a la larga, vencer el movimiento natural de la piedra. Para Aristóteles el movimiento era un proceso que inevitablemente termina en el reposo o en la ausencia de movimiento.

Es importante analizar este punto de vista sobre el movimiento, porque proviene de una de las mentes más lúcidas que han existido en la historia de la humanidad. El modelo aristotélico parecía predecir y explicar tantas cosas, que fue aceptado por grandes estudiosos durante los dos mil años siguientes. Al parecer, las dudas sobre su teoría empiezan a surgir cuando se detectan contradicciones internas. A pesar de ellas, los defensores de la teoría aristotélica encontraron argumentos para contrarrestarlas; de este modo, el modelo sobrevivió muchos años.

En la actualidad, la forma definitiva de someter a prueba una teoría consiste en obtener una conclusión necesaria de ella y luego confrontarla con el experimento tan exactamente como sea posible.

Los griegos, al igual que la mayoría de los estudiosos medievales europeos, aparentemente se contentaban con la belleza lógica de las teorías que formulaban, no les preocupaba someterlas a la prueba experimental. Algunas excepciones notables en estos aspectos experimentales fueron Arquímedes (nacido alrededor del año 287) y Herón (quien vivió en el siglo I de nuestra era).

2.5.2 El modelo de Galileo

De manera especial, el pensamiento desarrollado a partir del Renacimiento europeo empezó a basarse en la necesidad de la experimentación. Es decir, se trataba de la observación del comportamiento de los objetos, de forma planeada de antemano y también bajo ciertas condiciones de control. La experimentación sobre el movimiento implicaba la medición de lo observado, a diferencia del método aristotélico que se basaba en la descripción cualitativa. A su vez, con la medición se hace uso del lenguaje matemático, de modo que la nueva física pasa de ser sólo cualitativa a ser también cuantitativa. En este aspecto fue notable la contribución de Galileo Galilei.

Galileo, a diferencia de Aristóteles, realizó experimentos concretos, dejando caer objetos bajo ciertas condiciones de control, como el que describe en uno de sus libros sobre el rodamiento de un balín en un canal inclinado de madera, forrado de cuero pulido. Pero también hizo experimentos pensados; éstos se refieren a situaciones ideales, inalcanzables en la realidad, como son el deslizamiento de balines en canales sin fricción o la caída de objetos en ausencia de aire.

En relación con la caída de los *graves* (o sea, los objetos que son atraídos hacia el centro de la Tierra), Galileo manejó conceptos, hipótesis y métodos muy alejados de los propuestos por Aristóteles. Galileo supuso que, en el vacío, todos los cuerpos caen con la misma rapidez hacia el centro de la Tierra; y que si se les deja caer al mismo tiempo desde la misma altura, llegarán al suelo simultáneamente. Lo anterior quiere decir que, en el vacío, las bolas de madera, metal o plumas de ave, caerán al mismo tiempo al suelo. Pero fue más adelante, pues pudo descubrir la forma del movimiento de caída de los graves en el vacío. Primero supuso, erróneamente, que la rapidez de caída era proporcional a la distancia. Varios años estuvo estacionado en ese error, hasta que se dio cuenta de que el movimiento

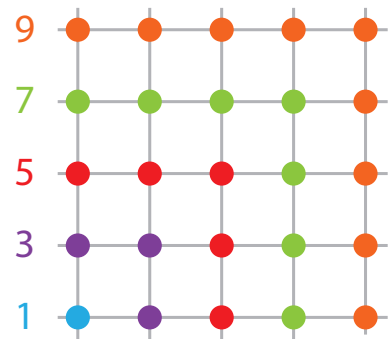
se caracteriza, de manera "natural", por el tiempo en vez de por la distancia. Esto le llevó a proponer que la rapidez de caída es proporcional al tiempo, resultado aparentemente ya conocido por Leonardo da Vinci (1452-1519).

También determinó que el movimiento de caída libre es uniformemente acelerado; la distancia y el tiempo de caída se relacionan de una manera muy especial. Esto lo expresó con la siguiente frase un tanto oscura, en una carta dirigida a un colega veneciano, el 16 de octubre de 1604:

> Los espacios atravesados por el movimiento natural están en proporción doble del tiempo y, por consiguiente, los espacios atravesados en tiempos iguales son como los números impares *ab unitate*.

¿Qué quiso decir con esto el famoso físico? Para descubrirlo, habrá que observar la figura siguiente:

Figura 14. Serie pitagórica.



Entre un elemento de la serie y el siguiente transcurre un intervalo igual de tiempo. La distancia recorrida es proporcional a la cantidad de puntos de cada cuadrado. Cada cuadrado se construye sumando al anterior un número de la serie de los números impares, empezando por la unidad (*ab unitate*, dice Galileo en latín).

Y esto es lo sorprendente: que la relación entre distancias y tiempos de caída se ajusta exactamente a la descripción geométrica pitagórica. Esto Galileo lo expresaba diciendo que "el lenguaje de la naturaleza es la geometría".

De este modo propuso un método o procedimiento para acercarse a las verdades de la naturaleza: observar el fenómeno en cuestión (como el movimiento vertical de graves), hacer alguna hipótesis respecto a dicho movimiento (por ejemplo, que la rapidez de caída es proporcional al tiempo y no depende del peso del objeto), derivar una consecuencia de la hipótesis o predicción (que la distancia recorrida cambia, como en la figura anterior), someter la hipótesis y sus consecuencias a la prueba experimental.

La hipótesis de Aristóteles era que los objetos pesados caerían más rápido que los ligeros, mientras que la suposición de Galileo era que el peso no influye. Cuál de las dos hipótesis es la verdadera, es una cuestión que la naturaleza nos responde cuando (de acuerdo con el nuevo método) el científico le pregunta mediante un experimento bien pensado y controlado, tomando medidas cuantitativas y no sólo con argumentos cualitativos. Así es que la naturaleza tiene la última palabra y se expresa en el lenguaje matemático.

Al hacer el experimento, seguramente Galileo se dio cuenta que el tiempo de caída vertical era demasiado rápido como para medirlo con el pulso de su corazón o algún otro procedimiento conocido. Entonces intentó modificar las condiciones de control de tal forma que el tiempo de caída disminuyera, sin cambiar por ello la esencia del fenómeno.

Fue así como concibió el experimento en que balines de pesos diferentes se dejaban caer por un canal situado en un plano inclinado. En la descripción del experimento, se observa que el tiempo de recorrido del balín sobre el canal lo midió cuantificando el agua que salía de un recipiente. Con ese reloj de agua Galileo pudo comprobar la predicción extraída de su hipótesis: que la rapidez de caída variaba directamente con el tiempo.

2.6 LA ACELERACIÓN

Una vez estudiado el movimiento rectilíneo uniforme, se revisará el movimiento de un cuerpo cuya velocidad cambia; se hablará de la aceleración.

En química se habla de acelerar una reacción cuando se menciona que se realiza en menor tiempo del ordinario; cuando se hace referencia al cambio climático, se indica que se está acelerando el deshielo de los glaciares; en economía, se habla de aceleraciones y de desaceleraciones para referirse a variaciones en la velocidad de crecimiento de algunos indicadores económicos de los países. Así pues, la aceleración en la vida cotidiana está relacionada con el cambio en la rapidez de algún fenómeno y con el tiempo en que ocurre dicho cambio. Este concepto de aceleración concuerda con el que se tiene en física.

En física, se define la aceleración media como el cociente del cambio de velocidad entre el tiempo empleado en dicho cambio. Esto es, si el cambio de velocidad se define como $\Delta v = v_2 - v_1$ y el intervalo de tiempo empleado como $\Delta t = t_2 - t_1$, entonces la aceleración media, a_m, se expresa así:

$$a_m = \frac{\Delta v}{\Delta t} \text{ (figura 15).}$$

Las unidades para la aceleración, en el Sistema Internacional de Unidades, son unidades de velocidad entre unidades de tiempo; esto es:

$$\frac{m/s}{s} = \frac{m}{s^2}.$$

Así pues, la aceleración está directamente relacionada con el cambio en la velocidad e inversamente con el tiempo empleado.

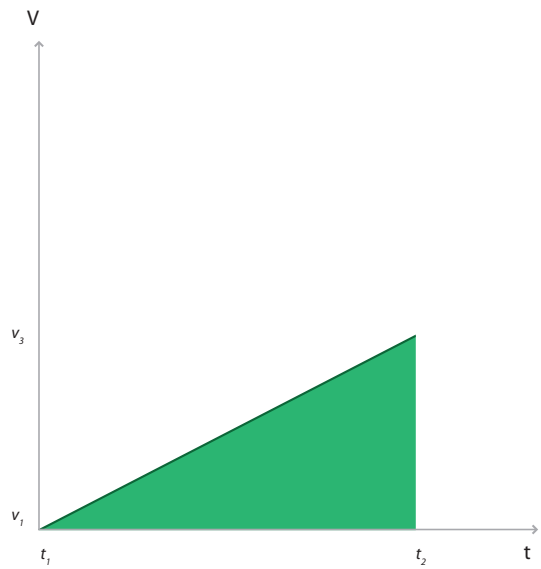


Figura 15. Velocidad contra tiempo.

De aquí se tiene que:

$$a(t_2 - t_1) = v_2 - v_1, \text{ y considerando } t = t_2 - t_1$$

y de la figura anterior se tiene que la distancia recorrida en este intervalo de tiempo es:

$$x = \frac{1}{2}(v_1 + v_2)t,$$

pero

$$t = \frac{(v_2 - v_1)}{a},$$

sustituyendo este valor en x se tiene:

$$x = \frac{1}{2}(v_1 + v_2)\left(\frac{v_2 - v_1}{a}\right),$$

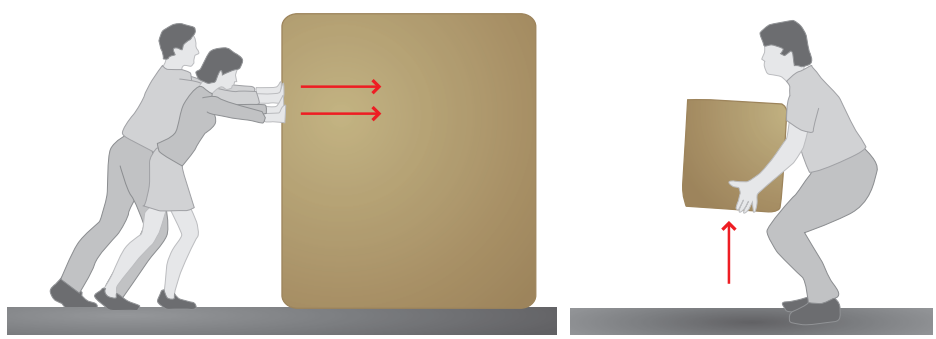
de donde

$$2ax = v_2^2 - v_1^2.$$

Al estudiar el movimiento de los cuerpos, Galileo no se preocupó por sus causas; es decir, el tratamiento de Galileo a los cuerpos que caían a la Tierra era puramente cinemático.

2.7 LA MEDICIÓN DE LA FUERZA

En la siguiente figura se presentan algunos ejemplos de aplicación de fuerzas, donde una persona realiza acciones como deslizar una silla sobre el piso, empujar un pesado bloque, levantar una almohada y una pesada caja.



Estas personas están aplicando una fuerza sobre cada uno de los objetos descritos; la fuerza aplicada es, en cada caso, de diferente magnitud.

Ahora bien, la fuerza aplicada sobre un objeto es una magnitud física y, por lo tanto, debemos contar con una unidad en el Sistema Internacional de Unidades (SI). Así, como hemos visto, la unidad para medir el tiempo es el segundo (s); la empleada para medir longitud es el metro (m) y la unidad para medir la fuerza es el newton (N). El instrumento que se utiliza para medir fuerzas se llama dinamómetro (figura 16).

En la figura 16 se presenta un dinamómetro, constituido por un resorte y una escala en newtons. ¿Por qué se emplea un resorte para medir fuerzas? Se utilizan los resortes como dinamómetros porque son cuerpos elásticos que modifican su tamaño al aplicarles una fuerza. En ellos, la longitud varía de forma directamente proporcional con la magnitud de la fuerza aplicada, como se muestra en la figura 17: x es el cambio de longitud en el resorte, es decir, la cantidad en la que se ha incrementado la longitud del resorte al aplicarle una fuerza. Si se tienen dos resortes diferentes, se les aplica los mismos valores de fuerzas a cada uno y se grafica la fuerza que marca el dinamómetro contra la longitud de elongación del resorte en cada caso, resulta que la pendiente de la recta roja es mayor que la de la azul. Esto indica que el resorte representado por la línea roja es más difícil de estirar que el correspondiente a la azul, es decir, es "más duro".

Figura 16. Dinamómetro

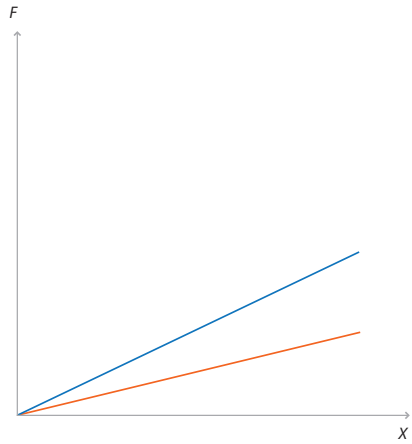


Figura 17. F contra x para dos resortes con diferente constante de elasticidad.

La fuerza de la atracción gravitacional de la Tierra sobre un objeto que cuelga de un dinamómetro, se puede relacionar con su elongación. Si la masa del objeto es un kilogramo (la masa de un litro de agua), entonces a la elongación que presenta el dinamómetro le llamaremos un kilogramo-fuerza. Esta unidad de fuerza equivale a 9.81 newtons.

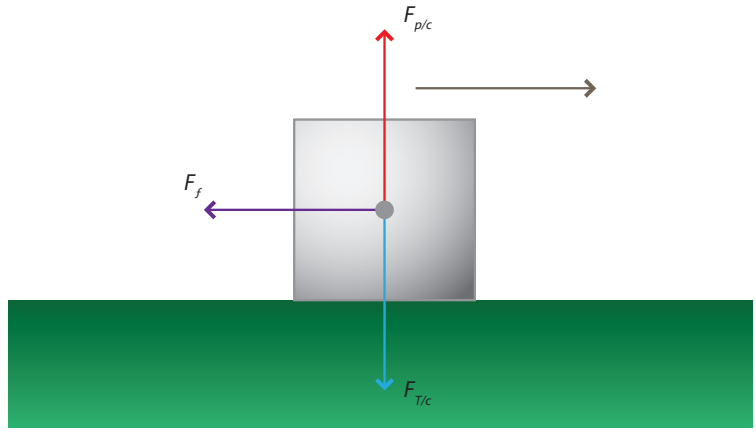
2.8 LA GRAN APORTACIÓN DE ISAAC NEWTON: LA IDEA DE INERCIA

Cuando se estudia el movimiento de los objetos sobre la superficie de la Tierra, se encuentra que existe una fuerza que siempre se opone a su movimiento: ésta depende de las características de las superficies en contacto; a esta fuerza la conocemos como *fuerza de fricción*.

Existen dos tipos de fuerzas de fricción: la *estática* y la *dinámica*. La primera es la fuerza que se opone a que se inicie el movimiento de un cuerpo sobre una superficie. Cuando el objeto ya está en movimiento, se presenta una fuerza de fricción entre las superficies (el piso y la base del objeto) que se denomina fricción dinámica.

Veamos con detenimiento qué fuerzas actúan sobre una caja que se desliza en el suelo y eventualmente se detiene. Sobre la caja se están aplicando las siguientes fuerzas: la fuerza gravitacional de la Tierra ($F_{T/c}$); la fuerza que ejerce el piso sobre el cuerpo ($F_{p/c}$) y la fuerza de fricción entre la superficie inferior de la caja y el piso (F_f) (figura 18).

Figura 18. Diagrama de fuerzas.



Como vemos en esta figura, la fuerza gravitacional de la Tierra $F_{T/c}$ y la fuerza que ejerce el piso sobre el cuerpo $F_{p/c}$ son verticales de sentidos opuestos y de igual magnitud, por lo que se anulan. Así que

la única fuerza no balanceada, es decir, la *fuerza neta*, es la fricción, la cual actúa oponiéndose al movimiento de la caja.

Debido a esta fuerza, la velocidad se hace cada vez menor hasta que finalmente la caja se detiene. ¿Qué sucede si disminuimos la fuerza de rozamiento, poniéndole ruedas o lubricando el piso? La caja tardará más en detenerse y, en consecuencia, recorrerá una mayor distancia. Si lográramos eliminar totalmente el rozamiento, es decir, si no hay fuerza de fricción que se oponga al movimiento, no existiría ninguna razón para que la velocidad disminuya; entonces la caja seguiría moviéndose indefinidamente en la misma dirección y con la misma velocidad inicial.

En resumen, cuando la fuerza neta sobre un objeto es cero, su velocidad se mantiene constante en magnitud y dirección; si un cuerpo se mueve con velocidad constante, entonces podemos decir que la fuerza neta sobre él es cero. Este resultado es el que establece Newton en su primera ley o principio de inercia.

Quizá las condiciones más parecidas a la ausencia de fricción las encontramos en los llamados *rieles* y *mesas de aire*; estas últimas, a veces se ubican en las salas de juegos. Se puede ver que el disco, después de que se le golpea, viaja a velocidad constante hasta que choca con alguna de las paredes de la mesa y entonces cambia su dirección.

2.9 LA RELACIÓN DE LA MASA, LA ACELERACIÓN Y LA FUERZA. SEGUNDA LEY DE NEWTON

Cuando sobre un objeto actúa una fuerza neta diferente de cero, entonces la velocidad del objeto cambia; el cambio puede ser aumento o disminución en su magnitud, en su dirección o en ambas. Podemos decir también que todo cambio en la velocidad de un objeto es debido a la acción de una fuerza no balanceada que actuó sobre él.

En la segunda mitad del siglo XVII, Isaac Newton analizó esta relación entre fuerza neta y cambio de velocidad, pasando a formar parte de las leyes del movimiento, las cuales son la base de su magna obra *Philosophiae Naturalis Principia Mathematica*, publicada en 1687. Newton, para algunos el más grande científico de la historia, declaró: "Si he visto lo que muchos otros no han podido ver, es porque he estado montado en hombros de gigantes." Esta declaración es un reconocimiento a que sus trabajos e investigaciones tuvieron como base los aportes de algunos sabios que lo precedieron, como Galileo Galilei, Johannes Kepler y Tycho Brahe.

Newton supo de los experimentos de Galileo y de las reflexiones de Descartes sobre la inercia, esa tendencia de los objetos a conservar su estado de movimiento rectilíneo o de reposo. Observó que la inercia de los objetos dependía directamente de su masa; los objetos muy masivos presentan una inercia muy grande, es decir, una gran tendencia a conservar su estado de movimiento rectilíneo o de reposo. Esto lo llevó a razonar que la aceleración que experimenta un objeto depende tanto de la fuerza neta aplicada como de su masa.

Experimentando encontró que la velocidad de un objeto varía de manera inversa con la masa. Es decir:

$$\Delta v \propto \frac{1}{m}.$$

Esto es, cuando la masa aumenta, el cambio en velocidad disminuye. Además, el cambio en la velocidad de un objeto varía directamente con la fuerza neta aplicada.

$$\Delta v \propto F_{\text{neta}}.$$

Es decir, el cambio en la velocidad aumenta cuando la fuerza neta aumenta. De estas dos relaciones de proporcionalidad, concluyó que:

$$\Delta v \propto \frac{F_{\text{neta}}}{m}.$$

De esta expresión se desprende que si se fija el valor de la fuerza neta, entonces el cambio de velocidad varía de manera inversa con la masa. Ahora, si se fija la masa y se varía la fuerza neta, entonces el cambio en la velocidad varía directamente con ésta.

Pero mientras esté actuando la fuerza neta sobre el objeto, la velocidad de éste estará cambiando; por lo tanto, el cambio total que experimente la velocidad tendrá que ser grande, si el intervalo de tiempo durante el cual actúa la fuerza es grande. Esto es, Δv varía de manera directa con el intervalo de tiempo Δt, es decir:

$$\Delta v \propto \Delta t\,.$$

Se pueden reunir las dos relaciones de proporcionalidad anteriores, en una sola:

$$\Delta v \propto \frac{F_{\text{neta}}}{m}\Delta t.$$

Esta relación nos dice que el cambio de velocidad de un objeto varía de manera directa con la fuerza neta aplicada, en el intervalo de tiempo durante el cual dicha fuerza se aplica y de manera inversa con la masa del objeto.

Ahora, en esta última relación multiplicamos a ambos lados por m:

$$m\Delta v \propto F_{\text{neta}}\Delta t.$$

Entonces, si llamamos k a la constante de proporcionalidad:

$$m\Delta v = kF_{\text{neta}}\Delta t.$$

Si las cantidades involucradas en esta relación de proporcionalidad se expresan en las unidades del Sistema Internacional, resulta que $k = 1$ y por lo tanto:

$$m\Delta v = F_{\text{neta}}\Delta t.$$

Esta ecuación es muy parecida a la que usó Newton para relacionar a la fuerza neta que actúa sobre un objeto con el cambio en su estado de movimiento, y nos permite cuantificar el cambio en velocidad que experimenta un objeto de masa m al aplicarle una fuerza neta conocida durante el intervalo de tiempo Δt.

Recordando que todo cambio en la velocidad es un indicador de la existencia de aceleración y utilizando este concepto, se obtiene una expresión que relaciona la *fuerza neta* con la aceleración. Dividiendo la ecuación anterior entre Δt, queda:

$$m\frac{\Delta v}{\Delta t} = F_{\text{neta}}.$$

En esta ecuación aparece el cociente del cambio de velocidad Δv, entre el intervalo de tiempo que dura ese cambio, Δt. Como es sabido, Δv / Δt es la aceleración del objeto en ese intervalo de tiempo, así:

$$F_{\text{neta}} = ma.$$

De esta ley se puede concluir que:

> Dado el movimiento, se puede encontrar la fuerza que lo produce o dada la fuerza, se puede encontrar el movimiento.

La ecuación anterior es la representación más conocida de la llamada segunda ley de Newton. En el SI, m se mide en kilogramos (kg), a se mide en metros sobre segundo al cuadrado (m/s^2) y F_{neta}, se mide en newtons (N). Así, se puede decir que un newton es la fuerza que, aplicada a un objeto cuya masa es de 1 kg, le produce una aceleración de 1 m/s^2.

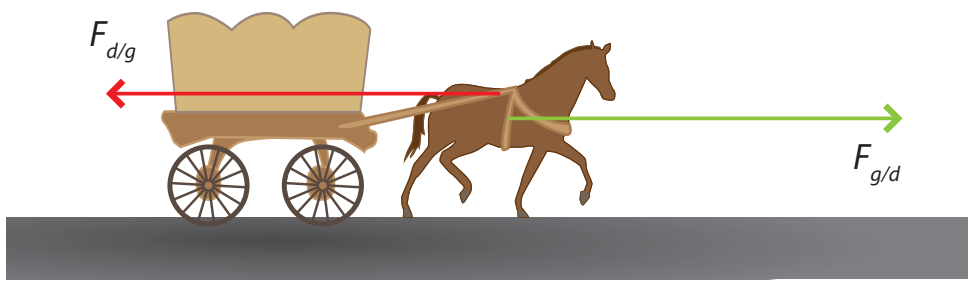
2.10 LA ACCIÓN Y LA REACCIÓN. TERCERA LEY DE NEWTON

Finalmente, la tercera ley de Newton se refiere a la interacción entre pares de cuerpos, que llamaremos A y B. Si un cuerpo *A* ejerce una fuerza sobre un cuerpo *B*, entonces el cuerpo *B* ejerce simultáneamente una fuerza sobre el cuerpo *A*. Se ha convenido en llamar a una de ellas *fuerza de acción*, y a la otra, *fuerza de reacción*.

Esta ley dice que la fuerza que ejerce el cuerpo A sobre el cuerpo B es de igual magnitud que la que ejerce B sobre A y que, además, son de sentidos opuestos:

$$F_{A/B} = -F_{B/A.}$$

Profundizando en lo relativo a estos pares de fuerzas *acción-reacción*, se tiene un par de *fuerzas*: $F_{d/g}$ y $F_{g/d}$:



- Forman un par de fuerzas.
- Se presentan de forma simultánea.
- Actúan sobre cuerpos diferentes.
- A una la llamamos *fuerza de acción* y a la otra *fuerza de reacción*.

Pero proporcionan más información:

$F_{d/g}$ y $F_{g/d}$ actúan en sentidos opuestos.

Otra característica de este par de fuerzas, que no es tan evidente, es que:

magnitud de $F_{d/g}$ = magnitud de $F_{g/d}$.

Esto es, que ambas fuerzas son de igual tamaño.

Aunque esta afirmación puede parecer errónea, pues la situación nos hace creer que actuó una fuerza de mayor magnitud sobre uno de ellos, no es así. $F_{d/g}$ y $F_{g/d}$ son de igual magnitud, sólo que la aceleración a_d es mayor que la aceleración a_g.

Estas tres leyes nos ayudan a saber si la suma de fuerzas sobre un objeto es cero o diferente de cero, según si el objeto está en reposo, se mueve con velocidad constante o en forma acelerada. Además, junto con la ley de la gravitación universal, que se abordará más adelante, forman el gran aporte de Isaac Newton al entendimiento de la relación entre fuerzas y movimientos de los cuerpos, conocida como la *mecánica newtoniana*.

Esta teoría explica tanto el movimiento de los cuerpos sobre la superficie terrestre, como el movimiento de los cuerpos celestes en el Sistema Solar. La *mecánica newtoniana* ha permitido diseñar y desarrollar una gran variedad de dispositivos mecánicos y grandes construcciones; también ha hecho posible los vuelos en órbitas terrestres y aún más los exitosos viajes a la Luna y el lanzamiento de las cápsulas espaciales que descendieron sobre la superficie de Marte.

2.11 LA LEY DE LA GRAVITACIÓN UNIVERSAL

Fue Newton al que se le ocurrió relacionar el movimiento de los cuerpos que caen a la superficie de la Tierra y el movimiento circular de la Luna en torno de la Tierra. Este razonamiento genial llevó a Newton a realizar cálculos basados en experimentos que culminaron con encontrar que ambos eran movimientos acelerados y que la causa de esta aceleración era la atracción gravitatoria dirigida hacia el centro de la Tierra.

¿Cómo se le ocurrió relacionar estos movimientos? ¿Cuál es la diferencia entre el movimiento del cuerpo que cae y el de la Luna? En efecto, ambos son movimientos acelerados y la causa de la aceleración es la misma: la atracción gravitatoria de la Tierra. Si los efectos no son los mismos, es debido a que las condiciones iniciales no son las mismas; la Luna tuvo alguna vez una velocidad inicial particular, que la hace describir una órbita circular, en cambio el cuerpo que cae, lo hace verticalmente hacia la Tierra. En ambos casos se trata de una caída provocada por una fuerza dirigida hacia el centro de la Tierra.

El sentido común indicaría que al lanzar un cuerpo hacia arriba, llegará a una cierta altura y después caerá; si lo lanzamos con más velocidad, llegará más alto y tardará más en caer. Si aumentamos y aumentamos la velocidad, llegará más lejos, pero siempre caerá, a menos que llegue a la velocidad para escapar de la Tierra, llamada velocidad de escape. ¿Qué ocurre si disparamos un proyectil en dirección paralela a la superficie de la Tierra, desde una altura y con una velocidad dadas? Si aumentamos la velocidad, el proyectil llegará cada vez más lejos. Si se considera el efecto de la curvatura de la Tierra, las órbitas que se obtienen son círculos, elipses o parábolas.

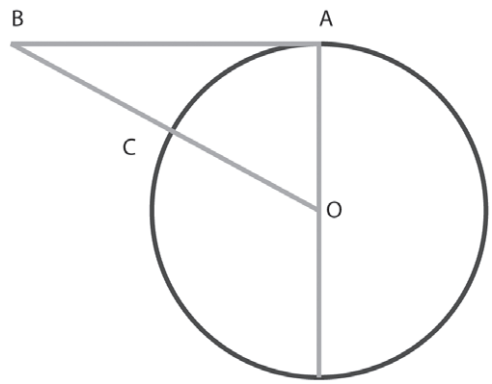
Por otro lado, ¿cómo podemos determinar la magnitud de la aceleración circular?

Desde luego esta aceleración será radial, esto es, hacia el centro del círculo. Christiaan Huygens (1629-1695) calculó esto. Supuso un cuerpo que se encuentra en A y que en su movimiento describe el arco de círculo AC en un tiempo t. Si no existiera una aceleración a, el cuerpo seguiría de A hasta B en movimiento rectilíneo uniforme, de modo que $AB = vt$.

Pero, por causa de dicha aceleración, se produce la "caída" BC que, de acuerdo con la Ley de Galileo, vale:

$$BC = \frac{1}{2}gt^2.$$

En el triángulo rectángulo *OAB* de la siguiente figura, se tiene que $OA = OC = R$, que es el radio del círculo.



$$OB^2 = OA^2 + AB^2.$$
$$(R + \frac{1}{2}at^2)^2 = R^2 + (vt)^2,$$
$$R^2 + aRt^2 + \frac{1}{4}(a^2t^4) = R^2 + (vt)^2,$$
$$aRt^2 + \frac{1}{4}(a^2t^4) = v^2t^2.$$

Dividiendo por t^2

$$aR + \frac{1}{4}(a^2t^2) = v^2.$$

Para tiempos muy pequeños, lo que es necesario para explicar una curva continua, se puede despreciar el término que contiene a t^2 y llegar al resultado:

$$a = \frac{v^2}{R}.$$

Ahora se puede calcular el valor de la aceleración que hace girar a la Luna alrededor de la Tierra, sustituyendo su velocidad y la distancia del centro de la Tierra al centro de la Luna; con estos valores se obtiene que:

$$a = 2.74 \times 10^{-3}\ \mathrm{m/s^2}.$$

Después se le ocurrió a Newton que estos resultados se podrían generalizar al caso del Sistema Solar. Claro que este problema era mucho más complicado, ya que en este caso, ni las órbitas son circulares (sino que son elipses), ni la velocidad de los planetas es constante.

El movimiento de los planetas fue un problema que interesó a los astrónomos desde Ptolomeo hasta Copérnico, y fue Johanes Kepler quien, analizando las observaciones planetarias de Tycho Brahe, pudo resumir dicho movimiento en tres leyes que llevan su nombre.

2.11.1 Leyes de Kepler

Primera ley | La línea que va del Sol al planeta barre áreas iguales en tiempos iguales.

Segunda ley | Los planetas describen órbitas elípticas en las que el Sol se encuentra en uno de sus focos.

Tercera ley | Los cubos de los semiejes mayores de las órbitas elípticas planetarias son proporcionales a los cuadrados de los periodos de los respectivos planetas.

Con estas leyes y su verificación llevada a cabo con los satélites de Júpiter y Saturno, Newton formuló la ley de la gravitación universal, proponiendo que se cumple para cualquier cuerpo que esté en presencia de otro u otros.

Si se considera el caso especial de un planeta que recorre una trayectoria circular alrededor del Sol, entonces, la aceleración centrípeta será como se mencionó anteriormente:

$$a = \frac{v^2}{R};$$

para un círculo, por lo que $v = \frac{2\pi R}{T}$,

$$a = \frac{4\pi^2 R}{T^2}.$$

Ahora, de la tercera ley de Kepler:

$$\frac{R^3}{T^2} = K,$$

de la cual obtenemos que

$$\frac{R}{T^2} = \frac{K}{R^2}.$$

Así que

$$a = K\frac{4\pi^2}{R^2}.$$

Entonces, enunció su ley diciendo que:

$$a = \frac{GM}{R^2}.$$

En esta ley, *M* representa la masa del cuerpo atractor y *G* es la constante de gravitación universal. Es decir, la aceleración gravitacional que un cuerpo le imparte a otro es inversamente proporcional al cuadrado de la distancia que los separa y directamente proporcional a una característica del cuerpo atrayente, a la que Newton le llamó "carga gravitatoria". Él observó que esta característica era mayor cuanto mayor era la "cantidad de materia" del cuerpo atrayente, y en consecuencia, la identificó con la masa gravitacional.

Esta ley fue comprobada en todos los casos conocidos y permitió a Simon Laplace (1749-1827) construir su trascendental *Mecanique Celeste*. Pero no sólo esto, sino que sentó las bases para predecir la existencia de otros planetas, como Neptuno por Leverrier (1811-1877). En la actualidad la ley explica satisfactoriamente el comportamiento de las estrellas binarias y la existencia de planetas en estrellas.

2.11.2 El campo gravitatorio

Un tratamiento alternativo para describir la dinámica gravitacional, es decir, el movimiento de un objeto debido a la fuerza de atracción gravitacional de otro, es mediante el concepto de *campo gravitatorio*. El campo gravitatorio E_g alrededor de un objeto de masa m_1 se define como la fuerza que ejerce sobre un objeto de masa unitaria situado a una distancia r de él. Esto es:

$$E_g = G\frac{m_1}{r^2}.$$

Si en un punto del espacio se coloca otro objeto con masa m_2 éste experimentará una fuerza de atracción gravitatoria dada por:

$$F = E_g m_2.$$

Nótese que al sustituir E_g en esta ecuación, se recupera la Ley de la Gravitación Universal de Newton:

$$F = G\frac{m_1 m_2}{r^2}.$$

En este contexto, es claro que, en contraste con el hecho de que la fuerza de atracción gravitatoria sólo se presenta entre dos objetos, el campo es una propiedad de un solo objeto que modifica a todo el espacio que lo rodea.

2.12 LA CANTIDAD DE MOVIMIENTO LINEAL

La cantidad de movimiento, momento lineal, ímpetu o *momentum* es una cantidad vectorial que se define como el producto de la masa del cuerpo y su velocidad en un instante determinado. En cuanto al nombre, Galileo en su obra *Discursos sobre dos Nuevas Ciencias* usa el término italiano *ímpeto*, mientras que Isaac Newton usa en *Principia Mathematica* el término latino *motus* (=movimiento) y *vis* (=fuerza). *Momentum* es una palabra directamente tomada del latín *momentum*, derivada del verbo *movere,* mover.

$$\vec{p} = m\vec{v}.$$

La idea intuitiva de esta definición estriba en que la "cantidad de movimiento" dependía tanto de la masa como de la velocidad. Para ilustrar este concepto pensemos en la cantidad de movimiento de un insecto y un tren que se mueven a la misma velocidad. Es obvio que, a pesar de que su velocidad es la misma, será mucho más fácil detener al insecto que al tren. Como se verá más adelante, esta cantidad es muy útil, principalmente cuando se estudian colisiones entre objetos.

2.13 EL CONCEPTO DE TRABAJO MECÁNICO

Para desplazar un cuerpo de un lugar a otro se necesita aplicar una fuerza a lo largo de la dirección en que se le desea mover. Es claro que cuanto más lejos se lleva, más trabajo hay que aplicar. Esta idea intuitiva, en mecánica se escribe:

$$W = Fx,$$

donde x es la distancia que se desplazó el objeto.

Las unidades en el SI serán newton por metro; a esta unidad se le llama *joule*, en honor a P. Joule (1818-1889), quien realizó experimentos encaminados a establecer el *principio de conservación de la energía.*

2.14 LA ENERGÍA: UNA IDEA FRUCTÍFERA Y ALTERNATIVA A LA FUERZA

Introduciremos ahora el concepto de energía, partiendo de la idea de que la energía se manifiesta en una amplia variedad de formas: eléctrica, mecánica, térmica, electromagnética, química, atómica, acústica, luminosa, etcétera.

Son diferentes manifestaciones de la energía: un rayo haciendo contacto con la copa de un árbol, un tornado levantando partes de la construcción de una granja, un volcán en erupción, la caída de agua llegando a una planta hidroeléctrica, una persona trabajando con soldadura eléctrica, un automóvil saliendo de una gasolinera, una estación radiodifusora emitiendo su señal, un radiorreceptor captándola y emitiendo sonido, o un boxeador conectando un golpe en la cara de su oponente.

¿Qué forma o formas de energía están presentes en cada una de las situaciones anteriores?, ¿en cuáles de ellas interviene sólo la naturaleza y en cuáles la tecnología inventada por el hombre?

En las situaciones antes descritas nos percatamos de que en todos los casos se presentan transformaciones de energía, de una forma a otra y de transferencia de energía de un cuerpo o sistema de cuerpos a otro. Éstas son características propias de esta magnitud física. En efecto, la energía se puede transformar y se puede transferir, aunque la propiedad más importante de la energía es que se conserva.

Si lo reflexionamos, se puede estar de acuerdo en que la energía es una magnitud física que caracteriza un estado del cuerpo o de un sistema de cuerpos y depende de ese estado. Ahora sabemos que la cantidad de energía que posee un cuerpo o sistema puede variar al intercambiar esta energía con otro u otros cuerpos o sistemas.

Con estas nociones sobre el concepto de energía se puede marcar la diferencia entre el uso que se da al término energía en ciencia y el que se le da en el lenguaje cotidiano, en donde algunas veces se usa el término energía con connotaciones muy diferentes a la aceptada en ciencia.

Así, por ejemplo, cuando una persona habla a otra con decisión y firmeza se dice que lo hace con energía; o también, cuando en un grupo de personas que realizan alguna actividad priva un ambiente de optimismo y buena relación se dice que hay "energía positiva" en ese grupo. Estas acepciones cotidianas del término *energía* no están relacionadas con el concepto científico, en donde se define a la energía como la magnitud fundamental de todo trabajo, considerando que no hay creación de energía, sino transformación de una forma a otra.

Con respecto a la idea de que la energía no se crea ni se destruye, sólo se transforma, consideremos el ejemplo de la energía almacenada en un recipiente de gasolina. La podemos consumir en el motor de un automóvil que nos transporta de un lugar a otro; esta energía se ha transferido al automóvil para transportarnos. En este caso podemos medir lo que nos costó el proceso, ya que conocemos la cantidad de combustible utilizado.

Cuando el automóvil se mueve de un lugar a otro, como en este ejemplo donde se ejerce una fuerza y se produce desplazamiento, se dice que se ha realizado un trabajo.

Si quisiéramos que el coche recorriera el doble de la distancia, tendríamos que duplicar la cantidad de gasolina; así, decimos que el trabajo que le costó al motor recorrer esa distancia será proporcional a ella. Entonces, como se mencionó anteriormente:

Trabajo = Fuerza por distancia recorrida.

Cuando la fuerza tiene la misma dirección que el movimiento del objeto, le transmite energía y por lo tanto se realiza trabajo sobre él. Si la fuerza es perpendicular a la dirección del movimiento, ni se transmite energía, ni se realiza trabajo. Pero, ¿qué sucede cuando la fuerza forma un ángulo agudo con la dirección del movimiento del cuerpo? La fuerza entonces posee dos componentes, una en la dirección del movimiento del cuerpo y otra perpendicular a esta dirección. La componente de la fuerza en la dirección del movimiento es la que realiza trabajo, mientras que la perpendicular no.

Así que podemos decir: el trabajo realizado sobre un cuerpo es igual al producto de la componente de la fuerza a lo largo de la dirección de movimiento multiplicado por la distancia recorrida.

2.15 LAS LEYES DE CONSERVACIÓN. LA CONSERVACIÓN DE LA CANTIDAD DE MOVIMIENTO O ÍMPETU

Se ha mencionado la ley que dice que la fuerza aplicada sobre un cuerpo de masa m es igual al producto de esta masa por la aceleración. Ahora se necesita de otra que explique cuáles son los efectos sobre el movimiento que produce esa fuerza.

Considerando un sistema aislado compuesto de un conjunto de canicas en movimiento, se puede hablar de las velocidades que llevan, pero para completar su descripción es necesario hacer referencia a sus masas.

Conociendo la masa y la velocidad de un objeto en movimiento, es posible conjuntarlas en la magnitud cinética más simple que se puede formar, que es el vector cantidad de movimiento:

$$\vec{p} = m\vec{v},$$

y es una característica que poseen los cuerpos en movimiento, que perdura aunque haya cesado la causa que la produjo.

Si sobre un cuerpo no actúa ninguna fuerza, su velocidad permanecerá constante y, por lo tanto, su ímpetu también será constante:

$$\vec{p} = m\vec{v} = \text{constante}.$$

Ésta es otra forma de presentar la *Ley de la Inercia*.

Si pensamos en dos cuerpos que chocan con velocidades v_1 y v_2, antes del choque y v'_1 y v'_2 después del choque y las medimos, el experimento muestra que:

$$-\frac{(v_1 - v'_1)}{(v_2 - v'_2)} = k_{12},$$

donde k_{12} es una constante positiva.

Si se repite el experimento con otras velocidades iniciales, obtenemos que:

$$-\frac{(u_1 - u_1')}{(u_2 - u_2')} = k_{12},$$

tiene el mismo valor que antes y este valor es

$$k_{12} = \frac{m_2}{m_1}.$$

Por lo tanto,

$$\begin{aligned} m_1 v_1 - m_1 v_1' &= m_2 v_2' - m_2 v_2 \\ m_1 v_1 + m_2 v_2 &= m_1 v_1' + m_2 v_2' \\ p_1 + p_2 &= p_1' + p_2' = \text{constante.} \end{aligned}$$

Si tenemos un conjunto de cuerpos

$$p_1 + p_2 + p_3 + \cdots = P = \text{constante.}$$

Y ésta es la ley de la conservación de la cantidad de movimiento para un sistema aislado, es decir, en ausencia de fuerzas externas.

2.16 LA ENERGÍA CINÉTICA

Descartes pensaba que la cantidad cinética que explicaba el efecto de una fuerza sobre un cuerpo, era el ímpetu comunicado a éste.

Por su parte, Leibniz (1646-1716) propuso que el "efecto de una fuerza" era proporcional, no a v, sino a v^2 y que esta cantidad se conservaba en los fenómenos mecánicos.

Siguiendo sus ideas, definiremos la energía cinética de un cuerpo como:

$$T = \frac{1}{2} m v^2,$$

y la de un sistema de n cuerpos como

$$T = \sum_i \frac{1}{2} m_i v_i^2 .$$

En términos de este concepto, se define un choque elástico como aquél en el cual se conserva la energía cinética, e inelástico en el que no. Sin embargo, la cantidad de movimiento se conserva en ambos tipos de choque.

2.16.1 Teorema Trabajo-Energía Cinética

Ahora bien, como $F = ma$, el trabajo realizado por la fuerza sobre el cuerpo será:

$$F\,x = ma\,x,$$

pero como vimos anteriormente, $v^2 = 2ax$, así que

$$ax = \frac{v^2}{2}.$$

Sustituyendo en la ecuación anterior

$$Fx = m\frac{v^2}{2}.$$

Por lo tanto, el trabajo realizado para acelerar un cuerpo de masa m desde el reposo hasta la posición *x* será igual a la energía cinética ganada por él.

2.17 ENERGÍA POTENCIAL

Una vez revisado lo relativo a la energía asociada al movimiento: la energía cinética, se analizará una energía asociada a la posición de un objeto respecto a otro. Si se estira una liga o se levanta del suelo un objeto a una cierta altura, se almacena una cierta cantidad de energía que llamamos *energía potencial*.

En el caso gravitacional, es decir, lo que sucede cuando dejamos caer el objeto que levantamos del suelo a una altura h_1, la aceleración es $(-g)$, y x es igual a h_1; entonces:

$$-gh_1 = \frac{v_1^2}{2}.$$

Multiplicando ambos lados por la masa *m* del cuerpo, se tiene:

$$-mgh_1 = m\frac{v_1^2}{2}.$$

El lado izquierdo de la ecuación es la energía potencial del cuerpo en reposo antes de soltarlo, y el lado derecho es su energía cinética que tendrá al llegar al suelo.

Considerando ahora la suma de la energía potencial y la cinética en cualquier punto de su trayectoria de caída:

$$mgh + m\frac{v^2}{2} = E = \text{ Energía mecánica total.}$$

Para el punto más alto, cuando $h=h_1$, la velocidad del objeto es cero; por lo tanto, la energía mecánica total es:

$$E = mgh_1.$$

En el punto más bajo, $h = 0$:

$$E = m\frac{v_1^2}{2}.$$

Por lo tanto, $-mgh_1 = mv_1^2/2$ indica que la energía mecánica total se conserva en cualquier punto de la trayectoria. Este teorema de conservación es válido sólo en ausencia de fuerzas disipativas o de fricción.

La conservación de la energía total permite calcular la velocidad dada la altura en cualquier punto de la trayectoria, o viceversa. Una aplicación inmediata de esta ley de conser-

vación podría ser el cálculo de la velocidad que llevará un carrito en una montaña rusa, dada la altura h_1 de la cima inicial:

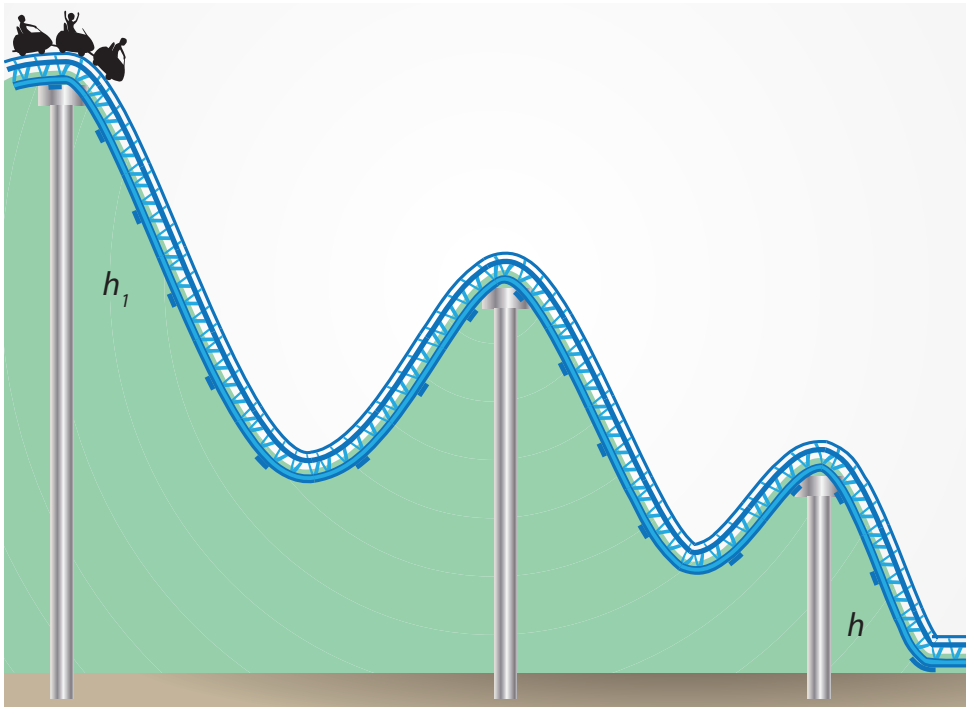
$$E = mgh_1 = mgh + m\frac{v_1^2}{2}.$$

Para una altura h cualquiera tendremos que la velocidad es:

$$v^2 = 2g(h_1 - h);$$
$$v = [2g(h_1 - h)]^{1/2}.$$

Esto es válido independientemente de las cimas que haya subido o bajado entre su posición inicial y la final a la altura h.

Nótese que si $h = 0$, recuperamos el valor de la velocidad de caída $v_1 = (2gh_1)^{1/2}$, y si $h = h_1$, entonces $v = 0$.



Montaña rusa.

ELECTRICIDAD Y MAGNETISMO

TEMA

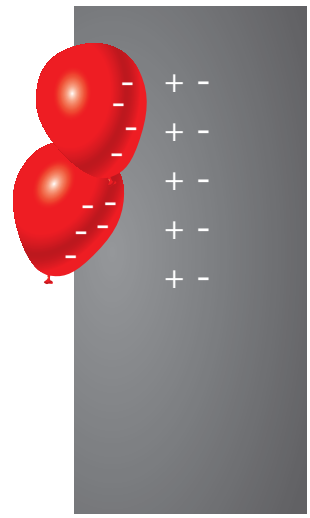
3

© Latin Stock México.
[Véase video en CD: "Introducción a la electrostática".]

3.1 CARGA ELÉCTRICA

¿Por qué, a veces, al tocar a otra persona o algún objeto, sentimos "un toque"? ¿Por qué caen los rayos? Para responder estas y muchas otras preguntas se debe entender cómo se comporta la electricidad en ciertas condiciones, para lo que es necesario comenzar por analizar el motivo por el cual dos objetos se atraen.

3.1.1 Conservación de la carga



Globos cargados. [Véase video en CD: "Fenómenos electrostáticos".]

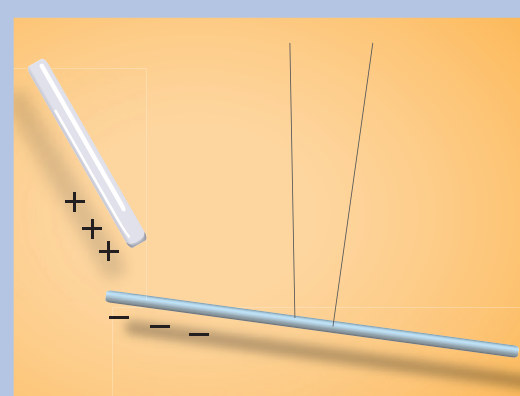
En el año 600 a.n.e., Tales de Mileto, uno de los más grandes pensadores de la antigua Grecia, observó la atracción que ejercía el ámbar (una resina amarilla y dura) sobre cuerpos ligeros después de frotarlo con pieles. Él pensó que el ámbar adquiría un "alma" que le confería la propiedad de atraer objetos, descubriendo así lo que ahora llamamos electricidad (que proviene de la palabra griega *electrón*, que precisamente significa ámbar).

Para reproducir su experimento, se puede sustituir el ámbar por globos y las pieles por nuestro cabello. Al frotar el globo con el cabello y acercarle pedacitos de papel se observa que éstos son atraídos por él; si después se amarran dos globos con un hilo, se frotan y se sostiene el hilo por la mitad dejando que los globos cuelguen, se observa que se repelen (véase figura).

Las observaciones de Tales de Mileto no se formalizaron científicamente sino hasta muchos siglos después. Stephen Gray (1666-1736) descubrió, frotando diferentes materiales, que existían dos clases de electricidad: la vítrea (que se producía al frotar vidrio) y la resinosa (que se producía al frotar resinas, como el ámbar). También concluyó que cuerpos cargados con el mismo tipo de electricidad se repelen y los cargados con diferente tipo se atraen. Eso significa que hay una fuerza entre ellos que los jala o los empuja, aunque no estén en contacto.

Experimento A

a] Suspender con un hilo un popote por su punto medio, permitiéndole girar libremente.
b] Frotar en el cabello para cargarlo.
c] Ahora frotar un tubo de ensayo y acercarlo a uno de los extremos del popote.
¿Qué pasa?
d] Ahora frotar un tubo de plástico (u otro popote) y acercarlo al popote.
¿Qué se observa?
¿Qué conclusiones se pueden sacar del experimento?
[Véase video en CD: "Experimento de Gray".]



Se puede observar que con la varilla de vidrio ocurre el fenómeno contrario que con el tubo de plástico. Se concluye, entonces, que existen dos comportamientos opuestos en el experimento que claramente dependen del material que se cargó. Se puede pensar que el tubo de plástico se cargó igual que el popote suspendido, por ser del mismo material, y se repelió. Con el vidrio ocurrió lo contrario, es decir, lo atrajo. Entonces, los cuerpos cargados con el mismo tipo de electricidad se repelen y los cargados con diferente tipo se atraen, como lo concluyó Stephen Gray.

Experimento B

Para entender mejor esta clasificación, se propone la siguiente experiencia:

a] Colgar de un hilo una bolita de unicel forrada de papel aluminio, poniéndola en contacto, en un extremo, con una regla metálica y acercando un globo cargado al otro extremo. Se observará que la esfera es repelida por la regla.
b] Cambiar la regla metálica por una de plástico y repetir el experimento. Se observará que ahora la bola no se mueve.

Lo único que se cambió en la segunda actividad del experimento fue la regla metálica por la de plástico, pero ambos materiales tuvieron diferente comportamiento. La diferencia se debe a que en el metal las partículas eléctricas son capaces de desplazarse a través de él, lo que no ocurre con el plástico.

A los materiales que en el experimento se comportan como el metal se les llama conductores, y a los que se comportan como el plástico se les llama aislantes.

[Véase video en CD: "Conductores y aislantes".]

¿Cómo se relaciona todo esto con la electricidad? Esto se explica porque la materia está formada por átomos, los cuales a su vez están conformados por partículas eléctricas que siempre existen por pares: una negativa unida a una positiva. En algunos materiales es más fácil arrancar las partículas negativas que las positivas, quedando el objeto con exceso de éstas y entonces se dice que el objeto quedó cargado positivamente.

Que sea más fácil arrancar un tipo de partículas en un caso que en otro, se debe a las características de cada material. En el experimento anterior, al vidrio se le pueden arrancar partículas diferentes (negativas) de las que se les puede arrancar al plástico (positivas). Esto ha permitido clasificar a los materiales en conductores y aislantes (experimento B).

Para saber cómo está cargado eléctricamente un material se puede construir un sencillo aparato (electroscopio) que detecta la carga (experimento C). Esto se explica porque existe el mismo número de cargas positivas y negativas distribuidas de manera uniforme en el aluminio, pero al acercar el globo, éste atrae las cargas de signo contrario y repele a las del mismo signo hacia el extremo de la "L" (experimento C, p. 65) y, por lo tanto, hacia el rectángulo de aluminio, lo que explica que ambos tengan el mismo tipo de carga.

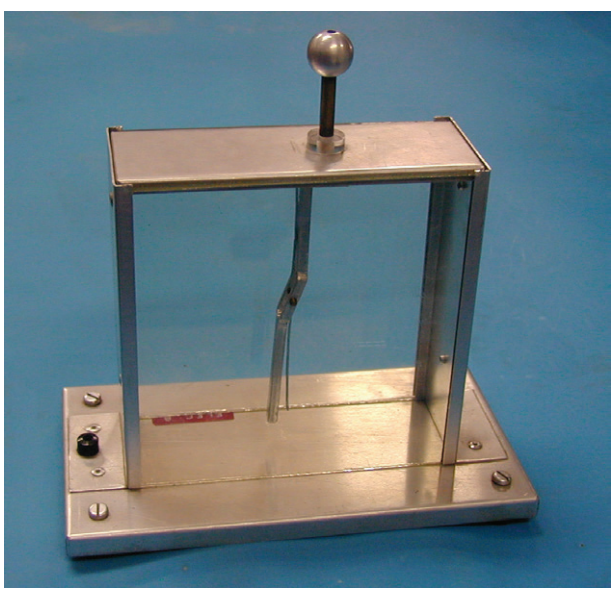


Figura 1. Carga por inducción en un electroscopio | © Latin Stock México.

El aparato construido es un electroscopio muy rudimentario. En la foto se muestra un electroscopio como el que se utiliza en los laboratorios y sirve para detectar la presencia de cargas eléctricas (figura 1).

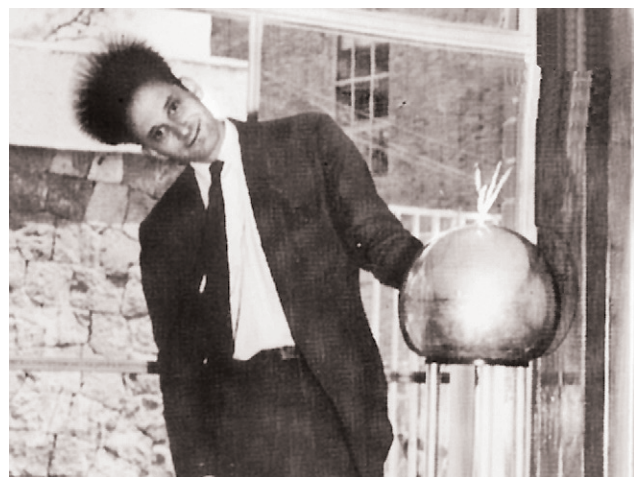
Hasta ahora sólo se han cargado los cuerpos frotándolos. Otra forma de hacerlo es si se acerca el globo cargado al aparato que se construyó (electroscopio), sin tocarlo. Cuando la hoja de aluminio esté levantada, se toca momentáneamente con el dedo y se retira el globo. ¿Qué ocurre? ¿Cómo es que la hoja de aluminio permanece separada aun al retirar el globo?

Experimento C

a] Recortar un rectángulo de cartón de unos 5 por 15 cm y forrarlo con papel aluminio.
b] Doblarlo en forma de "L" y pegarlo con cinta adhesiva a un vaso desechable.
c] Cortar otro rectángulo de papel aluminio de unos 3 por 4 cm.
d] Hacerle un pequeño doblez en la parte superior y colocarlo como se muestra en la ilustración.
e] Frotar un globo con el cabello y acercarlo a la base de la "L".
f] Se podrá observar que la lámina de papel aluminio se separa, formando un ángulo.
[Véase video en CD: "Construyendo un electroscopio".]

Con lo que se explicó hasta ahora se puede entender un principio: había el mismo número de cargas negativas y positivas, por lo que, al acercar el globo cargado al electroscopio se atraen las cargas positivas y se alejan las negativas. Cuando se toca la hoja de aluminio, estas cargas pasan a nuestro cuerpo, quedando así un exceso de cargas positivas en la "L" y, ya que están en contacto, tendrán el mismo signo, por lo que se repelerán, formando un ángulo. A este fenómeno se le llama "carga por inducción" (figura 1, p. 352).

Como es posible cargar objetos frotándolos con otro material, se desarrollaron máquinas que producen electricidad llamadas generadores Van de Graff. Por medio de un motor eléctrico se hace girar una banda de hule en contacto con cerdas, transportando la carga a una bola hueca de metal. Se pueden ver demostraciones de estos aparatos en algunos museos científicos. ¿Por qué se le eriza el cabello al brillante científico mexicano, fallecido en 1988, doctor Tomás Brody (en la foto)? Al igual que ocurre con el electroscopio, la bola de metal se carga eléctricamente. Cuando la mano de Brody entra en contacto con ella, todo su cuerpo se carga, incluyendo el cabello, que tiene la misma carga, por lo que las puntas se repelen unas a otras y se erizan.



Generador Van de Graff. [Véase video en CD: "Acelerador Van de Graff".]

3.1.2 Ley de Coulomb

Para los físicos no es suficiente con saber que existe una fuerza eléctrica debida a las cargas, sino que se debe conocer cuantitativamente cuál es su magnitud. Para esto se realizó un experimento simple que consistió en considerar dos cargas y la distancia entre ellas. El primero en estudiar cuantitativamente las fuerzas eléctricas fue el físico francés Charles Coulomb (1736-1806), quien en 1785 realizó su experimento construyendo una balanza de torsión (véase figura 2, p. 354), aparato cuyo principio es similar al que se usó en el experimento de los popotes para recuperar las observaciones de S. Gray.

Cuando la esfera cargada 1 se sitúa a una distancia determinada de la esfera cargada 2, la fuerza eléctrica sobre la esfera 2 hace girar el brazo horizontal del aparato, el cual queda

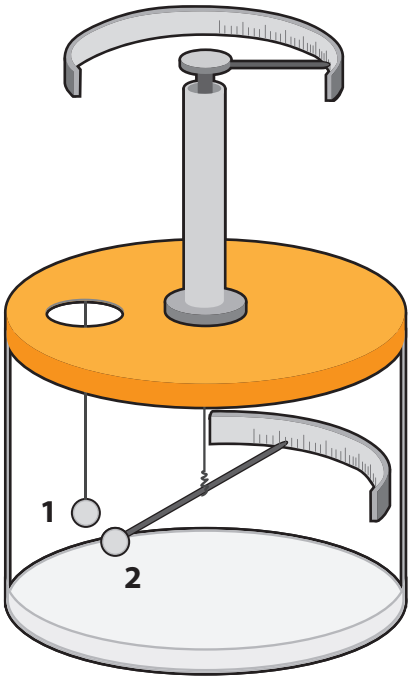


Figura 2. Experimento de la balanza de torsión, de Charles Coulomb. [Véase simulación en CD: "Ley de Coulomb".]

en reposo en una nueva posición, con el hilo retorcido. A mayor torsión corresponde una mayor fuerza. De esta forma, Coulomb podía medir la fuerza eléctrica en función del ángulo Coulomb de torsión. Variando la separación entre las esferas cargadas, determinó la fuerza como función de la separación (figura 2).

Utilizando esferas cargadas, positivas y negativas, Coulomb demostró que la fuerza es siempre inversamente proporcional al cuadrado de la distancia que existe entre ellas. Esto lo llevó a enunciar: "La fuerza de atracción o de repulsión entre dos cargas puntuales es directamente proporcional al producto de las dos cargas e inversamente proporcional al cuadrado de la distancia que las separa." Este enunciado se conoce ahora como la Ley de Coulomb. Matemáticamente se puede escribir así:

$$F = k\frac{q_1 q_2}{r^2}.$$

Dicha ley se puede relacionar con la ley de los signos, que afirma que si se multiplican dos números que tienen el mismo signo, el resultado será positivo, y si su signo es diferente se obtendrá un número negativo; en el caso de la ley de Coulomb: si las cargas son del mismo signo, se obtiene una fuerza positiva (se repelen) y si tienen signo diferente el resultado es una fuerza negativa (se atraen). Entonces la fuerza eléctrica tiene magnitud y dirección.

En el Sistema Internacional de Unidades, las cargas se miden en coulombs (C), la distancia en metros (m) y la fuerza en newtons (N); así, las unidades de la constante k son:

$$\text{newtons} = [k]\frac{\text{coulombs}^2}{\text{metros}^2},$$

en donde, despejando la k, se tiene:

$$[k] = \frac{\text{newtons metros}^2}{\text{coulombs}^2},$$

y su valor es de:

$$k = 9 \times 10^9 \frac{\text{Nm}^2}{\text{C}^2}.$$

En el capítulo anterior se estudió la Ley de la Gravitación Universal, que explica el movimiento de los planetas alrededor del Sol, que depende de las masas de los mismos. Matemáticamente se expresa de la siguiente manera:

$$F = G\frac{m_1 m_2}{r^2}.$$

Como se puede observar, ambas leyes presentan expresiones matemáticas similares y son ejemplos de fuerzas de acción a distancia, es decir, que los cuerpos o cargas de los que hablamos no están en contacto.

3.1.3 Campo eléctrico

Así como se habla del campo gravitacional, se puede hablar del campo eléctrico que se aplica a los objetos cargados eléctricamente. Como se sabe, alrededor de la Tierra existe un campo gravitacional, que es el que nos mantiene unidos a ella. De la misma forma existe, alrededor de un cuerpo cargado, un campo eléctrico. Una persona puede percibir el campo eléctrico de un generador Van de Graff o de la pantalla de un televisor, si acerca la mano a ellos.

Si se coloca una carga q_1 en el campo eléctrico E, experimenta una fuerza dada por:

$$F = q_1 E,$$

despejando:

$$E = \frac{F}{q_1}.$$

Si se sustituye el valor de F se obtiene:

$$E = \frac{kq_1 q_2}{r^2}\frac{1}{q_1} = k\frac{q_2}{r^2}.$$

En otras palabras, la fuerza eléctrica que una carga ejerce sobre la otra se puede describir como la interacción entre la carga y el campo eléctrico producido por la otra. En el Sistema Internacional de Unidades, las unidades de E son:

$$[E] = \frac{\text{newtons}}{\text{coulombs}}.$$

3.1.4 Potencial eléctrico

Cuando se estira una liga, en términos físicos, se dice que se almacena en ella energía potencial, es decir, el trabajo que costó estirarla queda "guardado" en forma de energía dentro de las partículas de la liga (al vencer la fuerza de restitución, que es la que hace que la liga recupere su forma) y está listo para que, en cuanto se suelte, se transforme ahora en lo que se llama energía cinética, que es la energía que posee un cuerpo al moverse.

Lo mismo ocurre cuando, en una montaña rusa, suben los carros a lo alto de una pendiente. Mientras ascienden están acumulando energía, producto de la fuerza que se está venciendo (que ahora es la fuerza de gravedad de la Tierra). Otros ejemplos son el resorte que se comprime en un juego de pinball para lanzar la bola, la cuerda de un reloj o un juguete, etcétera. De esta forma, cuando existe una fuerza entre dos cargas con diferente signo y se tratan de alejar, se sentirá resistencia porque, como en el caso de estirar la liga, se acumula energía potencial. A esto se le llama realizar trabajo.

Se puede entender el funcionamiento de una pila eléctrica pensando que está formada por muchas cargas que fueron empujadas unas contra otras acumulando energía, como ocurre en el resorte (figura 3, p. 356). En las figuras 4-6 (p. 356) se muestra la siguiente analogía: al subir a lo alto de una pendiente se acumula energía, la cual se libera al descender, lo mismo que en una pila eléctrica se acumula energía. Si la montaña rusa está

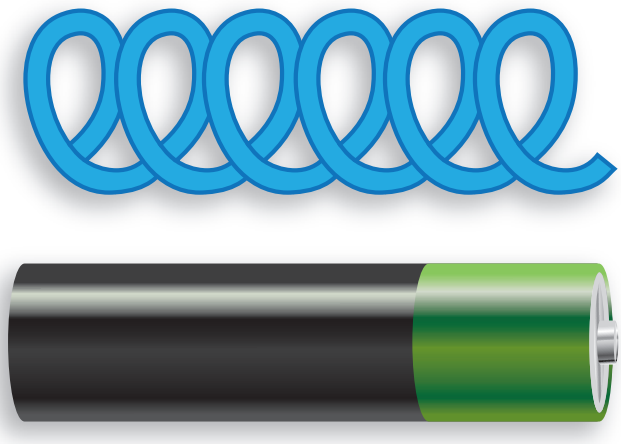


Figura 3. Representación de energía potencial de una pila como un resorte comprimido.

montada desde una pila, como se puede ver en la figura, el carrito descenderá y liberará su energía. ¿Qué pasa si se ponen dos pilas una junto a la otra? ¿Cuál será la diferencia de energía liberada al descender? (figuras 4-6).

Si se piensa en dos pilas, pero una junto a la otra, de las cuales desciende una montaña rusa de cada una, ¿cuál es la diferencia de energía liberada?

Si se tiene una pila sobre la otra, la energía liberada será el doble que la que existía cuando se tenía una sola. Cuando se tienen dos pilas una junto a otra, dado que se tienen dos pendientes con un carrito cada una, se puede transportar el doble de personas. Si se piensa en ellas como partículas eléctricas, se habrán transportado el doble de éstas, a diferencia del caso en el que se tenía una pila sobre otra, y en este caso bajan más rápido ya que la altura inicial es mayor.

Cuando el carrito se desplaza desde una altura mayor (lo que equivale a una energía potencial mayor) hasta una altura menor (una energía potencial menor), a la diferencia de energías potenciales (relacionadas en el ejemplo con la diferencia de alturas) se le conoce como diferencia de potencial.

En las pilas se puede leer un número, el más común es 1.5 V, que representa el valor de la diferencia de potencial acumulado en ellas. La v se refiere a volts (en honor a Alessandro Volta (1745-1827), que es la unidad de medida en el Sistema Internacional para diferencia de potencial.

Con base en lo anterior, podemos decir que cuando algo nos dio "toques" es por el exceso de carga que se libera en el momento en que tocamos el objeto que tiene un potencial menor. Lo mismo sucede con los rayos que caen sobre la tierra, aunque con un potencial mayor, por el frotamiento entre las nubes y el aire.

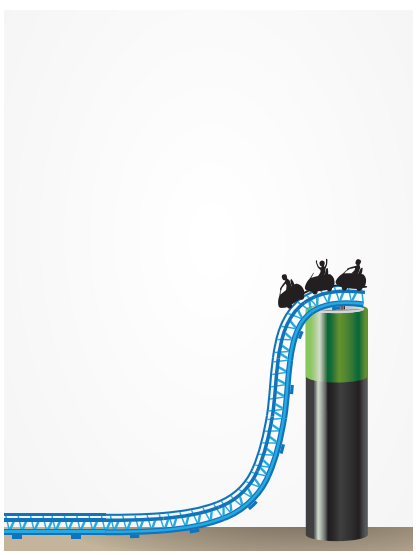


Figura 4. Montaña rusa con una pila.

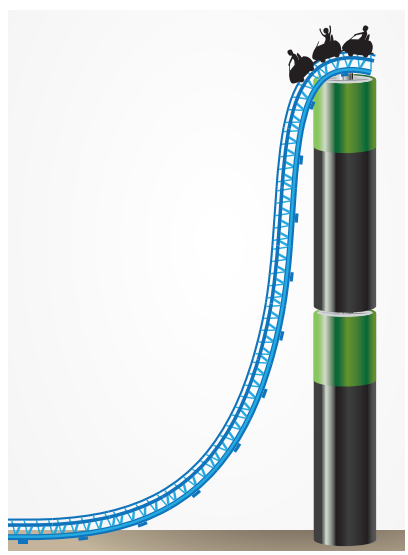


Figura 5. Montaña rusa con dos pilas, una encima de la otra.

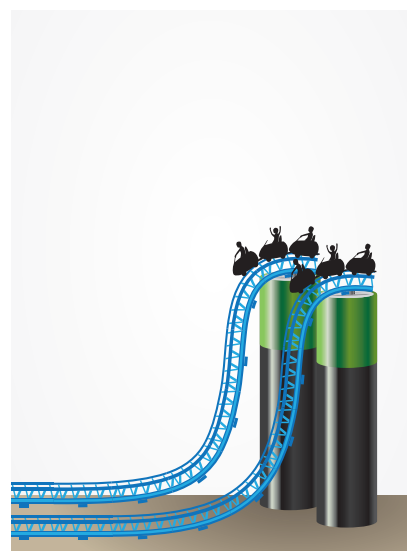


Figura 6. Montaña rusa con dos pilas, una junto a la otra.

3.2 NOCIONES DE CIRCUITOS SIMPLES

Hasta aquí se ha aprendido que existen partículas eléctricas que pueden moverse a través de materiales conductores y que éstas pueden adquirir energía potencial, como en el ejemplo de la montaña rusa. Ahora imaginemos la montaña rusa completa, es decir, con una cima principal, algunos rizos, una salida y una llegada: un recorrido cerrado, esto es, un circuito. [Véase simulación en CD: "Interruptores serie y paralelo".]

3.2.1 Circuitos

De manera intuitiva se pueden imaginar muchas partículas cargadas moviéndose juntas y formando un flujo, como el torrente de un río. Si se define una cantidad que llamaremos corriente, como la cantidad de carga q por cada unidad de tiempo, entonces:

$$I = \frac{q}{t}.$$

Esta corriente fluye por un material conductor (metal) y, como se vio, para que exista esta corriente se requiere de una diferencia de potencial en el circuito. A principios del siglo XIX, el científico George Ohm (1789-1854) descubrió que la corriente en los metales es proporcional a esta diferencia de potencial, a la que llamó voltaje.

$$V \propto I,$$

Adicionalmente descubrió que la constante de proporcionalidad es una propiedad del conductor, a la que llamó resistencia. Se puede interpretar como la oposición al flujo de corriente, enunciando así la ley que lleva su nombre:

$$V = RI.$$

En el Sistema Internacional, las unidades para el voltaje V son los volts, y las unidades de la corriente I son los amperes. Entonces, las unidades de la resistencia R serán volts/amperes. A esta unidad se le denomina ohm, en honor a su descubridor. En la vida cotidiana nos encontramos con una gran cantidad de aplicaciones en el diseño de aparatos eléctricos, esto es, combinando en un circuito diferentes elementos, como pilas, resistencias, capacitores, transistores, etc., en diferentes arreglos. Por ahora sólo se hablará de pilas y resistencias conectadas en serie y en paralelo, como se puede ver en la figura 7 (p. 70).

Se llama "resistencia" a los dispositivos conductores con un valor característico de R en la ley de Ohm y se representa por una línea en zigzag, en la figura por R_1, R_2 y R_3.

De la ley de Ohm, la resistencia equivalente en el circuito, que es una cantidad que representa la contribución de todas las resistencias involucradas, está dada por:

$$R = \frac{V}{I}.$$

En un circuito con resistencias conectadas en serie circula la misma corriente; entonces, como se vio anteriormente, el voltaje total será $V = V_1 + V_2 + V_3$, así que:

$$\frac{V}{I} = \frac{V_1}{I} + \frac{V_2}{I} + \frac{V_3}{I}.$$

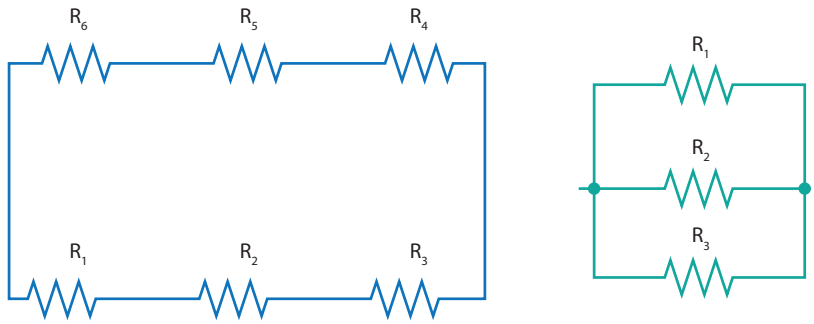
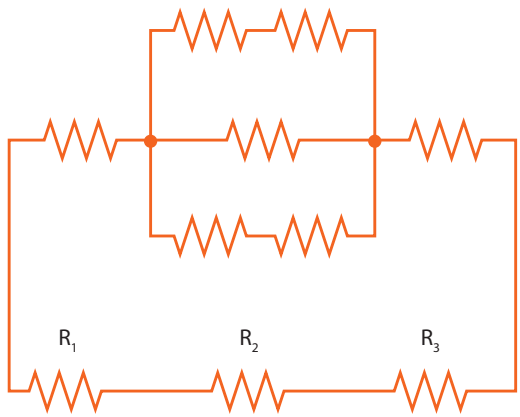


Figura 7. Resistencias en serie y en paralelo. [Véase simulación en CD: "Cálculo de resistencias en serie y en paralelo".]



$Rt = R_1 + R_2 + R_3$

Figura 8. Resistencias en serie y en paralelo.

Entonces, cuando las resistencias están conectadas en serie, la resistencia equivalente será igual a la suma de cada resistencia, es decir:

$$R_e = R_1 + R_2 + R_3.$$

Cuando se tienen resistencias conectadas en paralelo, la corriente total que circula es la suma de las corrientes que fluyen por cada resistencia. Imaginemos la tubería de una casa, donde hay una toma principal de la cual se derivan varias tuberías más pequeñas para el lavabo, la regadera, el fregadero. Entonces:

$$I = I_1 + I_2 + I_3.$$

Pero como la pila es la misma, el voltaje será igual, por lo que:

$$\frac{I}{V} = \frac{I_1}{V} + \frac{I_2}{V} + \frac{I_3}{V}.$$

Así, cuando las resistencias están conectadas en paralelo, el inverso de la resistencia equivalente es la suma del inverso de cada resistencia. De modo que:

$$\frac{1}{R_e} = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3};$$

entonces se pueden simplificar circuitos complicados dividiéndolos en combinaciones de circuitos sencillos conectados en serie y en paralelo, como se puede ver en la figura 8.

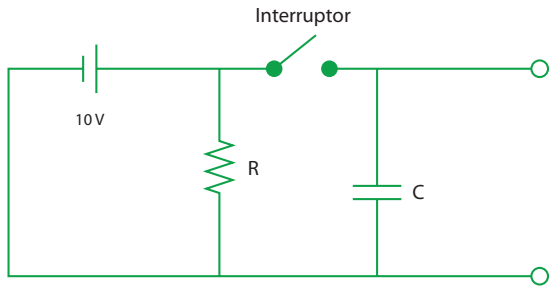


Figura 9. Circuito RC. [Véase simulación en CD: "Circuito RC".]

Otro elemento que se puede estudiar es el capacitor, el cual no es más que un arreglo de dos placas conductoras paralelas separadas por un aislante (generalmente plástico); entonces, si cada placa se carga eléctricamente con carga opuesta, las partículas cargadas se atraerán unas con otras, por lo que se quedarán almacenadas en las placas tratando de acercarse, hasta que exista algún medio por el cual se conecten eléctricamente ambas placas.

Un circuito muy conocido es el llamado circuito RC, donde se combina una resistencia R con un capacitor C, como se observa en la figura 9. Al cerrar el interruptor, se carga el capacitor; se puede imaginar como un tanque que se llena de agua; al abrir nuevamente el interruptor, la resistencia actuará como un tubo con un diámetro pequeño por el cual se vacía el tanque más lentamente de lo que se llenó.

3.2.2 Potencia eléctrica

Siempre que una carga eléctrica se mueve en un circuito, realiza un trabajo, el cual se utiliza para hacer funcionar algún elemento del mismo, como prender un foco o hacer girar un motor. Así que la potencia eléctrica es la rapidez con la que se realiza este trabajo, o bien, la energía que consume una máquina o cualquier dispositivo eléctrico en un segundo.

Para encontrar la potencia eléctrica recordemos la analogía de la montaña rusa que ejemplifica la energía potencial. Cuando se sube al carrito a la cima a una altura h sobre el nivel del suelo, tendrá una energía potencial y podrá bajar por el riel realizando un trabajo equivalente. Es decir:

$$T = E_p.$$

Entonces, si se considera que el voltaje o la diferencia de potencial se puede expresar como la energía potencial por cada unidad de carga, tendríamos:

$$V = \frac{E_p}{q},$$

entonces, $V = T/q$, y por lo tanto, $T = Vq$.

Ahora, la potencia es la rapidez con la cual se realiza un trabajo, es decir:

$$P = \frac{T}{t} = V\frac{q}{t},$$

y, recordando que la corriente $I = q / t$, llegamos a que la potencia eléctrica en watts es el producto de la corriente en amperes por el voltaje en volts, esto es:

$$P = VI.$$

3.3 NOCIONES DE ELECTROMAGNETISMO

Hace aproximadamente dos mil años, en la ciudad griega de Magnesia, se encontraron unas piedras de imán que atraían pedazos pequeños de hierro a las cuales se les dio el nombre de magnetitas. A este fenómeno de atracción se le conoce como magnetismo.

3.3.1 Campo magnético

Si se tienen dos imanes de barra y se acerca uno al otro, se observará que por un extremo se atraen ambas puntas, pero si se voltea un imán entonces se repelen. Con los mismos imanes, si se une uno al otro, como se muestra en la figura 10, marcando sus extremos con blanco y negro alternadamente y ahora se juntan a lo ancho, ¿cómo quedan las marcas? Si los extremos se ponen alternados se atraen, pero si se ponen blanco con blanco y negro con negro se repelen.

Experimento D | Magnetismo

Para ver lo que ocurre con la brújula, se puede hacer el siguiente experimento: imantar un clip frotándolo con un imán; después, colocarlo sobre un pedazo de unicel a modo de flotador y el conjunto en un recipiente con agua. Se podrá observar que se orienta espontáneamente en una dirección, al extremo que apunta hacia el norte le llamaremos N y al otro S. Lo que atrae al clip es otro gran imán que forman los polos Norte y Sur de la Tierra.

[Véase video en CD: "Jugando con imanes".]

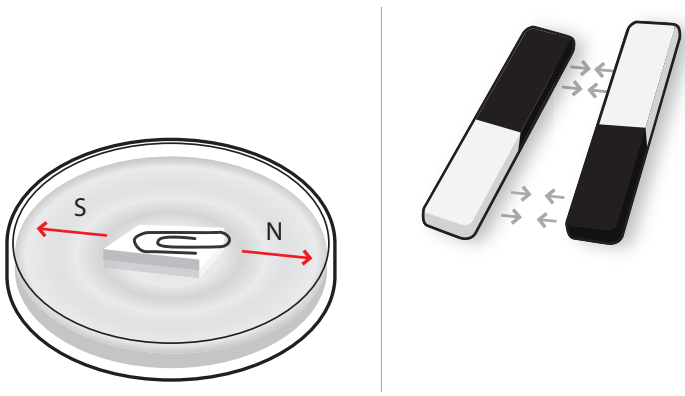


Figura 10. Brújula casera y atracción de imanes.

En el siglo XI se descubrió que una aguja imantada montada libremente se orientaba. Entonces, una brújula es un imán formado por una aguja ligera de acero imantada que se apoya sobre un soporte con muy poca fricción. La invención de la brújula como tal se atribuye a los chinos. Existe una leyenda que dice que el emperador Huang-ti, durante una batalla en la niebla en el año 2634 a.n.e., utilizó un carro con una figura humana que señalaba siempre el sur para orientar a las tropas. En realidad, la primera referencia escrita del uso de la brújula por los chinos data del siglo XI. Entre los árabes se menciona por primera vez en 1220. Probablemente fueron ellos quienes la introdujeron en Europa, donde no tardó en ser adoptada por los vikingos.

Brújula | © Latin Stock México.

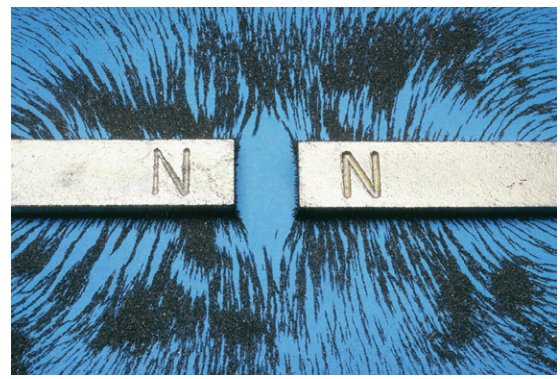


Figura 11. Campos magnéticos | © Latin Stock México.

En cualquier región en donde una aguja imantada se orienta, se dice que existe un campo magnético. Así como se puede observar el campo eléctrico con su movimiento, también se pueden observar las líneas de fuerza del campo magnético. Si sobre una mesa se colocan un imán y sobre él una cartulina, y después se esparce sobre la cartulina limaduras de hierro, se observará el efecto del campo magnético sobre las mismas al formar líneas cerradas (figura 11).

3.3.2 Materiales ferromagnéticos, paramagnéticos y diamagnéticos

No todos los metales son atraídos por un imán. Si se hace la prueba de acercar varios metales a un imán, uno se dará cuenta de que el aluminio de una lata de refresco no es atraído por éste. Entonces, al igual que existen materiales conductores y aislantes, existen materiales ferromagnéticos, paramagnéticos y diamagnéticos.

Los materiales ferromagnéticos y paramagnéticos son aquellos que son atraídos por el imán, mientras que los diamagnéticos no; sin embargo, los paramagnéticos pueden ser atraídos con menor intensidad, al grado que en ocasiones es difícil diferenciarlos de los diamagnéticos.

3.3.3 Bobinas, campos magnéticos y corrientes eléctricas

El efecto de un campo magnético sobre una carga en movimiento se presenta de forma similar a la fuerza que siente una partícula cargada en un campo eléctrico; se calcula como:

$$F = qE.$$

Es posible calcular la fuerza que siente ahora una partícula cargada en movimiento debida a un campo magnético; ésta estará dada por:

$$F = q\ v\ B \text{ sen } \theta,$$

donde v es la velocidad de la partícula, q su carga, B la magnitud del campo magnético y θ el ángulo que forman la velocidad con las líneas de campo.

Generalmente en los aparatos que usan este principio, el ángulo θ es de 90°, de modo que el sen θ es igual a 1, así que es posible emplear la expresión

$$F = qvB.$$

Experimento E | Motor

Se puede construir un motor usando dos clips, alambre de cobre, un imán y una pila. Se enrolla el alambre de cobre en un plumón dándole unas 30 vueltas, se sujetan las puntas dejando un pedazo de alambre, el cual se montará sobre los clips y éstos sobre el vaso de unicel, como se puede ver en la figura.

Se raspa el barniz aislante a una de las puntas, completamente; y a la otra se le raspa sólo la mitad, como se muestra. Ahora se coloca el imán debajo de la bobina y se conectan los extremos de una pila a los clips.

El motor está listo. Es necesario dar un impulso inicial a la bobina para que arranque.

[Véase video en CD: "Construcción de un motor eléctrico".]

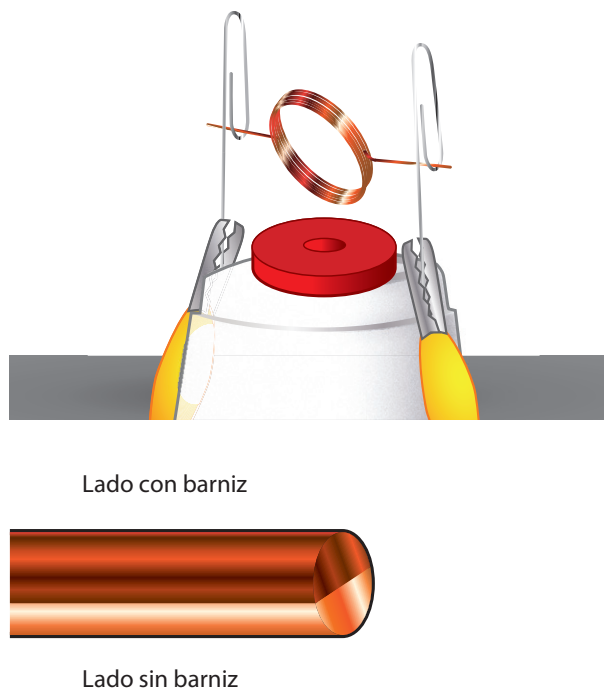


Figura 12. Experimento que muestra que la corriente eléctrica genera un campo magnético.

Esta fuerza hará que la trayectoria de la partícula se desvíe. Este fenómeno es el responsable, por ejemplo, del funcionamiento del cinescopio de una televisión. De manera muy simplificada, se puede describir pensando que se tiene un generador de partículas cargadas negativamente que salen de éste con una cierta velocidad y que pasan por un campo magnético variable (generado por un par de bobinas, las cuales se analizarán más adelante) que permite desviar a las partículas con mucha precisión, de modo que incidan en la pantalla para formar una imagen. [Véase simulación en CD: "Partícula en campo magnético".]

3.3.4 Generación de un campo magnético por una corriente eléctrica

En el año de 1820, Hans Oersted realizó un experimento para demostrarles a sus alumnos que las cargas en movimiento y los imanes no interactuaban. El experimento consistía en colocar una brújula cerca de un alambre, y la hipótesis fue que si existiera algún tipo de interacción, al hacer circular una corriente eléctrica por el alambre, la aguja de la brújula se movería. Para su sorpresa, la aguja comenzó a moverse hasta que se orientó perpendicularmente al alambre. Lo que sucedió fue que, hasta ese momento, Oersted siempre había realizado el experimento poniendo el alambre perpendicular a la aguja de la brújula, por lo que no había observado ningún movimiento en la aguja, dado que ya estaba orientada. Cuando lo realizó frente a sus alumnos, por casualidad puso el alambre y la aguja de la brújula paralelos y entonces notó el efecto, el cual cambia dependiendo de la orientación entre la brújula y el alambre. Con este experimento se demuestra que una corriente eléctrica (que es un flujo continuo de cargas eléctricas en movimiento) genera un campo magnético. Más tarde, Andre-Marie Ampere demostró que el polo norte de la aguja de la brújula se desvía siempre a la izquierda de la dirección que lleva la corriente.

Un experimento parecido al de Oersted (figura 13) consiste en usar un alambre largo enrollado, es decir, una bobina. Para construirla, se enrolla un pedazo de alambre de cobre

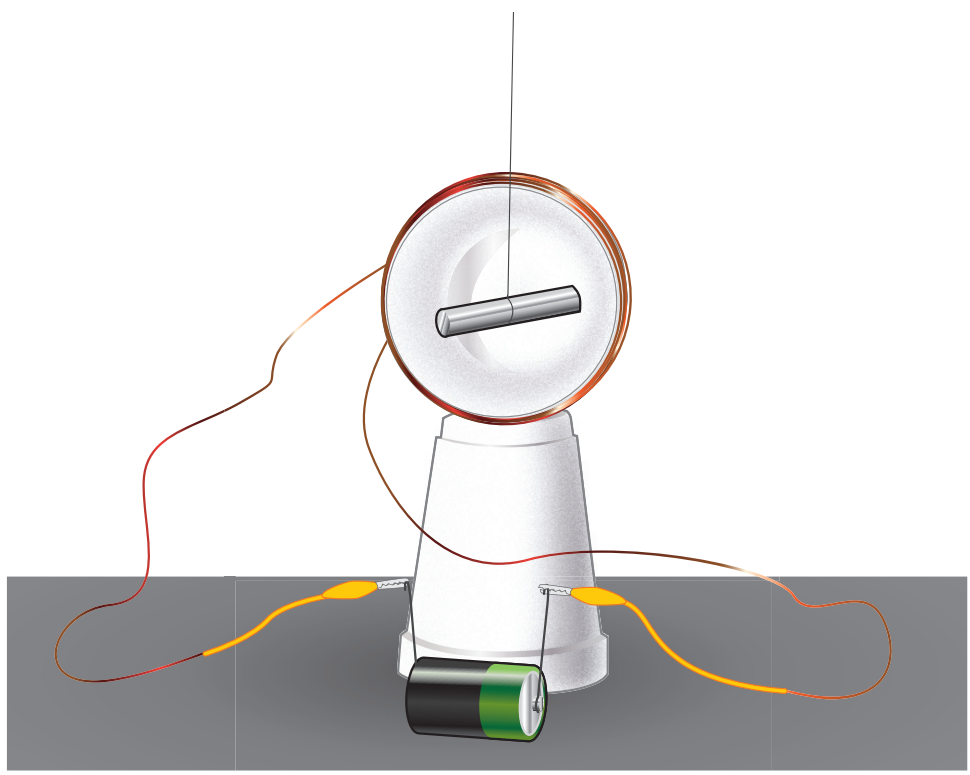


Figura 13. Modulación del sonido.

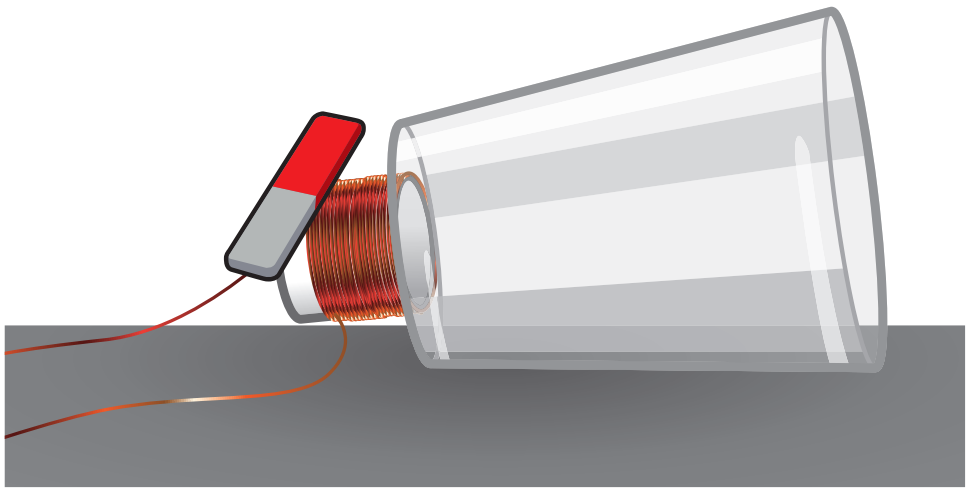


Figura 14. Bocina.

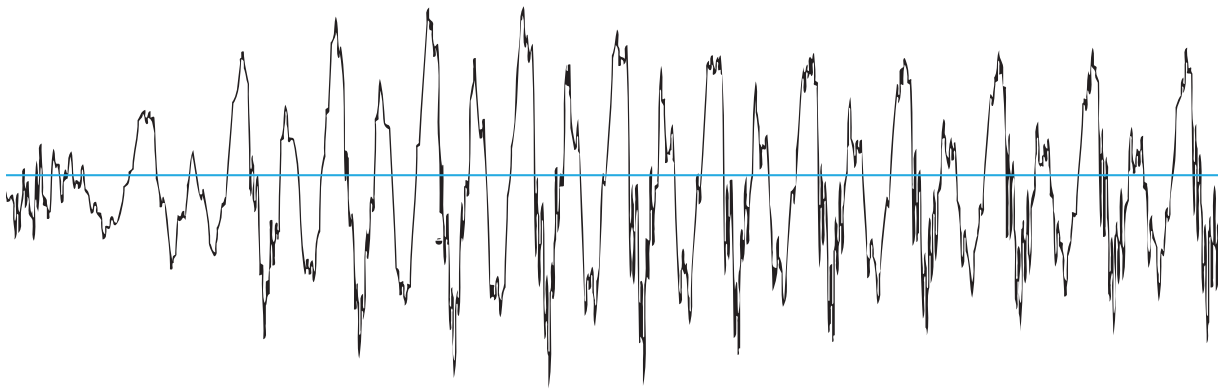


Figura 15. Modulación del sonido.

en la parte más ancha de un vaso de unicel, dándole 150 vueltas, más o menos. Se raspa con una lima de uñas el barniz aislante de las dos puntas de la bobina y se cuelga un imán de barra por su parte central. Se le acerca a la bobina, que debe estar conectada a una pila tamaño D. ¿Qué se espera que ocurra? [Véase video en CD: "Experimento para generar un campo magnético".]

Ahora, si se desconecta la pila, ¿qué sucede? Se puede repetir esta actividad identificando con etiquetas de colores los dos extremos o polos del imán. Al conectar la pila a uno de los polos del imán quedará apuntando hacia afuera del vaso. Si se intercambian las conexiones de la pila, se observa que es el otro extremo el que apunta hacia afuera. La orientación del campo magnético que genera la bobina depende de la dirección de la corriente eléctrica, tal y como demostró Ampere (figura 12, p. 361).

Una aplicación directa de este fenómeno es la fabricación de una bocina. Si se hace otra bobina más pequeña, se enrrolla el alambre a un marcador grueso dándole unas 30 vueltas, pegándola al fondo de un vaso desechable de plástico y se conectan las puntas de ésta en lugar de la bocina de un radio. Cuando se acerque un imán a la bobina, la bocina sonará (figura 14). [Véase video en CD: "Construyendo una bocina".]

Lo anterior sucede porque al hacer pasar una corriente eléctrica por una bobina, se genera un campo magnético que ejerce una fuerza sobre el imán. El radio está construido para modular una corriente eléctrica con el sonido (en la figura 15 se muestra una representación gráfica). Esta corriente eléctrica al pasar por la bobina genera una fuerza que al interactuar con el imán hace vibrar al vaso.

La explicación completa de este fenómeno (el que una corriente eléctrica genere un campo magnético) es algo que se estudia en cursos de física avanzados.

El principio de funcionamiento de un motor es el mismo que el de la bobina que orienta al imán, es decir, se genera una fuerza magnética que hace que un embobinado montado en el eje, girando libremente, se oriente con el imán y la corriente que alimenta al embobinado se conecta y desconecta selectivamente, dependiendo de la orientación del eje (experimento C, p. 353).

Aunque el motor propuesto para construir es muy sencillo, el principio de su funcionamiento es el mismo que el de los motores eléctricos más grandes que se usan en aparatos electrodomésticos, elevadores, bombas hidráulicas, etcétera.

3.3.5 Generación de una corriente eléctrica por un campo magnético

Al igual que una corriente eléctrica genera un campo magnético, un campo magnético puede generar una corriente eléctrica. Este fenómeno se puede observar usando la bobina que se construyó para el experimento de Oersted.

Si se conecta la bobina a un voltímetro y se acerca el imán a la bobina sin moverlo, la aguja del voltímetro no se moverá porque que no se generó ninguna corriente. Ahora acérquese el imán a la bobina y retírese rápidamente; en este caso se observará un movimiento en la aguja indicando que se generó una corriente. De esto se concluye que es el cambio en el campo magnético lo que genera la corriente eléctrica (figura 16).

Por otro lado, para generar una corriente constante habrá que girar rápidamente el imán cerca de la bobina. Un generador funciona de forma similar a un motor eléctrico, donde se tiene una bobina que puede moverse cerca de un imán, de modo que, al hacerlo girar, genera una corriente. [Véase video en CD: "Principio del generador eléctrico".]

Entonces, la electricidad que llega a las casas se genera de forma similar al experimento anterior, pero a una escala mucho mayor, es decir, dependiendo de la tecnología de la planta de luz, se hacen girar grandes generadores que funcionan con este principio.

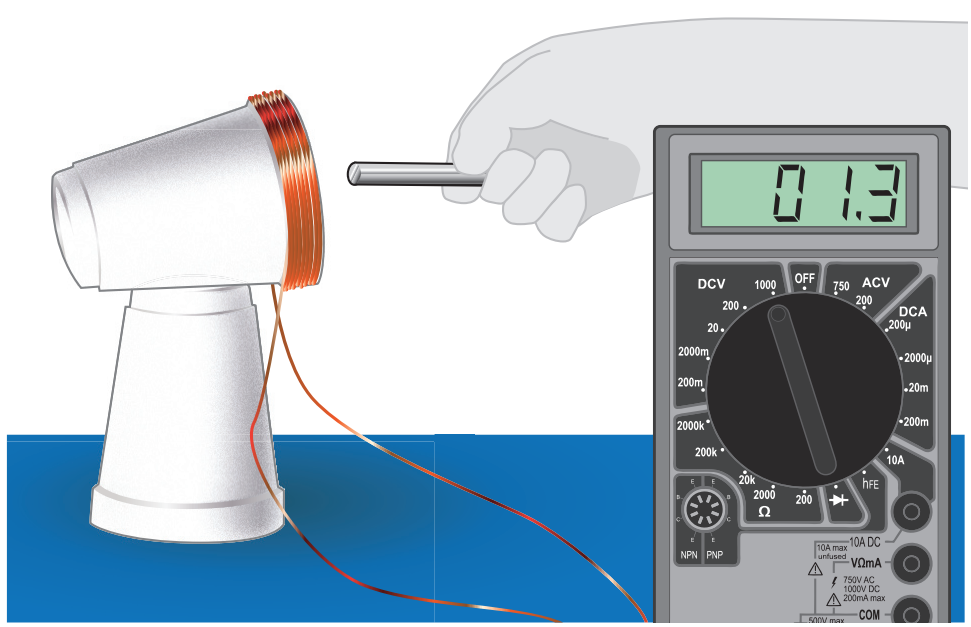
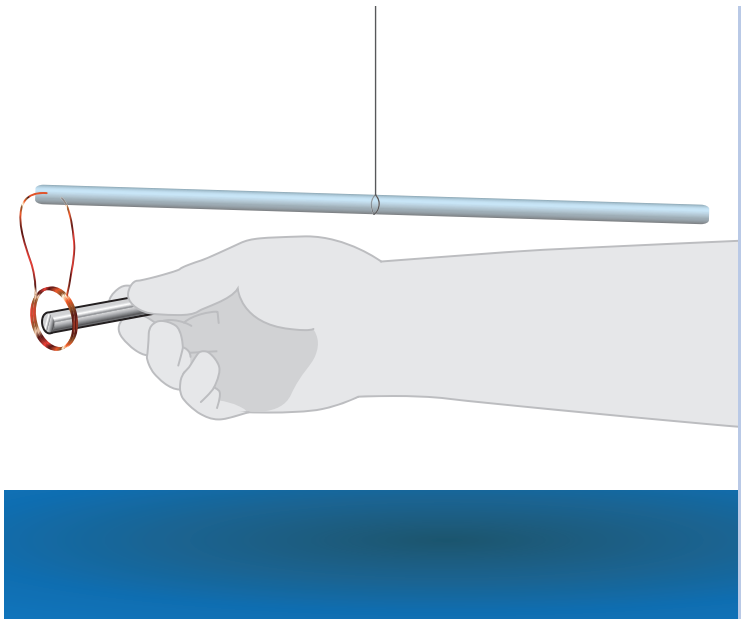


Figura 16. Medición de cambio magnético.



Experimento F

a] Con un imán de barra y una bobina de pocas vueltas, con un diámetro sólo un poco mayor que el ancho del imán, se raspa el barniz aislante de las puntas y se unen.

b] Se sujeta la bobina a uno de los extremos de un popote y se colocan algunos clips al otro extremo, como contrapeso.

c] Se cuelga el popote como se muestra en la figura y se pasa el imán a través de la bobina, sin tocarla.

d] Obsérvese lo que ocurre: ¿hacia dónde gira el popote?, ¿qué pasa si no se unen las puntas de la bobina?

[Véase video en CD: "Demostración de la ley de Lenz".]

Hasta ahora se ha aprendido que una corriente eléctrica circulando en una bobina genera un campo magnético, y además, que un campo magnético genera una corriente eléctrica en una bobina. Entonces ¿qué pasaría si se conecta la bobina consigo misma? Es decir, ¿se puede usar la corriente generada para producir un campo magnético por la misma bobina? En realidad sí es posible, pero se debe tener cuidado al interpretar esta respuesta, ya que se podría llegar a conclusiones equivocadas: si se imagina un imán cerca de una bobina conectada consigo misma, si se mueve un poco el imán se generaría un campo magnético. Y si dicho campo atrajera al imán, lo movería, induciendo a su vez una corriente eléctrica mayor a la anterior y, por lo tanto, se crearía un campo magnético mayor que a su vez atraería al imán con más fuerza, y así sucesivamente. Entonces se habría creado energía y el imán saldría disparado. Pero se sabe que no existen procesos en la naturaleza en donde se genere energía espontáneamente; entonces, ¿dónde está el error en el experimento imaginario?

El error está en suponer que el campo que se genera atrae al imán. Lo que realmente sucede es lo contrario, es decir, lo repele. De hecho, en el experimento donde se acerca y retira rápidamente el imán de la bobina existe una fuerza que se opone al movimiento que se realiza. Aunque esta fuerza es pequeña y casi imperceptible, es posible diseñar un experimento que permita apreciarla (experimento F).

Obsérvese en el experimento que el popote gira como si el imán "empujara" a la bobina, es decir, se genera una fuerza que se opone siempre al movimiento del imán, independientemente de su orientación. El físico ruso Heinrich Friedrich Lenz (1804-1865) explicó este fenómeno y se puede resumir de la siguiente manera: la corriente inducida en la bobina es tal, que el campo magnético producido por ella se opone al campo magnético del imán que la genera

Esto no es otra cosa que una consecuencia más de la conservación de la energía.

3.3.6 Ley de Faraday

Como se vio anteriormente, al hacer pasar una corriente variable (alterna) por una bobina, se genera un campo magnético; ahora, si se acerca otra bobina, en ésta se inducirá entonces una corriente.

Visto de otra forma, si se hace pasar una corriente variable por un circuito cerrado inducirá un voltaje en otro circuito cerrado. Este fenómeno, que consiste en la producción de un voltaje en otro circuito precisamente por medio de una corriente variable en algún otro lugar, fue descubierto por Michael Faraday (1791-1867) y por Joseph Henry de forma independiente en 1830. Una de las aplicaciones más comunes de la ley de Faraday son precisamente los transformadores, que constan de dos bobinas, una a la que se le llama primaria y otra secundaria. Por detalles técnicos se construyen de tal forma que la secundaria está enrollada sobre la primaria, de manera que por ambas pase la misma cantidad de líneas de campo magnético. El cociente de los voltajes en una y otra bobina está relacionado de la siguiente manera:

$$\frac{V_s}{V_p} = \frac{N_s}{N_p},$$

y por conservación de la energía, la potencia en la bobina primaria y en la secundaria son iguales:

$$P_p = P_s.$$

Pero $P = VI$, entonces,

$$V_p I_p = V_s I_s,$$

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} = \frac{I_p}{I_s},$$

donde V_s y V_p son los voltajes secundario y primario, respectivamente, N_s y N_p son el número de vueltas de las bobinas secundaria y primaria, I_s e I_p son lo propio para las corrientes como se muestra en el siguiente esquema.

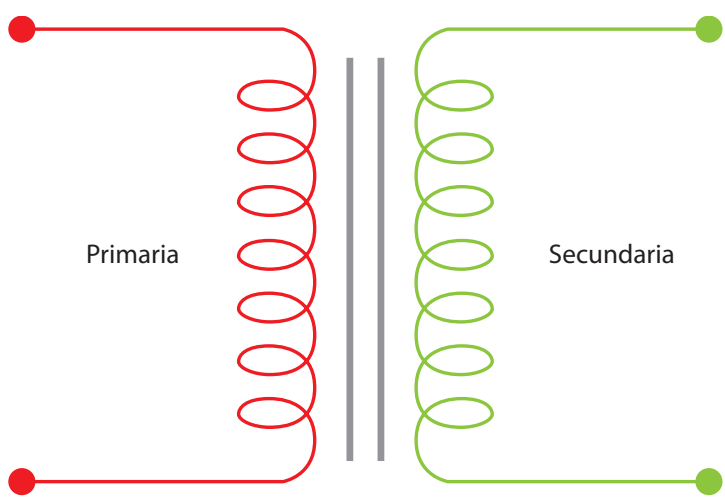


Figura 17. Inducción de corriente con dos bobinas.

Ya sabemos que para que una bobina genere una corriente, ésta debe estar en presencia de un campo magnético que cambie con el tiempo, para lo cual, debe pasar por la primaria una corriente que cambie con el tiempo, esto es, debe ser alimentada con lo que llamamos corriente alterna, la cual, a diferencia de la corriente directa que se obtiene de una pila, pasa de ser positiva a negativa 60 veces por segundo.

De aquí que si se aplica un voltaje de corriente alterna a la bobina primaria, el voltaje inducido en la bobina secundaria puede ser mayor o menor, de acuerdo con el número de vueltas de cada bobina. Lo anterior se puede aplicar también a las corrientes de los transformadores de las calles que alimentan de electricidad a las casas, los cuales se rigen por esta sencilla ecuación.

Se sabe que se pierde una gran cantidad de energía al transportar la electricidad desde la planta que la genera hasta las casas, por lo que es necesario generar una mayor cantidad de energía en las plantas; sin embargo, si se aumentara la corriente, los cables se calentarían demasiado, por lo que, para transmitir energía eléctrica a grandes distancias, la corriente debe ser pequeña y el voltaje alto. Entonces, en la planta el voltaje es de 120 000 v, aproximadamente; al llegar a la ciudad se reduce el voltaje a 2 200 v con un primer transformador y, finalmente, con el transformador que está fuera de las casas, se reduce a 120 v.

Lo que se observa en la ecuación es que los voltajes y las corrientes están en relación inversa. Entonces, al llegar al transformador, se reduce el voltaje y se aumenta la corriente.

Por lo anterior, los científicos trabajan en encontrar materiales para hacer más eficiente el transporte de electricidad, como los superconductores o materiales que no pre-

Figura 18. Tren Maglev | © Latin Stock México.

sentan resistencia eléctrica, con lo cual se perdería muy poca energía en transportarla. Hasta la fecha, los materiales que se han encontrado para este propósito sólo funcionan a temperaturas muy bajas (aproximadamente 200°C bajo cero, que es la temperatura del nitrógeno en su estado líquido), por lo que los cables se tendrían que enfriar a esta temperatura, lo cual sería muy costoso.

Otra propiedad de estos materiales es que expulsan el campo magnético de su interior, de modo que levitan en presencia de un imán. Un ejemplo de la aplicación de este fenómeno es el tren de levitación magnética Maglev, el cual tiene bobinas de material

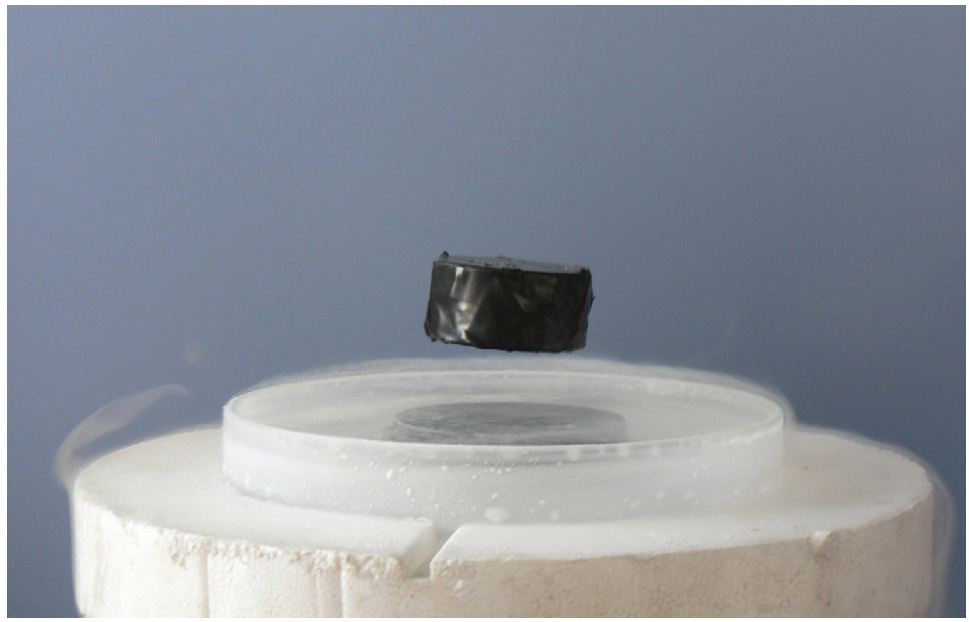


Figura 19. Levitación magnética, superconductores, efecto Meissner. [Véase video en CD: "Superconductores".]

superconductor en la parte inferior y, conforme se desplaza, éstas inducen corriente en las bobinas fijas en el carril y actúan como imanes "espejo" que hacen levitar el tren a unos centímetros del riel, lo cual hace que no exista fricción entre el tren y el riel, haciéndolo muy rápido y eficiente (véase figura 19).

3.4 ONDAS ELECTROMAGNÉTICAS. ESPECTRO ELECTROMAGNÉTICO

Las microondas son ondas electromagnéticas que están formadas por campos eléctricos y magnéticos oscilantes. Este tipo de ondas son producidas por cargas eléctricas que oscilan a muy alta frecuencia. Esa rapidísima vibración es la que da lugar a la aparición de los campos eléctricos y magnéticos.

Un campo eléctrico genera un campo magnético y un cambio en este último genera, a su vez, otro campo eléctrico. De ese modo, del punto en que se generan se expanden radialmente viajando a la velocidad de la luz (300 000 km/s en el vacío).

En el caso de la luz, en la fuente que la produce existe una infinidad de átomos excitados vibrando a altas y muy variadas frecuencias, con sus electrones saltando entre niveles atómicos y también con infinidad de electrones libres brincando de un átomo a otro; todo ello da lugar a la emisión de ondas electromagnéticas en una amplia gama de frecuencias; incluidas, desde luego, las frecuencias de la luz visible, del infrarrojo (ondas electromagnéticas asociadas al calor) y del ultravioleta.

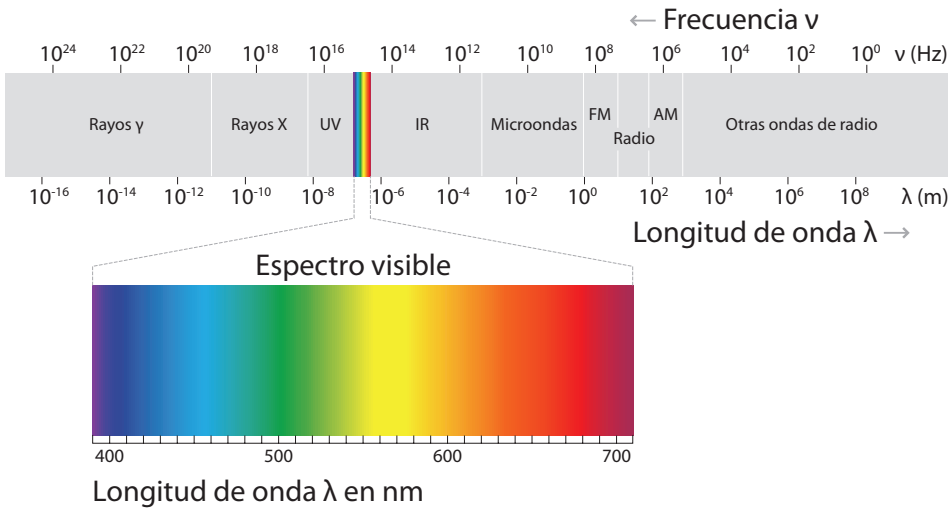


Figura 20. Espectro de luz.

El ser humano ha diseñado y elaborado dispositivos electrónicos para la generación y control de ondas de radio, microondas, infrarrojo, luz visible, ultravioleta y rayos X y se ha diseñado un esquema para representar a las ondas electromagnéticas ordenándolas por su frecuencia y de mayor a menor longitud de onda, como se aprecia en la imagen de arriba.

Como se puede ver, en este espectro la parte del visible es la parte que se puede percibir con los ojos y es relativamente angosta; los humanos pueden ver luz cuya frecuencia está entre 4.0×10^{14} Hz y 7.5×10^{14} Hz, que comprende longitudes de onda de entre 7.5×10^{-7} m y 4.0×10^{-7} m (esto es, entre 400 y 750 nanómetros). Por otro lado, la longitud de onda de las microondas es de centímetros; su frecuencia está entre 10^9 y 10^{10} Hz.

Desde fines del siglo XIX se hizo la primera transmisión telegráfica sin hilos entre Gran Bretaña y Francia, mediante ondas electromagnéticas, obra de Guglielmo Marconi (1874-1937), lo que representó el antecedente inmediato de las radiotransmisiones; también, por aquellos años Wilhelm Conrad Roentgen (1845-1923) descubrió los rayos X y de inmediato se vislumbró su aplicación en los diagnósticos médicos, práctica que es común en la actualidad.

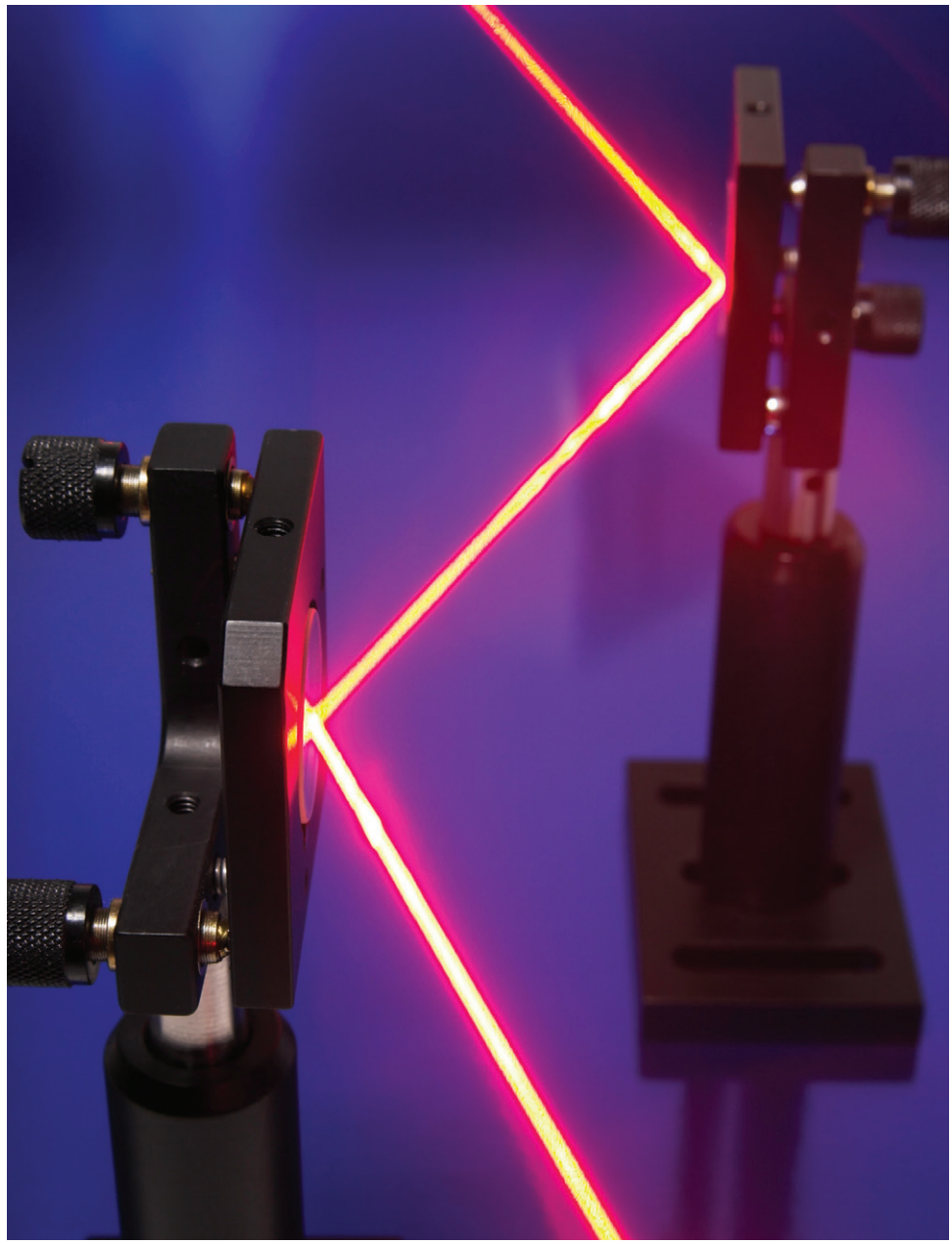
Luego, a principios del siglo XX, se iniciaron los estudios y experimentos que desembocaron en la transmisión de imágenes, esto es, la televisión, tan presente en nuestra vida cotidiana. En la actualidad contamos con dispositivos como el teléfono celular, el horno de microondas o el láser, que tiene usos muy variados como, por ejemplo, en la industria de la construcción, en intervenciones oftalmológicas o en los discos compactos de reproducción de imágenes y sonidos.

El horno de microondas posee un dispositivo denominado magnetrón, que genera ondas electromagnéticas cuya frecuencia es de alrededor de 10^{10} Hz, las cuales viajan a través de los alimentos e interactúan con las moléculas de agua. El campo eléctrico origina que las moléculas de agua oscilen de un lado a otro ganando energía. Las microondas tienen justamente la frecuencia correcta para que las moléculas oscilen de manera que resuenen absorbiendo una gran cantidad de energía. A través de los choques con las moléculas vecinas, su energía hace que se eleve la temperatura de los alimentos que las contienen.

ÓPTICA

TEMA

4



INTRODUCCIÓN

Este tema se ocupará del estudio del comportamiento de la luz y de otros fenómenos ondulatorios, pues, como se verá más adelante, la luz también se comporta como una onda.

La física, junto con las otras ciencias naturales y las matemáticas, proveen conocimientos, habilidades y actitudes que contribuyen a explicar cómo funciona la naturaleza. Esos conocimientos, además, han dado y seguirán dando las bases para el diseño de nuevos recursos tecnológicos o para explicar el funcionamiento de otros ya en uso, recursos que facilitan y hacen más eficientes muchas actividades del ser humano.

Desde tiempos remotos los seres humanos se han maravillado ante las manifestaciones de la naturaleza. Tal es el caso de la aparición del arcoiris, el intenso azul del cielo de un día despejado o la presencia de las auroras boreales o australes en lugares situados en los círculos polares del planeta. Algunos se han interesado por estudiar cómo es que ocurren o cómo se producen.

Desde la Edad Media se diseñaron y empezaron a usar anteojos para mejorar la capacidad visual de las personas. Los primeros microscopios y telescopios se ubican a inicios del siglo XVII, con lo que se dio un gran salto en el estudio de los microorganismos y se sentaron las bases para los avances en biología y medicina. Con el uso del telescopio se aceptó por fin la teoría heliocéntrica propuesta por Copérnico (1473-1543), que explica muchos de los aspectos de lo que hoy conocemos como Sistema Solar y se revolucionó la concepción del Universo. Posteriormente, a fines del siglo XVII, Isaac Newton dio explicación a la formación del arcoiris y, a fines del siglo XIX, barón de Rayleigh explicó el azul del cielo.

Estos avances en el conocimiento del carácter electromagnético de la luz, junto con los avances en electrónica, han sido el punto de partida para el diseño de dispositivos tecnológicos que en la actualidad usamos de manera cotidiana, como los aparatos de rayos X, para el diagnóstico médico en clínicas y hospitales, la telefonía inalámbrica, la emisión-recepción de señales de radio y televisión, los hornos de microondas, el rayo láser, los discos compactos de reproducción de sonidos e imágenes, etcétera.

4.1 ÓPTICA GEOMÉTRICA

La óptica geométrica es el estudio de las trayectorias que siguen los rayos de luz al incidir en espejos o a través de cuerpos transparentes, en particular algunas lentes delgadas. Si se considera a la luz como un fenómeno ondulatorio, los rayos de luz representan la dirección de propagación de las ondas luminosas, lo que permitiría explicar la formación y características de las imágenes de los objetos iluminados colocados frente a espejos y lentes.

Reflexión de la luz en espejos planos

Con ayuda de un espejo plano, una hoja de papel blanco, dos lápices nuevos e iguales, una regla, una barra de plastilina, escuadra y transportador, se puede hacer un sencillo experimento:

Primero se pega sobre una mesa la hoja de papel blanco, usando la cinta adhesiva. Con la regla se traza una línea a lo largo de la parte central de la hoja. Ahora, hay que fijar el espejo plano con la plastilina sobre la línea trazada. Debe quedar en posición vertical. Enseguida hay que dibujar una equis en la hoja de papel de manera que se refleje en el espejo, marcando la equis con la letra O (objeto).

Colocar la punta de uno de los lápices sobre ese punto y pedir a otra persona que marque con la letra A su punto de observación frente al espejo. Luego se colocará el otro lápiz detrás del espejo, justo en el punto que, desde su posición (el punto A), parezca que es la prolongación de la imagen que se ve en el espejo (de esta manera se estará ubicando la posición de la imagen). Localizado este punto detrás del espejo, se señalará con una X e identificará con la letra O' (O prima). En la hoja de papel quedarán señalados tres puntos.

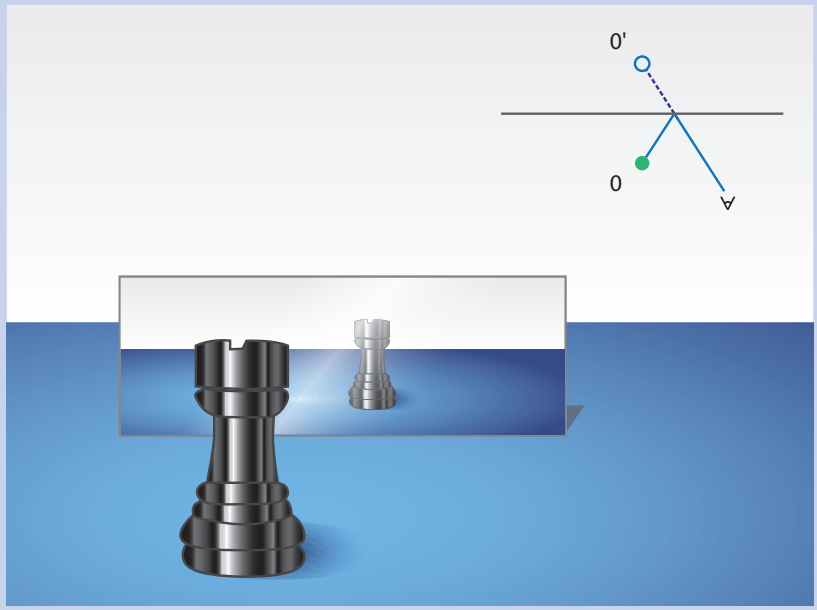
Ahora se escoge otro punto de observación B y se repite la actividad. Se hará evidente que O' coincide también con la prolongación de B detrás del espejo, lo que se comprueba al retirar el espejo, quedando ahora cuatro puntos.

Con ayuda de la regla, se traza una línea de O' a A y, con otro color, de O' a B. Los segmentos de línea delante del espejo se trazan con una línea continua y los de atrás del espejo con una línea punteada.

Finalmente, con la ayuda de una escuadra, se dibujan líneas perpendiculares a la línea del espejo en cada punto de intersección hacia la parte de enfrente del espejo; con ello se estará dibujando la línea normal para cada uno de los puntos de observación. Luego se trazan las rectas que van del punto O a cada uno de los puntos de intersección de las primeras rectas con el espejo.

Analizando el dibujo obtenido se puede comprobar que los ángulos de incidencia y de reflexión son iguales, lo que permitiría concluir que para los rayos de luz que inciden y se reflejan en espejos planos *el ángulo de incidencia es igual al ángulo de reflexión, medidos éstos respecto a la normal* (línea perpendicular al espejo), que es, en esencia, la ley de la reflexión de la luz en espejos planos. Se verifica también que la distancia de la imagen al espejo es igual a la distancia del espejo al objeto.

[Véase animación en CD: "Reflexión en espejos planos".]



Rayos en espejos planos.

Para verificar experimentalmente lo que representa la segunda ley de reflexión de la luz en espejos planos que establece que el ángulo de incidencia, el ángulo de reflexión y la normal, están en un mismo plano, se puede realizar lo siguiente:

Colocar horizontalmente un lápiz frente a un espejo plano, iluminado por alguna fuente luminosa (véase la figura 1, p. 373). Se observará que, desde todos los puntos de la superficie del lápiz, salen rayos de luz reflejada y una parte de ellos va a incidir al espejo. Así que al espejo llegan infinidad de rayos de luz, los cuales cumplen las leyes de la reflexión.

En la figura que ilustra el experimento anterior se muestran sólo dos rayos originados en la punta del lápiz que se reflejan en el espejo hacia el ojo del observador. Como se puede ver, los rayos divergen a partir de la punta del lápiz y se prolongan a partir del es-

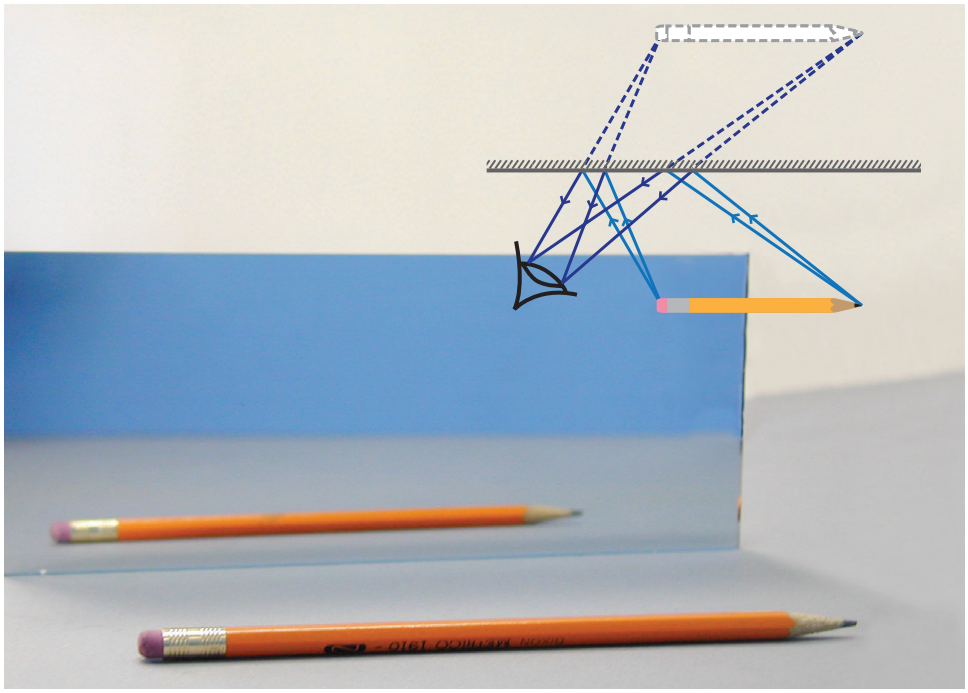


Figura 1. Comprobación de la segunda ley de reflexión.

pejo al reflejarse. Así, estos rayos divergentes parecen provenir de un punto ubicado atrás del espejo.

Un análisis semejante puede hacerse para los rayos de luz procedentes de cualquier otro punto de la superficie del lápiz, que inciden y se reflejan en el espejo para luego llegar a los ojos del observador. La imagen del lápiz que el observador ve en el espejo se denomina *imagen virtual*, porque en realidad la luz no pasa por la posición de la imagen, pero se comporta como si de ella proviniera. Esto es, como si en realidad hubiese un lápiz en esa posición. Como se puede ver, coincidiendo con el experimento, la imagen está atrás del espejo, a la misma distancia que el objeto real frente al espejo. También se aprecia que la imagen y el objeto son de igual tamaño.

A partir de lo anterior, se puede explicar por qué cuando una persona se mira al espejo su imagen es de su mismo tamaño y a una distancia igual a la que se encuentra frente a él; esto es, si la persona está a un metro del espejo, su imagen la ubica a un metro "dentro" del mismo. Se observa también que si la persona levanta la mano derecha, su imagen

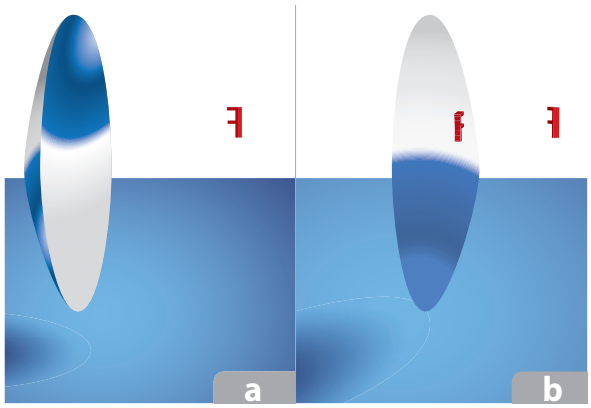


Figura 2. Espejos esféricos: (a) cóncavo y (b) convexo.

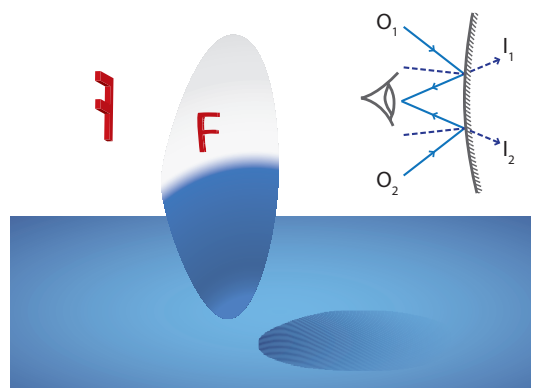


Figura 3. Imagen formada en un espejo convexo.

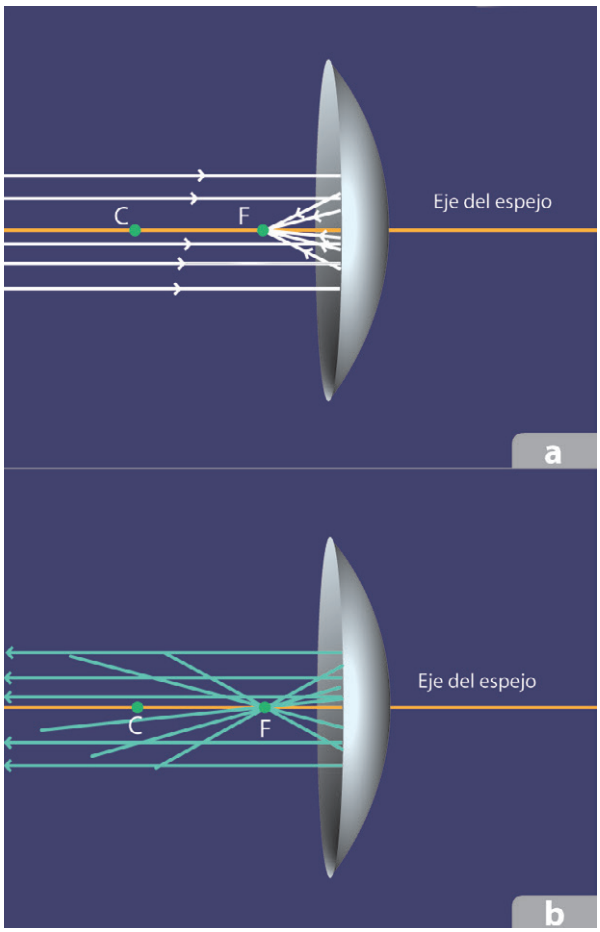


Figura 4. Rayos en espejo cóncavo. (a) Cuando los rayos incidentes son paralelos al eje; (b) cuando los rayos incidentes pasan por el foco.

levanta su mano izquierda y, si levanta la izquierda, su imagen levanta la derecha.

De este modo, parece que los espejos planos invierten las imágenes izquierda-derecha; pero no es así. Esto se confirma con la imagen obtenida para el lápiz, en la figura 1 (p. 85). Otra manera de concluir que no hay tal inversión es poniendo frente al espejo una letra "E", recortada en cartoncillo. No se observará en la imagen ninguna inversión.

4.1.1 Imágenes en espejos curvos

Pero lo anterior no ocurre en algunos espejos curvos; en éstos, algunas veces las imágenes sí se ven invertidas (arriba-abajo). Cabe destacar que en nuestra vida diaria es frecuente el uso de espejos curvos; tal es el caso de algunos espejos retrovisores de automóviles o los espejos que se usan en tiendas departamentales, como un recurso auxiliar en la vigilancia, o en algunas esquinas de las calles como un recurso para prevenir accidentes automovilísticos. También en los parques de diversiones podemos encontrar espejos con curvaturas particulares que dan imágenes curiosas.

Usualmente se estudia la formación de imágenes en espejos curvos empezando por los que son un casquete de esfera hueca; esto es, espejos esféricos. Así, se obtienen dos tipos de espejos: los cóncavos y los convexos.

Una esfera de Navidad es un buen ejemplo de un espejo esférico convexo; refleja imágenes derechas y de menor tamaño que el objeto real, sin invertirlas. Y aunque las leyes de la reflexión de la luz siguen siendo válidas para espejos curvos, en este caso, las líneas "normales" cambian de dirección en cada punto del espejo y no son paralelas entre ellas, como se muestra en la figura 3 (p. 85). La imagen es virtual.

Un espejo esférico cóncavo da imágenes invertidas o derechas, dependiendo de la distancia a la que se encuentre el objeto. Estos espejos se comportan de forma similar a los espejos parabólicos, cuya geometría permite simplificar el análisis. Los rayos de luz que llegan al espejo y que son paralelos al eje del mismo, se reflejan pasando por el punto focal "F" y los rayos que llegan al espejo, pasando por el punto focal, se reflejan en dirección paralela al eje del espejo.

Aquí, por supuesto, también se cumplen las leyes de la reflexión de la luz que se analizaron anteriormente. El punto focal se encuentra en el punto medio entre el espejo y su centro de curvatura (en el caso de los espejos esféricos), esto es, $F = r/2$, como se muestra en la figura 4.

Si el objeto puesto frente al espejo se encuentra a una distancia mayor que la distancia focal, entonces, la imagen se formará donde convergen los rayos reflejados y se verá invertida.

En el diagrama expuesto en la figura 5 (p. 375) sólo se han dibujado tres de los rayos que proceden del objeto, los cuales bastan para describir a la imagen. A esta imagen que se forma donde convergen los rayos de luz reflejados se le denomina "imagen real", y

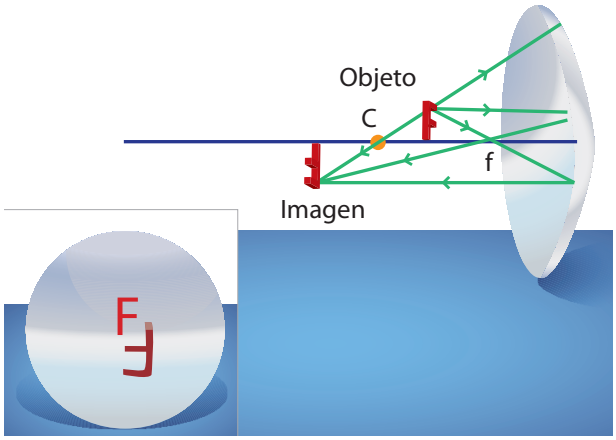


Figura 5. Imagen formada en espejo cóncavo. [Véase animación en CD: "Espejo cóncavo".]

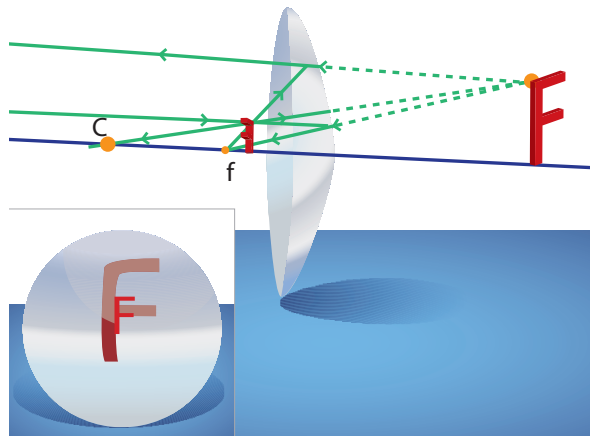


Figura 6. Formación de imagen virtual en espejo cóncavo.

puede verse en una pantalla o en una película fotográfica colocada en ese lugar. Así, se puede explicar por qué algunos espejos presentan imágenes invertidas.

Ahora, si el objeto se encuentra frente a un espejo cóncavo, a una distancia menor que la distancia focal, entonces la imagen será derecha y virtual; esto último dado que la luz no pasa realmente por la imagen, ya que ésta se forma detrás del espejo.

Como se puede ver en el diagrama de la figura 6, la imagen, además de ser virtual y derecha, es más grande que el objeto real; de ahí que este tipo de espejos se utilicen para obtener imágenes amplificadas.

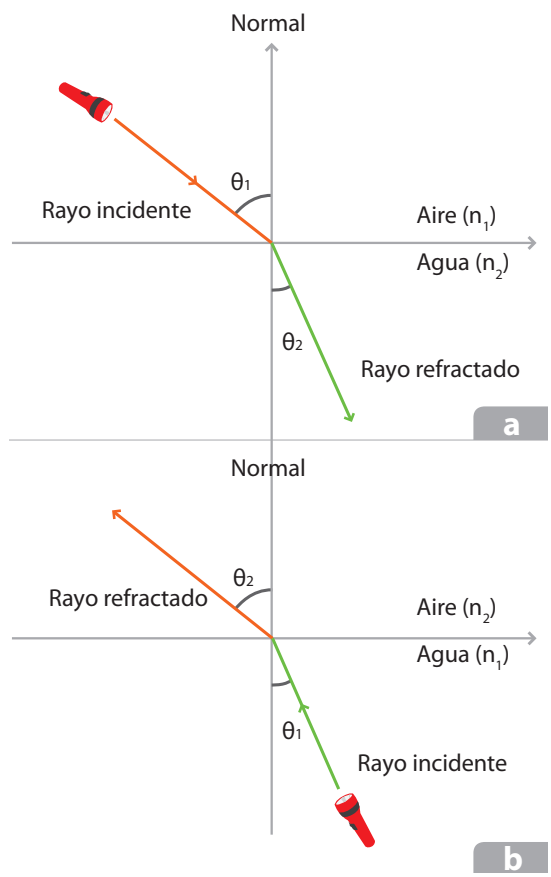
4.1.2 Refracción de la luz. Ley de Snell

Cuando la luz pasa de un medio transparente a otro, parte de la luz incidente se refleja en la frontera de los dos medios, y el resto, pasa al otro medio, sufriendo una desviación. Ésa es la explicación del porqué a las personas metidas en una alberca, con el agua hasta la cintura, se les ven las piernas más cortas; o porqué los peces en un estanque se ven a una profundidad menor de aquella a la que realmente están.

La velocidad de una onda cambia al pasar de un medio a otro y la luz no es la excepción su velocidad también cambia al pasar de un medio a otro. Si un rayo de luz incide formando un ángulo diferente de 90° con la frontera entre los dos medios, se desviará al entrar al segundo medio, precisamente por el cambio que sufre en su velocidad, y a esta desviación se le conoce con el nombre de *refracción*.

Así, cuando un rayo de luz pasa del aire al agua, el rayo refractado se acerca a la normal (línea imaginaria perpendicular a la frontera entre los dos medios y que pasa por el punto en que el rayo incide), lo que ocurrirá siempre que un rayo de luz pase de un medio de bajo índice de refracción a otro más alto. El índice de refracción de cierto medio "n" se define como el cociente de la velocidad de la luz en el vacío "*c*" entre la velocidad de la luz en el medio en cuestión "v", así:

Figura 7. Ley de Snell.



$$n = \frac{c}{v}.$$

El índice de refracción no tiene unidades por ser un cociente entre velocidades. El valor de *n* siempre es mayor que uno, pues en cualquier medio transparente v < *c*. El índice de refracción de un medio varía de manera inversa con la velocidad de la luz al propagarse en él. Ahora, si el rayo de luz pasa de agua a aire, entonces el rayo refractado se aleja de la normal, y esto ocurre siempre que un rayo de luz pasa de un medio con índice de refracción mayor a otro con índice de refracción menor.

Sustancia	*Índice de refracción (n)*
Aire (0° C ,y 1 atm)	1.000293
Agua	1.333
Alcohol etílico	1.361
Hielo	1.309
Vidrio sin plomo	1.52
Diamante	2.419

Cuando una persona se encuentra en una alberca, con el agua a la cintura, los rayos de luz procedentes de sus pies, pasan del agua al aire, alejándose de la normal (figura 10), por lo que un observador fuera de la alberca verá que los pies del bañista están más cerca de la superficie del agua de lo que realmente se encuentran, por eso las piernas se ven más cortas. Con un razonamiento semejante podemos concluir que, efectivamente, vemos a los peces en un estanque como si estuvieran a una profundidad menor que aquella en que realmente están.

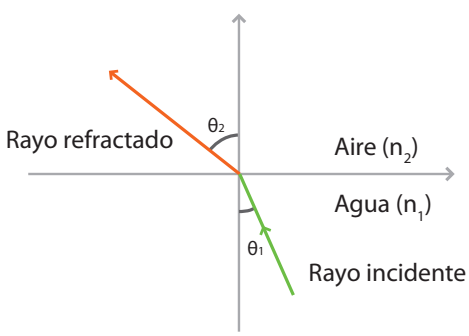


Figura 8. Deformación de una imagen por la refracción en el agua.

En el siglo XVII, Willebrord Snell (1591-1626) encontró experimentalmente la relación entre el ángulo de incidencia "θ_1"y el ángulo de refracción "θ_2":

$$n_1 \operatorname{sen}\theta_1 = n_2 \operatorname{sen}\theta_2,$$

donde n_1 y n_2 son los índices de refracción de los materiales transparentes en donde se desplazan los rayos incidente y refractado, respectivamente. A esta relación se le conoce con el nombre de Ley de Snell y es la ley fundamental de la refracción de la luz.

Ahora se sabe que n_1 y n_2 están relacionados con la velocidad con que se desplaza la luz en cada uno de los medios transparentes, lo que en el tiempo de Snell se ignoraba, pues en los años en que él vivió aún no se medía la velocidad de la luz.

Dado que la Ley de Snell establece que $n_1 \operatorname{sen}\theta_1 = n_2 \operatorname{sen}\theta_2$, entonces, si $n_2 > n_1$, necesariamente debe cumplirse que $\theta_2 < \theta_1$, para que se cumpla la igualdad; esto es, en relación con el fenómeno óptico, si la luz entra en un medio donde n es mayor (y por ello, su velocidad menor), entonces, el rayo se desvía acercándose

a la normal. Pero si $n_2 < n_1$, entonces $\theta_2 > \theta_1$, para que la igualdad se cumpla, con relación al fenómeno, quiere decir que el rayo de luz se desvía alejándose de la normal.

Otro ejemplo de lo anterior es cuando una vara está sumergida en agua, la cual veríamos como doblada debido a que la luz procedente de la parte de la vara sumergida se refracta al salir al aire.

Cuando un rayo de luz llega de manera perpendicular a la frontera entre los dos materiales transparentes, ¿cuál es el ángulo de incidencia?, ¿cuál el de refracción?

Ya que estos ángulos se miden respecto a la normal, y el rayo incide perpendicularmente, $\theta_1 = 0^\circ$; por lo tanto, $\theta_2 = 0^\circ$; y el rayo de luz pasa sin desviarse.

4.1.3 Formación de imágenes con una lente delgada biconvexa

Una lupa es una lente biconvexa usada en la vida cotidiana. Un microscopio también es conocido en amplios sectores por estudiantes, técnicos y profesionistas, quienes lo usan con regularidad. Es interesante conocer y comprender cómo estos instrumentos ópticos dan imágenes amplificadas de los objetos colocados frente a ellos. Para entender cómo se forman estas imágenes, es necesario estudiar las trayectorias que siguen los rayos de luz que inciden en una lente biconvexa.

Si se considera una lente delgada (se denomina así a una lente que es muy delgada en comparación con su diámetro) biconvexa, como la mostrada en la figura 14; el eje de la lente es una recta que pasa por su centro y es perpendicular a sus superficies. Todos los rayos de luz que inciden en ella y que son paralelos a su eje se desvían de acuerdo con la Ley de Snell, convergiendo en el punto focal, que se representa con la letra "F". La distancia medida del punto focal al centro de la lente se conoce como *distancia focal* y se representa con la letra "*f*".

Si a la izquierda de esa primera lente se coloca una segunda lente biconvexa, de manera que su foco coincida con el foco de la primera, entonces, de la segunda lente y de acuerdo con la Ley de Snell, los rayos de luz saldrán paralelos al eje de la lente.

Para una lente biconvexa, un rayo de luz que incida en ella en dirección paralela a su eje se refractará pasando por su foco; un rayo de luz que incida en ella habiendo pasado por su foco se refractará saliendo de ella en dirección paralela a su eje, y un rayo de luz que incida en el centro de la lente pasará por ella sin desviarse.

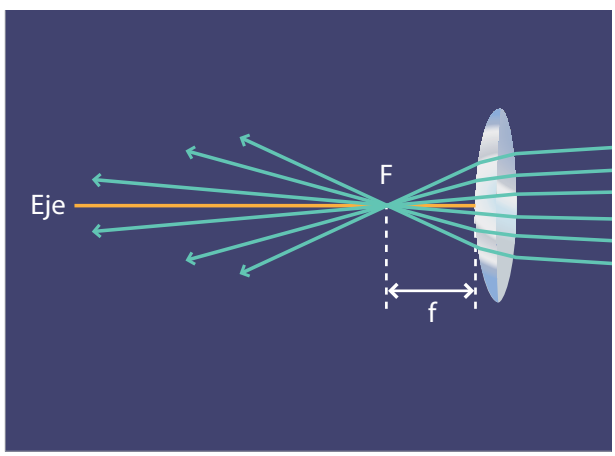


Figura 9. Rayos de luz paralelos, incidiendo en lente convergente. Luego de atravesar la lente, converjen en el punto focal.

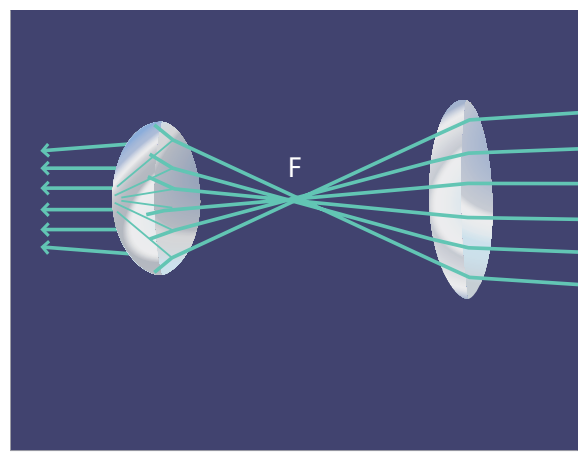


Figura 10. Arreglo de dos lentes convergentes.

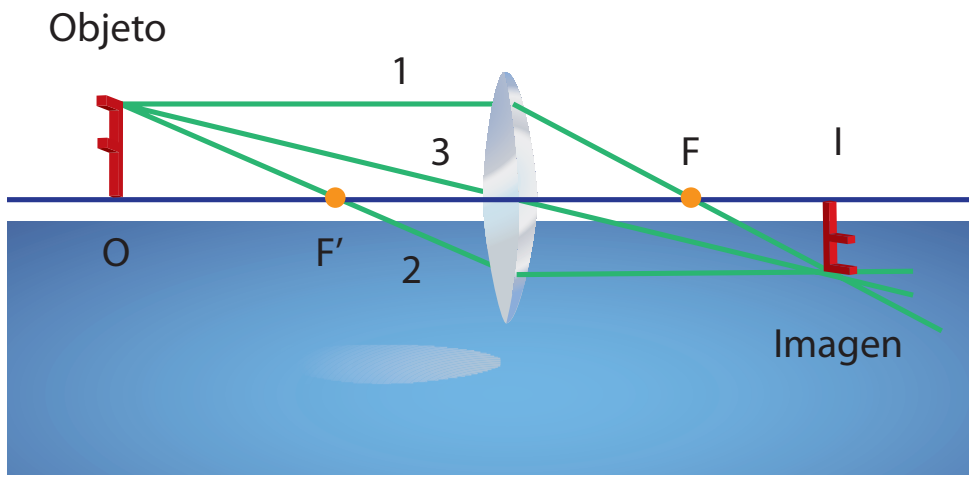


Figura 11. Ubicación de la imagen formada por una lente convergente. [Véase animación en CD: "Lente convergente".]

Con base en el comportamiento de estos tres tipos de rayos de luz que pasan por una lente biconvexa, es posible predecir algunas características importantes de las imágenes que pueden formarse con una lente de este tipo. En la figura 11 se muestran una letra F como objeto y una lente biconvexa que forma una imagen a la derecha. Seguiremos la trayectoria de sólo tres rayos de luz que salen de la parte superior de la F.

El rayo 1 se dibuja paralelo al eje; por lo tanto, al ser refractado por la lente, pasará por el punto focal F, situado detrás de ella.

El rayo 2 se dibuja pasando por el punto F, del mismo lado de la lente en que está el objeto; por lo tanto, emerge de ella paralelo al eje (si este tipo de lentes tienen dos puntos focales simétricos).

El rayo 3 incide en el centro de la lente y pasa sin desviarse.

Con dos de estos rayos basta para localizar el punto correspondiente a la imagen del punto del objeto (la esquina superior de la F, en este caso). Ahí donde se intersectan los rayos de luz, se localiza el punto imagen. Los puntos imagen correspondientes a los demás puntos del objeto pueden hallarse de la misma manera, hasta determinar la imagen completa del objeto. En este caso la imagen está invertida, respecto a la posición del objeto. Una imagen como ésta se llama imagen real, dado que los rayos de luz pasan realmente por la imagen y ésta puede ser captada en una pantalla, en una película fotográfica, aunque también se puede ver directamente.

El tamaño de la imagen obtenida dependerá de la distancia a la que se encuentra el objeto de la lente. Con estos diagramas de rayos se puede explicar por qué se ven algunas veces imágenes invertidas con la lupa.

El rayo 1 se dibuja, análogamente, sin mayor problema. Luego, el rayo 3 se dibuja fácilmente pasando por el centro de la lente sin desviarse. Como puede verse en la figura 12, estos dos rayos, al salir de la lente divergen, esto es, no se juntan en ningún punto, de modo que no pueden formar una imagen real. Pero si estos dos rayos divergentes se prolongan hacia atrás, se encuentra el punto en que estas líneas coinciden y en ese punto se forma la imagen de la esquina de la letra F. De hecho, se determina así la posición en que se forma una imagen y, como puede verse, se trata de una imagen virtual, pues los rayos de luz no pasan por la imagen; además, es derecha y de mayor tamaño que el objeto.

Para encontrar la distancia focal de alguna lente se necesita una fuente luminosa cuyos rayos sean paralelos. Puede ser el Sol, ya que la distancia a la que se encuentra de la Tierra es tan grande que se puede considerar que los rayos que pasarían por la lente son paralelos. También se puede quemar un papel usando una lupa y la luz del Sol; lo que se

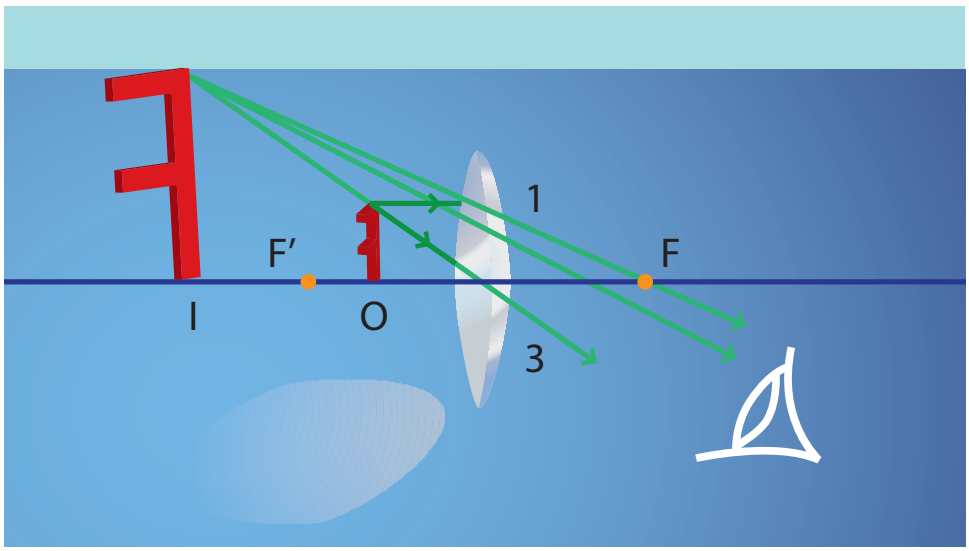


Figura 12. Un objeto situado entre el punto focal y la lente convergente produce una imagen virtual.

hace es justamente encontrar el foco. Por lo tanto, la distancia que hay entre la lupa y el papel cuando éste se empieza a quemar es justamente la distancia focal.

Una vez conocida la distancia focal, se pueden realizar diferentes experimentos colocando una lámpara a diferentes distancias de la lupa. Una hoja de papel al otro lado de la lente captará en ella la imagen de la lámpara. Se encontrará que si se pone la lámpara en el foco no se observa ninguna imagen, pero al colocarla entre el foco y dos veces la distancia focal, la imagen se proyectará amplificada e invertida, y la distancia a la que aparece la imagen nítida estará más allá del doble de la distancia focal.

¿Qué sucede si se pone la lámpara exactamente al doble de la distancia focal? La imagen aparecerá del mismo tamaño de la lámpara y la distancia a la que aparece la imagen invertida será el doble de la distancia focal. Si se pone la lámpara de dos veces más allá de la distancia focal se obtendrá una imagen reducida e invertida entre la distancia focal y dos veces la distancia focal.

Ecuación de la lente. La ley de los puntos conjugados | Se puede obtener una ecuación que relaciona la distancia al objeto con la distancia a la imagen y con la longitud

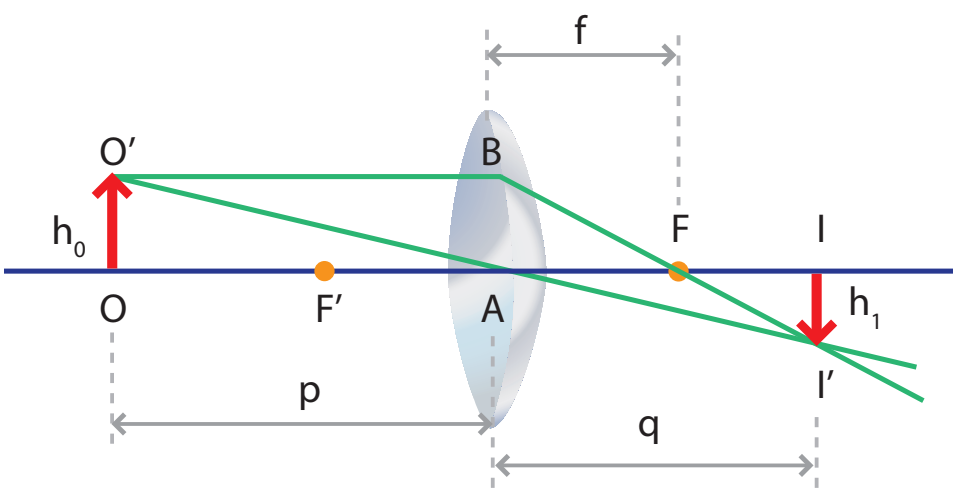


Figura 13. Esquema para obtener la ecuación de la lente.

focal de una lente. Esta ecuación permite determinar la posición de la imagen de manera más rápida y precisa que mediante el trazado de rayos.

Considérese que "p" representa la distancia al objeto, esto es, la distancia del objeto al centro de la lente; "q" la distancia a la imagen o la distancia del centro de la lente a la imagen; y que "h_o" y "h_i" sean las alturas del objeto y de la imagen, respectivamente.

Considérense los dos rayos mostrados en la figura anterior, correspondientes a una lente delgada biconvexa.

Los triángulos FI'I y FBA son semejantes, pues el ángulo AFB es igual al ángulo IFI', además de tener un ángulo recto cada uno de ellos. Así pues, por ser triángulos semejantes se cumple que:

$$\frac{h_i}{h_o} = \frac{q-f}{f}, \quad \text{pues} \quad AB = h_o\,.$$

Por otro lado, los triángulos IAI' y OAO', también son semejantes y por ello:

$$\frac{h_i}{h_o} = \frac{q}{p},$$

de manera que, igualando estas dos expresiones, se tiene que:

$$\frac{q}{p} = \frac{q-f}{f} = \frac{q}{f} - 1.$$

Dividiendo entre q:

$$\frac{1}{p} = \frac{1}{f} - \frac{1}{q},$$

entonces:

$$\frac{1}{p} + \frac{1}{q} = \frac{1}{f}.$$

Esta ecuación es conocida como *ley de los puntos conjugados*, o ecuación de la lente, que relaciona la distancia a la imagen con la distancia al objeto y con la longitud focal. Entonces, si se conocen dos de estos datos, se puede calcular el tercero; ésta es la ecuación más útil de la óptica geométrica.

Microscopio compuesto | En realidad, una lupa es un microscopio simple, y se pueden ver con ella imágenes amplificadas de los objetos; puede tomarse como objeto una imagen real formada por otra lente.

Cuando se proyecta la imagen de un objeto en una pantalla, se coloca a ésta en la posición donde inciden los puntos luminosos correspondientes a puntos del objeto; pero aun cuando no se colocara la pantalla, los rayos de luz continuarían su camino, exactamente igual que si estuvieran saliendo de un objeto real. Entonces se puede tomar una imagen real como objeto de otra lente y conseguir así una mayor amplificación.

Considérese el siguiente diagrama, en cuyo arreglo se podría proyectar una imagen derecha de un objeto y de su mismo tamaño.

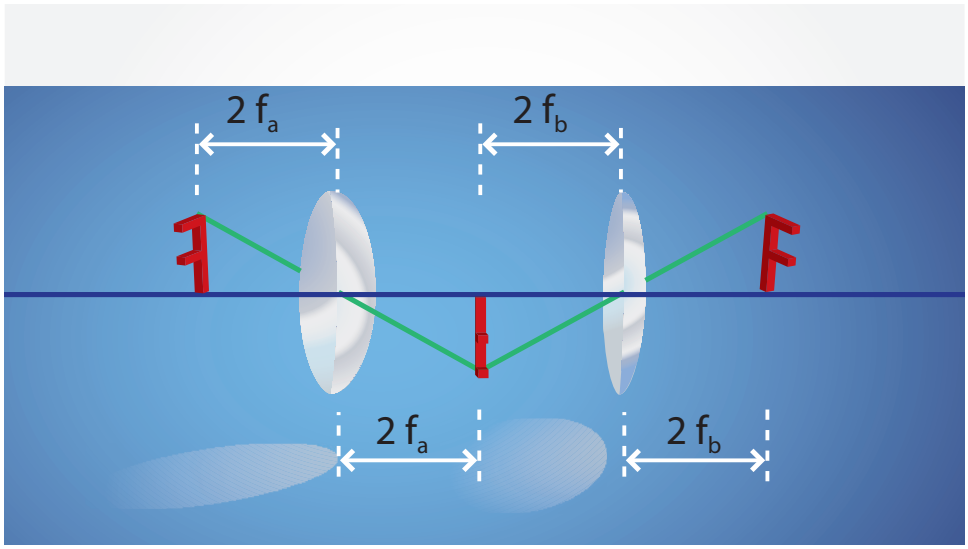


Figura 14. Proyección de una imagen derecha del mismo tamaño del objeto. [Véase simulación en CD: "Formación de imágenes con dos lentes convergentes".]

En este caso se tienen dos lentes biconvexas: la primera de longitud focal f_a, y la segunda, de longitud focal f_b. Se coloca el objeto a una distancia $2f_a$ de la primera lente, de manera que se forma una imagen real, invertida y del mismo tamaño, a una distancia $2f_a$. Entonces, se coloca la segunda lente exactamente a una distancia $2f_b$ de la primera imagen, de manera que se forma una segunda imagen (derecha respecto al objeto real) a una distancia $2f_b$ de la segunda lente. Con esto se obtiene una imagen real derecha y de igual tamaño que el objeto real usando dos lentes biconvexas. Pero si se acerca la segunda lente a la primera imagen, de modo que la distancia sea menor que su distancia focal, y se observa a través de la segunda lente (llamada ocular) se verá una imagen virtual y de mayor tamaño que el objeto.

Se puede obtener una mayor amplificación si las distancias del objeto real a la primera lente son pequeñas, pero mayores que la distancia focal, de manera que con la primera lente se tendrá un primer aumento; esto es, que se genere una imagen real de mayor tamaño y luego se utilice el ocular para conseguir un aumento aún mayor. Éstas son las bases del microscopio compuesto.

Galileo (1564-1642) fue uno de los primeros científicos en utilizar un microscopio, incluso en Italia se considera que él fue el inventor de este instrumento, hacia el año 1610. Sin embargo, las primeras publicaciones importantes en el campo de la microscopía óptica aparecen a mediados del siglo XVII por el médico italiano Marcelo Malpighi (1628-1694), quien probó la teoría del médico inglés William Harvey (1578-1657) sobre la circulación sanguínea, al observar al microscopio los capilares sanguíneos, y también por el científico inglés Robert Hooke (1635-1703), quien, al observar con un microscopio un delgado corte de corcho, descubrió y nombró *células* a los segmentos que formaban este material; de hecho esa fue la primera observación de células muertas.

A mediados de ese siglo, el biólogo y comerciante holandés Anton van Leewenhoek (1632-1723), utilizando microscopios de manufactura propia, describió por primera vez protozoarios, bacterias, espermatozoides y glóbulos rojos, de manera que este invento fue un pilar importante para el desarrollo de la biología y la medicina.

Al paso de los años los microscopios ópticos se perfeccionaron y surgió una gran variedad de éstos con mayor resolución, hasta llegar a los modernos microscopios electrónicos.

Otro instrumento óptico cuyo funcionamiento se basa también en el empleo de dos lentes delgadas es el telescopio; su uso ha contribuido al crecimiento y desarrollo de la sociedad. Aunque su invención alrededor de 1608 se atribuye al comerciante holandés Hans Lippershey, quien era dueño de una fábrica de anteojos en Middleburg, Holanda, se reconoce que Galileo fue uno de los primeros en estudiar el cielo con este tipo de instrumentos. Esto ocurrió en 1609, en Italia, con un telescopio fabricado por él mismo, para el cual utilizó una lente cóncava como ocular. Galileo observó la superficie de la Luna, encontrando que tenía cráteres y montañas, esto es, no era aquella esfera perfecta que muchos habían imaginado; observó también algunos satélites girando en torno a Júpiter y que Venus tenía fases como las de la Luna. Todas estas observaciones de Galileo le dieron elementos para apoyar la tesis del sistema heliocéntrico de Nicolás Copérnico, lo que causó que fuera enjuiciado por la Santa Inquisición. Pero, gracias a sus contribuciones, los sectores ilustrados de la sociedad, y luego la población en general, comenzaron a comprender y aceptar la posición de nuestro planeta en el Sistema Solar y en el Universo.

Cabe destacar que de los tiempos de Galileo a la fecha actual el telescopio se ha perfeccionado para captar imágenes de cuerpos ubicados a muchos años luz de nosotros. Hoy se dispone de un potente telescopio puesto en órbita terrestre: el telescopio espacial Hubble.

Los anteojos | Una lente cóncava, es decir, aquella que es más delgada en el centro que en los bordes, es una lente divergente puesto que hace que los rayos de luz, que inciden paralelos a su eje, diverjan (figura 15).

En este tipo de lentes el punto focal F es el punto del cual parecen emerger los rayos refractados que se originan en los rayos incidentes paralelos y, desde luego, la distancia entre el punto F y la lente, es la distancia focal f.

Los anteojos son instrumentos ópticos de uso común que basan su funcionamiento en el empleo de lentes cóncavas y convexas. Desde tiempos tan remotos como los de la Roma imperial, Nerón, debido a su miopía, utilizaba una esmeralda moldeada en forma cóncava, de media luna, para mirar las peleas de gladiadores.

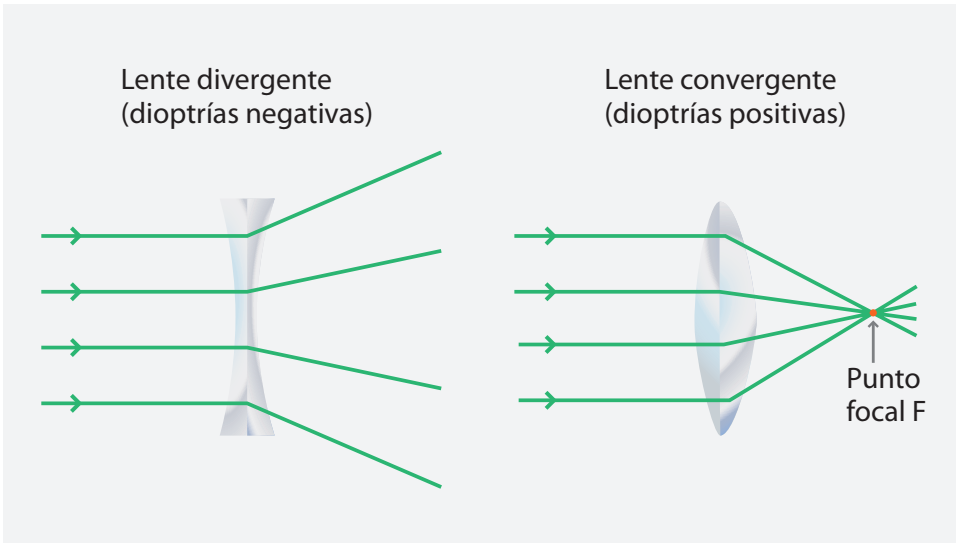


Figura 15. Rayos de luz paralelos al eje de la lente cóncava incidiendo en ella y saliendo divergentes.

Alrededor del año 1000 de nuestra era, el físico y matemático árabe Alhazen escribió un amplio tratado sobre óptica en el cual describió cómo se forma la imagen en la retina humana debido al cristalino del ojo, que es una lente convexa natural. Después, hacia 1266, el fraile franciscano inglés Roger Bacon talló los primeros lentes con forma de lenteja que ahora se conocen (de ahí su nombre de lentes) y describió las propiedades de una lente para amplificar la letra escrita. Así, a finales del siglo XIII en el norte de Italia, zona en que estaba muy desarrollada la tecnología del pulido de cristales, aparecen las primeras lentes convergentes.

Las primeras lentes se fabricaron para corregir la presbicia; eran convexas y se idearon inicialmente para un solo ojo; luego se unieron dos de esos lentes en una sola armadura y se les agregó un mango, para mayor comodidad. La armadura se colocaba sobre la nariz.

Las lentes cóncavas para miopes aparecen, aproximadamente, cien años más tarde. Se dice que ya las usaba el poeta y humanista italiano Petrarca (1304-1374). Con la invención de la imprenta en el siglo XV se incrementó la demanda de anteojos. Los primeros anteojos bifocales se inventaron en la segunda mitad del siglo XVIII. Se dice que Benjamín Franklin fue de los primeros en usarlos. Así pues, aquellos primeros estudios sobre lo que hoy denominamos óptica geométrica, realizados por el físico árabe Alhazen, dieron las bases para esos dispositivos tan útiles a la sociedad en su necesidad de ver mejor.

Los optometristas y oftalmólogos utilizan el recíproco de la distancia focal (1/f) para referirse a la graduación de las lentes —ya sean convergentes o divergentes— para anteojos. La unidad empleada es la dioptría (D); $1D = 1/m = 1\ m^{-1}$. Así, por ejemplo, una lente que tenga 50 cm de distancia focal tiene dos dioptrías ($1/0.5\ m = 2\ m^{-1} = 2$).

La miopía es la incapacidad de ver con claridad objetos lejanos; ocurre en un ojo demasiado alargado y, por lo tanto, la imagen de los objetos se forma delante de la retina. Una lente divergente, que hace que los rayos de luz que llegan a ella paralelos se separen, permite que los rayos se enfoquen y formen la imagen de los objetos lejanos en la retina, corrigiendo así el defecto (figura 16).

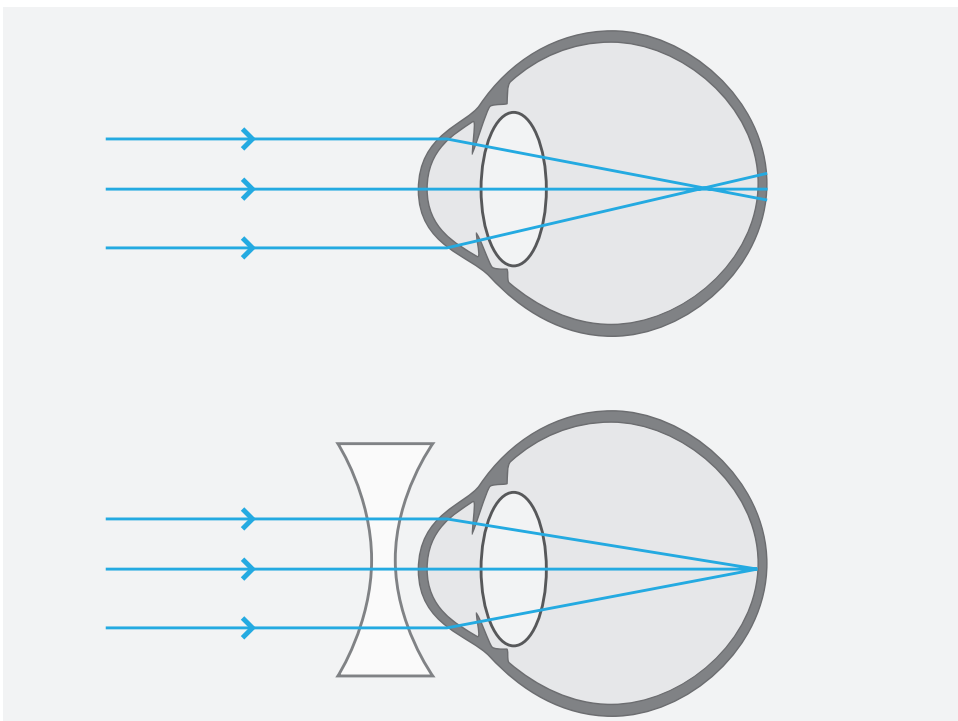


Figura 16. Esquemas de dos casos de ojo miope. En el superior, los rayos de luz llegan paralelos y convergentes dentro del ojo antes de llegar a la retina. En el inferior se antepone una lente cóncava al ojo, de manera que los rayos que inciden en ella de forma paralela, se divergen un poco para que, luego de pasar por el cristalino, converjan en la retina.

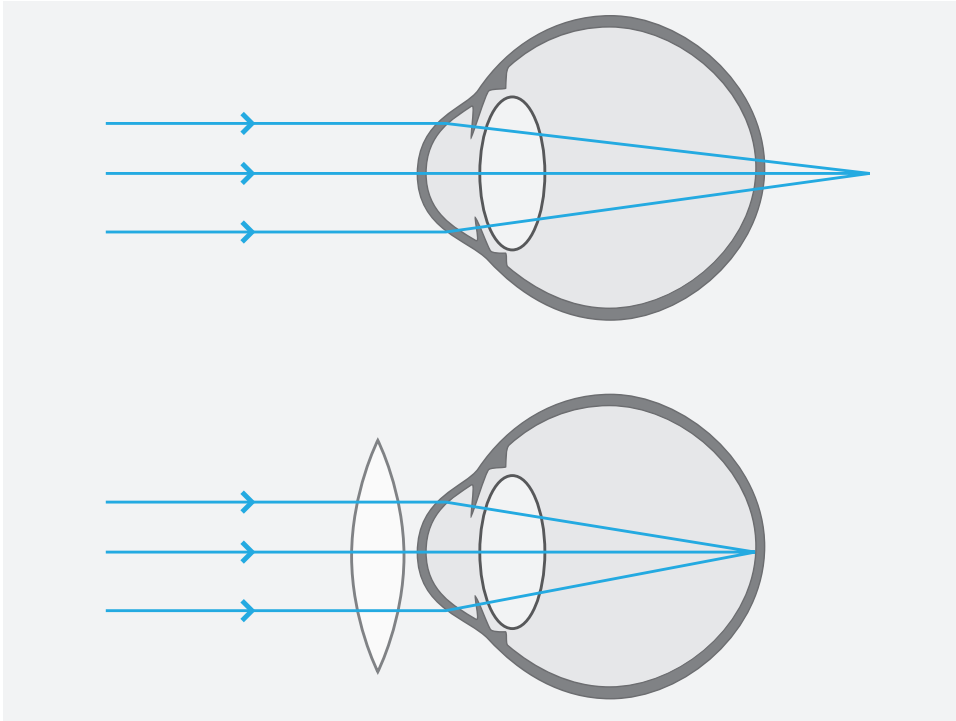


Figura 17. Esquema de ojo hipermétrope. El esquema superior muestra la convergencia de los rayos de luz atrás de la retina; el esquema inferior muestra la corrección usando una lente convergente para que los rayos de luz converjan en la retina.

La hipermetropía consiste en la incapacidad de ver con claridad los objetos cercanos, lo que hace difícil la lectura. Este defecto ocurre en un ojo demasiado corto y los rayos de luz convergen atrás de la retina. Esta falla en la visión se puede corregir con una lente convergente (figura 17).

Con base en los primeros estudios de Alhazen se han construido también otros dispositivos que hoy son de uso cotidiano, como los binoculares, las cámaras fotográficas y los diferentes tipos de proyectores de imágenes. Por otro lado, con las aportaciones de Alhanzen, probablemente enriquecidas con los estudios del físico neerlandés Willebrord Snell (1591-1626) sobre la refracción de la luz de Isaac Newton, con sus propios descubrimientos sobre la dispersión de la luz blanca en los colores que la integran, dio explicación al maravilloso fenómeno natural del arcoiris.

4.2 NATURALEZA DE LA LUZ

Hasta principios del siglo XIX, la mayoría de los hombres de ciencia consideraban que la luz era una corriente de partículas (corpúsculos) emitidas por una fuente luminosa, que estimulaba el sentido de la vista. Esta teoría había sido impulsada por Isaac Newton, por lo menos cien años atrás, y con ella explicaba la reflexión y la refracción de la luz.

El holandés Christiaan Huygens (1629-1695) planteó que la luz tenía comportamiento ondulatorio y explicó la reflexión y la refracción de la luz; sin embargo, esta teoría no tuvo aceptación inmediata, por diversas razones: una, el gran prestigio del que en esos años disfrutaba Newton, pero además, se daba por hecho que todas las ondas habían de viajar en un medio material y la luz viajaba desde el Sol a la Tierra, a través del espacio vacío; por otro lado, se decía que si la luz fuera una propagación ondulatoria, esas ondas habrían de desviarse al llegar a un obstáculo y no se observaba tal desviación; entonces la luz siempre viajaba en línea recta.

Hoy se sabe que esa desviación de la luz al llegar a la orilla de un objeto sí ocurre; se manifiesta mediante la difracción. De cualquier modo, la mayoría de los científicos de esa época rechazaron la teoría ondulatoria de la luz y se adhirieron a la teoría corpuscular impulsada por Newton, por más de un siglo.

En 1801, Thomas Young (1773-1829) demostró que la luz manifestaba un comportamiento ondulatorio llamado interferencia. Pocos años después, el francés Augustin-Jean Fresnel (1778-1829) realizó experimentos de interferencia y difracción de la luz. Luego, en 1850, Jean Foucault (1819-1868) demostró que la velocidad de la luz en líquidos es menor que en el aire, con lo cual daba un golpe mortal a la teoría corpuscular, que pregonaba que la velocidad de la luz sería mayor en líquidos y cristales que en el aire.

Fue así que en el curso del siglo XIX se generalizó la aceptación de la teoría ondulatoria de la luz. Vino luego la aportación teórica de James Clerk Maxwell (1831-1879) quien, en 1873, demostró que la luz era una propagación ondulatoria; ondas electromagnéticas de alta frecuencia que, de acuerdo con su teoría, debían viajar a una velocidad aproximada de 3×10^8 m/s, que coincidía con las mediciones experimentales de la velocidad de la luz que ya se conocían para entonces. La primera medición exitosa, en ese sentido, fue la del danés Ole Roemer (1644-1710).

Todo apuntaba a considerar a la luz como un fenómeno ondulatorio. Pero, a fines del siglo XIX, el físico experimental Heinrich Hertz (1857-1894), el mismo que en 1887 confirmó experimentalmente la teoría de Maxwell, descubrió el efecto fotoeléctrico; la expulsión de electrones de un metal expuesto a la luz. Una explicación de este fenómeno fue propuesta por Einstein, en 1905, utilizando el concepto de fotones: "paquetes" discretos de energía. De acuerdo con esta propuesta, el efecto fotoeléctrico es la transferencia de energía de un fotón a un electrón del metal, y la energía de los electrones expulsados depende de la frecuencia de la luz incidente.

Así pues, la luz debe considerarse como un fenómeno de naturaleza dual. Por un lado, la teoría electromagnética ondulatoria explica la propagación de la luz, los fenómenos de interferencia y difracción; por otro lado, un modelo corpuscular ofrece una mejor explicación del efecto fotoeléctrico.

En 1871 el físico inglés John William Strutt, barón de Rayleigh, publicó un artículo dando explicación al azul del cielo, con base en la característica ondulatoria de la luz y, particularmente, hablando de la dispersión que ésta sufre al interactuar con las moléculas del aire, dependiendo de la longitud de onda de la luz incidente. Encontró que la dispersión variaba de manera inversa con la longitud de onda de la luz, elevada a la cuarta potencia; es decir, a menor longitud de onda mayor dispersión; y que de los colores que integran a la luz blanca, el azul es el de menor longitud de onda, de ahí que sea el color que sufre más dispersión y, por ello, el cielo se ve azul.

FÍSICA DE FLUIDOS

TEMA

5

INTRODUCCIÓN

En la naturaleza, la materia se manifiesta en cuatro distintos estados: sólido, líquido, gaseoso y plasma. Los procesos y fenómenos relacionados con los fluidos (líquidos y gases) han sido objeto de estudio de destacados físicos, quienes centraron su atención en el agua y el aire, por ser básicos para la vida. Gracias a estos estudios se puede comprender por qué flotan los barcos a pesar de estar construidos con materiales más densos que el agua, o por qué las personas pesan menos al estar sumergidas parcial o totalmente en

ella. También se puede entender por qué el agua de los ríos se mueve con más rapidez en las zonas donde el cauce es más angosto, y más lentamente en las zonas en las que el cauce es más ancho, entre otros ejemplos.

A partir de esos estudios se han diseñado y construido equipos y vehículos de uso frecuente, como los submarinos, los batiscafos o los equipos de buceo autónomo, que permiten permanecer bajo el agua durante tiempos prolongados; o para viajar por el aire, como los aviones y helicópteros.

5.1 NOCIONES DE HIDROSTÁTICA

La hidrostática es la rama de la física que tiene como objeto de estudio los fenómenos físicos inherentes a los fluidos en reposo, concretamente, el aire y el agua.

5.1.1 Presión atmosférica

Nuestro planeta está cubierto por una capa de aire: la atmósfera. Todo lo que está sobre la superficie terrestre está inmerso en un océano de aire, lo que hace que las personas no se percaten de su peso; es más, generalmente no se cree que el aire pese.

Con un poco de ingenio se puede evidenciar que el aire pesa. Una manera de comprobarlo es con ayuda de una balanza graduada en décimas de gramo y dos jeringas de 60 ml. Deben ser jeringas sin aguja y con tapón; además, deben tener, en la parte interior del émbolo, un orificio en el cual pueda introducirse un clavo de, digamos, 4 cm de longitud, como puede verse en la figura 2.

Para realizar el experimento se nivela en ceros la balanza, se toma una de las jeringas, se le retira el tapón, se empuja el émbolo hasta el fondo y se coloca de nuevo el tapón firmemente apretado; posteriormente se jala el émbolo sin sacarlo y se coloca el clavo para impedir que el émbolo regrese hacia adentro. Así se tendrá vacío en el interior de la jeringa.

Una vez hecho lo anterior, se coloca dicha jeringa sobre la balanza. Se repite la operación con la segunda jeringa y se nivela la balanza. Enseguida, se toman una a una las jeringas y se les retira el tapón (se oirá un chasquido). Colocando luego los tapones y las jeringas sobre la balanza, se notará que hay un desequilibrio, pues las jeringas pesan más porque ahora están llenas de aire (por eso se escuchó ese chasquido al retirar el tapón de

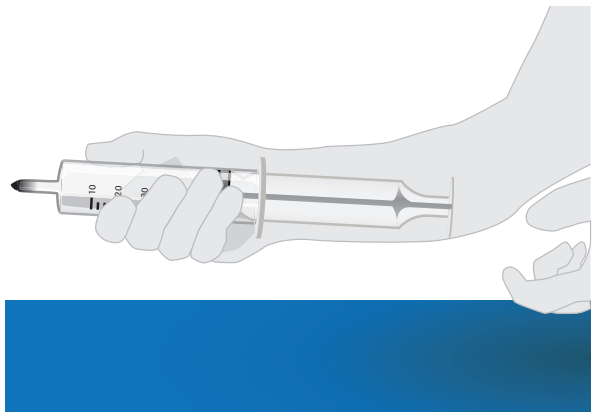


Figura 1. Preparación del material. [Véase video en CD: "Preparación del experimento para pesar el aire".]

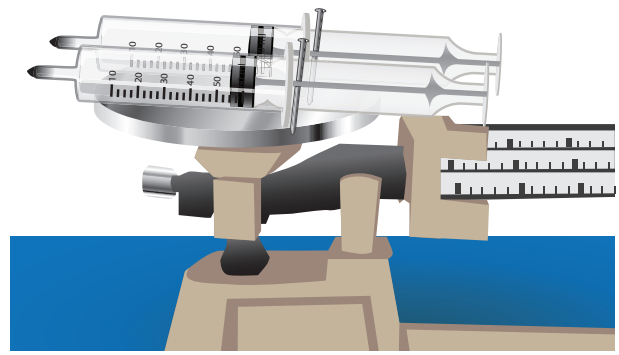


Figura 2. Peso del aire. [Véase video en CD: "Pesando el aire".]

la jeringa). Esta experiencia permite concluir que el aire pesa; de hecho, está ejerciendo su peso sobre todos los cuerpos en la superficie terrestre. A este "peso" se le conoce como *presión atmosférica.*

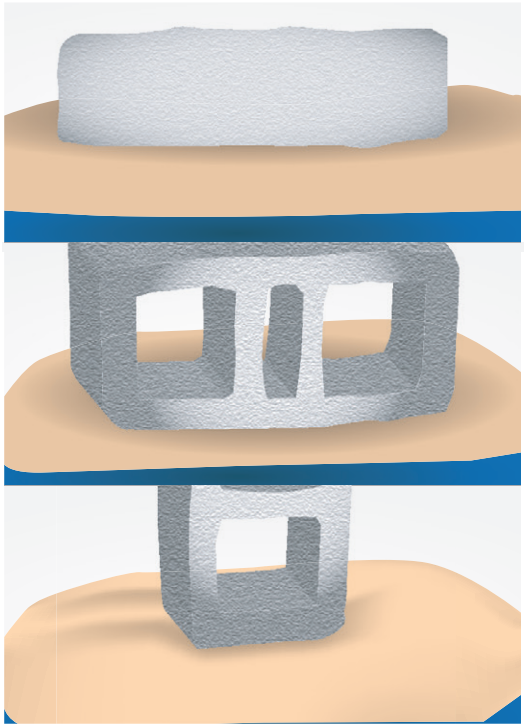


Figura 3. Presión. [Véase video en CD: "Presión".]

5.1.2 Unidad de la presión

¿Qué es la presión?, ¿en qué unidades se mide? Decir presión es hablar de apretar, oprimir o aplicar una fuerza sobre la superficie de un cuerpo. Así, por ejemplo, se aplica presión sobre un teclado de computadora, el colchón de la cama cuando una persona se acuesta o sobre el suelo al estar de pie; también se aplica presión sobre el líquido contenido en una jeringa al empujar el émbolo. En resumen, se aplica presión cuando se ejerce una fuerza sobre una superficie. De modo que la presión está relacionada con la fuerza y la superficie sobre la que se aplica.

Un experimento en el que se aprecia el concepto de presión es utilizando un bloque de hule espuma (como el que se usa para elaborar cojines) y dos o tres tabiques. Se coloca el hule espuma sobre una mesa; sobre él, en forma horizontal, un tabique; el hule espuma se hundirá debido a la fuerza aplicada sobre él; la fuerza es, en este caso, el peso del tabique.

El tabique ejerce una fuerza (su peso) sobre el hule espuma y la presión se manifiesta en el hundimiento. Si se coloca otro tabique encima y después un tercero, seguramente se observará cada vez un mayor hundimiento pues, a más tabiques colocados uno sobre el otro, el hundimiento será cada vez mayor; esto es, la presión sobre el hule espuma aumentará al aumentar la fuerza aplicada sobre él. En resumen, *la presión es directamente proporcional a la fuerza aplicada* (figura 3).

Representando a la presión con P y a la fuerza con F, tendríamos:

$$P \propto F.$$

Ahora, se retiran todos los tabiques colocados sobre el hule espuma, permitiéndole recuperar su forma original. Una vez hecho esto, se toma uno de los tabiques y se coloca horizontalmente, como antes, pero esta vez observando el hundimiento, el cual ha de ser proporcional a la presión ejercida sobre él. Enseguida, ese mismo tabique se coloca de canto, sobre el hule espuma; el hundimiento del hule espuma respecto al caso anterior será mayor, aun cuando la fuerza aplicada es la misma que en el caso anterior (el peso del tabique), pero el área sobre la que actúa esa fuerza es ahora menor.

Finalmente, se coloca el tabique en posición vertical sobre el hule espuma. Se observará que el hundimiento será aún mayor, debido a que, aunque sigue siendo la misma fuerza aplicada, ahora actúa sobre un área aún menor. De manera que cuando la fuerza se ejerce sobre un área grande, la presión es pequeña, y cuando se ejerce sobre un área pequeña, la presión es grande. Esto es, la presión varía de manera inversa con el área "A" y de manera directa con la fuerza:

$$P \propto \frac{F}{A}.$$

Lo anterior representa el modelo matemático para la presión, que indica:

$$P = \frac{F}{A},$$

en donde "F" es la fuerza, que se mide en newtons y "A" el área, cuya unidad es el metro cuadrado; las unidades de la presión son los pascales (1 pascal = 1 newton/m^2).

5.1.3 Variación de la presión atmosférica

La presión atmosférica no es la misma en todos los lugares del planeta**.** Los cuerpos están sometidos a la presión que el aire ejerce sobre ellos y esa presión es mayor cuando la profundidad dentro de ese "océano de aire" es mayor. El aire tiene su mayor profundidad a nivel del mar y disminuye en las montañas. Por ejemplo, la presión atmosférica en Xalapa es menor que en el puerto de Veracruz, pero en la ciudad de México es menor que en Xalapa, y en Toluca es aún menor, ya que al pasar de Xalapa a la ciudad de México y luego a Toluca se va a lugares cada vez más altos sobre el nivel del mar, es decir, cada vez menos profundos en el "océano de aire".

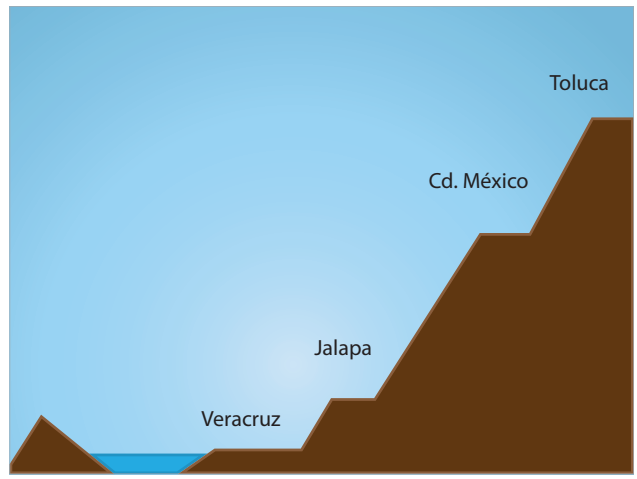


Figura 4. Alturas de diferentes ciudades.

5.1.4 Presión hidrostática. Principio de Pascal

La presión que el agua en reposo (o cualquier otro fluido) ejerce sobre los cuerpos sumergidos en dicho fluido se conoce como presión hidrostática. Depende de la profundidad "h" a la que se encuentra dentro del fluido en reposo, así como de la densidad del mismo. De esta forma, la presión hidrostática no es la misma en agua que en mercurio, aun estando a la misma profundidad, debido a la densidad del líquido.

La *densidad* es una propiedad específica de cada cuerpo o sustancia y su magnitud se obtiene del cociente de su masa entre su volumen (figura 5). Se denota con la letra ρ (ro) del alfabeto griego y en el Sistema Internacional de Unidades se mide en kg/m^3:

$$\rho = \frac{m}{V}.$$

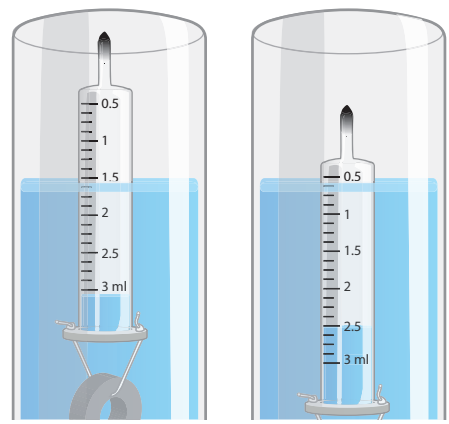


Figura 5. En un cilindro lleno de agua se introduce una jeringa con un pequeño volumen de aire. Al ir descendiendo, a mayor profundidad en el agua, el volumen de aire disminuirá debido al incremento en la presión hidrostática. De hecho, el experimento servirá para encontrar una relación entre ésta y la profundidad. [Véase video en CD: "Experimento para medir presión hidrostática".]

El modelo matemático para la presión hidrostática se puede obtener mediante un sencillo experimento (figura 5), así como a través del siguiente y sencillo desarrollo teórico: en un tanque lleno de agua, considérese una columna cilíndrica de base "A" y altura "h" (figura 6, p. 390). La presión que esta columna de agua ejerce sobre su base, de acuerdo con la definición de presión es:

$$P = \frac{\text{Peso de la columna de agua}}{\text{Área}} = \frac{m_a\, g}{A};$$

donde m_a es la masa de la columna de agua.

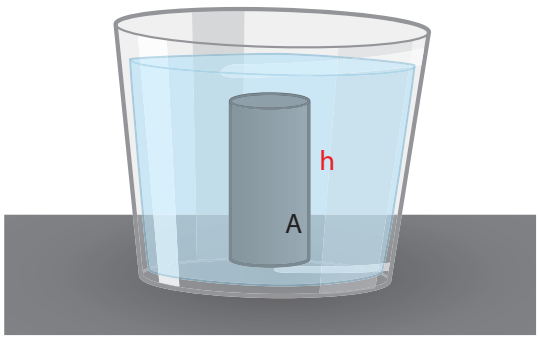


Figura 6. Presión hidrostática.

La definición de densidad es: $m_a = \rho V_a$, donde ρ es la densidad del agua y V_a el volumen de la columna de agua.

De ahí que:

$$P = \frac{\rho V_a}{A} g.$$

Sustituyendo el volumen, $V_a = A\,h$, se tiene

$$P = \frac{\rho A h g}{A},$$

por lo que finalmente

$$P = \rho g h.$$

Si ρ se mide en kg/m^3, g en m/s^2 y "h" en metros, se obtiene que las unidades de presión son N/m^2, que es igual a pascales (Pa).

Si una persona se sumerge en agua, ¿estará sujeta a la presión hidrostática o seguirá siendo afectada por la presión atmosférica? De acuerdo con el Principio de Pascal, la presión atmosférica siempre será ejercida. Para comprobarlo se pueden realizar los siguientes experimentos:

1. Se utiliza una jeringa de 60 ml con tapón, un globo que quepa holgadamente en esta jeringa y agua. Se extrae totalmente el émbolo de la jeringa, se verifica que el tapón esté firmemente apretado, se introduce luego el pequeño globo inflado en la jeringa y se coloca el émbolo sin introducirlo.

Al empujar el émbolo, el aire se comprime y se observa que el volumen del globo se reduce, pero conserva su forma (figura 7). Esto permite concluir que la presión ejercida se transmitió al aire dentro de la jeringa y al globo; además, la presión que actúa sobre el globo es perpendicular a cada punto de su superficie, pues conservó su forma y al regresar el émbolo a su posición inicial, se observa que el globo regresa a su volumen original.

Ahora se extrae totalmente el émbolo de la jeringa, se coloca firmemente el tapón, se introduce en ella el globo y se le coloca encima un pequeño lastre de plomo; ahora se llena la jeringa totalmente de agua y se coloca el émbolo, igual que antes (figura 8).

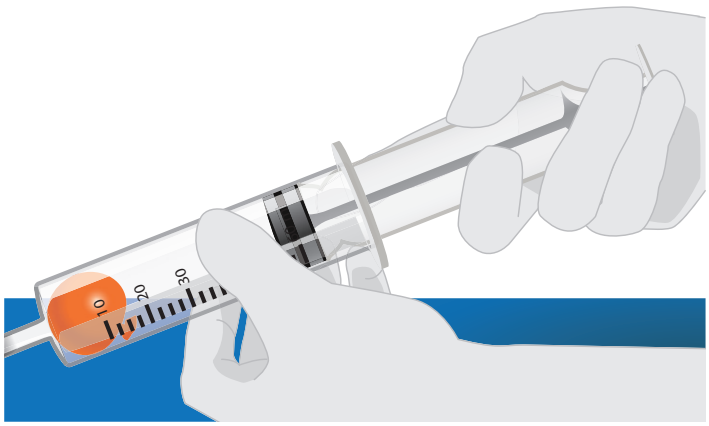


Figura 7. Experimento presión de aire 1. [Véase video en CD: "Compresión y expansión del aire".]

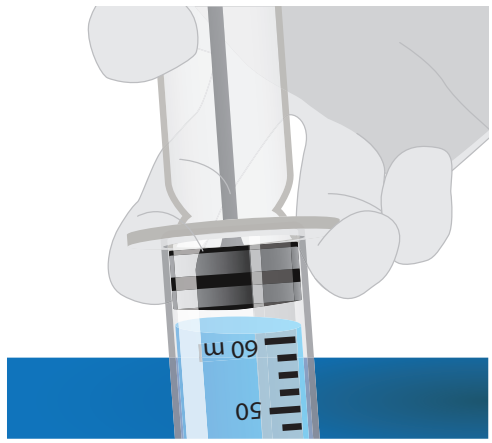


Figura 8. Experimento presión de aire 2.

Al empujar el émbolo se observará que no puede ser introducido; prácticamente no se logra que avance y el globo, otra vez, se ve reducido en su volumen, conservando su forma. La conclusión es que la presión ejercida se transmitió al agua y al globo, actuando sobre el globo de manera perpendicular en cada punto de su superficie, pues también conservó su forma. Se observa igualmente que al dejar de ejercer presión sobre el agua y por lo tanto sobre el globo, éste regresa a su volumen original.

2. Otro experimento consiste en extraer totalmente el émbolo de la jeringa del experimento anterior y retirar un poco de agua, para luego colocarle el émbolo igual que antes. Esta vez se tiene agua dentro de la jeringa, y sobre ella aire; además, dentro del agua está el globo mantenido en el fondo con el lastre (figura 9).

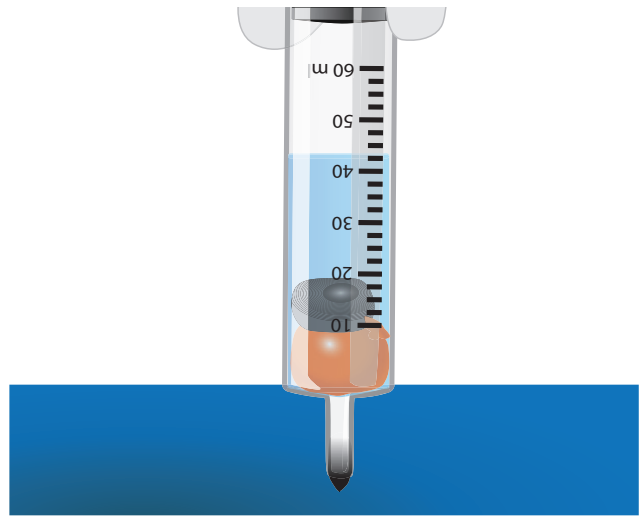


Figura 9. Experimento presión de aire 3.

Al empujar el émbolo se observa que éste se introduce parcialmente y el globo, otra vez, reduce su volumen, conservando su forma. De ahí se concluye que la presión ejercida se transmitió al aire y del aire al agua y al globo, y que la presión actuó sobre el globo de manera perpendicular en cada punto de su superficie. Asimismo, al dejar de ejercer presión sobre el émbolo, éste regresa a su posición inicial y el globo recupera su volumen original.

Estos experimentos y los resultados observados contribuyen a la comprensión del enunciado del Principio de Pascal, que dice: *La presión ejercida sobre un fluido encerrado en un recipiente se transmite íntegramente a todo ese fluido y a las paredes del recipiente que lo contiene*.

Se puede agregar que la presión actúa en todas direcciones. Este es el principio con el que funcionan las prensas hidráulicas y fue establecido por Blas Pascal (1623-1662), con cuyo apellido se ha acordado denominar a las unidades de presión en el Sistema Internacional de Unidades. Por esto —y en particular por los resultados del último experimento— queda claro ahora que si una persona se sumerge hasta cierta profundidad en el agua, le afectará la presión hidrostática y también la presión atmosférica del lugar donde se encuentre. De hecho, existe el concepto de presión total o absoluta, que se define como:

Presión absoluta = presión hidrostática + presión atmosférica.

$$P_a = \rho g h + P_o.$$

5.1.5 Medición de la presión atmosférica

La presión atmosférica en un lugar depende de su altura sobre el nivel del mar y se mide con un instrumento llamado barómetro, cuyo principio fue descubierto por el sabio italiano Evangelista Torricelli (1608-1647). Su experiencia fue la siguiente:

> Un tubo cilíndrico de vidrio cerrado en uno de sus extremos se llena totalmente de mercurio. A continuación, tapándole el extremo abierto, se le da un giro para que quede boca abajo y se coloca en un recipiente con mercurio, de manera que quede en posición verti-

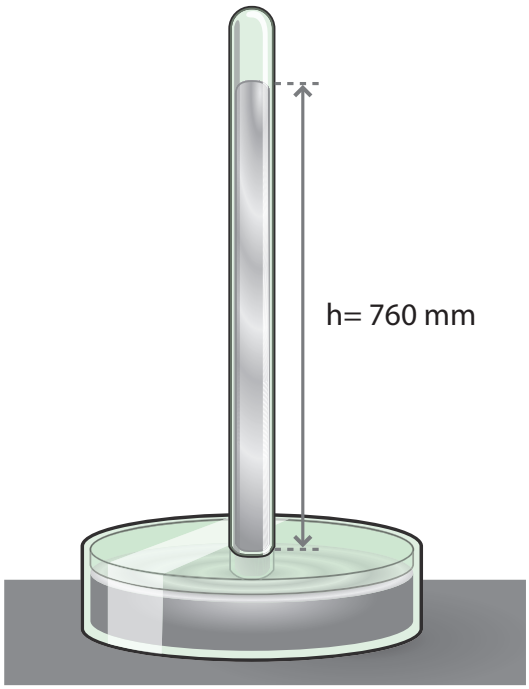


Figura 10. Barómetro de mercurio.

cal. Después se destapa el extremo abierto del tubo permitiendo que el mercurio contenido en él empiece a salir. Finalmente, se observa que el mercurio deja de salir y la altura de la columna de mercurio se estabiliza (figura 10).

La columna de mercurio deja de bajar en el momento en que la presión atmosférica (que está actuando sobre la superficie del mercurio en la bandeja) se equilibra con la ejercida por la columna. Como la presión dentro del tubo, arriba de la columna de mercurio, es prácticamente cero, la altura de la columna sobre el nivel de la bandeja indica la presión atmosférica. Cuando este experimento se realiza a nivel del mar, la altura de la columna de mercurio es de 76 cm; por ello se dice que la presión atmosférica a nivel del mar es de 760 mm de mercurio.

Cabe destacar que en el siglo XVII, Pascal reforzó la propuesta de Torricelli para medir la presión atmosférica a partir de la altura de la columna de mercurio, realizando mediciones barométricas a diferentes alturas. De esta experiencia, a cada milímetro de mercurio en el tubo se le ha denominado torr, en honor a Torricelli. También se le considera una unidad para la presión, de manera que es válido decir que la presión atmosférica a nivel del mar tiene una magnitud de 760 torr. En la actualidad se calcula la magnitud de la presión atmosférica en pascales al nivel del mar. Esto se puede hacer con los datos del experimento de Torricelli y con ayuda del modelo matemático para la presión hidrostática. Tomando en cuenta que la densidad del mercurio, es 13.6×10^3 kg/m^3, la presión hidrostática ejercida por la columna de mercurio, a nivel del mar, es:

$$\begin{aligned} P_h &= \rho g h \\ &= (13.6 \times 10^3 \text{ kg/m}^3)(9.8 \text{ m/s}^2)(0.76m) \\ &= 101.3 \times 10^3 \text{ N/m}^2 = 1.013 \times 10^5 \text{ Pa.} \end{aligned}$$

Al valor de la presión atmosférica a nivel del mar se le conoce también como *una atmósfera* (1 atm); de manera que se tienen las siguientes equivalencias:

$$1\,\text{atm} = 1.013 \times 10^5\,\text{Pa} = 760\text{mm Hg.}$$

Es importante mencionar que el barómetro es un instrumento indispensable en los trabajos de meteorología y que en ese campo de actividades a la presión atmosférica se le mide en *bar*; un *bar* equivale a 1×10^5 Pa.

5.1.6 Presión debajo de la superficie del agua

Para calcular la presión a la que se ven sometidos los buzos en las profundidades del mar y explicar el motivo por el cual procuran que su ascenso a la superficie del mar sea lento o por etapas, se presenta a continuación un ejercicio.

¿Podría una persona sumergirse a cinco metros de profundidad en el agua, y desde ese punto respirar a través de un tubo rígido de cinco centímetros de diámetro?

Se puede calcular la fuerza que actúa sobre la caja torácica de una persona sumergida a cinco metros de profundidad en el mar, considerando que el área de su caja torácica es de 0.35 m^2 y que la densidad del agua de mar es de 1.03×10^3 kg/m^3. La presión absoluta sobre esta persona sería:

$$\begin{aligned} P_a &= P_o + \rho\, gh \\ &= P_o + (1.0^3 \times 10^3\ \text{kg/m}^3)(9.8\ \text{m/s}^2)(5\ \text{m}) = P_o + 50.47 \times 10^3\ \text{Pa} \\ &= 1.013 \times 10^5\ \text{Pa} + 0.5047 \times 10^5\ \text{Pa} \\ &= 1.5177 \times 10^5\ \text{Pa}. \end{aligned}$$

Ahora:

$$\begin{aligned} P &= \frac{F}{A} \\ F &= PA \\ &= (1.5177 \times 10^5\ \text{N/m}^2)(0.35\ \text{m}^2) \\ &= 0.53 \times 10^5\ \text{N} = 53000\ \text{N}. \end{aligned}$$

Ésta es la fuerza ejercida sobre la caja torácica de la persona sumergida a cinco metros de profundidad en el mar y, aunque el cuerpo humano está en condiciones de soportar la presión atmosférica, en este caso, tanto la presión como la fuerza sobre la caja torácica están incrementadas en, aproximadamente, un 50%. Esto impide a la persona poder ensanchar su caja torácica para respirar. La solución es que los buzos respiren aire a alta presión, el cual llevan en sus tanques de aire comprimido.

Se puede comprobar que, en el mar, la presión se ve incrementada en una atmósfera aproximadamente por cada diez metros de profundidad. Entonces, si un buzo realiza su trabajo a diez metros de profundidad, estará sometido a una presión de aproximadamente dos atmósferas y el aire que respire deberá entrar a sus vías respiratorias con esa presión.

Cuando los buzos ascienden a la superficie deben hacerlo lentamente o por etapas, para dar oportunidad a que las moléculas de oxígeno y de nitrógeno, componentes del aire, sean absorbidas por su sangre. Si ascienden con rapidez, pasando de zonas de alta presión (lugares profundos) a zonas de baja presión (lugares de menor profundidad), esas moléculas dan lugar a la formación de burbujas que crecen al ascender rápidamente, las cuales, en los vasos sanguíneos, pueden causar embolias.

5.1.7 Principio de Arquímedes. Peso relativo o aparente

Es claro que los materiales menos densos que el agua flotan en ella y los materiales más densos se hunden, pero ¿cómo se puede explicar el que los barcos floten en el agua, a pesar de estar construidos con materiales densos? La respuesta se fundamenta en el principio de Arquímedes (287 a.n.e.-212 a.n.e.), el cual establece que *todo cuerpo sumergido, parcial o totalmente, en agua experimenta un empuje vertical hacia arriba igual al peso del líquido desalojado*.

Figura 11. Principio de Arquímedes 1. [Véase video en CD: "Principio de Arquímedes".]

Mediante un experimento, empleando algunas pesas de masa conocida, un dinamómetro, una probeta graduada y agua, se ilustra este principio (experimento A). La magnitud del empuje vertical hacia arriba del que habla el principio de Arquímedes, se puede determinar a partir de la expresión para la presión absoluta, haciendo algunas suposiciones sencillas.

Experimento A

Si se coloca un objeto cilíndrico en posición horizontal dentro de un tanque con agua, se observará que su cara inferior está a una profundidad "h_i", la superior a una profundidad "h_s" y la presión absoluta sobre la cara superior, "P_s" es:

$$P_s = P_o + \rho g h_s,$$

y la presión absoluta sobre la inferior, P_i:

$$P_i = P_o + \rho g h_i,$$

dado que h_i es mayor que h_s, se tiene que P_i es mayor que P_s, y habrá una diferencia de presiones $(P_i - P_s)$,

$$\begin{aligned} P_i - P_s &= (P_o + \rho g h_i) - (P_o + \rho g h_s) \\ &= \rho g h_i - \rho g h_s \\ &= \rho g (h_i - h_s). \end{aligned}$$

Esta diferencia de presiones da lugar a una fuerza vertical hacia arriba, cuya magnitud es:

$$F = (P_i - P_s) A,$$

donde A es el área de la base del cilindro. Entonces, la fuerza vertical hacia arriba que experimenta será:

$$F = \rho g (h_i - h_s) A,$$

pero, $(h_i - h_s) A = V$ es el volumen del cilindro, que es el volumen del agua desplazada. Entonces:

$$F = \rho V g.$$

Ahora, dado que ρ es la densidad del agua y V el volumen del agua desplazada, entonces (ρV) es su masa, $m_{ad.}$ La fuerza o el empuje vertical hacia arriba que experimenta el cuerpo sumergido en el agua, es igual al peso del agua desplazada por él.

$$F = m_{ad} g.$$

Otro experimento permite verificar que el principio de Arquímedes explica la flotación de los barcos. Los materiales a emplear son un pedazo de plastilina, un recipiente transparente con boca ancha y de un litro de capacidad, aproximadamente, un plumón de punta fina y agua.

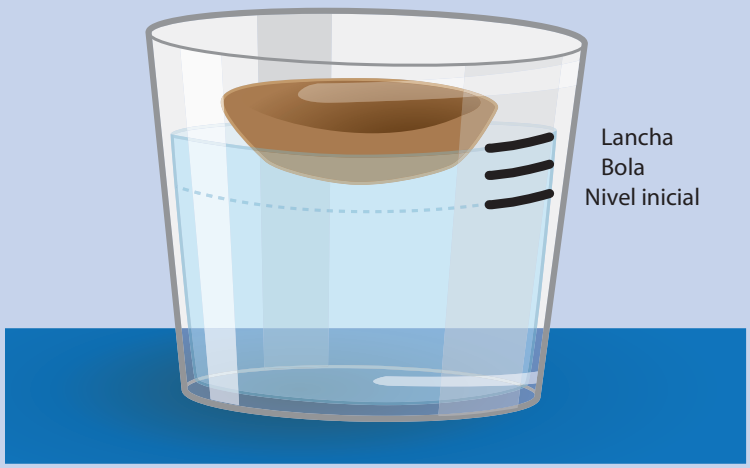


Figura 12. Principio de Arquímedes.

Se vierte agua en el recipiente hasta dos o tres centímetros debajo de su borde y en la cara exterior del recipiente se marca, con el plumón, el nivel del agua. Con la plastilina se elabora una especie de plato hondo o "lancha", cuyo diámetro quepa holgadamente en el recipiente y se coloca suavemente sobre el agua del recipiente, de manera que quede flotando. Con el plumón, junto a la marca anterior, se señala el nuevo nivel del agua.

La diferencia entre las dos marcas del plumón indica el volumen de agua desplazada por la lancha de plastilina. Ya se sabe, por el principio de Arquímedes, que la lancha está recibiendo un empuje vertical hacia arriba, cuya magnitud es igual a lo que pesa el volumen del agua desplazada.

La lancha está en reposo, ya que al estar flotando en el agua hay dos fuerzas verticales contrarias actuando sobre ella: su peso hacia abajo y el empuje vertical hacia arriba. Estas dos fuerzas verticales ejercidas sobre la lancha están equilibradas; es decir, son de igual magnitud y actúan en sentidos opuestos. Si se retira la lancha del agua, ésta regresará al nivel original.

Ahora, se desbarata la lancha haciéndose una bola de plastilina, la cual se deposita en el agua; se observará que irá al fondo del recipiente y que el agua sube de nivel, el cual será marcado junto a las dos señales anteriores. Se observará que el volumen de agua desplazada por la pelota de plastilina es menor que el volumen de agua desplazado cuando la misma cantidad de plastilina tenía forma de lancha, lo que permite concluir que el empuje vertical hacia arriba es menor.

Sobre la plastilina hecha pelota e introducida en el agua también actúan dos fuerzas verticales contrarias: su peso hacia abajo y el empuje vertical hacia arriba. El peso de la plastilina hecha pelota es el mismo que el de la lancha, pues la cantidad de plastilina es la misma, pero en la forma de lancha se desplaza un volumen mayor de agua. El peso de la plastilina hecha bola es de mayor magnitud que la fuerza producida por el agua desplazada y se va al fondo del recipiente.

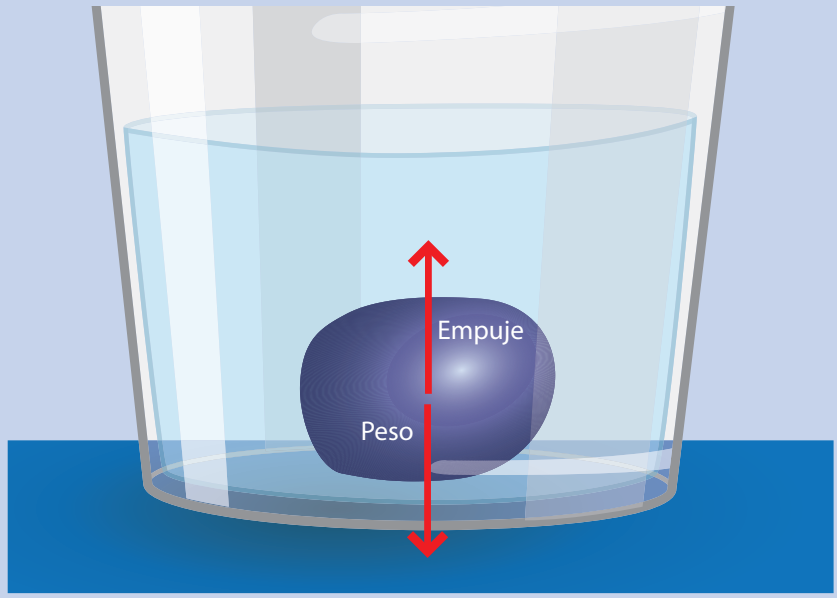


Figura 13. Esquema de fuerzas.

Entonces, los barcos están construidos con materiales más densos que el agua, pero de forma que sobre el mar desplazan un gran volumen de agua y por eso reciben un empuje vertical hacia arriba, cuya magnitud es igual a lo que pesa el volumen del agua que han desplazado. Ese empuje vertical equilibra el peso del barco y por ello flota sobre el agua.

[Véase video en CD: "Por qué flotan los barcos?"]

5.1.8 Peso aparente o relativo

Los cuerpos pesan menos al estar parcial o totalmente sumergidos en el agua, sensación que prácticamente todas las personas han experimentado al estar dentro del agua. Para calcular el peso de un objeto parcial o totalmente sumergido en el agua, también llamado *peso relativo o aparente*, se tiene:

$$P_{\text{rel}} = P - F_{\text{flot}},$$

donde P es el peso del cuerpo fuera del agua y F_{flot} es la fuerza de flotación que, calculada de acuerdo con el principio de Arquímedes, es:

$$F_{\text{flot}} = m_{ad}g = \rho V_{ad}g;$$

donde ρ es la densidad del agua y V_{ad} es el volumen de agua desplazada; entonces, el peso relativo es menor que el peso real del cuerpo:

$$P_{\text{rel}} = P - \rho V_{ad}g.$$

Considérese de nuevo un estanque con agua y dentro de ella, totalmente inmerso, un cuerpo de masa m, sobre el cual actuarán dos fuerzas verticales: su peso P, que sería su peso fuera del agua, estaría apuntando hacia abajo y el empuje vertical hacia arriba, F_{flot}.

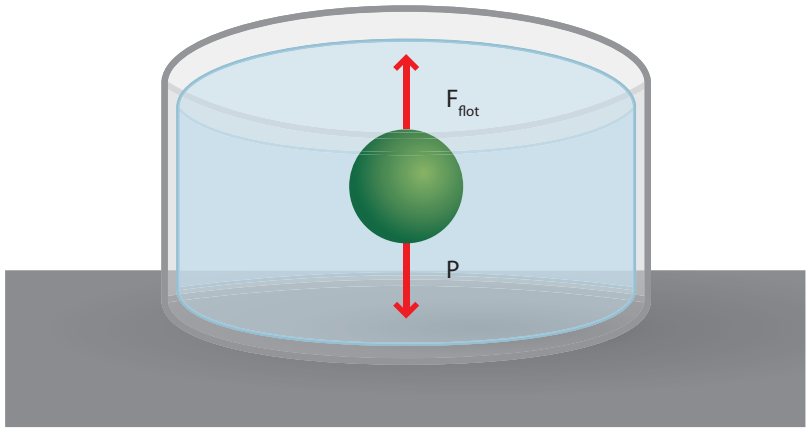


Figura 14. Esquema de fuerzas.

A la fuerza resultante de estas dos se le conoce como peso relativo o peso en el agua P_{rel}. Así:

$$P_{\text{rel}} = P - F_{\text{flot}},$$

pero la fuerza de flotación es igual al peso del volumen de agua desplazada.

$$\begin{aligned} F_{\text{flot}} &= m_{ad}g \\ &= \rho V g \ , \end{aligned}$$

donde ρ es la densidad del agua y V es el volumen de agua desplazada, que coincide con el del objeto sumergido. El peso del cuerpo u objeto, fuera del agua es:

$$P = mg = \rho_0 V g,$$

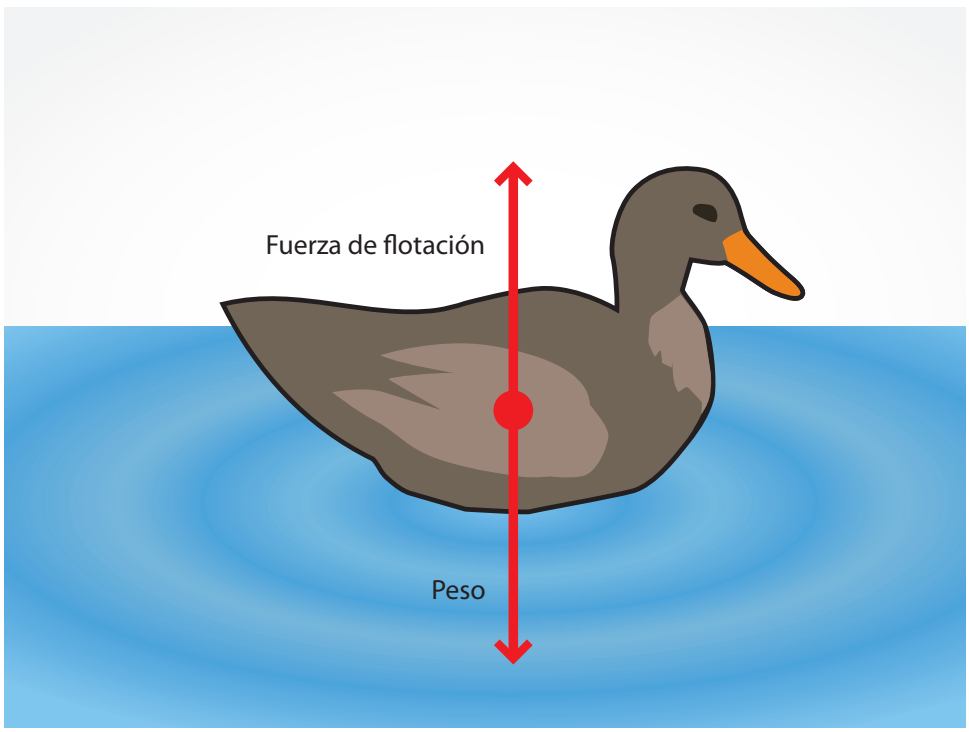


Figura 15. Esquema de fuerzas de flotación.

donde ρ_0 es la densidad del objeto y V es el volumen del mismo; entonces:

$$P_{\text{rel}} = \rho_0 V g - \rho V g = (\rho_0 - \rho) V g.$$

De ahí que

$$P_{\text{rel}} = (\rho_0 - \rho)\, V g,$$

y de aquí puede verse que:

si $\rho_0 > \rho$ entonces, $P_{rel} > 0$, y el objeto se hunde.
si $\rho_0 < \rho$ entonces, $P_{rel} < 0$ y el objeto sube a la superficie del agua y flota.
si $\rho_0 = \rho$ entonces, $P_{rel} = 0$ y el objeto se queda dentro del agua, en el punto en que se deje.

5.2 NOCIONES DE HIDRODINÁMICA

Cuando nos preguntamos: ¿por qué el agua de los ríos fluye con mayor velocidad donde el cauce es más angosto y lo hace lentamente donde el cauce es más ancho?, seguramente encontraremos respuesta gracias al estudio de los fluidos en movimiento, lo cual compete a la hidrodinámica.

5.2.1 Ecuación de continuidad

En la hidrodinámica, un concepto importante es el de "gasto", el cual se define como el volumen "V" de un fluido que atraviesa una determinada superficie por unidad de tiempo y generalmente se denota con la letra Q. Su unidad es el metro cúbico sobre segundo "m^3/s".

$$Q = \frac{V}{t}.$$

Con el concepto de gasto es posible cuantificar, por ejemplo, cuál es el volumen de agua que entra a la ciudad de México cada segundo, que es de varias decenas de metros cúbicos por segundo. O bien, la magnitud del gasto cardiaco en nuestro cuerpo, que en un hombre adulto en reposo es de aproximadamente cinco litros de sangre por minuto. Otro ejemplo es el gasto de gasolina de un coche que va en una carretera a 100 km/h, que es de aproximadamente 2 cm^3/s.

Para estudiar el flujo de agua a través de tubos cilíndricos rígidos, considérese que éste es laminar y continuo; esto significa un flujo suave y sin turbulencias.

Supóngase agua fluyendo a lo largo de un tubo cilíndrico como el de la figura 16 (p. 399), en la que se observa que a partir de cierto punto, disminuye el diámetro del tubo. En la figura se muestra un mismo volumen de agua en dos regiones diferentes del tubo. Siendo igual el volumen, el que se encuentra en la región con menor área transversal tendrá mayor longitud. Si se denomina al volumen de la parte ancha del tubo como V_1 y al volumen de la parte angosta como V_2, se tendrá la siguiente igualdad $V_1 = V_2$, es decir, $A_1L_1 = A_2L_2$, de donde:

$$\frac{A_1}{A_2} = \frac{L_2}{L_1}.$$

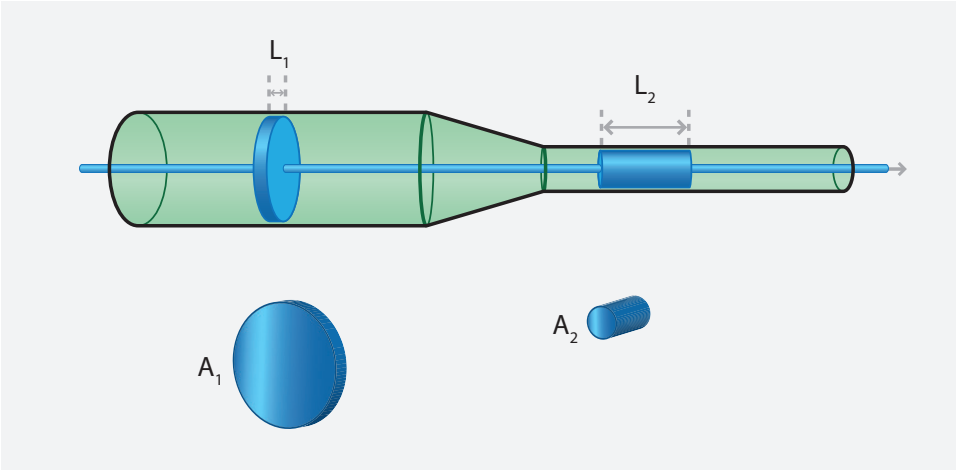


Figura 16. Tubo con reducción de diámetro. [Véase animación en CD: "Cambio de velocidad de un fluido en un tubo con reducción de diámetro".]

Si el gasto del agua que fluye por el tubo es constante y los volúmenes V_1 y V_2, son iguales, puede afirmarse que el tiempo que tarda el agua en recorrer el segmento L_1 es igual al tiempo en que recorre L_2. Como la velocidad es:

$$v = \frac{L}{t},$$

se puede escribir que:

$$t = \frac{L_1}{v_1} = \frac{L_2}{v_2}$$

donde v_1 y v_2 son las velocidades con las que el agua se desplaza en ambas partes del tubo. De esta última expresión se obtiene que:

$$\frac{L_2}{L_1} = \frac{v_2}{v_1}.$$

como:

$$\frac{A_1}{A_2} = \frac{L_2}{L_1},$$

entonces:

$$\frac{A_1}{A_2} = \frac{v_2}{v_1},$$

por lo tanto:

$$A_1 v_1 = A_2 v_2 = \text{cte.}$$

Esta ecuación es conocida como la *ecuación de continuidad* y representa la relación entre la velocidad de un fluido incompresible dentro de un tubo cilíndrico y el área transversal del mismo en dos puntos cualesquiera. Esto significa que el gasto es constante en cualquier parte del tubo. Si el área de una sección del tubo se reduce, la ecuación de continuidad predice que la velocidad necesariamente aumenta.

La respuesta a la pregunta de por qué el agua de los ríos fluye más rápidamente en las zonas donde el cauce se hace angosto y más lentamente en las zonas donde el cauce se hace ancho, se puede responder con base en que el gasto de agua en el río es constante y se puede aplicar la ecuación de continuidad.

5.2.2 Ecuación de Bernoulli

Si nos preguntamos: ¿por qué vuelan los aviones?, en la hidrodinámica también podremos encontrar respuesta, ya que el aire es también un fluido.

Para poder elevarse, un avión debe adquirir cierta velocidad en tierra.

En el siglo XVIII el científico suizo Daniel Bernoulli (1700-1782), estudiando el flujo de agua en tubos cilíndricos rígidos, descubrió que la presión en las paredes laterales del tubo era menor cuando el agua se movía con mayor rapidez, que cuando el agua se movía más lentamente. Esto ocurría también con los gases. Sus descubrimientos los publicó en su obra *Hidrodinámica*, en 1738.

Para explicar la relación entre la velocidad del fluido y la presión que éste ejerce sobre las paredes del tubo en el que se desplaza, consideremos un flujo incompresible (puede ser de agua) constante, suave y sin turbulencias, en un tubo que cambia de diámetro y de altura. La región 1 es la parte ancha del tubo y la 2 es la parte delgada. Los volúmenes de agua son iguales en ambas partes (figura 17).

Sobre el volumen de la parte inferior del tubo se ejerce la fuerza $F_1 = P_1A_1$ (recuérdese que $P = F/A$), donde P_1 es la presión en la región 1. Ahora, el trabajo realizado por esta fuerza al desplazar a este volumen es:

$$W_1 = F_1L_1 = P_1A_1L_1 = P_1V$$

donde V, es el volumen.

De manera semejante, el trabajo realizado sobre el volumen de agua en la parte superior del tubo, en ese tiempo, es:

$$W_2 = P_2A_2L_2 = P_2V.$$

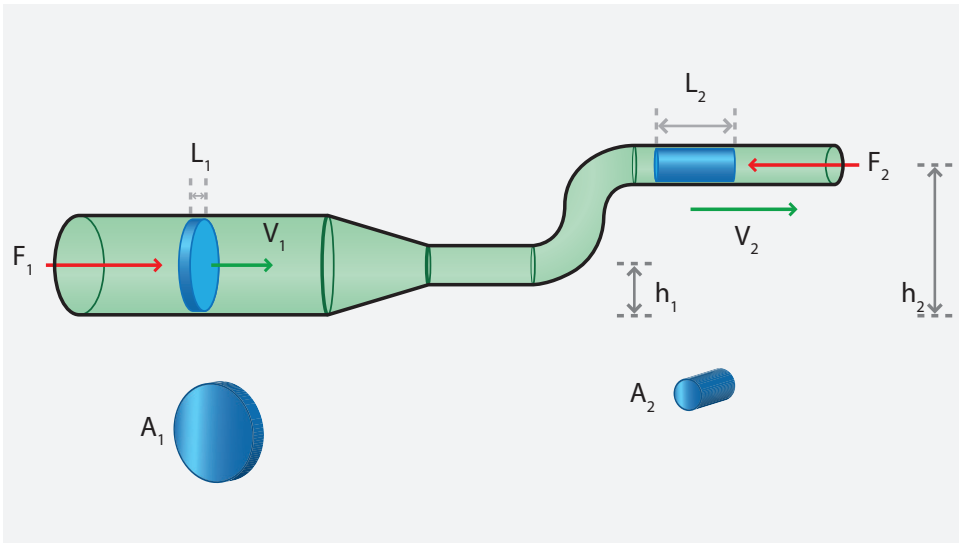


Figura 17. Tubo con cambio de diámetro y altura.

por conservación de la energía, ésta es la misma en ambas regiones:

$$\frac{1}{2}mv_1^2 + mgh_1 + P_1V = \frac{1}{2}mv_2^2 + mgh_2 + P_2V.$$

Entonces, como ρ = m/V (ρ es la densidad del agua), dividiendo cada término entre V, la expresión anterior se reduce a:

$$P_1 + \frac{1}{2}\rho v_1^2 + \rho g h_1 = P_2 + \frac{1}{2}\rho v_2^2 + \rho g h_2.$$

Ésta es la *ecuación de Bernoulli* para el flujo estacionario de fluidos incompresibles, como el agua. Esta ecuación establece que la suma de la presión, más la energía cinética por unidad de volumen, más la energía potencial por unidad de volumen, tiene el mismo valor en todos los puntos a lo largo del recorrido:

$$P + \frac{1}{2}\rho v^2 + \rho g h = \text{constante}.$$

Para el caso en que el tubo es horizontal: $h_1 = h_2$ (figura 18), la ecuación de Bernoulli se reduce a:

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2.$$

Esta expresión establece que cuando un fluido que se está desplazando dentro de un tubo pasa de una región 1 a otra región 2, con diferente diámetro, ocurre que si v_2 es mayor que v_1, entonces necesariamente P_2 será menor que P_1, y si v_2 es menor que v_1, entonces P_2 será mayor que P_1. Esto nos conduce al significado fenomenológico de la expresión: *Para un fluido en movimiento, donde la velocidad se incrementa, la presión disminuye y viceversa.*

El comportamiento de líquidos en movimiento se ha obtenido considerando que el fluido es incompresible, constante y sin turbulencias, a través de tubos cilíndricos rígidos. Sin embargo, se ha encontrado experimentalmente que funciona también, con muy bue-

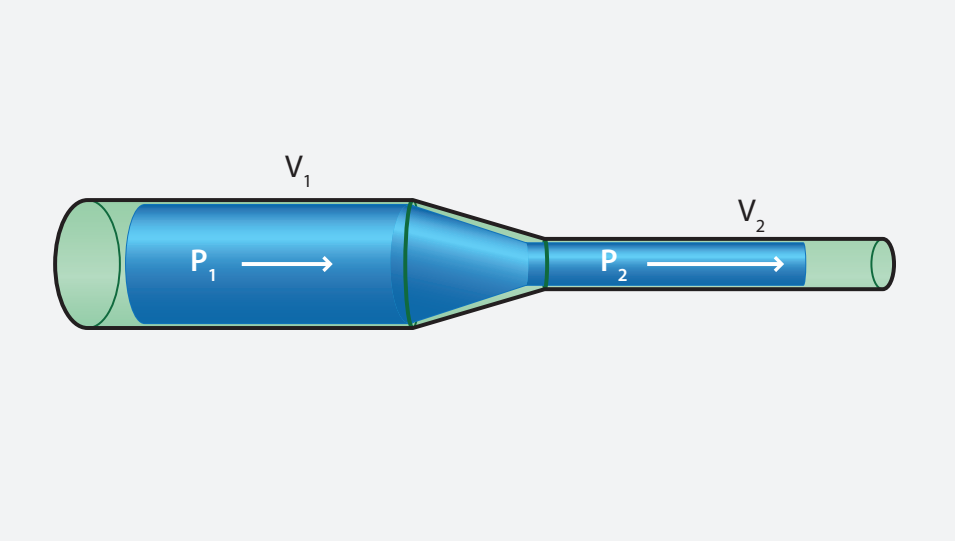
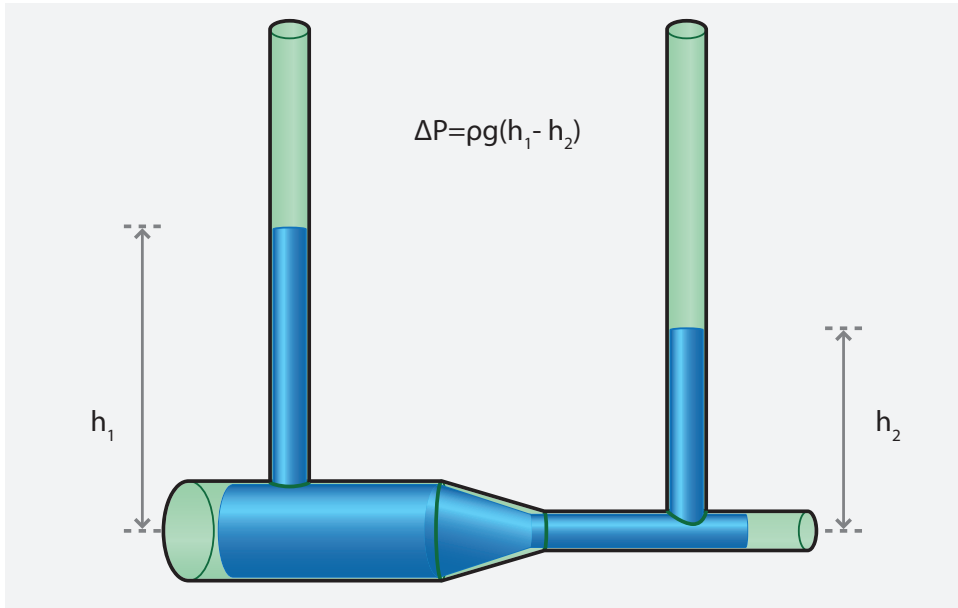


Figura 18. Presiones y velocidades en un tubo horizontal.

na aproximación, para todo tipo de fluidos (incluidos los compresibles como el aire) y a través de tubos no rígidos.

Una aplicación de la ecuación de Bernoulli se encuentra en el tubo de Venturi, estudiado por el físico italiano Giovanni Battista Venturi (1746-1822). Este tubo se ilustra en la figura 19 y con él se pueden medir velocidades de un fluido incompresible, como el agua. Así, puede determinarse la velocidad del flujo en la región 2, si se conoce la diferencia de presiones $\Delta P = P_1 - P_2$ y las áreas transversales de ambas secciones (figura 19).

Figura 19. Tubo de Venturi.



De acuerdo con las alturas del líquido en ambos tubos verticales, se tiene que, $P_1 > P_2$, y por lo tanto $v_1 < v_2$. Usando la ecuación de Bernoulli:

$$P_1 + \frac{1}{2}\rho v_1^2 = P_2 + \frac{1}{2}\rho v_2^2.$$

Además, de la ecuación de continuidad, $A_1v_1 = A_2v_2$, se tiene que $v_1 = A_2v_2/A_1$. Sustituyendo en la ecuación anterior, se obtiene:

$$P_1 + \frac{1}{2}\rho \left(\frac{A_2}{A_1}\right)^2 v_2^2 = P_2 + \frac{1}{2}\rho v_2^2.$$

de donde:

$$v_2 = A_1\sqrt{\frac{2(P_1 - P_2)}{\rho(A_1^2 - A_2^2)}}.$$

Así que, conociendo las secciones transversales del tubo, en las regiones 1 y 2, la diferencia de presiones $(P_1 - P_2)$ y la densidad del fluido, se determina la velocidad del flujo en la región 2. Aplicando este resultado y la ecuación de continuidad, se puede conocer el valor de la velocidad de flujo en la región 1.

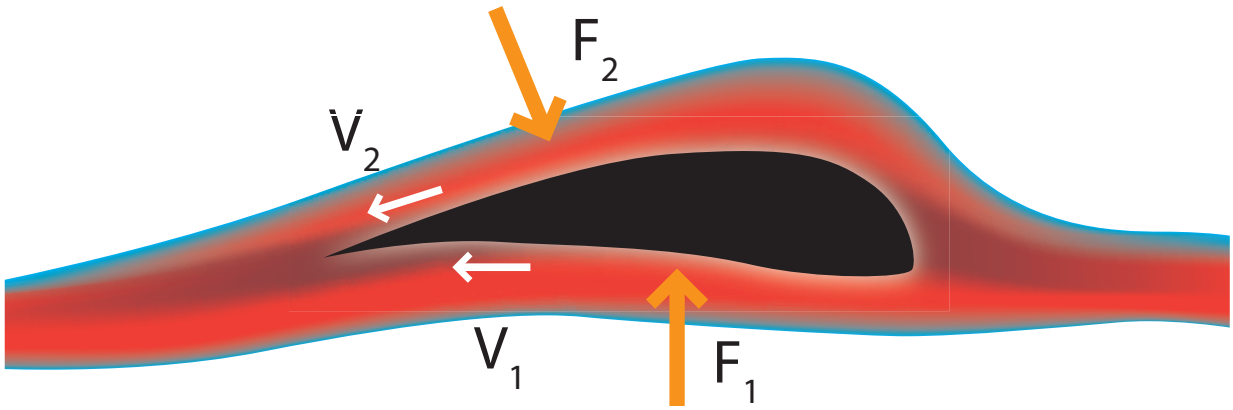


Figura 20. Ala de un avión.

Es así que, ante la pregunta de por qué vuelan los aviones, se debe considerar la forma del ala, la cual ha de ser tal que mantenga un flujo uniforme del aire (figura 20).

El aire, en la región superior del ala, recorre una distancia mayor que la que recorre el aire en la parte inferior de ella, en el mismo tiempo. Por lo tanto, la velocidad del aire arriba del ala es mayor que por debajo de ella y, debido a esta diferencia de velocidades, la presión del aire en la parte superior del ala es menor que la presión del aire en la parte inferior.

Es así que la diferencia de presiones produce una fuerza neta hacia arriba, llamada *fuerza de sustentación*, la cual depende de la velocidad del avión y del ángulo entre el ala y la horizontal.

Ahora queda claro que para poder elevarse y mantenerse en vuelo, sobre los aviones debe actuar una fuerza de sustentación y, para ello, deben adquirir cierta velocidad.

Otra aplicación de la ecuación de Bernoulli son los atomizadores y las pistolas de aire para pintar. En éstos, la corriente de aire pasa perpendicularmente por un extremo del tubo, inmerso en el líquido. El paso de aire a alta velocidad reduce la presión en la parte superior del tubo, y esta disminución hace subir al líquido hacia donde está pasando la corriente de aire y se dispersa formando una nube de pequeñas gotas. Desde luego, en este proceso físico, la presión atmosférica desempeña un papel importante para que el líquido suba por el tubo (figura 21).

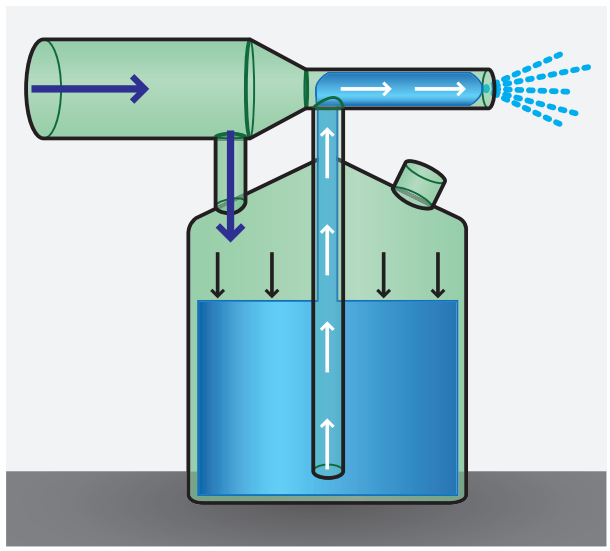


Figura 21. Diagrama de un atomizador.

Un sencillo experimento que ejemplifica lo anterior es cuando, al sostener dos hojas de papel en posición vertical y paralelas frente a la boca de una persona y al soplar suavemente entre ellas, observamos que tienden a juntarse (figura 22).

La explicación es que al soplar suavemente entre las hojas se tiene entre ellas una zona de baja presión y, por ello, la presión atmosférica que afecta a las hojas por la parte externa las empuja, haciendo que se junten.

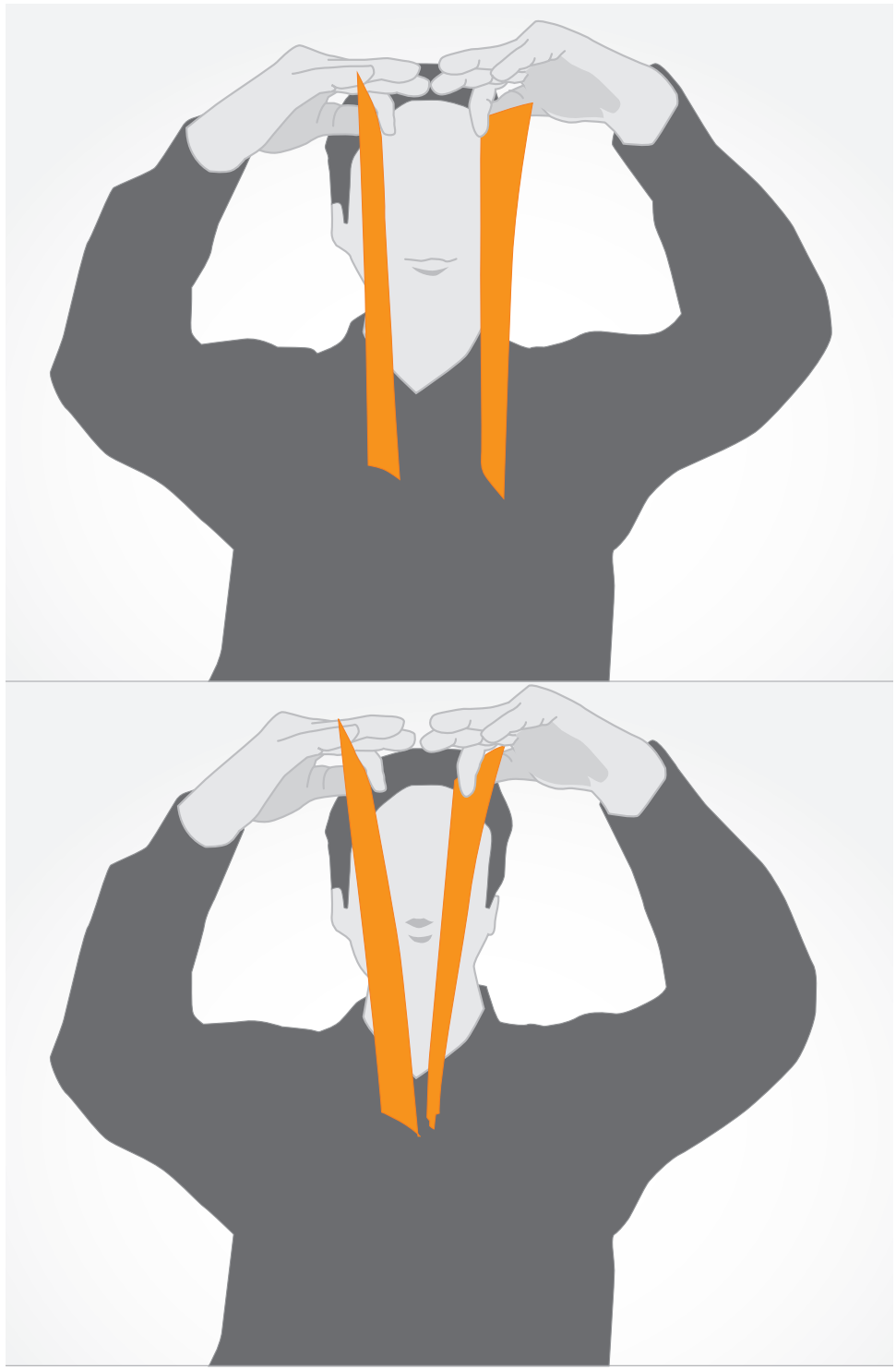
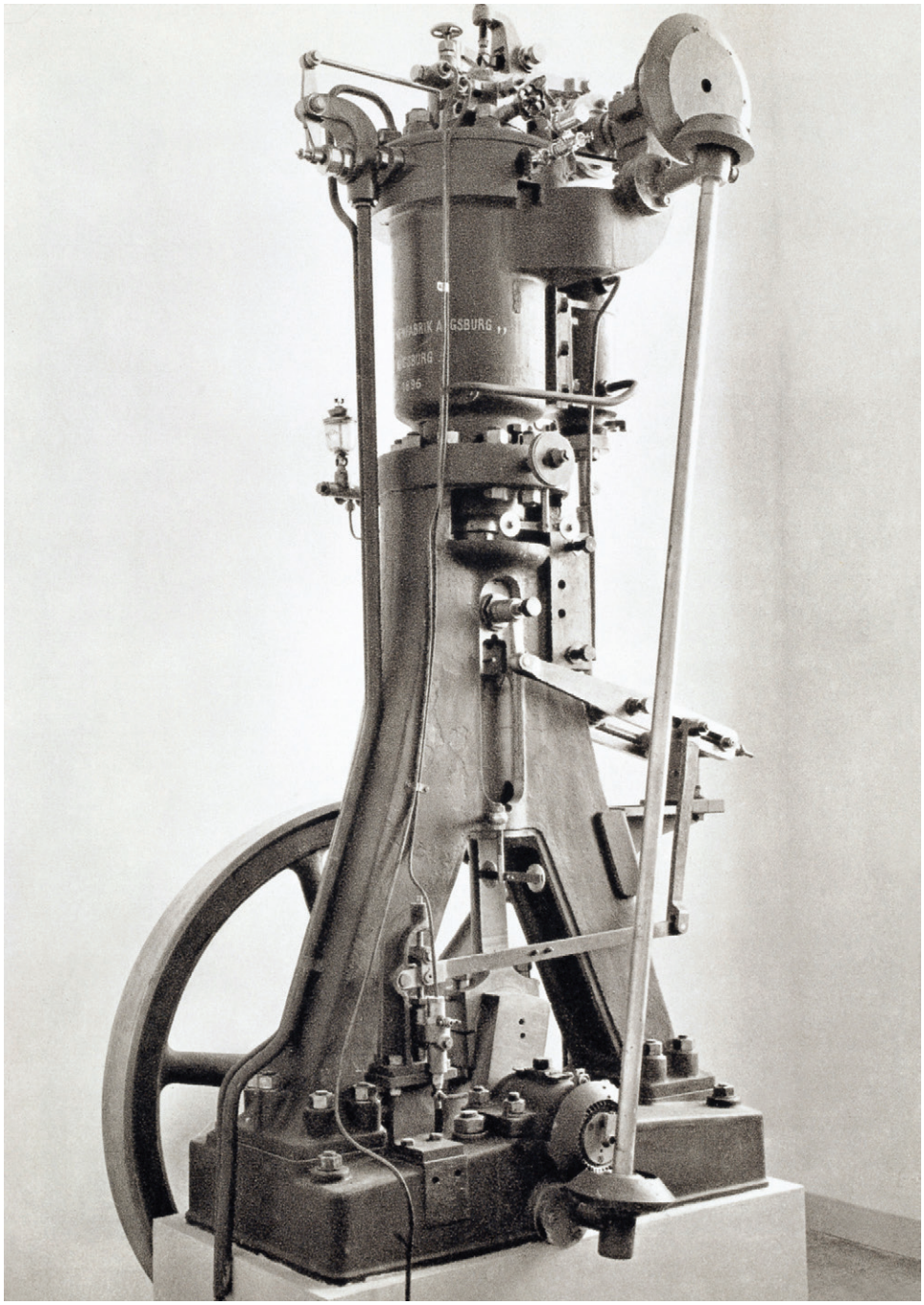


Figura 22. Persona soplando entre dos hojas.

TERMODINÁMICA

TEMA

6



© Latin Stock México.

INTRODUCCIÓN

En el presente capítulo se analizarán los aspectos fundamentales de la termodinámica, considerando sus aplicaciones a la vida cotidiana, en la que es frecuente preguntarse: ¿cómo protegerse del frío en invierno y del calor en verano?, ¿cómo disminuir el gasto de gas, leña y electricidad en casa? y, más recientemente, ¿cómo disminuir la contaminación y acceder al desarrollo sustentable?

Desde luego, las tres preguntas anteriores no fueron formuladas así en los orígenes del desarrollo histórico de la termodinámica. La noción de desarrollo sustentable, por ejemplo, no era una preocupación tan generalizada como lo es ahora, cuando los límites del crecimiento poblacional y productivo, así como su huella ambiental desbocada se avizoran en el futuro cercano, lo cual ha motivado avances notables en la disciplina.

Algunos historiadores indican que la termodinámica nació hace unos cuatro siglos, asociada a la resolución de problemas sociales prácticos, tales como la predicción del tiempo atmosférico, que permitió el entendimiento de las variaciones en el volumen de los gases al cambiar la presión y la temperatura; o por el uso generalizado de los motores térmicos a base de carbón mineral, lo cual hizo necesario entender las relaciones entre los conceptos de trabajo mecánico y calor.

Más adelante, el agotamiento de los yacimientos superficiales de carbón obligó a su extracción del subsuelo, con los consiguientes problemas de inundaciones en las minas ocasionadas por el encuentro con los mantos freáticos, presentándose así la necesidad de construir motores térmicos cada vez más eficientes en cuanto al gasto de carbón y la realización de trabajo.

Hoy en día, la extracción eficiente de combustibles fósiles como petróleo, gas y carbón, sigue siendo tan prioritaria como en aquellos lejanos tiempos, pero agravada por el hecho del agotamiento cercano de los dos primeros. Si a ello agregamos los problemas evidentes para la salud y el ambiente, ocasionados por su quemado o su utilización generalizada en múltiples actividades sociales, queda inevitablemente en el orden del día la disminución de la contaminación. Para lograr esto, al menos para reducirla notoriamente, es imprescindible replantearse el uso que hacemos de los combustibles fósiles, la electricidad y la leña en casa.

La termodinámica ayuda a enfrentar los problemas mencionados, que han sido una preocupación permanente desde que los motores térmicos hicieron su aparición. De esto se dará cuenta el lector al estudiar las siguientes páginas. Asimismo, se mostrará que a partir de la formulación de los conceptos y las leyes de la termodinámica es posible entender los aspectos básicos de nuestro mundo con las descripciones de las dimensiones humanas, inclusive las micro y macroscópicas.

La termodinámica es una parte importante de la física que comprende el estudio de las transformaciones de energía por calor y por trabajo, a través de los cambios de estado de un objeto de tamaño macroscópico. Analiza también la forma en que los objetos se acercan al equilibrio, cuando inicialmente parten de un estado de desequilibrio.

La ley cero de la termodinámica revela que la temperatura es una de las propiedades básicas de todo objeto o sistema termodinámico, que indica si un objeto está en equilibrio térmico con otro.

La primera ley de la termodinámica establece la existencia de la energía interna como una propiedad fundamental de todo sistema termodinámico y estipula la forma en que cambia esta energía al interactuar el objeto con otros cuerpos, ya sea por trabajo o por calor.

La segunda ley de la termodinámica se refiere a que los objetos tienen además otra propiedad, llamada entropía, la que siempre aumenta cuando un objeto aislado experimenta un cambio de estado, marcando de esta manera la dirección en que el proceso de transformación ocurre.

La termodinámica, como ciencia moderna, nació en el siglo XVII en Europa, asociada a la solución de problemas técnicos referentes al estudio del estado del tiempo atmosférico y su predicción, así como al estudio de la eficiencia de los motores térmicos.

6.1 ¿CÓMO PROTEGERNOS DEL FRÍO EN INVIERNO Y DEL CALOR EN VERANO?

6.1.1 Nociones preliminares sobre temperatura: paredes adiabáticas y diatérmicas. Conductividad térmica

La sensación de frío o de calor se asocia al hecho de que la temperatura de nuestro cuerpo está por encima o por debajo de la temperatura del ambiente que nos rodea.

Si estamos en un lugar donde cae nieve y el agua de los lagos, ríos o el mar se congela, sentiremos frío porque nuestro cuerpo tiene una temperatura muy superior a la de los objetos que nos rodean. Esto mismo sucede en lugares de alta montaña en México o en sitios ubicados a gran altura sobre el nivel del mar, en los meses de invierno.

Si la temperatura corporal es mayor que la del ambiente, entonces la energía de nuestro cuerpo tenderá a escapar hacia fuera, dándonos la sensación de frío. Lo que se hace para evitar que la energía escape al exterior por causa de esa diferencia de temperatura, generalmente, es taparse.

Por el contrario, si estamos en un lugar de nuestro país durante el verano, a baja altura sobre el nivel del mar, lo más probable es que la temperatura de nuestro cuerpo llegue a ser menor que la del ambiente. Entonces sentimos calor, porque ahora la energía de nuestro cuerpo tiende a aumentar al absorberla del ambiente; lo que comunmente se hace es ponerse ropa delgada o usar algún objeto que nos abastezca de aire fresco.

La temperatura es un concepto que desempeña un papel clave en cuanto a sentirnos confortables en nuestro ambiente vital. ¿Qué es la temperatura? Iremos profundizando en lo que este concepto significa, hasta llegar a su definición científica.

En la vida cotidiana sabemos que dejamos de tener frío si nos ponemos ropa caliente, si activamos un calentador en el interior del cuarto o si aislamos nuestro cuarto con lambrín (que es un material hecho de madera, y es un buen aislante térmico). O las tres cosas, si el calentador, la ropa y el aislamiento de las paredes y suelo con madera son insuficientes de manera separada.

Imaginemos el siguiente experimento, en el que se compara el efecto de ponerse un abrigo grueso con el de ponerse varias prendas delgadas hasta que sienta uno que se sofoca. ¿Cómo se explican las semejanzas y diferencias?

¿Se puede calentar una olla de agua envolviéndola con "ropas calientes"? Si se envuelve una olla de agua con un sarape o un abrigo y se siente la temperatura del agua con uno de los dedos, antes y después de cubrir la olla, ¿qué se percibiría con el dedo? ¿Se ha calentado el agua? Si ahora se envuelve a la olla de agua con una cobija eléctrica y se sumerge uno de los dedos en el agua, ¿se percibiría alguna diferencia con lo planteado en el párrafo anterior? Veremos que la olla envuelta con el sarape o abrigo no se calienta, aunque nosotros sí sentimos calor cuando nos tapamos con ellos.

Entre la olla y nosotros hay una gran diferencia y es que, a diferencia de los seres humanos, la olla no es un ser vivo. Nuestro metabolismo toma energía de los alimentos y parte de ella la irradiamos en forma de calor; usar ropa abrigadora impide que esta energía escape. Se puede consultar un libro de biología para ver las diferencias entre los llamados animales de "sangre caliente" y de "sangre fría".

Se introducen algunos conceptos que ayudarán a entender lo que sucede y que permitirán explicar, con más propiedad científica, las preguntas y los experimentos propuestos.

No es que haya ropas calientes o frías, en el sentido de que al contacto con otro cuerpo lo calienten o lo enfríen; lo que sí sucede es que algunas ropas impiden más que otras la fuga de la energía de nuestro cuerpo.

Imaginemos otro experimento: se calientan tres ollas con agua, una se rodea con un sarape o abrigo, la segunda con una delgada camiseta y la tercera no se envuelve. Si se revisa periódicamente el enfriamiento del agua de cada olla tocándola con el dedo, quizá se detecte que el agua caliente se enfría menos rápidamente cuando se cubre con un sarape que cuando se deja a la intemperie o se cubre con una camiseta.

Este hecho es inequívoco cuando se utiliza el dedo como termómetro. Suponiendo que "el dedo es un buen termómetro", concluiremos que la olla cubierta con el sarape tarda más en enfriarse que cuando se cubre con una camiseta, y aún más que cuando se deja a la intemperie sin protección alguna. De aquí se concluye que se puede hablar de mejores y peores aislantes de energía para retrasar el enfriamiento de los cuerpos calientes.

A los materiales aislantes de energía les llamaremos *adiabáticos* y a los buenos transmisores (buenos conductores) de energía los denotaremos como *diatérmicos*.

En otro experimento en el que se trata de pegar clips con parafina en una barra de metal y en una de asbesto, calentando los extremos con un mechero, se observará que el metal es un buen conductor de energía por calor, mientras que el asbesto es un mal conductor (figura 1).

Figura 1. Conducción por calor en una barra de cobre y en una de asbesto. [Véase video en CD: "Conducción de energía por calor".]

La *conductividad térmica* es una propiedad de los materiales que permite determinar si se trata de buenos o malos conductores. El valor de conductividad es muy bajo para los materiales adiabáticos y muy alto para los diatérmicos (figura 2). Como se observa en la figura, las escalas vertical y horizontal no son lineales. De lo analizado hasta aquí se podría concluir que para protegerse del frío se debe cubrir con materiales adiabáticos, como un sarape de lana grueso o muchas prendas delgadas.

6.1.2 Energía interna, calor y equilibrio térmico

Al principio de este capítulo se mencionó que otra posibilidad para no sentir frío en el invierno consiste en activar un calentador en la habitación, ya que éste libera energía

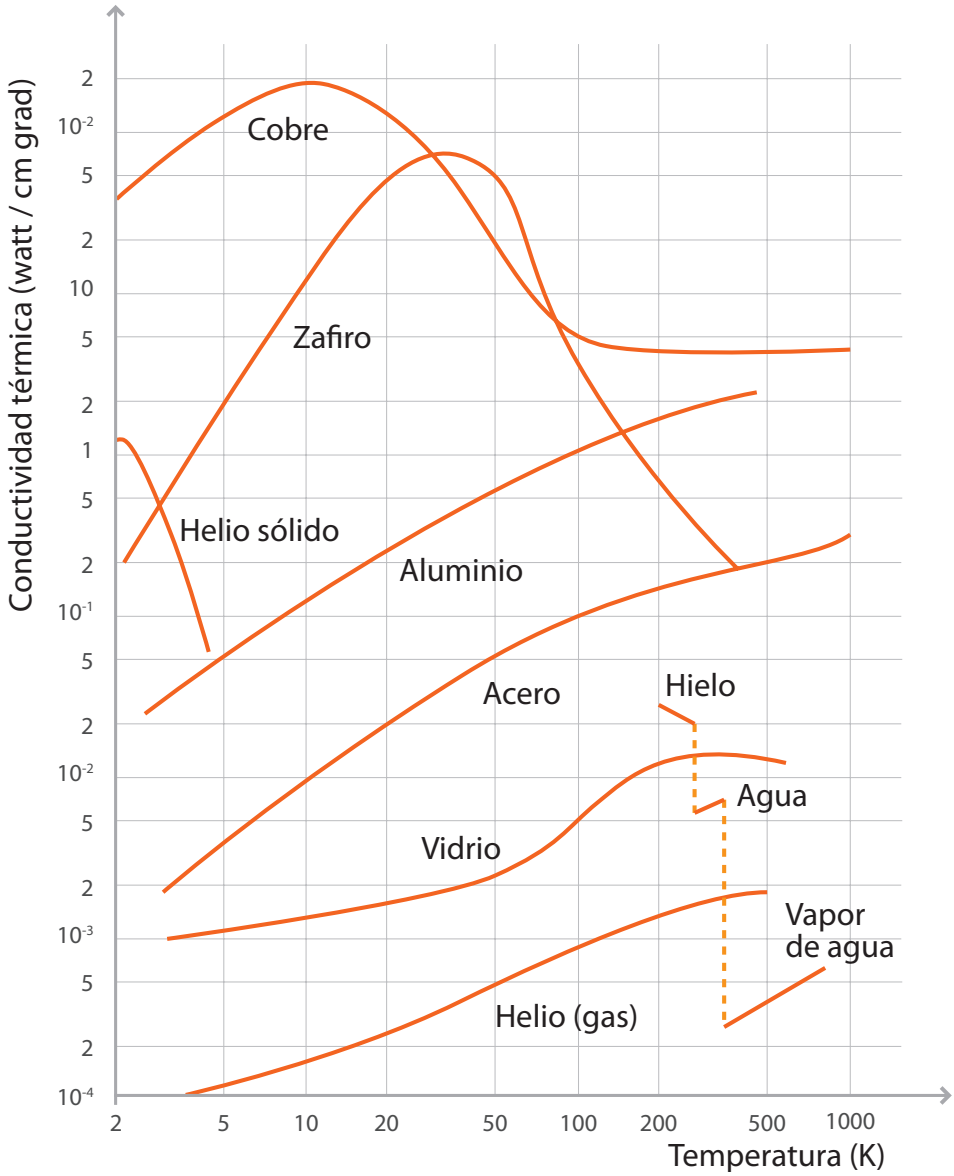


Figura 2. Conductividades térmicas de diferentes materiales.

dentro del cuarto, la que se toma de la corriente eléctrica o quemando algún combustible. El calentador transforma la energía eléctrica de la corriente en *energía interna* del aire; como resultado, el aire aumenta su temperatura (se calienta). El aumento de la temperatura del aire provoca que se pase energía a los otros cuerpos que están dentro de la habitación: las cobijas, la mesa, el librero, etc. A este paso de energía por diferencia de temperatura se le llama *calor* (se denota por la letra Q).

Se dice entonces que la energía pasa por calor, Q, del aire a los demás objetos. El calor, como proceso, deja de existir cuando la temperatura del aire y la de los demás objetos se igualan, en cuyo caso hablamos de que los objetos han llegado al *equilibrio térmico* con el aire.

Resumiendo lo anterior, diremos que:

El equilibrio térmico entre dos objetos es equivalente a que dos o más objetos tengan la misma temperatura.

Paredes forradas con una cubierta adiabática | © Latin Stock México.

El calor es el proceso mediante el cual se intercambia energía entre los objetos que tienen diferentes temperaturas. Cuando las temperaturas se igualan, el proceso de calor termina, y Q es igual a cero. Si ponemos en contacto diatérmico un cuerpo de alta temperatura con uno de menor temperatura, el primero disminuirá su energía interna mientras que la del segundo aumentará. Se dice, entonces, que el sentido en que Q procede es del cuerpo de alta al de baja temperatura.

Con el calentador se consigue que suba la temperatura del cuarto. Para que no disminuya, se tiene que hacer como con la olla de agua caliente; es decir, debemos proteger las paredes de la habitación con una cubierta adiabática que no deje escapar la energía suministrada por el calentador. De lo contrario, saldría energía del cuarto por calor, al estar la temperatura del exterior a un valor menor que la alta temperatura ya conseguida en el cuarto. El lambrín es, como antes se dijo, un material hecho de madera prensada que funciona relativamente bien como aislante adiabático.

Cuando la energía que escapa del cuarto al exterior, en un lapso de tiempo mayor que la energía suministrada por el calentador, la temperatura del cuarto tenderá a disminuir, y a aumentar en caso contrario. Si el calentador funciona al máximo de potencia y no consigue mantener una temperatura confortable a pesar de utilizar el lambrín, entonces tendremos también que vestirnos con ropas adiabáticas adecuadas.

Un razonamiento semejante lleva a utilizar "aires acondicionados", para enfrentar los climas cálidos en los que la temperatura del ambiente es mayor que la corporal. Tales aparatos lo que hacen es disminuir la energía interna del aire del espacio habitable, a fin de bajar su temperatura.

6.1.3 ¿Por qué los objetos de metal y madera se sienten a diferente temperatura?

En esta sección vamos a explicar la razón de por qué los objetos de metal y de madera se sienten a temperaturas diferentes, lo que da origen a malentendidos como los que afirman que en una habitación hay objetos calientes y objetos fríos. Es común encontrarse con afirmaciones del estilo: "los metales son fríos, las telas o la madera son calientes".

Para ello hay que reconsiderar el experimento de conducción de energía por calor, por medio de una barra de cobre y una de asbesto. Al poner uno de los extremos en la flama, sube inmediatamente su temperatura, creándose una diferencia de temperatura con respecto al otro extremo, el alejado de la flama.

En la figura 3 (p. 124), $\Delta t = t_d - t_i$ es la diferencia de temperatura entre el extremo derecho y el extremo izquierdo de la barra de longitud $L = x_d - x_i$. Esta diferencia se establece como consecuencia del paso de energía por calor Q a través de la barra con área transversal A. Experimentalmente, se encuentra que la rapidez con la que la energía por calor se transmite está dada por:

$$\frac{Q}{\text{tiempo}} = \text{potencia calorífica.}$$

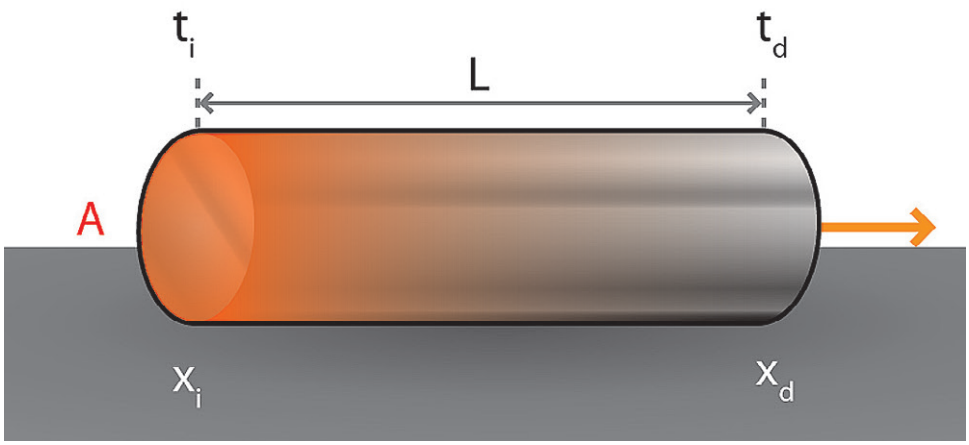


Figura 3. Conducción por calor en una barra metálica.

Esta potencia es más grande cuanto mayor son los tres factores siguientes:

- El grueso de la barra, o lo que es lo mismo, su área transversal A;
- la variación longitudinal de temperaturas en la barra, entre los extremos derecho e izquierdo $\frac{\Delta t}{L}$, y
- la conductividad térmica del material, K.

Es decir, $\frac{Q}{\text{tiempo}}$ es proporcional al producto de los tres factores anteriores:

$$A,\ \Delta t/L \text{ y } K.$$

Ahora bien, como la energía por calor Q pasa del extremo izquierdo al derecho, en la dirección positiva del eje de las *x*'s, mientras que la temperatura disminuye hacia la derecha (o sea, que Δt es negativo), tendremos que escribir un signo (–) enfrente de los tres factores:

$$\frac{Q}{\text{tiempo}} = -A\frac{\Delta t}{L}K.$$

Las unidades de Q/tiempo son joule/s = watt = W.

De esta expresión se ve que las unidades de la conductividad térmica K deben ser:

$$[K] = \left[\frac{Q}{\text{tiempo}}\right]\frac{[L]}{[A][\Delta t]} = \frac{\text{W m}}{\text{m}^2\ ^\circ\text{C}} = \frac{\text{W}}{\text{m}\ ^\circ\text{C}}.$$

El símbolo °C significa "grado centígrado o Celsius", que es la unidad de temperatura en la escala de Celsius, que se definirá más adelante.

La conductividad térmica K de la madera o de una colcha es mucho menor que la conductividad térmica de un metal (hierro, por ejemplo). Por esto, al tocar con la mano uno u otro material, la cantidad de energía que por calor Q se transfiere entre ambos (la mano y el material), en un tiempo dado, será muy diferente: mayor en el caso del metal que en

el de la madera. Por esta razón, las sensaciones térmicas provocadas en la persona serán muy diferentes, de acuerdo con la temperatura corporal, la cual puede ser mayor o menor que la temperatura del ambiente:

a] Primer caso: temperatura ambiente < temperatura corporal.
La madera y el metal están a la misma temperatura que el ambiente. En este caso, ambos materiales tienen menor temperatura que el cuerpo, por lo que se infiere que la energía por calor se transfiere de las manos a los materiales, pero más rápidamente hacia el metal que hacia la madera, dando así la sensación de que "el metal es más frío que la madera".

b] Segundo caso: temperatura ambiente > temperatura corporal.
Ahora, la energía por calor se transfiere de la madera y del metal hacia la mano; pero, como antes, la cantidad de energía por calor transferida del metal es mayor que la de la madera, dando así la sensación de que "el metal es más caliente que la madera".

En resumen, el metal se siente más frío que la madera cuando la temperatura ambiental es menor que la corporal; en caso contrario, el metal se siente más caliente que la madera.

6.1.4 Noción científica de la temperatura

La temperatura de un cuerpo es una noción científica que se refiere al "grado de calentamiento" de un objeto. Suele decirse que un cuerpo está "caliente" porque su temperatura es "alta", o que el cuerpo está "frío" porque su temperatura es "baja". Lo alto y lo bajo de la temperatura es relativo, generalmente, a nuestro estado de confort. Un cuerpo está caliente si quema nuestra mano al tocarlo, y está frío si se siente algo parecido cuando se está en contacto con el hielo.

Pero la temperatura, científicamente, nos indica si un cuerpo está en equilibrio con otro. Dos cuerpos tienen la misma temperatura cuando están en equilibrio mutuo. Esto quiere decir que si ponemos a los dos cuerpos en contacto por medio de una pared diatérmica, al conjunto lo forramos con una pared adiabática y no se observan cambios en sus propiedades, como su volumen, su presión o cualquier otra variable, entonces decimos que los objetos están en equilibrio térmico.

Nieve y lava | © *Latin Stock México.*

Si el estado físico de los cuerpos se modifica al intercambiar energía a través de la pared diatérmica, hasta que se llega al equilibrio térmico, decimos que alcanzan el estado de equilibrio. Si los cuerpos, al contacto diatérmico inicial, no cambian de estado, o sea, no intercambian energía entre ellos, quiere decir que estaban originalmente en equilibrio.

En un cuarto con diferentes objetos, nos damos cuenta de que nuestro cuerpo está "más caliente" que la colcha de la cama o que el metal del anaquel, porque nuestra temperatura es mayor que la de la colcha y los metales. Podríamos preguntarnos cuál tiene mayor temperatura. Y si tocamos ambos objetos con la mano para determinar su temperatura, seguramente se sentirá frío el metal y tibia la colcha, a pesar de que ambos están en equilibrio térmico; pero sin duda surgiría la pregunta de si las manos son buenos termómetros.

Dos cuerpos intercambiando energía.

¿Cómo se comprueba que la colcha y el metal tienen la misma temperatura? En física, al igual que en las otras llamadas ciencias exactas, los conceptos están asociados en general a cantidades que se pueden medir experimentalmente.

El siguiente experimento está encaminado a mostrar si las manos son termómetros confiables, para luego construir un termómetro más adecuado. Se colocan enfrente tres cubetas con agua, la de la izquierda con agua fría, la del centro con agua tibia y la de la derecha con agua caliente. Se sumerge una mano en la cubeta izquierda y la otra en la cubeta derecha, hasta que alcancen el equilibrio térmico con el agua correspondiente. Acto seguido, se introducen ambas manos en la cubeta de en medio. ¿Qué se siente ahora en la mano izquierda y en la mano derecha? ¿Se siente que el agua de la cubeta de en medio está a una temperatura única? ¿Sirven las manos como termómetros?

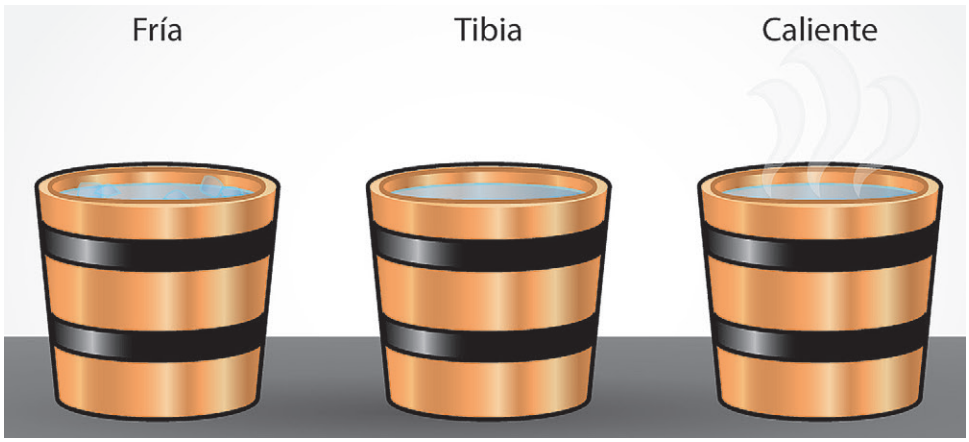


Figura 4. Experimento de las tres cubetas.

El resultado de este experimento llevaría a la conclusión de que las manos no son termómetros fiables, pues la percepción de las manos será distinta. Podríamos concluir que la piel de las manos y sus terminales nerviosas no son buenos sensores termoscópicos. Y como en física debe haber una medida precisa no contradictoria de las propiedades de los cuerpos (como la que dan las manos), conviene utilizar los termómetros de mercurio. Con este tipo de termómetro la temperatura del agua de en medio podría ser de 37 grados centígrados. ¿Qué quiere decir esta cantidad? ¿Qué son los grados centígrados?

Para contestar a estas preguntas no hay mejor respuesta que la construcción de un termómetro propio.

6.1.5 Construcción de un termómetro

La construcción de un termómetro se parece a la construcción de una regla para medir longitudes. La regla tiene divisiones de un tamaño que se fija arbitrariamente, mediante convenciones que son adoptadas mundialmente. En el sistema métrico decimal la distancia entre las divisiones más pequeñas se llama "milímetro"; diez milímetros son un centímetro y cien centímetros son un metro.

En ambos casos hay un origen desde el cual empiezan las mediciones: en la regla hay una señal en el cero, a partir del cual se inicia el conteo de la longitud; en el termómetro hay un origen también, aunque la primera cifra que aparece grabada en el vidrio es 35, porque la temperatura corporal en estado saludable es de 36.5 grados centígrados; esto quiere decir que el origen (o el cero) estaría situado 35 divisiones a la izquierda. Cada grado se compone de diez subdivisiones, o sea, de una décima de grado cada subdivisión.

La regla se gradúa aceptando que no hay longitudes negativas. ¿Qué querría decir, en caso de que se aceptaran longitudes negativas, "menos un metro de hilo"? El cero corresponde al inicio de las longitudes. Como de las longitudes se componen áreas y volúmenes, la convención de longitudes positivas implica que las áreas y los volúmenes tendrán sentido físico solamente si son positivos. Por convención, no habrá cuerpos con volúmenes negativos. El cuerpo de mínimo volumen será un punto geométrico.

Figura 5. Construcción de un termómetro y escala de temperatura. [Véase video en CD: "Construcción de un termómetro".]

De manera semejante se diseña un termómetro, aunque una lectura de temperatura está asociada con el valor que toma alguna propiedad de la sustancia de que se compone el termómetro; en un termómetro clínico, esa propiedad es el volumen del mercurio, el cual cambia cuando varía la temperatura.

Se puede construir un termómetro casero colocando un líquido coloreado en un recipiente de aproximadamente un litro, que bien puede ser agua. Después llenar con ese líquido una botella de plástico de un tercio de litro o menos y hacerle un agujero al tapón de hule o corcho para que pase por él un popote, de modo que sobresalga unos centímetros. A continuación se inserta el tapón con el popote en el cuello de la botella y se sellan los bordes con parafina derretida de una vela (procurando inclinar la botella en el proceso de sellado). Empleando un gotero se llena el popote con el líquido coloreado, de modo que sobresalga unos centímetros desde la parte superior del tapón y se agrega una gota de aceite comestible con el gotero sobre la superficie libre del líquido, para que no se evapore. ¡Ya está listo el termómetro! La propiedad que mide la temperatura es el cambio del volumen del líquido coloreado, pues cuando la temperatura varía, el volumen cambia también proporcionalmente.

Para graduar el termómetro* se tiene que recurrir a un sistema cuya temperatura en dos estados sea fácil de reproducir; por ejemplo, agua en estado de congelación y agua en estado de ebullición. ¿Qué quiere decir esto?

* Consúltese el disco compacto.

El agua se encuentra en el estado de congelación cuando una porción en estado líquido está en equilibrio con otra porción de agua en su estado sólido (o hielo).

Para poner aproximadamente el agua en el estado de congelación, se ponen varios cubos de hielo en un recipiente con poca agua y se espera a que empiecen a derretirse, mezclando agua y hielos. Al cabo de un rato, el agua y el hielo alcanzan el equilibrio térmico, quedando así hasta que los hielos se derriten por completo (por esto a tal estado se le llama "punto fijo", porque la temperatura no cambia, se queda "fija" mientras ocurre el proceso de derretimiento).

El agua en el punto de ebullición se consigue poniendo agua a hervir en una olla; este estado constituye también un punto fijo, pues mientras el agua se evapora, la temperatura no varía. Para graduar el termómetro, se sumerge la botella de líquido coloreado en la mezcla de hielos con agua; se observa que la superficie inferior del aceite en el popote se sitúa en una altura que se marca con un plumón sobre el popote y se escribe el número 0.

Enseguida se sumerge la botella con el líquido en agua hirviendo y se pinta una raya en el popote, a donde llega el nivel del líquido justamente por debajo de la superficie del aceite y se escribe el número 100. Se divide el intervalo de 0 a 100, en 100 subdivisiones; cada una corresponderá a un grado centígrado, que denotaremos por °C. Este es el grado, que ahora se llama grado Celsius, por convención universal.

Y ahora surge la pregunta: ¿qué tan bien lee este termómetro la temperatura de un objeto?, por ejemplo, la temperatura del agua tibia en el experimento de las tres cubetas.

Para esto, hay que comparar la lectura del termómetro casero con la de un termómetro de mercurio comercial al sumergirlos en la cubeta con agua fría. ¿Qué se observa? Las lecturas, ¿son iguales o son diferentes? Si son diferentes, ¿en cuánto lo son? ¿Cuál de las dos lecturas es el valor correcto de la temperatura del agua fría?

Al sumergir los dos termómetros en la cubeta con agua caliente se observa una nueva diferencia en las lecturas. Al compararla con la diferencia de las lecturas previas del agua fría, se notará que las diferencias, aproximadamente, se mantienen (aunque la pregunta sobre la temperatura correcta persiste).

¿Se podría concluir que, a pesar de las diferencias en las lecturas, es válido decir que estos termómetros son mejores medidores de la temperatura del agua tibia que las manos? ¿Acaso la temperatura del agua tibia medida por ambos termómetros difiere tanto como para decir que uno de los termómetros lee "alta" temperatura, mientras que el otro lee "baja" temperatura (como sucede con las manos)?

Hemos dicho que la temperatura nos permite saber si un objeto está en equilibrio térmico con otro. Así que, si ponemos en contacto a dos objetos con distintas temperaturas y esperamos a que lleguen al equilibrio, el valor de sus temperaturas finales será el mismo. Además, ésta puede ser medida tanto por el termómetro de mercurio como por el termómetro de líquido coloreado, pero, al margen de que estos dos termómetros lean o no la temperatura "correcta", las lecturas de cada objeto en equilibrio deben coincidir.

En una sección posterior se presentará un termómetro que sí lee la temperatura correcta de los objetos. Pero, a riesgo de resultar insistentes, enfatizaremos que, aunque se utilice un termómetro que no la lee, de todos modos cuando dos objetos están en equilibrio térmico las lecturas hechas con este termómetro "imperfecto" deben coincidir.

Con la finalidad de dar un significado más profundo al concepto de temperatura, presentaremos enseguida la llamada ley cero de la termodinámica. Se llama cero porque cuando fue formulada ya habían sido postuladas las leyes primera, segunda y tercera y, dada su importancia, fue necesario introducirla pero sin cambiar el orden de las otras leyes. Como debería ser la inicial, se la llamó cero.

6.1.6 Ley cero o de transitividad de la termodinámica

Pensemos ahora en tres cuerpos que podemos llamar A, B y C. Si se miden las temperaturas de A y B con un termómetro y las lecturas son iguales, y el termómetro da también la misma lectura para los cuerpos A y C, entonces al poner en contacto diatérmico a los cuerpos B y C estarán en equilibrio térmico (en el termómetro se leerá la misma temperatura).

Se podría decir que esto es obvio, porque si $t_A = t_B$ y, al mismo tiempo, $t_A = t_C$, pues se tiene que cumplir que $t_B = t_C$.

Además, A, B y C guardan una relación entre ellos al hallarse en equilibrio mediante una parte diatérmica. A esta relación le llamaremos R. *Si A está relacionado con B y A está relacionado con C, entonces B y C también están relacionados a través de R.* Cuando se cumple lo anterior, se dice que la relación R es *transitiva*, lo que indica que tienen una propiedad en común que, además, tiene el mismo valor. James Clerk Maxwell (1831-1879), el gran científico escocés que logró formular las leyes básicas del electromagnetismo, dice, en un libro que escribió sobre termodinámica:

> Ley de la igualdad de temperaturas: cuerpos cuyas temperaturas son iguales a aquella del mismo cuerpo tienen a su vez la misma temperatura. Esta ley no es una perogrullada, sino que expresa el hecho de que si al sumergir un pedazo de hierro en una cubeta de agua está en equilibrio con el agua, y si la misma pieza de hierro, sin alterar su temperatura, se transfiere a una cubeta de aceite, y se encuentra que también se halla en equilibrio térmico con el aceite, entonces si el aceite y el agua se ponen en la misma cubeta estos estarán en equilibrio, siendo este resultado válido para cualesquiera otras sustancias. Esta ley, en consecuencia, dice mucho más que el axioma de Euclides: Cosas que son iguales a la misma cosa son iguales entre sí, y es el fundamento de toda la ciencia termométrica.

Se podría establecer inversamente: si el cuerpo A está en equilibrio térmico con B, pero también con C, entonces al poner en contacto B y C estarán a su vez en equilibrio térmico; es decir, que los tres cuerpos guardan entre sí una relación R transitiva con respecto al equilibrio térmico, y esto es así porque tienen algo en común: su temperatura.

6.1.7 El termómetro de gas a volumen constante y la lectura "correcta" de la temperatura de un objeto

Cuando las lecturas de los termómetros de mercurio y del líquido coloreado son diferentes, entonces, a pesar de que se sabe que la temperatura es una propiedad intrínseca de los cuerpos, podría preguntarse si existe un termómetro que pueda medir la temperatura con exactitud.

El primer termómetro que la historia de la ciencia registra con la propiedad de medir la temperatura correcta es el conocido como "termómetro de gas a volumen constante", que se representa en la siguiente figura. En este termómetro, que trabaja a volumen constante, se usa la presión del gas para medir la temperatura. Cuando la temperatura del gas aumenta también lo hace su energía interna y lo mismo ocurre con la presión.

La presión del gas se debe a los choques de las moléculas contra las paredes del recipiente que lo contiene; en cada choque la molécula empuja a la pared con una cierta fuerza, que es debida al cambio de la cantidad de movimiento (masa × velocidad) en el impacto. Como la velocidad de las moléculas aumenta con la energía que se comunica al gas, se produce un incremento en la presión.

Como se ve en la figura 6, el bulbo de volumen V está conectado a un manómetro de mercurio (instrumento que mide la presión) por medio de un tubo delgado, de modo que el aire del bulbo que se filtra por este tubo no altera apreciablemente el volumen del gas en el bulbo. Puede considerarse, entonces, que el volumen del gas es el encerrado en el volumen V y las medidas de presión serán a volumen constante.

La presión del gas se determina por la diferencia de altura *h* de las columnas de mercurio mediante la expresión:

$$p = p_a \pm \rho g h,$$

donde p_a es la presión atmosférica, ρ es la densidad del mercurio y g es la aceleración de la gravedad.

El signo positivo o negativo en la ecuación anterior se determina de la siguiente manera: si la altura de la columna izquierda de mercurio es menor que la de la derecha, se utiliza el signo (+) y la presión del aire del bulbo es mayor que la presión del exterior. Si la altura de la columna izquierda es mayor que la de la derecha, se utiliza el signo (–) y, por lo tanto, la presión del interior del bulbo es menor que la del exterior.

En este termómetro de gas a volumen constante, cuando la temperatura t aumenta en Δt, la presión aumenta en Δp. Así, a mayor temperatura en el bulbo, mayor presión, y a menor temperatura del gas en el bulbo, menor presión. Se puede, entonces, definir la relación entre la temperatura y la presión con una relación de proporcionalidad, de modo que:

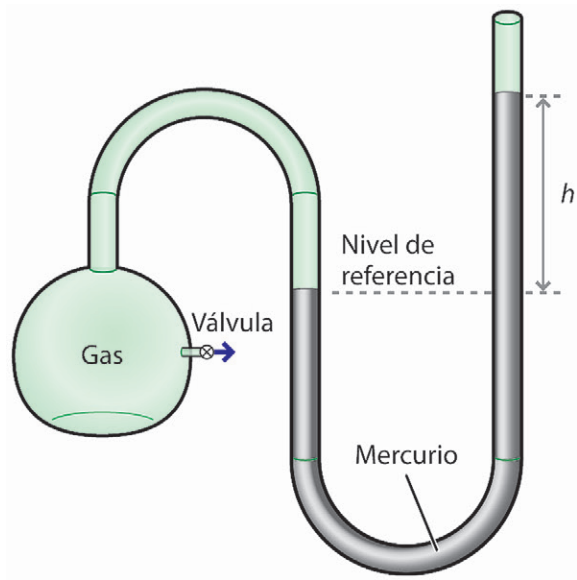


Figura 6. Termómetro de gas a volumen constante.

$$\Delta t = A\ \Delta p, \text{ a volumen constante,}$$

donde A es una constante igual a tan ϕ, como se puede ver en esta figura.

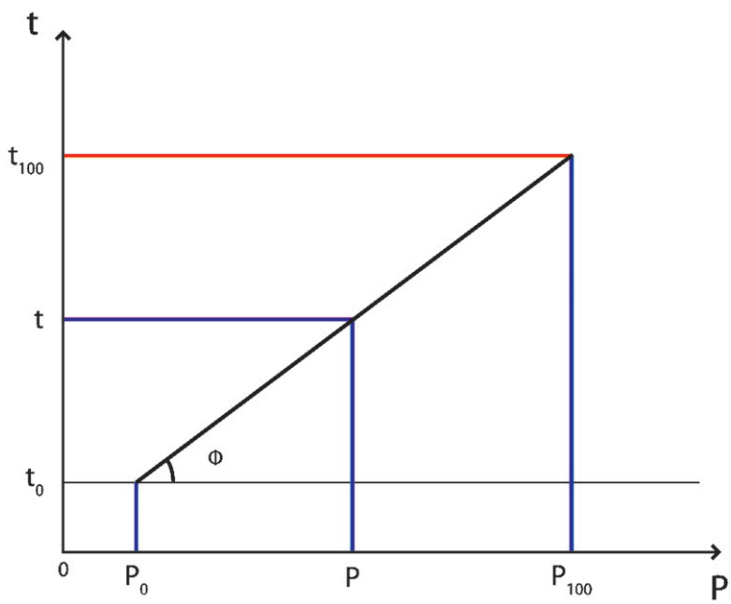


Figura 7. Construcción de una escala termométrica.

El procedimiento para construir la escala de temperatura con este termómetro se parece al procedimiento ya descrito para graduar el termómetro de líquido coloreado, visto anteriormente; es decir, primero se sumerge el bulbo en agua en equilibrio con hielo y se mide la altura del mercurio, que llamaremos h_o, calculando la presión correspondiente p_o según la ecuación:

$$p_0 = p_a \pm \rho g h_0.$$

Enseguida se sumerge el bulbo en agua hirviendo y se mide la nueva altura, que llamaremos h_{100}, calculando la presión correspondiente $p_{100:}$

$$p_{100} = p_a \pm \rho g h_{100}.$$

A cada una de estas presiones le asociamos las temperaturas correspondientes t_0 y t_{100}. Entonces, de la figura se tiene la siguiente relación trigonométrica:

$$\tan \varphi = \frac{t_{100} - t_0}{p_{100} - p_0} = \frac{t - t_0}{p - p_0};$$

de aquí se obtiene:

$$t = \frac{t_{100} - t_0}{p_{100} - p_0}(p - p_0) + t_0.$$

Si se asigna, como hizo Celsius, los valores $t_0 = 0°C$ y $t_{100} = 100°C$, se obtiene:

$$t = \frac{100}{p_{100} - p_0}(p - p_0).$$

Ahora bien, si se quiere medir la temperatura del agua tibia en la cubeta con este termómetro, se coloca el bulbo dentro de ella, se mide *h* y se determina p, de:

$$p = p_a \pm \rho g h.$$

Este valor de *p* se sustituye en la ecuación anterior y se calcula t, que es el valor de la temperatura del agua tibia en la cubeta de en medio.

Se puede comprobar en la ecuación de t que $t = 0$ si $p = p_0$, y que $t = 100°$ C cuando $p = p_{100}$. En relación con estas medidas, es interesante notar que:

1] Si se emplea aire en el bulbo, el valor de t calculado cambia cuando se extrae aire por la válvula, es decir, cuando se baja la densidad.
2] Lo mismo sucede si en el bulbo se emplea otro gas, por ejemplo dióxido de carbono (el gas de un refresco). En general, la lectura de t depende de la naturaleza del gas y de su densidad.
3] Si ahora se extrae más gas del bulbo mediante la válvula, de modo que la densidad del gas dentro del bulbo sea aún más baja, las diferencias en los valores de t calculados con gases diferentes, pero cada vez menos densos, tienden a disminuir y a aproximarse a 0, conforme la densidad tiende a 0.

Este hecho indica que la temperatura medida con un gas cualquiera, siempre y cuando su densidad sea cercana a 0, nos dará el valor "verdadero" de la temperatura del agua tibia; y también será el valor "verdadero" para cualquier cuerpo cuya temperatura se mida con este tipo de termómetro de gas.

Entonces, efectivamente, los científicos han encontrado al menos un termómetro confiable, el termómetro de gas a volumen constante de muy baja densidad. Este hecho fue fundamental para el establecimiento de la termodinámica como ciencia exacta.

Cualquier otro termómetro —por ejemplo el clínico de mercurio— se graduó de modo que su escala coincidiera con la del termómetro de gas muy diluido. Así sucedió históricamente en las primeras etapas del desarrollo de la termodinámica moderna.

Otras escalas de temperatura

Hay una escala que se llama centígrada, porque entre el agua en equilibrio con hielo (t_{cong}) y el agua en ebullición (t_{eb}) hay exactamente, por construcción, cien grados. Es una escala muy vieja, pues fue diseñada por el físico sueco Anders Celsius (1701-1744), alrededor de la primera mitad del siglo XVIII. Hoy en día esta escala recibe el nombre de su inventor, por lo que es conocida como escala Celsius.

En la escala Celsius:

$$\begin{aligned} t_{\text{eb}} - t_{\text{cong}} &= 100^\circ\text{C}; \\ t_{\text{cong}} &= 0^\circ\text{C}; \\ t_{\text{eb}} &= 100^\circ\text{C}. \end{aligned}$$

En la escala Fahrenheit (1686-1736), en cambio, se escoge:

$$\begin{aligned} t_{\text{eb}} - t_{\text{cong}} &= 180^\circ\text{F}; \\ t_{\text{cong}} &= 32^\circ\text{F}; \\ t_{\text{eb}} &= 212^\circ\text{F}. \end{aligned}$$

Se obtiene así la relación entre las lecturas Celsius y Fahrenheit:

$$^\circ\text{C} = \frac{5}{9}(^\circ\text{F} - 32).$$

6.1.8 Ecuación de estado de un "gas muy diluido" o gas ideal o perfecto

¿Cómo varía el volumen con la temperatura, si ahora se deja fija la presión? El conocimiento de esta relación tiene gran importancia práctica, por ejemplo en el diseño de sistemas de ventilación, en que la dilatación de los gases por el incremento de la temperatura provoca corrientes de aire.

Por otra parte, el conocimiento del cambio de volumen con la temperatura a presión constante es relevante en materiales sólidos, como el hierro con el que se construyen las vías de ferrocarril.

En las vías se deja un espacio libre entre los rieles, para que al aumentar la temperatura tengan espacio para dilatarse; de lo contrario, la vía se levanta y se tuerce en los puntos de contacto, como se aprecia en la figura 8 (p. 420).

Figura 8. Vías torcidas en el punto de unión | © Latin Stock México.

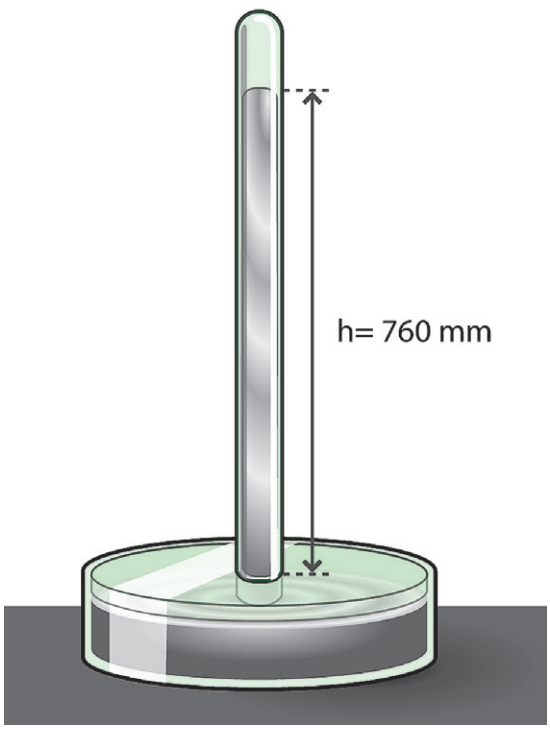


Figura 9. Barómetro de mercurio.

Al inicio se mencionó que la necesidad de saber el comportamiento de los gases cuando cambian las tres propiedades —presión, volumen y temperatura— surgió del diseño de barómetros, útiles en la predicción del tiempo atmosférico, pero también del diseño de los motores térmicos.

El barómetro es un instrumento que mide la presión atmosférica y el más común consta, básicamente, de una columna de mercurio (figura 9).

Se requiere conocer el cambio del volumen con la presión a temperatura constante, que es la ambiental. Una presión alta hace subir la columna de mercurio e indica "mal tiempo", y una presión baja, "buen tiempo".

En un motor térmico (véanse más adelante las figuras de los motores térmicos de Newcomen y Watt, pp. 443-444) el vapor de agua mezclado con el aire se dilata y contrae en el cilindro, haciendo que el pistón, junto con la acción del aire atmosférico, se mueva de arriba a abajo. En estos movimientos, la presión, temperatura y volumen de la mezcla dentro del cilindro cambian continuamente. Entender la relación entre estos cambios es básico para conseguir un motor eficiente, es decir, un motor que optimice el carbón que se quema en la caldera.

A la relación que hay entre las variables de un gas (volumen, presión y temperatura) se le conoce como *ecuación de estado*. A los físicos y químicos les tomó cientos de años

descubrir la ecuación de estado de los gases. Por su simplicidad, la primera ecuación de estado que se obtuvo fue la de los gases poco densos o muy diluidos, también llamados gases perfectos o ideales.

Leyes de Boyle-Mariotte y Gay Lussac-Charles

La ley de Boyle-Mariotte especifica la relación entre el volumen y la presión de un gas, cuando la temperatura se mantiene constante. La ley se llama así porque fue obtenida de manera independiente por los científicos Robert Boyle (1627-1691), en 1660, en Inglaterra, y Edme Mariotte (1620-1684), en 1676, en Francia. La relación se puede determinar con una jeringa a la que se le quita la aguja y se sella el orificio. Ésta se mantiene en posición vertical, como se ilustra en la figura 10.

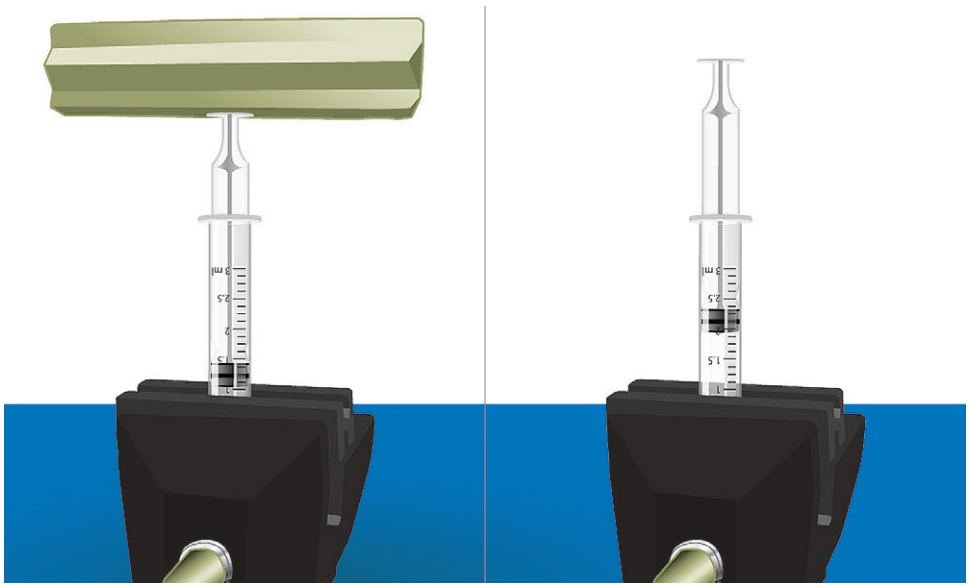


Figura 10. Experimento que ilustra la ley de Boyle-Mariotte. [Véase video en CD: "Experimento ley de Boyle".]

Después se agregan pesos variables sobre la base del émbolo de la jeringa y se analiza lo que pasa con el volumen del aire encerrado en su interior.

Las medidas se grafican en un plano cartesiano, donde el eje vertical corresponde al volumen y el eje horizontal a la presión, obtenida de dividir cada peso entre el área del émbolo. El experimento se realiza a temperatura constante, ya que el aire del interior de la jeringa está en contacto diatérmico con la atmósfera, de modo que:

$$pV = C \text{ (ley de Boyle-Mariotte)},$$

donde C es una constante relacionada con la temperatura del cuarto; se espera que cambie de valor al variar la temperatura a la que se realiza el experimento.

Si a temperatura constante se introduce más gas a la jeringa, el volumen tendrá que aumentar proporcionalmente, de modo que C no sólo depende de la temperatura, sino que también varía proporcionalmente con la cantidad de gas en el interior. Es decir:

$$C = NA,$$

donde ahora A es el único factor que depende de la temperatura, y N es la cantidad de moléculas contenidas en el gas atrapado en la jeringa.

Siendo A proporcional a la temperatura (lo que ha sido determinado experimentalmente), se puede llamar k a la constante de proporcionalidad, de manera que:

$$A = kt.$$

Sustituyendo A en la ecuación para C, queda:

$$C = Nkt.$$

Reemplazando este valor de C en la ley de Boyle-Mariotte, nos da que:

$$PV = Nkt.$$

Es claro que t, en esta ecuación, no puede estar expresada en ninguna de las escalas de temperatura hasta ahora definidas.

Hay que tener en cuenta que en ambas escalas, Celsius y Fahrenheit, el origen de cada una se fija arbitrariamente, por lo que el cociente:

$$\frac{PV}{Nk}$$

dependerá de dicho origen, mientras que p, V, N y k son cantidades que no son arbitrarias, sino que se determinan objetivamente.

Como en muchas otras cuestiones que demandan una respuesta apropiada, la indagación histórica proporciona el camino correcto: el científico francés Gay-Lussac (1778–1850) encontró que, a presión constante, el volumen V variaba linealmente con la temperatura en la escala Celsius.

Además, Gay-Lussac y otros científicos hallaron que el cambio relativo de volumen, con la temperatura $\Delta t = t - t_0$,

$$\frac{1}{V_0}\frac{\Delta V}{\Delta t} = \frac{1}{V_0}\frac{V - V_0}{\Delta t};$$

tiene el mismo valor para todos los gases, cuando su densidad es muy baja, es decir, cuando están muy diluidos o se comportan "idealmente"; es decir:

$$\frac{1}{V_0}\frac{\Delta V}{\Delta t} = B_0 = 0.003661\frac{1}{^\circ\mathrm{C}};$$

donde B_0 es el llamado *coeficiente de dilatación volumétrico a presión constante*.

En la ecuación anterior, V_0 es el valor del volumen a la temperatura inicial $t_0 = 0°$ C.

De la última ecuación, despejando ΔV se obtiene:

$$\Delta V = V - V_0 = B_0 V_0 \Delta t = B_0 V_0 (t - t_0) = B_0 V_0 (t - 0) = B_0 V_0 t.$$

De aquí:

$$V = V_0 + B_0 V_0 t = V_0(1 + B_0 t) = B_0 V_0 \left(t + \frac{1}{B_0}\right).$$

Ésta es la ley de Gay-Lussac, igualmente llamada de Charles (1746-1823), por haber sido descubierta también por este físico francés, de manera independiente.

Hay que recordar que en esta última ecuación, t es la temperatura en la escala Celsius y que, por eso, se trata de una temperatura que no es objetiva debido a que su origen ha sido fijado arbitrariamente, al asignar el valor 0 °C al punto de congelamiento del agua. Sin embargo, dado que el volumen tiene que ser mayor o igual a 0 m^3, debe cumplirse que:

$$t + \frac{1}{B_0} \geq 0,$$

es decir,

$$t \geq -\frac{1}{B_0} = -\frac{1}{0.003661\left(\frac{1}{°C}\right)} = -273.15°C.$$

En otras palabras, t tiene que ser mayor o igual que –273.15 °C; de lo contrario, para temperaturas menores, el volumen de los gases sería negativo. El valor mínimo que se puede alcanzar en la escala Celsius es, por lo tanto, de –273.15 °C; es decir, el valor mínimo es un valor no arbitrario, que bien puede ser el origen objetivo de otra escala de temperatura también objetiva. Esta temperatura se define por:

$$T = t + \frac{1}{B_0} = t(\text{en } °C) + 273.15°C.$$

T es la escala llamada de Kelvin, en honor de William Thomson (1824-1907), quien más tarde sería Lord Kelvin. También es llamada escala de temperatura objetiva o absoluta. En esta escala se define el tamaño del grado Kelvin igual al tamaño del grado Celsius. En términos de ella, la ley de Gay Lussac-Charles queda como:

$$V = B_0 V_0 T.$$

Obviamente V > 0 si T > 0. Pero siempre se tendrá que T > 0, porque t > –273.15 °C. El valor T = 0 K no se puede alcanzar, pues en tal punto el volumen de cualquier gas sería 0, lo que físicamente no tiene sentido.

En conclusión, la temperatura T que aparece en la ecuación debe ser la temperatura Kelvin. Con esto se tiene que los gases ideales cumplen con la siguiente relación entre sus variables p, V, T y N:

$$pV = NkT.$$

De modo que, tanto del lado izquierdo como del lado derecho de la ecuación se tienen cantidades objetivas. N es un número extraordinariamente grande, pues se trata de la cantidad de átomos o moléculas que hay en el volumen *V*, por ejemplo, un litro.

Conviene recordar ahora que la materia está formada de átomos; pero si los átomos se agregan en unidades mayores, entonces forman moléculas.

El agua, por ejemplo, está constituida de moléculas, formada cada una de ellas por dos átomos de hidrógeno y un átomo de oxígeno, por lo que su fórmula constitutiva se escribe como H_2O.

El agua puede estar en estado de vapor, en estado líquido o en estado sólido, dependiendo de la distancia relativa entre las moléculas. En el estado gaseoso las moléculas estarán más alejadas unas de otras que en los estados líquido o sólido.

La *cantidad de sustancia* es la cantidad de átomos o moléculas que una cierta porción de aquélla contiene. La unidad de cantidad de sustancia es el *mol*, que es el número de átomos que hay en 12 gramos del isótopo 12 de carbono.

En un mol de átomos de carbono 12 hay 6.02×10^{23} de estos átomos. Asimismo, la cantidad de moléculas H_2O que hay en un mol de moléculas de agua es también 6.02×10^{23}. De hecho, por definición, todos los mol de cualquier sustancia contienen la misma cantidad de átomos o moléculas: 6.02×10^{23}. Y el *número de Avogadro* es la cantidad de elementos estructurales (átomos o moléculas) que hay en un mol. Ese número es 6.02×10^{23} y se le representa por N_A:

$$N_A = 6.02 \times 10^{23}\ \text{mol}^{-1}.$$

Después del número 6.02×10^{23} aparece la unidad

$$\text{mol}^{-1} = \frac{1}{\text{mol}},$$

porque es el número de elementos estructurales por cada mol. Así, ¿cuántos elementos estructurales habrá en 2 mol de moléculas de agua? Pues:

$$N = 2\text{mol} \times 6.02 \times 10^{-23} \times \text{mol}^{-1} = 2 \times 6.02 \times 10^{-23} \quad \text{moléculas de } H_2O.$$

Si en una sustancia hay N elementos estructurales, podemos expresar tal número en mol, dividiendo entre N_a. A la cantidad de moles la representaremos por la letra n. Por ejemplo, en el caso anterior:

$$n = \frac{N}{N_A} = \frac{2 \times 6.02 \times 10^{23}}{6.02 \times 10^{23}\text{mol}^{-1}} = 2\,\text{mol}.$$

Siguiendo el razonamiento, se puede concluir que en 2 mol de moléculas de agua hay 4 mol de átomos de hidrógeno, 2 mol de átomos de oxígeno, 20 mol de protones, 20 mol de electrones y 16 mol de neutrones.

La cantidad de mol de electrones se obtiene así: en una molécula de agua H_2O hay 10 electrones y, como se tienen 2 mol de moléculas de agua, habrá 20 mol de electrones.

$$\text{De } n = \frac{N}{N_A}, \text{ se sigue que } N = nN_A.$$

por lo que la ecuación de estado de los gases perfectos o ideales queda como:

$$pV = nN_AkT.$$

Pero el producto $N_A k$ no es más que la llamada *constante de los gases*, que se representa por R, que tiene un valor de 8.315 J/mol K. Y, con esto, finalmente obtenemos la ecuación de estado de los gases ideales:

$$pV = nRT.$$

El aire, por ejemplo, se comporta siguiendo aproximadamente la ley de los gases ideales. Esto quiere decir que, para valores de p, T y n como los del ambiente, el valor calculado de V no difiere en más de 5% del valor correcto, experimentalmente medido.

Volvemos a la cuestión de las protección contra las altas temperaturas. Para combatir el calor veraniego habrá que tomar acciones tanto en lo personal como en las viviendas. En lo personal, podemos modificar la vestimenta en grosor y color, o emplear algún dispositivo sencillo como un abanico o un sombrero. En cuanto a la vivienda, se podrá utilizar algún dispositivo técnico, funcionando con electricidad o con algún combustible fósil, capaz de bajar la temperatura de las habitaciones, o bien construyendo las casas y edificios de modo que los vientos y el bloqueo de la iluminación solar permitan bajar la temperatura.

Es notoria la reducción en temperatura que en un día soleado producen las nubes al interponerse entre el Sol y el suelo. Siguiendo esta idea, es posible reducir apreciablemente la temperatura dentro de los edificios si se orientan en direcciones apropiadas, o si entre ellos y el Sol se interponen árboles u otros objetos que proyecten sombra, es decir, que impidan que la radiación solar llegue a sus superficies. La radiación solar es energía luminosa que nos llega del Sol; se trata de radiación electromagnética que transporta energía.

Algunos pueblos, sometidos a regímenes de radiación solar elevados, emplean ropas de color blanco. La razón se puede encontrar realizando un experimento casero con tres latas de refresco vacías. Una se pinta de negro, otra de blanco y la tercera se deja con su color original. Se llenan las tres latas con agua y se exponen al Sol en un día despejado. Se podrá apreciar que, al cabo de unos minutos, el agua de mayor temperatura está en la lata negra, seguida de la que no se pintó, siendo el agua de menor temperatura la de la lata blanca.

Ocurre que los cuerpos negros absorben mejor la radiación solar que los pintados de un solo color, y aun mejor que los de color blanco. Pero los cuerpos negros no pueden absorber toda la radiación, pues si así sucediera, su energía interna y su temperatura aumentarían tanto que llegarían a fundirse; en lugar de esto, los cuerpos negros emiten la radiación con la misma facilidad con que la absorben.

Los cuerpos blancos, por el contrario, son pésimos absorbedores de la radiación solar, la cual reflejan en gran proporción. Y de aquí que las telas de color blanco sean usadas en los lugares de alta irradiación solar.

Debido a que los colores oscuros absorben casi toda la radiación que les llega, las superficies de los calentadores solares se pintan de negro. Si debajo de la capa de pintura se pone un material de alta conductividad térmica, por ejemplo, un tubo de cobre que contiene agua, la energía de la radiación pasará fácilmente del cobre hacia el agua, calentándola por conducción.

El uso de ventiladores manuales, como un abanico, logra bajar la "sensación de calor"; es decir, la sensación de una alta temperatura, al acelerar el proceso de evaporación de las gotas de sudor en la piel.

6.2 ¿CÓMO AHORRAR ENERGÉTICOS EN EL HOGAR?

Ya se vio que para evitar el frío en invierno y el calor en verano se recurre a la ley cero de la termodinámica, así como a los conceptos de temperatura, ecuación de estado y conductividad térmica. Ahora, para explicar el ahorro de "energéticos" en las casas y la contaminación asociada, se verá que es imprescindible recurrir a la primera ley de la termodinámica y a los conceptos de trabajo, calor y capacidades térmicas, entre otros.

La palabra energético denota a un objeto que tiene la potencialidad de desarrollar trabajo o generar calor al interactuar con otros sistemas u objetos, debido a que está en desequilibrio con ellos. Ejemplos: *i*] un objeto a cierta altura del suelo, en *desequilibrio gravitatorio* con el suelo, puede clavar un clavo al caer o generar electricidad para cargar una batería; *ii*] un litro de gas butano o un leño, al quemarse en la atmósfera gracias al *desequilibrio químico* con ella, puede calentar el agua en una olla; *iii*] una corriente eléctrica, generada por un *desequilibrio de potencial eléctrico*, puede accionar un motor.

Para tratar este aspecto de la economía familiar y del ambiente, desde el punto de vista de la física, tenemos que:

1] El gas se quema para: la cocción de alimentos; el calentamiento de agua para el aseo personal; el calentamiento de las casas durante el invierno.
2] La leña tiene los mismos usos que el gas, además de la iluminación.
3] La electricidad se utiliza para: la cocción de alimentos, mediante parrillas eléctricas u horno de microondas; la iluminación de espacios interiores y exteriores; la refrigeración de comestibles y bebidas; el calentamiento de las casas y edificios; el enfriamiento de las casas y edificios; el bombeo de agua de la cisterna al tinaco; la televisión, los videorreproductores, las computadoras personales y el radio; la lavadora, la lavadora de trastes, la secadora, la aspiradora, el secador de pelo; la licuadora y extractor de jugos; el tostador de pan; el cobertor eléctrico; la rasuradora; la regadera eléctrica para baño, etcétera.

El consumo de los tres energéticos —gas, leña y electricidad— es el más común, tanto en las zonas urbanas como en las rurales de nuestro país; sin embargo, en el campo la leña es el energético principal, en tanto que en las zonas urbanas lo son el gas y la electricidad.

6.2.1 Conservación de energía

El consumo de gas en un hogar común urbano es de aproximadamente 20% para la cocción de alimentos y 80% para el calentamiento de agua. Pero, para calcular con exactitud la proporción del gasto de gas destinado a la cocción de alimentos se deben considerar varios factores: la dieta, la cantidad de habitantes en el hogar, los hábitos alimenticios, el tipo de estufa, etc. Los cálculos del ahorro se basarán, por lo tanto, en consideraciones generales.

Aunque los alimentos se pueden cocinar de diferentes formas —friéndolos en una sartén con aceite, poniéndolos en una olla con agua a hervir, entre otras—, aquí se considerará solamente el segundo caso, por ser el de mayor gasto de gas.

Cocer un alimento significa elevar su temperatura desde la del ambiente hasta la del punto de ebullición del agua (la que, como vimos, es de 100 °C a la altura del nivel del mar, pero disminuye con la altura del lugar). En este proceso se tienen las siguientes transferencias de energía: por un lado, la flama del gas transfiere energía por calor al recipiente de cocción, el agua recibe esa energía y empieza a hervir. Cuando hierve, el agua ya no aumenta

su temperatura ya que comienza a evaporarse, pero al mismo tiempo continúa pasando energía por calor al alimento crudo, hasta que lo cuece.

Si una cantidad de agua, de masa *m*, estaba al inicio del calentamiento a la temperatura t_0 y ésta se eleva a *t*, tenemos un incremento, denotado por Δt, de modo que,

$$\Delta t = t - t_0.$$

Este incremento de temperatura es ocasionado por un aumento ΔU en la energía interna del agua, dado por:

$$\Delta U = Q.$$

En esta ecuación, *Q* es la energía que por calor (quemando gas) pasa de la flama a la masa *m* de agua, hasta que su temperatura se eleva al valor *t*; es decir, es la energía transferida debido a la diferencia de temperatura entre la alta temperatura de la flama y la menor temperatura del agua. Esta ecuación se basa en el *principio de conservación de la energía*, que establece que la energía que recibe la masa *m* de agua por calor de la flama se invierte en aumentar su energía interna. La energía pasa de la flama al agua, de modo que se conserva. La cantidad de calor *Q* puede medirse por el efecto que produce en el agua, al elevar su temperatura en:

$$\Delta t = t - t_0.$$

Si para una masa *m* fija de agua, la cantidad de calor Q se duplica, entonces el incremento de energía interna también se duplica, ocurriendo lo mismo con el cambio de temperatura de la sustancia que se está considerando. Es decir, Q y Δ*t* son directamente proporcionales:

$$\Delta t = \propto Q, \text{ con } m \text{ fija.}$$

Pero ahora hay algo interesante: si la misma cantidad de calor Q hubiese aumentado la energía interna de una mayor cantidad de agua, el aumento de temperatura provocado hubiese sido menor. Esto se puede comprobar experimentalmente calentando en la flama medio litro de agua durante tres minutos y enseguida un litro de agua durante el mismo tiempo.

Es decir, para una sustancia determinada el cambio de temperatura, ante un incremento en su energía interna por calor, es inversamente proporcional a la cantidad de sustancia. Si la cantidad de sustancia aumenta en cierta proporción, el aumento de temperatura es inversamente proporcional.

6.2.2 Capacidad térmica

Uno se puede imaginar que el cambio de temperatura de una alberca que se calienta con la misma flama, por tres minutos, es prácticamente de 0 °C, pero una pequeña cantidad de agua hasta puede hervir.

Entonces, para una cantidad de calor Q fija,

$$a = \frac{v^2}{R}.$$

La propiedad que expresa el aumento de temperatura de una cierta cantidad m de sustancia, al incrementar su energía interna por una cantidad de calor Q, se llama *capacidad térmica* y se representa por la letra C. La capacidad térmica es proporcional a la cantidad de sustancia, que medimos por su masa m, es decir:

$$C \propto m.$$

Si Q es fija,

$$\Delta t \propto \frac{1}{C},$$

de modo que:

$$\Delta t = \frac{Q}{C}.$$

El aumento de temperatura de una sustancia por calor se sintetiza en la fórmula anterior, porque si una cantidad fija de sustancia se energiza por calor, o sea a C fija, el aumento de temperatura es proporcional a la cantidad de calor Q ($\Delta t \propto$ Q). Es el caso de someter medio litro de agua a la acción de la flama por tres minutos y luego por seis.

Si se energiza la sustancia por una cantidad fija de calor Q, el cambio de temperatura es inversamente proporcional a su masa m ($\Delta t \propto$ 1/C). Es el caso en que medio litro de agua se pone sobre la flama por tres minutos, y luego un litro de agua se somete a la flama el mismo tiempo.

Si se despeja C, quedará

$$C = \frac{Q}{\Delta t}.$$

Las unidades de C son las de Q (joules, ya que el calor es energía), divididas entre las de Δt, que son grados Celsius; es decir:

$$[C] = \frac{\text{J}}{^\circ\text{C}}.$$

Ahora bien, si se aumenta la energía interna por calor de un kilogramo de agua y uno de mercurio, con la misma cantidad de calor Q, ¿qué sustancia aumentará más su temperatura? En el experimento, tanto las masas como Q son los mismos, pero la temperatura a que llega el agua es mucho menor que la del mercurio. Por lo tanto, el cambio de temperatura depende de la naturaleza química de la sustancia.

Entonces, para comparar adecuadamente la respuesta térmica de una sustancia con otra ante una energización por calor, definimos una nueva capacidad térmica, que sea para la misma cantidad de masa:

$$c = \frac{C}{m}.$$

A c se le llama *capacidad térmica específica*, queriendo decir con *específico* que mide la respuesta térmica de una masa fija unitaria (un kilogramo) de dicha sustancia. La c depende, así, sólo de la naturaleza de la sustancia.

Las unidades de *c* son las mismas que las de *C*, pero divididas por la unidad de la masa que es 1 kg; es decir,

$$[c] = \frac{\text{J}}{\text{kg}^{\circ}\text{C}}.$$

La capacidad térmica específica del agua es una de las más grandes; debido a esto, el aumento de su temperatura es menor que el de otras sustancias, cuando a masas iguales se les transfiere la misma cantidad de energía por calor. En la tabla siguiente se ofrecen datos comparativos de *c*:

Sustancia	$c\left(\frac{\text{cal}}{\text{g}\,^{\circ}\text{C}}\right)$
Agua	1.00
Hielo	0.55
Vapor de agua	0.50
Aluminio	0.22
Vidrio	0.20
Hierro	0.11
Latón	0.094
Cobre	0.093
Plata	0.056
Mercurio	0.033
Plomo	0.031

Las unidades en que está expresada *c* en la tabla anterior difieren de las empleadas en el SI; ello obedece a razones históricas que se explican a continuación.

La sustancia que se tomó como patrón para medir la cantidad de calor fue el agua, observando su incremento de temperatura ante una energización por calor conocida. El experimento es parecido al que se realiza en los laboratorios de las escuelas: en un vaso de material aislante se vierte una cierta cantidad de agua, a la temperatura de 14.5 °C. Luego se sumerge en el agua un alambre por el que pasa una corriente, hasta que la temperatura del agua suba un grado Celsius (véase figura 11, p. 430).

El alambre se calienta con el paso de la corriente, subiendo su temperatura por arriba de la del agua. Esto ocasiona que el agua se energice por calor y, como consecuencia, eleve su temperatura.

Se dice que la cantidad de calor con que se energiza el agua es de *1 caloría*, si la temperatura de 1 gramo de agua sube de 14.5 °C a 15.5 °C.

Para una masa de 1 gramo, se tendrá que la cantidad de calor involucrada es de:

$$1 \text{ caloría} = 1 \text{ cal} = c\Delta t = c\,(15.5^{\circ}\text{C} - 14.5^{\circ}\text{C}) = c.$$

El valor de 1 *cal*, que coincide numéricamente con la capacidad térmica específica *c* del agua a 14.5 °C, como se vio en la ecuación anterior, se calcula a partir de la energía eléctrica que circuló por el alambre.

Esta energía se mide en las unidades del SI y es de 4.185 joules, por lo que:

$$1 \text{ cal} = 4.185 \text{ J}.$$

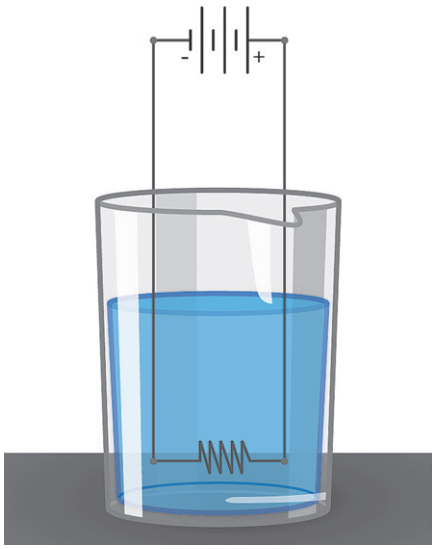


Figura 11. Experimento para determinar la caloría.

La cantidad de calor Q que calienta una masa *m* de agua, de 14.5 °C a 15.5 °C, será:

$$Q = mc\Delta t = C\Delta t = C \text{ J}.$$

En calorías, el calor necesario para energizar una masa m' cualquiera de otra sustancia diferente del agua y elevar su temperatura en Δt, será de:

$$Q' = m'c'\Delta t = C'\Delta t.$$

En la expresión anterior, las cantidades relativas a la otra sustancia se marcan con una prima. En el caso del experimento con un kilogramo de agua y uno de mercurio (es decir, m = 1 kg), el aumento de temperatura ante la energización por la misma cantidad de calor Q será:

$$\Delta t = \frac{Q}{c}, \text{ para el agua, y}$$

$$\Delta t' = \frac{Q}{c'}, \text{ para el mercurio.}$$

Como $c > c'$, se sigue que $\Delta t < \Delta t'$; es decir, el mercurio "se calienta" más. La tabla anterior (p. 142) de *c* para diferentes sustancias muestra que:

$$\frac{\Delta t'}{\Delta t} = \frac{Q/c'}{Q/c} = \frac{c}{c'} = \frac{1}{0.033} = 30.3,$$

es decir, la masa de 1 kg de mercurio aumenta más de treinta veces su temperatura que la masa de 1 kg de agua, ante la misma cantidad de calor Q. Se dice, por semejanza con la masa mecánica, que el agua tiene mayor "inercia térmica" que el mercurio (y que muchas otras sustancias).

La relación cuantitativa entre las unidades en que se puede medir *c* se deduce de las siguientes relaciones:

$$\frac{1 \text{ cal}}{\text{g°C}} = \frac{4.185 \text{ J}}{10^{-3}\text{kg°C}} = 4185\frac{\text{J}}{\text{kg°C}}.$$

Empleando esta conversión es posible reescribir la tabla anterior en *joule/kg°C*, simplemente multiplicando los valores por 1 000.

Las sustancias son de dos clases: elementos y compuestos. Los elementos no pueden descomponerse en sustancias químicas más simples por métodos ordinarios de la química, pero sí un compuesto. En general, los elementos están formados por unidades estructurales microscópicas llamadas *átomos*, mientras que los compuestos están formados por unidades estructurales llamadas *moléculas* (que son agregados de átomos).

La *cantidad de sustancia* se determina especificando el número de unidades estructurales microscópicas; por ejemplo, en una muestra de hierro se puede especificar que hay 3.0×10^{24} átomos de Fe. En otro ejemplo, la cantidad de sustancia en un cristal de sal común, NaCl, contiene 7×10^{21} pares de iones Na^+ Cl^-.

En el SI, la unidad de cantidad de sustancia es el *mol*. El mol es la cantidad de átomos que hay en 12 gramos del isótopo carbono 12, que es igual a una cantidad constante llamada *número de Avogadro*, $N_A = 6.23 \times 10^{23}$.

Un mol de moléculas de agua, H_2O, también tiene N_A moléculas, al igual que un mol de cualquier otra sustancia. Si N fuera la cantidad de moléculas de agua en un recipiente, la cantidad de moles, representada por la letra *n*, sería de $n = N/N_A$. Entonces, la *capacidad térmica molar*, denotada por c^*, es:

$$c^* = \frac{C}{n}.$$

La respuesta térmica molar comparativa de una sustancia se refiere al cambio de temperatura cuando se energiza, por una cantidad fija de calor, una misma cantidad de átomos o moléculas. La respuesta térmica dependerá, entonces, de la forma de cómo esos N_A elementos estén estructurados microscópicamente en 1 mol.

En la tabla siguiente se muestra la capacidad térmica molar de algunos gases, agrupándolos en monoatómicos, diatómicos y poliatómicos. La respuesta térmica de 1 mol de cada tipo de gas es muy parecida cuantitativamente.

Gas	caloría / mol °C
Monoatómico	
He	20.8
Ar	20.8
Diatómico	
H_2	28.8
N_2	29.1
O_2	29.4
Poliatómico	
CO_2	37.0
NH_3	36.8

Al energizar una porción de sustancia por calor, generalmente aumenta su temperatura; un efecto adicional es que también puede aumentar su volumen. Si se vierte agua en un recipiente y se calienta, su volumen aumenta al recibir energía por calor de la flama; la expansión del agua se efectúa en contra del aire.

Supóngase que la superficie del agua se comporta como un pistón sin peso, de sección circular, que empuja el aire hacia arriba. El diagrama de la figura adjunta muestra que el agua vence en su dilatación la fuerza externa del aire. La fuerza total que el agua ejerce contra el aire es el producto de su presión *p*, por el área *A* de la superficie de contacto con el pistón:

$$F = pA.$$

Si el pistón se mueve hacia arriba una distancia $\Delta h = h_f - h_i$, siendo h_i la altura inicial y h_f la altura final de la superficie del pistón imaginario, el agua habrá realizado un trabajo, W, contra la atmósfera dado por:

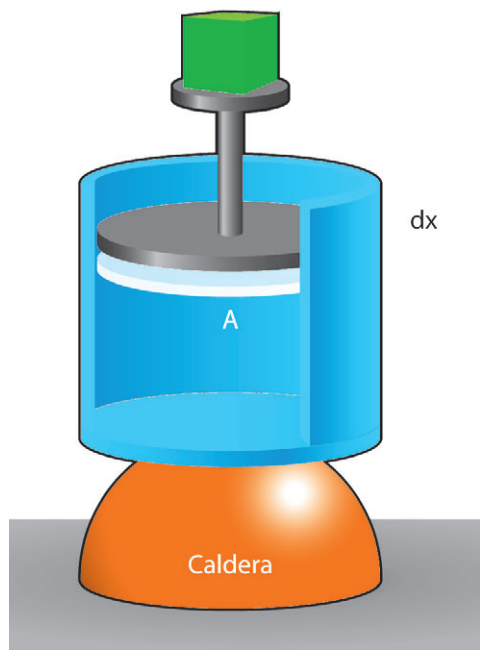


Figura 12. Trabajo por presión.

$$W = F\Delta h = pA\Delta h = p\Delta V.$$

El producto $A\Delta h$ no es más que el volumen de dilatación del agua, ΔV.

6.2.3 Primera ley de la termodinámica

Se tienen dos procesos que afectan la energía interna del agua, que la hacen cambiar en ΔU: por un lado, el calor $Q = C\Delta t$ por la interacción con la flama que la aumenta y, por otro, el trabajo $W = p\Delta V$ que el agua efectúa contra la atmósfera al dilatarse, que la disminuye. Ambos efectos se toman en consideración en la siguiente ecuación:

$$\Delta U = Q - W.$$

Esta ecuación, que gobierna los intercambios de energía de un objeto por calor y por trabajo, es la primera ley de la termodinámica. En realidad, la ecuación anterior es válida para cualquier otro objeto, sea sólido, líquido o gaseoso, de cualquier naturaleza, siempre y cuando sea un objeto mesoscópico. En particular, la ecuación se aplica para un fluido descrito por sus variables de volumen, presión y temperatura.

Puede ser, por ejemplo, vapor de agua o aire encerrado en un cilindro provisto de un pistón. En este caso, la presión interna del gas debe ser suficiente para generar una fuerza capaz de contrarrestar el peso del pistón y del aire por encima de él. El trabajo de expansión será nuevamente calculado por $p\Delta V$.

El signo negativo de W corresponde a la convención que asocia un signo positivo al trabajo cuando el sistema lo realiza sobre el exterior, y uno negativo en caso contrario; es decir, cuando un agente externo efectúa trabajo sobre el sistema.

En este caso, el agua al expandirse hace trabajo contra la atmósfera, por lo que W es positivo y, al restarse en la ecuación, significa que dicho trabajo se efectúa a expensas de su energía interna. Asimismo, Q se considera negativo si el sistema pierde energía interna por calor con otro objeto, y positivo si el sistema gana energía por calor de otro cuerpo.

Desde el punto de vista microscópico clásico, la energía interna U de todo sistema termodinámico, en particular del agua, se compone de términos asociados al movimiento de sus componentes moleculares, aunque también de los relacionados con la energía potencial de unas moléculas respecto a las otras.

Entonces, cuando U cambia, por calor o por trabajo, varían tanto la energía de movimiento de las moléculas (influyendo en el cambio de temperatura) como el valor promedio de la posición de ellas (cambio asociado, a su vez, con la variación del volumen del objeto).

En un gas monoatómico (formado por moléculas de un solo átomo), U se compone de la suma de las energías cinéticas $1/2\ m\nu_i^2$, en donde m es la masa de la molécula y ν_i es la rapidez con que se mueve. En el agua, U consta de la energía cinética de las moléculas de H_2O, sumada a la energía potencial asociada a las fuerzas intermoleculares.

Cuando un campo electromagnético de microondas hace vibrar a las moléculas de agua, es necesario sumar a U la energía asociada al movimiento vibratorio de las moléculas. El agua en una olla sobre una flama se calienta por transferencia de energía por calor.

Al hervir el agua, la temperatura de ebullición permanece constante, ya que la energía que la flama le proporciona por calor se invierte en romper los "amarres" de las moléculas del líquido. Las moléculas de agua escapan en forma de vapor, efectuando también trabajo expansivo en contra de la atmósfera; mientras quede líquido por evaporar, su temperatura se mantiene constante.

Ocurre también que la temperatura de ebullición depende de la altura sobre el nivel del mar. La presión del aire por encima de la olla impide el escape de las moléculas; como la capa de aire es menor en la ciudad de México que en Acapulco, el agua hervirá a una temperatura menor en la capital (94 °C) que en la costa (100 °C).

Si ahora se calienta agua en una olla cerrada (olla de presión), la presión aumenta dentro de ella, de modo que el agua alcanza temperaturas mayores a la de ebullición al aire libre. Si se logra mantener la temperatura alta, a fuego lento, dejando escapar un mínimo de vapor para que el recipiente no explote, se puede cocer un alimento en menos tiempo y empleando menos combustible que cocinando en una olla abierta.

6.2.4 Ahorro de gas

El gas se puede ahorrar de varias maneras al cocinar; son cada vez menos las estufas de gas que utilizan "piloto"; se estima que cerca de 10% del gas para cocinar se ahorra apagando esa pequeña flama. Pero este gasto inútil subsiste en los calentadores de gas que calientan agua, a menos que se adquiera la costumbre de apagar el piloto o sustituir el calentador de gas por uno de encendido electrónico o uno solar.

La cocción que consume más gas es el agua hirviendo. Pero la preparación de frijoles "de la olla" es, sin duda, la que toma mayor tiempo y consumo de gas. Una opción de ahorro consiste en el empleo de una olla de presión. En este caso el ahorro se puede calcular aproximadamente, si se toma en cuenta que los frijoles se cuecen en olla abierta en un lapso de 2.5 a 3 horas, mientras que en una olla a presión el tiempo de cocinado es entre 30 y 40 minutos solamente.

Supóngase los valores inferiores de tiempo de cocción de 2.5 horas y 30 minutos para la olla abierta y la olla cerrada, respectivamente. El porcentaje del ahorro de gas se calcula como sigue.

Sea R la rapidez con que se quema el gas, es decir, $R\left(\frac{g}{\min}\right)$. Su valor típico es de unos cuantos gramos de gas por minuto.

El gas quemado en 2.5 horas, a olla abierta, es:

$$G_1 = 2.5 \times 60\,\min \times R\frac{\mathrm{g}}{\min} = 150\,\min \times R\frac{\mathrm{g}}{\min} = 150 \times R \times g.$$

El gas quemado en la olla a presión vale:

$$G_2 = 30\,\min \times R\frac{\mathrm{g}}{\min} = 30 \times R \times g.$$

Así que el ahorro de gas es:

$$\Delta G = G_1 - G_2 = (150 - 30) \times R \times g = 120 \times R \times g.$$

La cantidad exacta en g no se puede saber, a menos que se conozca la rapidez de quemado, R. Pero se puede calcular el porcentaje de ahorro:

$$\frac{\Delta G}{G_1} \times 100\% = \frac{100 \times R \times g}{150 \times R \times g} \times 100\% = 66\%.$$

(Hay que recordar que el porcentaje se calcula de la siguiente regla de tres:

$$\frac{\Delta G}{G_1} = \frac{X}{100}.$$

De aquí se obtiene el porcentaje desconocido *X*.)

El ahorro es aún mayor porque la rapidez de quemado de una olla a otra difiere: es menor en la olla a presión que en la ordinaria, o sea, $R_2 < R_1$. Como antes, el subíndice 1 se refiere a la olla abierta y el 2 a la cerrada.

Pero hay otras formas para ahorrar todavía más gas. Para esto, es necesario mostrar las interacciones energéticas por calor entre la flama de la estufa, la olla y el aire de la cocina. El calor q'_{evap} es el que se logra reducir drásticamente en la olla a presión.

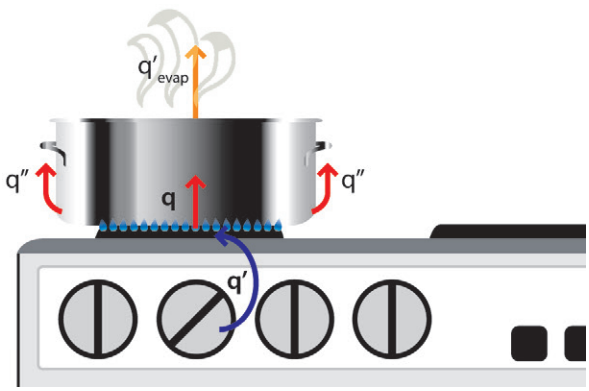


Figura 13. Esquema de procesamiento de calores por unidad de masa.

En la figura 13, *q* es la energía por calor suministrada a la olla.El calor *q*" es la energía que se pierde hacia el aire circundante, mientras que *q*' se traduce en la energización efectiva del agua en la olla. En una estufa bien diseñada se minimiza *q* y se maximiza *q*'. Una medida cuantitativa del buen diseño de una estufa se obtiene del cociente de las dos cantidades anteriores, que se denomina *eficiencia térmica* y se denota por la letra *e*:

$$e = \frac{q'}{q}.$$

Expresada en forma porcentual, es:

$$e = \frac{q'}{q} \times 100\%.$$

Aunque varias de las características físicas del quemado son desconocidas, pero pueden medirse, se pueden ofrecer recomendaciones para el ahorro del gas en la cocción de alimentos por ebullición de agua:

1] Sustituir la olla abierta por una olla de presión. En este tipo de olla q'_{evap}/s, o potencia de evaporación, es casi 0. De esta manera se puede observar que el ahorro de gas es grande.
2] Aumentar la eficiencia *e* de la estufa por ajuste de la rapidez de quemado, o por modificación de las características geométricas del quemador (cerrar más el quemado, a modo de disminuir el escape de energía lateral).

El calentamiento de agua para baño es responsable de hasta 80% del consumo de gas doméstico. El proceso de quemado del gas es parecido al de la ebullición del agua en la cocción de algunos alimentos, excepto porque el proceso ocurre en el calentador de gas y la cantidad de agua es mucho mayor. El calentamiento de agua para diversos usos es grande, pero el calentador solar puede sustituir ventajosamente al calentador de gas.

1] Características económicas del calentador solar: alto costo relativo de inversión (pero con tendencia a la baja, al aumentar el número de usuarios); nulo costo del "energético" empleado.

2] Características económicas del calentador de gas: bajo costo relativo de inversión; alto costo del "energético" empleado (con tendencia al alza, al disminuir las reservas).

Lo anterior quiere decir que, al cabo de cierto tiempo, el uso del calentador solar permite la recuperación de la inversión inicial. Esto se puede ver con el siguiente cálculo simplificado:

Supóngase una familia de cinco personas que gastan unos 6 000 pesos de gas al año, de los cuales 80% es para calentamiento de agua. El gasto anual por calentamiento de agua = ($6 000 / año) × (0.8) = $4 800 / año, es decir, el costo diario del calentamiento de agua por gas es de unos 13.5 pesos.

Las necesidades de calentamiento de agua de la familia se pueden satisfacer con un calentador solar de 200 litros, que cuesta 15 000 pesos. En 3.12 años [= $15 000 / ($4 800/ año)], el dinero gastado en gas se iguala con el costo del calentador solar y su instalación, por lo que si la vida útil del calentador solar es de 15 años, como afirma el fabricante, el consumidor se ahorrará en pago de gas:

$$(11.88 \text{ años})(\$4\,800/\text{año}) = \$57\,024.$$

A esta cantidad hay que agregar el ahorro del costo de inversión del calentador de gas, que puede ser de 1 000 pesos, dando un total para el ahorro de 58 024 pesos. Una cantidad digna de consideración (al momento de escribir esto, el salario mínimo es cercano a los 1 500 pesos mensuales). Hay que tomar también en cuenta que el precio del gas sube, simplemente por tratarse de un recurso no renovable, de manera que la ventaja económica del calentador solar es aún mayor.

Calentadores solares

Desde el punto de vista ambiental, el calentador solar no contribuye a la contaminación del aire, durante al menos quince años. Este ahorro de contaminación tampoco es despreciа-

Figura 14. Calentador solar | © Latin Stock México.

ble, sobre todo si se toma en cuenta que la quema de gas en los hogares es la tercera fuente de contaminación atmosférica en las grandes ciudades, después del transporte y la industria. Para entender cualitativamente el funcionamiento de un calentador solar, es necesario presentar brevemente las distintas formas en que se procesa calor.

Como se mencionó en otro apartado, el intercambio energético entre los sistemas a distintas temperaturas se efectúa a través de la pared diatérmica, sin que se deforme apreciablemente ni haya transferencia de sustancia. Así, dos cuerpos que están a diferente temperatura pueden cambiar sus energías internas por calor de conducción, y si se conectan mediante una barra de cobre, ésta permanece inalterada durante el proceso.

Hay diferentes diseños de calentadores solares. El de la figura 14 (p. 435) permite tener agua caliente día y noche, a temperaturas superiores a los 70 °C y un poco menores en días nublados. El calentador solar funciona de la siguiente manera: el agua fría del tinaco desciende a una serie de tubos paralelos. Éstos consisten de dos cilindros concéntricos; el exterior es de vidrio y el interior de metal pintado de negro. Entre ambos cilindros hay vacío, para evitar la pérdida de energía por la diferencia de temperatura entre el tubo y el aire de la atmósfera. El tubo interior, pintado de negro, absorbe eficientemente la radiación solar directa y difusa y se calienta, transmitiendo energía por calor al agua contenida en su interior. El agua caliente, por ser menos densa, sube a un recipiente cilíndrico horizontal aislado térmicamente. Se construye así un ciclo de circulación de agua por convección, en el que agua fría baja del tinaco y el agua caliente sube al cilindro aislado, almacenándose ahí para su uso posterior.

Uso de calentadores de gas

"Calentar un cuarto" significa aumentar la temperatura del aire del cuarto en $\Delta t = t - t_0$, siendo t una temperatura confortable, generalmente superior a 21 °C, y t_0 la temperatura inicial 0 °C. Con estos valores,

$$\Delta t = t - t_0 = (21 - 0)^\circ\mathrm{C} = 21^\circ\mathrm{C}.$$

La energía del quemador de gas que se debe transferir por calor al aire del cuarto está dada por:

$$Q = C_A \Delta t(\mathrm{J}/^\circ C)(^\circ C) = C_A \Delta t(\mathrm{J}).$$

donde C_A es la capacidad térmica total de la masa m_A de aire del cuarto, o del número total de moles n_A. C_A es el producto de la cantidad n_A de moles en el aire por su capacidad térmica específica molar, c^*_A, $C_A = n_A c^*_A$. Si el volumen del aire no cambia, entonces c^*_A es la capacidad térmica molar del aire a volumen constante. Así que:

$$Q = n_A c^*_A \Delta t.$$

De esta forma, cuanto más grande es la cantidad de aire, mayor es Q y, por consiguiente, mayor es la cantidad de gas que habrá que quemar para subir la temperatura del aire de t_0 a t. Para conocer la cantidad de gas m_G quemada, hay que tomar en cuenta que el gas tiene un "calor de combustión por kilogramo" $q_G = 55 \times 10^6$ (Joules/kg). Esto quiere decir que quemar un kilogramo de gas genera una cantidad de energía por calor a la atmósfera de $q_G = 55 \times 10^6$ joules. Entonces:

$$Q_G = m_G(\text{kg})q_G(\text{joule/kg}) = m_G q_G(\text{kg joule/kg}) = m_G q_G(\text{joule}).$$

Por conservación de energía:

$$Q = Q_G,$$

o sea que:

$$Q = n_A c_A^* \Delta t = m_G q_G.$$

De esta expresión se puede despejar m_G:

$$m_G = n_A c_A^* \frac{\Delta t}{q_G}.$$

Solamente faltaría conocer la cantidad de los mol de aire en el cuarto para saber la cantidad de gas que hay que quemar para subir su temperatura hasta 21 °C. Si el aire se comporta como gas ideal, y se considera un cuarto de 3 m de altura, 4 m de ancho y 4 m de largo, su volumen será de 48 000 litros. Puesto que un mol de gas ideal ocupa 22.4 litros a 0 °C, a la presión de 1 atmósfera, el número de mol es:

$$n_A = \frac{48000 \text{ litros}}{22.4 \text{ litros/mol}} = 2.1 \times 10^3 \text{mol}.$$

Entonces,

$$m_G = \frac{(2.1 \times 10^3 \text{ moles})(20.8 \text{ J/(mol°C)})(21°\text{C})}{55 \times 10^6 \text{ J/kg}}$$
$$= 16.7 \times 10^{-3} \text{ kg} = 16.7 \text{ g}.$$

Esta cantidad de gas m_G se necesita quemar inicialmente para elevar la temperatura del aire del cuarto de 0 °C a 21 °C; pero si el gas deja de quemarse la temperatura tenderá a bajar de nuevo. Supongamos, para propósitos de ilustración, que en una hora la temperatura *t* bajaría de nuevo a la temperatura t_0.

Esto es, que tendríamos que estar quemando una masa de gas m_G cada hora, para vivir cómodamente en nuestra habitación. Si se usa el calentador 4 horas al día, entonces la masa total de gas M_G quemada por día será:

$$M_G = 4(\text{horas})m_G(\text{kg/hora}) = 4 \times 16.7 \times 10^{-3} \text{ kg} = 66.8 \times 10^{-3} \text{ kg}.$$

En un mes: $30 \times M_G = 30 \times 66.8 \text{ x}10^{-3}$ kg = 2,004 × 10^{-3} kg = 2 kg. Si se calientan tres habitaciones de las mismas dimensiones que la anterior:

Gas quemado al mes = 6 kg.

Costo del gas quemado por mes = 6 kg × (costo del kilogramo de gas)

= 6 kg × 280 pesos/30 kg = 56 pesos.

Esta cantidad de gas y dinero se puede ahorrar al mes si se aíslan las paredes con lambrín y se utiliza ropa adecuada en el interior de la casa, en vez del calentador de gas.

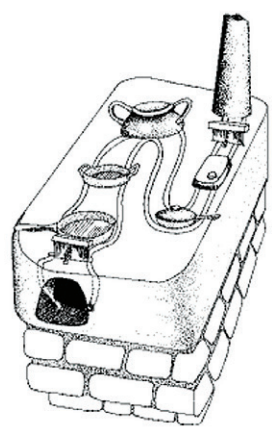
Uso de leña

La leña se utiliza como combustible sobre todo en las zonas rurales, en donde habitan cerca de 27.5 millones de personas en nuestro país (año 2000). Se estima que 89% de esa población emplea la leña como principal combustible para la cocción de alimentos, mientras que los usuarios de leña en zonas urbanas representan el 11 por ciento.

Su uso principal es para cocinar alimentos, calentamiento de agua y, en menor proporción, para iluminación. En algunas comunidades rurales la elaboración de tortillas es la tarea que requiere el mayor consumo de leña, con poco más de 42%, según un estudio realizado por la Facultad de Ciencias de la UNAM. El consumo per cápita es de 2 kg al día, y más de 80% de los hogares utiliza el tradicional fogón de *tres piedras*, es decir, a fuego abierto por los lados.

El quemado de leña a fuego abierto tiene inconvenientes: la eficiencia termodinámica es baja y ocasiona enfermedades respiratorias por la aspiración de los gases de la combustión. Varios estudios muestran que la contaminación dentro del hogar, debida al quemado de leña es incluso mayor que en las grandes ciudades; ocasiona mundialmente alrededor de 1.2 millones de muertes prematuras anuales, de niños menores de 5 años.

Figura 15. Estufa Lorena.



Una opción técnica alternativa al fogón de tres piedras son las estufas de lodo y arena (llamadas "lorenas", nombre compuesto por ambas palabras).

Al cerrar el compartimiento donde ocurre la combustión de la leña y poner el utensilio encima del fuego, evitando fugas laterales, se aumenta la cantidad de calor que energiza directamente el alimento. La chimenea adicionada expulsa los gases residuales a la atmósfera, lo que evita daños a la salud y contribuye a un aumento en la eficiencia, al regular la entrada de aire en el proceso de combustión.

Las pruebas de campo efectuadas por un grupo de investigación de la Facultad de Ciencias de la UNAM muestran que, en promedio, se ahorra 40% de la leña, siendo posible, en algunos casos, ahorros de hasta 75%. En lugares desarbolados estos ahorros significan una liberación del tiempo de trabajo de recolección, así como una disminución de la desertificación.

Ahorro por la reducción del consumo de electricidad

En este apartado se analizan tres usos de la electricidad: iluminación, calefacción y calentamiento de agua para el aseo personal.

En la actualidad hay dos tipos de focos: los tradicionales y los "ahorradores", éstos últimos conocidos así porque con el mismo consumo de energía eléctrica generan más iluminación. Un foco ahorrador de 20 watts genera tanta iluminación como un foco tradicional de 100 watts. Aunque los focos ahorradores son más caros que los tradicionales, en corto tiempo el ahorro de electricidad compensa su costo.

Lo anterior quiere decir que, en una hora, el foco ahorrador disipa una energía de 20 W × 1 hora = 20 J/s × 3 600 s = 72 000 joules, mientras que el foco tradicional disipa en el mismo tiempo una energía 100/20 = 5 veces superior, o sea 360 000 joules. En los cálculos aritméticos anteriores se ha utilizado:

$$\text{energía} = \text{potencia} \times \text{tiempo de operación.}$$

En los recibos de la luz el consumo de electricidad no viene expresado en joules, sino en kilowatt-hora, abreviada como kW h (recordar que k = 1 000). Ésta sería la energía que un aparato de una potencia de 1 000 watt disipa en una hora (o cualquier combinación del producto potencia y tiempo que equivalga a 1 000 watt × 3 600 s = 3.6 millones de joules; por ejemplo, 10 focos de 100 W de potencia cada uno, funcionando durante una hora), es decir:

$$1\ \text{kWh} = 1000\ \text{W} \times 3600\ \text{s} = 3.6 \times 10^{6}\ \text{J}.$$

o, alternativamente:

$$1\ \text{J} = 0.28 \times 10^{-6}\text{kWh}.$$

Entonces, el ahorro de electricidad mediante la sustitución de los focos tradicionales por los "ahorradores", en una casa donde se utilizan 6 focos, durante 4 horas al día, se calcula como sigue:

Energía eléctrica disipada = 6 focos × potencia de cada foco × 4 horas
= 24 × potencia de cada foco × hora.

Energía diaria disipada por 6 focos tradicionales = 24 × 100 W × h = 2.4 kWh.
Energía diaria disipada por 6 focos ahorradores = 24 × 20 W × h = 0.48 kWh.
Energía diaria ahorrada por 6 focos = 2.4 kWh – 0.48 kWh = 1.92 kWh.
Energía diaria ahorrada por cada foco = 0.32 kWh.
Pesos diarios ahorrados por cada foco = 0.51 pesos.

Dado que el costo del kWh, en la tarifa de alto consumo, es de 1.6 pesos (abril de 2006), el ahorro diario es de 3.1 pesos. En un bimestre, el ahorro total de electricidad para iluminación es de: 1.92 × 60 días = 115.2 kWh y el bimestral es de 184 pesos.

La diferencia de costo en el mercado entre un foco ahorrador y un foco tradicional de 100 W es de 30 pesos – 3 pesos = 27 pesos. O sea, en unos 53 días de operación (a razón de 4 horas por día) el ahorro por gasto de electricidad se iguala a la diferencia de costo. En este tiempo se recupera el costo del foco ahorrador.

Bomba de calor contra calentador de resistencia

En México son comunes los calentadores eléctricos de resistencia, los cuales convierten cada joule de electricidad en un joule de calor. Esto podría dar la impresión de que por ello son muy eficientes, pero se trata de una falsa impresión.

Por otro lado, las llamadas "bombas de calor", que son algo parecido a un refrigerador empotrado en la pared, por cada joule de electricidad gastado transfieren cinco joules de

Calentador eléctrico |
© Latin Stock México.

energía al cuarto; éstas bombean calor del aire exterior al aire interior de la habitación. Son, por ello, cinco veces más eficientes que los calentadores eléctricos comunes y corrientes.

Para calcular el ahorro al sustituir a los calentadores eléctricos por las bombas de calor, se considera que dos calentadores de 1 000 W de potencia cada uno se utilizan dos horas diariamente. La energía diaria disipada en el aire será:

Energía eléctrica consumida diariamente =
2 000 W × 2 h = 4 kWh.
Energía eléctrica consumida bimestralmente =
60 × 4 kWh = 240 kWh.
Costo de 4 kWh = 1.6 pesos/kWh × 4 kWh = 6.4 pesos.
Costo al bimestre = 60 × 6.4 pesos = 384 pesos.

Dado que la bomba de calor produce cinco veces más calor que el calentador eléctrico, por cada kWh de energía eléctrica consumida, se tiene que:

Energía eléctrica consumida al día por una bomba de calor = 4 kWh/5 = 0.8 kWh.
Energía eléctrica consumida al bimestre = 60 × 0.8 kWh = 48 kWh.
Energía ahorrada al día = 3.2 kWh.
Energía ahorrada en un bimestre = 60 × 3.2 kWh = 192 kWh.
Ahorro diario = 3.2 kWh × 1.6 pesos/kWh = 5.12 pesos.
Ahorro bimestral = 5.12 pesos × 60 días = 307 pesos.

Si la única opción existente son los calentadores eléctricos convencionales, entonces el uso tanto de ropas adecuadas como de paredes recubiertas con materiales aislantes daría un ahorro bimestral de electricidad de 307 pesos.

Calentamiento de agua para aseo personal y limpieza: calentamiento eléctrico contra calentamiento solar

Si se desconoce el gasto por calentamiento de agua con regaderas eléctricas, en el recibo de luz se puede calcular indirectamente, de manera aproximada, conociendo la cantidad de agua que se utiliza y el incremento de temperatura.

Si m es la masa diaria de agua que se calienta por persona, digamos de 15 °C a 40 °C (unos 40 litros); la energía empleada por calor será:

$$Q_1 = mc_p \Delta t = 40\ (\text{kg})\ 4182\ (\text{J})/(\text{kg °C})\ 25\ °\text{C} = 4\ 182\ 000\ \text{J} = 1.17\ \text{kWh}.$$

Para cinco personas:

$$Q_5 = 5.85\ \text{kWh}.$$

En el bimestre:

$$Q_{\text{aseo bimestral}} = 60 \times 5.85\ \text{kWh} = 351\ \text{kWh}.$$

Ahorro bimestral en pesos:

$$= 351 \times 1.6 \text{ pesos/kWh} = 562 \text{ pesos.}$$

Los ahorros calculados anteriormente, tanto en gas como en leña y electricidad, son apenas una muestra de la importancia que el análisis termodinámico tiene en ellos.

6.3 ¿CÓMO REDUCIR LA CONTAMINACIÓN PARA UN DESARROLLO SUSTENTABLE?

Los casos que se han analizado de ahorro de los energéticos —gas, leña y electricidad— en el hogar no sólo contribuyen a aligerar los costos de manutención de una casa, sino que también disminuyen la contaminación del ambiente.

Se han analizado dos clases de medidas para ahorrar dinero y disminuir la contaminación: aumentando la eficiencia de los dispositivos técnicos y sustituyendo los energéticos consumidos en casa. De todos modos, los problemas económicos y de contaminación ambiental son más amplios; participan en ellos otros sectores de la sociedad, como el transporte, la industria, la agricultura y la misma generación de electricidad, entre otros.

Se podría pensar que, con la introducción de focos ahorradores o bombas de calor, la contaminación se reduciría al mínimo, pero no es así, pues tanto los focos ahorradores como las bombas de calor funcionan con electricidad de la red y, en México, más de 65% de ella se genera en plantas *termoeléctricas* que queman combustibles fósiles (petróleo, gas y carbón). Mundialmente, se calcula que la electricidad y la calefacción de interiores produce 24.6 % de los gases de efecto invernadero (GEI).

En los dos casos mencionados la contaminación casi se podría eliminar si, además de introducir dispositivos más eficientes para iluminar y calentar, éstos operaran con electricidad no generada en plantas termoeléctricas, sino que utilizaran los energéticos provenientes del Sol.

Los energéticos solares son de dos tipos: indirectos y directos. Los primeros son las caídas de agua (mediante presas hidroeléctricas), el viento (con los aerogeneradores, como en La Ventosa en el istmo de Tehuantepec), el bagazo de caña y otros residuos agrícolas (a través de su combustión en termoeléctricas). En el caso del bagazo de caña, la contaminación neta de GEI se reduce apreciablemente porque los gases de la combustión son recapturados de la atmósfera, en el siguiente cultivo de la planta.

La conversión directa de energía solar en electricidad se lleva a cabo mediante el *efecto fotovoltaico*, o bien, concentrándola mediante espejos en una caldera; el vapor resultante mueve la turbina de un generador.

Un problema grave es que los combustibles fósiles no sólo se utilizan en la generación de electricidad, sino que también se queman en los motores de combustión interna del transporte, industria, agricultura, etc. Tan sólo el transporte es responsable de 13.5% de las emisiones de GEI en el mundo.

Lo interesante —y por eso constituyen una opción real— es que los energéticos solares pueden utilizarse en todos estos sectores y así emitir menos contaminantes a la atmósfera. Por esta razón es inevitable la transición de los energéticos agotables (petróleo, gas y carbón) a los energéticos inagotables (solares y geotérmicos), aspecto central del llamado *desarrollo sustentable*.

6.3.1 Generación de electricidad por combustibles fósiles

Una instalación generadora de electricidad (hidroeléctrica, termoeléctrica, aerogenerador o panel de celdas fotovoltaicas) tiene como característica principal su *capacidad instalada*, que en el SIU se especifica en watt, además del tipo de energético utilizado. La capacidad instalada de las actuales plantas generadoras convencionales de electricidad con gas, petróleo, uranio, carbón y las grandes presas, es muy grande; tanto, que se mide en miles de millones de watt. Mil millones de watt = 10^9 watt = 1 GW= y se lee "un giga watt".

Para tener una idea del significado de 1 GW se puede comparar con la capacidad de Chicoasén, nuestra mayor hidroeléctrica, que se encuentra en el estado de Chiapas; su capacidad instalada es de 1.2 GW. La capacidad de las termoeléctricas que usan derivados del petróleo puede ser de varios cientos de millones de watt (1 millón de watt = 1 MW = 0.001 GW, y se lee "un megawatt").

Por su gran capacidad y el tipo de energético empleado (10^9 watt = 1 GW), estas plantas generan mucha contaminación ambiental. Las grandes hidroeléctricas, que utilizan las caídas de agua —un recurso energético proveniente del Sol—, también tienen considerables efectos en el ambiente por la tierra que inundan, la población que desplazan y la producción de gases de efecto invernadero como el metano, que se produce por la vegetación que queda debajo de las aguas. Las cifras siguientes aportan una idea de la magnitud de la generación de electricidad. En el año 2003, la capacidad eléctrica instalada mundial era de unos 3 641.3 GW, distribuida de la siguiente manera: 2 469.9 GW (68%) de termoeléctricas de combustibles fósiles; 739.8 GW (20%) de hidroeléctricas; 368.5 GW de nucleoeléctricas (10%) y el resto (63.1 GW, 17.3%) repartido en aerogeneradores, celdas fotovoltaicas y geotermoeléctricas. En unos cuantos años, este espectro de capacidades ha cambiado a favor de las fuentes renovables de electricidad.

6.3.2. Motores térmicos

Las termoeléctricas modernas son descendientes lejanas de los primeros motores térmicos europeos del siglo XVII. Esos motores transformaban calor en trabajo, quemando carbón. La figura 16 muestra una representación del motor de Newcomen (1663-1729), inventor inglés, quien se dedicó a resolver el problema de cómo efectuar trabajo mecánico a partir de la combustión del carbón para aplicarlo en la extracción del mismo y del agua de las minas, así como para utilizarlo en las múltiples tareas de las fábricas de hilados, tejidos, etcétera.

El funcionamiento del motor de Newcomen es el siguiente: si las válvulas B y C están cerradas y se abre la válvula A, entrará vapor caliente de la caldera. Éste empuja al pistón hacia arriba, moviendo el balancín para subir una carga de agua o de carbón desde el fondo de la mina. Cuando el pistón se encuentra en la parte superior del cilindro, se cierra la válvula A y se abre la C, vertiendo agua fría sobre el cilindro y el pistón. Este enfriamiento condensa al vapor del interior del cilindro, disminuyendo su presión por debajo de la presión atmosférica, haciendo que el pistón se mueva hacia abajo. Se abre la válvula B y el vapor residual escapa al exterior. El ciclo comienza de nuevo al abrirse la válvula A, para dejar entrar vapor otra vez, estando las válvulas B y C cerradas, y así sucesivamente (p. 443).

Antiguamente la eficiencia del motor de Newcomen no pasaba de 1 a 2% por varias razones: el ajuste entre el pistón y las paredes del cilindro era tan deficiente que una moneda podía caber en el espacio intermedio. Pero fue Watt quien descubrió la causa principal de la ineficiencia. Él estudió un modelo a escala del motor de Newcomen y descu-

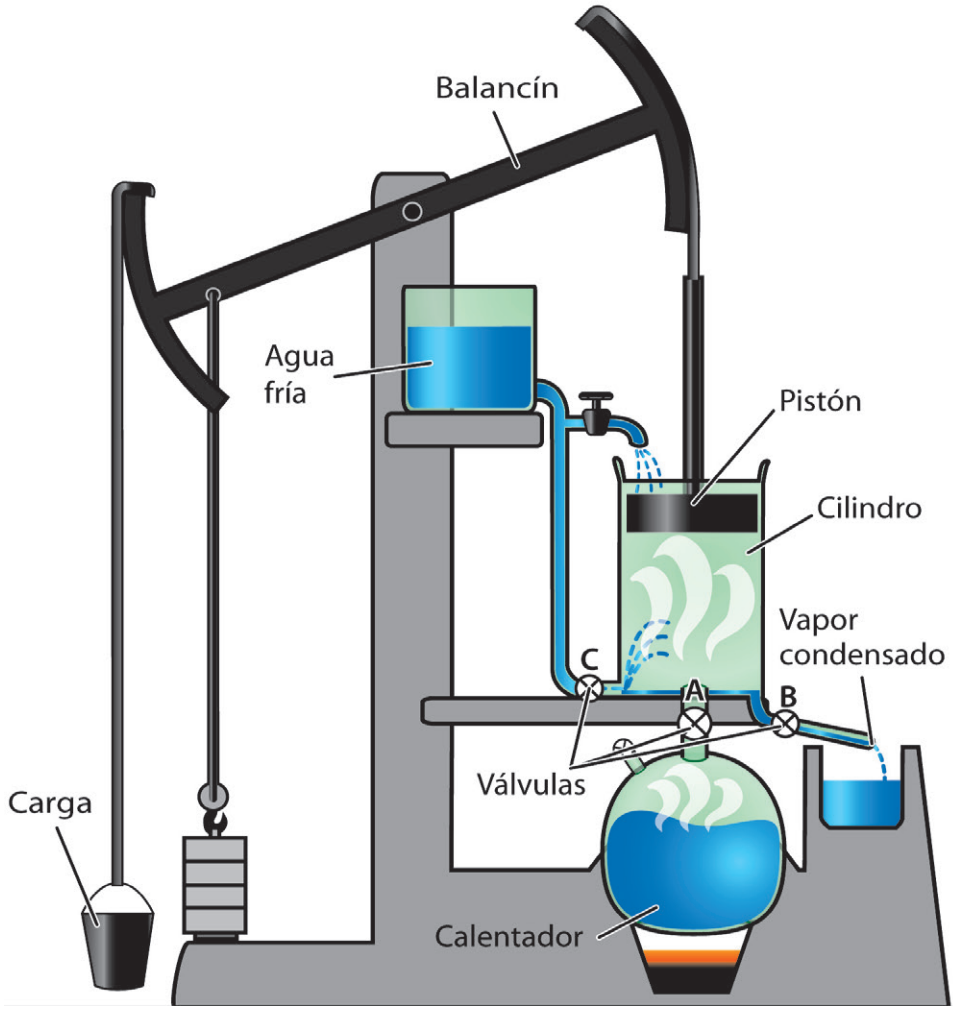


Figura 16. Motor de Newcomen. [Véase animación en CD: "Motor de Newcomen".]

brió que los calentamientos y enfriamientos sucesivos del cilindro y el émbolo podían evitarse separando la operación de condensación, evitando así la necesidad de que en una misma parte del motor se efectuaran las dos operaciones de calentamiento y enfriamiento. Su propuesta se esquematiza en la figura 17 (p. 444).

Con la válvula B cerrada se abre la A; el vapor caliente entra al cilindro, lo que ocasiona que el émbolo se mueva hacia arriba, empujando al balancín para elevar la carga o realizar otras tareas. Después se cierra la válvula A y se abre la B, dejando salir vapor al condensador. Enseguida se abre de nuevo la válvula A, cerrando la B y el ciclo se repite. Watt, con su motor, fue capaz de elevar la eficiencia del motor de Newcomen hasta un valor del 7%, es decir, un factor superior a 3. El francés Sadi Carnot, gran teórico de los motores térmicos, en 1824, indicó a propósito de la importancia de su desarrollo:

> El servicio más relevante que el motor térmico ha hecho en Inglaterra es, sin duda, el de haber reanimado la explotación de sus minas de hulla, que había disminuido y que amenazaba por extinguirse completamente a causa de la dificultad siempre creciente para el desagüe y la extracción del combustible (la extracción de hulla se multiplicó por diez, sucediendo algo semejante con otros minerales, tanto en Europa como en el Nuevo Mundo). Se deben colocar en segundo lugar los servicios prestados a la fabricación del hierro, tanto por la hulla que se ofrecía en abundancia como sustituto de la madera cuando ésta

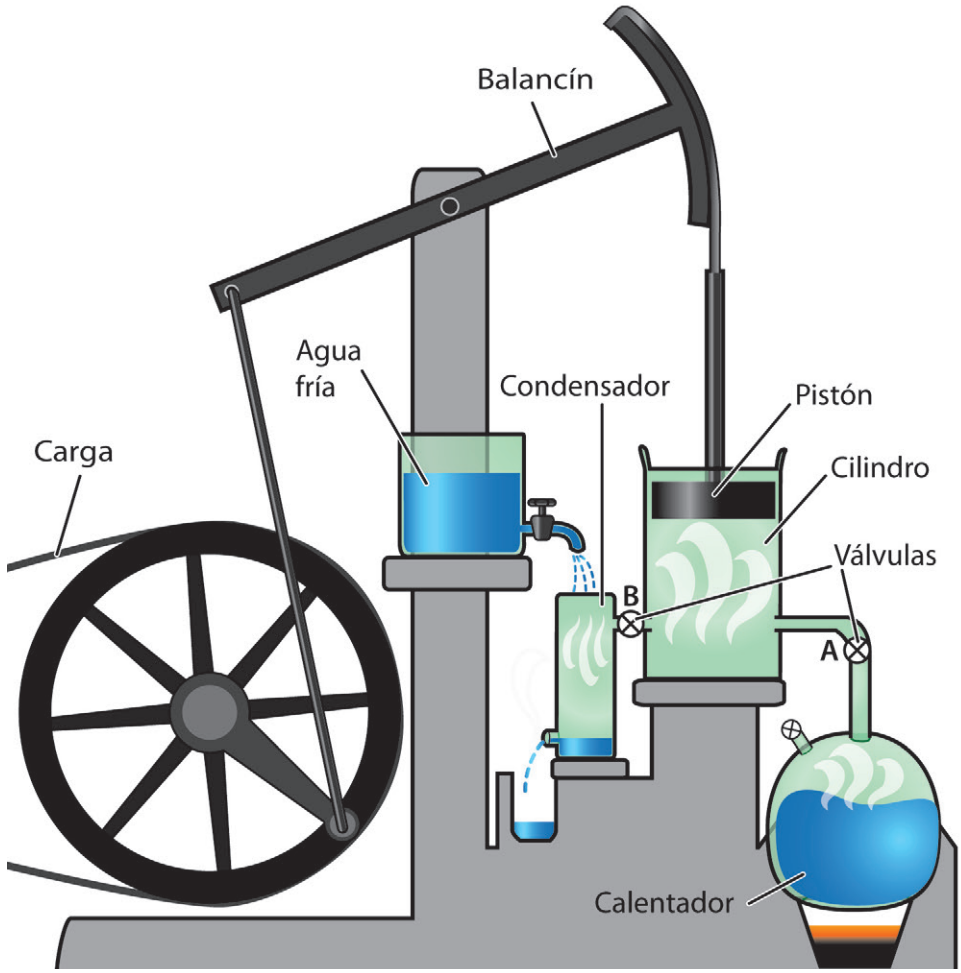


Figura 17. Motor de Watt. [Véase animación en CD: "Motor de Watt".]

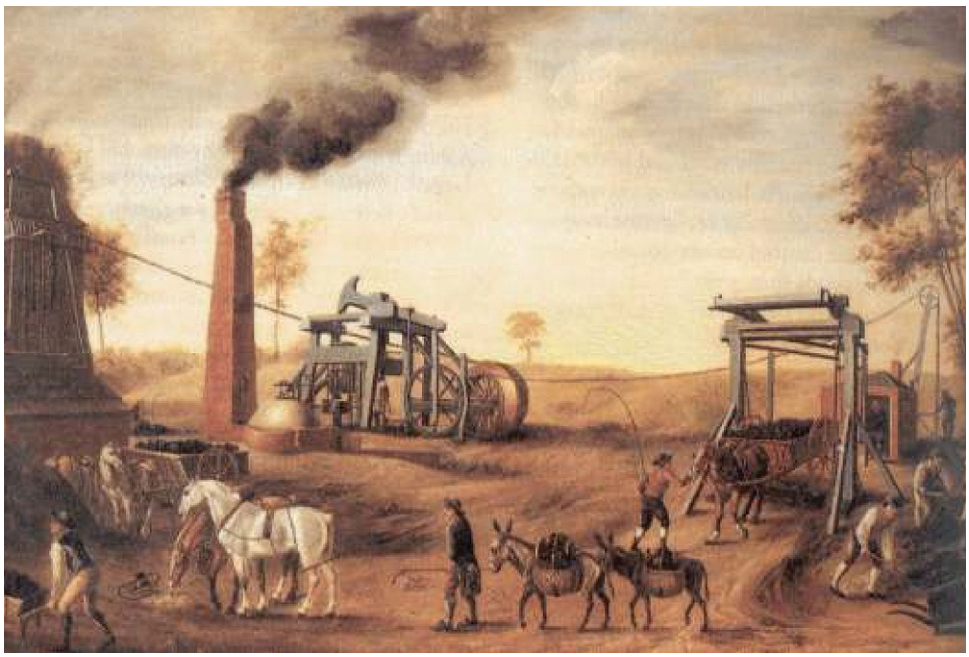
empezaba a agotarse, como por las potentes máquinas de toda clase, cuyo uso permitió o facilitó el empleo del motor térmico.

Más adelante, dice Carnot:

Quitar hoy a Inglaterra sus motores térmicos sería arrebatarle la hulla y el hierro al mismo tiempo; esto equivaldría a agotar todas sus fuentes de riqueza y arruinar todos sus medios de prosperidad; eso significaría aniquilar esta potencia colosal.

En aquellos tiempos, uno de los problemas era que la extracción del carbón de las minas se tenía que hacer desde el subsuelo, pues para entonces ya se había agotado el carbón superficial, con el agravante de que la perforación penetraba las capas freáticas y las minas se inundaban. Era necesario extraer el agua y el carbón simultáneamente, con rapidez suficiente; es decir, se necesitaba elevar desde el fondo de la mina (a una profundidad h) una cierta cantidad de agua de masa m, en un tiempo t, para permitir la extracción de carbón. Según lo expuesto en el capítulo de mecánica, se debería poder desarrollar una potencia:

$$P = \frac{\text{energía}}{\text{tiempo}} = \frac{mgh}{t}.$$



Extracción de carbón | © <www.uk.filo.pl/uk_history_8.htm>.

Para tener una idea cuantitativa, se tenía que elevar una tonelada de agua desde una profundidad de 20 metros en un minuto. La potencia mínima que los animales o el motor tendrían que desarrollar sería de:

$$P = (1\,000\ \text{kg} \times 9.8\ \text{m/s}^2 \times 20\ \text{m})/60\ \text{s} = 3\,267\ \text{W}.$$

Un aparato que puede desarrollar una potencia comparable sería un automóvil de unos 100 *caballos de potencia* (HP). La conversión de HP a watt es de:

$$1\ \text{HP} = 746\ \text{W}.$$

Un carro con seis caballos de potencia | © Latin Stock México.

Por lo tanto, la potencia en watt del motor de un automóvil de 100 HP es de

$$100 \text{ HP} = 100 \times 746 \text{ W} = 7\,460 \text{ W},$$

casi el doble de la potencia que se necesita en la tarea del desagüe.

El término "caballo de potencia" probablemente esté relacionado con una cláusula del contrato que Boulton y Watt hacían firmar a los clientes a quienes vendían los servicios de sus motores térmicos. Boulton se había asociado con Watt para vender lo que, según ellos, *todos querían*: *"energía"*. Una parte medular del contrato decía lo siguiente:

> Nuestra firma, Boulton y Watt, instalará la máquina, libre, gratis y por nada en su mina. La haremos funcionar durante los primeros cinco años y todo lo que pedimos a cambio es una tercera parte de la diferencia entre el costo del carbón para nuestra máquina y el costo del forraje para los caballos que tuviesen que realizar la misma cantidad de trabajo.

Lo que pedían Boulton y Watt era:

$$\frac{1}{3}(\text{costo del forraje} - \text{costo del carbón}), \quad \text{para realizar el trabajo } W.$$

En el párrafo aparecen dos términos que requieren de una definición precisa, además de un procedimiento riguroso para medirlos con exactitud: el trabajo desarrollado por un caballo y el trabajo desarrollado por el motor térmico construido por la compañía, así como las cantidades de carbón y forraje consumidas cuando ambos trabajos son iguales. Mientras que la cantidad de carbón necesaria para elevar una cierta cantidad de agua (o de carbón) de una cierta profundidad podía calcularse mediante pruebas directas con el motor, no era tan fácil calcular para un caballo.

Watt realizó el siguiente experimento: ató una cuerda a un cuerpo de masa m, la introdujo en un pozo de profundidad *h*, hizo pasar la cuerda por una polea situada en la boca del pozo y luego la amarró al arnés de un caballo. Ajustó *m*, *h* y el tiempo de elevación para saber la potencia máxima ($P_{máx}$) a la que el animal podía trabajar durante un cierto tiempo *t* sin cansarse, llegando así a que la $P_{máx}$ calculada adquirió un valor cercano al que conocemos ahora de 746 W.

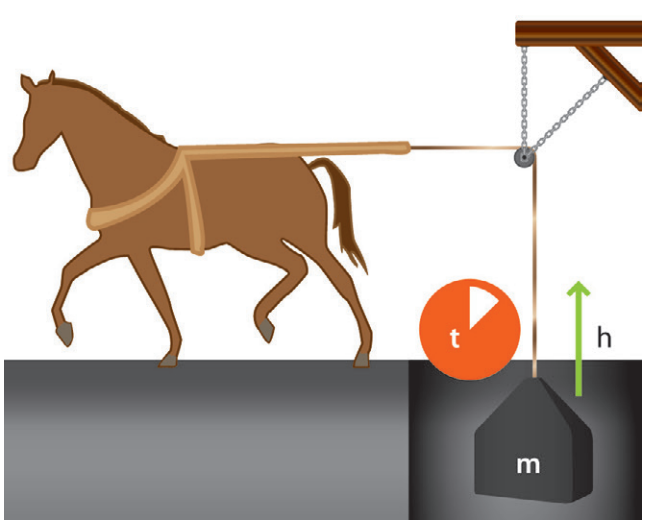


Figura 18. Experimento de Watt para medir el "caballo de potencia" (HP).

Quienes vendían y compraban tenían que comprender los términos de la transacción; por eso los conceptos de *cantidad de trabajo*, *potencia* y *energía* (ahora comunes en la física) fueron inventados para la aplicación de los motores térmicos en la sociedad.

Por otro lado, un concepto fundamental necesario para definir la operación de un motor térmico es su *eficiencia* (representada por la letra griega η, eta). La eficiencia es el cociente del trabajo obtenido por el motor entre la cantidad del calor que recibe de la caldera.

La operación del motor de Newcomen satisfizo desde el principio a sus usuarios. Sin embargo, en la medida en que el costo del carbón aumentaba, se hacía necesario obtener más trabajo con la misma

cantidad de carbón, o sea, era necesario aumentar la eficiencia η. Por eso, el desarrollo de motores térmicos continuó y James Watt (1736-1819) produjo una revolución en su funcionamiento.

Sadi Carnot (1796-1832), estudiando los motores térmicos de su época, descubrió los tres elementos universales del motor térmico: la caldera, el ciclo de la sustancia que trabaja y el condensador. Se propuso estudiar científicamente los motores térmicos, porque eran parte vital de la actividad económica y productiva de las naciones altamente industrializadas, en particular Inglaterra. Carnot observó que el aumento de la eficiencia, con las sucesivas modificaciones a los motores térmicos, era cada vez menor. Watt había obtenido al principio un salto espectacular de 300% con respecto a la eficiencia del motor de Newcomen; pero en los cambios subsiguientes la eficiencia parecía tener un límite del cual no podría pasarse, independientemente de que se mejorara el diseño o la sustancia con la que trabajara el motor. Carnot se preguntó si la eficiencia podía aumentarse indefinidamente hasta 100%, en cuyo caso todo el calor de la caldera se convertiría en trabajo, ya fuera por el cambio de diseño, la naturaleza de la sustancia con que trabajara o cualquier otro factor.

La pregunta tenía sentido porque, como se puede observar en la figura 19, el límite de η quedaba lejos del valor 1, es decir, de 100 por ciento.

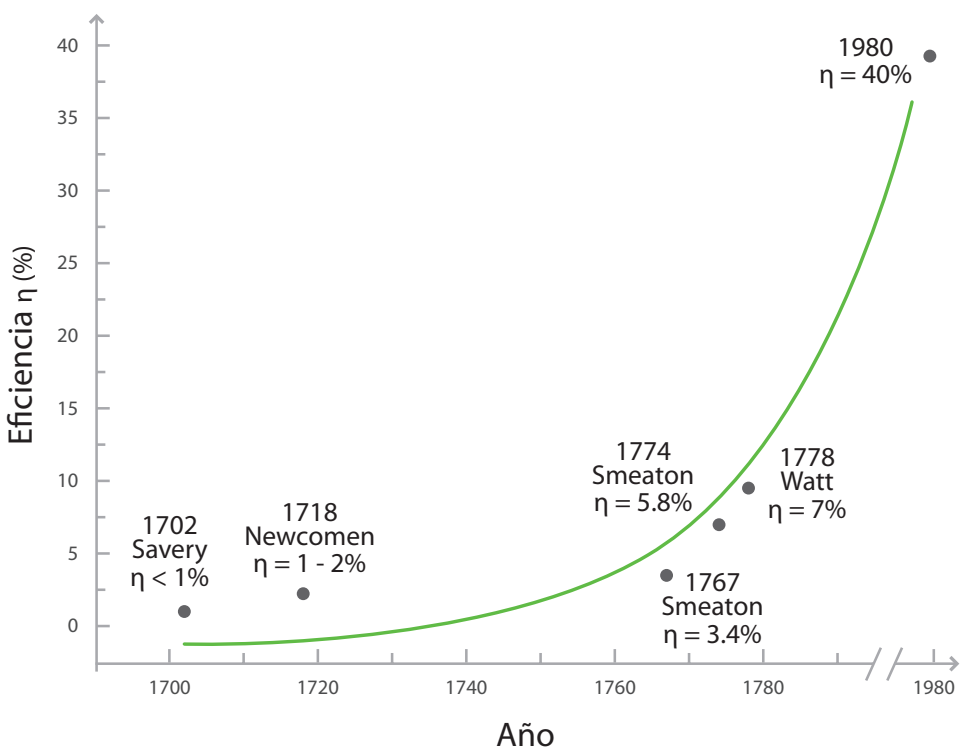


Figura 19. Eficiencia de varios motores térmicos a lo largo del tiempo.

En la figura 20 (p. 448) se muestra un esquema del motor térmico universal. En primer lugar, la caldera donde se quema el carbón y se genera el calor Q_1 que actúa sobre la "sustancia que trabaja" (el vapor de agua que ingresa al cilindro); en segundo lugar, la sustancia que trabaja en el cilindro, que lo hace en ciclos, y, por último, se tiene el condensador, que recibe la energía disipada por calor Q_2. Con estos elementos, Sadi Carnot logró establecer que:

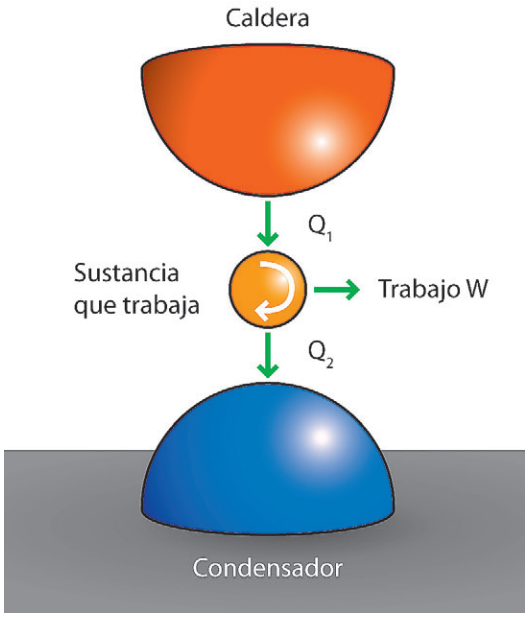


Figura 20. Motor térmico universal. [Véase video en CD: "Motor térmico".]

1] En el motor térmico era tan importante el condensador (a baja temperatura) como la caldera (a alta temperatura) para poder desarrollar trabajo, porque preveía que si el ambiente estuviera todo a la temperatura de la caldera, el calor no podría "fluir", pues no habría la posibilidad de "pasar" a un lugar con menor temperatura. En otras palabras, Carnot utilizó el modelo "sustancialista", el cual supone que el calor es similar a un líquido que necesita pasar de un nivel superior a uno inferior para moverse, al igual que el agua en una cascada.

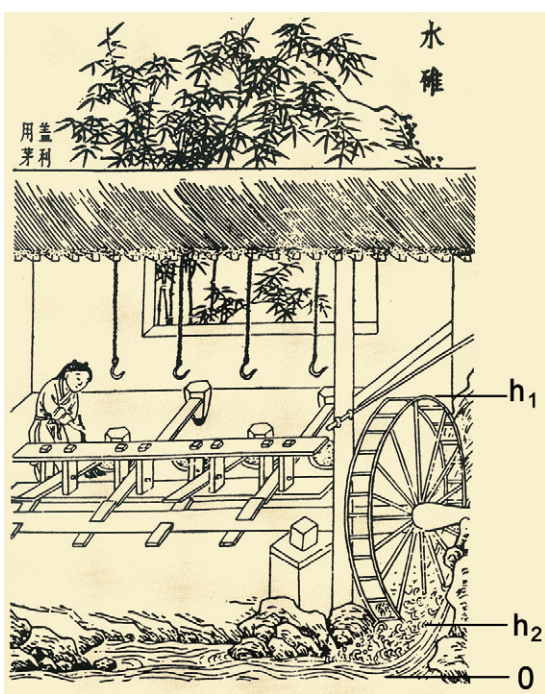
2] El máximo trabajo que se puede obtener de Q_1 es mediante la utilización de un motor que opere reversiblemente, es decir, sin que haya fricción entre sus partes y que opere muy lentamente. Este importante enunciado es el ahora conocido *teorema de Carnot*, el que también puede enunciarse en términos de la eficiencia, es decir: *Ningún motor térmico funcionando entre una caldera y un condensador fijos puede tener eficiencia mayor que la de un motor reversible funcionando entre la misma caldera y el mismo condensador.*

3] El trabajo debía depender de la diferencia de temperatura o "caída de temperatura" $(t_1 - t_2)$ entre la caldera a t_1 y el condensador a t_2.

4] El trabajo máximo no solamente depende de la "caída" en temperatura $(t_1 - t_2)$, sino que aumenta al disminuir la temperatura t_1 de la caldera; es decir, el calor Q_1 produce más trabajo en la caída de 50 °C a 10 °C, que de 90 °C a 50 °C, aunque en ambos casos la diferencia de temperaturas sea de 40 °C.

Rueda hidráulica | *© Latin Stock México.*

Reversibilidad mecánica. Motores hidráulicos



En la demostración de estos resultados, Sadi Carnot se guió por la analogía hidráulica que años antes su padre, Lázaro Carnot, había establecido con un teorema: *Los motores hidráulicos, es decir, las ruedas de agua verticales de máxima potencia, son las que funcionan reversiblemente.*

La rueda de agua funciona reversiblemente si una masa m de agua, que se recibe en un cajón de la rueda a la altura h_1, baja al rotar la rueda a la altura h_2 y regresa de nuevo a la altura inicial. Cuando la masa de agua se encuentra a la altura h, la conservación de energía nos dice que:

Energía de rotación de la masa m + Energía de rotación de la rueda + Energía perdida por fricción + $mgh = mgh_1$.

Llamemos ER a la suma de las energías de rotación de la masa m y la rueda, y EF a la energía perdida por fricción, por ejemplo en el eje de rotación de la rueda. Tenemos, por la conservación de la energía total:

$$ER + EF + mgh = mgh_1.$$
$$ER = mg(h_1 - h) - EF.$$

Si no hay fricción en la rueda, que es reversible, es decir, EF = 0, entonces:

$$ER = mg(h_1 - h).$$

O sea, la energía de rotación del sistema (rueda y masa de agua m) aumenta conforme disminuye h, es decir, conforme m baja al rotar la rueda. El valor mínimo de h al que m puede bajar, que es h_2, corresponde al valor máximo de ER, que denotaremos por $ER_{máx}$, por lo que:

$$ER_{\max} = mg(h_1 - h_2).$$

Si la rueda opera sin fricción y no se pierde agua durante todo el trayecto, es decir, si la rueda opera reversiblemente, la masa m seguirá subiendo después de alcanzar el punto más bajo de la rueda, hasta llegar nuevamente a la altura superior h_1, de la cual partió. En este punto otra vez ER = 0. Si hubiera fricción en alguna parte, por ejemplo en el eje de

Caja sobre reversibilidad. Parte 1.
La reversibilidad desde el punto de vista mesoscópico

Otro ejemplo de reversibilidad mecánica es un balín de acero que se deja caer de cierta altura contra el piso; el balín rebota a la posición original al cumplirse que:

1] el aire no presenta resistencia;
2] el impacto contra el suelo es elástico, lo que quiere decir que la energía cinética del balín no varía en el choque.

Si se toma un video del movimiento del balín, no se distingue este movimiento cuando se pasa de adelante hacia atrás o de atrás hacia adelante.

En cambio, si se graba en video el movimiento del balín de modo que las condiciones 1 y 2 no se cumplan, es posible entonces distinguir el sentido en que se pasa el video, ya que se vería que el balín comienza a rebotar, cada vez a mayor altura.

Así pues, la fricción con el aire y el hecho de que los choques del balín contra el suelo no sean elásticos, nos permite distinguir si el video está en reversa; cuando esto sucede, es porque el proceso grabado es irreversible. Con una cámara de video o un teléfono celular se pueden filmar las siguientes acciones:

1] un balín rebotando;
2] movimiento contra el suelo de una pelota blanda;
3] piedra cayendo en un estanque;
4] difusión de un perfume en el aire, o de una sustancia coloreada al caer en agua.

Con el balín y con la piedra, la fricción interviene en el intercambio de energía de tipo termodinámico entre estos objetos, el aire y el objeto de impacto: el suelo y el agua, respectivamente. Algo semejante sucede con la tinta expandiéndose en el agua; la fricción es, en este caso, entre las moléculas de la tinta y las del agua.

La fricción implica cambios en las energías internas de los sistemas por calor y por trabajo, de modo que ellos deben calcularse tomando en cuenta la primera ley de la termodinámica Uf – Ui = ΔU = Q – W, así como la segunda ley.

giro de la rueda o se presentara una fuerte corriente de aire, la rueda se detendría a una altura h menor que h_1; de hecho, hasta podría ser que m no llegara hasta abajo.

Ahora bien, si se quiere hacer funcionar algún otro dispositivo técnico usando la rueda en rotación, el trabajo máximo que se puede obtener será igual al valor máximo de la energía de rotación. En otras palabras, el máximo trabajo que puede realizar la rueda, llamémosle $W_{máx}$, será el que la deja quieta, con la masa m en la posición más baja h_2. Por lo tanto:

$$W_{\max} = mg(h_1 - h_2).$$

Esta ecuación nos lleva a que $W_{máx}$ depende de la "caída" de altura $(h_1 - h_2)$.

Los motores se caracterizan por su eficiencia, que es el trabajo que pueden desarrollar en relación con la energía suministrada, y está definida como:

$$\eta = \frac{\text{Energía desarrollada como trabajo W}}{\text{Energía invertida}} = \frac{W}{mgh_1}.$$

La eficiencia es máxima cuando para una inversión de energía potencial de la masa m a la altura h_1, se obtiene el trabajo máximo, que es el desarrollado por la rueda de agua reversible:

$$\eta_{\max} = \frac{W_{\max}}{mgh_1} = \frac{mg(h_1 - h_2)}{mgh_1} = \frac{h_1 - h_2}{h_1} = 1 - \frac{h_2}{h_1}.$$

El máximo trabajo también se puede expresar en términos de la máxima eficiencia:

$$W_{\max} = mgh_1\eta_{\max} = mgh_1\left(1 - \frac{h_2}{h_1}\right).$$

Una propiedad interesante de $W_{máx}$ es que aumenta al disminuir h_1. Es decir, el máximo trabajo es mayor a pequeña altura que a gran altura, aunque la caída $(h_1 - h_2)$ sea la misma.

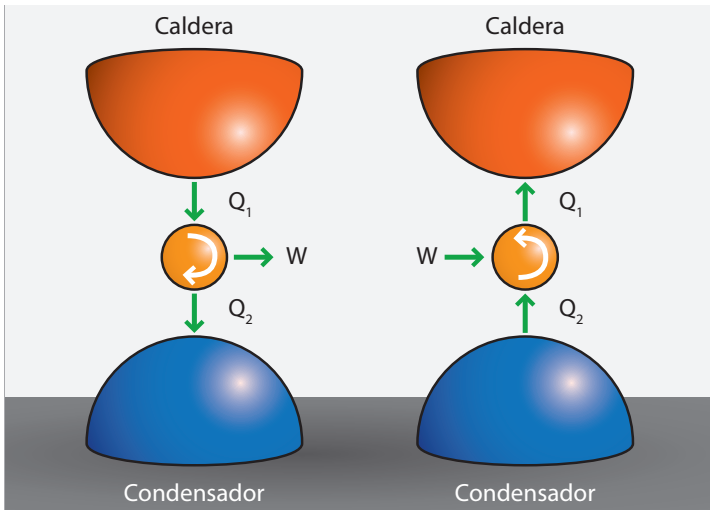


Figura 21. Motor reversible y su inverso: un refrigerador reversible.

Un motor térmico funciona reversiblemente si los intercambios de energía de la sustancia que trabaja con la caldera y el condensador son muy lentos, sin fricción y sin disipación, de tal manera que al invertir las "entradas" y "salidas" de energía en el motor se invierten exactamente.

El motor reversible funcionando al revés es un refrigerador; en la rueda hidráulica correspondería a la bomba hidráulica que transporta agua del nivel inferior al superior, para lo cual habría que realizar trabajo sobre la bomba. En un refrigerador, de manera semejante, también hay que invertir trabajo (eléctrico) para disminuir la energía interna y, por lo tanto, la temperatura de su interior (figura 21).

Sadi Carnot, de acuerdo con la analogía hidráulica, establece un teorema semejante al de su padre:

> Ningún motor térmico funcionando entre una caldera y un condensador fijos puede tener eficiencia mayor que un motor reversible operando entre la misma caldera y el mismo condensador.

Las condiciones de operación cíclicas de un motor térmico reversible son análogas a las de operación de una rueda hidráulica, siguiendo el ciclo de Lázaro Carnot. Es decir, el *ciclo térmico de Sadi Carnot* está compuesto de cuatro pasos reversibles:

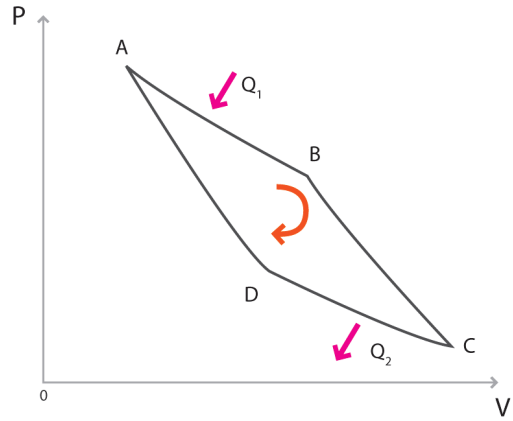


Figura 22. Ciclo de Sadi Carnot.

1] La sustancia que trabaja dentro del cilindro recibe una cantidad de energía por calor Q_1 de la caldera, sin fricción ni disipación. Esto significa que Q_1 debe aumentar la energía de la sustancia muy lentamente, siempre muy cerca del equilibrio, a la temperatura constante de la caldera T_1. Esto corresponde, en la rueda hidráulica, a que el agua se recibe en el cajón con velocidad relativa cero, a la altura h_1.

Si el proceso no fuera isotérmico, sino que se realizara a una diferencia finita de temperatura, sería semejante a que se recibiera el agua a una altura superior a h_1, produciéndose rebotes y pérdidas, violándose las condiciones de reversibilidad. El aumento de la energía interna de la sustancia que trabaja provoca en ella una expansión isotérmica, denotada por el paso de A a B en la figura 22.

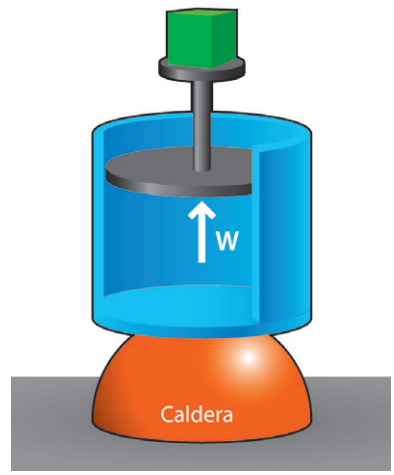
2] La expansión de B a C se efectúa adiabáticamente, es decir, sin cambios por calor ($Q = 0$) en la energía interna de la sustancia, que en la rueda corresponde a la bajada del agua sin pérdidas de líquido.

3] En la etapa de compresión de C a D la sustancia pierde energía por calor Q_2 y la cede al condensador, de manera isotérmica.

4] Finalmente, la sustancia que realiza el trabajo regresa a la posición inicial, por un paso semejante al 2, pero ahora de compresión.

Sadi Carnot había notado que la analogía del motor térmico con el motor hidráulico no es exacta, porque $Q_1 \neq Q_2$, y el trabajo desarrollado $W = Q_1 - Q_2$ es distinto de cero. En el motor hidráulico, en cambio, la cantidad de líquido que cae se conserva.

Figura 23. Gas en un cilindro con pistón en contacto con una caldera.



Entropía

El ciclo de Sadi Carnot se compone de dos isotermas y dos adiabáticas reversibles. Para mayor comprensión, en la figura 23 se dibuja un gas encerrado en un cilindro con un pistón, en contacto diatérmico con una caldera, recibiendo energía por calor isotérmicamente.

La eficiencia de un motor térmico será la máxima y, por analogía con la rueda reversible en que la máxima eficiencia es $1 - h_2 / h_1$, tendrá el valor:

$$\eta_{\max} = 1 - \frac{T_2}{T_1}.$$

Como:

$$\eta_{\max} = \frac{W_{\max}}{Q_1} = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{Q_2}{Q_1} = 1 - \frac{T_2}{T_1}.$$

se tiene que el máximo trabajo térmico vale:

$$W_{\max} = Q_1 \eta_{\max} = Q_1 \left(1 - \frac{T_2}{T_1}\right) = \frac{Q_1}{T_1}\left(T_1 - T_2\right).$$

Es decir, se cumple que el trabajo máximo depende de la "caída", o "diferencia" de temperaturas ($T_1 - T_2$), pero aumenta conforme disminuye T_1, tal y como había anticipado Sadi Carnot. Esto mismo sucede en un motor hidráulico.

De la expresión para la máxima eficiencia térmica:

$$\eta_{\max} = \frac{W_{\max}}{Q_1} = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{Q_2}{Q_1} = 1 - \frac{T_2}{T_1} \Rightarrow \frac{Q_2}{Q_1} = \frac{T_2}{T_1} \Rightarrow$$
$$\frac{Q_2}{T_2} = \frac{Q_1}{T_1}.$$

Al cambio de estado físico, calculado por el cociente Q/T, se le llama *cambio de entropía* de la caldera o del condensador. Si la entropía se denota por *S*, y su cambio por Δ*S*, la igualdad

$$\frac{Q_2}{T_2} = \frac{Q_1}{T_1}$$

implica que

$$\Delta S_1 = \Delta S_2.$$

En esta igualdad los valores de los cambios de entropía son absolutos, es decir, positivos. Pero si tomamos en cuenta que Q_1 es en realidad negativo, porque se extrae energía de la caldera, se tiene que el cambio de entropía de ésta será también negativo:

$$\Delta S_1 = \frac{Q_1}{T_1} < 0.$$

Por el contrario, el cambio de entropía del condensador ($Q_2 > 0$) es positivo:

$$\Delta S_2 = \frac{Q_2}{T_2} > 0.$$

Por lo tanto, en vez de $\Delta S_1 = \Delta S_2$, que se refiere a valores absolutos, se debe tener:

$$-\Delta S_1 = \Delta S_2.$$

$$\Delta S_1 + \Delta S_2 = 0.$$

Como la sustancia que trabaja cambia cíclicamente de estado como consecuencia de los intercambios energéticos con la caldera y el condensador, su cambio de entropía es cero.

Así que, si sumamos todos los cambios de entropía involucrados en la operación de un motor reversible, se concluye que:

$$\Delta S_{\text{total}} = 0,$$

si los procesos en el motor son reversibles.

Si, por el contrario, los intercambios energéticos entre la sustancia que trabaja, la caldera y el condensador son irreversibles, entonces la cantidad $\Delta Q/T$ no se conserva, sino que en valor absoluto será mayor la del condensador que la de la caldera.

Supongamos ahora un motor irreversible que procesa calor Q_1 de la caldera y Q_2 del condensador. El teorema de Carnot afirma que su eficiencia será menor que la de un motor reversible funcionando entre dicha caldera y condensador; por lo tanto:

$$\eta_{\text{irrev}} < \eta_{\text{rev}}.$$

Es decir:

$$\eta_{\text{irrev}} = \frac{W}{Q_1} = \frac{Q_1 - Q_2}{Q_1} = 1 - \frac{Q_2}{Q_1} < 1 - \frac{T_2}{T_1} = \eta_{\text{rev}}.$$

De aquí se deduce que:

$$\frac{Q_2}{T_2} > \frac{Q_1}{T_1}.$$

En términos de cambios de entropías y si, como antes, se toman en cuenta sus signos:

$$\Delta S_2 > -\Delta S_1 \Rightarrow \Delta S_1 + \Delta S_2 > 0.$$

El cambio de entropía de la sustancia que trabaja sigue siendo nulo porque trabaja en ciclos completos. En consecuencia, es cierto ahora que:

$$\Delta S_{\text{total}} > 0,$$

si los procesos en el motor son irreversibles.

Principio del incremento de entropía del universo termodinámico

Los dos resultados anteriores, que sólo se refieren a la operación de motores térmicos, se pueden expresar de otra manera. Para esto se define el *universo termodinámico* como el conjunto de objetos que participan en un proceso; en el caso del motor térmico, el universo está compuesto de la sustancia que trabaja, la caldera y el condensador.

Principio del incremento de la entropía: *El cambio de entropía del universo termodinámico es positivo siempre que ocurre un proceso irreversible en su interior.*

Si el proceso es reversible, entonces la entropía del universo permanece constante y no cambia. Ahora bien, es posible demostrar que el principio del incremento de la entropía es válido para cualquier tipo de proceso irreversible que ocurre en el interior del universo termodinámico, y no sólo para procesos involucrados en la operación de motores térmicos.

Una característica de todos los procesos que ocurren en el universo termodinámico es que se dan porque existe un desequilibrio, ya sea entre un objeto y otros, o en el interior del objeto. En el motor térmico, el desequilibrio se da entre la caldera y el condensador.

Los objetos evolucionan entonces hacia el equilibrio, de manera irreversible, de tal modo que la entropía del universo siempre crece. Una vez que este proceso hacia el equilibrio termina, se llega al reposo, que es el estado en que la entropía del universo llega a su valor máximo. El principio del incremento de la entropía marca así una dirección, que los procesos hacia el equilibrio deben seguir, es decir, hacia el estado del universo en que la entropía aumenta. Se dice también que el principio del incremento de la entropía fija un sentido en el tiempo, del pasado al futuro. Una consecuencia importante de lo expuesto es que los motores térmicos contaminan aun cuando funcionen reversiblemente, sin fricción ni disipación de energía.

El calor Q_2 procesado necesariamente en el condensador, que generalmente es el ambiente, se traduce en contaminación; es decir, provoca un impacto ambiental que, ciertamente, es mínimo, pero no despreciable.

Si el motor es irreversible, tenemos que:

$$Q_2 > Q_1 \frac{T_2}{T_1};$$

de modo que la contaminación térmica mínima es :

$$Q_{2,\ \text{mínimo}} = Q_1 \frac{T_2}{T_1}.$$

Esta energía por calor eleva la temperatura de algún cuerpo de agua (río, mar, etc.) o el aire, si la termoeléctrica cuenta con torres de enfriamiento. La contaminación térmica que genera suponiendo que Q_2 es el calor de una termoeléctrica de 1 GW, que el condensador se enfría con el agua de un lago y que trabaja con una eficiencia de $\eta = 0.36$, entonces:

$$Q_1 = \frac{\mathrm{W}}{\eta} = \frac{1\ \mathrm{GW} \times \mathrm{s}}{0.36}.$$

Por otro lado, la temperatura ambiente es de 300 K, mientras que la de la caldera es típicamente de 825 K, así que

$$\frac{T_2}{T_1} = \frac{300\ \mathrm{K}}{825\ \mathrm{K}} = 0.36.$$

Entonces:

$$Q_2 = \frac{1\ \mathrm{GW} \times \mathrm{s}}{0.36} \times 0.36 = 1\ \mathrm{GW} \times \mathrm{s}.$$

En un segundo, la termoeléctrica energiza el agua del lago con una energía de:

$$Q_2 = \text{potencia} \times 1\ \mathrm{s} = 1\ \mathrm{GW} \times 1\ \mathrm{s} = 10^9\ \mathrm{W} \times \mathrm{s} = 10^9\ \mathrm{J}.$$

Si el lago tiene 10 km de diámetro y una profundidad promedio de 10 m, su volumen de agua es de:

$$V = \text{arca} \times \text{profundidad} = \pi\, r^2 \times h = 3.1416 \times (5000\ \text{m})^2 \times 10\ \text{m} = 7.854 \times 10^8 \times \text{m}^3.$$

y la masa de agua es de:

$$M = \text{densidad} \times \text{volumen} = 1\frac{\text{kg}}{\text{litro}} \times \text{volumen} = \frac{1000\ \text{kg}}{1 \times \text{m}^3} \times 7.854 \times 10^8 \times \text{m}^3$$
$$= 7.854 \times 10^{11}\ \text{kg}.$$

Por lo tanto, la energía de la termoeléctrica vertida por calor en un segundo en el lago eleva su temperatura en:

$$\Delta t = \frac{Q_2}{c_p M} = \frac{10^9\ \text{J}}{4.185 \times 10^3 \frac{\text{J}}{\text{kg} \times {}^\circ\text{C}} \times 7.854 \times 10^{11}\ \text{kg}} = 3 \times 10^{-7}\ {}^\circ\text{C}.$$

Se ha supuesto que la energía por calor es recibida de manera uniforme en un segundo por toda el agua del lago. El resultado anterior parece mostrar que el efecto térmico en el agua en un segundo es insignificante.

Sin embargo, la elevación de la temperatura en 5 °C es capaz de alterar severamente las condiciones de vida en el lago. Esta elevación se alcanza, según los cálculos simplificados anteriores, en apenas 190 días.

Obviamente, la contaminación térmica en los alrededores del desagüe del agua de enfriamiento que proviene del condensador, es mucho mayor de los 5 °C, por lo que su impacto es aún más intenso.

La contaminación térmica de las aguas provoca que disminuya la cantidad de oxígeno disuelto, pero que al mismo tiempo aumente la utilización de oxígeno por los organismos vivos del medio acuático, que se estratifiquen las aguas por diferencias en densidad y que se aceleren las reacciones químicas. Estos cambios hacen que disminuya la reproducción y la capacidad de supervivencia de los peces, que se limiten los patrones de migración, que se modifiquen los procesos que dependen de los ritmos biológicos, que se incremente la susceptibilidad a las enfermedades y el aumento desproporcionado de algunos organismos. Es decir, en apenas un año, un lago del tamaño considerado queda destruido. La única forma de no deteriorar térmica y químicamente el ambiente por causa de las termoeléctricas es evitando su uso. Los energéticos renovables son la alternativa. Esta contaminación inevitable es una consecuencia de la segunda ley de la termodinámica.

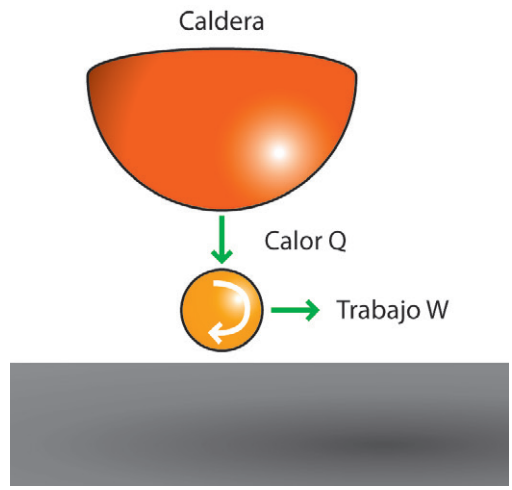


Figura 24. Motor que viola la segunda ley de la termodinámica.

6.3.3 La segunda ley de la termodinámica

La afirmación básica de Carnot de que en todo motor (reversible o no) son indispensables tanto la caldera como el condensador, fue tan importante que, decenios después, Max Planck la postuló como la segunda ley de la termodinámica: *Ningún motor térmico trabajando en un ciclo puede convertir íntegramente calor en trabajo*. En otras palabras, es imposible construir un motor que funcione según la figura 24.

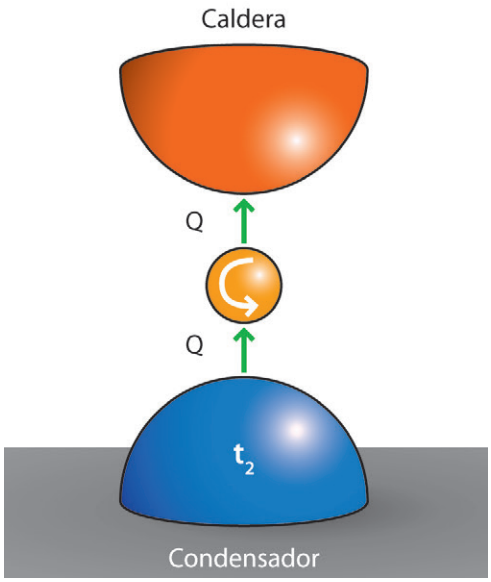


Figura 25. Refrigerador que viola la segunda ley de Clausius.

El diagrama de operación en ciclos del motor que viola la segunda ley de la termodinámica elimina precisamente el condensador. El precio a pagar, sin embargo, es la inevitable y ya mencionada contaminación térmica y química, ya que el calor Q_2 proviene mayoritariamente de la quema de combustibles fósiles en las termoeléctricas, al igual que Q_1 (figura 25).

Se ha mencionado que un motor trabajando en sentido inverso es un refrigerador, de manera semejante a como una rueda trabajando al revés es una bomba de agua.

Clausius formuló, a su vez, la segunda ley de la termodinámica en términos de refrigeradores, aunque se puede demostrar que es equivalente al enunciado de Planck: *Ningún refrigerador trabajando en un ciclo puede enfriar un sistema, pasando energía a otro sistema más caliente, sin que se efectúe trabajo sobre él, por lo que es* imposible construir un refrigerador como el de la figura 25.

El sistema que se enfría es el foco indicado por T_2. Así pues, no hay refrigeradores que operen en ciclos sin que se haga trabajo sobre ellos. No hay refrigeradores gratis: necesariamente hay que pagar el precio de la electricidad consumida.

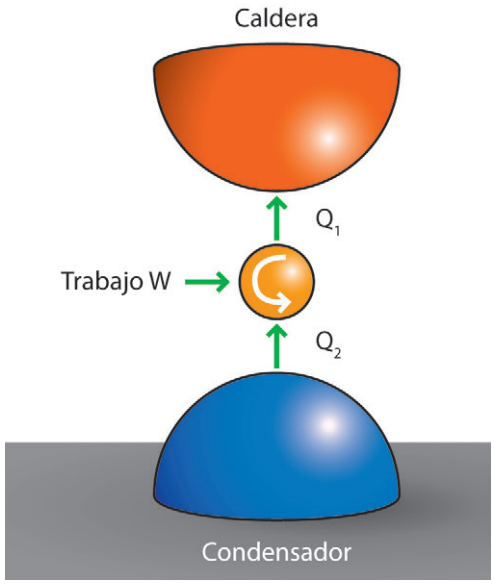
La pregunta de Carnot sobre la eficiencia máxima de los motores térmicos condujo a avances significativos de la termodinámica, pero también llevó al cálculo de su contaminación mínima. Este valor mínimo sirve como patrón para medir lo que se puede ahorrar en la quema de combustibles fósiles.

Hoy en día hay una gran variedad de dispositivos técnicos: refrigeradores, bombas de calor, acondicionadores de aire, etc. Si se quiere disminuir los costos de operación y la contaminación que ocasionan por funcionamiento ineficiente, hay que saber cuáles son sus eficiencias máximas posibles. Estas preguntas requieren la actualización del estudio de Carnot de los motores térmicos, a estos dispositivos.

Respecto a la eficiencia máxima de un motor térmico, Carnot indicó que:

$$\eta_{\max} = 1 - \frac{T_2}{T_1}.$$

Figura 26. Refrigerador doméstico | © Latin Stock México.
A la derecha, su diagrama.



Caja sobre reversibilidad. Parte 2. La irreversibilidad desde el punto de vista microscópico

Se ha analizado el caso del balín que se deja caer desde una cierta altura y el de la piedra que cae en un estanque. Si es una gota de tinta en un vaso de agua la tendencia es a ocupar uniformemente el recipiente, la probabilidad de que el balín comience a moverse del reposo para alcanzar la altura inicial, que la piedra salga del fondo del estanque hasta ocupar la posición inicial o de que la tinta regrese espontáneamente a ocupar el volumen de la gota de la que partió, es extraordinariamente pequeña.

Lo asombroso es que si estos procesos obedecieran solamente a las leyes de Newton de la mecánica, los tres sistemas podrían regresar a su estado inicial, pues nada hay en esas leyes que impida que, al invertir el sentido del tiempo, los sistemas regresen a sus estados iniciales.

Si tuviéramos un gas coloreado o perfumado que constara solamente de unas tres moléculas, esperaríamos un regreso simultáneo de ellas hacia la mitad de donde salieron, en unos minutos; pero si consideramos 6, después 60, luego 600 y al final 10^{23} moléculas, la probabilidad del regreso simultáneo de todas ellas a la mitad inicial del recipiente es increíblemente pequeña. [Véase video en CD: "Procesos reversibles e irreversibles".]

Así, la probabilidad de que una molécula de perfume gaseoso se encuentre en la mitad inicial es ½; la probabilidad de que dos moléculas regresen simultáneamente a la mitad inicial es de $(½)^2$; $(½)^3$ es la probabilidad para que tres moléculas concurran al mismo tiempo en la mitad inicial. Generalizando, $(½)^{10^{23}}$ es la probabilidad de que 10^{23} moléculas coincidan al mismo tiempo en la mitad inicial, una vez que han ocupado uniformemente todo el recipiente. Esta probabilidad es, para todo propósito práctico, igual a cero.

Con cuántos ceros, después del punto decimal, se escribe el número $(½)^{10^{23}}$.

Tómese en cuenta que 2 es aproximadamente $10^{0.3}$, entonces:

$$\left(\frac{1}{2}\right)^{10^{23}} \approx \left(\frac{1}{10^{0.3}}\right)^{10^{23}} = 10^{-0.3\times 10^{23}} = 10^{-3\times 10^{22}} = 0.0\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 01.$$

(aquí hay 3×10^{23} ceros)

Si para escribir cada cero usamos medio centímetro en el papel, la longitud del número completo mediría 10^9 (o sea, mil millones) de veces la distancia de la Tierra al Sol, que es de 150 millones de kilómetros.

El cálculo anterior indica que la distribución uniforme de moléculas del gas perfumado es mucho más probable que cualquier otra distribución. La distribución no uniforme inicial, en que el perfume está en una sola mitad, lleva irreversiblemente a la distribución uniforme de moléculas en las dos mitades. Este proceso ocurre porque hay un desequilibrio de presión entre una mitad y la otra. En el estado final de equilibrio las presiones se igualan.

Desde el punto de vista macroscópico, la entropía del estado no uniforme es menor que la entropía del estado uniforme. La dirección en que el proceso transcurre es tal que la entropía crece, hasta llegar a su valor máximo en el estado de equilibrio. Esta afirmación es la versión microscópica de la segunda ley de la termodinámica: *Los procesos ocurren en el sentido en que la probabilidad del estado final aumenta al máximo.*

Parecería extraño que la formulación de Planck o la de Clausius se pudieran entender en estos términos microscópicos, sin embargo, es cierto. Si, partiendo del estado uniforme, se tuviera a mano un refrigerador que violara el enunciado de Clausius, entonces se podría usar para "enfriar" la mitad del recipiente, hasta dejar esa mitad vacía de moléculas.

El resultado sería equivalente a que, sin realización de trabajo alguno, se hiciera que todas las moléculas ya difundidas en ambas mitades regresaran a una sola. Esto indudablemente sería imposible, porque la probabilidad, como ya vimos en el cálculo numérico, es prácticamente cero.

Si una termoeléctrica opera con una eficiencia de 0.36, entre una caldera a 825 K y un condensador a 300 K, su eficiencia máxima es 0.64. Es decir, su eficiencia puede mejorarse casi al doble, y su contaminación térmica puede reducirse considerablemente. ¿Cuál es la eficiencia máxima de un refrigerador? En un refrigerador de cocina, se trata de extraer energía por calor Q_3 de la parte fría del interior, que está a la temperatura T_3, invirtiendo para ello trabajo W, como se ve en la figura 26 (p. 456). Por eso su eficiencia, que se llama coeficiente de desempeño, COD, es:

$$COD = \frac{Q_3}{W} = \frac{Q_3}{Q_2 - Q_3} = \frac{1}{\frac{Q_2}{Q_3} - 1}.$$

El coeficiente de desempeño máximo es el de un refrigerador reversible. Al invertir el refrigerador en su operación, se convierte en un motor reversible, en donde Q_2 se procesa con la habitación "tibia" a T_2 (la cocina), y Q_3 es el calor procesado con el condensador a T_3 (el interior del refrigerador). Entonces,

$$\frac{Q_2}{Q_3} = \frac{T_2}{T_3}.$$

Sustituyendo este cociente en el COD queda que el valor máximo es:

$$COD_{\max} = \frac{1}{\frac{T_2}{T_3} - 1}.$$

Razonamientos semejantes permiten calcular las eficiencias máximas de todos los dispositivos técnicos termodinámicos y así disminuir los gastos económicos y de contaminación.

6.3.4 Tarea termodinámica y su eficiencia. Contraste termodinámico y exergía

El embargo petrolero de 1973 por parte de la Organización de Países Exportadores de Petróleo (OPEP) en contra de los países que habían apoyado a Israel en la guerra del Yom Kipur y que cuadruplicó el precio del barril de petróleo, obligó al mundo industrializado a pensar sobre la manera ineficiente con que se venía utilizando tan preciado recurso energético.

En 1975, la American Physical Society (Sociedad Americana de Física) publicó un estudio en el que se ofrecía un marco conceptual para enfrentar una mejor utilización no sólo del petróleo, sino de todos los recursos energéticos de los que puede disponer una sociedad cualquiera para el sostenimiento de todas sus actividades (industria, transporte, agricultura, comercio, vivienda, oficinas, etcétera). Este estudio resultó ser la actualización del libro publicado en 1824 por Sadi Carnot, *Sobre la potencia motriz del fuego*, pero con muy importantes novedades.

En primer lugar, el libro de Carnot se centraba en identificar la máxima cantidad de trabajo que se puede obtener de un motor térmico y determinar su eficiencia máxima. Carnot sólo incluía el caso de los motores térmicos, pues no había ningún otro tipo de dispositivos o aparatos técnicos cuyo funcionamiento se basara en energetizaciones por calor. Pero en el siglo xx la idea era lograr más eficiencia en el funcionamiento de otros aparatos termodinámicos, como refrigeradores, aires acondicionados, bombas de calor y otros utensilios técnicos, de amplia utilización en la sociedad.

Es evidente que si se conoce teóricamente la eficiencia máxima alcanzable por cualquier aparato, se puede calcular el potencial de ahorro de los energéticos empleados para su funcionamiento. Sólo hay que hacer una comparación entre la eficiencia máxima posible y aquella con la que los aparatos operan.

Sadi Carnot dejó fuera de sus investigaciones el análisis de la eficiencia de una tarea termodinámica, como el calentamiento de interiores.

En esta tarea se trata de la creación de un contraste de temperatura entre el recinto y el ambiente, para lo cual debe aumentarse la energía interna del espacio a "calentar". Otra tarea es el inflado de una llanta, donde se genera un contraste de presión entre ella y la atmósfera. En este caso, el contraste es bárico. Para efectuar las tareas es necesario emplear un dispositivo técnico, que requiere realización de trabajo para su funcionamiento. El calentamiento de interiores se logra con una bomba de calor, que opera con trabajo eléctrico de la red, y la presurización de la llanta se consigue con una bomba de aire, que también puede funcionar con trabajo eléctrico o incluso con trabajo manual.

En el caso del motor térmico, éste puede efectuar trabajo debido a que hay un contraste de temperatura entre una parte caliente, que es la caldera, y una parte fría, que es el condensador. Un aerogenerador puede efectuar trabajo en virtud del contraste de presión entre una porción del aire de la atmósfera y otra; el aire, al pasar de la región de alta a baja presión, puede mover las aspas del aerogenerador y producir una corriente eléctrica, o bien hacerla de molino triturador de algún grano.

Para crear el contraste térmico entre la caldera y el condensador del motor térmico se tuvo que quemar combustible, lo cual implicó, a su vez, la destrucción del *contraste químico* entre dicho combustible y la atmósfera. En la atmósfera, los contrastes de presión entre unas masas de aire y otras se logran por la destrucción de otros contrastes nucleares en el Sol, que se traducen en la emisión de radiación.

En resumen, la satisfacción de tareas implica la creación de contrastes, destruyendo para ello contrastes en otras partes.

6.3.5 La exergía

La exergía es definida como la máxima cantidad de trabajo que se puede obtener de un contraste entre un sistema y su ambiente. El concepto, tal cual, ya lo había definido Clausius en el siglo XIX como *trabajo disponible*. Una tarea termodinámica se puede comprender mejor con los siguientes ejemplos:

a] El calentamiento o enfriamiento de una habitación (tarea térmica).
b] El enfriamiento o calentamiento de un alimento (tarea térmica).
c] La elevación de un peso, por ejemplo de agua a un tinaco o del subsuelo para riego (tarea gravitatoria).
d] La puesta en movimiento de traslación de un objeto, como el caso de un automóvil (tarea cinética de traslación).
e] La puesta en movimiento rotatorio de un objeto, por ejemplo la turbina en una termoeléctrica (tarea cinética de rotación).
f] La elevación de la presión de un gas en un recipiente, en el caso de las bombas de aire (tarea bárica).
g] La carga de una batería (tarea eléctrica).
h] La refinación del petróleo crudo para la obtención de gasolina (tarea química).

En todos estos casos, que corresponden a acciones o tareas necesarias para la satisfacción de necesidades sociales, se produce como consecuencia una diferencia, contraste o desequilibrio entre alguna variable de un objeto y su entorno, que se ha designado por C_Y, en donde el subíndice indica la variable intensiva que cuantifica el tipo de contraste.

Entonces, en cada tarea anterior se crean los contrastes:

a] $C_T = T_2 - T_0$, entre la temperatura de la habitación, T_2, y la temperatura T_0 del ambiente.

b] $C_T = T_3 - T_2$, entre la temperatura T_3 de un alimento dentro del refrigerador, y la temperatura T_2 de la cocina.

c] $C_h = h - 0 = h$, entre la altura del suelo, al que asignamos el valor de $h = 0$, y el nivel h al cual se eleva un objeto.

d] $C_\omega = \omega - 0 = \omega$, entre el objeto que pasa de velocidad de rotación *0* al valor ω.

e] $C_p = p - p_0$, cuando se eleva la presión de un gas de la inicial p_0 a la presión superior p.

f] $C_{fem} = fem - 0 = fem$, al cargar una batería con una carga Z, creándose una fuerza electromotriz *fem* desde el valor *0*.

g] $C\mu = \mu - \mu_0$, al obtener gasolina, con una diferencia de composición química respecto de los gases de la atmósfera, formándose una diferencia de potencial químico, Δμ (véase en un texto de química el concepto de potencial químico).

Aerogenerador | © *Latin Stock México.*

Para cada contraste creado $C_y = y_1 - y_2$ entre un objeto 1 y otro 2 habrá una exergía asociada Ex_y y, por lo tanto, hay potencialmente acumulada entre los dos sistemas la posibilidad de realizar un trabajo.

El máximo trabajo que se puede producir es la exergía, que es una propiedad conjunta de los dos objetos en desequilibrio.

La exergía, a diferencia de la energía, es una propiedad de al menos dos objetos en desequilibrio; también, a diferencia de ella, no se conserva. La exergía desaparece cuando los dos objetos alcanzan el equilibrio termodinámico, mecánico o de otra índole ($C_Y = 0$).

La exergía resulta también una medida cuantitativa (en términos de la posibilidad de realizar trabajo) de cualquier recurso de los llamados "energéticos", por ejemplo, las caídas de agua, la geotermia, los combustibles fósiles, etc. Cada uno de estos energéticos es un objeto en contraste con el ambiente y, por ello, tiene exergía almacenada.

El consumo de los energéticos se mide por el consumo de su exergía en el ambiente. Interesa consumir la menor cantidad de energéticos, es decir, de exergía, por lo que muchos gobiernos y empresas instituyen "comisiones de ahorro de energía".

6.3.6 Ahorro de exergía

¿Tiene significado físico la frase "ahorro de energía"?, o ¿debe decirse "ahorro de exergía"? Tómese en cuenta que, por lo antes dicho, la energía siempre se conserva, pero la exergía se consume.

Hay dos casos de contraste en que ya se conoce el $W_{máx}$ o la exergía, el gravitatorio y el térmico. Las exergías son:

$$Ex_{\text{gravitatoria}} = W_{\max} = mg\,(h_1 - h_2) = mgC_h,$$

$$Ex_{\text{térmica}} = W_{\max} = Q_1\eta_{\max} = Q_1\left(1 - \frac{T_2}{T_1}\right) = \frac{Q_1}{T_1}\,(T_1 - T_2) = \frac{Q_1}{T_1}C_T$$

6.3.7 Eficiencia de la segunda ley de la termodinámica

Lo importante para el ahorro de energéticos es que la realización de cada una de las tareas expuestas, es decir, la creación de un contraste y su exergía, demanda la destrucción de otros contrastes y, por lo tanto, de otras exergías existentes entre otros objetos. Se define la eficiencia de una tarea, también llamada *eficiencia de la segunda ley de la termodinámica*, denotada por la letra griega ε (epsilon), como:

$$\varepsilon = \frac{\text{Ex creada en la tarea}}{\text{Ex destruida al realizar la tarea}}$$

Si los contrastes (exergías) que se crean son de la misma naturaleza que los contrastes (exergías) que se destruyen, siempre se ahorran energéticos.

Por ejemplo, la realización de la tarea mecánica de elevar un peso (creación de un C_h y una Ex_h), se puede efectuar de varias maneras: la mejor sería bajar otro peso (destrucción de otro C_h y otra Ex_h) empleando una palanca (véase figura 27), o poniendo a operar un motor eléctrico funcionando con energía eléctrica generada por medios también mecánicos: caídas de agua (destrucción de otro C_h y otra Ex_h), aerogeneradores (destrucción de un C_p y una Ex_p).

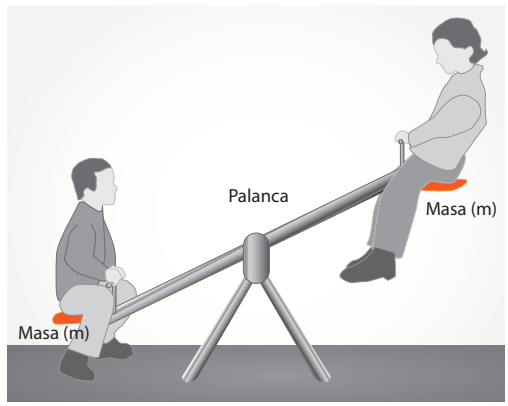


Figura 27. Palanca.

Cuando la exergía creada en la tarea es de la misma naturaleza que la exergía destruida, la eficiencia toma su valor máximo, que es 1.

Sin embargo, con frecuencia se emplean exergías de naturaleza diferente a la que exige la tarea; en el caso que se discute, empleando un motor eléctrico alimentado con la electricidad generada en una termoeléctrica. La exergía consumida para la generación de electricidad proviene del contraste químico del combustible con el ambiente. Si a ésta sumamos la exergía perdida en la transmisión de la corriente eléctrica y las pérdidas en la transformación a exergía gravitatoria, resulta que la *Ex* consumida total es mucho mayor que la *Ex* mínima requerida para efectuar la tarea. De este modo, resulta que $\varepsilon < 1$.

Otro ejemplo: la tarea de elevar la temperatura de una habitación de T_0 a T_2 (creación del contraste térmico $C_T = T_2 - T_0$) se cumple incrementando, por ejemplo por calor Q_2, la energía interna de la habitación hasta llegar a la temperatura confortable T_2.

La cantidad mínima de trabajo (igual a la exergía mínima creada) que hay que invertir en esta tarea se puede calcular como sigue: el trabajo mínimo es igual, a su vez, al máximo trabajo que se puede obtener de Q_2, procesado de una caldera a la temperatura T_2 frente al condensador a temperatura T_0. Sabemos que este trabajo máximo es $Q_1/T_2\ C_T$, de modo que el numerador en la última ecuación es esta cantidad, entonces:

$$\varepsilon = \frac{Q_2}{T_2}\frac{C_T}{\text{Ex consumida}}.$$

La cantidad Q_2 de calor mínima necesaria para elevar la temperatura del cuarto se calcula, como antes se hizo, tomando en cuenta la capacidad térmica del aire (por ejemplo a presión constante) y el incremento de temperatura propuesto, así:

$$Q_2 = C_P \Delta T = C_P\,(T_2 - T_0) = C_P C_T.$$

Q_2 es la cantidad mínima de calor necesaria para calentar el cuarto en ΔT, porque el aire, a su vez, pasa energía a los demás objetos del cuarto. Por simplicidad, al no tomar este efecto en cuenta queda:

$$\varepsilon = \frac{\frac{C_P C_T^2}{T_2}}{\text{Ex consumida}}.$$

El valor final de ε depende de la exergía consumida, que es mucho menor empleando una bomba de calor que un calentador de resistencia, y aún menor si se consume la exergía de la radiación solar.

6.3.8 Desarrollo sustentable

Es posible definir cuantitativamente el *desperdicio* que se tiene al realizar tareas termodinámicas; este desperdicio o basura tiene que ver con la utilización inadecuada de la exergía propia para la realización de la tarea en cuestión. Si se consume otro tipo de exergía, habrá un desaprovechamiento y, al final, los efectos se manifiestan como contaminación.

El desperdicio de los recursos energéticos es una medida de la contaminación debida a su empleo. Puede reducirse y aun eliminarse en los casos en que la *Ex* mínima requerida para la tarea es la misma que la consumida. De lo contrario la diferencia es el desperdicio (o basura).

Basura, contaminación o desperdicio = Ex consumida – Ex requerida.

La *Ex* consumida es basura química si se consume un recurso químico, como el carbón o el petróleo; es basura radiactiva si se consume un recurso nuclear, como el uranio o el torio; es basura térmica si proviene de una caldera. Por esta razón, para evitar la contaminación, se deben tratar de cumplir las necesidades termodinámicas evitando el consumo de estos tipos de exergías. ¿Es posible satisfacer todas las necesidades de una sociedad solamente con fuentes de exergía mecánicas? Si pensamos en el caso del calentamiento de agua, concluimos que la tarea se puede cumplir con la exergía gravitatoria de las presas hidroeléctricas o con la exergía del viento, pero la más adecuada sería la exergía de la radiación solar.

Las exergías acumuladas en el viento, las corrientes, las presas, son las más limpias en su consumo, siempre y cuando su utilización cumpla con algunas condiciones. Por ejemplo, las presas no deben alterar la biodiversidad del entorno de manera irreversible, ni deben competir con el empleo de la tierra para la agricultura o implicar el desplazamiento de una población a la que no se le ofrecen alternativas apropiadas. En fin, la satisfacción de tareas termodinámicas no es una cuestión simple de resolver y, en última instancia, debe sujetarse a normas compatibles con el llamado *desarrollo sustentable*.

Antes de ver el concepto de desarrollo sustentable, conviene considerar otro aspecto de la energía, que tiene que ver con su degradación. Esto es, aunque la energía es una cantidad que se conserva cuando un objeto experimenta un proceso, puede suceder que se degrade. El concepto de degradación se refiere a la pérdida de la capacidad para realizar trabajo sobre el exterior del objeto. Si se considera un sistema aislado del exterior, en el estado inicial (i), compuesto de dos partes separadas por una pared removible, una con gas y la otra al vacío, y se elimina la pared, el gas se precipita sobre el vacío; después de un tiempo se llega al equilibrio, en el estado final (f). La primera ley de la termodinámica, aplicada al sistema que pasa del estado inicial al final, establece que:

$$\Delta U = U_f - U_i = Q - W = 0 - 0 = 0,$$

por lo que

$$U_f = U_i.$$

La energía interna del gas es la misma en el estado inicial que en el final. Pero la gran diferencia es que en el estado inicial *i*, el sistema puede realizar un trabajo sobre el exterior, en tanto que en el estado final *f*, tal posibilidad ha desaparecido, se ha destruido o consumido. En otras palabras, en *i*, el sistema posee exergía, pero en *f* se ha consumido.

En resumen, se puede decir que mientras la energía se conserva, la exergía se destruye. Pero también se puede establecer que del estado *i* al estado *f* la energía se ha degradado. Esta última afirmación constituye el llamado *principio de degradación de la energía*.

Suele decirse que "energía es la capacidad para realizar un trabajo", pero el ejemplo demuestra que tal afirmación no es válida; sólo es válida para la exergía. La energía inicial es igual a la final, pero en la primer situación tiene capacidad para realizar trabajo (exergía), mientras que en la segunda no.

En 1987 la Comisión Internacional sobre Ambiente y Desarrollo de la Organización de las Naciones Unidas definió como *desarrollo sustentable* el que "satisface las necesidades y aspiraciones del presente sin comprometer la capacidad para satisfacer las del futuro".

Para conseguir la sustentabilidad es necesario que la contaminación exagerada que el planeta sufre actualmente sea drásticamente reducida, lo que implica un cambio en la forma como se procesan los recursos minerales, energéticos, nutrientes (nitratos y fosfatos, entre otros) y el agua.

6.3.9 Uso lineal y cíclico de los recursos exergéticos

Para entender el meollo de la sustentabilidad, supóngase que un estudiante, al inicio de sus estudios, es sorprendido por dos noticias: la mala es que el pariente cercano, que le apoyaba sistemáticamente con los gastos de manutención y escolares de sus estudios, fallece; la buena es que le deja una cantidad que, como inversión en manos de una institución financiera, le genera intereses suficientes para seguir cubriendo los gastos que anteriormente se le asignaba. Ante el estudiante se abren dos opciones: la sustentable y la no sustentable. En la primera adapta el gasto para la satisfacción de sus necesidades a los intereses que recibe por su dinero. En la opción no sustentable (o insostenible) adquiere necesidades que implican gastos mayores a los intereses y comienza a extraer dinero del capital; al cabo de un tiempo su situación es insostenible.

En la figura 28 se muestra el paradigma actual hegemónico de la humanidad, en que se vive del capital y no de los intereses. El capital fijo (el depósito B) son los combustibles fósiles (carbón y petróleo) y el uranio, así como los depósitos de minerales; es decir, B son los recursos agotables, mineros y energéticos.

El capital fluyente (los intereses) está constituido por los ciclos de nutrientes (como el nitrógeno y el fósforo), agua y energéticos inagotables (o renovables) movidos por el Sol (es el ciclo A, que se ve muy deteriorado por la contaminación). La humanidad afecta a A, por el uso masivo de energéticos agotables.

El paradigma es *lineal* en la utilización de los recursos porque las actividades humanas, representadas por la persona de la figura, recogen recursos del capital fijo (2) y del variable (3), los procesan y los devuelven al ambiente (1) en forma de contaminantes gaseosos (5), líquidos y sólidos (4). A este esquema de insostenibilidad, se opone el paradigma de la utilización *cíclica* de los recursos, como se indica en la figura 29.

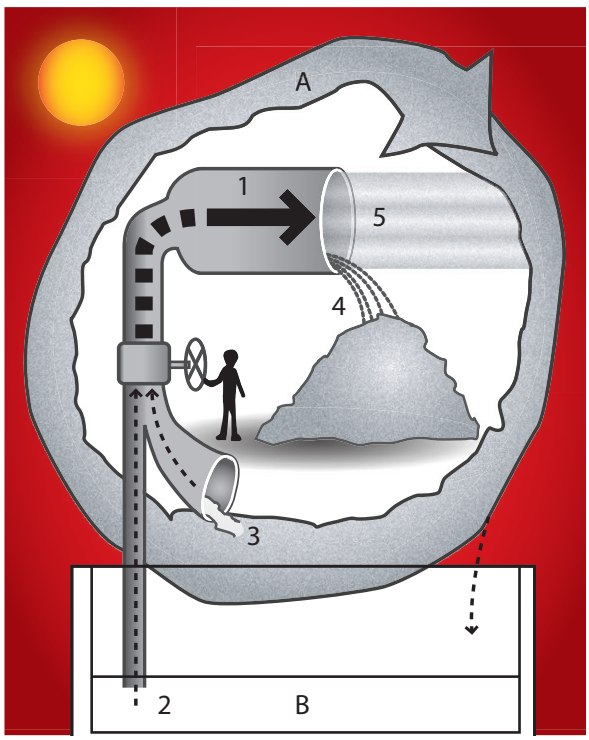


Figura 28. Vivir del capital o el paradigma de la utilización lineal de los recursos.

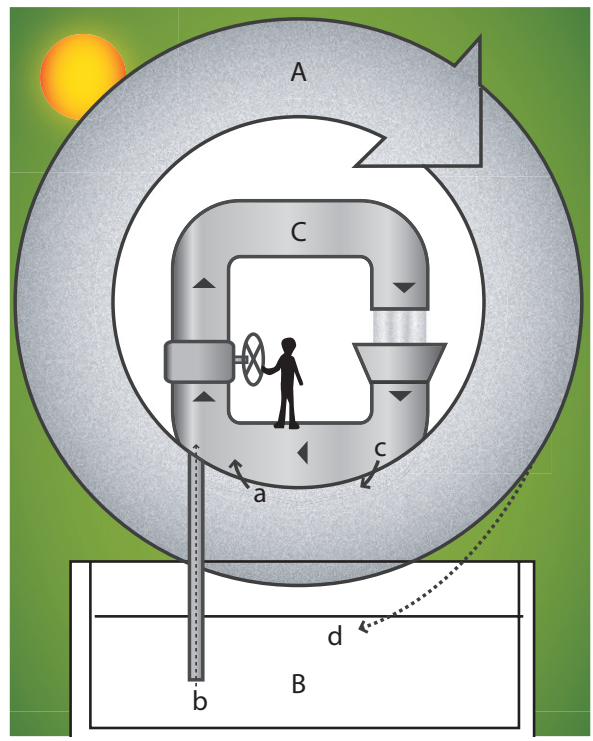


Figura 29. Vivir de los intereses o el paradigma de la utilización circular de los recursos.

En este paradigma la actividad humana se "engarza" en el punto (a) con los ciclos naturales-solares de circulación de nutrientes, agua y energéticos (A) que, al ser reciclados en C luego de ser utilizados, regresan a A en (c), sin dañarlo. El uso que se hace de los recursos agotables de B es mínimo. De este modo se logra que la humanidad tenga un modo sostenible de subsistencia con el ambiente. La satisfacción de las necesidades humanas es resuelta con el consumo de exergías suministradas en los accesos (a), provenientes de las exergías naturales A creadas a diario por el Sol.

Un ejemplo es la utilización de la exergía del agua: el Sol evapora el agua del suelo y lo eleva a la atmósfera. Por lluvia el agua se deposita a una altura superior a la del suelo, adquiriendo exergía hidráulica. Ésta es almacenada en una presa y es utilizada para el riego, o para generar electricidad, volviendo después a niveles inferiores, de donde el Sol vuelve a evaporarla y así sucesivamente. Sin embargo, si el agua regresa al mar o a los lagos con una carga excesiva de nutrientes químicos, provenientes de la fertilización de los suelos agrícolas o de contaminantes de los herbicidas e insecticidas, el ciclo no se cierra limpiamente y se ve maltrecho, como en la figura 28.

6.3.10 Huella ecológica

Otra forma práctica de especificar científicamente la sustentabilidad de un conglomerado humano en la superficie del planeta es mediante el concepto de *huella ecológica.* Ésta se establece, como se hizo por primera vez en el municipio (comuna) de Malmöhus, en Suecia, al comparar los recursos consumidos y el desperdicio producido en la satisfacción de las necesidades del municipio, con la capacidad para absorber estas demandas y regenerarse, manteniendo un mínimo de 12% de la superficie del municipio para preservar la biodiversidad natural.

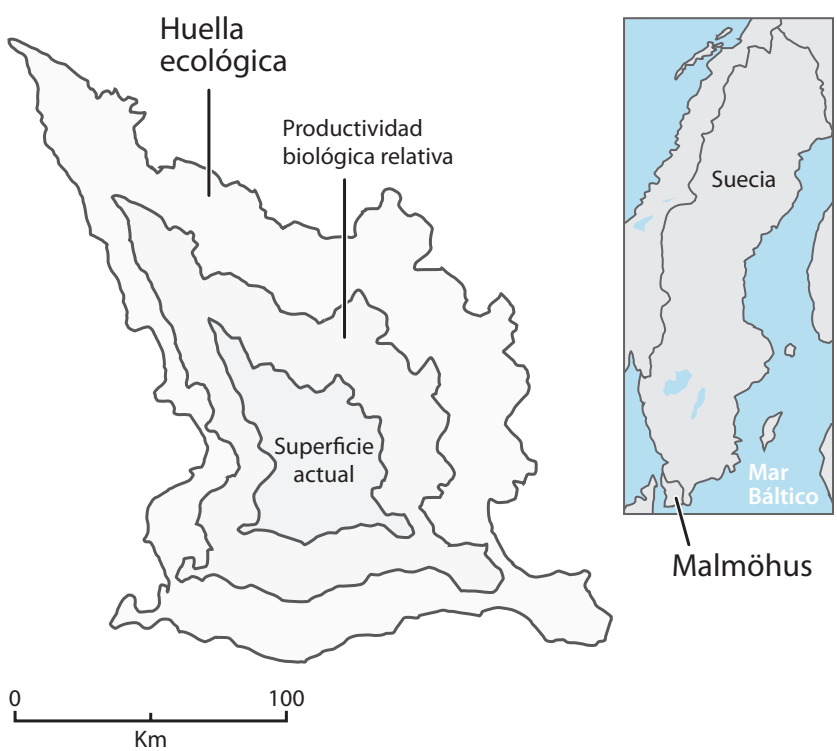


Figura 30. La huella ecológica del municipio sueco de Malmöhus.

El concepto así definido es verdaderamente multidisciplinario, ya que integra conocimientos de física, biología, química, etc. Esto es debido a lo complejo del sistema sociedad-ambiente, que es una unidad indisoluble. En este sentido, la huella ecológica ofrece un método de evaluación de la capacidad del conglomerado humano para sostener un estilo de vida, compatible con un estado saludable de plantas y animales del ambiente, acorde con el paradigma de la utilización circular de los recursos.

La huella incluye el consumo de energía, alimentos, vestido y otras necesidades de la población, así como los desperdicios que se generan. Los desperdicios son principalmente dióxido de carbono, absorbido por el crecimiento de bosques permanentes y nitratos y fosfatos de los drenajes y residuos de los fertilizantes agrícolas, filtrados por los humedales y asimilados por plantas y microorganismos que protegen los suministros de agua dulce y los ecosistemas acuáticos. La huella, igualmente, toma en cuenta el balance entre la importación y exportación de consumibles y el desperdicio producido por ellos, de manera que incorpora el uso de tierra y océano fuera de las fronteras de Malmöhus.

El número de hectáreas del municipio entre el número de habitantes es de 1.2 hectáreas, y son tan productivas biológicamente que corresponden a 3.4 hectáreas de tierra de productividad global promedio. Cuando se quita el 12% de esta superficie para mantener la biodiversidad, quedan 3.0 hectáreas de tierra de productividad global promedio disponibles para cada habitante.

Sin embargo, la huella ecológica sueca es de 7.2 hectáreas de tierra de productividad global promedio, de modo que superficie de otras partes debe utilizarse para compensar la diferencia de 4.2 hectáreas por cada persona de Malmöhus. Esta diferencia representa el déficit ecológico de Malmöhus, como se expresa en la figura 30 (p. 465). Afortunadamente para los suecos, 8.2 hectáreas de tierra de productividad global promedio por persona están disponibles después de sustraer 12% para preservar la diversidad biológica. No obstante, si todas las naciones adoptaran el estilo de vida sueco de 1997, la capacidad de la Tierra sería excedida por un factor de tres.

Las estimaciones mundiales revelan que la huella ecológica actual está excedida en una tercera parte de la superficie de la Tierra.

6.3.11 Consumo de recursos energéticos agotables

La gran dependencia actual de los recursos energéticos agotables (es decir, de los recursos fósiles provenientes del depósito B de la figura 28 (p. 464) es una manifestación de que el paradigma insostenible está en la actualidad firmemente asentado en nuestra civilización. A continuación se analiza la secuencia temporal de la consolidación del paradigma de la utilización lineal de los recursos energéticos.

Alrededor de 1880 se extrajeron unas cuantas toneladas de petróleo; para el año 1950 llegaron a ser 50 millones de toneladas. A partir de entonces el crecimiento de la producción ha sido casi exponencial, hasta alcanzar poco más de 3 500 millones de toneladas en el año 2000. Además de la contaminación que el uso del petróleo produce, está el problema de su agotamiento. Si no hay un cambio pronto, el peligro de guerras por este energético estará presente.

Se trata de un escenario insostenible. A la tasa de explotación del año 2000, sólo quedaría petróleo para 38.4 años y gas para 62. Aunque el carbón alcanzaría para unos 227 años, al acabarse el petróleo y el gas, su tasa de explotación aumentaría y duraría mucho menos. Además, el carbón produce más gases de efecto invernadero que el petróleo y el gas.

La sujeción al paradigma de la utilización lineal de los recursos es la responsable de que la huella ecológica mundial exceda la superficie del planeta en una tercera parte. La huella ecológica promedio mundial es de 2.8 hectáreas por persona, en tanto que la biocapacidad disponible en hectáreas por persona es de 2.1.

Sin embargo, esta huella está muy diferenciada entre naciones. Por ejemplo, en 1997, Canadá utilizó por persona 7.0 hectáreas de tierra productiva promediada globalmente y 0.7 hectáreas de océano productivo, para un total de 7.7 hectáreas por persona, mientras que su sustentabilidad disponible es de 9.6 hectáreas por persona. En Estados Unidos, en cambio, la persona promedio requirió de 10.3 hectáreas, pero su sustentabilidad es de 6.7 hectáreas por persona. Así, Canadá tiene un margen positivo en sustentabilidad de 1.9 hectáreas por persona, en tanto que Estados Unidos sobrevive con un déficit de 3.6 hectáreas por persona, que explota de los demás.

En México también somos deficitarios, es decir, nuestro modo de vida es no sustentable, ya que nuestra huella ecológica es de 2.6 hectáreas por persona, en tanto que nuestra biocapacidad disponible es de apenas 1.4 hectáreas por persona.

6.3.12 Los energéticos renovables

Los energéticos renovables provenientes directa o indirectamente del Sol son los intereses o capital fluyente generados por el gran astro luminoso. La ventaja sobre las otras opciones es que no contaminan tanto, son potencialmente inagotables y además utilizables no sólo en la generación de electricidad, sino en el transporte, la industria y la agricultura. Su difusión masiva tendría un efecto amplio en la disminución drástica de los gases de efecto invernadero y, sin duda, implicarían una gran reducción de la huella ecológica planetaria.

El ritmo de crecimiento de la capacidad instalada de energéticos renovables generadores de electricidad presenta un panorama optimista, tanto que, de desarrollarse a ritmos tan acelerados, puede significar el abandono de los combustibles fósiles en un futuro no muy lejano. En la tabla siguiente se muestran las contribuciones de los energéticos renovables en la generación de electricidad para los años 2005, 2006 y 2007, en GW, en el mundo. Recuérdense dos aspectos:

1] que 1 GW de capacidad instalada es aproximadamente la capacidad de nuestra hidroeléctrica más grande, Chicoasén, en Chiapas;

2] la capacidad total eléctrica instalada en el mundo en 2003 era de 3 641.3 GW, repartidos así: 2 469.9 GW (68 %), en termoeléctricas de combustibles fósiles; 368.5 GW (10 %) en nucleoeléctricas; 739.8 GW (20 %) en hidroeléctricas).

Energético renovable	*2005*	*2006*	*2007 (estimado)*
Capacidad total instalada (excluye hidro)	182	207	240
Capacidad total instalada (incluye hidro)	930	970	1010
Capacidad eólica instalada	59	74	95
Fotovoltaica conectada a la red eléctrica	3.5	5.1	7.8
Producción fotovoltaica	1.8	2.5	3.8

A los valores de esta tabla se pueden agregar, al menos para el año 2006, otras fuentes renovables e inagotables (como la geotermia) de energéticos generadores de electri-

cidad (también en GW): pequeñas hidroeléctricas, 73; biomasa, 45; geotérmica, 9.5; solar térmica, 0.4; energía de mareas, 0.3. Con éstas y las de la tabla, el total de renovables y geotermia es de 207 GW. Las grandes hidroeléctricas contribuyen con 770 y la capacidad eléctrica total mundial, para ese año de 2006, fue de 4300 GW.

La energía eólica ha sido la de mayor crecimiento, en proporción, en cuanto a capacidad instalada, como se muestra en la tabla anterior. Se espera que tal impulso se mantenga, pues se supuso un avance de unos 15 GW instalados anualmente hasta el final de 2006, a 33.5 GW anuales para 2011. Aunque la proporción de la capacidad instalada de renovables respecto al total es de una cuarta parte si se incluyen las grandes hidroeléctricas, el ritmo de avance es muy rápido y sostenido.

Desde el decenio de los setenta varios grupos e investigadores propusieron que las grandes reservas probadas de hidrocarburos de México (unos 70 000 millones de barriles de petróleo) podrían ser utilizadas para financiar y apoyar energéticamente la transición mexicana a las fuentes inagotables de energéticos, desarrollando para ello, nacionalmente, las técnicas correspondientes.

Las técnicas de los energéticos eólicos y fotovoltaicos son desarrolladas por los países altamente industrializados, entre los que aparecen China e India, pero no México, por desgracia. Se perdió así una oportunidad especial para acceder a la independencia técnica en energéticos. Ni China ni India tienen suficiente petróleo que les sirva de apoyo a una transición a los energéticos inagotables, como se propuso para nuestro país, pero sí cuentan con una política científica y técnica decidida que los tiene en el lugar que ocupan. Esto lo podría lograr México, aun sin petróleo, de contar con una política científica, técnica y energética pertinente.

En una opción, las instituciones de educación superior e investigación científica (universidades y tecnológicos) bien podrían intentar desarrollar prototipos técnicos energéticos y difundirlos socialmente, bien a través de empresas propias, en asociación con gobiernos y empresarios o con cooperativas.

6.3.13 Ecoaldeas y ecomunicipios

Desde hace cuatro decenios se han difundido en varios países el concepto y la construcción de ecoaldeas y ecomunicipios. Entre estos intentos resalta por su perseverancia y claridad de planteamiento el caso sueco, en el que explícitamente se busca poner en práctica, como paradigma de vida, la utilización cíclica de los recursos exergéticos.

A pesar de los años transcurridos, los habitantes de las ecoaldeas aún en funciones reconocen no haber llegado a la meta; sin embargo, persisten en ello por los logros alcanzados, en que la solidaridad es una ganancia efectiva y enriquecedora.

Al parecer, en las ecoaldeas se pone de manifiesto que las relaciones armónicas entre la sociedad y el ambiente conducen a relaciones humanas armónicas en el interior de la sociedad.

Una de las ideas centrales del paradigma de la utilización circular de los recursos es el reciclaje de nutrientes, agua y energía. Su consecución requiere de la efectiva cooperación entre los habitantes de la ecoaldea. En el sitio de internet de Tuggelite, el lector encontrará la historia del movimiento de ecoaldeas y ecomunicipios (en Suecia, la comuna es el equivalente del municipio de nuestro país).

Se sugiere que una de las primeras acciones para transformar cualquier municipio en un municipio ecológico sea la de reciclar la basura en una planta construida ex profeso, con biodigestor incorporado.

Las consecuencias serían inmediatas: educación ecológica de los habitantes, ofrecimiento costeable de abono orgánico a los agricultores de la región, producción de gas metano para estufas domésticas y, eventualmente, para calentadores de agua. En realidad, habría que analizar la factibilidad del establecimiento del paradigma de la utilización circular de los recursos en varios niveles de acción. Los niveles tienen dos componentes: espacial geográfica (e_i) y temporal (t_i).

El componente espacial tiene cinco instancias:

a] hogar;
b] aldea, pueblo o municipio;
c] región;
d] país;
e] planeta.

El componente temporal tiene tres:

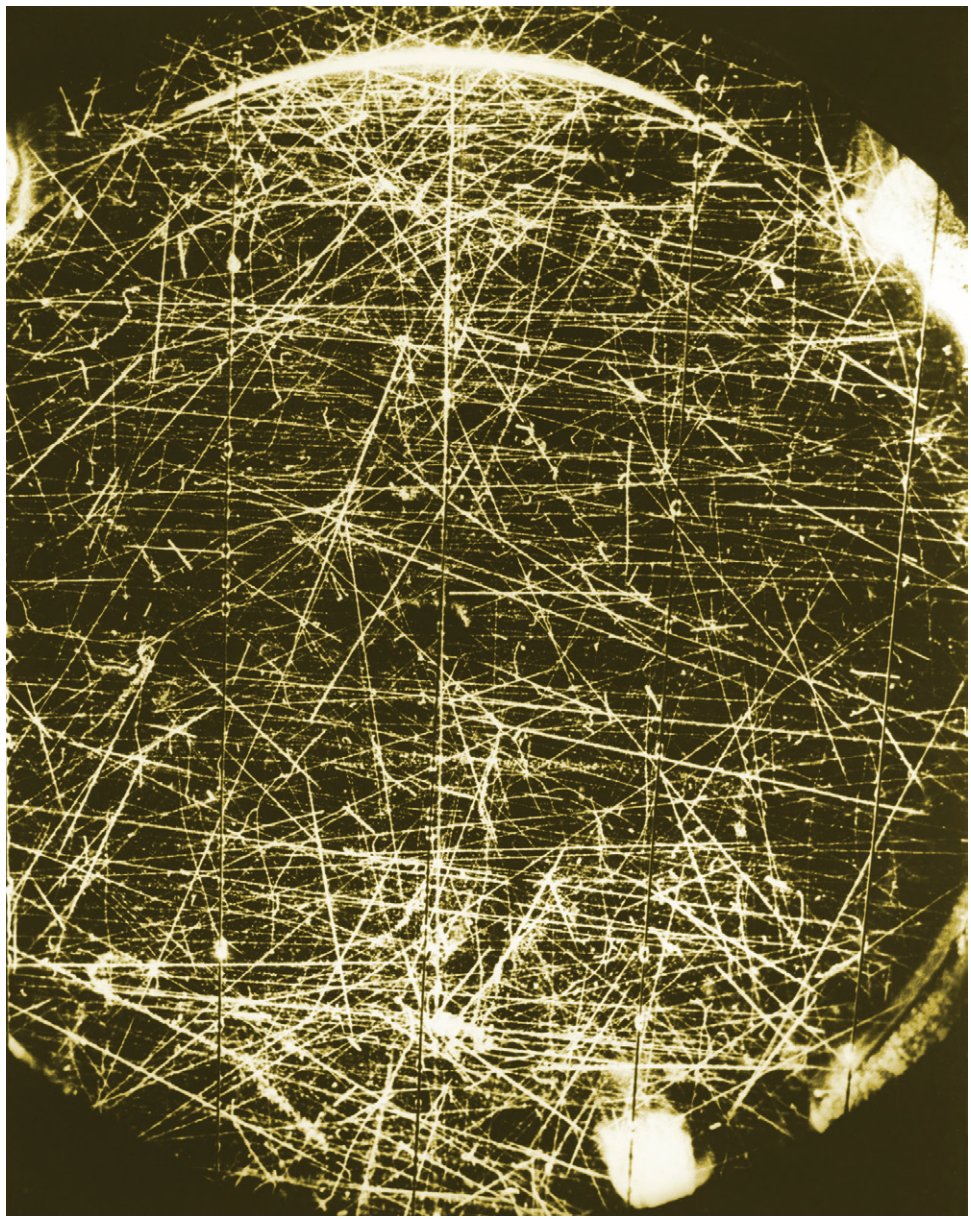
a] corto;
b] mediano;
c] largo plazo.

De ambos componentes se definen 15 niveles de acción hacia la sustentabilidad, los cuales están interconectados, pues no se puede tener un planeta sustentable a largo plazo si no se consigue que todos los hogares lo sean a corto plazo. Algo parecido se puede decir de otra manera: "pensar globalmente y actuar localmente".

LO MÁS PEQUEÑO

TEMA

7



© *Latin Stock México.*

INTRODUCCIÓN

A Demócrito (470-380 a.n.e.) le llamaban el filósofo risueño, por su eterna sonrisa. Nació en la ciudad griega de Abdera. Sus conciudadanos parecían calificar sus ideas y forma de vida como expresiones de locura, pues se preocupaba por cosas que, para quienes le rodeaban, no tenían sentido, como, por ejemplo, "hasta dónde podría dividirse una gota de agua".

Demócrito pensaba que se podrían obtener gotas cada vez más pequeñas, hasta casi perderlas de vista. Pero Leucipo (~450 a.n.e.), maestro de Demócrito, había intuido que

esa división tenía un límite. Como Demócrito hizo suya esa idea, enunció que cualquier sustancia podía dividirse hasta ese límite y no más.

Al trozo más pequeño —o partícula— indivisible de cualquier material lo llamó *átomo*. Según Demócrito, todo el Universo estaba constituido por esas partículas *indivisibles* y entre ellas no había nada, es decir, había espacio vacío.

La mayoría de los filósofos griegos rechazaron la idea del *átomo* pues la consideraban absurda. Y aunque de todos los libros que escribió ninguno se conserva, algunos de los pensadores que aceptaron la idea de la partícula indivisible fundaron escuelas de importancia que preservaron su pensamiento hasta nuestros días.

Epicuro (342-272 a.n.e.) fundó una escuela de gran popularidad y desarrolló la corriente filosófica conocida como epicureismo, que permaneció por más de 700 años, hasta que la avalancha del cristianismo barrió las escuelas de pensamiento "pagano". La escuela de Epicuro era mecanicista y consideraba al placer como el don humano más importante. Adoptó el átomo de Demócrito como una explicación satisfactoria de la estructura del Universo. Estas ideas fueron resumidas por pensadores como Aristóteles (384-322 a.n.e.):

> Se le llama Elemento a la materia primitiva que entra en la composición de los objetos, y que no puede ser dividida en partes heterogéneas […] Los que tratan los elementos de los cuerpos, dan también este nombre a las últimas partes que no se pueden dividir en otros cuerpos de especies diferentes. Esto es lo que llaman ellos elementos, ya admitan un solo elemento, ya admitan muchos.[1]

A pesar de que del legado de Demócrito casi nada sobrevive, prevaleció el tiempo suficiente para poder influir sobre pensadores romanos como Lucrecio (95 a.n.e. a 55 a.n.e.). En los tiempos antiguos, los libros se copiaban a mano, por lo que las grandes obras se podían confeccionar en unos pocos ejemplares y sólo eran accesibles a personas o instituciones económicamente poderosas.

La invención de la imprenta, hacia el año 1450 de n.e., hizo un gran cambio. Fue posible entonces contar con grandes tirajes. Uno de los primeros libros considerados para esto fue *Sobre la naturaleza de las cosas*, de Lucrecio. Las ideas de Demócrito habían encontrado el camino hacia las nuevas generaciones:

> Llamamos elementos a los cuerpos primeros y compuestos a los que resultan de ellos. Los elementos son indestructibles, porque su solidez triunfa de todo […] los principios que componen el gran todo creado tienen un cuerpo sólido y eterno […] No puede disolverlos choque externo, ni puede penetrar extraña fuerza a su tejido; ni de acción extraña puede recibir daño, como he dicho […] Si no fuesen eternos, a la nada todo el mundo se hubiera reducido […] luego, los principios la simplicidad sólida contienen, porque sin ella no hubieran podido durante tantos siglos conservarse […] Como un cuerpo más pronto se destruya que lo que tarda el mismo en rehacerse, las pérdidas que hubiera padecido en la edad precedente, irreparables fueran sin duda alguna en las siguientes […] La división de la materia tiene límites invariables y precisos.[2]

No fue sino hasta el siglo XVII cuando el filósofo francés Pierre Gassendi (1592-1655) se pronunció como epicúreo, defendiendo la teoría de las partículas indivisibles. En 1660,

[1] Aristóteles, *Metafísica* (libro V), México, Porrúa, 1987, p. 78.
[2] Lucrecio, *De la naturaleza de las cosas*, Buenos Aires, Orbis, 1984, pp. 111-113.

el físico inglés Robert Boyle (1627-1691), discípulo de Gassendi, hizo estudios sobre las características del aire, observando que se le puede comprimir hasta un cierto límite. Concibió la idea de que el aire está compuesto de partículas minúsculas que dejaban grandes espacios entre ellas.

Comprimir el aire —reflexionaba el científico inglés— equivale a juntar más las partículas reduciendo el espacio vacío que hay entre ellas. El agua podría consistir, entonces, de partículas tan juntas que no se las puede acercar más; por lo tanto, su conclusión era que no es posible comprimirla. Al separarle las partículas, el agua se convierte en vapor, sustancia parecida al aire.

7.1 CONCEPCIÓN DE ÁTOMO

En 1756 Benjamin Franklin intuía que la materia estaba formada de gránulos portadores de cargas eléctricas. Al contemplar el fenómeno de la inducción electrostática, afirmaba: "La materia eléctrica consiste en partículas extremadamente sutiles, capaces de atravesar la materia ordinaria, incluso la más densa, con tal libertad y facilidad que no encuentra la menor resistencia."

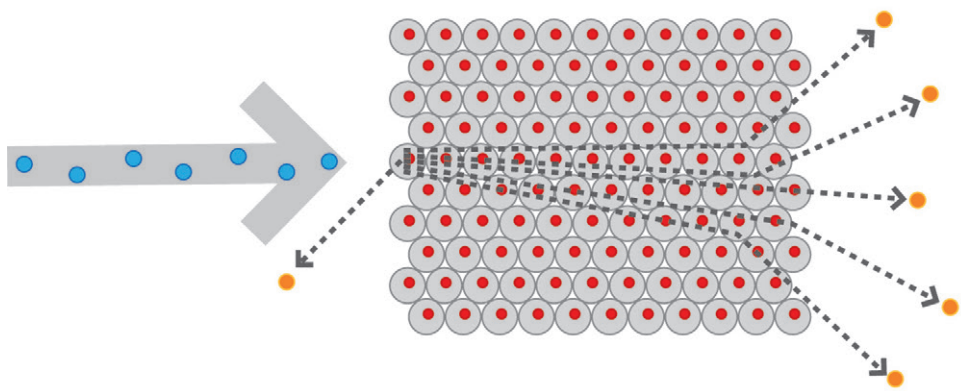


Figura 1. Haz de partículas interactuando.

Sin embargo, opiniones de este tipo eran marginales, incapaces de conmover a la comunidad científica de la época. Claude Louis Berthollet (1748-1822) consideraba que una reacción química dependía de las cantidades de sustancias que reaccionaban y que éstas, a su vez, actuaban sobre la velocidad de la reacción y sobre la naturaleza del compuesto final.

Realizando elaborados experimentos, Joseph Louis Proust (1754-1826) demostró que el carbonato de cobre contenía proporciones fijas en peso de carbono, oxígeno y cobre, sin influir el método de preparación en el laboratorio o de la obtención de los elementos. La proporción siempre era cinco partes de cobre, cuatro de oxígeno y una de carbono.[3] Esto lo encontró también para otros compuestos, concluyendo que: "Todo compuesto contiene sus elementos en proporciones definidas, sin influir en absoluto su modo de obtención".

John Dalton (1766-1844) trabajó las propiedades de los gases, aceptando las teorías de Boyle e Isaac Newton (1643-1727) de que éstos están formados por partículas. No obstante, Dalton fue mucho más lejos, diciendo que no sólo los gases están constituidos por estas pequeñas partículas sino todos los estados de la materia. Señaló que la *ley de propor-*

[3] Joseph Louis Proust, "Investigaciones sobre el cobre", *Ann. chim.* 32, pp. 26-54 (1799).

ciones definidas de Proust se explica fácilmente al suponer que cada compuesto está formado por partículas indivisibles:

> Tres clases hay de cuerpos o tres estados de los cuerpos, que de manera especial han llamado la atención de los químicos filósofos: a saber, los denominados fluidos elásticos, líquidos y sólidos.
>
> En el agua tenemos un caso conocidísimo de un cuerpo que en ciertas circunstancias puede adquirir cualquiera de dichos tres estados. En el vapor hallamos un fluido perfectamente elástico, en el agua un líquido perfecto y en el hielo un sólido cabal.
>
> Estas observaciones han llevado tácitamente a la conclusión, al parecer universalmente aceptada, de que todos los cuerpos de magnitud sensible, ya fueran sólidos o líquidos, están constituidos por un inmenso número de partículas en extremo pequeñas, o átomos de materia, unidos entre sí por la fuerza de la atracción; la cual es más o menos poderosa, según las circunstancias.[4]

Hielo, agua y vapor | *© Latin Stock México.*

Dalton reconoció la similitud existente entre sus teorías y las que había enunciado Demócrito, más de 2 mil años antes, por lo que llamó a estas partículas átomos. Sostuvo que todos los elementos están formados por átomos pequeñísimos, indivisibles e indestructibles, y que todas las sustancias conocidas están formadas por distintas combinaciones de dichos *átomos*.

Una sustancia —decía Dalton— se puede convertir en otra al deshacer su combinación específica de átomos y formar una nueva y distinta. Los átomos de un mismo elemento son exactamente iguales —agregaba—, aunque diferentes a los de otro elemento.

7.1.1 Rayos catódicos

En 1853, un desconocido científico francés llamado Masson hizo saltar una chispa eléctrica desde una bobina de inducción de alto voltaje a través de un tubo cerrado de vidrio, al cual se le había extraído el aire, y descubrió que en lugar de la típica chispa que se observa en el aire, el tubo se llenaba de una luminosidad brillante.

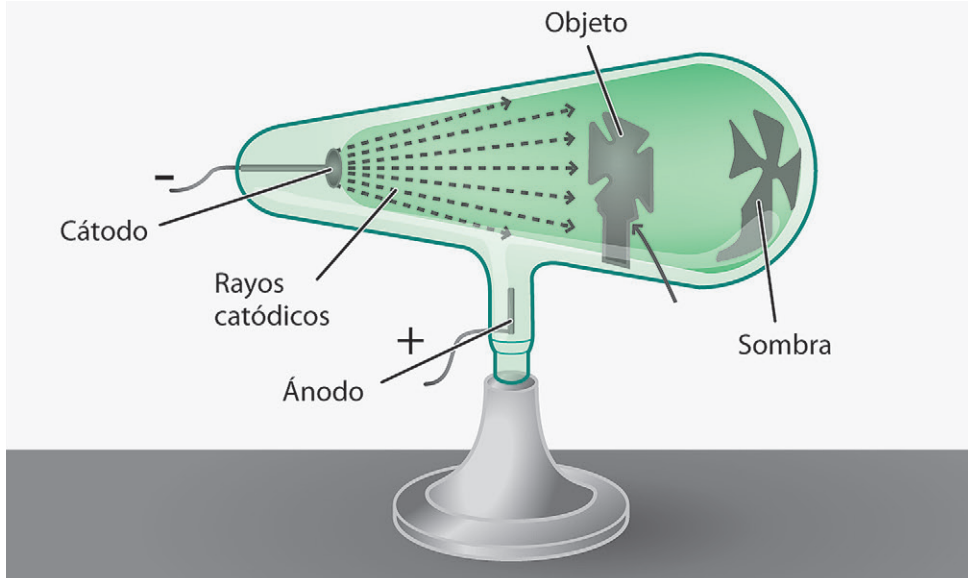
Algunos años más tarde, el alemán Heinrich Geissler (1814-1879), un soplador de vidrio en Tubinga, desarrolló y empezó a fabricar tubos de descarga gaseosa, similares a los modernos tubos de neón y argón utilizados en publicidad. Pronto se descubrió que la luminiscencia verde en el tubo, lograda a presiones internas muy bajas, se debía a que rayos invisibles que surgían del polo negativo (cátodo) hacían brillar al vidrio, por lo que se les llamó *rayos catódicos*.

En 1869, Johann Wilhem Hittorf (1824-1914) introdujo objetos en el camino de la descarga. Allí donde los rayos inciden directamente en las paredes del tubo, el vidrio adquiría un brillo verde, mientras que en donde no golpean permanecía oscuro. Observando que la sombra definía claramente el perfil del objeto seleccionado, concluyó que los rayos catódicos *se mueven en línea recta* (véase figura 2, p. 474).

William Crookes (1832-1919), en 1870, introdujo en el tubo un pequeño molinete movible en el camino de los rayos. Al encender el tubo, el molinete empezaba a rodar hasta llegar al otro extremo del tubo. Crookes explicó este fenómeno diciendo que son los rayos

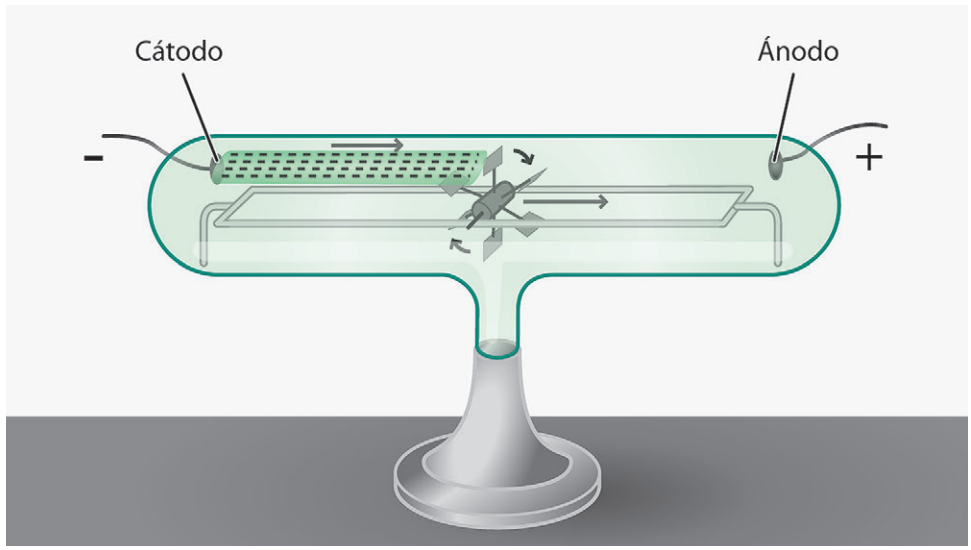
[4] John Dalton, *Un nuevo sistema de filosofía química*, Manchester, Londres, 1808.

Figura 2. Tubo de rayos catódicos.



los que, al chocar con las paletas de plástico, hacían rodar al molinete. Crookes concluyó de sus observaciones que los rayos catódicos *tienen cantidad de movimiento (ímpetu)*, lo que Newton llamaba *momento*, y por lo tanto tienen masa, velocidad y energía cinética (figura 3).

Figura 3. Tubo de rayos catódicos con molinete.



En 1895, Jean Baptiste Perrin (1870-1942) montó un aparato con el que logró obtener un haz fino, haciéndolo visible al permitirle chocar contra una tira metálica pintada con un pigmento fluorescente, sulfuro de zinc. Cuando colocó un imán de herradura observó que los rayos se curvaron hacia abajo (figura 4).

Al darle vuelta al imán de herradura, es decir, al cambiar su polaridad, observó que los rayos ahora se curvaban hacia arriba. Perrin notó que el comportamiento coincidía con la regla de la mano izquierda de las ya conocidas corrientes eléctricas, concluyendo que los rayos catódicos debían tener carga negativa.

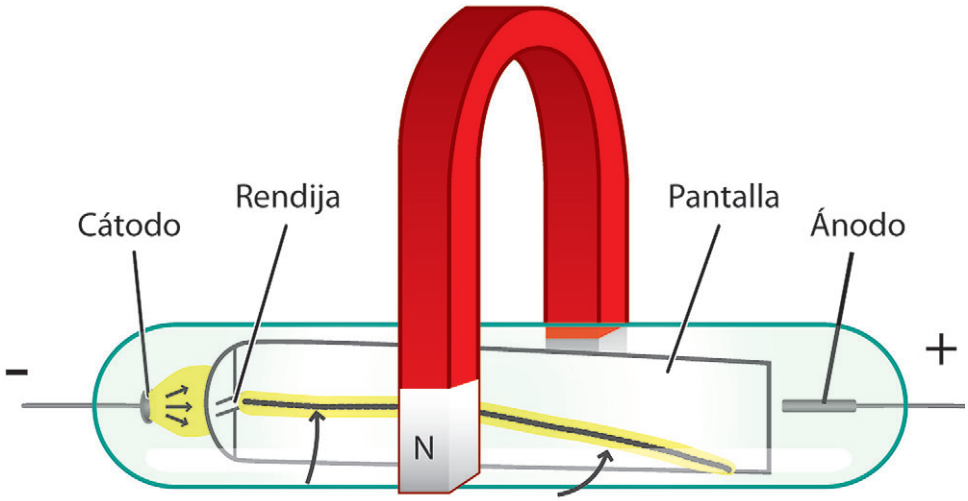


Figura 4. Tubo de rayos catódicos interactuando con un campo magnético.

7.1.2 El corpúsculo llamado electrón

Cuando en 1895 se concluyó que los rayos catódicos se comportaban como la corriente eléctrica, cobró fuerza la idea de que debían ser partículas, como lo enunció Benjamin Franklin, siglo y medio atrás. Se asumía, igualmente, que mostrar la existencia de una unidad fundamental en carga y masa sería definitivo para comprobar la idea de que se trataban de partículas, por lo que los físicos de la época se dieron a esa tarea.

En 1897, Joseph John Thomson (1856-1940) diseñó un tubo de rayos catódicos en el cual, después de colimar (afinar) el haz, lo hacía pasar entre dos placas de carga controlada, cuyo campo eléctrico atraía al haz en una dirección, mientras un campo magnético lo atraía en la dirección contraria. Por último, chocaba con una pantalla fluorescente al final del tubo (figura 5).

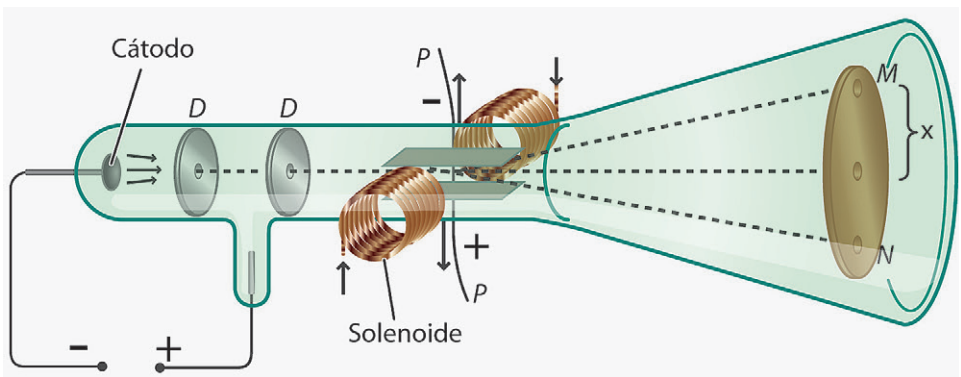


Figura 5. Tubo de rayos catódicos colimados.

Al aplicar simultáneamente los campos eléctrico E y magnético B, graduándolos de tal modo que la desviación debida a un campo se anulaba por el otro, Thomson podía asumir que las fuerzas eléctrica y magnética eran de igual magnitud pero de sentido contrario y, como consecuencia, pudo calcular la velocidad de los rayos catódicos.

La fuerza eléctrica está dada por:

$$F_e = Eq,$$

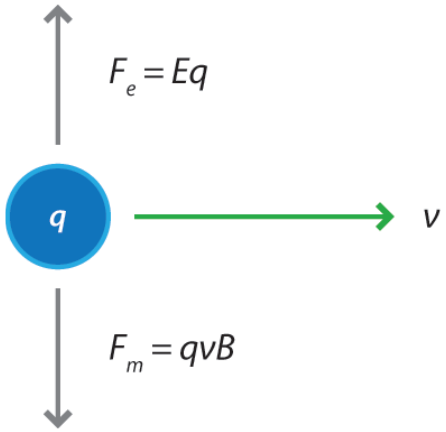


Diagrama de fuerzas.

y la magnética por:

$$F_m = qvB,$$

donde E es la magnitud del campo eléctrico; q la carga de la partícula constituyente de los rayos catódicos y v su velocidad. Dado que estas fuerzas deben ser iguales:

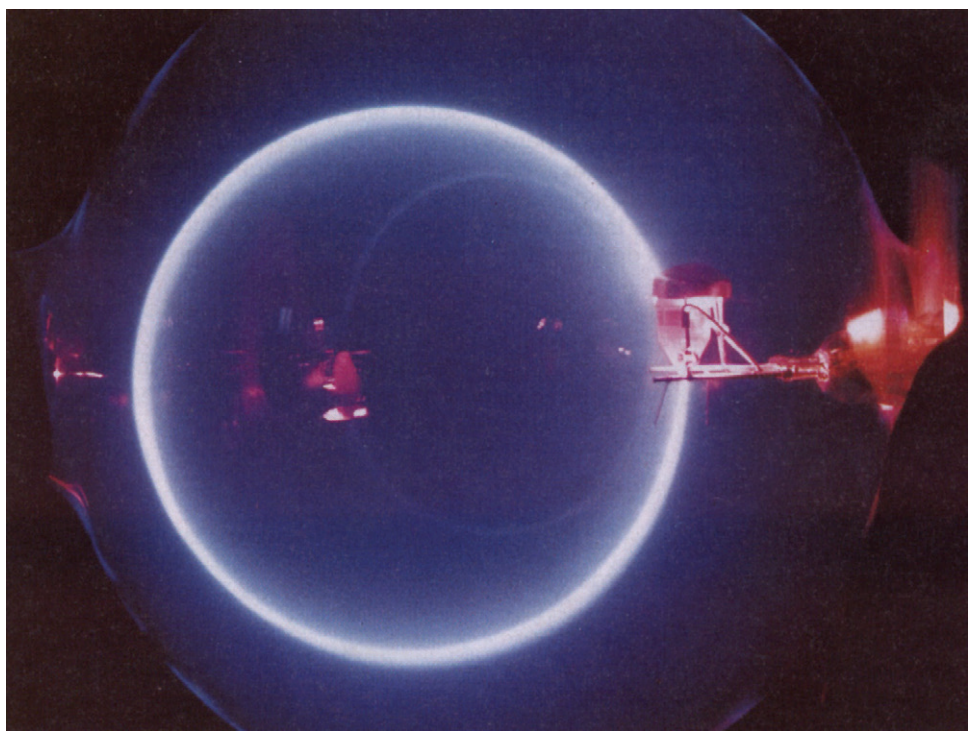
$$Eq = qvB,$$

de donde:

$$v = \frac{E}{B}.$$

Cuando Thomson sustituyó los valores respectivos de los campos que estaba usando, vio que esta velocidad era de varios miles de kilómetros por segundo, aproximadamente un quinto de la velocidad de la luz, y que dependía del campo eléctrico entre ánodo y cátodo, como se desprende de la ecuación.

Haz cerrado sobre sí mismo | © Latin Stock México.



En otro experimento, el cual consistía en aplicar un campo magnético uniforme, observó que el haz se cerraba sobre sí mismo formando un círculo. Esto le permitió asumir que la fuerza generada por el campo magnético actuaba siempre perpendicular al haz y al campo magnético, lo cual obligaba a tener una trayectoria circular, con la fuerza magnética apuntando hacia el centro. Es decir:

$$F_m = qvB \quad \text{y} \quad F_{\text{centrípeta}} = m\frac{v^2}{r}.$$

Así, igualando estas ecuaciones:

$$qvB = m\frac{v^2}{r}.$$

Sustituyendo la velocidad medida anteriormente, la magnitud del campo magnético utilizado B y el radio r de la trayectoria observada, concluyó que:

$$\frac{q}{m} = 1.7589 \times 10^{11}\,\frac{\mathrm{C}}{\mathrm{kg}}.$$

Bastaría conocer la carga para poder calcular la masa de la partícula de los rayos catódicos, ya que Thomson desconocía la carga y la masa del corpúsculo, pero conocía su cociente.[5] Robert A. Millikan, en 1909, diseñó un aparato que le permitía "balancear" en el aire una pequeña gota de aceite ionizada por un haz de rayos x, que caía a través de un campo eléctrico.

Ajustaba el campo E hasta que la fuerza eléctrica y el peso eran iguales en magnitud pero de sentido contrario, anulándose la fuerza total sobre la gota que caía con *velocidad constante*. Esto le permitió calcular la carga de cada gota.

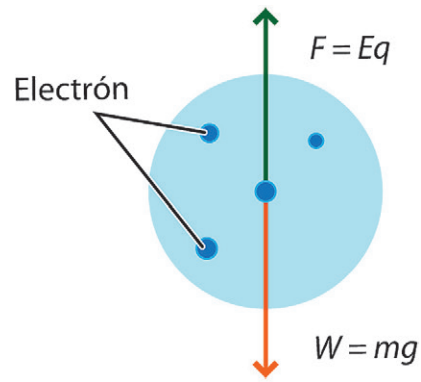


Diagrama de fuerzas.

$$Eq = mg,$$

en donde la masa se determinaba multiplicando el volumen de la gota (considerándola esférica) cuando caía, por la densidad del aceite (véase figura 6, p. 478).

Usando las cargas medidas para diferentes gotas y suponiendo que eran múltiplos enteros de una carga fundamental e, Millikan fue capaz de determinar que:

$$e = -1.602 \times 10^{-19}\ \mathrm{C}.$$

Una vez establecida por Millikan la carga fundamental, que sería entonces la del electrón, fue posible determinar su masa usando la relación de Thomson antes mencionada,

$$m = 9.1072 \times 10^{-31}\ \mathrm{kg}.$$

Más adelante, Thomson escribiría en sus *Recuerdos y reflexiones*, publicados en 1936:[6]

> Tras largas meditaciones acerca de los experimentos, me pareció que eran ineludibles las siguientes conclusiones:
>
> Los átomos no son indivisibles; de ellos pueden arrancarse partículas cargadas de electricidad negativa, mediante la acción de fuerzas eléctricas, el choque de átomos que se mueven con rapidez, la luz ultravioleta o el calor.
>
> Todas esas partículas son idénticas en cuanto a la masa y llevan idéntica carga eléctrica negativa, sea cual fuere la especie de átomos de que salgan, y son elementos constitutivos de todo átomo.

[5] J. J. Thomson, 1897, "Rayos catódicos", *Philosophical Magazine*, núm. 44, p. 293.
[6] J. J. Thomson, "Recuerdos y reflexiones", *Antología de física*, Lecturas Universitarias, núm. 9, p. 226.

La masa de dichas partículas es menos de un millonésimo de la masa del átomo de hidrógeno.

Al principio di a esas partículas el nombre de "corpúsculos", pero ahora se designan con el más apropiado de "electrones".

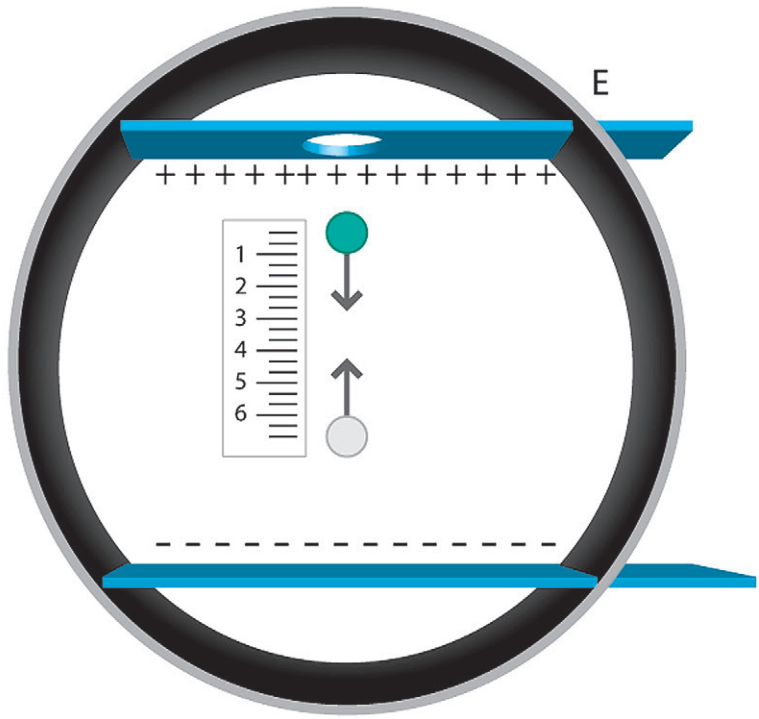


Figura 6. Esquema del experimento de Millikan.

Con esta conclusión, la idea de átomo de John Dalton fue severamente golpeada, ya que no podía considerarse más una partícula indivisible, pues se había encontrado ahora un corpúsculo mucho más pequeño en su interior: el *electrón*.

7.1.3 El modelo atómico de Thomson

En 1909, el electrón era la única partícula atómica que se conocía. Para explicar la neutralidad del átomo, Thomson sugirió que los electrones estaban embebidos en una nube con carga positiva distribuida homogéneamente en el átomo. Este modelo parecía razonable, pero había que probar la consistencia de las suposiciones que lo sustentaban. Había mucho por averiguar.

Ernest Rutherford (1871-1937) llevaba diez años estudiando las partículas emitidas por ciertos materiales radiactivos, a las que llamaba α (alfa), y se disponía a probar el modelo propuesto por Thomson. Demostró que las partículas alfa tienen carga positiva, son más pequeñas que los átomos, son pesadas (tienen masa) y las emiten las sustancias radiactivas con una gran velocidad. Siendo así, estas partículas podían emplearse como proyectiles de alta energía para estudiar a los átomos.

Con su colaborador Hans Geiger (1882-1945), Rutherford enviaba haces de partículas α a través de diversos materiales y, mediante pantallas fluorescentes, similares a la de un televisor, detectaba el lugar de salida. Así, podían observar si se desviaban a causa de las posibles interacciones en su viaje al interior de átomo. Por aquellos días, ellos habían

disparado miles de proyectiles contra sus delgados objetivos y ninguno se había desviado más allá de unos pocos grados (figura 7).

Estas leves desviaciones parecían deberse a la influencia que ejercía la carga negativa de los electrones existentes en el átomo sobre la carga positiva de la partícula disparada (figura 8).

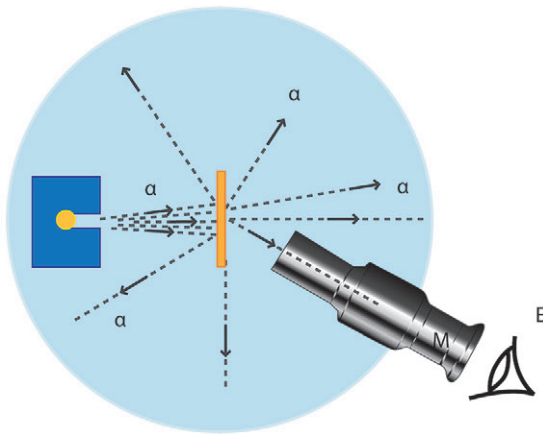


Figura 7. Retrodispersión.

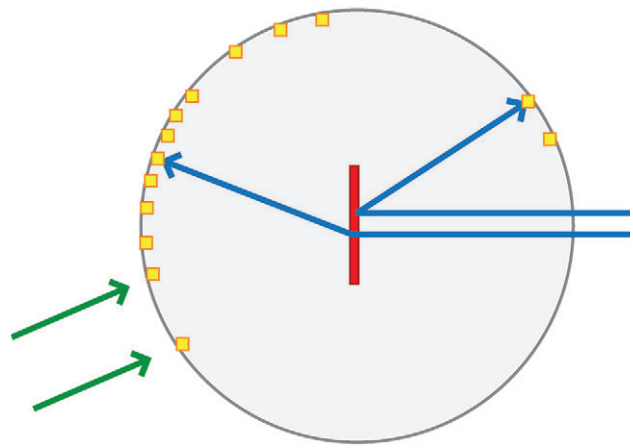


Figura 8. Marcas sobre la pantalla.

De acuerdo con cálculos probabilísticos, existía la posibilidad de que, al pasar a través de las muestras, la partícula se encontrara con un electrón, luego con otro y otro. El efecto de aquellos encuentros sucesivos podría dar, teóricamente, una desviación hasta de 45°, pero la probabilidad era pequeñísima. Sobraban razones para esperar que, más allá de los 45°, no se detectara nada. Sin embargo, había que probarlo experimentalmente (figura 9).

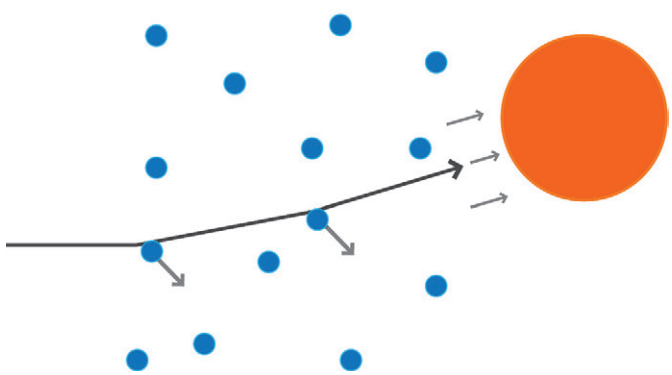


Figura 9. Diagrama de colisiones.

7.1.4 El modelo de Rutherford: el descubrimiento del núcleo atómico

De acuerdo con Rutherford, el dispositivo experimental era un tubo de vidrio que encerraba a la fuente de partículas α, que incidirían sobre una delgada lámina de oro, y fueron sustituidas por pantallas detectoras cilíndricas para cubrir las posibles desviaciones de 45° o más (véase figura 10, p. 480).

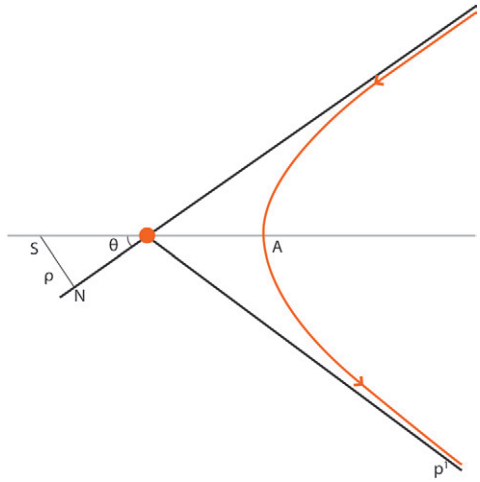


Figura 10. Diagrama de la dispersión de Rutherford.

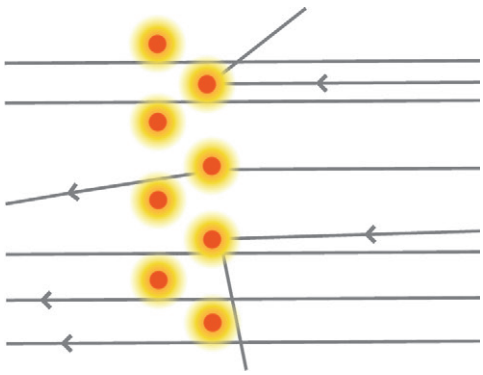


Figura 11. Desviación de las partículas.

A principios de 1911 Rutherford buscó a Hans Geiger. En contra de todo lo esperado, se encontró que de los millares de partículas α disparadas a través de la lámina de oro, algunas, muy pocas, sufrieron una gran desviación. Una o dos se habían desviado más de 90º, saliendo del blanco por el mismo lado que entraron. Rutherford estaba convencido de que tales rebotes no podían deberse a una serie de colisiones de una partícula α con los electrones. Era como si un tráiler rebotara al colisionar con una bicicleta (véase figura 11).

El sencillo modelo atómico de Thomson no explicaba nada de esto. Al llegar a este punto podría haberse concluido que los resultados eran incorrectos. Sin embargo, Rutherford tomó los resultados como buenos: "Hay algo en el átomo que puede hacer rebotar a las veloces y pesadas partículas α."

El cálculo le indicó que debían de haber encontrado un campo eléctrico muy fuerte. Semejante intensidad podría ser producida por una carga eléctrica concentrada en un espacio muy pequeño. Estaba tomando forma una nueva hipótesis: "La electricidad positiva del átomo no es —como creía J. J. Thomson— un fluido distribuido uniformemente por el átomo, sino que está concentrada en el centro, en un volumen muy compacto."[7]

Basándose en esta idea, Rutherford se planteó un problema: dada una carga eléctrica y una cantidad conocida de partículas α dirigiéndose hacia ella con una velocidad conocida, ¿cuál sería la dispersión más probable? ¿Cuántas partículas se acercarían lo suficiente al centro cargado para dispersarse con ángulos de 20°, 45°, 60° y 90°?

Rutherford calculó la respuesta y la comparó con sus observaciones y con los resultados de otros experimentos de dispersión realizados con anterioridad. Las observaciones mostraban una gran compatibilidad con los cálculos que surgían de su hipótesis. Aquello lo entusiasmó y le dio seguridad, pero su modelo debía ser puesto a prueba con mayor detalle. Con Ernest Marsden (1889-1970) y Geiger ideó nuevos experimentos de dispersión. Antes de dar por concluido su trabajo, se contaba con más de un millón de destellos para analizar.

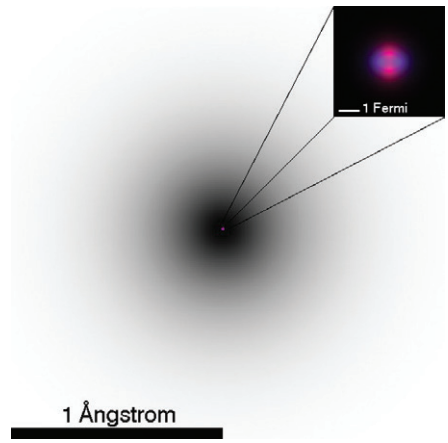
En mayo de 1911, Ernest Rutherford publicó su primer artículo acerca de lo que habían encontrado y anunció así el descubrimiento del *núcleo*, como llamó al centro de carga positiva del átomo. De sus experimentos, concluyó que el núcleo era 10 mil veces menor que el átomo; tan pequeño como una cabeza de alfiler en la sala vacía de una casa. No obstante, en el núcleo reside casi toda la masa del átomo. Fuera de este punto diminuto y pesado, en el centro del átomo, habría espacio vacío. Ahí están los electrones en número suficiente para compensar la carga positiva del núcleo.

> Puede que no carezca de interés el tratar de imaginarnos el concepto que hasta ahora nos hemos formado del átomo. Para ello elegiremos como ejemplo el átomo más pesado de

[7] Ernest Rutherford, "La dispersión de partículas α y β por materia, y la estructura del átomo", *Philosophical Magazine*, serie 6, vol. 21, mayo de 1911, pp. 669-688.

todos, el de uranio. En el centro del átomo hay un núcleo diminuto, en torno del cual se arremolina un conjunto de 92 electrones, los cuales se mueven recorriendo órbitas determinadas y ocupando, aunque de ninguna manera llenando, un volumen muy grande en comparación con el núcleo.

Algunos electrones recorren órbitas casi circulares alrededor del núcleo; otros, órbitas de forma elíptica, con ejes que giran con rapidez alrededor del núcleo.[8]



Representación del tamaño del átomo.

7.1.5 La atómica trinidad

De los experimentos de Rutherford se concluyó también que el núcleo, a su vez, está compuesto de partículas llamadas protones y neutrones. El descubrimiento del protón se le acredita a Rutherford. En 1918, él encontró que cuando se disparan partículas α contra un gas de nitrógeno, en sus detectores se registraron trazas de núcleos de hidrógeno.

Rutherford determinó que el único sitio del cual podían provenir estos núcleos era del nitrógeno y que, por lo tanto, el nitrógeno debía estar formado por núcleos de hidrógeno. Por estas razones Rutherford sugirió que el núcleo de hidrógeno debía ser una partícula fundamental, ahora conocida como protón. Para tratar de explicar que los núcleos no se desintegrasen debido a la repulsión electromagnética de los protones, el mismo Rutherford propuso en 1920, por primera vez, la existencia del neutrón.

Tabla periódica de los elementos.

1 H																	2 He
3 Li	4 Be											5 B	6 C	7 N	8 O	9 F	10 Ne
11 Na	12 Mg											13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
19 K	20 Ca	21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
37 Rb	38 Sr	39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
55 Cs	56 Ba	*	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
87 Fr	88 Ra	**	104 Rf	105 Ha	106 Sg	107 Ns	108 Hs	109 Mt									

*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

[8] "Ernest Rutherford, descubrimiento del núcleo", en Lovett C., Bárbara, *Los creadores de la nueva física* (cap. 1), México, Fondo de Cultura Económica, pp. 11-30.

En 1930, en Alemania, Walther Bothe (1891-1957) y H. Becker descubrieron que cuando se hacían incidir partículas α sobre berilio, boro o litio, se producía una radiación particularmente penetrante. Se pensó que eran rayos γ (gama), aunque los recién encontrados eran más penetrantes que los rayos γ hasta entonces conocidos, así que los detalles de los resultados experimentales eran difíciles de interpretar.

Dos años después esta teoría se desechó cuando, en París, Irène Joliot-Curie (1897-1956) y Frédéric Joliot-Curie (1900-1958) mostraron que esta radiación desconocida, al golpear parafina u otros compuestos que contenían hidrógeno, producía protones de alta energía. Eso era consistente con la suposición de que eran rayos γ.

Por último, a finales de 1932, el físico inglés James Chadwick (1891-1974), en Inglaterra, continuó con experimentos similares a los científicos anteriores de los que obtuvo resultados que no concordaban con los predichos por la teoría.

Para explicar tales resultados fue necesario suponer que la radiación estaba formada por corpúsculos, así que éstos quedaron explicados, aunque fue necesario aceptar que las partículas que formaban la radiación no tenían carga eléctrica. Tales partículas debían tener una masa muy semejante a la del protón, pero sin carga eléctrica. Así se identificó al neutrón como una nueva partícula.

La precaria identidad de los "átomos" de Dalton, que Dmitri Mendeléiev (1834-1907) ordenó en su célebre tabla periódica, empeoró con la aparición del protón y el neutrón. Los que convencionalmente conocemos como átomos, no son indivisibles, pero tampoco son los átomos de la definición de Demócrito ya que los podemos separar en electrones, protones y neutrones.

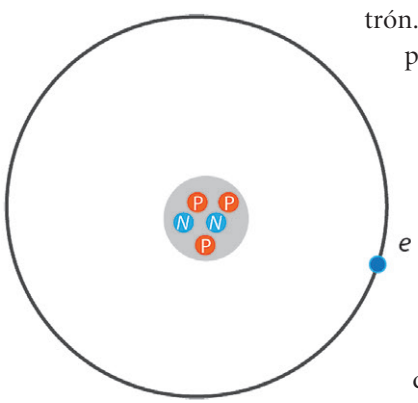


Figura 12. Representación del átomo.

No obstante, con estas tres partículas podemos reproducir todos los elementos de la tabla periódica y, por lo tanto, todo lo que nos rodea. La historia no requería más. Por segunda vez, la búsqueda parecía haber concluido: *estas tres partículas parecían ser los átomos de Demócrito* (véase figura 12).

Por supuesto había preguntas importantes, todavía sin respuesta. Por ejemplo: si el núcleo de un átomo puede tener más de cien protones, todos cargados y positivos, y por lo tanto, todos repeliéndose, ¿por qué no estallan los núcleos atómicos? ¿Qué les da estabilidad? O al revés: ¿por qué no vemos que los electrones se agrupen? Pero esto no se consideró trascendente.

7.1.6 La búsqueda de nuevas partículas

Los decenios que siguieron fueron de gran actividad, tanto teórica como experimental, para despejar la duda de si el electrón, protón y neutrón eran las partículas elementales de la materia. La sorpresa fue mayor cuando se fueron encontrando cada vez más partículas. El número de ellas y sus propiedades requieren de teorías complejas de la física. Sin embargo, se debe estar consciente de que existe un número muy grande de partículas elementales con diversas propiedades que sólo se pueden entender haciendo uso de teorías sofisticadas, cuyos fundamentos se enseñan en programas de maestría y doctorado.

El esquema todavía es más complejo porque se ha descubierto que para cada partícula hay una antipartícula, de forma tal que al encontrarse se aniquilan, con una gran liberación de energía. En este tema se desconoce por qué esa aniquilación no sucede en forma espontánea en nuestro Universo. Esta incógnita, junto con muchas otras, todavía no tiene respuesta.

BIBLIOGRAFÍA BÁSICA

Alonso, M., E. J. Finn, *Física*, vol. I: *Mecánica*, vol. II: *Campos y ondas*, México, Addison-Wesley Iberoamericana, 1995.

Baird, D.C., *Introduction to experimentation*, Mass., Addison-Wesley/Read, 1995.

Dicke, R. H., P. J. E. Peebles, P. G. Roll, D. T. Wilkinson, *Astrophysical Journa*l, vol. 142, 1965, pp. 414-419.

Meiners, H. F., W. Eppenstein y K. H. Moore, *Experimentos de física*, México, Limusa, 1980.

Feynman, R., *Mecánica, movimiento ondulatorio y calor*, Madrid, Aguilar, 1987.

——, R. B. Leighton, M. Sands, *The Feynman lectures on physics*, vol. 1, Mass., Addison-Wesley/Read, 1987.

Hecht, Eugene y Alfred Zajac, *Optics*, Mass., Addison-Wesley/Read, 1997,

Hubble, Edwin P., "Una relación entre la distancia y la velocidad radial entre nebulosas extragalácticas", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 15, núm. 3, marzo de 1929, pp. 168-173.

Humason, M. L., "¿Está el Universo expandiéndose?", *Astronomical Society of the Pacific Leaflets*, vol. 2, 1936, p. 161.

——, "No. 531, La velocidad radial aparente de 100 nebulosas extragalácticas", *Contributions from the Mount Wilson Observatory*, Carnegie Institution of Washington, vol. 531, 1936 , pp.1-13.

Introduction to Modern Optics, 2a. ed., Holt-Rinehart-Winston, 1975.

Kittel, C., W.D. Knight, M.A. Ruderman, *Berkeley Physics Course*, vol. 1: *Mecánica*, México, Reverté, 1968.

Leavitt, Henrietta S., "1777 variables en las nubes de Magallanes", *Annals of Harvard College Observatory*, vol. 60, 1908, pp. 87-108.

Leavitt, Henrietta S., y Edward C. Pickering, "Periodos de 25 estrellas variables en la pequeña Nube de Magallanes", *Harvard College Observatory Circular*, vol. 173, marzo de 1912, pp. 1-3.

López Cano, J.L., *Leyes, teorías y modelos*, Asociación Nacional de Universidades e Instituciones de Educación Superior (anuies), 1978.

Lloyd, W. T., *Introducción al estudio de la mecánica, materia y ondas*, México, Reverté, 1973.

Marroquín de la Rosa, J. D. y otros, *Conocimientos fundamentales de física*, México, Pearson/unam, 2006.

Malacara, D., *Óptica básica*, México, Secretaría de Educación Pública/Fondo de Cultura Económica, 1989.

Peimbert, Manuel, "Evolución química del Universo", en *Temas selectos de astrofísica*, México, unam, 1984, pp. 307-331.

Penzias, A., R. Wilson, *Astrophysical Journal*, vol. 142, 1965, pp. 419-420.

Purcell, E.M., *Berkeley Physics Course*, vol. 2: *Electricidad y magnetismo*, México, Reverté, 1992.

Rabinowicz, E., *Método e hipótesis científicos*, México, Trillas, 1970.

Resnick, R., D. Halliday, S. K. Krane, *Física*, vols. 1 y 2, 5a. ed., México, Patria, 2009.

Revista Mexicana de Física (sección de enseñanza), *American Journal of Physics*, *The Physics Teacher*, varios artículos en las revistas.

Slipher, V. M., "La velocidad radial de la nebulosa de Andrómeda", *Lowell Observatory Bulletin*, núm. 58, vol. II, 1913, pp. 56-57.

Wagoner, R. V., *Astrophysical Journa*l, vol. 179, 1973, pp. 343-360.

Zemansky, M. W., R. H. Dittman, *Calor y termodinámica*, México, McGraw-Hill, 1973.

APÉNDICE

FÍSICA

ARISTÓTELES
Obras

GALILEO GALILEI
Diálogo acerca de dos nuevas ciencias

ALBERT EINSTEIN
La relatividad

OBRAS
ARISTÓTELES

[Publicado en Francisco de P. Samaranch (ed.), *Aristóteles. Obras*, Madrid, Aguilar, 1964, pp. 749-757]

CAPÍTULO 13

La Tierra, su situación y su figura. Sobre si está o no en movimiento. Y por qué razones.

Nos queda ahora por hablar de la Tierra. Hay que decir dónde está colocada, si es uno de los seres que están en reposo o uno de los que se mueven, y también hemos de decir algo de su figura.

De su posición no todos tienen iguales opiniones: muchos dicen que está colocada en el centro, los que dicen que el cielo es un todo infinito; por el contrario, los que habitan una parte de Italia y se llaman pitagóricos, opinan al revés de esto. Dicen, en efecto, que en el centro está el fuego, mientras que la Tierra es una de las estrellas y se mueve en torno al centro, y que de esta manera se produce el día y la noche. Además, conciben e imaginan otra Tierra, contraria a esta, que llaman la "antitierra", no ya buscando las explicaciones y las causas para las cosas que se ven, sino llevando a ciertas opiniones suyas lo que se ve y procurando adornarlo. También a otros varios les parece o puede parecer que no es necesario darle a la Tierra el lugar del centro, apoyándose para ello no en las cosas aparentes, sino buscando más bien su argumentación y su creencia en algunas nociones abstractas. Creen, en efecto que un cuerpo en grado sumo notable debe necesariamente ocupar un lugar en grado sumo notable; y creen que el fuego es de más digna condición que la Tierra, y que el fin es más digno que los medios, y, finalmente, que el límite extremo y el centro son un fin. Razonando, pues, de esta manera, vienen a creer que la Tierra no está en el centro de la esfera, sino más es el fuego el que está allí. Además, los pitagóricos, fundados en que conviene sobre todo conservar lo que en el Universo es lo más importante, y en que el centro cumple esta condición, llaman por esta razón al fuego que ocupa aquel lugar la custodia de Júpiter; como quien habla absolutamente del centro y fuera este el centro de las magnitudes, de las cosas y de la Naturaleza.

Ahora bien: igual que en los animales no es lo mismo el centro del animal y el centro del cuerpo,[1] así es como hay que pensar, y con mayor razón, acerca del mismo cielo total. Por consiguiente, por esta misma causa no es conveniente que nos turbemos en la cuestión del Universo, ni que llevemos al centro esta guardia o custodia; antes debemos buscar qué es y cuáles son las cualidades del centro, y dónde puede estar según su aptitud. Aquel centro es, en efecto, un principio, y precioso; pero el centro es más semejante a un límite del lugar que a un principio. El centro, en efecto, constituye un término; y el término es él mismo un fin. Y es más apreciable o digno de aprecio lo que limita y contiene que lo que es limitado y contenido. Porque esto es la materia y aquello es la sustancia de la constitución misma del ser.

Esta es, pues, la opinión que algunos tienen respecto a la Tierra, y, análogamente, respecto del reposo y el movimiento. No todos, en efecto, piensan de igual manera,

[1] El centro lógico o real del animal es el corazón. Sobre su posición en el cuerpo cf. *De partibus animalium*, 665b 21, 666b 3.

sino que los que no la creen dicen que se mueve en torno al centro; y no tan sólo es esta la que se mueve, sino también la "antitierra", tal como hemos dicho antes. Para algunos, también todos estos varios cuerpos parece se pueden mover alrededor del centro, si bien no nos es manifiesto a causa de la interposición de la Tierra. Por esto dicen que en la Luna se verifican más eclipses que en el Sol, pues dicen que cada uno de los cuerpos son movidos o llevados en el espacio, se ocultan mutuamente y ocultan también a la Luna, no sólo a la Tierra. Pues al no ser la Tierra el centro, sino que dista de él todo un hemisferio, creen que nada impide que sucedan las cosas que se ven, de manera igual aunque nosotros no habitemos en el centro, como si la Tierra estuviera también en el centro; porque ahora tampoco se nos hace nada evidente o manifiesto, al distar la mitad de un diámetro. Otros autores dicen que la Tierra, fija en el centro, gira sobre sí misma, y se mueve en torno al mismo polo a través del Universo extenso, como se halla escrito en el *Timeo*.

De manera análoga se discute sobre su figura. A algunos, en efecto, les parece ser esférica; a otros les parece ser extensa como un tambor; de esto último aducen esta prueba: que cuando el Sol nace o se pone, parece esconderse por una línea recta, no por una línea curva, ya que sería necesario que la división del Sol tuviera lugar por una línea curva si la Tierra fuese esférica; dicen esto sin tener en cuenta la distancia del Sol a la Tierra y la magnitud de la circunferencia, ya que en estos círculos que parecen pequeños también aparece recta. Por tanto, por esta sola apariencia no es necesario que crean ellos que la Tierra no es esférica. Añaden además y dicen que ella debe tener necesariamente esta figura por su estado de reposo. Pues los modos del movimiento y del reposo, de que hemos hablado, son muchos.

Es necesario, pues, que todos hayan dudado de esto, pues quizá es propio de una mentalidad demasiado despreocupada no admirar por qué razón una pequeña parte de la Tierra, elevada en el aire, si es arrancada, se mueve y no quiera estar en reposo, y siempre más rápidamente si es mayor, mientras que si es la Tierra entera la que uno suelte en el aire, una vez elevada a lo alto, no se mueve. Ahora, este peso tan grande está en reposo. Y si alguien, mientras se mueven las partes de ella y antes que caigan, quita la Tierra, estas partes serán llevadas hacia abajo si nada se lo impide.

De esta manera, a todos se les ocurre dudar a causa de la filosofía; pero, sin duda, se puede uno admirar de que las soluciones no parecen más absurdas que la misma duda. Algunos, en efecto, por este motivo dicen que la parte inferior de la Tierra es esférica, diciendo que ella está enraizada en el infinito, como dijo, por ejemplo, Jenófanes de Colofonia, para que no tengan molestias los que buscan insistentemente las causas. Por esta razón, Empédocles le increpó con estas palabras:

> Puesto que la profundidad de la tierra es infinita y el éter es abundante, ya que, por obra de las bocas y lenguas muchas cosas dichas en vano se derramaron, por bocas, digo, de los que no veían más que una mínima parte del universo…

Otros dicen que está echada encima del agua. Esta es la sentencia más antigua que hemos recibido, la cual se atribuye a Tales de Mileto; es decir, que la Tierra está en reposo porque, igual que si fuera un madero o algo equivalente, flota o nada. Porque ninguna de estas cosas es apta para permanecer en el aire, pero sí sobre el agua. Como si no fuera la misma la noción acerca de la tierra y acerca del agua que lleva la Tierra. Porque tampoco el agua es apta para permanecer en lo alto, antes siempre está encima de algo.

Por otra parte, igual que el aire es más ligero que el agua, el agua es más ligera que la tierra. Así pues, ¿cómo pueden pensar que lo que es más ligero está debajo de lo que es más pesado por naturaleza? Además, si toda ella es apta para permanecer encima del agua, es evidente que también cualquier parte de ella será igualmente apta para ello. Ahora bien: la experiencia nos enseña que esto no ocurre, sino que una parte cualquiera de la tierra se va al fondo del agua, y tanto más aprisa cuanto mayor es.

Pero parecen llevar su inquisición tan sólo hasta cierto límite y no hasta donde se puede llevar la investigación. Porque todos nosotros tenemos esta costumbre, a saber, de no llevar las preguntas hasta la misma cuestión, sino tan sólo de cara al que opina lo contrario. En efecto, uno busca en sí mismo hasta el punto o momento en que él mismo no se puede contradecir a sí mismo. Por esta razón, es conveniente que el que haya de investigar las cosas con exactitud sea apto y esté preparado para inferir todas las objeciones correspondientes a un género. Y está en estas condiciones el que ha considerado y meditado todas las diferencias.

Anaxímenes, por su parte, Anaxágoras y Demócrito, dicen que la causa de que la Tierra esté en reposo es su anchura. Porque pretenden que la parte inferior no corta el aire, sino que lo tapona completamente; que es lo que, al parecer, hacen los cuerpos que tienen anchura o extensiones. Estos, en efecto, pueden moverse con dificultad por los vientos a causa de su capacidad de resistencia. Así,

pues, dicen que esto mismo hace que la Tierra esté junto al aire sujeto, y que éste, al no disponer de lugar suficiente al que trasladarse, reposa simultáneamente con ella en la parte inferior, como hace el agua en las clepsidras. Ahora bien: demuestran con abundantes pruebas y señales que el aire, si se le somete a una gran presión, puede llevar o sostener sobre sí un gran peso.

Primero, pues, si la figura de la Tierra no es ancha y aplanada, no por este motivo dejará de estar en reposo la Tierra. No obstante, según lo que dicen, no es esta la única causa del estado de reposo, sino más bien el tamaño de la Tierra, porque al no tener paso el aire, debido a la estrechez, reposa la Tierra a causa de su pluralidad (del aire). Y ese aire es abundante o copioso gracias a que está sometida a presión por la enorme magnitud de la Tierra. De manera que ocurría lo mismo que dicen, aunque la Tierra sea esférica y sea tan grande su masa: estará en reposo, según el parecer de aquellos.

Por otra parte, a los que hablan así acerca de la cuestión del movimiento no hay que discutirles sobre lo que ocurre en las partes, sino sobre lo que pasa en el Universo y la totalidad.[2] Hay que determinar, en efecto, desde el comienzo, si existe algún movimiento que corresponda naturalmente a los cuerpos, o no existe este tal movimiento; y si en caso de no existir este movimiento natural existe, sin embargo, un movimiento violento. Pero puesto que ya hemos definido lo que hay que pensar acerca de este punto, según nos lo permitió la oportunidad de entonces, es preciso que ahora hagamos uso de estas conclusiones como verdaderas. Pues si a los cuerpos no les corresponde ningún movimiento natural, tampoco les corresponderá ningún movimiento violento. Y si ninguno de los dos existe, nada absolutamente estará en movimiento. Hemos ya determinado antes, en efecto, que era necesario que ocurriera esto.

Además tampoco nada podría estar en reposo, pues igual que un movimiento es natural y otro es movimiento impuesto por la violencia, también ocurre así en el estado de reposo.[3] Pero si realmente existe un movimiento natural, no sólo no habrá movimiento violento, sino tampoco habrá reposo violento. De manera que si ahora la Tierra descansa por la fuerza, fue también llevada al centro por una rotación violenta. Todos, en efecto, sostienen esta causa, apoyados en las cosas que se mueven en medios líquidos y en las cosas que ocurren respecto del aire. Ya que en estos momentos las cosas mayores y más pesadas siempre se mueven hacia el centro mismo de la rotación. Por otra razón, todos los que dicen que el cielo nació alguna vez, dicen que la Tierra llegó al centro. Pero buscan la causa por la cual está en reposo. Y la explican unos de esta manera, es decir, a base de que la extensión o anchura y la magnitud sean la causa de su estado de reposo; otros, como cree Empédocles, por ejemplo, opinan que el cielo, con su rotación, que es más rápida que la traslación de la Tierra, impide la traslación de esta misma, como ocurre con el agua que está en los cazos. Ésta, en efecto, cuando el cazo se mueve en órbita, aunque con frecuencia el cazo se vuelve de tal manera que el fondo queda arriba y el agua queda abajo, no obstante, el agua no se mueve hacia abajo, con ser apta para moverse hacia allí, y ello por la misma causa sin duda.

Ahora bien: si no lo impide su anchura ni la rotación del cielo —pregunto yo—, ¿hacia dónde se moverá desde allí, en el caso de que el aire ceda y se marche? En efecto, por hipótesis, ha sido llevada hacia el centro por la violencia, y permanece en él por la violencia. Ahora bien: es necesario que posea entonces algún movimiento de traslación propio de su naturaleza. ¿Es este el que se dirige hacia arriba, el que se dirige hacia abajo u otro cualquiera? Es necesario que sea alguno. Y si no es más el que tiende hacia arriba que el que tiende hacia abajo, y el aire que está encima no le impide el movimiento que la lleve hacia arriba, tampoco, sin duda, el aire que está abajo le impedirá el movimiento que la lleve hacia abajo. Porque es necesario que, respecto de los mismos seres, sean las mismas las causas de los mismos efectos.

Además, se le podría también decir a Empédocles lo siguiente: ¿cuál era la causa del reposo de la Tierra cuando los elementos estaban sujetos a la acción de la discordia y separación? Porque no va a decir que entonces la causa de ello era la rotación. Es también absurdo no entender que al principio las partes de la Tierra fueron llevadas por la rotación hacia el centro, pero ahora, ¿por qué razón son arrastradas hacia ella todas las cosas que tiene peso? Pues la rotación no se realiza en dirección a nosotros.

Más aún: ¿por qué razón el fuego es llevado hacia arriba? Ciertamente no será por rotación. Y si éste es apto para ser llevado a algún lugar, evidentemente hay que creer que también la Tierra es, de igual manera, apta para dirigirse a algún lugar determinado. Ahora bien: lo ligero y lo pesado no vienen determinados por la rotación, sino que de los seres que antes eran pesados o ligeros se dirigen

[2] Es decir, la conducta de un elemento particular, la Tierra, por ejemplo, no se puede considerar aisladamedente, sino tan sólo como una parte del cosmos, con sus leyes universales.

[3] La conexión estrechísima entre movimientos y reposo, naturales o violentos, es debido a la definición de ambos, en relación con el lugar natural.

los unos hacia el centro, y los otros se colocan por encima de todos los demás a causa del movimiento. Había, por tanto, ya antes de que tuviera lugar la rotación, un ser ligero y un ser pesado. ¿Y de qué manera se diferenciaban estas cosas y de qué manera y a dónde eran aptas para ser llevadas? Porque si existe el infinito, no existe el lugar superior ni el inferior; y lo pesado y lo ligero vienen definidos por estos lugares.

Una gran mayoría de los autores dan vueltas en torno a estas causas. Los hay que dicen que la Tierra está en reposo, como por una indiferencia; por ejemplo, sostenía esto, entre los antiguos, Anaximandro.[4] Dicen, en efecto, que aquello que está colocado en el centro y equidista de los extremos, no es necesario que sea llevado hacia arriba más que hacia abajo, o bien hacia los lados, y que al mismo tiempo no puede moverse hacia los contrarios, de manera que vienen a decir que ello está necesariamente en reposo.

Ahora bien: esto encierra ciertamente una elegancia en su formulación, pero no es verdadero. Pues, según esta opinión es necesario que todo lo que se coloca en el centro permanezca en él en estado de reposo. Así pues, también el fuego estará en reposo. Porque lo que se ha dicho no es sólo propio de la Tierra.

Pero es que ni tan siquiera es necesario,[5] porque no parece tan sólo estar en reposo en el centro, sino ser llevada también hacia el centro. Pues, al punto a que es llevada cada una de sus partes, es necesario sea llevada también ella en su totalidad; y en el punto a que según su naturaleza es llevada, también allí permanece naturalmente en reposo. No es, pues, la razón de su reposo el estar equidistante de los extremos. Esto es, en efecto, común a todos los seres, mientras que el ser llevado hacia el centro es algo exclusivo y propio de la Tierra.

Es también absurdo inquirir por qué razón permanece la Tierra en el centro y no preguntarse o inquirir por qué razón el fuego halla su reposo en el extremo terminal. Porque si al fuego le corresponde por naturaleza aquel lugar superior y extremo, es necesario que también a la Tierra le corresponda algún lugar natural. Y si este lugar no le corresponde naturalmente a la Tierra, sino que permanece en él estado de reposo a causa de la necesidad de la semejanza o simetría —como afirma aquella opinión que se aduce a raíz del cabello, extendido con fuerza pero homogéneamente hacia todas partes, dice, en efecto, que no se romperá, e igual lo que dice de un ser que estuviera muy sediento y hambriento, cuando dista una distancia igual de lo que se bebe y lo que se come, pues es necesario que esté quieto—, deben también ellos inquirir acerca del reposo del fuego en los lugares últimos. Ahora bien: es digno de admiración que se pregunte uno por el estado de reposo de estos seres y no se pregunte, en cambio, por su movimiento de traslación; es decir, que no se pregunte por qué causa uno de ellos se mueve hacia arriba y el otro es llevado hacia abajo, hacia el centro, si nada lo impide.

Pero es que tampoco es verdadero lo que se dice de él. Es, con todo, verdadero accidentalmente que sea necesario que todo ello permanezca en reposo en el centro; es decir, aquel ser a quien le corresponde no moverse más hacia una parte que hacia otra. Ahora bien; según esta opinión no estará en reposo, antes se moverá: no todo, sin embargo, sino totalmente desgarrado o separado. La misma teoría, en efecto, tendría su aplicación al fuego. Porque si se le colocara en el centro, será necesario que esté en reposo igual que la Tierra, porque será equidistante respecto de cualquier extremo de los puntos cardinales; no obstante, se apartará del centro, igual que vemos que es llevado hacia el límite extremo, de no haber nada que lo impida; pero no será llevado todo él a un solo punto —ya que esto es necesario suceda tan solo por la noción que se deduce de la analogía—, sino que una parte será llevada a una parte del límite, hacia una parte sin duda proporcional a él; por ejemplo, una cuarta parte a una cuarta parte del continente. Porque ningún cuerpo es puntual. Pero así como un ser puede pasar de un sitio mayor a un sitio menor, por condenación, también puede pasar de uno menor a uno mayor, por rarefacción. De manera, pues, que la Tierra se movería de esta manera, por la noción de su semejanza o similitud, si este lugar no le correspondiera por naturaleza.[6]

[4] El significado de "δμοιότητα" se ve claro por el contexto. Hemos adoptado, con otros autores, la versión de la palabra por "indiferencia".

[5] Algunos comentaristas creen que este *necesario* se refiere a una necesidad lógica. Pero parece algo difícil de entender que un argumento no es verdadero ni es necesario, en este sentido. El sentido del fragmento es más bien éste: "su argumento es falso, pero en todo caso no es necesario, porque la inmovilidad de la Tierra está suficientemente justificada por el hecho de la moción natural, que implica un natural lugar de reposo".

[6] El argumento queda un tanto oscuro, porque sólo está parcialmente desarrollado. Vendría a ser más o menos así: La razón real, por la que la Tierra permanece en el centro es que éste es un lugar natural. Resultado accidental de esta razón esencial es que la Tierra, una vez en el centro, no tenga incentivo que la impulse a moverse más en una dirección que en otra, y es verdad que este hecho impide su movimiento. Para probar que ha discernido exactamente lo que es esencial, Aristóteles demuestra que la doctrina de la "indiferencia", sin apoyarse en la doctrina del lugar natural, es insuficiente para justificar la inmovilidad: a) la indiferencia se aplicara por igual al fuego que a la Tierra, y la experiencia muestra que, en el caso del fuego, este no puede ser reducido a una total inmovilidad en el centro. Colocado en

Éstas parecen ser, en suma, las opiniones que corren acerca de la figura, el lugar, el reposo o el movimiento de la Tierra.

CAPÍTULO 14

La tierra, conclusiones positivas

Digamos primero si tiene o no movimiento, y si en este caso está en reposo. Como dijimos, algunos creen que es una de las estrellas; otros la suponen colocada en el centro y que se mueve y gira en torno al polo. Es evidente, no obstante, que esto es imposible, tomando pie de lo que sigue: si es llevada, sea que esté en el centro, sea que esté fuera del centro, es necesario que ella sea llevada por este movimiento de manera violenta. No es, en efecto, propio de la misma Tierra, ya que cada una de sus partes tendría también como propio este mismo movimiento de traslación; ahora, en cambio, todas las cosas terrestres son llevadas en línea recta al centro. Con lo cual, puesto que el movimiento es violento, y contrario a la Naturaleza, es imposible que el movimiento sea eterno. En cambio, el orden del mundo es eterno. Además, todas las cosas que son llevadas en movimiento de rotación parecen quedar atrás y moverse, excepto la esfera primera, con más de un tipo de traslación. Luego también la Tierra, sea que esté colocada en el centro, sea que esté alrededor del centro, es necesario que se mueva con dos tipos de traslación. Y si ello es así, es necesario que se produzcan los cambios y evoluciones en las estrellas fijas. Pero no se ve que esto tenga lugar; antes las mismas estrellas nacen y se ponen en los mismos lugares de la Tierra.[7]

Además, la traslación naturalmente propia de las partes de la Tierra tiende al centro mismo del Universo. Ya que precisamente por esto está ahora en reposo en el mismo centro.

Podría plantearse la dificultad al ser el mismo el centro de unos y otros, de a cuál de las dos partes son llevadas las cosas que tienen peso, según su naturaleza. Si es porque es el centro del Universo o porque es el centro de la Tierra. Porque las cosas ligeras y el fuego, al tender al lugar contrario de los pesos de la Tierra, son llevados al extremo de aquel lugar que contiene el mismo centro. Pero sucede que es el mismo el centro de la Tierra y el del Universo, pues los pesos son llevados también hacia el centro de la Tierra, pero accidentalmente, sencillamente porque la Tierra tiene su centro en el centro mismo del Universo. Por su parte, de que las cosas pesadas son llevadas al centro de la Tierra esto es señal: los cuerpos, en efecto, que son llevados a ella, no son llevados según distancias iguales, sino según ángulos semejantes. De manera que son llevados a un mismo punto central del Universo y de la Tierra.

Por tanto, es evidente que la Tierra está necesariamente en el centro y que es inmóvil, tanto por las causas que ya hemos explicado como también porque los pesos que son arrojados hacia arriba vuelven de nuevo por el mismo sitio, aunque la fuerza que los lanza los lance hasta el infinito o hacia lo indeterminado. Así pues, queda con ello en claro que la Tierra ni se mueve ni está situada fuera del centro.

También queda en evidencia, por lo dicho, cuál es la causa del estado de reposo. Pues si es apta para ser llevada al centro desde cualquier parte, por naturaleza, según vemos ocurre, el fuego es de manera semejante para moverse desde el centro hasta el último límite o extremidad, es imposible que ninguna de sus partes sea llevada fuera del centro sin que se le infiera una violencia. Una sola es, en efecto, la traslación de un único ser, y es simple la traslación de un ser simple; pero no las contrarias. La traslación que parte del centro es contraria a la que se dirige hacia él. Si, pues, no es posible que ninguna parte de ella sea sacada del centro, es evidente que la totalidad será aún más imposible de ser llevada fuera. Pues el todo es apto para ser llevado al mismo sitio a que es llevada la parte. Por consiguiente, si es imposible que ella misma se mueva a

el centro, el fuego se dispersa, buscando cada particula el punto más cercano de la circunferencia. La tierra entonces haría lo mismo. Sería indiferente tan sólo de palabra. El error de los adictos a la doctrina de la indiferencia estaba en que no tuvieron en cuenta, en el movimiento que se aleja del centro, más que una sola cosa, es decir, un ser que se movía hacia afuera, en una dirección, engendrado por la tierra como una totalidad. Esto, pues, sería incompatible con la indiferencia, pero no así la dispersión de un cuerpo, desde el centro de la circunferencia, en todas direcciones instantáneamente. Si, pues, la tierra está en el centro, ello debe ser porque, de manera distinta al fuego, no tiene un impulso natural a apartarse de él, antes, por el contrario un impulso natural hacia él. b) Hay otra clase de movimiento que es compatible con la doctrina de la indiferencia, y por tanto ayuda a demostrar que la doctrina, si bien es verdadera, no exige completamente de necesidad esta total inmovilidad. Este es el movimiento de expansión uniforme en todas direcciones, que acompaña al proceso de rarefacción.

[7] La crítica depende de la analogía o semejanza con los planetas, siguiendo lo que admite Aristóteles de que, si la Tierra se moviera por su propio movimiento, de igual manera a como es llevada en órbita con el movimiento del primer cielo, su movimiento propio estaría en el plano de la eclíptica y no en el del ecuador. Si ello fuera así, las estrellas fijas nos mostrarían las irregularidades, que él describe con las palabras «παρόδοος χαι τρπας»: el polo de cada estrella parecería describir un círculo en el cielo, y las estrellas no estarían sujetas a la fases de nacimiento y puesta, como vemos nosotros.

no ser por fuerzas superiores, es necesario que ella permanezca en reposo en el centro.

Atestigua también esto lo que dicen los matemáticos sobre la astrología: en efecto, las cosas que se ven, cuando cambian las figuras, por las que se ha definido el orden de las estrellas, suceden porque la Tierra está situada en el centro.

Baste, pues, con esto en la cuestión del lugar de la Tierra y de las condiciones con que se dan su reposo y su movimiento.

Es, por otra parte, necesario que la Tierra tenga figura o forma esférica. Cada una de las partes, en efecto, posee un peso dirigido hacia el centro; y una parte menor, al ser empujada por una parte mayor, no puede salir,[8] antes bien se comprime, y una cede el lugar a la otra, hasta que esta llega al centro. Y conviene entender lo que se dice, como si tuviera lugar lo que muchos naturalistas dicen del modo como la Tierra tuvo su origen. Solo que aquéllos dicen que la causa de la Tierra fue la violencia de la traslación hacia abajo. Pero es mejor asentar bien la verdad y decir que esto sucede así porque lo que tiene un peso tiene por naturaleza la capacidad de ser llevado hacia el centro. Por tanto, cuando existía la confusión en potencia, las cosas que eran separadas eran llevadas hacia el centro, desde todas partes análogamente. Por tanto, sea que las partes divididas se congregan en el centro procedente de los extremos, sea que ello ocurriera de cualquier otra manera, harán sin duda lo mismo.

Es, pues, evidente que lo que, de una manera semejante es llevado desde los extremos hacia el centro, es necesario que de igual manera sea ello una masa o volumen en todos sentidos. Porque si en todas dimensione se hace una adición equivalente, es necesario que el extremo equidiste del centro. Ahora bien: esta figura es una figura esférica. En nada difiere el razonamiento, aunque las partes de ella no concurran al centro de una manera semejante desde todos los sentidos. Porque siempre es necesario que la parte mayor empuje a la parte menor que está delante de ella, poseyendo ambas un movimiento hacia el centro y empujando el cuerpo mayor al que es menor en peso. La dificultad que aquí podría aducirse tiene, en efecto, la misma solución. Pues si la Tierra, situada en el centro y dotada de una figura esférica, aumentara en un peso múltiple, añadido en uno de los hemisferios, no sería el mismo el centro del Universo y el de la Tierra. Con lo cual, o bien no estará en reposo en el centro, o bien, si reposa allí, reposará sin tener el centro, al que ahora es naturalmente llevada por aptitud propia.

Esto es, pues, el motivo de las dificultades. No es difícil ver, por poco que ello se medite y se distinga, de qué manera concebimos nosotros el que una magnitud cuan grande se quiera que posee un peso, es llevada hacia el centro. Es evidente que no es llevada hasta que toque el centro mismo, sino que es necesario que salga vencedora la parte mayor, hasta que con su centro comprenda el mismo centro. Pues tan sólo hasta este punto tiene movimiento. No importa que esto se diga de una gleba o una partícula mínima, o que se diga de toda la Tierra. Porque lo que hemos explicado que sucede no ocurre a causa de la pequeñez o grandeza, sino se ha dicho de todo aquello que tiene un movimiento tendencial hacia el centro. De manera que, tanto si se mueve desde algún lugar la Tierra toda como si tan sólo se mueve por partes, es necesario que ella sea llevada hasta tanto que ocupe el centro, de una manera proporcional desde todas partes, una vez se hayan equilibrado las partes menores con las mayores, según el empuje de la inclinación.

Por tanto, si ha sido hecha, es necesario que haya sido hecha de esta manera, con lo que es evidente que su generación debió de ser esférica; y si es ingénita y siempre está en reposo, reúne las mismas condiciones que hubiera tenido al comienzo, si hubiera tenido comienzo. Por este motivo, pues, es necesario que su forma sea esférica, y también porque todos los seres pesados son llevados hacia ella según ángulos semejantes, aun no equidistantes. Y esto es apropiado a lo que por naturaleza es esférico. En consecuencia, o bien es esférica, o bien es esférica por naturaleza. Y es conveniente expresar cada ser tal como suele ser por naturaleza y lo que es por naturaleza, pero no aquello que es en contra de la naturaleza y por violencia.

También según aquello que se percibe sensorialmente. Ya que de lo contrario los eclipses o fases de la Luna no tendrían las divisiones que tiene. Mientras que ahora, en las distintas figuras o formas que toma a lo largo del mes, recibe todas las divisiones. Viene a ser, en efecto, recta y luego curva por una y otra parte (es decir, convexa), y luego cóncava. En las fases, con todo, siempre conserva una línea convexa que la distingue. Por tanto, puesto que se eclipsa o adquiere sus fases por la interposición de la Tierra, sin duda la causa de esta forma es la circunferencia de la Tierra.

Más aún: según lo que la vista nos enseña de las estrellas, es evidente que no sólo es esférica, sino que además su volumen o su mole no es grande. Pues si se

[9] El movimiento significa el empuje de las olas, según la raíz del verbo.

produce una ligera desviación hacia el mediodía o el Sur y hacia la Osa, el límite de la órbita se manifiesta distinto: de manera que las estrellas que tenemos encima de la cabeza sufren un gran cambio y no parecen las mismas yendo hacia el mediodía que yendo hacia la Osa. En efecto, algunas estrellas se ven en Egipto y cerca de Chipre, mientras que en los lugares que están hacia las Osas no se ven; y las estrellas que se ven siempre en lugares cercanos o hacia la parte de la Osa se ponen en Egipto y Chipre. Por consiguiente, es por todo ello evidente que no sólo es esférica la Tierra, sino también que su mole esférica no es muy grande. Porque no tendría lugar tan rápidamente este cambio con sólo haber efectuado una desviación o desplazamiento tan breves.

Así pues, los que creen que el lugar aquel que está cerca de las columnas de Hércules está unido al lugar que hay cerca de la región índica, y que de esta manera afirman que hay un solo mar, no parecen creer cosas tan increíbles. Dicen esto aventurando una conjetura que derivan de la existencia de los elefantes, ya que su género existe en ambos lugares, puesto que ambos extremos están así relacionados por una unión.

También los matemáticos que han intentado medir la circunferencia de la Tierra dicen que la Tierra está ceñida por cuatrocientos mil estadios. De lo cual se deduce, si se tiene en cuenta la conjetura, que la masa de la Tierra no sólo es esférica, sino que además no es necesario que su magnitud, comparada con la magnitud de las demás estrellas, sea grande.

DIÁLOGO ACERCA DE DOS NUEVAS CIENCIAS

GALILEO GALILEI

[Publicado en Galileo Galilei, *Diálogo acerca de dos nuevas ciencias*, anotado por el doctor Teófilo Isnardi, traducción de José San Román, Buenos Aires, Losada, 1945, pp. 197-207]

JORNADA TERCERA

En torno de *los movimientos locales*
Interlocutores: Salviati, Sagredo, Simplicio

[190]

Vamos a instituir una ciencia nueva sobre un tema muy antiguo. Tal vez no haya, en la naturaleza, nada más antiguo que el movimiento; y acerca de él son numerosos y extensos los volúmenes escritos por los sabios (*philosophis*). Sin embargo, entre sus propiedades (*symptomatum*), que son muchas y dignas de saberse, encuentro yo no pocas que todavía no han sido observadas ni demostradas hasta ahora. Se ha fijado la atención en algunas que son de poca importancia, como por ejemplo, que el movimiento natural [libre] de los graves en descenso se acelera continuamente; sin embargo, no se ha hallado hasta ahora en qué proporción se lleve a cabo esta aceleración pues nadie, que yo sepa, ha demostrado que los espacios, que un móvil en caída y a partir del reposo recorre en tiempos iguales, retienen entre sí la misma razón que tiene la sucesión de los números impares a partir de la unidad. Se ha observado que las armas arrojadizas o proyectiles describen una línea en cierto modo curva; sin embargo, nadie notó que esa curva era una parábola. Yo demostraré que esto es así, y también otras cosas muy dignas de saberse; y, lo que es de mayor importancia, dejaré expeditos la puerta y el acceso hacia una vastísima y prestantísima ciencia, cuyos fundamentos serán estas mismas investigaciones, y en la cual ingenios más agudos que el mío podrán alcanzar mayores profundidades.

Dividiremos este tratado en tres partes: en la primera parte consideraremos lo referente al movimiento constante (*aequabilem*)* o uniforme; la segunda versará sobre el movimiento naturalmente acelerado; en la tercera se tratará del movimiento violento o sea de los proyectiles.

[191]

Del movimiento uniforme

Acerca del movimiento uniforme tenemos necesidad de una sola definición, que yo enunciaré del modo siguiente:

DEFINICIÓN

Entiendo por *movimiento uniforme* aquel cuyos espacios recorridos por un móvil en cualesquiera (*quibuscunque*) tiempos iguales, son entre sí iguales.

ADVERTENCIA

Me ha parecido bien añadir a la antigua definición (que llama simplemente movimiento uniforme, a aquel en que espacios iguales son recorridos en tiempos iguales) el vocablo *quibuscunque*, o sea en tiempos cualesquiera igua-

* Galileo usa indistintamente los vocablos *aequabilis* (acuable) y *uniformis* (uniforme) para designar el movimiento uniforme. A veces, como sucede en esta ocasión, usa los dos juntos: "aequabilis sea uniformis". En tales casos los traduciremos simplemente por uniformes. [N. del T.]

les; porque puede suceder que el móvil recorra espacios iguales durante tiempos iguales, y que sin embargo no sean iguales los espacios recorridos durante algunas fracciones más pequeñas, aunque entre sí iguales, de esos mismos tiempos. De la definición precedente dependen cuatro axiomas, a saber:

AXIOMA I

Tratándose de un mismo movimiento uniforme, el espacio recorrido durante un tiempo más largo, es mayor que el espacio recorrido durante un tiempo más breve.

AXIOMA II

Tratándose de un mismo movimiento uniforme, el tiempo en que un espacio mayor es recorrido, es más largo que el tiempo en que es recorrido un espacio menor.

AXIOMA III

Un espacio recorrido con mayor velocidad durante un mismo tiempo, es mayor que el espacio recorrido con menor velocidad.

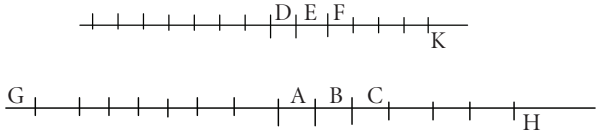
[192]

AXIOMA IV

La velocidad con que, durante un mismo tiempo, es recorrido un espacio mayor, es mayor que la velocidad con que es recorrido un espacio menor.

Teorema I. PROPOSICIÓN I

Si un móvil, que marcha con movimiento uniforme y con velocidad constante, recorre dos espacios, los tiempos de los trayectos son entre sí como los espacios recorridos.



[193]

Sea un móvil que marche con movimiento uniforme y que recorra con velocidad constante los dos espacios AB, BC; y sea DE el tiempo del movimiento por AB; y el tiempo del movimiento por BC sea EF. Digo que el espacio AB es al espacio BC, como el tiempo DE es al tiempo EF. Extiéndanse en ambas direcciones los espacios y los tiempos hacia G, H e I, K respectivamente; en AG tómese un número cualquiera de espacios iguales al mismo AB, y de modo semejante en DI un número igual de tiempos iguales al tiempo DE; en CH tómese también un número cualquiera de espacios iguales al mismo CB, y en FK un número idéntico de tiempos iguales al tiempo EF; en tal caso, el espacio BG y el tiempo EI serán múltiplos iguales (*aeque*) del espacio BA y del tiempo ED, tomados según factores cualesquiera, y del mismo modo el espacio HB y el espacio KE serán por igual múltiplos, en una multiplicación cualquiera, del espacio CB y del tiempo FE. Y como DE es el tiempo del recorrido por AB, toda la EI será el tiempo de toda la BG, dado que el movimiento se supone ser uniforme, y hay en EI tantos tiempos iguales al DE como espacios iguales al BA hay en BG; de modo semejante se concluye que KE es el tiempo de traslación por HB. Y dado que se supone ser uniforme el movimiento, si el espacio GB fuera igual al BH, también el tiempo IE sería igual al tiempo EK; y si GB es mayor que BH, también IE será mayor que EK; y si es menor, menor. Hay, por consiguiente, cuatro cantidades, AB la primera, BC la segunda, DE la tercera, EF la cuarta; y de la primera y de la tercera, es decir del espacio AB y del tiempo DE, se han tomado como múltiplos iguales, según una multiplicación cualquiera, el tiempo IE y el espacio GB; y se ha demostrado que éstos o son a la vez (una) iguales, o a la vez menores o a la vez mayores que el tiempo EK y que el espacio BH, que son por igual múltiplos de la segunda y de la cuarta; luego la primera respecto a la segunda, es decir el espacio AB respecto al espacio BC, tiene la misma razón que la tercera y la cuarta, es decir el tiempo DE respecto al tiempo EF: que es lo que se quería demostrar.

Teorema II. PROPOSICIÓN II

Si un móvil recorre dos espacios en tiempos iguales, esos espacios serán entre sí como las velocidades. Y si los espacios son como las velocidades, los tiempos serán iguales.

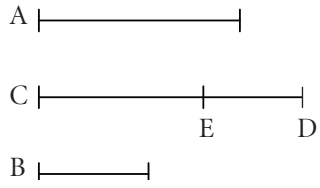
Tornando, pues, a la figura anterior, sean dos espacios AB, BC, recorridos en tiempos iguales; pero el espacio AB con la velocidad DE, y el espacio BC con la velocidad EF. Digo que el espacio AB es al espacio BC, como la velocidad DE es a la velocidad DF. Tomados, pues, de una y otra parte, como se hizo antes, equimúltiplos para factores cualesquiera, tanto de los espacios como de las velocidades; es decir GB e IE de los AB y DE, e igualmente HB, KE de los BC, EF, se deducirá, del mismo modo que antes, que los múltiplos GB, IE o son a la vez mayores, o iguales, o menores, que los múltiplos por igual

BH, EK. Por consiguiente, queda demostrado lo que pretendíamos.

Teorema III. PROPOSICIÓN III

Los tiempos de dos móviles que recorren un mismo espacio con velocidades desiguales, están en razón inversa con las velocidades.

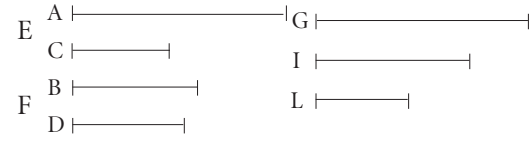
[194]
Sean las velocidades desiguales, A la mayor, B la menor. Y según una y otra efectúese el movimiento por el mismo espacio CD. Digo que el tiempo en que un móvil a la velocidad A recorre el espacio CD, es al tiempo en que a la velocidad B recorre el mismo espacio, como la velocidad B es a la velocidad A. Sea CD a CE como A es a B; será, pues, de acuerdo a la que precede, el tiempo con que la velocidad A cumple CD, igual al tiempo con que la velocidad B cumple CE. Pero el tiempo con que a la velocidad B cumple CE, es al tiempo con que a la misma cumple CD, como CE es a CD; por consiguiente, el tiempo con que a la velocidad A recorre CD, es al tiempo con que a la velocidad B recorre el mismo CD, como CE es a CD, esto es, como la velocidad B es a la velocidad A: que era lo intentado.



Teorema IV. PROPOSICIÓN IV

Si dos móviles marchan con movimiento uniforme, pero con velocidades desiguales, los espacios recorridos por los mismos en tiempos desiguales tendrán una razón compuesta de la razón de las velocidades y de la razón de los tiempos.

Sean los móviles E, F que se mueven con movimiento uniforme, y la razón que tiene la velocidad del móvil E a la velocidad del móvil F, sea como A es a B; pero la razón entre el tiempo con que se mueve E, y el tiempo con que se mueve F, sea como C es a D. Digo que el espacio recorrido por E con velocidad A en tiempo C, respecto al espacio recorrido por F con velocidad B en tiempo D, tiene la razón compuesta de la razón de la velocidad A a la velocidad B, y de la razón del tiempo C al tiempo D. Sea G el espacio recorrido por E con velocidad A en tiempo C; y sea G a I como la velocidad A es a la velocidad B; además sea I a L como el tiempo C al tiempo D. Se infiere que I es el espacio por el que se mueve F durante el mismo tiempo en que E se mueve por G, dado que los espacios G, I son entre sí como las velocidades A, B. Y siendo I a L como el tiempo C es al tiempo D; al ser, además, I el espacio recorrido por el móvil F en el tiempo C; será L el espacio recorrido por F en el tiempo D con velocidad B. Pero la razón de G a L se compone de las razones de G a I y de I a L, esto es de la razón de la velocidad A a la velocidad B, y de la del tiempo C al tiempo D: luego queda de manifiesto lo propuesto.



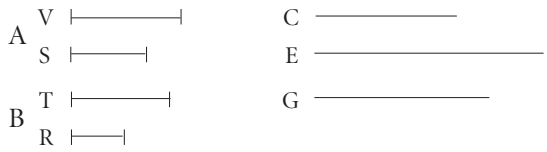
[195]

Teorema V. PROPOSICIÓN V

Si dos móviles marchan con movimiento uniforme, pero son desiguales las velocidades y desiguales los espacios recorridos, la razón de los tiempos será igual a la razón de los espacios por la razón inversa de las velocidades.

Sean dos móviles A, B, y sea la velocidad de A a la velocidad de B como V a T; y los espacios recorridos sean como S a R. Digo que la razón que hay entre el tiempo con que se ha movido A y el tiempo con que se ha movido B, es la razón compuesta de la velocidad T a la velocidad V, y de la razón del espacio S al espacio R. Sea C el tiempo del movimiento A, y sea el tiempo C al tiempo E como la velocidad T es a la velocidad V; y al ser C el tiempo en que A con velocidad V recorre el espacio S, y siendo el tiempo C al tiempo E como la velocidad T del móvil B es a la velocidad V, será el tiempo E aquél en que el móvil B recorrería el mismo espacio S. Sea, ahora, el tiempo E al tiempo G como el espacio S es al espacio R: se deduce, que G es el tiempo en que B recorrería el espacio R. Y como la razón de C a G se compone de las razones de C a E y de E a G; y como la razón de C a E es idéntica con la razón de las velocidades de los móviles A, B tomadas inversamente, esto es con la razón de T a V; y como la ra-

zón de E a G es la misma que la razón de los espacios S, R; queda de manifiesto lo que nos proponíamos.

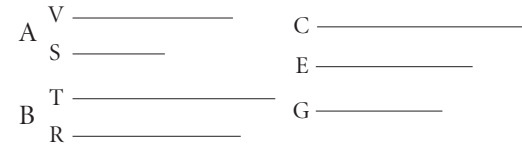


[196]

Teorema VI. PROPOSICIÓN VI

Si dos móviles marchan con movimiento uniforme, la razón de sus velocidades será igual a la razón compuesta de la razón de los espacios recorridos y de la razón de los tiempos tomados inversamente.

Sean los móviles A, B, que marchan con movimiento uniforme; y estén los espacios recorridos por ellos en la razón de V a T, y los tiempos estén como S a R. Digo que la velocidad del móvil A respecto a la velocidad del B, tiene una razón compuesta de la razón del espacio V al espacio T, y del tiempo R al tiempo S. Sea la velocidad C aquélla con que el móvil A recorre el espacio V en el tiempo S, y la velocidad C tenga respecto a E la misma razón que tiene el espacio V al espacio T; será E la velocidad con que el móvil E recorre el espacio T en el mismo tiempo S. Ahora bien, si la velocidad E es a la velocidad G como el tiempo R al tiempo S, será la velocidad G aquélla con la cual el móvil B recorre el espacio T en el tiempo R. Por consiguiente, tenemos la velocidad C, con la que el móvil A recorre el espacio V en el tiempo S, y la velocidad G, con la que el móvil B recorre el espacio T en el tiempo R; y la razón de C a G está compuesta de las razones de C a E y de E a G; y la razón de C a E es la misma que la del espacio V al espacio T; y la razón de E a G es idéntica con la razón de R a S: por consiguiente, queda de manifiesto lo propuesto.



SALVIATI. Esto que hemos visto, es cuanto escribió nuestro Autor acerca del movimiento uniforme.

LA RELATIVIDAD

ALBERT EINSTEIN

[Publicado en Albert Einstein, *La relatividad*, cap. 7, traducción de Ute Schmidt de Cepeda, México, Grijalbo, 1970, pp. 33-46]

Capítulo 7
LA APARENTE INCOMPATIBILIDAD ENTRE LA LEY DE LA PROPAGACIÓN DE LA LUZ Y EL PRINCIPIO DE LA RELATIVIDAD

Difícilmente hay en la física una ley más simple que la de la propagación de la luz en el espacio vacío. Cualquier alumno de una escuela media sabe, o cree saber, que la luz se propaga en línea recta con una velocidad *c* de 300 000 km/s. En todo caso, sabemos con una gran exactitud que esta velocidad es la misma para todos los colores; ya que, si no fuera así, el mínimo de la luz emitida por el objeto luminoso de una estrella fija doble no se podría observar simultáneamente para los diferentes colores, en el momento en que dicho objeto luminoso es ocultado por su acompañante opaco. Por medio de una consideración análoga, refiriéndose a las observaciones hechas acerca de las estrellas dobles, el astrónomo holandés De Sitter ha podido demostrar que la velocidad de propagación de la luz no puede depender de la velocidad con la cual se mueve la fuente luminosa. La hipótesis de que esa velocidad de propagación depende de la dirección "en el espacio" es intrínsecamente improbable.

Brevemente, supongamos por ahora que nuestro alumno de enseñanza media tiene razones para aceptar la simple ley de la propagación de la luz con una velocidad constante (en el vacío). ¿Quién podría pensar que esta simple ley ha colocado al físico consciente y reflexivo en las mayores dificultades? Esas dificultades han surgido de la manera siguiente.

El proceso de la propagación de la luz, como cualquier otro proceso, tiene que estar vinculado naturalmente con un cuerpo de referencia rígido (con un sistema de coordenadas). Podemos volver a escoger como sistema de referencia a nuestro terraplén de ferrocarril y suponemos ahora que el aire que se encontraba arriba de dicho terraplén ha sido extraído por medio de una bomba de succión. Supongamos también que, a lo largo del terraplén, se envía un rayo de luz que se propaga, con respecto al terraplén, con la velocidad *c*. Supongamos, además, que sobre la vía del ferrocarril se desplaza nuestro vagón con la velocidad *v*, en el mismo sentido en el que se propaga el rayo de luz; pero, naturalmente, con una velocidad mucho menor que la de la luz. Entonces, podemos preguntarnos, ¿cuál es la velocidad de propagación del rayo luminoso con respecto al vagón? Es fácil advertir que aquí se puede aplicar la consideración hecha en el capítulo 6, porque el hombre que se desplaza a lo largo del vagón del tren en marcha, en el mismo sentido que dicho tren, desempeña el papel del rayo de luz. Su velocidad *W* con respecto al terraplén es sustituida aquí por la velocidad de la luz con respecto al propio terraplén; la velocidad de la luz que se busca, con respecto al vagón, es *w*, cuyo valor es:

$$\omega = c - \upsilon$$

Por consiguiente, la velocidad de propagación del rayo luminoso, con respecto del vagón, resulta ser menor que *c*.

Sin embargo, este resultado se encuentra en contradicción con el principio de la relatividad expuesto en el capítulo 5. De acuerdo con el principio de la relatividad, la ley de la propagación de la luz en el vacío, al igual que cualquier otra ley general de la naturaleza, debería ser la misma, independientemente de que se escoja como cuerpo de referencia el vagón o la vía del ferrocarril. No obstante, de acuerdo con nuestra reflexión, tal cosa parece imposible. Ya que, si todo rayo luminoso se propaga con la velocidad *c*, con respecto al terraplén, entonces, por eso mismo, la ley de la propagación debería ser diferente con respecto al vagón, lo cual se encuentra en contradicción con el principio de la relatividad.

Ante este dilema, parece inevitable renunciar al principio de la relatividad, o bien, abandonar la ley simple de la propagación de la luz en el vacío. El lector que haya seguido atentamente nuestra exposición hasta aquí, esperará seguramente que se mantenga el principio de la relatividad, que aparece ante nuestro espíritu como algo enteramente natural, simple y casi ineluctable; y que la ley de la propagación de la luz en el vacío sea sustituida

por una ley más complicada, que resulte compatible con el principio de la relatividad. Sin embargo, el desarrollo de la física teórica ha demostrado que ese camino no es practicable. Las investigaciones teóricas sumamente originales de H. A. Lorentz, con respecto a los procesos electrodinámicos y ópticos que se producen en los cuerpos en movimiento, han demostrado, en efecto, que las experiencias realizadas en ese dominio conducen necesariamente a una teoría de los procesos electromagnéticos que tiene, como consecuencia inevitable, la constancia de la velocidad de la luz en el vacío. Por eso los teóricos más eminentes se han inclinado más bien a rechazar el principio de la relatividad, aun cuando no se haya podido encontrar un solo resultado experimental que lo contradiga.

En este punto es en donde intervino la teoría de la relatividad. Mediante un análisis de los conceptos físicos de tiempo y de espacio, dicha teoría ha demostrado que, *en realidad, no existe incompatibilidad alguna entre el principio de la relatividad y la ley de la propagación de la luz*; y que, por lo contrario, manteniendo de una manera firme y sistemática esos dos principios, se llega a establecer una teoría lógica que se encuentra a salvo de cualquier objeción. A dicha teoría le hemos dado el nombre de "teoría de la relatividad restringida", para distinguirla de la teoría más general que trataremos más adelante. Ahora vamos a exponer las ideas fundamentales de la teoría de la relatividad restringida.

Capítulo 8
SOBRE EL CONCEPTO DE TIEMPO EN LA FÍSICA

Supongamos que en dos puntos *A* y *B*, muy distantes el uno del otro, de nuestra vía de ferrocarril, ha caído un rayo; y que además, afirmamos que esos dos rayos han sido "simultáneos". Si ahora te pregunto, querido lector, si esa afirmación tiene un significado, me responderás convencido que "sí". Pero, si te insisto y te pido que me expliques de un modo más preciso el significado de esa afirmación, entonces advertirás después de cierta reflexión, que la respuesta a esa pregunta no es tan simple como parece a primera vista.

Después de algún tiempo, quizás te venga a la mente la respuesta siguiente: "El significado de esa afirmación es claro en sí mismo y no necesita de mayores aclaraciones; pero, por lo demás, tendría que reflexionar bastante si estuviera encargado de establecer por medio de observaciones si, en ese caso concreto, los dos acontecimientos se realizaron simultáneamente o no." Sin embargo, esa respuesta no me satisface, por las razones siguientes: supongamos que un meteorólogo haya encontrado, a través de reflexiones penetrantes, que los rayos deben caer siempre simultáneamente en los puntos *A* y *B*; entonces, es necesario comprobar si este resultado teórico corresponde o no corresponde a la realidad. Semejante comprobación se requiere para todos los enunciados físicos en los cuales desempeña algún papel el concepto de "simultaneidad". Dicho concepto existe para el físico solamente cuando encuentra la posibilidad de verificar, en el caso concreto de que se trate, si el concepto es o no es exacto. Por lo tanto, es necesaria una definición de la simultaneidad tal que nos suministre un método por medio del cual podamos decidir, en el caso en cuestión, a través de experimentos, si los dos rayos han sido o no han sido simultáneos. Mientras no se cumpla con esa exigencia, soy víctima como físico (y también no siendo físico) de una ilusión, al creer que se puede asociar un significado a la afirmación de la simultaneidad. (Si no quedas de acuerdo con esto, querido lector, con plena convicción, entonces es inútil que sigas adelante.)

Después de cierto tiempo de reflexión, podrías hacerme la siguiente proposición para comprobar la simultaneidad. Podemos medir la recta AB a lo largo de la vía férrea y colocar en el punto medio *M* de dicha recta un observador provisto de un aparato (que podría consistir, por ejemplo, en dos espejos que formen un ángulo de 90°) que le permita observar simultáneamente los dos puntos *A* y *B*. Si este observador percibe los dos relámpagos al mismo tiempo, entonces dichos rayos son simultáneos.

Aunque estoy muy satisfecho con esta proposición, sin embargo, no puedo considerar que el problema esté completamente aclarado, porque me siento obligado a presentar la siguiente objeción: "Tal definición sería enteramente correcta, si yo supiera ya que la luz que comunica al observador situado en *M* la percepción de los dos relámpagos se propaga con la misma velocidad sobre la recta $A \rightarrow M$ que sobre la recta $B \rightarrow M$. Pero, una verificación de esta suposición únicamente sería posible si ya dispusiéramos de un medio para medir el tiempo. Entonces, parece que nos movemos aquí en un círculo vicioso."

Después de hacer algunas otras reflexiones, me mirarías con razón con un cierto aire desdeñoso, diciéndome: "A pesar de todo mantengo mi definición anterior, ya que, en realidad, no presupone nada sobre la luz. La definición de la simultaneidad debe cumplir una sola condición, que es la de suministrar, en cada caso real, un medio empírico para decidir si el concepto por definir se confirma o no se confirma. Es indiscutible que mi definición

cumple con esta condición. La afirmación de que la luz necesita el mismo tiempo para recorrer la recta $A \rightarrow M$ que sobre la recta $B \rightarrow M$, no es realmente una suposición o una hipótesis sobre la naturaleza física de la luz, sino una convención que yo puedo fijar libremente, para establecer una definición de la simultaneidad."

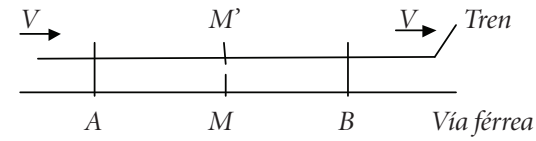
Es claro que esta definición puede ser usada no sólo para dar un significado exacto a la simultaneidad de dos acontecimientos, sino de un número cualquiera de acontecimientos, independientemente de la posición relativa que ocupen los lugares en donde se producen dichos acontecimientos, con respecto al cuerpo de referencia (que en este caso es el terraplén).[1] De esta manera se llega también a establecer una definición del "tiempo" en la física. En efecto, imaginémonos que en los puntos *A*, *B*, *C*, de la vía férrea (sistema de coordenadas) estén colocados relojes de la misma construcción y ajustados de tal modo que las posiciones respectivas de sus manecillas sean simultáneas (en el sentido antes indicado). Entonces, se entiende por "tiempo" de un acontecimiento la indicación (posición de las manecillas) del reloj que se encuentre en la vecindad inmediata del acontecimiento. De este modo, a cada acontecimiento se encuentra asociado un valor del tiempo que es, en principio, observable.

Esta convención contiene una hipótesis física cuya validez casi no puede ser puesta en duda, además, mientras no haya alguna prueba empírica en contrario. Se supone, en efecto, que todos esos relojes "marchan al mismo ritmo" si son de la misma construcción. En términos más precisos, si dos relojes en reposo, colocados en lugares distintos del cuerpo de referencia, están ajustados de tal modo que la posición de las manecillas del otro son simultáneas (en el sentido antes indicado), entonces las posiciones iguales de las manecillas son siempre simultáneas (en el sentido de la definición dada anteriormente).

[1] Suponemos además que, si tres acontecimientos, *A*, *B* y *C*, se realizan en tres lugares diferentes, de tal manera que *A* y *B* son simultáneos y que *B* y *C* también son simultáneos (simultáneos en el sentido de la definición anterior), entonces el criterio de la simultaneidad se cumple igualmente para la pareja de acontecimientos *A* y *C*. Esta suposición es una hipótesis física con respecto a la ley de la propagación de la luz; y debe ser absolutamente verdadera, si se quiere tener la posibilidad de conservar la ley de la constancia de la velocidad de la luz en el vacío.

Capítulo 9
LA RELATIVIDAD DE LA SIMULTANEIDAD

Hasta aquí hemos referido nuestras reflexiones a un cuerpo de referencia particular, al que hemos designado como "vía de ferrocarril". Ahora bien, supongamos que en esa vía avanza un tren muy largo, con una velocidad constante v y en el sentido indicado en la figura 1. Los viajeros que se encuentran a bordo de este tren pueden utilizar con ventaja al propio tren como cuerpo de referencia rígido (sistema de referencia), para referir a éste todos los acontecimientos. Por lo tanto, cualquier acontecimiento que ocurra a lo largo de la vía férrea tiene lugar también en un punto determinado del tren. La definición de la simultaneidad también puede ser formulada exactamente de la misma manera con respecto al tren, que con respecto a la vía férrea. Entonces, se plantea de un modo natural la cuestión siguiente:



¿Dos acontecimientos (por ejemplo, los dos relámpagos *A* y *B*) que son simultáneos *con respecto a la vía*, también son simultáneos *con respecto al tren*? Mostraremos en seguida que la respuesta debe ser negativa.

Cuando decimos que los relámpagos *A* y *B* son simultáneos con respecto a la vía férrea, eso significa que los rayos luminosos que parten de los puntos *A* y *B* se encuentran en el punto medio *M* de la distancia *A–B*, situada sobre la vía. Pero a los acontecimientos *A* y *B* corresponden también los lugares *A* y *B* en el tren. Sea *M'* el punto medio de la recta *A–B* del tren en marcha. Es cierto que este punto *M'* coincide con el punto *M* en el instante en que se producen los relámpagos,[2] pero, en el diagrama, dicho punto *M'* se desplaza hacia la derecha con la velocidad *v*. Si un observador colocado en el punto *M'* del tren no se estuviera moviendo con esa velocidad, entonces se mantendría permanentemente en *M* y los rayos luminosos que parten de *A* y *B* lo alcanzarían simultáneamente, es decir, que los dos rayos se encontrarían justamente en el punto en donde está colocado el observador. Sin embargo, el observador (visto desde el terraplén) avanza en realidad hacia el rayo de luz proveniente de *B*), mientras que se adelanta al rayo de luz prove-

[2] Visto desde el terraplén.

niente de *A*. Por consiguiente, el observador verá el rayo de luz proveniente de *B* antes que el rayo luminoso proveniente de *A*. Los observadores que utilizan el tren como cuerpo de referencia, deben llegar a la conclusión de que el relámpago *B* se produjo antes que el relámpago *A*. Arribamos, por lo tanto, al importante resultado que sigue:

Dos acontecimientos, que son simultáneos con respecto a la vía férrea no son simultáneos con respecto al tren, y recíprocamente (relatividad de la simultaneidad). Cada cuerpo de referencia (sistema de coordenadas) tiene su tiempo propio: una indicación de tiempo sólo tiene significado cuando indica el cuerpo de referencia al que se refiere.

Antes de la teoría de la relatividad, la física suponía siempre tácitamente que la indicación del tiempo tenía un valor absoluto, es decir, que era independiente del estado de movimiento del cuerpo de referencia. Pero ahora acabamos de demostrar que esa suposición es incompatible con la definición tan natural de la simultaneidad; si se la rechaza, entonces desaparece el conflicto expuesto en el capítulo 7, entre la ley de la propagación de la luz en el vacío y el principio de la relatividad.

A este conflicto llevan en efecto las reflexiones hechas en el capítulo 6, que ahora ya no se pueden sostener. Basados en el hecho de que el viajero recorre la distancia *w en un segundo* con respecto al vagón del tren, hemos llegado a la conclusión de que recorrerá esa distancia igualmente *en un segundo* con respecto a la vía. Pero, de acuerdo con las reflexiones que acabamos de hacer, la duración de un acontecimiento determinado con respecto al vagón no puede ser igual a la duración de ese mismo acontecimiento con respecto a la vía considerada como cuerpo de referencia; sin que se pueda sostener que el viajero al andar haya recorrido la distancia *w*, con respecto a la vía, en un tiempo que —medido con respecto a la vía— es igual a un segundo.

El razonamiento del capítulo 7 se apoya, además, en otra suposición que, de acuerdo con una reflexión estricta, parece arbitraria, a pesar de haber sido hecha siempre (tácitamente) antes de la formulación de la teoría de la relatividad.

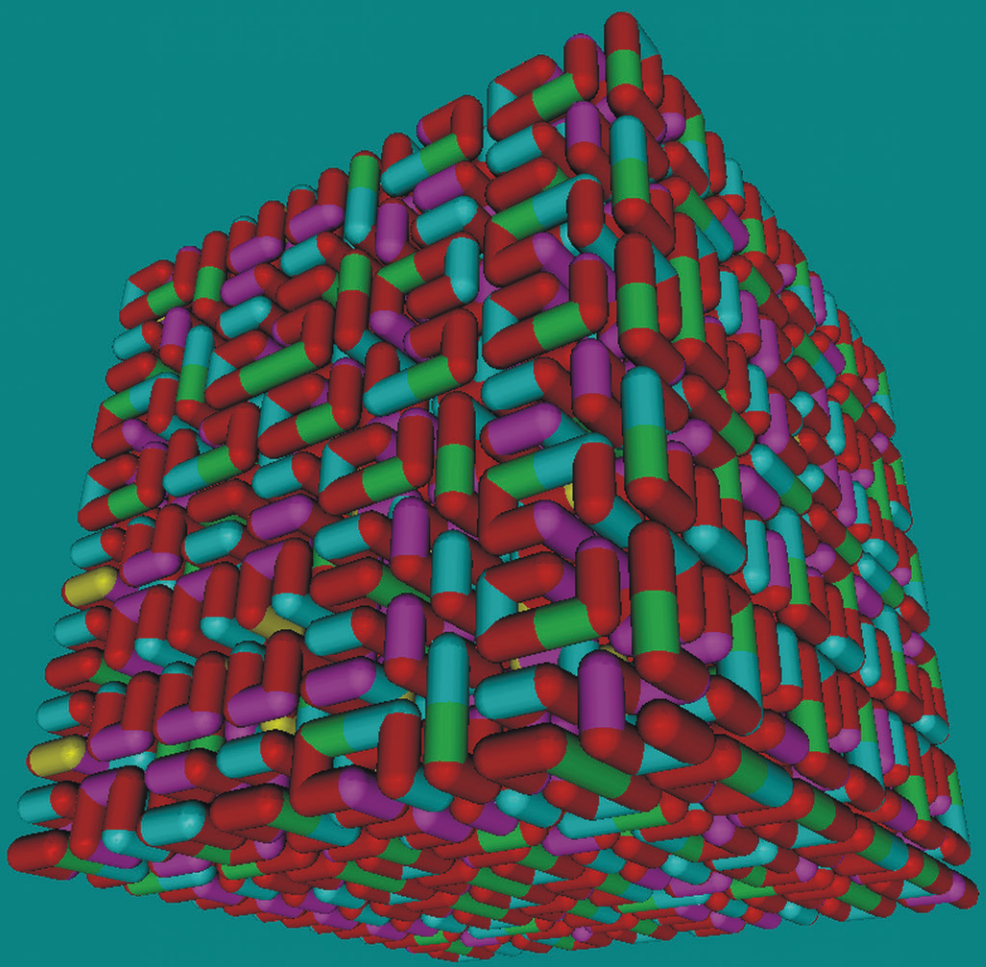
Capítulo 10
LA RELATIVIDAD DEL CONCEPTO DE DISTANCIA ESPACIAL

Consideremos dos puntos determinados del tren[3] que avanza con la velocidad *v* a lo largo del terraplén y preguntémonos cuál es su distancia. Ya sabemos que, para medir una distancia, se necesita un cuerpo de referencia con respecto al cual se mide la distancia. El modo más simple es el de utilizar el tren mismo como cuerpo de referencia (sistema de coordenadas). Un observador que viaja en el tren mide la distancia colocando sucesivamente su regla de medir, en línea recta, a lo largo de los pisos de los vagones, tantas veces como sea necesario para llegar desde un punto determinado hasta otro igualmente determinado. El número que indica cuántas veces se ha colocado la regla de esa manera representa la distancia buscada.

Cuando se trata de medir esa misma distancia sobre el terraplén, el problema es completamente distinto. En tal caso, se puede emplear el siguiente método: se denominan *A*' y *B*' a los puntos del tren cuya distancia se trata de determinar y que se desplazan a lo largo del terraplén con la velocidad *v*. Desde luego, nos preguntamos cuáles son los puntos *A* y *B* del terraplén ante los cuales pasan los puntos *A*' y *B*', justamente en un tiempo determinado *t* (con respecto al terraplén). Esos puntos *A* y *B* del terraplén pueden ser determinados gracias a la definición del tiempo dada en el capítulo 8. Después se mide la distancia entre esos puntos *A* y *B*, colocando sucesivamente un cierto número de veces la unidad de medida a lo largo del terraplén.

No se puede demostrar *a priori* que esta última medición dará el mismo resultado que la primera. La longitud del tren, medida desde el terraplén, puede ser diferente de la longitud medida en el tren mismo. Esta circunstancia suscita una segunda objeción en contra del razonamiento, aparentemente tan evidente, del capítulo 6. Si el viajero recorre en el vagón la distancia ω en la unidad de tiempo, medida en el tren, esa distancia, cuando es medida sobre el terraplén, no necesariamente es igual a ω.

[3] Por ejemplo, el punto medio del primer vagón y el punto medio del centésimo vagón.



LOS AUTORES

ERNESTO BRIBIESCA CORREA

Tiene la licenciatura en ingeniería en comunicaciones y electrónica del Instituto Politécnico Nacional y obtuvo el grado de doctor en ciencias en el Departamento de Matemáticas en la Universidad Autónoma Metropolitana de Iztapalapa (donde le otorgaron la Medalla al Mérito Universitario). Ha trabajado en diferentes lugares, principalmente en centros de investigación, entre los que se destacan: el Centro Científico IBM, el INEGI e IIMAS-UNAM; pertenece al Sistema Nacional de Investigadores, con el nombramiento de Investigador Nacional Nivel 3. Entre sus estancias de investigación se distinguen dos: Laboratorio de Inteligencia Artificial del MIT y el Earth Resouces Laboratory de la National Aeronautics and Space Administration (NASA).

JOSÉ GALAVIZ CASAS

Obtuvo el título de matemático, el grado de maestro y el de doctor en ciencias de la computación en la UNAM (este último con mención honorífica). Es profesor de la Facultad de Ciencias de la misma universidad desde 1998. Ha escrito, previos a esta colaboración, dos libros: uno de texto a nivel licenciatura y otro de divulgación acerca de la historia de la computación. En 2007 obtuvo la Distinción Universidad Nacional para Jóvenes Académicos en el área de docencia en ciencias exactas.

SERGIO RAJSBAUM

Obtuvo el grado de ingeniero en computación en la Universidad Nacional Autónoma de México (UNAM) y el de doctor en ciencias de la computación por el Instituto Tecnológico de Israel. Desde 1991 es investigador del Instituto de Matemáticas de la UNAM. Ha tenido varias estancias de investigación en el Massachusetts Institute of Technology (MIT), los laboratorios de investigación de HP y de IBM, y otras universidades. Su área principal de investigación es el cómputo distribuido, especialmente problemas relacionados con coordinación, sincronización, tiempo y tolerancia a fallas, muchos de los cuales tienen que ver con internet y la web. También trabaja en algoritmos y otros problemas matemáticos relacionados con la computación y sus fundamentos. Ha publicado más de 70 trabajos de investigación, en colaboración con investigadores de Estados Unidos, Francia e Israel, principalmente.

FRANCISCO SOLSONA Estudió ciencias de la computación en la Facultad de Ciencias de la UNAM, obteniendo el promedio más alto de su generación. Actualmente es coordinador de Servicios de Cómputo de la Facultad de Ciencias y coautor y editor del libro en línea: *Scheme Cookbook*. Desde 2002 es editor de los Scheme Request for Implementation (SRFI) y coordinador del libro *Manual de supervivencia Linux*, publicado por las Prensas de Ciencias en 2007. Desde 2005, ha liderado diversos proyectos de vinculación: auditorías de riesgo tecnológico y capacitación para programadores, entre otros.

AGRADECIMIENTOS

Suponer que la escritura de un texto como el que ponemos a consideración del lector es una empresa exclusiva de los autores es una trivialización. El texto y los inevitables errores contenidos en él los debemos, ciertamente, a los autores, pero la labor de éstos no hubiera sido posible sin la generosidad de muchas personas e instituciones. En el rubro de estas últimas están el Instituto de Matemáticas, el de Investigaciones en Matemáticas Aplicadas y en Sistemas, y la Facultad de Ciencias, donde se encuentran adscritos los autores; también debemos agradecer al Instituto de Astronomía, la Dirección General de Servicios de Cómputo y de Tecnologías de Información y Comunicación, la Secretaría de Desarrollo Institucional y la Dirección General de Publicaciones, quienes ofrecieron el apoyo logístico y técnico necesario. En el rubro de las personas están, por supuesto, en primer lugar nuestras familias, a quienes tuvimos que robarles tiempo durante la labor. A Luis A. Martínez, por su asistencia en la coordinación y porque se mantuvo junto a los autores y editores realizando una denodada e invaluable labor, determinante para alcanzar la meta. Y, finalmente, a quienes hicieron valiosas aportaciones al contenido del libro: al doctor Jesús Savage por sus valiosas contribuciones, que se convirtieron en la sección de "Robótica" y "Los robots en la literatura" en el módulo de Aplicaciones; al maestro en ciencias Elio Vega Munguía por la información proporcionada con relación a la Sala Ixtli, y a la doctora Graciela Bribiesca Correa, profesora de la Facultad de Contaduría y Administración de la UNAM, por sus sugerencias en el tema de la "Computación en los negocios" del módulo de Aplicaciones.

p. 502, Cuarto estado de la curva de Hilbert, representado el código de cadenas 3D |
© Ernesto Bribiesca.

COMPUTACIÓN

INTRODUCCIÓN

DE LOS AUTORES

El ser humano tiene dos modos de interactuar con el mundo. Con su cuerpo puede empujar, cortar, construir, hacer hoyos y diques. Usando este modo de interacción, logró defenderse de las inclemencias de la naturaleza y procurarse alimentos en los albores de la historia. Pero los animales lo superan por mucho en habilidades físicas. Siempre hemos soñado con volar como un pájaro, correr con la gracia y velocidad de una gacela, tener la fuerza de un elefante, el olfato de un perro, la vista de un águila. Es otro tipo de habilidades, sin embargo, en las que el hombre es el rey de la naturaleza; éstas no se pueden ver, ni tocar, ni oler, ni oír, pero le permiten al hombre, con su pequeñez y debilidad, hacer grandes proezas. En este mundo no sirve de nada ser muy fuerte. En este mundo un médico puede hacer una operación quirúrgica en un paciente que se encuentra a cientos de kilómetros de distancia, sin mover un solo dedo. El médico recibe imágenes y otros datos del paciente, y emite instrucciones que indican la manera en que distintos instrumentos deben moverse. Lo esencial no es la manera en que el médico expresa estas instrucciones —podría ser mediante movimientos de sus dedos sobre una palanca o pulsando teclas, pero también hablando— sino la *información* que produce a partir de lo que observa o escucha, y la cual se envía al quirófano. La habilidad del médico a la que nos referimos es la de procesar información. Para comunicar esta información el médico requiere de *lenguajes*, ya sea para hablar con otros médicos o para interactuar con instrumentos. Y su herramienta principal es su cerebro, no sus músculos. El médico utiliza su cerebro para procesar la información, ayudándose del conocimiento acerca de la medicina, aprende asimilando la experiencia de cada operación realizada, ejecuta algoritmos, métodos mediante los cuales estima la presión sanguínea del paciente, evalúa las implicaciones de distintas posibilidades a través de razonamientos lógicos. El médico resuelve los distintos problemas que se presentan durante la operación, algunos relacionados con redes de trabajo en equipo, por ejemplo para coordinar la ejecución de varias tareas simultáneamente. Éste es el mundo de la computación: información, conocimiento, lenguajes, problemas, algoritmos, abstracción, lógica, redes. Así como el mundo de la medicina no son los bisturíes, ni el de los astrónomos los telescopios, el de la computación no son las computadoras. En cada mundo el hombre crea instrumentos para ampliar sus capacidades y poderes. En el de la computación, estos instrumentos son las computadoras, y las capacidades que aumenta son almacenar información y conoci-

miento, procesar números y otros datos a enormes velocidades, comunicarse e interactuar con el mundo y otros hombres.

Al inicio, el hombre operaba en un mundo mayoritariamente físico, apenas lograba una comunicación rudimentaria mediante algunas señas o sonidos. A lo largo de los años el hombre ha desarrollado sus habilidades mentales, en una lucha constante por dejarle las tareas físicas a las máquinas, para concentrarse él en tareas mentales. Quizás el momento decisivo fue llegar a la cima que representó la revolución industrial, en la cual el ser humano alcanzó grandes éxitos con las máquinas. Hoy vivimos la era de la información. La ubicuidad de los sistemas de cómputo es la característica principal del mundo moderno, y éstos abarcan casi todos los ámbitos de nuestra vida. Nuestros medios de transporte y comunicación son controlados mediante tales sistemas; la exploración de otros planetas es realizada por robots; sistemas que involucran satélites predicen el clima cada vez con mayor exactitud; la investigación científica, la operación de los sistemas financieros y nuestras transacciones comerciales son inconcebibles sin el uso de computadoras; el acceso a bibliotecas digitales y acervos de información se simplifica y extiende constantemente, etcétera.

La omnipresencia de los sistemas de cómputo en el mundo moderno hace indispensable que seamos capaces de adaptarnos al entorno tecnológico que nos rodea. No sólo en el sentido evidente de poder operar e interactuar con las tecnologías de la información sino también en el de comprender las ideas detrás de éstas. Se trata de una labor que va mucho más allá de aprender a usar diversas plataformas y programas de cómputo. Es comparable a la labor realizada en otras áreas científicas y de humanidades: mirar el mundo a la luz de principios fundamentales. Al igual que en otras disciplinas, el objetivo final es llegar a un mejor entendimiento de nosotros mismos y de nuestro entorno por medio de una perspectiva particular; en nuestro caso, la del computólogo. Sólo que, en cuanto a la computación, se trata de un mundo que nos involucra de manera especialmente cercana: ¡es un mundo del cual nuestras mentes forman parte! En una u otra forma, nuestros procesos mentales, la manera en que almacenamos recuerdos y los utilizamos mediante asociaciones y otras búsquedas, resolvemos problemas, entendemos lo que otra persona nos dice, escribimos y leemos, jugamos ajedrez, son todos procesos de cómputo, en el sentido amplio de la palabra al que antes nos referimos. El objetivo de este libro es presentar al mundo desde esta perspectiva, la de la computación. Cuando nos referimos a "cómputo", lo hacemos en este sentido amplio y, por tanto, el computólogo es una persona que estudia el cómputo, ya sea natural, artificial o imaginario.

Nuestra propuesta evidencia que la computación es ciencia e ingeniería: pretende entender y explicar el mundo que nos rodea por un lado y, por el otro, demostrar que los conocimientos adquiridos pueden derivar en la construcción de mecanismos útiles que permitan expandir las capacidades del ser humano. Impulsamos a los lectores a desarrollar habilidades, adquirir conocimientos de cómputo, y razonar a diversos niveles de abstracción, desde lo cotidiano del tráfico vehicular hasta las matemáticas, pasando por la programación, el funcionamiento de una computadora, internet y la web. ¡Bienvenidos al viaje!

COMPUTACIÓN

TEMA

1

© *Pamela Parker.*

Las creaturas que habitan el mundo de la computación son los datos. Éstos se manipulan para resolver problemas mediante algoritmos. Algunos problemas se pueden resolver eficientemente y otros se cree que no, aunque no se sabe a ciencia cierta. La gran mayoría no se pueden resolver en un tiempo razonable o simplemente son irresolubles.

Mantén dos verdades en tu bolsa y sácalas de acuerdo con la necesidad del momento. Por mí el mundo fue creado. Soy polvo y cenizas.
SIMCHA BUNAM, SIGLO XVIII.

1.1 INTRODUCCIÓN: ENTRE EL POLVO Y LA DIVINIDAD

Hasta antes de la Edad Media, el hombre estaba cómodamente sentado en un trono en el centro del universo, desde el cual creía reinar sobre todas las creaturas, creía haberlo hecho desde el inicio de los tiempos, lo que se suponía habría sucedido unos 5 000 años atrás. Fue a partir del Medioevo, como narra Stephen Jay Gould en *Time's Arrow Time's Cycle,* cuando el hombre inició un doloroso proceso, en el cual, golpe tras golpe, se derrumbarían sus ilusiones de ocupar un lugar trascendente en el universo. Este gran paleontólogo cita a Sigmund Freud, y señala que las principales disciplinas científicas han contribuido a la reconstrucción del pensamiento humano:

¡Oh Dios! Podría estar encerrado en una cáscara de nuez y tenerme por rey del espacio infinito.
HAMLET, II, 2, CA. 1600.

Golpe 1 (astronomía)
La Tierra no es el centro del universo. A pesar de que, siglos antes, sabios hindúes, griegos y musulmanes ya lo habían deducido, fue Copérnico quien sentó uno más de los pilares de la ciencia: el Sol es el centro del Sistema Solar, no la Tierra.

Golpe 2 (biología)
La teoría de la evolución formulada por Darwin dice que todas las especies evolucionaron de unos cuantos ancestros comunes, mediante un proceso de selección natural.

Golpe 3 (psicología)
El psicoanálisis de Freud habla del subconsciente, los procesos mentales que no se llevan a cabo conscientemente, que conducen a dudar de algunos de los razonamientos de las personas.

> La humanidad ha tenido que soportar dos fuertes golpes de la ciencia a su inocente autoestima. El primero sucedió cuando reparó que nuestra Tierra no es el centro del universo, sino sólo un granito de polvo en un universo de magnitud inimaginable. El segundo fue despertar del sueño de sentirse privilegiado, creado de forma especial, y percatarse de que también pertenece al reino animal.

Freud asevera, en una de las afirmaciones menos modestas que se han pronunciado, que su propio trabajo constituye un tercer golpe que posiblemente derribará el último pedestal de la confianza del ser humano en sí mismo: el consuelo de que, a pesar de haber evolucionado del mono, por lo menos posee una mente racional.

El desplome de la confianza que el hombre tenía sobre su supremacía fue provocado por el avance de ciencias como la física, la biología y la psicología, nos dice Gould; habría que agregar a la lista la contribución de la geología y el descubrimiento del tiempo geológico denominado "tiempo profundo", que privó al hombre de la confortable situación de pertenecer a una Tierra joven, dominada por él. Quedó desamparado en la inmensidad del tiempo, donde la existencia de la humanidad se condensa en una fracción de segundo del último instante. Como bien lo ilustró Mark Twain:

> El hombre ha estado aquí unos 32 000 años. Que se hayan invertido cien millones de años en preparar al mundo para él, es prueba de que el mundo fue hecho para eso. Supongo, no estoy seguro. Si la torre Eiffel representara la edad del mundo, la capa de pintura de la esfera en su punta representaría la parte del tiempo en la que el hombre ha existido; cualquiera percibiría que la torre fue construida para esa capa. Supongo, no estoy seguro.

Una metáfora interesante acerca del tiempo profundo aparece en el libro *Basin and Range* de John McPhee: la historia de la Tierra equivale a la distancia de la nariz del rey a la punta de su brazo extendido, como la antigua medida de la yarda inglesa. Con frotar una vez su dedo medio con una lima de uñas, borra toda la historia de la humanidad.

Valdría la pena recordar que la edad de la Tierra se estima hoy en día en unos 4 570 millones de años, aproximadamente la tercera parte de la edad del universo, estimada en 13 700 millones de años.

La Tierra | © *Reto Stökcli* NASA *Goddard Space Flight Center.*

En 1830, uno de los padres de la geología, Charles Lyell, querido amigo de Darwin, explicó acertadamente la liga metafórica entre el tiempo profundo y la amplitud del espacio en el cosmos de Newton:

> Las visiones sobre la inmensidad del tiempo pasado, desenmascaradas por la filosofía newtoniana respecto al espacio, fueron demasiado vastas como para despertar ideas de lo sublime sin que éstas se mezclaran con un doloroso sentimiento de nuestra incapacidad para concebir un plan de tan infinito alcance. Mundos que alcanzan a verse más allá de otros y que se encuentran separados por distancias inmedibles y, aún más allá, otros innumerables sistemas que apenas se alcanzan a vislumbrar en los confines del universo visible.

Golpe 4 (geología)
El padre de la geología moderna, Hutton, sintetizó la noción de tiempo geológico con su famosa frase "no encontramos vestigio de un inicio, ni prospecto de un fin".

Si se observa el macrocosmos, la Luna, el Sol, la Tierra y todo el mundo visible, el hombre no es más que un punto insignificante en la enorme cavidad del universo. Y cuando se dirige la mirada hacia abajo, al microcosmos, se encuentra otro universo en la infinitud de lo pequeño. En el espesor de un cabello humano cabe un millón de átomos de carbono. Una sola gota de agua puede contener 2 000 trillones de átomos de oxígeno (un 2 seguido de 21 ceros). Más aún, si comparamos un átomo con una catedral, el núcleo no sería más grande que una mosca. Entre estos dos infinitos, el hombre, que antes creía dominar el universo, percibe su insignificancia.

Golpe 5 (física)
Newton es considerado por muchos el mayor científico de todos los tiempos. Entre sus numerosas contribuciones, la mecánica clásica muestra que los cuerpos celestes y los de la Tierra cumplen las mismas leyes, y abre las puertas para comprender la inmensidad del universo.

En 1900, la autoestima de la humanidad recibe otro golpe con las investigaciones de Max Planck, que conducen posteriormente a la formulación de la mecánica cuántica en 1926, por un grupo de científicos, entre los que se encuentran Werner Heisenberg y Erwin Schrödinger.

La teoría cuántica reveló una limitación fundamental en la capacidad del hombre para predecir el futuro: el universo es probabilístico y no determinístico, como se pensaba. Es decir, antes del surgimiento de la mecánica cuántica se creía que cada acción ocasionaba una reacción predecible que era posible determinar en un principio si se conocían todos los detalles de una situación en un momento dado. La mecánica clásica afirma que si se conoce el valor de todas las variables que influyen, por ejemplo, en el movimiento de un automóvil, tales como velocidad, dirección, peso, volumen, estado de los neumáticos y temperatura ambiente, entre otras, es posible predecir con precisión dónde estará en el futuro y en qué condiciones. El universo no es más que una enorme máquina con un comportamiento complejo, sin duda, pero predecible. La mecánica cuántica, en cambio, postula que los fenómenos observables poseen un comportamiento influido permanentemente por el azar, y que el hecho mismo de pretender medir las variables que influyen sobre ellos los altera irremediablemente. "Dios no juega a los dados con el universo", afirmaría Albert Einstein, al pretender refutar la teoría que su propio trabajo contribuyó a fundar.

Golpe 6 (física cuántica)
La infinitud de lo pequeño y la imposibilidad de predecir el futuro.

Kurt Gödel.

En la primera mitad del siglo XX, el ego homocéntrico recibe una nueva lección. Ahora son las matemáticas, en especial la lógica, las que hacen su aparición en el trabajo de Kurt Gödel. Su legado ha tenido un enorme impacto en el pensamiento científico y filosófico. En 1931, a la edad de 25 años, demostró sus "teoremas de incompletitud", que señalan que ni siquiera en el mundo de las matemáticas es posible que el hombre lo sepa todo. Existen verdades que no se pueden demostrar, y cualquier sistema formal, lo suficientemente poderoso como para hablar de números y operaciones, se verá necesariamente limitado en algún punto.

Golpe 7 (matemáticas)
No existe ningún sistema formal suficientemente poderoso, que sea a la vez completo y consistente.

A los golpes al ego humano asestados por la física, la biología, la psicología, la geología y las matemáticas se añaden los de la computación. En 1936, mientras el hombre empieza a soñar con robots y supercomputadoras, otro jovencito de 24 años, Alan Turing, pu-

Alan Turing.

blica el artículo "On Computable Numbers, with an Application to the Entscheidungs Problem", que cuestiona las capacidades omnipotentes del ser humano para resolver problemas. Con este trabajo, y otros de Alonzo Church y Stephen Kleene, se descubre que ni siquiera con la ayuda de estas sorprendentes máquinas, en apariencia invencibles, el hombre volverá a sentir que reina sobre el universo. A pesar de que existen muchos problemas que se pueden resolver con la ayuda de las computadoras, el universo de problemas sin solución es infinitamente más grande, no sólo para las computadoras del presente sino aun para las que se inventen en el futuro.

En 1965, también en el campo de la computación, un estudio de Juris Hartmanis y Richard Stearns se concentró en determinar el tiempo requerido para resolver un problema, independientemente de si éste es computable o no. Es entonces cuando descubrimos que del pequeño mundo de los problemas computables, la gran mayoría está fuera de nuestro alcance, ya que tomaría en resolverse más tiempo que la edad del universo, inclusive con las más veloces computadoras que el ser humano posee y con las mejores que pueda llegar a inventar.

Golpe 8 (computación)
Existen límites a lo que podemos computar, y la mayoría de los problemas computables, debido a su enorme dificultad, jamás los podremos resolver.

Hoy en día la humanidad se encuentra muy lejos de la imagen antropocéntrica del mundo que prevaleció hasta el Medioevo, donde el hombre se situaba a sí mismo en el núcleo del universo, con el suelo firme de la Tierra bajo sus pies y las esferas celestes contemplándolo desde su inmensidad, habitadas por Dios y sus ángeles. Pascal, aterrorizado, decía: "El eterno silencio de estos espacios infinitos me llena de temor." Pero ésta es justamente la grandeza del hombre: ser capaz de observar estos abismos, de estudiarlos, de temerles y de maravillarse con ellos.

El centro de la Vía Láctea | © NASA/JPL-Caltech/S. Stolovy (Spitzer Science Center/Caltech).

1.2 PROBLEMAS

Se ha mencionado que los datos son las creaturas que habitan el mundo de la computación y que, por lo tanto, los problemas que en él se encuentran tienen que ver con el

procesamiento y almacenamiento de datos. Este tema comenzará tratando de entender lo que es un problema y cómo se presenta en la vida cotidiana.

1.2.1 El problema de los regalos de Arcadio

Arcadio no tenía aún una idea clara acerca de qué regalo sorpresa debía comprarle a Úrsula, pero pensó que, una vez en la tienda, algo se le ocurriría. Desde que entró, en efecto, la isla de joyería de fantasía le dio varias ideas. Podía comprarle una pulsera de cristales, un dije de corazón o un par de aretes. De hecho, pensó que no necesariamente tenía que ser un solo regalo; podrían ser varios, siempre y cuando no excedieran los mil pesos que le prestó su madre.

Ahora el problema era decidir cuál o cuáles eran los regalos adecuados. Arcadio no tardó en percatarse de la primera dificultad. Había muchas posibilidades: elegir tres y comprar uno (tres posibilidades), comprar dos de ellos (tres posibilidades, dependiendo de cuál elija no comprar), comprar los tres (una posibilidad), e inclusive la salida fácil de no comprar ninguno. Para cada una de estas ocho posibilidades, tendría que calcular si le alcanzaban sus mil pesos.

Mientras reflexionaba cuánto tiempo le tomaría evaluar tantas opciones, le surgió una duda: ¿no le había regalado su ex novio a Úrsula un par de aretes en su cumpleaños pasado? Fue entonces cuando se dio cuenta de que no tenía mucho caso evaluar cada una de las posibilidades antes de conocer los gustos de Úrsula y estar seguro de qué regalo apreciaría más.

Después de 20 minutos de no saber qué hacer, decidió hablarle a Diana, la mejor amiga de Úrsula, para preguntarle su opinión. En vez de ayudar a Arcadio a decidirse por alguna de las opciones que había considerado, Diana le dio nuevas alternativas: "¿Qué tal un CD de Robbie Williams o el DVD de la última película de Harry Potter?", sugirió. Bueno, ahora la cosa estaba peor. En efecto, la idea de Diana era buena y Arcadio encontró fácilmente el DVD de la última película de la saga de Harry Potter y los últimos tres discos de Robbie Williams. Ahora el número de opciones era mayor. En medio del vértigo que comenzó a sentir, se le ocurrió que quizá sería mejor comprarle cosas pequeñas, pero muchas. Y ya totalmente confundido, pensó que, dado que no tenía idea de qué hacer, mejor aprovecharía para comprarse unos audífonos que le hacían falta y, con lo que sobrara, le compraría a Úrsula el regalo más costoso.

La visión del computólogo

Arcadio se dio cuenta de que lo primero que tenía que hacer era discernir cuál era exactamente el **problema** que debía resolver. Antes que nada, cuando se presenta un problema, éste viene acompañado de un conjunto de datos que lo determinan: un número, un mapa, algo sobre lo que se trabajará para encontrar la solución. Lo anterior se denomina la *entrada* al problema. En segundo lugar, resolver un problema significa producir una *salida* que indica la solución; por ejemplo: tomar una decisión, encontrar un camino en un mapa o calcular un número. Finalmente, no se trata de calcular cualquier número o de decidir cualquier cosa; se tiene una *relación* entrada/salida acerca del objetivo que se quiere alcanzar. Esta relación explica cuál es el vínculo entre la salida que se debe producir y la entrada que se recibe. Como ejemplos de problemas se pueden mencionar los siguientes:

Concepto
Un problema consiste en la especificación de un conjunto de posibles entradas, y una relación de salidas para cada entrada.

1] *Elevar un número al cuadrado.* La entrada es un número x, la salida un número y. La relación entre el número de entrada x y el de salida y es que y debe ser igual a x^2.
2] *Encontrar la salida de un laberinto.* La entrada es un laberinto que tiene marcado un lugar por donde entrar y uno por donde salir. La salida es un camino. La relación específica: que el camino producido debe comenzar en la entrada del laberinto, terminar en la salida y ser un camino válido del laberinto (no brincar las paredes).
3] *Colorear mapas.* La entrada es un mapa de países. La salida es un color para cada país. La relación específica: que cada país debe ser coloreado con un solo color, de manera que si dos países tienen una frontera común sus colores sean diferentes. Además, debe usarse el menor número de colores posible.

Curiosidades
Probablemente el primer laberinto fue una prisión en la isla de Creta. Según la mitología griega, ahí vivía el Minotauro. El algoritmo que utilizó Teseo para recorrer el laberinto consistió en el uso de una cuerda que desenrollaba conforme iba avanzando por el laberinto y que luego enrollaba para salir.

Para exponer puntualmente el problema de Arcadio, se comienza por identificar los datos con los que se trabajará.

Entrada:
1] Una cantidad de dinero inicial de mil pesos.
2] Un conjunto de regalos posibles, S = {pulsera, dije, aretes}. Debido a la recomendación de Diana, se define el conjunto S1 = {pulsera, dije, aretes, CD, DVD}.
3] Cada regalo posible de S1 tiene un precio en pesos y un valor emocional distinto para Úrsula. La cuestión entonces consiste en procurar una adquisición óptima; elegir objetos cuyo costo sea accesible y que el aprecio de Úrsula sea considerable. Así, los regalos poseen un valor emocional para Úrsula que se puede expresar como un valor numérico entre 0 (no le gusta) y 10 (le encanta).

Supóngase que los valores están dados como en la siguiente tabla:

Pulsera	Dije	Aretes	CD	DVD
$ 800	$ 300	$ 400	$ 500	$ 100
♥ 8⁻	♥ 10⁻	♥ 5⁻	♥ 4⁻	♥ 1⁻

Tabla 1.

Una vez decidida la entrada al problema, puede pensarse en lo que se desea obtener como salida.

Salida:
1] Uno o más regalos para Úrsula. Es decir, un subconjunto T de S1 (o de S).
2] Quizá con lo que sobre de los mil pesos Arcadio pueda comprarse algo para él.

Arcadio debe aclarar qué opción es la más conveniente para el regalo de Úrsula: comprarle la mayor cantidad de regalos posible; evitar a toda costa comprarle algo que un ex novio le obsequió en el pasado; comprarle un solo regalo que le guste mucho; mostrarle, a través de su elección, algo acerca de su personalidad y sensibilidad; o asegurarse de que le sobre dinero para comprarse algo él mismo. Cada una de estas posibilidades implica un problema distinto. El problema de Arcadio está dado por lo que se plantea a continuación.

Relación entrada/salida:

Los regalos elegidos, es decir, los del conjunto T, no deben costar más de mil pesos en total. Además, no debe existir otra elección de regalos que coincida con esta característica y que tenga un valor emocional total mayor. Es decir, los regalos elegidos deben tener el mayor valor emocional posible y costar máximo mil pesos.

Una salida posible al problema sería {aretes, CD}, ya que el costo de estos regalos es de $400 + $500 = $900 con un valor emocional de ♥5 + ♥4 = ♥9. Pero esta salida no resuelve el problema; hay mejores salidas, como {aretes, CD, DVD}, que tampoco rebasa los mil pesos, pero que tiene un valor emocional mayor para Úrsula, de ♥10. El problema se habrá resuelto una vez que Arcadio elija los regalos con mayor valor emocional, sin que rebasen los mil pesos. ¿La solución es {aretes, CD, DVD}? No, porque si se cambian los aretes por el dije, el costo total baja a $900 obteniéndose una combinación de mayor valor emocional, es decir, ♥15. Aparentemente, la opción {dije, CD, DVD} es una buena solución al problema. Pero, ¿cómo se puede estar seguro de ello?

1.3 PROBLEMAS DE LA VIDA COTIDIANA

En la vida cotidiana, definir de manera precisa un problema es en sí un problema. Incluso, definir la entrada al problema requiere de toda la atención. Se podría decir que la entrada al problema es simplemente el deseo de Arcadio de comprarle un regalo a Úrsula, pero esto es demasiado vago. Para obtener una buena solución a un problema, es necesario pensar cuidadosamente en el objetivo, e ir precisando poco a poco lo que se desea lograr. Arcadio comenzó por acotar el universo de regalos posibles, primero a tres y luego a cinco regalos específicos. Sin embargo, otras opciones pudieron ser visitar varias tiendas o buscar regalos por internet.

Esto pudo haber implicado otro problema, o en otras palabras un "preproblema", consistente en definir un conjunto de opciones para buscar los regalos, que funcionara como entrada al problema original. Otra dificultad que tuvo Arcadio fue definir qué perseguía con la compra: tratar de satisfacer las preferencias de Úrsula o decidir qué regalos le gustaban más a él. Y finalmente estaba la dificultad de asignar valores emocionales o preferencias a los regalos.

En resumen, Arcadio tenía que definir tanto la entrada al problema (qué regalos, qué valores emocionales) como la relación entrada/salida.

1.3.1 El significado de resolver un problema

Es importante notar que para resolver el problema es necesario concentrarse en su *especificación* —es decir, en términos de entrada, salida y relación entrada/salida—, y descartar cualquier cosa que aparte la atención del objetivo. Se trata de encontrar una salida que satisfaga los requerimientos de la relación entrada/salida. Quizá si Arcadio le hubiera preguntado a Úrsula qué regalo le gustaría, ella le hubiera explicado que le acababan de regalar un dije y que ya no quería otro. Si éste hubiera sido el caso, los valores emocionales de entrada al problema serían incorrectos. No es que la solución fuera incorrecta, lo que no era correcto es la especificación del problema. Esto ilustra lo que sucede con frecuencia en la vida real, tanto a nivel personal como profesional. Antes de comenzar a trabajar en la resolución de un problema, es importante identificar el problema que interesa resolver.

1.4 ALGORITMOS: RESOLVIENDO EL PROBLEMA

Una vez definido el problema, hace falta describir un último elemento antes de pensar en resolverlo: las reglas del juego, las herramientas disponibles y el uso que se puede hacer de éstas. A esto se le denomina ***modelo de cómputo***.

Concepto
El modelo de cómputo especifica las operaciones que la máquina puede ejecutar para resolver un problema, así como el costo de ejecutar cada una de las operaciones.

Considérese el problema de cambiar una llanta. Hay un abanico de opciones para lograrlo: llamar por teléfono a un servicio de grúas; utilizar las herramientas que vienen en la cajuela; comprar otras herramientas; pedir ayuda, entre otras. Sin embargo, se debe tener en mente que la solución dependerá de las herramientas disponibles.

En cuanto a la solución del problema de Arcadio, una opción podría ser hablarle a su mamá y pedirle que le pregunte a Úrsula qué regalos prefiere. Pero esto no resuelve el problema planteado, ya que involucra valores emocionales distintos (específicamente los de Úrsula) y no los estimados por Arcadio. Tampoco sirven las soluciones basadas en la telepatía o en contratar un programador que resuelva el problema. Es necesario ponerse en el lugar de Arcadio, que está en la tienda y tiene que analizar las posibles soluciones, y sólo tiene acceso a lápiz y papel.

1.4.1 Una solución: búsqueda exhaustiva

Lo primero que Arcadio debe hacer es analizar los posibles regalos. Tiene que decidir, uno por uno, si comprarlos o no, y una vez que ha elegido un regalo en particular, pensar en las mismas opciones para cualquier otro.

Antes de las sugerencias de Diana, Arcadio tenía las opciones que se enlistan en la tabla 2. Aunque algunas se pueden descartar rápidamente, por ejemplo la primera, que implica no comprar el regalo, o la última, dado que no le alcanza para comprar las tres cosas.

Más en detalle, el método de solución o *algoritmo* que se le ocurre a Arcadio y que se seguirá paso a paso, es el siguiente:

1] Considerar todas las opciones, renglón por renglón, en la tabla 2 y, para cada una, calcular el costo total de los objetos y su valor emocional.
2] Tomar lápiz y papel, y reproducir la tabla 3.
3] Anotar a un lado de la última columna, fila por fila, el valor más grande, siempre y cuando el costo correspondiente dado por la penúltima columna no sea mayor a mil pesos.

Pulsera	Dije	Aretes
No	No	No
No	No	Sí
No	Sí	No
No	Sí	Sí
Sí	No	No
Sí	No	Sí
Sí	Sí	No
Sí	Sí	Sí

Tabla 2. Posibles combinaciones en la compra de tres regalos.

Es decir, primero anota ♥0, después ♥5, ♥10, ♥15, sucesivamente.

Después se descartan los que sobrepasan el costo de 1 000, y entre los que sobran se identifica al que tiene mayor valor, y se descartan los demás, descartando los valores ♥8 (puesto que es menor al ♥15), ♥13 (también menor al ♥15 y con costo de $1 200), ♥18 (que, aunque es mayor que el ♥15, implica un costo de $1 100), y por último el ♥23.

Pulsera	Dije	Aretes	Costo	Valor
No	No	No	$ 0	♥ 0
No	No	Sí	$ 400	♥ 5
No	Sí	No	$ 300	♥ 10
No	Sí	Sí	$ 700	♥ 15
Sí	No	No	$ 800	♥ 8
Sí	No	Sí	$ 1 200	♥ 13
Sí	Sí	No	$ 1 100	♥ 18
Sí	Sí	Sí	$ 1 500	♥ 23

Tabla 3. Posibles combinaciones de compra y sus valores correspondientes.

A partir de lo anterior, se deduce que la mejor opción es, en efecto, comprar el dije y los aretes.

1.4.2 Análisis de la solución de una búsqueda exhaustiva

Una vez propuesto un algoritmo, surgen dos preguntas: ¿es correcto el algoritmo? —es decir, si en verdad encuentra una solución al problema— y ¿es eficiente? —o sea, en el supuesto de que el algoritmo realmente sea correcto, se requiere saber si facilitará el trabajo en cuanto a tiempo, dinero o algún otro aspecto.

Corrección. El algoritmo de Arcadio es correcto, pues para obtenerlo consideró una a una todas las posibilidades existentes. Después de completar la tabla 2, revisó una por una las ocho posibilidades para emplear la de mayor valor que, además, no sobrepasó la cantidad de dinero con la que contaba.

Complejidad. De esta forma se denomina el costo que genera el algoritmo, ya sea en tiempo, dinero o alguna otra medida. En el caso de Arcadio, implicó el tiempo empleado para calcular la tabla 2 y recorrer la última columna para encontrar la combinación de regalos óptima en la tabla 3. Supóngase que Arcadio demoró 10 segundos en llenar cada renglón de la tabla. La tabla 1 indica que son ocho renglones, lo que da como resultado un tiempo de 8×10 segundos = 80 segundos = 1 minuto, 20 segundos. En el supuesto de que la revisión de los últimos dos elementos de cada renglón en la tabla 3 le haya tomado cinco segundos, el total sería $8 \times 5 = 40$ segundos. En resumen, Arcadio resolvió el problema en dos minutos.

Concepto
Un algoritmo es correcto si para cualquier entrada válida al problema, produce una salida de acuerdo con la relación entrada/salida.

1.4.3 Análisis del caso general de la búsqueda exhaustiva

Arcadio resolvió el problema en sólo dos minutos. Sin embargo, ¿qué tan bueno es su algoritmo? Si el algoritmo va a ser usado una sola vez, es difícil predecir qué tan bueno es su método de solución. ¿Servirá para resolver situaciones similares? ¿Arcadio podría enseñar el método a otras personas? Se puede pensar que el algoritmo va a ser empleado muchas veces. Obviamente, si Arcadio volviera a usar su algoritmo, no sería con la misma entrada que se dio en la tabla 3, puesto que ya conoce su solución. Entonces, considérese el **conjunto de entradas** posibles al problema, dado por:

Concepto
Un algoritmo tiene un conjunto de entradas válidas definido, de distintos tamaños. Cada vez que se ejecuta el algoritmo, es con un elemento de este conjunto.

Concepto
Cada elemento del conjunto de entradas tiene un tamaño. Mientras más grande sea la entrada, más recursos requerirá el algoritmo para producir una salida.

Curiosidades
Un camino que se bifurca una y otra vez indefinidamente, genera un número de posibilidades que crece exponencialmente. Éste es un tema recurrente en los libros de Jorge Luis Borges, al igual que los laberintos y los espacios de búsqueda inmensos.

Conjunto de entradas:

1] Una cantidad de dinero inicial: $ *X*.
2] Un conjunto de *n* regalos posibles: $R = \{R1, R2, \ldots, Rn\}$.
3] Para cada regalo posible *Ri* de *R*, un precio en pesos, $c(Ri)$, y un valor emocional, $v(Ri)$.

Cada vez que Arcadio resuelve el problema, ejecuta su algoritmo con una entrada particular, con valores específicos de *X*, de *R*, de precios y valores. Por ejemplo, cuando Diana le sugirió otros dos regalos tuvo que ejecutar otra vez su algoritmo, ahora para $n = 5$, $R = \{$pulsera, dije, aretes, CD, DVD$\}$, con sus respectivos precios y valores (la *X* no cambia, sigue siendo igual a mil pesos).

Para evaluar la complejidad del algoritmo, en cuanto al tiempo que toma ejecutarlo, se debe considerar el *tamaño de la entrada,* pues no es lo mismo resolver el problema para tres regalos que para cinco o ¡para 50! Si el algoritmo de Arcadio funciona, deberá servir para resolver casos más complejos. Nótese que inclusive en los casos de tres regalos, el algoritmo debe funcionar para cualquier entrada, es decir, para cualquier conjunto de tres regalos, especificado cada uno por su precio y su valor emocional. Pero ¿qué sucede cuando se incrementa el número de regalos, o sea, el tamaño de la entrada *n*, a partir de $n = 1$?

Si sólo se considera un regalo ($n = 1$), se tendrían dos opciones: comprarlo o no. Si fueran dos objetos se obtendrían cuatro opciones: comprarlos ambos, no comprar ninguno y comprar alternadamente alguno de ellos y el otro no; el número de opciones resultante es el doble de las que había para un solo objeto. Cuando entran en escena los tres objetos tenemos, como se aprecia en la tabla 2, ocho opciones en total (el doble de las que teníamos para el número de objetos anterior). Mientras el número de objetos crezca de uno en uno, el número de opciones siempre se duplicará. Por ejemplo, si con cierto número de objetos se tienen *x* opciones y aparece un objeto más como un CD, el número de opciones será ahora $2x$, ya que si originalmente había una opción: {pulsera no, dije sí, aretes no}, con el nuevo objeto se obtienen dos {pulsera no, dije sí, aretes no, CD no} y {pulsera no, dije sí, aretes no, CD sí}.

Inicialmente Arcadio tenía ocho posibilidades descritas en la tabla 2. Si añadiera el CD, el número de opciones sería $2 \times 8 = 16$; si incluyera el DVD, el número de opciones aumentaría a $2 \times 16 = 32$. Si añadiera otro regalo, $2 \times 32 = 64$, y uno más todavía, $2 \times 64 = 128$. En resumen, al ejecutar su algoritmo, Arcadio tiene que trazar una tabla cuyo número de filas crece conforme el número de regalos a considerar aumenta. A tres regalos corresponden ocho filas, y el tiempo para escribir en la tabla es de 8×10 segundos; con el CD son 16 filas y el tiempo es de 16×10 segundos, y así sucesivamente.

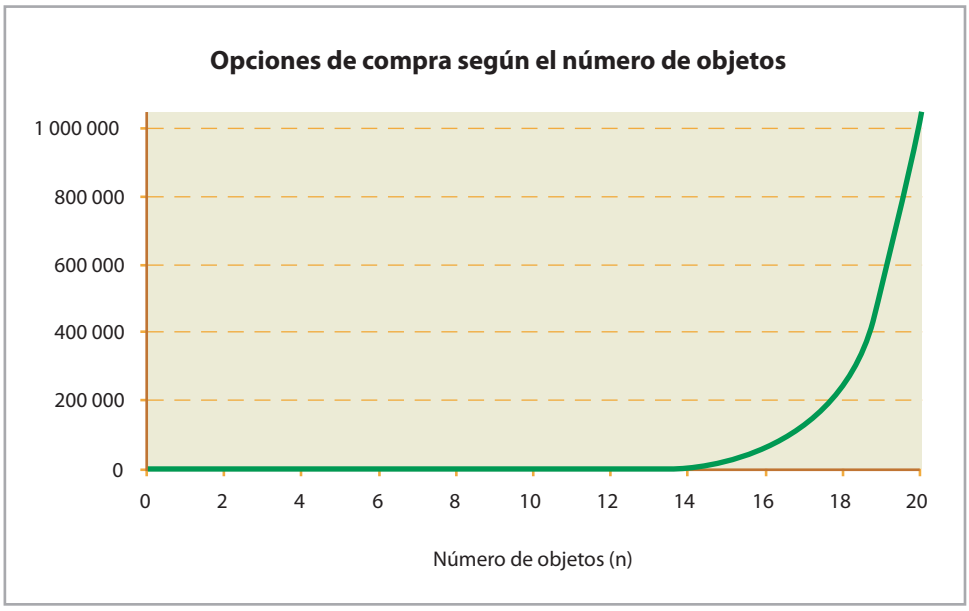
Tabla 4. Crecimiento del tiempo cuando las posibilidades aumentan.

N	Posibilidades	Tiempo
1	2	20 s
2	4	40 s
3	8	1 min, 20 s
4	16	2 min, 40 s
5	32	5 min, 20 s
6	64	10 min, 40 s
7	128	21 min, 20 s
8	256	42 min, 40 s
9	512	1 h, 25 min, 20 s
10	1 024	2 h, 50 min, 40 s
20	1 048 576	+ de 121 días
30	1 073 741 824	+ de 340 años

Si *n* representa el número de objetos, dependiendo del valor que adquiere, el número de posibilidades se comporta como se muestra en la tabla 4 y en la gráfica 1. Se observa que el número de posibilidades siempre es 2 elevado al número de objetos, es decir 2^n, lo que proviene de la multiplicación de $2 \times 2 \times 2 \ldots \times 2$, *n* veces, ya que cada vez que se añade un objeto se duplica el número de posibilidades. En términos de tiempo, Arcadio emplea 10 segundos para escribir en cada fila de su tabla, por consiguiente le tomará 10×2^n segundos llenar la tabla.

Basada en la fórmula anterior, la tabla 4 muestra el tiempo que emplearía Arcadio en escribir toda la tabla.

La n es la manera que se tiene para decir qué tan grande es la entrada al algoritmo. Al principio, Arcadio sólo consideró tres objetos de joyería; en ese momento n valía 3. Cuando entraron en escena los objetos restantes, producto de las sugerencias de Diana, Arcadio tenía $n = 7$ objetos.



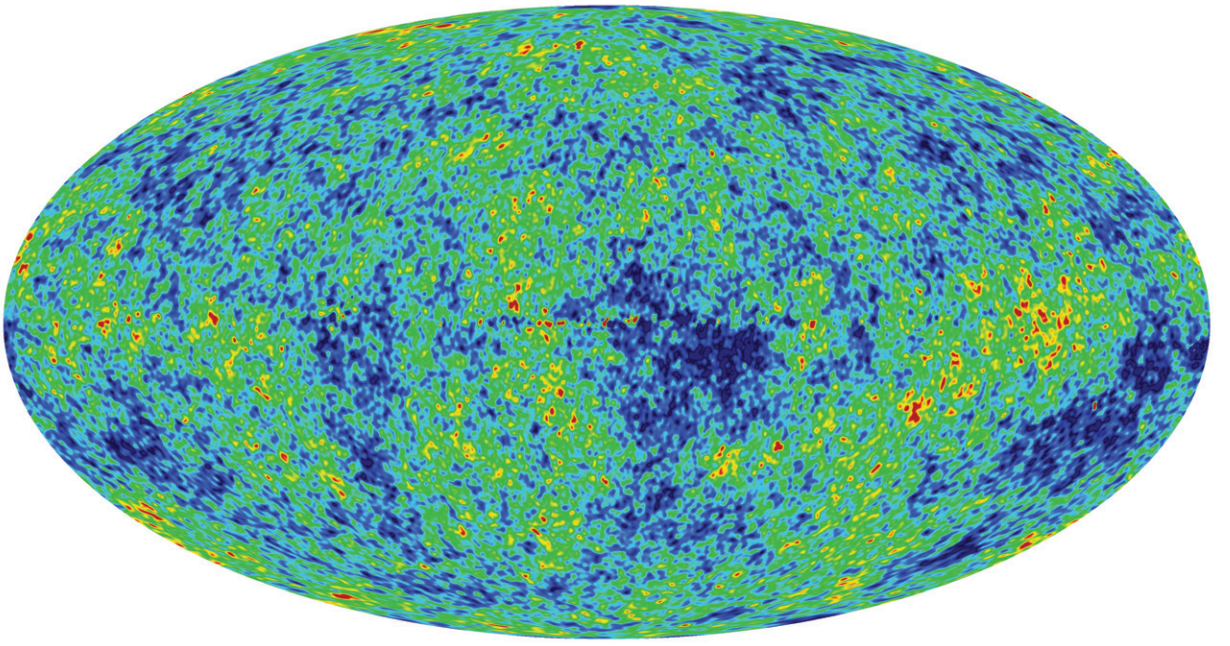
Gráfica 1. Crecimiento de las opciones según el número de objetos.

1.5 CRECIMIENTO EXPONENCIAL

Ya se ha visto cómo el tiempo de ejecución del algoritmo de Arcadio crece aproximadamente 2^n con el tamaño de la entrada n. A una relación como ésta, en general se le denomina *función exponencial* $f(n)$: la variable independiente, la n, forma parte del exponente al que se eleva algún número, en este caso el 2. Por ejemplo, al multiplicar el 3 por sí mismo n veces, se obtiene otra función exponencial, $f(n) = 3^n$.

1.5.1 Crecimiento exponencial en computación

El tiempo para resolver un problema crece de manera exponencial cuando, como en el caso de Arcadio, hay que buscar una solución en un conjunto muy grande, cuyo tamaño es exponencial. Si la entrada es de tamaño n, el conjunto sería de tamaño $f(n)$. Y más aún, no existen pistas que guíen la búsqueda, que eviten revisar, una por una, todas las opciones que surgen para determinar cuál es la óptima. Se enfrenta un problema complejo en el sentido computacional del término, cuya solución requiere de mucho, muchísimo tiempo, aun cuando el tamaño de la entrada no sea muy grande. Para entender esto, la información de la tabla 4 puede ser de gran utilidad, ya que representa el tiempo de ejecución del algoritmo de Arcadio. Supóngase que emplea 10 segundos en escribir cada fila. En la tabla se observa que considerar 20 regalos implica demasiado tiempo, puesto que se necesitarían más de 121 días; no se diga considerar 30 regalos, ¡llevaría más de 340 años completar la tabla!



Detalle de una imagen compuesta por datos acumulados durante 7 años por WMAP. La imagen revela 13 700 millones de años de fluctuaciones de temperaturas del universo temprano | © NASA/WMAP Science Team.

Ante el crecimiento exponencial, de poco sirve la tecnología. Puesto que el amor de Arcadio es enorme y no está dispuesto a tomar una decisión sin considerar al menos 20 posibles regalos, pero esperar 121 días está fuera de cuestión (pues para entonces Úrsula ya habría regresado con su ex novio), decide contratar a un programador que ejecute su algoritmo en una veloz computadora e imagina que será una herramienta útil para calcular cada fila de la tabla miles de veces más rápido que él; tal vez, un millón de veces más rápido que él. Es decir, en lugar de 10 segundos, 10/1 000 000 segundos = 10 µs. Dado que $2^{20} = 1\,048\,576$, para calcular el tiempo, habría que multiplicar este número por 10 microsegundos; de este modo todo se reduce a únicamente 10.48576 segundos. "Perfecto", piensa Arcadio.

Sólo que ahora decide que la belleza de Úrsula es aún mayor y que por lo menos debe considerar 30 regalos. Con esto el tiempo empleado se eleva a 10 073 741 824 segundos, o sea, 2.77 horas de ejecución en una veloz computadora. "Mmm…", reflexiona Arcadio, "¿será muy caro rentar la computadora por un poco más de tiempo y considerar por lo menos 35 regalos?" Arcadio realmente ama a Úrsula. El tiempo se dispara a más de 164 horas, es decir, 6.8 días; y para el cálculo de 40 regalos se emplearían años de ejecución en una veloz computadora.

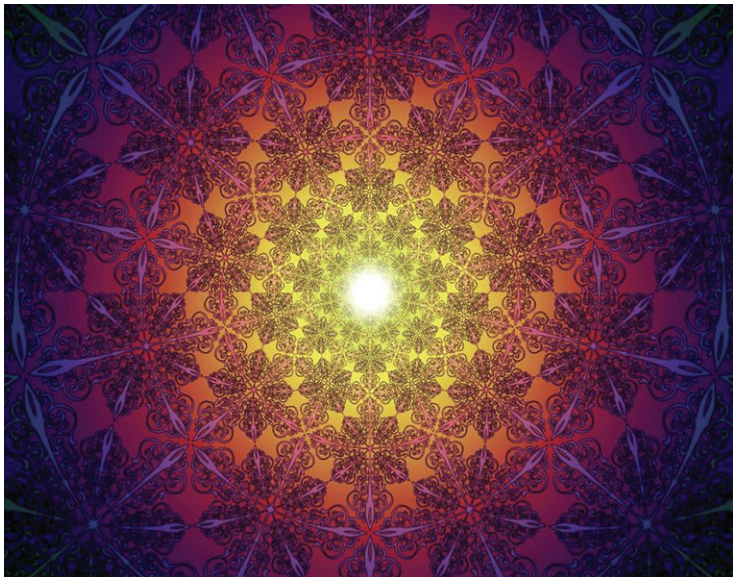
Curiosidades
Si a alguien le ofrecen un átomo de oro por cada segundo transcurrido desde el Big Bang, ¿se haría rico? Resulta que los átomos son tan increíblemente pequeños, que esa cantidad de átomos equivale solamente al pedacito de una moneda de oro que pesa 0.14 miligramos y no vale más que unos cuantos pesos.

1.5.2 Crecimiento exponencial en la sociedad y en la naturaleza

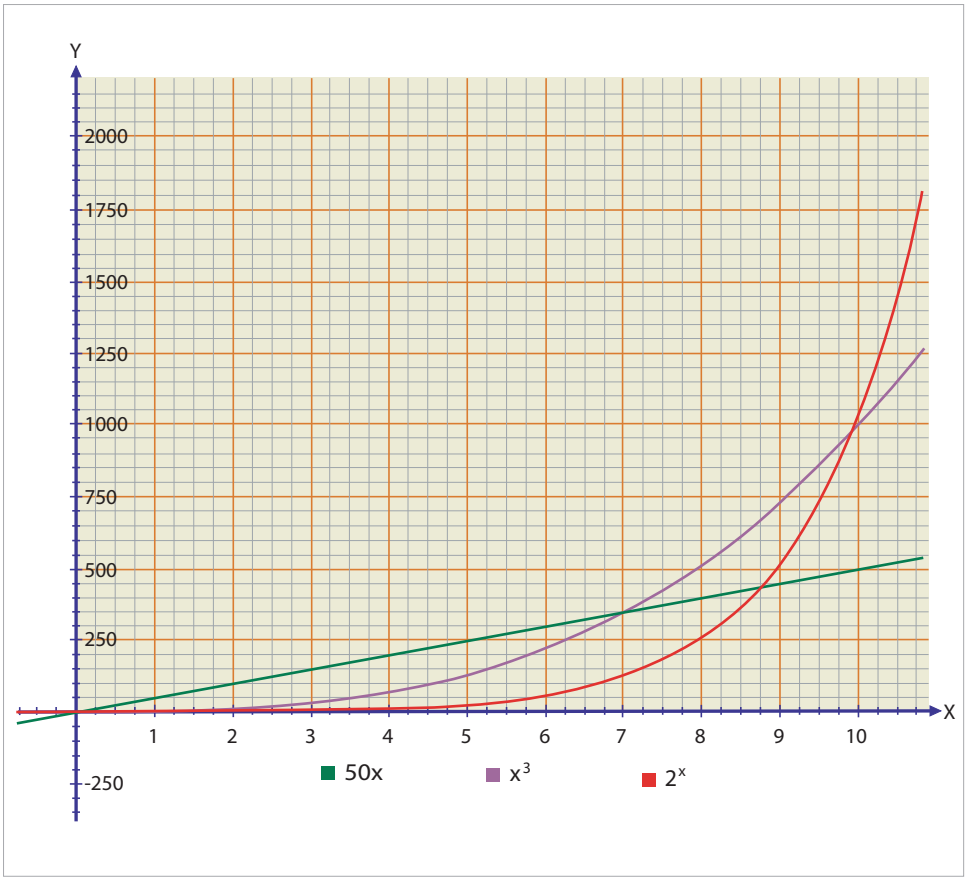
Con frecuencia se escucha hablar en los noticiarios de crecimiento exponencial para enfatizar que algo crece rápidamente. Hasta ahora se ha visto que el crecimiento exponencial tiene un significado preciso. Ocurre cuando se tiene una función de la forma $f(n) = c^n$ para algún número c. Nótese que, al principio, el factor tiempo no crecía atropelladamente; por ejemplo, en la tabla 4, con los pocos regalos que había escogido Arcadio, el tiempo aún no se disparaba. Hasta los ocho regalos, Arcadio todavía podía lidiar con el problema en menos de una hora. Incluso 10 regalos implicaron menos de tres horas. Sin embargo, como indica la gráfica 1, alrededor de los 18 regalos la función se dispara.

Velocidad de crecimiento

El crecimiento exponencial se puede caracterizar por la velocidad a la que crece el valor de una función. En la gráfica 2 se muestra una **función lineal**, cuya velocidad de crecimiento es constante, es decir, es la misma en cualquier momento. Se trata de una línea recta y su inclinación o pendiente siempre es la misma. En cambio, en una función exponencial, la velocidad de crecimiento es en sí exponencial, es decir, su valor crece a una velocidad proporcional al valor de la función. En el caso de 2^n la tasa de crecimiento es 2, se duplica su valor con el tiempo. O sea que mientras más grande es la cantidad, más rápido crece. Y tarde o temprano el valor de la función se dispara. En la gráfica también aparece la función n^3, cuya velocidad de crecimiento ya no se mantiene constante como en el caso de la función lineal. Aunque crece rápidamente, después de cierto valor ($n = 10$) será rebasada por la función exponencial.



Espiral de Fibonacci | *© Edward S. May.*



Gráfica 2. Funciones con diferentes velocidades de crecimiento.

Curiosidades
Cada número de Fibonacci es la suma de los dos anteriores 1, 1, 2, 3, 5, 8, 13, 21… Se les encuentra frecuentemente en la naturaleza, en particular en terrenos como el del crecimiento poblacional, con un crecimiento de 1.618^n aproximadamente. Imágenes como ésta se generan a partir de ellos.

Concepto
Una función $f(n)$ es lineal si es de la forma c^n, para alguna constante c, quizá más otra constante. Por ejemplo, $f(n) = 3n - 5$.

1.5.3 Ejemplos de crecimiento exponencial

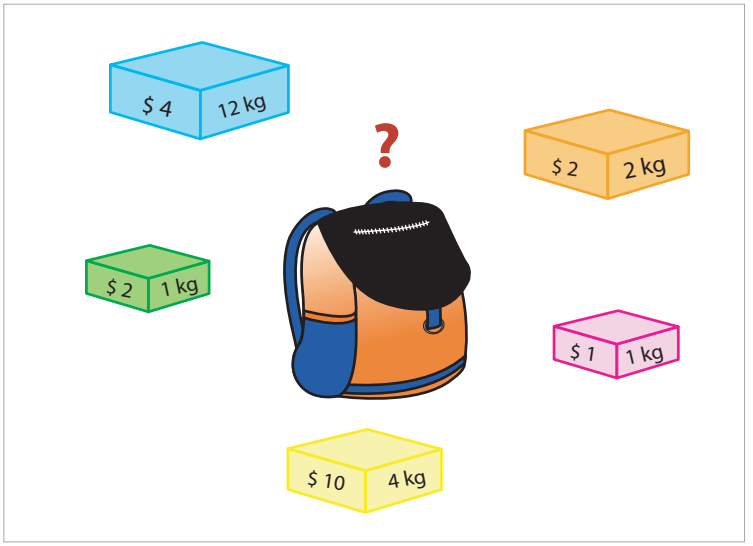
Variantes del problema de Arcadio

Curiosidades
El problema de la mochila aparece en diversas formas en negocios, criptografía, matemáticas aplicadas y muchas otras áreas, donde se resuelve de manera aproximada, ya que es un problema NP-completo.

Arcadio no es el único con este tipo de problemas; es muy frecuente encontrarse en situaciones similares. Dos ejemplos:

a] Una empresa tiene un presupuesto para contratar empleados nuevos y debe decidir a cuántos y cuáles contratar. Cada candidato posee una experiencia y un nivel de preparación distintos, por lo que recibirían un sueldo diferente. Por otro lado, cada uno tendría un valor diferente para la empresa. El objetivo de la empresa es contratar a las personas que ofrezcan el valor máximo total, pero sin rebasar la cantidad de dinero disponible para pagar los sueldos.

b] El problema de la mochila (del inglés *knapsack*). Arcadio se va de campamento y lleva una mochila de cierta capacidad. Su objetivo es empacar latas de comida en la mochila. Cada lata ocupa una cierta cantidad de espacio y tiene un valor nutricional determinado. El objetivo de Arcadio es empacar las latas aprovechando al máximo su valor nutricional total, cuidando que las latas elegidas no rompan su mochila.



Problema de la mochila.

Crecimiento poblacional

Quizá sea éste el problema más serio de las sociedades modernas, debido a que la población crece exponencialmente y, con ésta, el uso de los recursos como el petróleo, los metales, las carreteras, etc.[1] Pero los recursos no se regeneran exponencialmente: "No es una simplificación afirmar que la falta de entendimiento del concepto de crecimiento exponencial por parte de legisladores y planificadores es el mayor problema de todos los estudios del medio ambiente y de administración".[2]

Para saber por qué una población crece exponencialmente, considérese una coneja que se aparea y tiene cuatro crías: dos hembras y dos machos. Éstos a su vez se aparean y las hembras tienen cada una cuatro crías; y así sucesivamente, cada hembra concibe cuatro crías, dos de cada sexo. La tasa de natalidad es de 2, ya que cada pareja tiene cuatro descendientes; es decir, la población se duplica cada generación. En la primera generación, la coneja concibió dos crías. Para la segunda generación ya son cuatro conejos, porque cada hembra tiene dos crías. En la tercera generación son ocho, luego 16. En promedio, el número de conejos en la generación n es $2n$, ya que el número de conejos se duplica.

[1] Para simplificar la discusión no se consideró el efecto de la mortalidad en el crecimiento poblacional.
[2] Véase la página <http://zebu.uoregon.edu/1999/es202/l3.html>.

Si una pareja tiene un promedio de tres crías, la tasa de natalidad sería 1.5. Si se tuviesen, por ejemplo, dos peces con esta tasa de natalidad, después de una generación se tendrían tres, luego cuatro, luego seis, luego 10, y así sucesivamente. Si se les suministrara comida suficiente y la pecera contara con capacidad para 500 peces, para la décimocuarta generación habría que comprar otra pecera; es posible experimentar con distintas tasas de nacimiento en <www.otherwise.com/population/expo nent.html>.

Si alguien posee una pecera con capacidad para 8 000 peces y quiere mantenerlos por 12 generaciones, debe asegurarse de comprar una especie con tasa de natalidad de dos peces, ya que para entonces tendrá 8 192.

Crecimiento "virulento"

Los virus, tanto los informáticos como los que atacan a los organismos, pueden ser muy peligrosos si no se combaten a tiempo, sobre todo si tienen una tasa de natalidad mayor a 1, ya que su crecimiento sería exponencial.

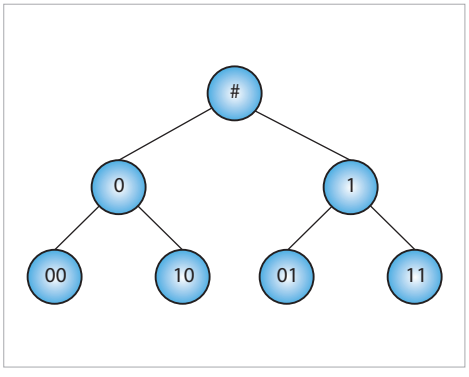
1.5.4 Árboles

Una representación muy útil del crecimiento exponencial es la figura del árbol. Éste consiste en dos tipos de elementos: vértices y aristas. Un vértice es un elemento que representa alguna cosa y que usualmente se dibuja como un punto o un círculo dentro del cual se puede escribir lo que éste representa. Las aristas conectan parejas de vértices, representando alguna relación entre ellos. Un árbol puede tener un vértice del cual descienden todos los demás vértices, llamado raíz. A los vértices que descienden de un vértice se les denomina hijos. Esto porque, en efecto, los árboles son útiles para representar relaciones de parentesco en una familia. A continuación se presentan algunos ejemplos.

1] *El problema de Arcadio.* Consideremos el caso de los tres regalos. En este caso la raíz representa la situación inicial, en la cual Arcadio aún no tomaba una decisión (la marcada con #). De la raíz se desprenden dos hijos (el primer regalo), la opción de comprarlo o la de no comprarlo, denotados como 0 y 1. En el siguiente nivel (correspondiente al segundo regalo), cada uno de los vértices tendrá a su vez dos hijos; y así sucesivamente.

En la figura, el primer nivel tiene un solo vértice, el segundo tiene dos y el tercero cuatro (2×2); en esta secuencia el cuarto nivel tendría ocho (dos hijos por cada uno de los del nivel anterior: 2×4). En general, si el árbol tiene n niveles, el enésimo nivel tendrá 2^{n-1} vértices.

Figura 1. Árbol de las posibles decisiones de compra para dos regalos.



1.5.5 Qué tan rápida es una computadora

Como se ha visto, el crecimiento exponencial del tamaño de un problema ocasiona que los esfuerzos por desarrollar computadoras más veloces sean inútiles para resolverlo. El

espacio de búsqueda de soluciones es enorme, aun para las computadoras más veloces. Para tener una idea más concreta de la situación y ver qué tan rápida es una computadora, antes que nada es importante resaltar que si bien se utiliza como ejemplo una computadora de la actualidad, para cuando este libro llegue al lector, seguramente existirá una más veloz. Sin embargo, las cifras que se mencionan a continuación proporcionan una buena idea de la magnitud de lo que implica la velocidad de una computadora. Bajo esta advertencia, considérese una computadora personal típica, como la que hay en muchos hogares. Probablemente sea una computadora con un procesador Pentium de Intel, corriendo Windows de Microsoft, o quizás una Macintosh. Una computadora como ésta ejecuta aproximadamente 100 millones de instrucciones por segundo, dependiendo del modelo.

Curiosidades
La UNAM adquirió en 2007 una supercomputadora, llamada KanBalam, que puede ejecutar 7.113 billones de operaciones por segundo (o Gflop/s, equivalentes a 7.113 teraflops). Tal velocidad se debe en parte a su gran cantidad de memoria: 3 016 gigabytes de RAM y 160 000 gigabytes en disco, y a que está conformada por 1 368 procesadores que trabajan en paralelo, cada uno como el de una PC moderna (Opteron Dual Core). En el ámbito internacional, KanBalam ocupa el lugar 126 de las 500 supercomputadoras más rápidas del mundo, y el 28 respecto a las supercomputadoras instaladas en universidades.

KanBalam | © Dirección General de Cómputo y de Tecnologías de Información y Comunicación, UNAM.

Una computadora muy personal: el cerebro

Los mejores programas informáticos de la actualidad tienen aún dificultad para entender todo lo que dice una persona cuando habla, incluso si lo hace despacio y sin ruido, mientras que el cerebro humano puede entender a muchas personas que hablan al mismo tiempo, con ruido y hasta con acentos diferentes. Hay cerebros que incluso entienden conversaciones donde varias personas hablan en distintos idiomas. Más aún, la parte del cerebro que procesa el lenguaje constituye sólo una pequeña parte. El cerebro puede procesar simultáneamente imágenes, controlar el cuerpo, recordar e imaginar cosas, todo esto al mismo tiempo que se conduce una bicicleta.

Un cerebro humano está constituido por alrededor de un billón de neuronas con 100 trillones de interconexiones entre ellas. Una estimación muy burda podría ser que procesa 10 billardos de instrucciones por segundo, aunque es muy difícil calcular esta cifra.

Curiosidades

Hay dos escalas de numeración, la larga y la corta. La escala larga está vigente en México, Francia, Alemania, Holanda, Suecia, Finlandia, Noruega, República Checa, Polonia, Rumania e Italia (con ciertos matices).

mil: $10^3 = 1000$
millón: $10^6 = 1\,000\,000$
millardo: $10^9 = 1\,000\,000\,000$
billón: $10^{12} = 1\,000\,000\,000\,000$
billardo: $10^{15} = 1\,000\,000\,000\,000\,000$
trillón: $10^{18} = 1\,000\,000\,000\,000\,000\,000$

La escala corta es la numeración vigente en Estados Unidos y se ha impuesto en todos los países de habla inglesa, así como en Rusia, Grecia y Brasil.

mil: $10^3 = 1\,000$
millón: $10^6 = 1\,000\,000$
billón: $10^9 = 1\,000\,000\,000$
trillón: $10^{12} = 1\,000\,000\,000\,000$
cuatrillón: $10^{15} = 1\,000\,000\,000\,000\,000$
quintillón: $10^{18} = 1\,000\,000\,000\,000\,000\,000$

Curiosidades

En el siglo XVII una minoritaria corriente francesa e italiana de matemáticos adoptó la denominación de billón para mil millones ($10^9 = 1000\,000\,000$), número que el resto llamaba millardo. Este significado es el que tiene el inglés estadunidense, el portugués brasileño y otras lenguas como el griego y el turco. El inglés británico mantuvo la denominación tradicional hasta hace poco. Debido a esto se suelen cometer errores al traducir artículos del inglés a otros idiomas. Incluso los británicos pueden desconocer si en un texto en inglés se habla de "mil millones" o "un millón de millones", ya que ambos usos coexisten.

1.5.6 Ejemplos de crecimiento exponencial benéficos

También hay ejemplos de crecimiento exponencial benéfico:

1] *Intereses bancarios.* Un ejemplo interesante es el de los intereses que paga un banco, pues aunque el dinero crece exponencialmente, la tasa que paga es tan pequeña que el dinero ahorrado crece mucho con sólo mantenerse en el banco muchos años.

2] *Contraseñas.* El uso de contraseñas para proteger información o acceder a una cuenta en un sistema de cómputo es común. Supóngase que se elige una contraseña de longitud n y alguien pretende adivinarla probando una por una todas las posibilidades. Si esta persona sólo toma en cuenta contraseñas que contengan alguna de las 26 letras del alfabeto, tendría que probar 26^n posibles contraseñas. Ahora bien, aunque esa persona consiguiera una computadora 25 veces más rápida para adivinar nuestra contraseña de n letras, si se elige otra de $n + 1$ letras, su computadora sería inútil porque no podría descifrarla, por lo que el crecimiento resulta benéfico. Este fenómeno se encuentra en el centro de la criptografía moderna: un villano debe gastar una cantidad de recursos (dinero, tiempo) que crece exponencialmente, mientras que alguien honesto sólo ocupa una cantidad de recursos polinomiales.

3] *Ley de Moore.* Como se sabe, la tecnología electrónica para la construcción de computadoras avanza vertiginosamente. Cada año se construyen computadoras más rápidas, más pequeñas y más baratas. Esto fue previsto por Gordon Moore, en 1965, cuando formuló lo que se conoce hoy en día como la Ley de Moore. Ésta señala que el número de componentes que la industria podría colocar en un chip

Gordon Moore.

(circuito integrado) de computadora se duplicaría cada año. En 1975 corrigió su predicción al indicar que lo anterior ocurriría cada dos años. En efecto, el número de transistores por pulgada cuadrada se ha duplicado cada año desde la invención del circuito integrado. Actualmente, la predicción indica la duplicación cada 18 meses. La mayoría de los expertos esperan que esta predicción se cumpla por lo menos 20 años más.

Año	Nombre	Marca	Procesador	Núm. procs.	Rendimiento numérico (Gflops)	Memoria (gigabytes)	Almacenamiento (gigabytes)
Nov. 1991	Sirio	CRAY	Vectorial	4	1	0.512	19
Abril 1997	Berenice	SGI	R10000	40	15.6	10	170
Marzo 2003	Bakliz	HP	Alpha EV67	32	80	32	1 000
Enero 2007	KanBalam	HP	Opteron Dual Core	1 368	7 113	3 016	160 000

Tabla 5. La Ley de Moore en la historia del supercómputo en la UNAM.

1.6 PROBLEMAS PROBABLEMENTE DIFÍCILES, SEGURAMENTE DIFÍCILES Y AUN PEORES

Un algoritmo que resuelve un problema en tiempo exponencial es de utilidad muy limitada. El tiempo de ejecución de este tipo de algoritmos crece exponencialmente con el tamaño de la entrada, de ahí que ocupe demasiado tiempo en resolver inclusive instancias pequeñas. Entonces, ¿por qué no buscar un algoritmo cuyo tiempo de ejecución no crezca tan rápido y que sea eficiente? Más adelante se verá que estos algoritmos eficientes son los llamados de tiempo polinomial, y se utilizan todo el tiempo. La clase de problemas que resuelven los algoritmos polinomiales se denomina tipo P.

Pero, ¿será posible que no exista un algoritmo eficiente para resolver cierto tipo de problemas? En efecto, los computólogos han logrado demostrar que existen problemas tan difíciles que no existe manera de resolverlos en tiempo polinomial mediante una computadora. Por cierto, también existe una clase muy importante de problemas, los llamados NP, para los cuales no se sabe si existen algoritmos polinomiales. Además, en el otro extremo, también existen problemas para los que no hay ningún algoritmo.

Estos últimos se tocarán posteriormente; en esta sección sólo se mencionarán problemas difíciles que sí se pueden resolver y para los cuales el tiempo requerido para resolverlos es exponencial.

Curiosidades
Flops es un acrónimo en inglés que significa operaciones de punto flotante por segundo (*floating point operations per second*). Es utilizado para medir el desempeño de las computadoras, especialmente en áreas científicas en las que se realiza una gran cantidad de cálculos de punto flotante. El punto flotante es un sistema para representar números reales en la computadora.

Curiosidades
Un niño nace con aproximadamente las mismas neuronas que tiene cuando es adulto. Lo único que cambia es el número de interconexiones, ya que el niño nace con pocas.

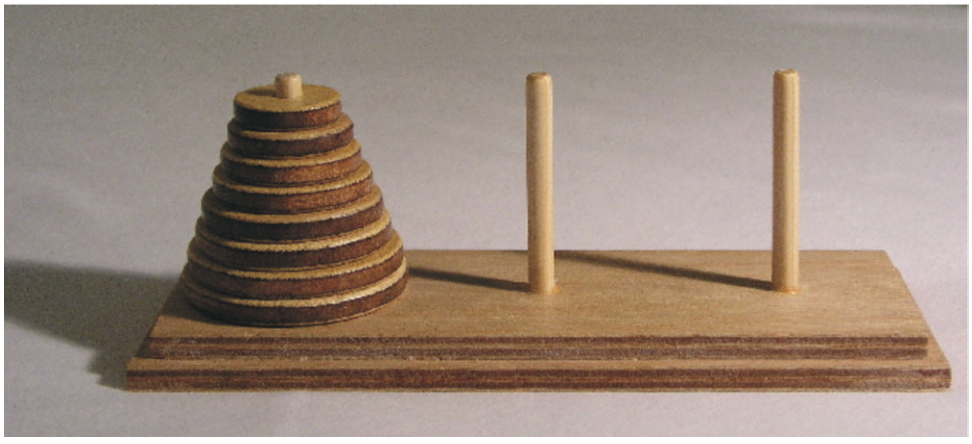
1.6.1 Problemas exponenciales

Los problemas que se pueden resolver mediante algoritmos cuyo tiempo de crecimiento es una función exponencial, se dice que están en la clase EXP. De hecho, se sabe que existen problemas en esta clase, es decir, que se pueden resolver en tiempo exponencial, y que

son imposibles de resolver en tiempo polinomial. El problema de las torres de Hanoi y muchos otros problemas que ocurren en la práctica son de este tipo.

1.6.2 Problemas NP-completos

Existe otra clase importante de problemas para los cuales no se conocen soluciones eficientes, pero no se ha logrado probar que no la tengan, y tienen la siguiente peculiaridad: los mejores algoritmos que se conocen corren en tiempo exponencial, pero se pueden verificar eficientemente. Estos problemas se denominan NP-completos.



Juego de las torres de Hanoi | © Ævar Arnfjörð Bjarmason.

Esto es, si alguien presenta una supuesta solución, se puede diseñar un algoritmo de tiempo polinomial que verifica si en efecto se trata de una solución al problema. El problema de Arcadio es justo de este tipo, si se toma en cuenta una pequeña variante.

Esta variante pide encontrar los regalos que tienen un valor emocional de al menos X unidades. Si Arcadio afirma que ya encontró una solución, por ejemplo, comprar el dije y los aretes por al menos ♥15, simplemente se suman los valores del dije y de los aretes, se verifica que no exceda el dinero con el que cuenta, se suman sus valores emocionales y se comprueba que sea de al menos ♥15.

Una situación fascinante en computación son los cientos de problemas que las personas han encontrado en la práctica y tienen que resolver, aunque sean NP-completos. Para todos esos problemas no se conocen soluciones eficientes y no se sabe si existan. Ésta es la llamada pregunta P = NP.

1.6.3 Problemas peores que los exponenciales

¿Qué puede ser peor que un problema exponencial? ¡Un doble exponencial! Es decir, tomar la función 2^n y en lugar de la n poner 2^n para obtener 2^{2^n}, o sea que, en lugar de multiplicar el número 2^n veces por sí mismo, hay que multiplicarlo 2^n veces por sí mismo. Es difícil darse una idea de lo rápido que crece esta función. Para $n = 5$ el valor de la función rebasa los 4 000 millones, mientras que en $n = 7$ el valor de la función rebasa el número de microsegundos que han pasado desde el Big Bang.

Curiosidades

El matemático francés Édouard Lucas publicó el acertijo de las torres de Hanoi en 1883 bajo el seudónimo de M. Claus, en el cuarto volumen de *Récréations mathématiques*. Quizás se inspiró en un problema similar que apareció en la edición de *De Subtilitate* del matemático italiano Girolamo Cardano. En 1884, otro matemático francés, De Parville, asoció al juego de las torres de Hanoi la siguiente leyenda: "En el gran templo de Benarés, debajo del domo que indica el centro del mundo, se encuentra un plato de metal que tiene fijas tres torres con diamantes. Al crear el mundo, Dios colocó sesenta y cuatro discos de oro en la torre de Brahma. Los discos son de diferente tamaño e inicialmente fueron colocados en orden decreciente de diámetros. Día y noche los monjes tienen que trasladar los discos desde la primera torre a una de las otras dos. La única operación permitida es mover un disco de una torre a otra cualquiera, pero con la condición de que no se puede situar encima de un disco otro de diámetro mayor. La leyenda dice que cuando los monjes terminen su tarea, las torres, el templo y los brahamaneses se convertirán en polvo y, después de un trueno, el mundo desaparecerá."

Curiosidades
La pregunta P = NP es la más importante en computación, y en general una de las más importantes en matemáticas.
En efecto, el Instituto Clay de Matemáticas la ha catalogado entre los siete problemas para celebrar las matemáticas del nuevo milenio.

Existen problemas que requieren de este tiempo para resolverse. Por ejemplo, algoritmos que deciden si un enunciado es falso o verdadero, cierto formalismo lógico en el cual se pueden hacer enunciados de los números enteros de la forma:

$$\text{Si } x = 16, \text{ entonces no existe un número } y, \text{ tal que } y + y + y = x$$

Asimismo, existen problemas que requieren un tiempo triple, cuádruple o quíntuple exponencial. Y por si fuera poco, existen problemas más difíciles que cualquiera de éstos. De hecho, para cualquier problema que se le ocurra a una persona, existe siempre uno todavía más difícil.

1.7 RESUMEN

En la introducción al módulo se ubicó la computación a lado de otras grandes disciplinas científicas y se describió cómo todas ellas atacaron la soberbia del ser humano. Se explicó que el mundo de la computación está hecho de datos. Asimismo, se expusieron problemas que surgen de éstos y los métodos que existen para resolverlos, es decir, los algoritmos. Se vio que cuando aparece un problema, lo primero que debe hacerse es definirlo. Los problemas del mundo de la computación se pueden clasificar en los que pueden resolverse eficientemente en la práctica: los de la clase P, y los que quizá puedan resolverse eficientemente en la práctica, pero no se sabe cómo: los llamados NP-completos. También se abordaron los problemas de tiempo exponencial, que definitivamente no es posible resolver eficientemente por su gran dificultad: los exponenciales, dobles exponenciales, etc., y aun problemas peores, cuyo tiempo para resolverse crece más rápido que un exponencial. Para cualquier problema siempre existe uno todavía más difícil. Por otro lado, se mostró el significado de que una función crezca exponencialmente, y cómo su magnitud se dispara incluso para valores pequeños.

ALGORÍTMICA

TEMA

2

© *Latin Stock México.*

2.1 INTRODUCCIÓN A LA ALGORÍTMICA

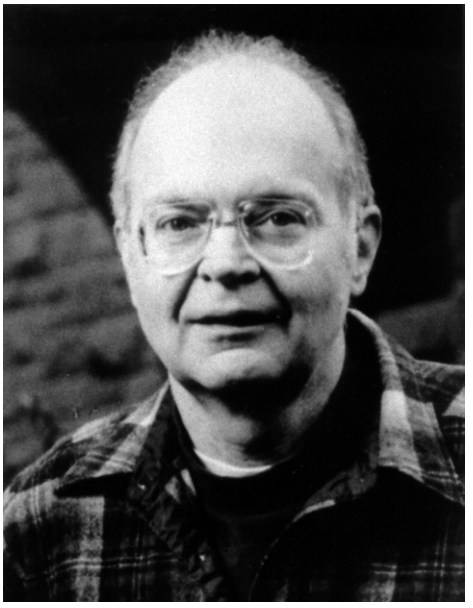
En el primer tema se vio que la gran mayoría de los problemas no se pueden resolver mediante una computadora en un tiempo razonable. Por el contrario, ahora se abordarán problemas para los cuales sí existen soluciones eficientes, y se estudiarán algoritmos para resolverlos. Más adelante, en el siguiente tema, se verá cómo implementar los algoritmos en un lenguaje de programación, para ejecutarlos en una computadora. Un algoritmo es la descripción de un método que resuelve un problema; un programa es la expresión de un algoritmo en un lenguaje de programación. La disciplina que interesa ahora es la *algorítmica*, cuya tarea es:

Dos ideas yacen brillando en el terciopelo del joyero: la primera es el cálculo; la segunda el algoritmo. El cálculo aunado al rico cuerpo de análisis matemático motivó que la ciencia moderna fuera posible; pero ha sido el algoritmo el que ha hecho posible el mundo moderno.
DAVID BERLINSKI, 2000.

1] Diseñar algoritmos correctos. Es decir, no sólo que resuelvan un problema, sino que además proporcionen maneras para cerciorarse de que lo hacen correctamente.
2] Comparar distintos algoritmos que resuelven el mismo problema, en relación con distintas medidas de eficiencia (qué tan rápido lo resuelve cada uno o qué tanta memoria utilizan), de manera que se pueda decir cuál es mejor respecto a alguna medida.
3] Buscar el mejor algoritmo para un problema. El reto incluye dos partes: diseñar un algoritmo y demostrar que no existe ninguno mejor.

Parafraseando a uno de los más grandes computólogos de la historia, Donald Knuth,[1] una persona bien entrenada en computación sabe de algoritmos: cómo construirlos, manipularlos, entenderlos y analizarlos. Este conocimiento es una preparación para mucho más que diseñar buenos programas de computadora: es una herramienta mental de propósito general que será de ayuda crucial para el entendimiento de otras disciplinas, ya sea la química, la lingüística, la música, etc. Al intentar formalizar un algoritmo, se llega a un entendimiento más profundo que la simple comprensión de las cosas de manera tradicional. En efecto, la algorítmica es más que una rama de la computación; constituye su núcleo, y es relevante en la mayor parte de la ciencia, los negocios y la tecnología, como bien dice David Harel.[2]

Donald Knuth (1938) Profesor emérito de la Universidad de Stanford, es considerado el padre del análisis de algoritmos. Galardonado con el Premio Turing en 1974, es autor del volumen *El arte de la programación de computadoras*, creador de TeX, aficionado a la teología y la música. En su casa tiene un órgano de dos pisos de altura.



Donald Knuth | *© Donald Knuth.*

Existen dos maneras de organizar un texto de algorítmica: por problemas y por paradigmas de diseño. La primera es apropiada para presentar la amplia gama de problemas que se estudian; es muy conveniente para el programador que necesita resolver alguno de ellos y busca un algoritmo que se ajuste a sus propósitos. En cuanto a este libro, dirigido al público en general, la segunda es más adecuada, ya que los paradigmas de diseño que se presentarán son mecanismos generales de solución de problemas, como "divide y vencerás", útiles en muchas disciplinas del conocimiento humano.

Galletas | *© Latin Stock México.*

2.1.1 Cocinar galletas

Úrsula se puso de acuerdo con su madre para cocinar la célebre receta de galletas de la abuela. Una vez adquiridos todos los ingredientes, los midieron según los requerimientos de la receta y procedieron a seguir las instrucciones de la abuela:

1] Precalentar el horno a 375 °F.
2] Pulverizar la avena en una licuadora.
3] Rallar el chocolate amargo.

[1] *Selected Papers on Computer Science*, Cambridge, U. Press, 1996.
[2] En su libro, *Algorithmics the Spirit of Computing*, 1992.

4] Hacer una mezcla con la mantequilla ablandada y los dos tipos de azúcar. Batir hasta incorporar todo.
5] Añadir a la mezcla los huevos y la esencia de vainilla. Batir hasta incorporar todo.
6] Añadir la harina y la avena. Batir hasta incorporar todo.
7] Añadir el chocolate rayado, la sal, el polvo de hornear y el bicarbonato. Batir hasta incorporar todo.
8] Agregar las chispas de chocolate y las nueces. Mezclar.
9] Hacer pequeñas bolitas del tamaño de una pelota de ping pong y colocarlas en una bandeja para hornear galletas. Deben separarse unos cuatro centímetros entre sí.
10] Hornear durante 10 minutos.

Luego de leer la receta, Úrsula pensó que era obvio que los pasos 4, 5, 6 y 7 se sintetizaran en uno: sólo había que poner juntos todos los ingredientes y proceder a batir.

—¡De ninguna manera! —exclamó su madre—. Si mezclas juntos todos los ingredientes vas a producir grumos en la masa. Lo mismo ocurre si alteras el orden de los pasos; la harina, por ejemplo, sólo puedes añadirla cuando hayas incorporado el huevo; de lo contrario, es imposible mezclarla.

—Por cierto, para hacer más obvia mi ignorancia, ¿qué significa "precalentar el horno"? —preguntó Úrsula.

—Significa prenderlo y poner el marcador de temperatura de acuerdo con lo que indique la receta y dejarlo así unos diez minutos al menos —respondió su madre.

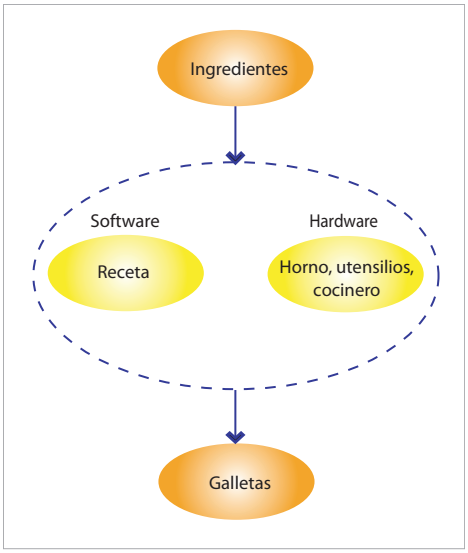
Úrsula se armó de paciencia y ayudó a su madre a mezclar y batir por turnos hasta que finalmente todo acabó en el horno.

2.1.2 Recetas de cocina *versus* algoritmos

Cualquier receta de cocina requiere ciertos ingredientes para su realización y produce, luego de seguir en orden una serie de instrucciones, el platillo que se desea preparar. Las instrucciones deben seguirse al pie de la letra para lograr el resultado deseado; en este caso, hacer 42 galletas.

Un algoritmo consiste en una secuencia finita de instrucciones, que se ejecutan cada vez que se corre el algoritmo, comenzando con los datos que recibe como entrada, y que termina al producir una salida. Cada vez que se corre puede recibir una entrada diferente.

Una receta de cocina es análoga al concepto de algoritmo en computación. Los ingredientes corresponden a los datos de entrada; las galletas juegan el papel de los datos de salida y los pasos de la receta el de las instrucciones del algoritmo, que transforman la entrada en la salida. Lo anterior constituye el software que es ejecutado por una máquina, entendida aquí, en un sentido muy general y abstracto, por Úrsula, su madre, el horno y demás utensilios de cocina. En resumen, el hardware lo constituyen las máquinas ejecutoras del algoritmo de las galletas de la abuela.



Analogía entre una receta de cocina y un algoritmo.

Nótese que este algoritmo tiene una sola entrada y produce una sola salida, aunque éstas sean compuestas: la entrada consiste en todos los ingredientes que especifica la receta, y la salida en las 42 galletas. Éste es un tipo básico de algoritmos, aunque los hay de muchos tipos, por ejemplo, los *algoritmos en línea*. Un algoritmo en línea es aquel que recibe entradas una y otra

vez a lo largo del tiempo, y genera las salidas correspondientes. Para aclarar cómo son, pensemos en la fluctuación del valor del peso respecto al dólar y en un algoritmo que monitorea esta relación: tomando una entrada el primer día de cada mes con los tipos de cambio se generan cálculos y se produce una salida, como el cambio promedio en lo que va del año.

La "máquina" que prepara las galletas debe saber qué significa batir las claras de huevo *a punto de turrón* o *tamizar la harina*, por ejemplo. Para un hardware diferente, quizá sea necesario dar instrucciones más detalladas, como: "Toma una cuchara, levántala 20 cm, inclínala 20 grados, realiza 35 giros a una velocidad de 15 por segundo, etcétera". O, por el contrario, una receta válida podría ser: "Prepara 42 galletas como las que hacía la abuela", pensada para que la hermana de Úrsula, quien vio a la abuela preparar las galletas muchas veces, la elabore. En otras palabras, las instrucciones deben ser adecuadas para la máquina que las ejecutará.

Esto permite establecer una diferencia importante entre recetas de cocina y algoritmos. La receta nos dice cómo preparar 42 galletas. Si queremos preparar 21 galletas, ¿qué hacemos? Si reducimos todos los ingredientes a la mitad, ¿en realidad saldrán 21 galletas?, ¿reducimos a la mitad el tiempo de horneado?, ¿a qué temperatura poner el horno? y ¿cómo modificamos la receta si requerimos 2 000 galletas para una boda? Inclusive si estamos interesados en exactamente 42 galletas, la receta no nos dice nada acerca de cómo preparar galletas de coco o de chocolate con nuez.

En realidad, la receta para preparar 42 galletas no enseña mucho acerca del arte de cocinar galletas. Si la receta dijera cómo preparar cualquier cantidad de galletas, se parecería más a un algoritmo, que acepta entradas de cualquier tamaño y produce una salida correspondiente. Un algoritmo especial para elaborar galletas diría mucho más acerca del proceso de hacer galletas; por ejemplo, indicaría cuál debe ser la temperatura del horno en función de la cantidad de masa de galletas utilizada, o cuántos huevos se requieren por cada galleta que se desee preparar.

Algunos algoritmos que se utilizan en comercio electrónico y cajeros automáticos son *probabilistas*. Es decir, incluyen instrucciones que deciden qué acción tomar al azar. Estos algoritmos pueden llegar a fallar, aunque con una probabilidad muy pequeña.

> **Un ejemplo:**
> Multiplíquense dos números. Recordar la tabla para multiplicar números del 1 al 9 es muy útil, ya que permite responder a cualquier multiplicación de números en este rango, rápido y sin necesidad de pensar. En este caso, un algoritmo válido para multiplicar 4×6 es simplemente regresar el resultado: 24. Sin embargo, saber multiplicar es más cercano a conocer un método como el que se aprende en la escuela, que facilita multiplicar cualquier pareja de números, de cualquier tamaño. Quizá el lector no se haya detenido a pensar lo maravilloso que es esto. Se tiene un algoritmo para multiplicar números arbitrariamente grandes, ¡de miles de billones de cifras! Claro, quizá sería imposible hacerlo porque tomaría millones de años, pero se sabe cómo se haría y que el algoritmo funciona siempre.

Justamente, los algoritmos son fascinantes por estar constituidos por una secuencia de pasos finita, es decir, un algoritmo debe tener un cierto número de instrucciones. En el caso de una receta para hacer 42 galletas probablemente esto no sorprende, pero en el caso de un algoritmo que debe funcionar para cualquier tamaño de entrada, sí. Esto hace que un algoritmo aporte mucha información acerca del problema.

Si se tuviese a disposición un número infinito de instrucciones, un algoritmo para multiplicar dos números consistiría en una tabla de multiplicar infinita, y para calcular el producto de dos números respondería con el valor correspondiente de la misma tabla de

multiplicar. Una tabla infinita como ésta no nos dice mucho acerca de la operación (multiplicación); de hecho, se podría resolver cualquier problema de esta forma con tan sólo una tabla infinita que para cualquier entrada indicara una salida válida al problema.

Un algoritmo es un procedimiento que puede seguirse mecánicamente, con el hardware adecuado, sin necesidad de interpretaciones, es decir, sin pensar qué hacer a cada momento. Por tanto, es importante que no exista ambigüedad en la interpretación de las instrucciones. Los términos usados deben significar exactamente lo mismo para cualquiera que lleve a cabo la receta. Lo anterior no significa que los algoritmos sólo hagan cosas sencillas; también hay los que juegan ajedrez al nivel de los mejores jugadores del mundo. Más aún, el que las instrucciones de un algoritmo deban estar exentas de ambigüedad, no significa que no se permita algo de libertad al cocinero. Un algoritmo puede incluir una instrucción no determinista que diga: "elige un número entre 1 y 10". Incluso podría incluir una instrucción probabilista, como indicar que se elija cualquiera de estos números con la misma probabilidad. Para algunos problemas, los mejores algoritmos son probabilistas, especialmente para criptografía. Más adelante se mostrará un algoritmo probabilista para ordenar números rápidamente. Pero cabe enfatizar que, aun cuando una instrucción pueda devolver distintos resultados, no debe existir ambigüedad en cuanto a su especificación.

Michael O. Rabin | *© Andrzej Lukaszewski.*

Michael O. Rabin (1931) Galardonado con el Premio Turing en 1976, ha contribuido, entre muchas cosas, con la introducción del concepto de no determinismo, así como con algoritmos probabilistas muy importantes, como el que determina si un número es primo y el criptosistema, que lleva su nombre.

2.1.3 Tipos de algoritmo

Los algoritmos se pueden clasificar de muchas formas: de acuerdo con el tipo de problema que resuelven, como algoritmos geométricos, algebraicos, de gráficas, para ADN, entre otros; de acuerdo con su tolerancia a las fallas, si las toleran o no, y qué tipo de fallas toleran.

A continuación se presentan algunos de los tipos de algoritmo más importantes. Según su procedimiento de ejecución, un algoritmo puede ser:

- *Secuencial*: ejecuta las instrucciones una por una y en orden.
- *Paralelo*: ejecuta al mismo tiempo distintas partes del algoritmo.
- *Distribuido*: consiste en varios algoritmos que corren en distintas computadoras, los cuales se comunican y colaboran para resolver un problema.

Según las instrucciones:

- *Deterministas*: cada instrucción realiza una sola operación; siempre la misma.
- *No deterministas*: algunas instrucciones tienen libertad para elegir qué resultado dar, entre varias opciones.
- *Probabilistas*: incluyen las instrucciones que pueden regresar distintos resultados, pero sin libertad de elegir; deben regresar el resultado según una distribución de probabilidades.

Curiosidades
Un fractal es una forma geométrica que, a pesar de tener un algoritmo recursivo simple que la define, es demasiado irregular para ser descrita por los métodos geométricos tradicionales, y tiene la característica de que puede dividirse en partes, cada una de las cuales es a su vez similar a la forma completa. El término fue inventado por Benoît Mandelbrot en 1975.

De acuerdo con las entradas y salidas:

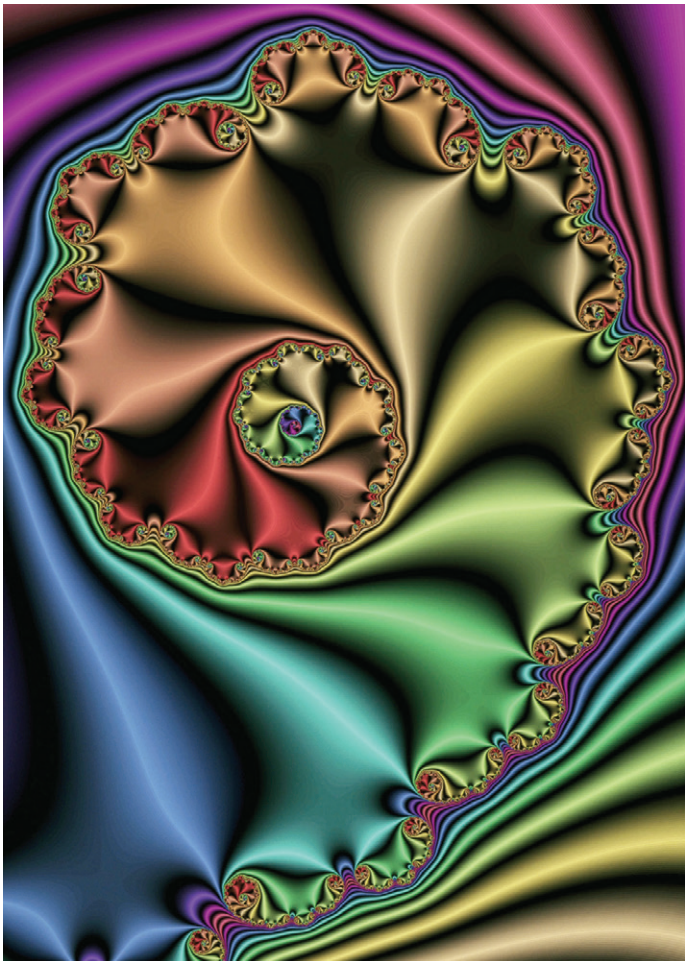
- Fuera de línea: recibe una entrada, ejecuta y produce una salida.
- En línea: recibe entradas una y otra vez y debe producir salidas conforme le llegan las entradas.

Por la máquina en la que corre:

- Tradicional: una computadora usual.
- Cuántica: hardware que aprovecha los principios de la física cuántica.
- Biológica: hardware basado en el cómputo del ADN.
- Otros: cómputo con regla y compás, computadoras analógicas, robots, etcétera.

Respecto a su precisión:

- Exacto: siempre devuelve una solución correcta.
- Aproximado: devuelve una solución cercana a un resultado óptimo o a un resultado correcto.
- Probabilista: devuelve una solución correcta con alta probabilidad.
- Heurístico: aparentemente funcionan bien, pero no sabemos con seguridad qué tanto o qué tan rápido.



Fractal | © *Latin Stock México.*

El algoritmo más sencillo, que se abordará con detalle en este capítulo, es el secuencial, determinista, fuera de línea.

En algoritmos deterministas cada instrucción realiza una sola operación, siempre la misma. Por lo tanto, la misma entrada, siempre que ejecutamos la secuencia de instrucciones del algoritmo, producirá la misma salida. Pero no hay que dejarse engañar: inclusive en este caso un algoritmo puede tener un comportamiento bastante complicado, con ejecuciones que nunca terminan y que producen múltiples salidas a lo largo de su ejecución. De hecho, en cierto sentido pueden llegar a ser impredecibles y artísticos.

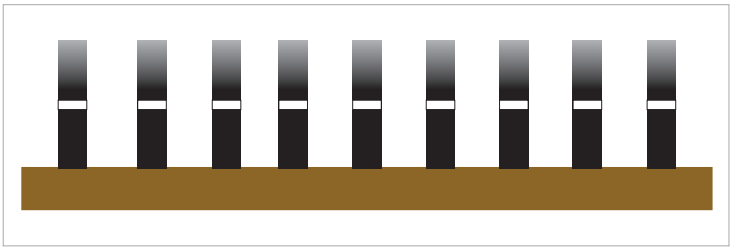
2.2 INDUCCIÓN Y GRÁFICAS

2.2.1 El método de inducción

La clave para analizar muchos problemas computacionales y matemáticos radica en una técnica llamada inducción. ¿Cómo poder cerciorarse de que un algoritmo funcionará siempre para cualquier entrada que se le proporcione, independientemente del tamaño de ésta? Es usual

en los algoritmos repetir un procedimiento una y otra vez hasta encontrar el resultado que se busca. La idea es probar que si el algoritmo funcionó bien hasta cierto punto, es decir, después de ejecutar el procedimiento cierto número de veces (n), al repetir el procedimiento una vez más (+1), seguirá trabajando bien, habiendo repetido el procedimiento $n + 1$ veces.

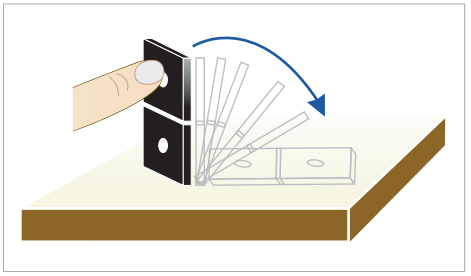
El argumento inductivo descrito en el párrafo anterior alude a lo que sucede cuando se colocan fichas de dominó en fila…

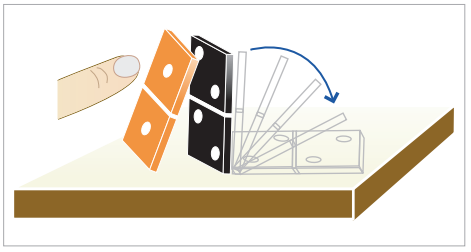


y se tira la primera ficha.

Se sabe que se caerán todas porque:

1] Si se empuja una ficha se cae.
2] Si una ficha se cae y está colocada correctamente junto a la ficha que le sigue, esta última también se caerá.





A partir de estos dos hechos se concluye que todas las fichas se caerán.

A continuación se mostrarán algunos ejemplos de cómo se utiliza un argumento de inducción para probar que un enunciado se cumple para todo valor de n, $n = 1, 2…$

Curiosidades
El Día del Dominó es un evento anual organizado por Robin Paul Weijers en Holanda, donde se construyen caminos de fichas de dominó, de manera que vayan cayendo una por una. En la celebración de 2006 los participantes que obtuvieron el récord mundial lograron colocar 4 400 000 fichas de las cuales cayeron 4 079 381.

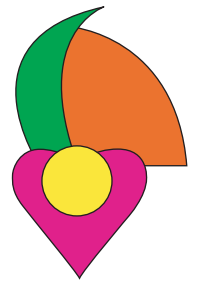
2.2.2 Colorear mapas

En general, el problema de colorear mapas consiste en elegir el color de cada país, de tal forma que siempre que dos países tengan una frontera común, posean colores diferentes (si dos países se tocan en un solo punto no comparten frontera). Esto es muy fácil de

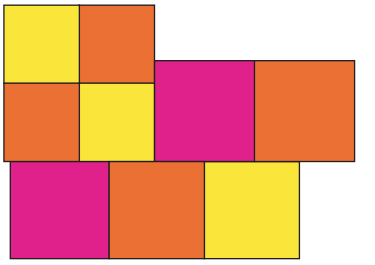
Curiosidades
En 1852, mientras Francis Guthrie intentaba colorear un mapa de los condados de Inglaterra, notó que cuatro colores eran suficientes para lograr su objetivo. En ese entonces Guthrie era estudiante, en la University College de Londres, de Augustus de Morgan, a quien pidió una explicación de por qué aparentemente es posible colorear cualquier mapa con cuatro colores. Así nació un problema que tomó más de 120 años de esfuerzo de muchos matemáticos para resolverlo. En 1976 se logró demostrar, con el auxilio de un programa de computadora que, en efecto, es posible colorear cualquier mapa con cuatro colores.

realizar, simplemente eligiendo un color distinto para cada país. El problema se complica cuando hay que usar pocos colores.

Hay muchos problemas de coloración que se estudian en matemáticas y cuyas aplicaciones en computación son de gran importancia, más que para colorear mapas para situaciones donde se necesita asignar recursos evitando conflictos. Más adelante se darán algunos ejemplos.



Este mapa no se puede colorear con menos de cuatro colores.



Éste se puede colorear con tres, pero no con dos.

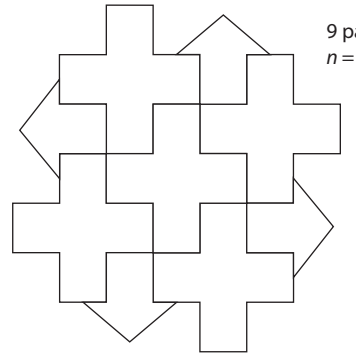
Tomó 123 años probar que cualquier mapa se puede colorear con cuatro colores y la prueba es complicadísima. Es mucho más sencillo demostrar que cinco colores son suficientes, pero aquí se probará algo aún más fácil: que es posible colorear cualquier mapa con seis colores como máximo.

Se analizará si este enunciado se cumple, mediante la técnica de inducción, para cualquier mapa de n países, así como para cualquier mapa de $n + 1$ países. Primero se abordarán casos pequeños.

Base de la inducción. Obviamente, el enunciado se cumple para mapas de máximo seis países, al designar a cada uno un color distinto.

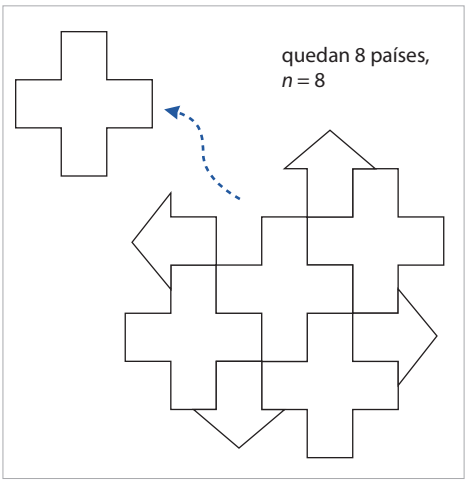
Paso de inducción. Supóngase que cualquier mapa de n países se puede colorear con seis colores, si se considera que cualquier mapa tiene, necesariamente, al menos un país que comparte frontera con cinco países como máximo.

Lo anterior se puede verificar con algunos ejemplos. Probarlo no es fácil,[3] pero vale la pena intentarlo. Tómese un mapa de $n + 1$ países.

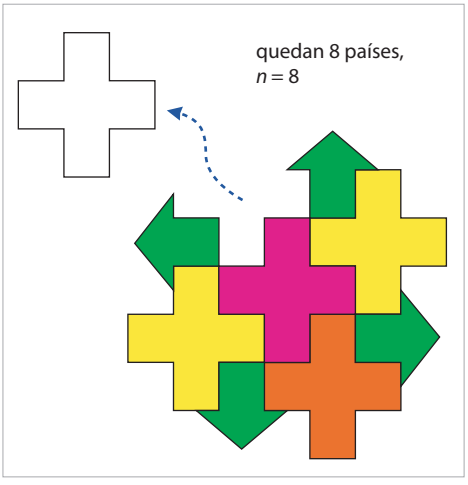


[3] Es un corolario de la fórmula de Euler que afirma que en cualquier gráfica plana el número de vértices menos el de aristas más el de caras siempre es igual a 2.

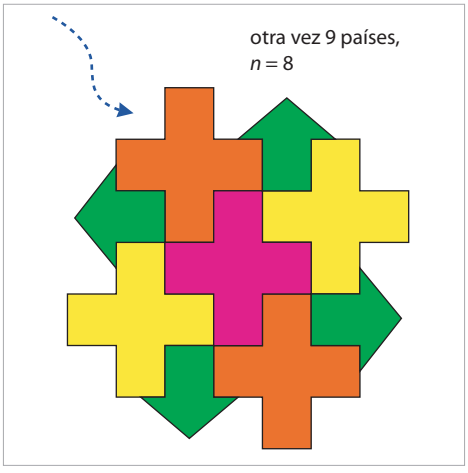
Después, elimínese un país que tenga frontera con cinco o menos países.



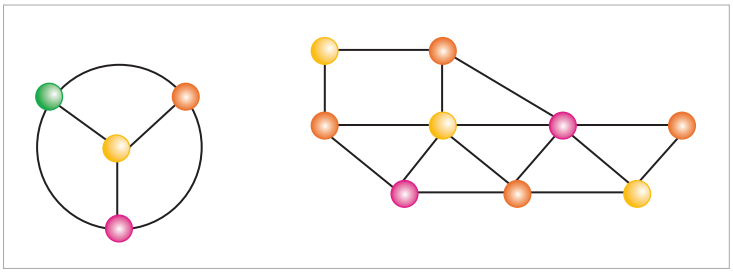
De acuerdo con la hipótesis de inducción, el mapa resultante se puede colorear con seis colores.



Regrésese el país a su lugar original. Como tiene cinco vecinos, se le puede asignar, de los seis colores disponibles, uno que no use ninguno de sus vecinos.



Con esto se ha concluido la prueba de que siempre se puede colorear un mapa con seis colores. Más aún, ahora se tiene una idea de cómo diseñar un algoritmo eficiente para colorear un mapa con máximo seis colores. Un reto interesante es diseñar uno antes de ver el algoritmo propuesto más adelante.



2.2.3 Gráficas

Concepto
Una gráfica es un conjunto de puntos llamados vértices, algunos de los cuales están conectados por líneas llamadas aristas.

Es claro que la forma de los países no es realmente lo que importa, sino con cuáles comparten frontera. Se puede abstraer la situación dibujando un punto que represente cada país y después conectar dos de ellos mediante una línea, cuando comparten frontera. A continuación se muestran las gráficas correspondientes a los mapas ubicados al inicio de la sección anterior.

Cabe señalar que ya se han usado estructuras de este tipo en el tema 1, en forma de árboles, es decir, sin ciclos (y conexas). Ahora se verán en una versión más general. Las gráficas aparecen por todas partes en computación y en otras disciplinas, ya que sirven para modelar un conjunto y las relaciones entre sus elementos.

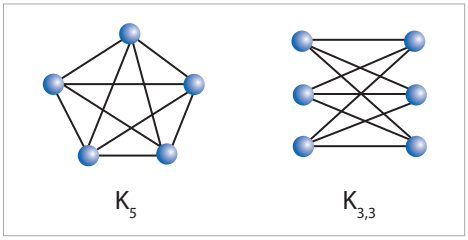
En el problema de coloración la entrada es una **gráfica** y la salida es una asignación de números (llamados colores) entre 1 y k a los vértices, de manera que dos vértices conectados por una arista reciban colores diferentes, y la k sea lo más pequeña posible.

La boda de Úrsula

"¡Arcadio, no es gracioso!", gritó Úrsula frustrada, al escuchar que éste no tenía aún la lista de invitados de su familia para la boda. "Ya sólo faltan tres meses, no podemos perder más tiempo", le explicó. En sus sueños, Úrsula ya está pensando en la boda, incluso tiene la lista de personas a las que va a invitar; sin embargo, el mayor problema que enfrenta es decidir la cantidad de mesas que deberá alquilar para la fiesta.

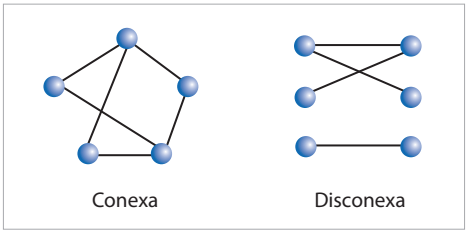
Supone que mientras menos mesas alquile, más económica resultará la fiesta, y que además hay mesas de todos tamaños. Es decir, lo ideal sería alquilar una sola mesa, donde se sienten juntos todos los invitados. Desafortunadamente, algunos invitados no mantienen buenas relaciones entre sí y no están dispuestos a sentarse en la misma mesa. Úrsula rápidamente se percata de que tiene que resolver un problema de coloración. Las personas corresponden a vértices, y dos personas tienen una arista si están disgustadas. Cada color representa una mesa, por lo que necesita un algoritmo que coloree a las personas y que utilice el menor número de colores posible. En este momento sonó el despertador y se despertó. Con una sonrisa dibujada en el rostro, corrió al teléfono para contarle a Arcadio su sueño.

Las gráficas que representan países de un mapa tienen la peculiaridad de que se pueden dibujar sin que se crucen las aristas. A estas gráficas se les llama gráficas planas. No todas las gráficas son planas. Por ejemplo, si intentas dibujar una gráfica de cinco vértices, en la que todos los vértices tienen aristas con todos los demás, verás que resulta imposible dibujar sin que se crucen aristas; ésta no corresponde a ningún mapa. Otro ejemplo básico es representar con puntos tres niños, tres niñas y aristas de todos los niños a todas las niñas. Se tienen así las gráficas denominadas K_5 y $K_{3,3}$.



¿Cuántos colores hacen falta para colo-rear K_5? Claramente cinco. ¿Y para K_6, la gráfica completa sobre seis vértices? Pues seis. Así que, mientras que una gráfica plana siempre se puede colorear con cuatro colores, otras gráficas requieren de un número arbitrariamente grande de colores.

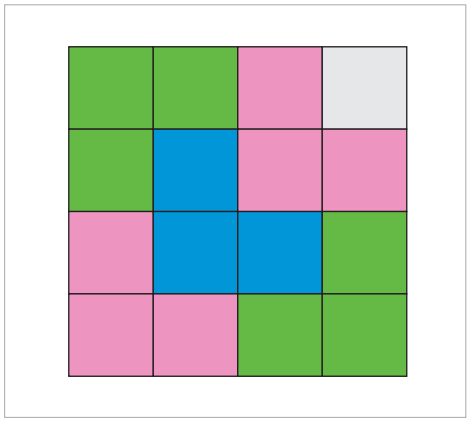
Una gráfica conexa tiene la propiedad de permitir llegar de cualquier vértice a cualquier otro caminando por las aristas. Una gráfica correspondiente a un mapa con dos continentes diferentes sería una gráfica disconexa.



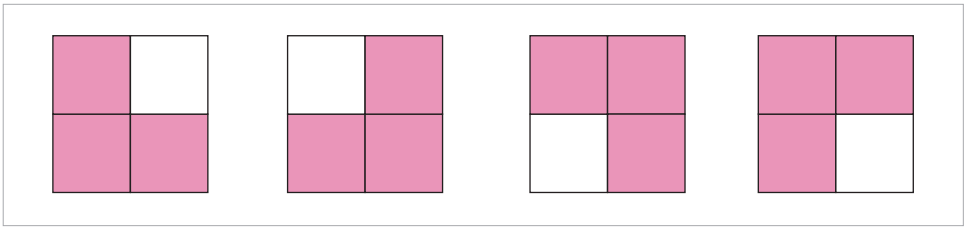
Por lo tanto, hay gráficas de varios vértices que se pueden colorear con un solo color: gráficas sin aristas (disconexas).

2.2.4 Acertijo de los tróminos

Inspirados en las fichas de dominó, están los tróminos. Un trómino es una ficha en forma de L, compuesta por tres cuadrados iguales. Se tiene un tablero de $2^n \times 2^n$ cuadros, uno de los cuales posee una marca. Es posible cubrir todo el tablero, menos el cuadro marcado, con tróminos. El siguiente ejemplo muestra cómo cubrir un tablero de 4×4 con cinco tróminos ($n = 2$), dejando libre el de arriba a la derecha.



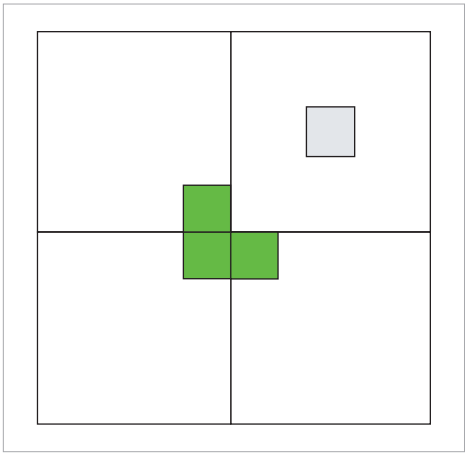
Para probar que siempre es posible realizar lo anterior, se usará la inducción. Primero la base de la inducción: se comprueba que para $n = 1$ se cumple el enunciado:

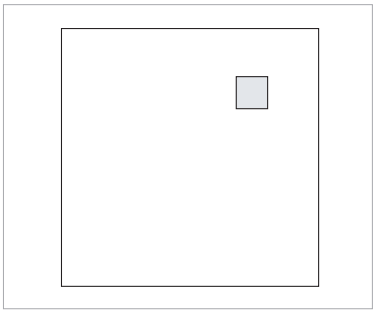


Curiosidades
Es posible usar una gráfica para representar un circuito electrónico. Para construirlo y almacenarlo en un chip (circuito integrado) es necesario que no se crucen cables. Con el fin de evitarlo, se usan algoritmos muy eficientes que dibujan la gráfica sin que se crucen las aristas, como el inventado por John Hopcroft y Robert Tarjan, dos grandes computólogos que recibieron, por esta y otras contribuciones a la algorítmica, el Premio Turing en 1986.

Ahora, el paso de inducción: la hipótesis consiste en suponer que es posible cubrir con tróminos cualquier tablero de $2^n \times 2^n$ cuadros, uno de los cuales tiene una marca. Posteriormente, se probará que es posible cubrir con tróminos cualquier tablero de $2^{n+1} \times 2^{n+1}$ cuadros, uno de los cuales tiene una marca:

Se divide en cuatro cuadrantes iguales, colocando un trómino en el centro:



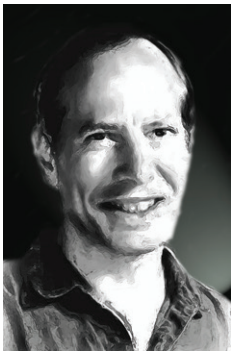


John Hopcroft.

Cada cuadrante es de $2^n \times 2^n$ y tiene un cuadro marcado, ya sea el original o uno cubierto por el trómino del centro. Por medio de la hipótesis de inducción, existe manera de cubrir cada cuadrante con tróminos; al final queda un solo cuadrado sin cubrir, justo el que estaba marcado originalmente.

2.2.5 Probando funciones por inducción

Funciones cuadráticas



Robert Tarjan.

La siguiente función aparece en análisis de algoritmos cuando una operación se repite una vez, luego dos veces, luego tres, n veces. La función está definida para todo número entero n mayor o igual a 1:

$$f(n) = 1 + 2 + 3 + \ldots + n$$

Es decir,

$$f(1) = 1, f(2) = 1 + 2 = 3, f(3) = 1 + 2 + 3 = 6$$

etcétera. Resulta que, en general,

$$f(n) = n(n + 1)/2$$

y es posible demostrarlo por inducción. Como en cualquier prueba por inducción, se comienza con la base.

Base: para $n = 1$, se corrobora que $f(1) = 1(1 + 1)/2 = 1$, lo cual corresponde a la definición de f.

Paso de inducción: supóngase que para un valor de n es cierto que:

$$f(n) = n(n + 1)/2$$

Entonces será igualmente cierto para el siguiente valor, n+1. Es decir, se debe probar que:

$$f(n + 1) = (n + 1)\ (n + 2)/2$$

En efecto, por definición,

$$f(n + 1) \text{ es igual a } 1 + 2 + 3 + \ldots + n + (n + 1)$$

Pero en la hipótesis de inducción se asume que los primeros n términos de esta suma, es decir $f(n)$, son iguales a $n(n + 1)/2$ y, por lo tanto, se les puede reemplazar por este valor:

$$f(n + 1) = n(n + 1)/2 + (n + 1)$$

lo cual, con un poco de álgebra, toma la forma esperada:

$$f(n) = n(n + 1)/2 + 2(n + 1)/2$$
$$f(n) = n(n + 1)/2 + 2(n + 1)/2$$
$$f(n + 1) = n(n + 1)\ (n + 2)/2$$

Función factorial

Una función similar que emplea la multiplicación en lugar de la suma es la siguiente:

$$f(n) = 1 \times 2 \times 3 \times \ldots \times n$$

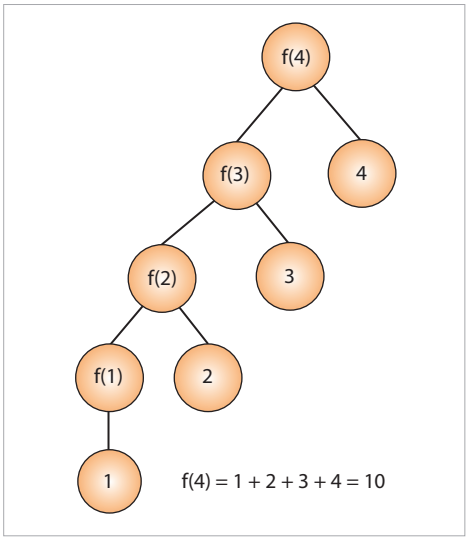
Ésta es la famosa función factorial. El número que resulta de estas multiplicaciones se denomina factorial de n y se denota $n!$, es decir que $f(n)$ y $n!$ es lo mismo. Por ejemplo, cuando n vale 4, $n! = 1 \times 2 \times 3 \times 4 = 24$. Nótese que esta función tiene la misma estructura que la función cuadrática. Ambas pueden ser definidas en términos de sí mismas, de manera recursiva. La forma recursiva es muy elegante y, como cualquier definición de este tipo, consta de dos partes: la base y el caso general.

Para la función cuadrática, se tiene que $f(n) = 1 + 2 + \ldots + (n - 1) + n$, pero los primeros $n - 1$ términos no son más que $f(n - 1)$:

$$f(n) = f(n - 1) + n$$

Lo mismo para $f(n - 1)$; ésta es igual a $f(n - 2) + n - 1$, y así sucesivamente hasta $f(1) = 1$.

Gráficamente se le puede representar mediante un árbol, como el que aparece a la derecha.



Y para la función factorial:

$$1! = 1$$
$$n! = (n - 1)\,! \times n$$

mediante la inducción puede mostrarse que la definición recursiva es igual a la iterativa. Una prueba por inducción de que la definición recursiva es igual a la iterativa para la función factorial es la siguiente:

Base: para $n = 1$ son iguales. Ambas definiciones dicen que $1! = 1$

Paso de inducción: Asúmase que ambas definiciones son iguales para un cierto valor, por ejemplo m. Ahora, se desea mostrar que también lo serán para el que sigue, $m + 1$. Según la definición recursiva:

$$(m + 1)! = m! \times (m + 1)$$

Pero se está suponiendo que $m!$ es lo que la definición iterativa plantea, es decir,

$$m! = 1 \times 2 \times \ldots \times (m - 1) \times m$$

así que es posible reemplazar esto en la fórmula anterior y obtener:

$$(m + 1)! = 1 \times 2 \times \ldots \times (m - 1) \times m \times (m + 1)$$

Imagen de un kilogramo | *© Latin Stock México.*

Función exponencial

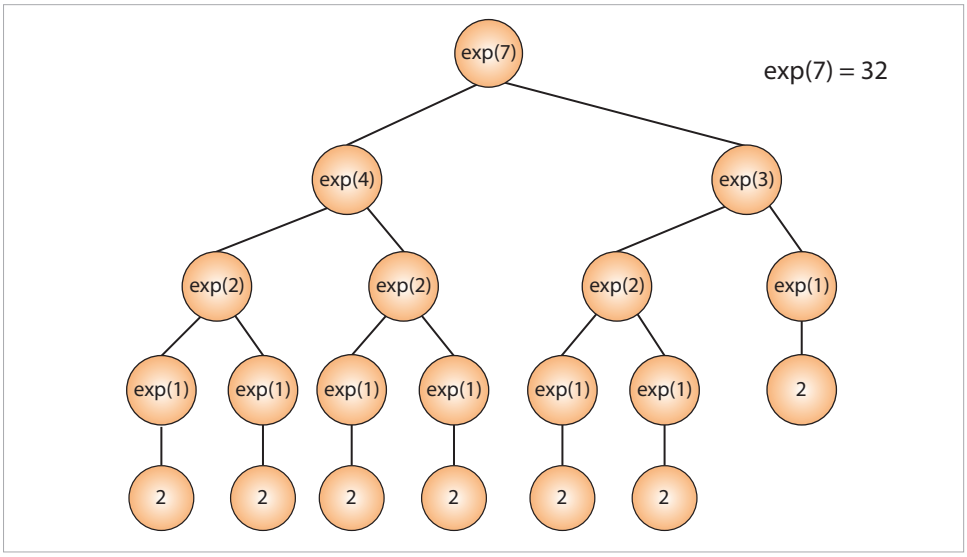
La función exponencial $\exp(n)$ se puede analizar de manera similar. Si se admite que es igual multiplicar 2 por sí mismo n veces, que hacerlo por partes (una parte de las multiplicaciones y luego el resto) se obtiene la definición recursiva:

$$\exp(n) = \exp(n_1) \times \exp(n_2)$$

donde n_1 y n_2 son números menores a n, tales que $n = n_1 + n_2$, hasta llegar a la base:

$$\exp(1) = 2$$
$$\exp(0) = 1$$

Un árbol correspondiente al caso de $n = 7$ es el siguiente:



Curiosidades
Una definición circular es aquella donde se expresa el concepto definido, y por lo tanto es inútil. Originalmente se definía el kilogramo como la masa de un litro de agua a la presión estándar y temperatura a la que tiene su máxima densidad (aproximadamente cuatro grados). ¡Pero la definición de presión se expresa en términos del kilogramo! Por esta razón se cambió la definición: un kilogramo es la masa de una cierta pieza de metal en la ciudad de Sèvres.

2.3 RECURSIVIDAD

Para comenzar se ilustrarán los tres aspectos de la algorítmica a través del problema de las torres de Hanoi: diseño, análisis y optimalidad. Se analizará la noción de algoritmo y cuáles son las maneras para medir su eficiencia. En el camino aparecerá un paradigma de solución de problemas muy importante: la recursividad, y se verá la utilidad de la técnica de inducción. Pero primero se trabajará con un ejemplo sencillo: colorear mapas.

2.3.1 Algoritmo para colorear mapas con seis colores

Recuérdese la prueba por inducción para el problema de colorear mapas. El argumento es éste: suponiendo que se puede colorear cualquier mapa de n países con seis colores, es posible hacer lo mismo para cualquier mapa de $n + 1$ países.

Así que a Arcadio se le ocurre un algoritmo muy sencillo para colorear un mapa, con ayuda de amigos. Cuando Arcadio desea colorear un mapa, primero busca un país, digamos X, que tenga a lo más cinco vecinos. Luego copia su mapa a otro papel, pero borra X. Le pasa el mapa a Úrsula y le pide que lo coloree con seis colores. Cuando ella se lo regresa, usa esos colores para colorear su mapa original y colorea a X con un color distinto al de sus vecinos. Éste es todo el algoritmo, ya que Úrsula hace lo mismo. Borra un país de igual manera y le pide ayuda a su hermanito para que coloree lo que queda del mapa. Eventualmente, después de borrar países del mapa, queda uno con sólo seis países, y para colorearlo no hace falta pedir ayuda a nadie.

> **Colorea con seis** (mapa de n países)
> Si n es seis o menos, asignar un color a cada país, entre uno y seis.
> De otro modo, **colorea con seis** (mapa de $n - 1$ países quitando algún país X que tenga a lo más cinco vecinos).
> Colorea X con un color distinto al de sus vecinos.

Éste es un ejemplo de **algoritmo recursivo**, que se ejecuta él mismo una y otra vez, pero cada vez con entradas más pequeñas, hasta llegar a una tan chica que se resuelve

trivialmente. Para probar que un algoritmo de este tipo es correcto, la técnica de la inducción es ideal, como se ve en este ejemplo, donde el algoritmo sigue la estructura de la prueba por inducción paso a paso.

¿Qué tan eficiente es el algoritmo? Al parecer bastante, ya que para colorear n países se pide ayuda n veces. El único detalle es que cada vez que se pide ayuda hay que localizar un país con cinco vecinos como máximo. Esto se puede evitar y obtener un algoritmo cuyo tiempo de ejecución sea proporcional a n.[4]

2.3.2 Un problema y un algoritmo: las torres de Hanoi

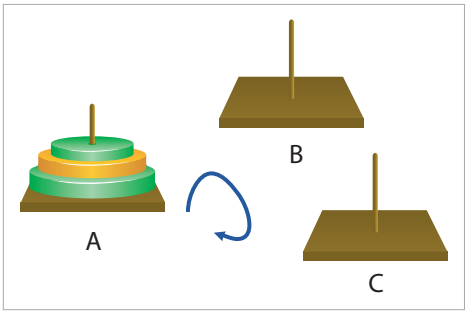
Ahora se verá otro ejemplo con más detalle, en el que es necesario pedir ayuda dos veces en lugar de una, para resolver problemas más pequeños.

El problema

Regresando al problema de las torres de Hanoi del capítulo 1, se tienen tres torres: A, B y C. Al principio, en una de las torres, por lo general la torre A, se encuentran apilados n anillos, digamos $n = 3$ anillos, de distintos diámetros, colocados de mayor a menor en forma ascendente. El objetivo del problema es, dando como entrada el número n, producir como salida los movimientos necesarios para mover todos los anillos a otra torre, uno por uno. Cada movimiento debe efectuarse de la siguiente forma: tomar el anillo que se encuentra hasta arriba de la torre X y colocarlo hasta arriba de la torre Y, donde X y Y pueden tomar cualquiera de los valores A, B o C. Este movimiento puede ejecutarse siempre y cuando el anillo que se coloque en la torre Y no descanse sobre un anillo de menor diámetro.

Antes de diseñar un algoritmo que resuelva el problema, es conveniente colocar las torres en círculo; si se colocaran en línea se establecería una diferencia entre las torres: una sería la primera y otra la última. En cambio, en un círculo todas son iguales y se pueden recorrer en el orden de las manecillas del reloj, por ejemplo.

El algoritmo

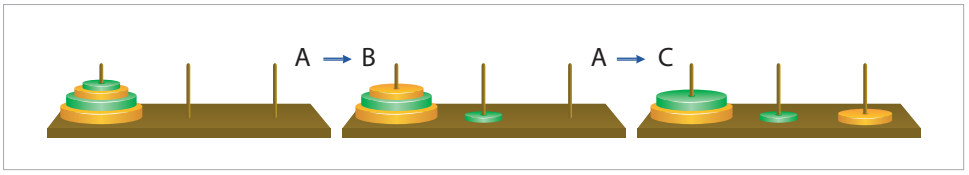


Obsérvese que un algoritmo es una descripción detallada y precisa de cómo resolver un problema, que consiste en una secuencia finita de instrucciones. Esta secuencia recibe algunos datos como entrada, los procesa a través de las instrucciones y produce datos como salida. Asimismo, las instrucciones de un algoritmo están diseñadas para ejecutarse a través de una máquina que entiende lo que significa cada instrucción, y sabe cómo ejecutarla.

[4] El algoritmo es muy simple, pero se requiere conocer las estructuras de datos para representar gráficas.

En el caso de las torres de Hanoi, las instrucciones indican que se debe tomar el anillo situado hasta arriba de la torre *X* y colocarlo hasta arriba de la torre *Y*, donde *X* y *Y* pueden adoptar cualquiera de los valores A, B o C.

El diseño del algoritmo utiliza una idea sencilla. Al principio no hay otra opción que mover el anillo más pequeño, por ejemplo, a la torre B. Una vez que se mueve, si $n = 1$, se ha resuelto el problema. Si $n > 1$, como no tiene caso desplazar ese anillo otra vez, sólo queda un anillo que se puede mover, es decir, el que le sigue en tamaño, y sólo hay una torre en donde es posible ubicarlo, la torre C.

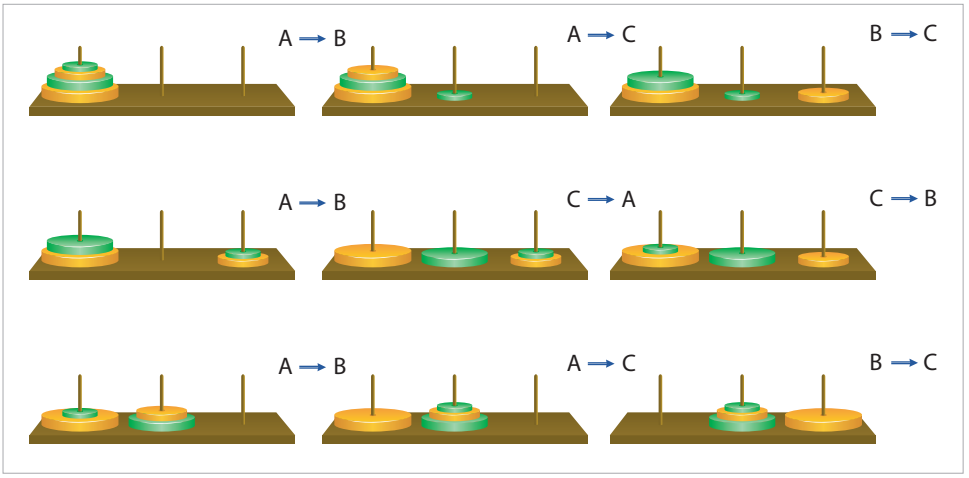


Si se piensa un poco, se observa que al que se debe mover ahora es al menor, y que conviene moverlo a la torre C, ya que si se mueve a la torre A, se terminaría en una configuración equivalente, pero habríamos ejecutado tres movimientos en lugar de uno.

Eventualmente es necesario trasladar el anillo más pequeño y debe hacerse en sentido circular, de la torre A a la B y posteriormente a la C, para luego regresar a la A, y así sucesivamente. Una vez que se ha movido, no tiene caso desplazarlo de nuevo, por lo que no hay otra opción que utilizar otro anillo; de hecho, sólo hay un anillo que se puede mover o ninguno si ya se terminó de mover todos los anillos. El algoritmo es el siguiente:

1] Repetir lo siguiente hasta que el paso (1.b) no se pueda ejecutar:
(1.a) Mover el anillo más pequeño a la torre siguiente en el orden de las manecillas del reloj.
(1.b) Ejecutar el único movimiento permitido posible que no involucre al anillo más pequeño.

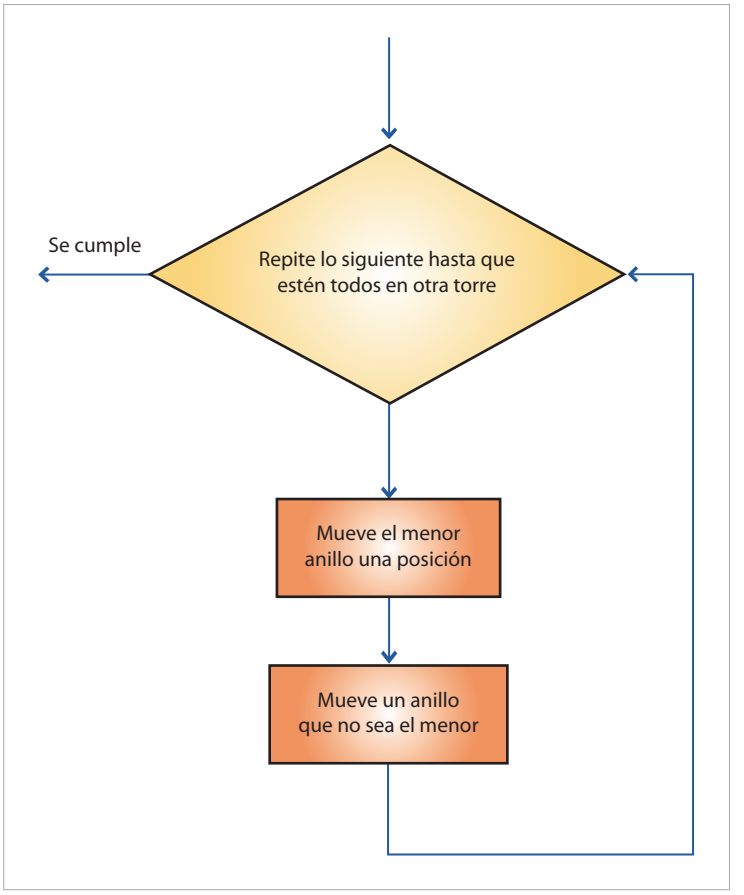
En la siguiente figura se muestra la ejecución de las ocho primeras instrucciones del algoritmo, con las torres A, B y C ordenadas en círculo.



Concepto
Un algoritmo resuelve un problema computacional, pero no necesariamente acerca de computadoras.

El modelo de cómputo

Como se vio anteriormente, para hablar de un algoritmo se requiere saber a qué máquina se dirige y qué instrucciones sabe ejecutar. Las instrucciones deben ser lo suficientemente precisas para que al utilizarlas en el diseño del algoritmo, se aprenda algo acerca del problema. Un algoritmo que diga: "Ve con un chamán y dile que adivine la secuencia de movimientos", no dirá mucho acerca de las torres de Hanoi.



Las instrucciones

Lo primero que se deja entrever es que el algoritmo consiste en una lista finita de instrucciones, y sin embargo puede resolver cualquier cantidad de entradas al problema; es decir, no importa con cuántos anillos *n* comencemos, ni en qué torre estén acomodados, el algoritmo resuelve el problema. Es claro, entonces, que cualquier algoritmo requerirá de algún tipo de instrucción que implique ejecutar repetidamente un bloque de instrucciones. Es posible visualizar al algoritmo anterior mediante un diagrama de flujo.

Nótese que no todos los algoritmos requieren instrucciones de repetición. Si el problema es muy simple, podría no hacer falta este tipo de instrucciones. Por ejemplo, si el problema pide tomar el anillo más pequeño, un algoritmo útil sería simplemente identificar en cuál de las tres torres están colocados los anillos y tomar el anillo situado en la parte superior de la pila.

El problema computacional

Ahora bien, incluso sin tener que analizar con detalle las instrucciones del algoritmo y sin necesidad de que el problema lo diga explícitamente, debe notarse que no está enfocado para resolver un problema de robótica o ingeniería mecánica. No se piensa en el hecho de que para resolver un problema de esta índole es necesario diseñar un dispositivo que pueda localizar, mediante la visión, dónde está un anillo y construir un brazo mecánico que lo sujete y mueva otra torre. Más bien, el interés se encuentra en qué secuencia de movimientos de discos es la que resolvería el problema. Ni siquiera se trata de un problema físico, no es necesario tener a la mano un juego de madera con tres varillas y anillos; es factible pensar en soluciones, dibujarlas en papel, inclusive mediante símbolos que indiquen cuántos anillos tiene cada torre. Por supuesto, aunque se pueden imaginar las soluciones y los movimientos de los anillos, tampoco se trata de un problema de psicolo-

gía. Es un problema de computación, a pesar de que se propone una solución y se analiza su eficiencia y corrección, sin necesidad de una computadora.

En efecto, la entrada al algoritmo en realidad no son los anillos ni las varillas de madera. La entrada es sencillamente el número *n*. Lo sorprendente es que la *n* no aparece en el algoritmo y sin embargo funciona para cualquier valor de *n*. Como siempre, la entrada debe ser válida; no sirve de nada un número como 3.45. En este caso, cualquier número entero mayor o igual a cero es una entrada válida. La salida es una lista de movimientos que, al ejecutarlos, permiten que los discos se desplacen a otra torre sin que se haya realizado un movimiento prohibido.

Curiosidades
Un algoritmo para el problema de las torres de Hanoi recibe como entrada un número entero (el número de discos) y produce como salida una lista de movimientos.

Un problema siempre supone un cierto nivel de detalle. En el caso del algoritmo de las torres de Hanoi, se supone, entre otras cosas, que se puede localizar la torre con el anillo más pequeño.

¿Pero cuáles son las instrucciones de este algoritmo? Se había mencionado que las instrucciones son: tomar el anillo ubicado en la parte superior de la torre *X* y colocarlo en la parte superior de la torre *Y*. Como hay un solo anillo hasta arriba de una torre y siempre se coloca hasta arriba de otra, es suficiente señalar de qué torre a qué torre se ejecuta el movimiento. Es decir, se puede simplemente usar instrucciones de la forma:

$$X => Y$$

Entonces, la salida del algoritmo debe ser una lista de este tipo de intrucciones. Por ejemplo, para $n = 2$ discos, una salida válida sería:

$$A => B, A => C, B => C$$

para que los dos discos queden acomodados en la torre C.

Se podría pensar que éste es el único tipo de instrucción requerido, además de las instrucciones de repetición mencionadas. Sin embargo, al analizar las instrucciones del algoritmo, se puede ver que son más complicadas:

- *Instrucciones de repetición*. Este algoritmo tiene una sola instrucción de este tipo, la 1: "Repite lo siguiente, hasta que el paso 1.b no se pueda ejecutar". Esta instrucción, como cualquier otra instrucción de repetición, tiene tres componentes: primero, su nombre, el que nos indica qué se pretende hacer, en este caso, "repite"; en segundo lugar, una manera de indicar cuál es el bloque de instrucciones que se deben repetir, que en este caso se trata simplemente de las dos instrucciones que siguen, la 1.a y la 1.b; finalmente, es necesario indicar cuántas veces se debe repetir el bloque de instrucciones. Se solicita repetirlo hasta que se cumpla determinada condición; el ejemplo indica repetir "hasta que el paso b no se pueda ejecutar", es decir, hasta que todos los anillos se encuentren en una sola torre.
- *La siguiente instrucción, la 1.a, es*: "Mueve el anillo más pequeño a la torre siguiente en el orden de las manecillas del reloj". Se supone que quien ejecuta esta acción puede localizar la torre con el anillo más pequeño, además de mover un anillo de una torre a otra.
- *La última instrucción, la 1.b, indica*: "Ejecuta el único movimiento permitido posible que no involucre al anillo más pequeño". Es de suponer que el ejecutor de la acción puede buscar una torre con el anillo que le sigue al más pequeño en tamaño, o bien, decir si no hay tal, debido a que todos se encuentran en una sola torre. De este modo se señala que las repeticiones solicitadas por la primera instrucción han concluido.

Concepto
Se puede evaluar qué tan bueno o eficiente es un algoritmo desde muchas perspectivas. Por ejemplo, cuánto tiempo emplea para resolver una tarea, cuánta memoria requiere, qué tan largo y difícil de entender es, etcétera.

La complejidad y el modelo de cómputo

El modelo de cómputo describe la máquina a la cual se hace referencia, el dispositivo, físico o imaginario, que sabe cómo ejecutar un conjunto de instrucciones dado. La algorítmica estudia los métodos para resolver problemas en un determinado modelo de cómputo, con especial interés en que las soluciones sean eficientes.

Para analizar la **eficiencia de un algoritmo** es necesario saber algo más acerca del modelo de cómputo; por ejemplo, los costos que conlleva ejecutar un algoritmo en el modelo. Se pueden evaluar muchos costos, pero el más importante es, por lo general, el tiempo que lleva ejecutar el algoritmo. Como no se conoce de antemano el dispositivo físico que lo ejecutará (o incluso si se pretende que algún dispositivo lo ejecute), lo que se hace con frecuencia es suponer que cada instrucción emplea una unidad de tiempo para ser ejecutada, sin necesidad de especificar el tipo de unidad (un segundo, un minuto o alguna otra).

Concepto
La complejidad de tiempo del algoritmo se define como el número de instrucciones a ejecutar, como función del tamaño de la entrada.

Por lo tanto, el tiempo utilizado para ejecutar un algoritmo equivale al número de instrucciones para resolver el problema. En el caso de las torres de Hanoi, simplemente se cuentan las instrucciones 1.a o 1.b, sin entrar en más detalles que no se conocen de antemano, como en cuántos segundos se ejecuta una instrucción en una computadora. Es de esperar que, mientras más grande sea la entrada al algoritmo, mayor será el número de instrucciones por ejecutar.

Respecto al algoritmo de las torres de Hanoi, se verá más adelante cómo demostrar $2^n - 1$. Por el momento, se comprobará con algunos ejemplos:

- $n = 1$, la fórmula da exactamente $2^1 - 1 = 2 - 1 = 1$. Es decir que en un solo movimiento se traslada el único anillo de una torre a la siguiente en el sentido de las manecillas del reloj, ejecutando la instrucción 1.a. Nótese que esto es precisamente lo que se logra después de ejecutar un movimiento, A => B, en la figura anterior.
- Para el caso de $n = 2$, se tiene que $2^2 - 1 = 4 - 1 = 3$ y, en efecto, en la misma figura se observa que después de tres movimientos se pueden mover dos anillos de la torre A a la C.
- Otro caso que se ilustra en la figura es el de $n = 3$, donde $2^3 - 1 = 8 - 1 = 7$. Obsérvese que después de siete movimientos se ha logrado mover tres anillos de la torre A a la B.

Posteriormente se verá cómo demostrar esta fórmula en general. De hecho, es posible demostrar que el algoritmo es óptimo en cuanto a su tiempo de ejecución. No hay manera de mover n anillos de una torre a otra usando menos de $2^n - 1$ movimientos.

Curiosidades
Los errores de software pueden ocasionar enormes pérdidas de dinero e incluso de vidas humanas. Por ejemplo, un error en uno de los algoritmos de división del procesador Intel Pentium ocasionó pérdidas a la compañía por más de 400 000 000 de dólares, pues tuvo que reemplazar el producto a sus clientes.

Versión recursiva del algoritmo: corrección, complejidad y optimalidad

Una vez diseñado un algoritmo se desea saber qué tan bueno es y, sobre todo, asegurarnos de que resuelve el problema correctamente.

¿Es correcto el algoritmo? El algoritmo de las torres de Hanoi es un ejemplo de lo difícil que puede resultar entender por qué resuelve el problema. Incluso tratándose de un algoritmo aparentemente tan simple como éste. La clave para entender por qué el algoritmo funciona correctamente radica en percatarse de que resuelve el mismo problema en casos más pequeños (valores de n menores), una y otra vez.

- Cada vez que se mueve un anillo por primera vez, se hace porque hay una torre vacía. Este anillo obviamente se encuentra en la torre A. Si no hubiera una torre vacía, en las

otras dos torres habría anillos más pequeños y el algoritmo indicaría mover alguno de ellos, en lugar de éste: el menor, de acuerdo con la instrucción 1.a, o el otro de acuerdo con la 1.b.

- Cuando se tiene una torre vacía como la del inciso anterior, se pueden presentar tres posibilidades: estar al inicio o al final del algoritmo con dos torres vacías; o bien, con sólo una torre vacía. En este caso, en la torre A se encuentran los anillos que nunca se han movido y en otra torre el resto, por ejemplo *m* anillos, que están apilados necesariamente en orden de mayor a menor. Es decir, el algoritmo resuelve el problema de las torres de Hanoi para *m* anillos.

Considérese la penúltima situación en la figura anterior, donde se tiene el anillo más grande en la torre A, que nunca se ha movido y los tres restantes en la torre B. Esto es, se ha logrado resolver el problema para tres anillos. El algoritmo indica ejecutar la instrucción 1.b y mover el único anillo más grande a la torre C para llegar a la última configuración de la figura. Nótese que este momento equivale a comenzar el algoritmo desde el principio, sin este anillo mayor, puesto que nunca se moverá. Es decir, el algoritmo moverá los tres anillos de la torre B a la C, como si el anillo mayor no existiera. Se procederá a revisar esto con más detalle a continuación.

Recurrencia: un paradigma de diseño de algoritmos

Muchas veces no se comprende bien una situación difícil o un problema agobiante, porque no se le observa desde la perspectiva adecuada. Para comprender mejor el algoritmo de las torres de Hanoi conviene seguir una línea de razonamiento y expresar de manera explícita cómo se resuelven problemas grandes, basados en la solución de problemas pequeños.

Supóngase que Arcadio recibe, como regalo de Úrsula, un juego de torres de Hanoi con cuatro anillos. Arcadio decide que para resolver este acertijo lo mejor es pedir la ayuda de su amigo Luis, gran aficionado a los acertijos, a quien le pide que resuelva un acertijo con sólo tres anillos, y que utilice las soluciones resultantes para resolver el acertijo con los cuatro anillos.

Arcadio muestra a Luis sus torres de Hanoi, pero con sólo los tres anillos más pequeños, y esconde el anillo más grande. Después de explicarle las reglas de los movimientos de los anillos, lo reta a que resuelva el problema de moverlos de una torre a otra. Le explica:

> Pongo los anillos en una torre, digamos la torre *X*, y te indico a qué torre los debes mover; por ejemplo, a la torre *Y*. Tienes que anotar en un papel la lista de movimientos que hay que hacer.

Arcadio le dice a Luis que es suficiente con que la lista de movimientos indique de qué torre a qué torre se realiza cada movimiento; es decir, cada elemento de la lista es de la forma *X* => *Y*. Para evitar confusiones o trampas, Arcadio decide mostrar un papel a Luis donde indica el problema a resolver:

```
Resuelve TH (3, Fuente, Libre, Destino)
```

Fuente, Libre y Destino toman alguno de los valores A, B, C. Al recibir el papel, Luis entiende que tiene que regresarlo a Arcadio con la serie de movimientos que hace falta

hacer para mover tres anillos de la torre Fuente a la torre Destino auxiliándose de la torre Libre.

Por ejemplo, Arcadio escribe:

```
Resuelve TH (3, C, A, B)
```

y Luis le regresa el papel con la siguiente anotación:

C = > B, C = > A, B = > A, C = > B, A = > C, A = > B, C = > B

Arcadio entrega a Luis varios papeles con cambios en la torre de inicio y de destino, y comprueba con satisfacción que Luis siempre responde con rapidez y precisión.

Arcadio resuelve fácilmente el acertijo para cuatro torres. Si desea moverlas de la torre A a la C, simplemente le pide a Luis:

```
Resuelve TH (3, A, C, B)
```

ejecuta los movimientos con los tres anillos menores y ejecuta:

A = > C

Luego pide a Luis:

```
Resuelve TH (3, B, A, C)
```

No hay como pedir ayuda para resolver un problema. Y Arcadio, muy feliz, se percata de que, al saber cómo resolver el acertijo para tres anillos, lo puede aplicar para cuatro, moviendo los anillos de una torre a cualquier otra y contestando cualquier pregunta de la forma:

```
Resuelve TH (4, X, Y, Z)
```

con los movimientos indicados:

```
Resuelve TH (3, X, Z, Y)
```

$X = > Z$

```
Resuelve TH (3, Y, X, Z)
```

Ahora, Arcadio claramente puede resolver el problema para cinco anillos sin ayuda de nadie más, ya que conoce el procedimiento para resolver los acertijos de cuatro anillos. En particular:

```
Resuelve TH (5, X, Y, Z)
```

para lo cual necesita ejecutar los movimientos indicados:

```
Resuelve TH (4, X, Z, Y)
```

$$X => Z$$

```
Resuelve TH (4, Y, X, Z)
```

Arcadio se percata de que no era necesario llamar a una eminencia en descifrar acertijos como Luis; hubiera sido suficiente pedirle a su hermanito de dos años que resolviera:

```
Resuelve TH (1, X, Y, Z)
```

es decir, que moviera un solo anillo de la torre *X* a la *Z*. Al utilizar la solución para este acertijo, resuelve el problema para dos anillos. Más aún, Arcadio podría recurrir a la ayuda de su sobrino de un año de edad. ¿Qué debe contestar el pequeño cuando le pida que resuelva lo siguiente?

```
Resuelve TH (0, X, Y, Z)
```

¡Nada! Claro, la lista de movimientos necesarios para resolver el acertijo con cero discos es una lista vacía, sin ningún movimiento.

En general, para cualquier número de anillos *n* se tiene el siguiente algoritmo recursivo:

```
Resuelve TH (N, Fuente, Libre, Destino)
```

Si N es 0…

Ningún movimiento

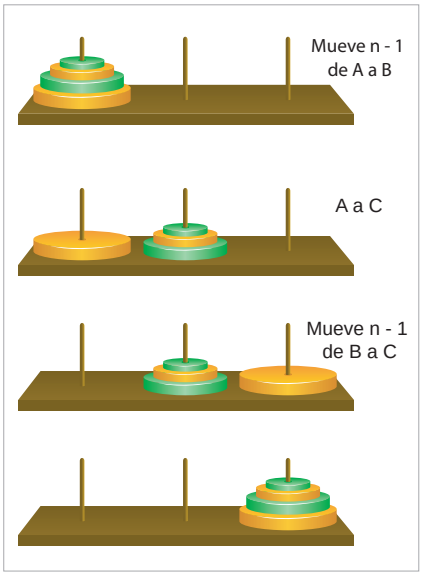
De otro modo,

```
Resuelve (N - 1, Fuente, Destino, Libre)

        Mueve Fuente = > Destino

Resuelve (N - 1, Libre, Fuente, Destino)
```

cuya operación se representa en la siguiente figura adjunta.



Corrección del algoritmo recursivo

Si se usa una simple inducción, se puede demostrar que el algoritmo recursivo de las torres de Hanoi es correcto. Primero se prueba el caso base de $n = 0$ anillos, donde, obviamente, se puede resolver sin ejecutar ningún movimiento. O el caso de $n = 1$ anillo, donde el algoritmo indica:

- Resolver el caso de 0 anillos y como resultado no hacer movimientos.
- Mover el anillo de la fuente al destino.
- Resolver el caso de 0 anillos y como resultado no hacer movimientos.

Concepto
Un algoritmo recursivo se define en términos de sí mismo. Dice algo así como: "para resolver el problema de tamaño *n* utiliza soluciones al problema de menor tamaño". De manera similar, una definición recursiva es aquella que está dada en términos de sí misma. Dice cómo obtener conceptos nuevos, usando el mismo concepto que desea describir. Al igual que con los algoritmos, para que no sea una definición circular, es necesario que esté planteada en términos de una versión más pequeña de ella misma.

Posteriormente, se demuestra el caso general. Supóngase que el amigo al que se le pide que devuelva la lista de instrucciones para resolver los casos de $n - 1$ anillos lo hace siempre correctamente. Es claro, entonces, que si se utilizan las soluciones de casos menores, se resolverá el problema como lo muestra la figura anterior.

Complejidad del algoritmo recursivo

La forma recursiva del algoritmo facilita el análisis de su complejidad. Denotando como Mov(n) al número de movimientos que ejecuta el algoritmo, se ha visto que

$$\text{Mov}(0) = 0$$

y en general, el algoritmo recursivo indica que para resolver el caso de n anillos se resuelve el de n – 1 dos veces y se ejecuta un movimiento entre ambas soluciones:

$$\text{Mov}(n) = 2 \times \text{Mov}(n - 1) + 1$$

Si se emplea la inducción podemos verificar fácilmente que

$$\text{Mov}(n) = 2^n - 1$$

¿Es esto lo mejor posible? Al inicio de este capítulo se mencionó que uno de los aspectos de la algorítmica tiene que ver con la búsqueda del mejor algoritmo posible para resolver un problema. Esta tarea es difícil, ya que para estar seguros de que un algoritmo es el mejor, de alguna manera hay que demostrar que no existe ningún algoritmo que lo supere. ¿Se puede estar seguro de que es imposible que alguien, en un futuro lejano, mediante un método ingenioso y conocimientos avanzados, diseñe un mejor algoritmo? En pocas palabras, hay que asegurarse de que no aparezca un "cisne negro".

Curiosidades
Un "cisne negro" es un evento de alto impacto, difícil de predecir y muy improbable que suceda, de acuerdo con el conocimiento actual. Su nombre proviene de la creencia en Occidente de que todos los cisnes eran blancos y el evento de encontrar uno negro se consideraba imposible antes del siglo XVII, hasta que se descubrió que existían en Australia.

Cisne negro | © *Corbis.*

El algoritmo recursivo para el problema de las torres de Hanoi es óptimo, en el sentido de que es imposible resolver el problema haciendo un número menor de movimientos de anillos. Para demostrarlo una vez más, se utilizará la inducción. Obviamente es óptimo para $n = 0$. Supóngaseque también para $n - 1$ anillos. Es decir, que es imposible mover $n - 1$ anillos de una torre a otra en menos de $\text{Mov}(n - 1)$ movimientos. Ahora se debe probar que para mover n anillos de una torre a otra, necesariamente hay que hacer $2 \times \text{Mov}(n - 1) + 1$ movimientos.

Considérese algún algoritmo hipotético que lo hace, y obsérvese en el momento justo en el que mueve el anillo mayor a su torre de destino final, por ejemplo la C. En ese momento, los $n - 1$ anillos restantes deben estar en la torre B, y por la hipótesis de inducción, para moverlos hacia ahí, se tuvieron que haber hecho al menos $\text{Mov}(n - 1)$ movimientos. De igual forma, ahora es necesario mover estos anillos a la torre C, para lo cual hacen falta otros $\text{Mov}(n - 1)$ movimientos, al menos. Contando el movimiento del anillo mayor, se tiene como resultado que se ejecutaron al menos $2 \times \text{Mov}(n - 1) + 1$ movimientos en total.

Memoria: otra medida de complejidad de un algoritmo

Este algoritmo recursivo para el problema de las torres de Hanoi es muy fácil de entender. Sin embargo, no es muy eficiente en cuanto a la cantidad de memoria que utiliza. Se verá más adelante cómo al ejecutar un programa en una computadora, se requiere que ésta tenga memoria para almacenar tanto el programa mismo como los datos que utiliza durante su ejecución. Por ahora sólo se hace notar que el algoritmo recursivo necesita mucha memoria, ya que cuando Arcadio quiere resolver el problema para n anillos, recurre a un amigo a fin de que anote los movimientos para resolver dos problemas de $n - 1$ anillos; éste a su vez pedirá a otra persona que resuelva dos problemas de $n - 2$ anillos, y así sucesivamente. La memoria a la que se ha hecho referencia son precisamente estas anotaciones. Así que se han visto dos algoritmos que resuelven el problema de las torres de Hanoi: uno iterativo y otro recursivo, ambos igualmente eficientes en cuanto al tiempo, pero el iterativo más eficiente en cuanto a memoria.

En general, la eficiencia de los algoritmos se mide de acuerdo con la cantidad de recursos que se consumen durante su ejecución. Si interesa determinar la complejidad de un algoritmo que debe lograr que A comunique algo a B, probablemente se desearía medirla en términos del número de mensajes que A debe enviar a B para lograrlo; si cada mensaje tiene un costo implícito, buscaríamos un algoritmo que requiriera la mínima cantidad de mensajes.

Si en cambio se estuviese interesado en un algoritmo que permita que un brazo robótico pinte un automóvil, a lo mejor la complejidad estaría dada en términos del número de movimientos del brazo. En fin, para cada algoritmo existen diferentes recursos susceptibles de cuantificarse; seguramente alguno o algunos de ellos serán los más importantes, según el criterio de quien lo analiza. Ciertamente existen dos recursos consumidos siempre por todo algoritmo: tiempo y memoria.

2.4 BÚSQUEDA EXHAUSTIVA

En el tema anterior se analizó el paradigma de recursividad para el diseño de algoritmos y en el capítulo 1 se presentó uno más sencillo: el de búsqueda exhaustiva, que no hace

más que explorar todas las posibles soluciones a un problema, con un costo usualmente enorme. En el caso del problema del regalo de Arcadio, el algoritmo simplemente consideraba una por una todas las combinaciones posibles de obsequios para Úrsula y elegía el mejor. Los algoritmos de búsqueda exhaustiva frecuentemente son de tiempo exponencial. Es necesario ser ingenioso para obtener algoritmos eficientes, pero no hay que olvidar que en ocasiones la búsqueda exhaustiva es la única posibilidad para resolver un problema.

2.4.1 Coloración de gráficas

Considérese el problema de colorear una gráfica con el menor número de colores. Un algoritmo muy sencillo sería simplemente probar todas las coloraciones posibles; sin embargo, esto tomaría tiempo exponencial. Por ejemplo, supóngase que se quieren probar todas las coloraciones posibles con cuatro colores en una gráfica de n vértices. Escójanse un vértice y uno de los cuatro colores para éste. Para el siguiente vértice se tienen otra vez cuatro posibilidades, y así sucesivamente. El número total de posibilidades a explorar es 4^n.

Desafortunadamente, para el problema de coloración, no se sabe si existe un algoritmo mejor que éste; se trata de un problema NP-completo (véase el tema 1). No se sabe siquiera si existe un algoritmo eficiente (polinomial) para decidir si un mapa requiere de los cuatro colores, o si se puede colorear con tres. Sin embargo, ya se ha descrito un algoritmo eficiente para cuando no se exige ser tan rigurosos, y es suficiente colorear un mapa con seis colores.

2.4.2 El problema de ordenamiento

Arcadio en la tienda

Mientras Arcadio elegía el regalo de Úrsula, sentía que la cabeza le daba vueltas de tanto analizar, caso por caso, todas las opciones que se le presentaban. La búsqueda exhaustiva puede, como ya lo estudiamos, encontrar la mejor solución, pero en un tiempo inaceptable para Arcadio, un tiempo exponencial. ¿Sería posible encontrar un regalo que fuera tan parecido al regalo ideal pero que no se notara la diferencia? Una vez decidido a ignorar el análisis caso por caso de todas las opciones, Arcadio asignó a los objetos una prioridad: qué tanto le gustaría a Úrsula cada objeto. Posteriormente, con los objetos "calificados" en esta forma, procedió a ordenarlos de acuerdo con este criterio y a comprarlos en orden de mayor a menor importancia hasta agotar su dinero.

El problema de ordenar una lista de objetos según algún criterio que utiliza Arcadio al final de su aventura comercial, es uno de los problemas más importantes que hay en computación. Muchas de las tareas que ejecutan las computadoras necesitan apoyarse en algoritmos de ordenamiento. Cada uno de los objetos que se quieren ordenar tiene asociado uno o más valores numéricos, como pueden ser el precio, el tamaño, el peso, etcétera, y se requiere ordenarlos apegándose a alguno de estos parámetros. Para el problema algorítmico se vuelven irrelevantes los otros parámetros y, de hecho, cualquier otra característica de ellos también, como podría ser su color. A veces se requiere ordenarlos de mayor a menor y, en otras, al revés; el mismo algoritmo funciona en ambos casos. De este modo, definimos formalmente: en el problema de ordenamiento, la entrada es una lista de números, y se pide que la salida sea una lista con los mismos números, pero

ordenados de menor a mayor. Por ejemplo, si la entrada es [17, 23, 17, 5], la salida debe ser [5, 17, 17, 23].

En busca de todos los ordenamientos

Un algoritmo de búsqueda exhaustiva para este problema sería inaceptable; correría en un tiempo peor que exponencial. Si se cuenta con una lista de tres números por ordenar crecientemente, por ejemplo [23, 17, 5], las posibles maneras de ordenar la lista son $3 \times 2 \times 1 = 6$; hay tres formas diferentes de elegir el primer elemento de la lista. Por cada una de ellas existen dos para el siguiente, puesto que no se puede repetir el primero y, finalmente, por cada una de las elecciones de los primeros dos elementos hay sólo una opción para el último, pues ya sólo queda exactamente uno por incluir.

En este ejemplo, los posibles ordenamientos son: [23, 17, 5], [23, 5, 17], [17, 23, 5], [17, 5, 23], [5, 23, 17] y [5, 17, 23]. Sólo una de estas opciones, la última, de hecho, tiene la lista ordenada de menor a mayor. Se usará n para designar el tamaño de la lista, es decir, el número de elementos de la lista por ordenar. Con $n = 5$, tenemos $5 \times 4 \times 3 \times 2 \times 1 = 120$.

En general, se tiene la función factorial que ya había sido presentada: el número de posibles ordenamientos es $n!$, lo cual es mucho peor que el problema inicial de Arcadio. Vale la pena recordar que esta función crece más rápido que cualquier función exponencial de la forma c^n. El número de opciones de compra para n objetos sólo duplica el de $n - 1$, pero el número de ordenamientos posibles es n veces el de $n - 1$.

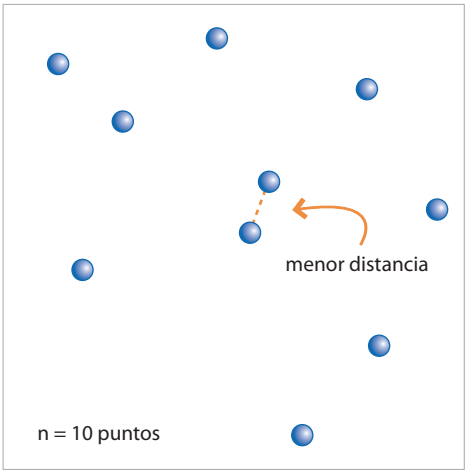
Afortunadamente, hay mejores maneras de resolver el problema, ya que no hace falta analizar todas las opciones, lo que, hasta donde sabemos, es imposible en el caso del problema de determinar los objetos a comprar. Más adelante se verán algunos de los algoritmos de ordenamiento más eficientes.

2.4.3 La pareja de puntos más cercanos

La búsqueda exhaustiva no siempre genera algoritmos exponenciales. A continuación se presentará un ejemplo básico de geometría computacional, un área con diversas aplicaciones de computación, especialmente relacionadas con imágenes y visión.

En el problema de los puntos más cercanos, la entrada es un conjunto de n puntos en el plano, y la salida es la menor distancia entre algún par de puntos.

El método de fuerza bruta para resolver el problema es simplemente considerar uno por uno todas las parejas de puntos; para cada pareja hay que calcular a qué distancia se encuentran, y elegir la menor de las distancias.



Algoritmo PC-exhaustivo
Para cada pareja de puntos **x, y:**
Sea **d** la distancia de **x** a **y**
Si **d** es menor a **d min**, la menor distancia vista hasta ahora,
Sea **d min** igual a **d**.

Para ejecutar este algoritmo, se supone que la máquina sabe calcular la distancia entre dos puntos, comparar dos de estas distancias para obtener la menor y entregar una por una todas las posibles parejas de puntos. En realidad, estas operaciones se pueden implementar fácilmente en cualquier lenguaje de programación.

Se ha visto en el tema de inducción que hay $n(n-1)$ parejas y que el tiempo de ejecución de este algoritmo crece de manera cuadrática conforme se incrementa el tamaño de la entrada. Es decir, mientras que en un algoritmo lineal, como el de coloración de mapas con seis colores, si el número de países del mapa se duplica, el tiempo para resolver el problema también se duplica; en el caso del algoritmo de los puntos más cercanos, el tiempo se cuadruplica. Esto se explicará con mayor detalle en el siguiente tema. Existe otro algoritmo mucho más eficiente que éste.

2.5 DIVIDE Y VENCERÁS

En muchas ocasiones es posible evitar la búsqueda exhaustiva y diseñar algoritmos más eficientes para resolver un problema, sin necesidad de considerar todas las posibles soluciones explícitamente. Una manera de lograrlo es mediante el método "divide y vencerás", que construye una solución poco a poco, cada vez más completa. El paradigma "divide y vencerás" es la clave para diseñar algoritmos muy eficientes para una variedad de problemas. La idea es dividir un problema en dos subproblemas del mismo tipo y tamaño; resolverlos de manera recursiva y combinar las soluciones para obtener el resultado deseado. Para que el paradigma resulte en un algoritmo eficiente es necesario que el procedimiento de combinar las soluciones con los subproblemas sea veloz. Aunque es habitual describir este tipo de algoritmos recursivamente, también se puede hacer de forma iterativa.

2.5.1 Ordenamiento por inserción

Para el caso del problema de ordenamiento, son sorprendentes los números tan grandes que se obtuvieron con una tarea aparentemente tan sencilla como ordenar una lista. Las amas de casa ordenan sus listas de compras cuando van al supermercado para que les sea más fácil encontrar los objetos; los carteros ordenan la correspondencia diaria para entregar cientos de cartas y paquetes; las bibliotecas mantienen ordenados sus libros; los profesores ordenan las listas de calificaciones de alumnos por sus apellidos; sería casi imposible encontrar un número telefónico en el directorio si los nombres no estuvieran ordenados.

Si se tiene una lista con n números enteros y se desea ordenarla crecientemente (de menor a mayor), no se hace una búsqueda sobre el conjunto de todos los posibles ordenamientos; es posible hacerlo de muchas otras maneras más eficientes. Cuando juegas cartas con tus amigos o tu familia, ¿cómo se podría ordenar un juego? Bueno, pues hay distintas maneras de hacerlo.

Supóngase que alguien reparte las cartas, una por una. Cada vez que te da una carta, la insertas en el lugar correcto entre las que ya tienes en la mano. Este algoritmo se conoce como *ordenamiento por inserción* y, aunque en general no es el más rápido de los ordenamientos, se puede utilizar el mismo algoritmo para ordenar un paquete de cartas de cualquier tamaño, incluso si no te dan una por una.

En términos abstractos, al trabajar con una lista de números en lugar de cartas, cada iteración del ordenamiento por inserción quita un elemento de la parte desordenada de la lista, el i-ésimo elemento, y lo inserta en la posición correcta en la parte ordenada de la lista. Repetir esto, con el $i + 1$-ésimo elemento, y así sucesivamente, hasta que no haya más elementos en la parte desordenada de la lista.

Ordenamiento por inserción
Entrada: lista L de n números
Salida: lista ordenada
Para i desde 1 hasta n, ejecutar:
Tomar el i-ésimo elemento de L, por ejemplo x
Insertar x en su lugar correcto dentro de las primeras i posiciones de L.

Queda definir el algoritmo para insertar un elemento en una lista ordenada. Es decir, se ha diseñado el algoritmo usando la noción de **subrutina**, y siguiendo una estrategia muy útil llamada de "arriba abajo" (que consiste en ir resolviendo un problema poco a poco, primero las dificultades generales, dejando algunas tareas más particulares para ser resueltas posteriormente). De esta manera, se resuelve el problema paulatinamente y damos libertad a la implementación de la subrutina. Al algoritmo de ordenamiento no le interesa de qué manera inserta el elemento. Con tal de que lo haga correctamente, puede buscar su posición de derecha a izquierda en la lista ordenada, esto es, de mayor a menor, o viceversa. Lo usual es que lo haga de derecha a izquierda.

Concepto
A la porción de un algoritmo que realiza una tarea específica y es relativamente independiente del resto del algoritmo, se le denomina subrutina o, dependiendo del lenguaje de programación y otras sutilezas, procedimiento, función o método. El uso de subrutinas dentro de un algoritmo o programa tiene muchas ventajas, como:

- reducir la duplicación de secciones que hacen la misma tarea;
- reutilizar las subrutinas en otros algoritmos;
- descomponer algoritmos complejos en partes más simples, y
- facilitar la comprensión y análisis del algoritmo.

Corrección

Para determinar que el algoritmo de ordenamiento es correcto, el primer paso será analizarlo por inducción, suponiendo que la subrutina de inserción funciona correctamente. Se presume que después de i iteraciones, ha ordenado de manera correcta los primeros i elementos de la lista, y se observa que después de una iteración, ordena correctamente los primeros $i + 1$ elementos. Ahora, se puede analizar de manera similar la subrutina y por inducción probar que funciona. ¿Cuál sería la hipótesis de inducción en cada caso?

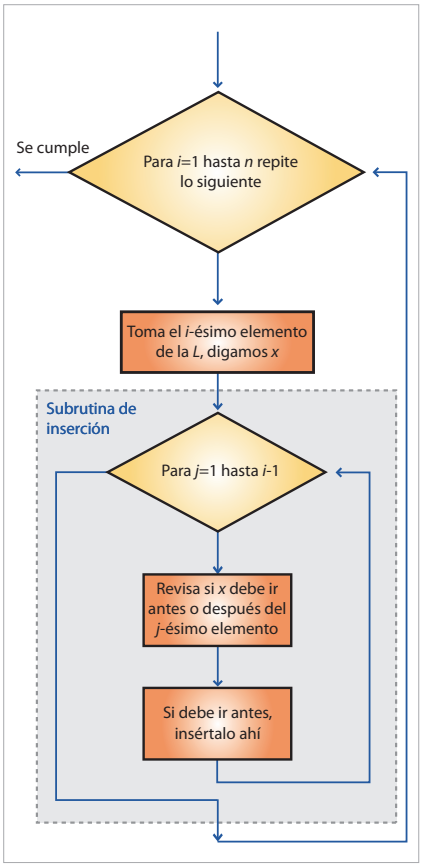
Complejidad

Para analizar el tiempo de ejecución del algoritmo es útil ver su diagrama de flujo, que incluye dos iteraciones anidadas: se repite la exterior n veces y, por cada una, la interna se repite varias veces, lo cual usualmente da lugar a algoritmos de tiempo cuadrático.

Obsérvese que ejecuta n iteraciones, y que en la primera debe insertar el primer elemento en la parte ordenada de L, que al inicio está vacía. En la segunda iteración debe insertar el segundo elemento en la parte ordenada de L, que ahora tiene un elemento. En general, en la i-ésima iteración debe insertar el i-ésimo elemento de L, por ejemplo x, en la parte ordenada de L, que ahora consiste de los primeros $i - 1$ elementos. Para hacerlo, debe revisarlos y comparar cada uno de ellos con x. Es decir, en la i-ésima iteración se pudieron haber realizado $i - 1$ comparaciones. Llega al último elemento, el n-ésimo, que inserta en una lista de $n - 1$ elementos, para obtener un total de:

$$1 + 2 + 3 + \ldots + n - 1$$

Concepto
Como se deduce de las actividades anteriores, el algoritmo de ordenamiento por inserción tiene un tiempo de ejecución que crece de manera cuadrática con respecto al tamaño de su entrada, n, en el peor caso (como cuando el orden de la lista está totalmente invertido). Sin embargo, para ciertas entradas (cuando la lista está ordenada), su tiempo de ejecución es lineal. En general, al analizar un algoritmo se considera su tiempo de ejecución en el peor caso, ya que éste muestra el tiempo máximo que emplea el algoritmo en resolver el problema tomando en cuenta todas sus entradas posibles (en algunas le tomará más tiempo y en otras menos).



comparaciones, en el peor caso (sobre todas las posibles listas de entrada). Y ya se vio que esta suma es igual a $n(n-1)/2$. Por ejemplo, si la lista original tiene dos elementos, con una comparación basta para saber si se dejan así o se invierten.

El **algoritmo de ordenamiento por inserción** es mucho mejor que buscar entre todos los $n!$ ordenamientos posibles. Sin embargo, existen algoritmos aún mejores, como se podrá ver más adelante. No obstante, este algoritmo funciona bien para listas pequeñas; conforme crece la lista, el tiempo de ejecución se incrementa rápidamente, y es necesario utilizar algoritmos más eficientes.

La característica de *estabilidad* de un algoritmo de ordenamiento permite ordenar, por ejemplo, una lista de alumnos con sus calificaciones: primero por calificación y luego por orden alfabético. Generalmente, permite ordenar listas de elementos con más de un criterio.

2.5.2 Ordenamiento de burbuja

Concepto
Cuando el tiempo empleado para resolver un problema depende de alguna potencia del tamaño de la entrada, como n^2 o n^3, se dice que el problema es de complejidad polinomial. Al conjunto de todos los problemas de complejidad polinomial los computólogos lo bautizaron como P.

Otro algoritmo de ordenación importante es el que se conoce como *ordenamiento de burbuja*, el cual recibe este nombre porque el efecto del algoritmo simula el movimiento de una burbuja hacia la superficie, sólo que aquí no es una burbuja sino el número más pequeño y no flota hacia la superficie, sino a la cabeza de la lista. Funciona así:

Repetidamente recorre la lista y compara cada posible pareja de números; si están en el orden equivocado, se intercambian. Cuando recorre la lista y no hace ningún intercambio, se detiene porque la lista está ordenada. ¿Sorprendente, verdad? Vale la pena intentarlo con la lista del ejemplo anterior: 31 25 12 22 11. Va la secuencia de operaciones:

31 25 12 22 11, se comparan 31 y 25; se intercambian
25 **31** 12 22 11, ahora 31 y 12; se intercambian
25 12 **31** 22 11, 31 y 22; se intercambian
25 12 22 **31** 11, 31 y 11; se intercambian
25 12 22 11 **31**…
12 **25** 22 11 31
12 22 **25** 11 31
12 22 11 **25** 31; aquí se comparan 25 y 31, pero no hay cambio
12 22 11 **25** 31
… etcétera.

El ordenamiento de burbuja también tiene complejidad cuadrática. Aquí cabe una pausa para reflexionar sobre lo que se ha visto respecto a los algoritmos de ordenamiento.

Los algoritmos que se revisaron tienen similitudes: son fáciles y, debido a su complejidad, del orden de n^2 es fácil percatarse de que son mejores en comparación con el viejo conocido $n!$; obsérvese que si empleamos un segundo por comparación, ordenar una lista de 20 elementos nos toma casi siete minutos y no cinco veces la edad del universo, como en el caso del factorial. Pero se puede lograr algo aún mejor.

2.5.3 Búsqueda binaria

Aquí presentamos un algoritmo de "divide y vencerás" sencillo, pero muy útil. De hecho, en este caso no hace falta el procedimiento de combinar las soluciones con los dos subproblemas, ya que sólo una de ellas aparece.

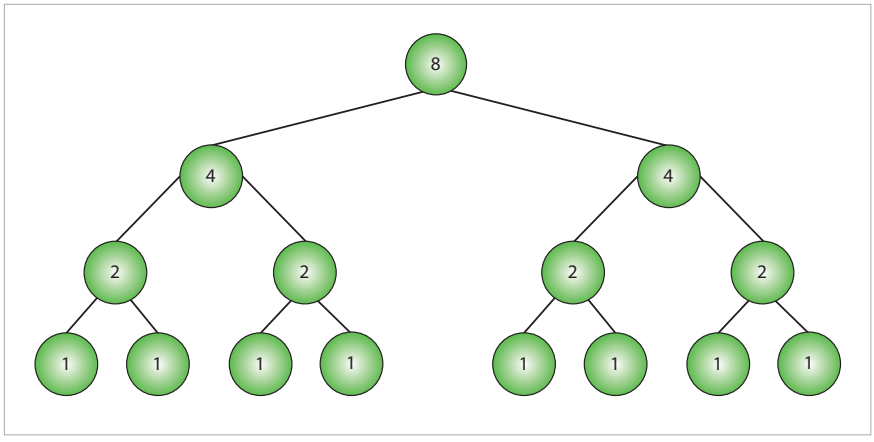
Si un directorio telefónico, como el de la Ciudad de México, que puede tener cientos de miles de números telefónicos, no estuviera ordenado alfabéticamente, no se tendría otra opción que buscar el número de una persona, nombre por nombre. Si tuviera 100 000 nombres, y tomara un segundo revisar cada uno, un usuario tardaría casi 30 horas en revisar todos los nombres. Afortunadamente, el directorio sí esta ordenado, y para buscar un nombre sólo se tiene que, por ejemplo, abrir el directorio a la mitad, y esa mitad dividirla en otra, guiándose por el orden alfabético de nombres.

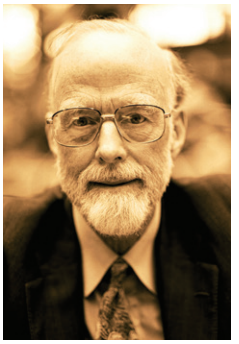
¿Cuánto tiempo le toma a una persona buscar el nombre de otra en el directorio? Tantas veces como se tenga que dividir 100 000 entre 2 una y otra vez, hasta llegar a un número muy pequeño. A esta operación se le conoce como logaritmo en base 2 de 100 000. En este caso, es suficiente hacer 17 comparaciones, en lugar de emplear 30 horas con 15 segundos.

El algoritmo de búsqueda exhaustiva recorre la lista de principio a fin, y su tiempo de ejecución es proporcional al tamaño de la lista, n. El paradigma "divide y vencerás" conduce al algoritmo de búsqueda binaria descrito arriba. El tiempo de ejecución de éste es proporcional al logaritmo en base 2 de n; como lo muestra la siguiente gráfica, llamada árbol binario, es igual al número de veces que hay que dividir n entre 2 para llegar a 1.

El problema de búsqueda ordenada tiene como entrada una lista de n números, ordenados de menor a mayor, y un número x. La salida es la posición de x en la lista, o es un "no", en caso de que x no se encuentre en la lista.

En la figura se presenta el caso de una lista de ocho números, donde se observa que el logaritmo en base 2 de 8 es igual a 3, representado por el número de aristas que hay que recorrer para llegar de la raíz, el vértice etiquetado con 8, hasta una hoja, con valor 1.





C.A.R. Hoare | © Frank Fremerey.

C. A. R. Hoare (1934)
Sir Charles Antony Richard Hoare recibió el Premio Turing en 1980 por sus contribuciones a los lenguajes de programación. Además, inventó QuickSort en 1960, que es el algoritmo de ordenamiento más utilizado en el mundo. Mientras era estudiante de intercambio en la Universidad Estatal de Moscú, le ofrecieron un puesto en el Laboratorio de Física Nacional en Teddington, para trabajar en un proyecto cuyo objetivo era traducir mecánicamente de ruso a inglés. Había la necesidad de ordenar todas las palabras en una oración, antes de poder localizarlas. Su primer intento le tomaba tiempo proporcional al cuadrado de la longitud de la oración, el segundo intento fue QuickSort. Sin embargo, tiempo después Hoare declinó el trabajo, porque la traducción mecánica de lenguajes naturales le pareció impráctica.

2.5.4 Ordenamiento por combinación

Si Arcadio necesita ordenar una lista de n números, no hay nada más sencillo que pedir ayuda a sus amigos. Pide a Úrsula que ordene una mitad de la lista, mientras que Luis ordena la otra. Cada uno le devuelve un papel con la parte de la lista ordenada y Arcadio, fácilmente, reúne ambas en una sola lista. Arcadio obtiene algoritmos diferentes, dependiendo de cómo decida dividir la lista. ¿Qué sucede si simplemente parte la lista a la mitad? Se obtiene el famoso algoritmo de ordenamiento por combinación (*mergesort* en inglés).

```
OrdenaM (L)
Si longitud de L es 1
Regresa L
De otro modo
Divide L en 2 del mismo tamaño (aproximadamente), L1 y L2
OrdenaM(L1)
OrdenaM(L2)
Mezcla(L1,L2)
```

Para finalizar, se describirá la subrutina de mezclado. El problema de mezclado consiste en combinar en una sola lista dos listas de números, cada una ordenada de la misma manera, de menor a mayor.

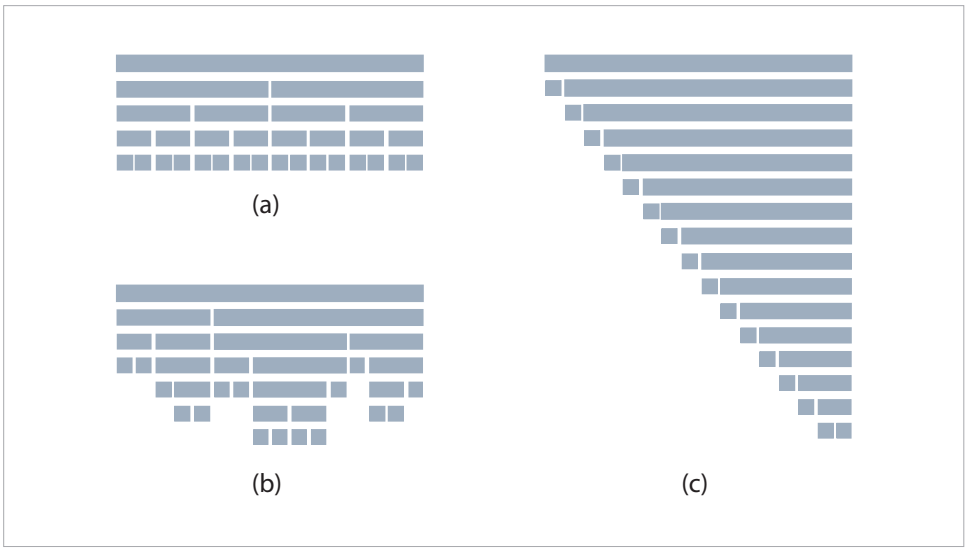
El tiempo de ejecución del algoritmo de ordenamiento por combinación, OrdenaM, es proporcional a n multiplicado por el logaritmo en base 2 de n, lo cual se denota como $n \log n$. Esto es debido a que se tiene un árbol binario que representa el número de veces que se divide la lista en 2. Como se mencionó anteriormente, la altura del árbol es $\log n$ y el número de comparaciones por nivel es n, ejecutadas por el procedimiento de mezclado.

2.5.5 Ordenamiento rápido

Es probable que el algoritmo de ordenamiento más utilizado en la práctica sea el de ordenamiento rápido (*quicksort*, en inglés), que tiene complejidad, en el peor de los casos, proporcional a n^2; sin embargo, la complejidad del caso promedio es de $n \log n$. Los pasos de este algoritmo son:

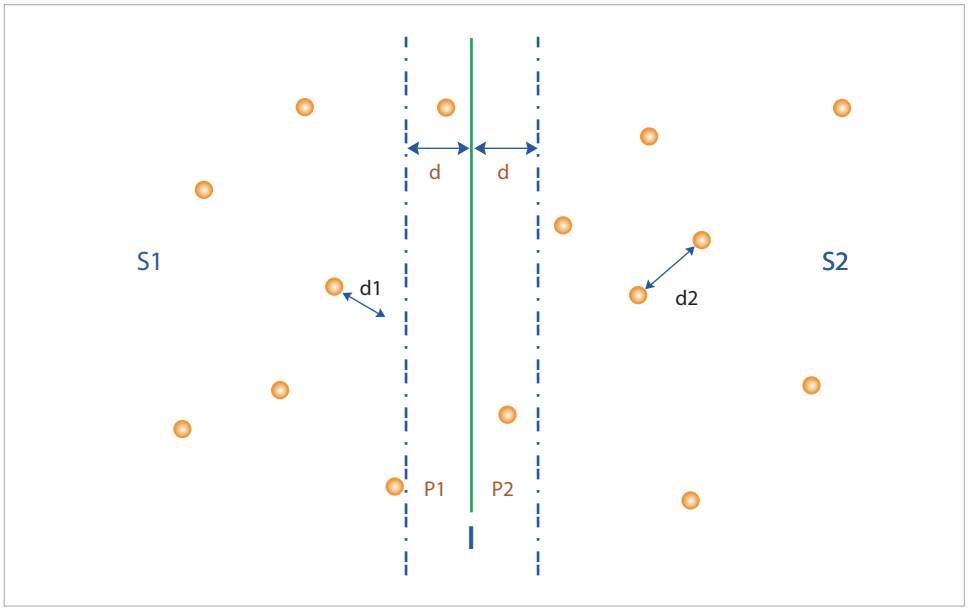
1] Seleccionar un elemento de la lista, al que se llamará "pivote".
2] Reordenar la lista para que todos los números menores que el pivote queden del lado izquierdo y los mayores del derecho. El pivote se encuentra en su posición final, que se denomina "la partición".
3] Ordenar la lista de elementos menores y la de elementos mayores, de manera recursiva.

El ordenamiento rápido es un ejemplo de algoritmo no determinista, ya que deja libertad en cuanto a la elección del pivote. En el mejor caso, en cada iteración la lista se divide a la mitad y la recursión se repite $\log n$ veces. Como en cada nivel hay que trabajar con n elementos para dividirlos de acuerdo con el pivote, el tiempo total es $n \log n$. Pero con mala suerte, el pivote podría ser el menor de la lista, y en cada llamada recursiva dividirla en dos partes, una con un solo elemento y otra con todos, en cuyo caso el tiempo de ejecución sería proporcional a n^2.



2.5.6 Parejas de puntos

Se vio un algoritmo de búsqueda exhaustiva para resolver el problema de encontrar una pareja de puntos lo más cercana posible. Dados n puntos, el algoritmo hace n^2 comparaciones. Mediante el paradigma "divide y vencerás" es posible reducir el tiempo a $n \log n$. Sin embargo, la solución no es obvia, por lo que sólo se describirá brevemente. La idea es dividir en dos conjuntos $S1$ y $S2$, usando una línea l, al conjunto de puntos de entrada. Después, calcular la distancia mínima en $S1$ y $S2$ recursivamente. Supóngase que estas distancias son $d1$ y $d2$, respectivamente.



Algunos puntos y sus distancias.

Ahora la dificultad radica en que podría haber dos puntos, de lados opuestos de la línea l, que estén muy cerca, a una distancia d, menor a ambas, $d1$ y $d2$. Pero este caso se

puede resolver rápidamente si se utiliza un argumento geométrico que explica por qué sólo puede haber seis puntos consecutivos en la franja problemática de ancho $2d$, indicada en la figura 2.

2.6 ÓRDENES DE CRECIMIENTO

Anteriormente se mostraron varios ejemplos de algoritmos de distintas complejidades. Se vieron dos algoritmos para el problema de búsqueda ordenada: el secuencial (de tiempo lineal), y el binario (de tiempo logarítmico). Asimismo, se revisaron algunos algoritmos de tiempo cuadrático, como el exhaustivo empleado para encontrar la pareja de puntos más cercana. Y finalmente se habló en distintas ocasiones sobre algoritmos de tiempo exponencial. ¿Qué significan exactamente estos términos? ¿Por qué son tan importantes?

El algoritmo de búsqueda lineal simplemente recorre la lista de entrada de n elementos en busca de un elemento dado. Se dice que su tiempo de ejecución es lineal en el tamaño de la lista porque, en el peor caso, debe analizar n elementos. Esto implica que, independientemente de cuáles sean con exactitud las operaciones que ejecuta sobre cada elemento, conforme crece el tamaño de la lista de entrada n, el tiempo de ejecución del algoritmo se incrementa linealmente. Si en una ejecución con una lista de n elementos el tiempo es T, con una lista de $2n$ elementos, el algoritmo emplea $2T$ unidades de tiempo para terminar la ejecución en la máquina. Si se corre en una máquina cuatro veces más rápida, la búsqueda en una lista de n elementos tomará $T/4$. Sin embargo, al duplicar el tamaño de la lista, otra vez se duplica el tiempo de ejecución a $T/2$.

En resumen, aunque no es posible decir cuál es el tiempo de ejecución de un algoritmo para una entrada de cierto tamaño, ya que éste depende de una variedad de factores como la computadora elegida para correr el algoritmo y la programación del algoritmo en un cierto lenguaje, entre otros, sí se puede estimar cómo se incrementa el tiempo de ejecución conforme el tamaño de la entrada crece. Es posible definir formalmente esto como *orden de crecimiento* del tiempo de ejecución, o en general de cualquier función $f(n)$.

En el caso de un algoritmo lineal, se afirma que su tiempo de ejecución es orden de n, y se escribe $O(n)$. Esto significa que para el computólogo, $2n$ o $28n$ equivalen a lo mismo, ya que ambas funciones crecen a la misma velocidad. Es decir, se agrupan todas las funciones lineales en una sola clase, la llamada clase de funciones lineales, denotada $O(n)$. Estas funciones no sólo incluyen cualquiera de la forma cn, para una constante c, sino también funciones como $4n + 7$; esto es, de la forma $cn + k$, lo cual implica que tampoco afecta el orden de crecimiento sumar una constante; quizá sólo empezar a correr el programa de búsqueda lineal suma un cierto tiempo, k, el mismo para cualquier entrada posible, pero esta k no afecta el orden de crecimiento.

Además de indicar cómo se comportará el algoritmo cuando trabaje con problemas más grandes, también ofrece una buena indicación de qué tan eficientemente se aprovecha la tecnología de cómputo. Si el algoritmo lineal emplea 60 segundos en buscar en una lista de 100 elementos en una computadora que corre a cierta velocidad, supóngase que le toma x unidades de tiempo ejecutar cada instrucción. Si se hace una inversión para que resuelva el problema en la mitad de tiempo, 30 segundos, con una computadora del doble de velocidad, $x/2$ unidades de tiempo por instrucción, lograremos esto.

Por otro lado, un algoritmo de orden exponencial, $O(2^n)$, por ejemplo, es el de la búsqueda exhaustiva que realiza Arcadio descrita antes, y para encontrar todos los ordenamientos que vimos arriba, el orden es $O = O(n!)$. En estos casos, duplicar la velocidad

de la computadora no ayuda prácticamente en nada; sólo se podrá buscar en listas de unos cuantos elementos más. Por ejemplo, si el algoritmo es exponencial y un minuto alcanzaba para buscar en una lista de 1 000 elementos, con una computadora del doble de velocidad, quizá lograría buscar en una lista de 1 002 elementos.

Si el algoritmo de la búsqueda binaria, cuyo tiempo de ejecución es $O(\log n)$, empleara, por ejemplo, 10 segundos en buscar en una lista de 1 000 elementos, para duplicar su tiempo de ejecución se tendría que llegar a una lista de hasta un millón de elementos. Este tipo de comportamiento establece una diferencia radical entre un algoritmo de tiempo lineal y uno logarítmico, independientemente de la computadora en la que se corran. De hecho, la diferencia es tan radical como cuando se pasa de un algoritmo polinomial a uno exponencial. Además, nótese que ya no hace falta especificar la base del logaritmo, ya que la diferencia entre una base y otra es sólo una constante.

La notación asintótica es aquella que expresa el orden de crecimiento de una función, sin tomar en cuenta cómo se comporta al inicio ni las constantes multiplicativas. Un ejemplo es la "O grande" descrita arriba. Es útil para simplificar las funciones analizadas, ignorando detalles como términos de crecimiento más rápido. Es decir, en la práctica, la notación permite afirmar que la función que analizamos crece como otra función más sencilla.

Richard Karp | © Richard Karp.

Richard Karp (1935) ganó el Premio Turing en 1985 por sus contribuciones a la teoría de algoritmos, y por identificar al crecimiento polinomial como lo deseable de un algoritmo práctico, además de desarrollar la metodología para probar que un problema es NP-completo.

Notación	Nombre	Ejemplo
$O(1)$	Constante	Determinar si un número es par o impar.
$O(\log n)$	Logarítmica	Realizar una búsqueda binaria.
$O(n)$	Lineal	Buscar un elemento en una lista desordenada.
$O(n \log n)$	Cuasilineal	Ordenar una lista lo más rápidamente posible.
$O(n^2)$	Cuadrática	Ordenar una lista con ordenamiento de inserción.
$O(n^c)$, $c > 1$	Polinomial	Multiplicación de matrices.
$O(c^n)$	Exponencial	Realizar los movimientos de las torres de Hanoi.
$O(n!)$	Factorial	Determinar si dos predicados lógicos son equivalentes.
$O(2^{c^n})$	Doble exponencial	Decidir si un formalismo lógico es falso o verdadero (véase en el módulo 1 la sección "problemas peores que exponenciales").

Tabla 1. Órdenes de crecimiento de funciones comunes.

Por ejemplo, se vio un algoritmo de ordenamiento con tiempo de ejecución $(n-1)n/2 = (n^2 - n)/2$, lo cual es simplemente $O(n^2)$, ya que la función anterior no crece más rápido que n^2. Se presentó un algoritmo para las torres de Hanoi, cuyo número de movimientos era $2^n - 1$.

Otros ejemplos de crecimiento exponencial pueden tener funciones de crecimiento del estilo: $2^n + 2^n$, $2^n + 356$ y $2^n + 3n^3 + 2^n$; empero, todas ellas tienen $O(n) = 2^n$, porque si la n es lo suficientemente grande, la aportación de los demás términos no afecta la velocidad de crecimiento. Sin embargo, es importante mencionar que 3^n crece estrictamente más rápido que 2^n. Esto es, aunque a cualquier función de la forma c^n se le llame exponencial, no todas tienen el mismo orden de crecimiento. Y lo mismo para $n!$, que crece más rápido que cualquier función exponencial.

De la misma manera, las funciones de crecimiento polinomial, $O(n^c)$, incluyen, por ejemplo, a n^2 y a n^7, aunque tienen distinto orden de crecimiento. En efecto, n crece me-

nos rápido que n^2, y ésta menos rápido que n^3. No obstante, $n \log n$ crece menos rápido que n^2, y aquí radica la complejidad de los mejores algoritmos para ordenar una lista.

¿Cuál sería el orden de crecimiento del tiempo de un algoritmo que no crece? Por ejemplo, para decidir si un número es par o impar, basta con examinar su dígito menos significativo, independientemente del tamaño del número. Es decir, el tiempo crece a la misma velocidad que una función constante (que no crece), como $f(n) = 23$, o simplemente $f(n) = 1$. Es decir, estas funciones son $O(1)$.

Entonces, ¿cuál es el orden de crecimiento más rápido? Pues, no existe tal. Para cualquier orden de crecimiento, por ejemplo $f(n)$, siempre hay uno aún más rápido. Por ejemplo $2^{f(n)}$. Todavía más sorprendente es que existen problemas tan difíciles que requieren de ese tiempo de ejecución.

En la tabla anterior se listan algunos de los órdenes de crecimiento que aparecen con frecuencia en el análisis de algoritmos, así como ejemplos de algoritmos cuyas funciones presentan ese orden de crecimiento.

2.7 RESUMEN

Detrás de todas las maravillas que hacen las computadoras están los algoritmos. En este capítulo se presentó una introducción a la disciplina que estudia estos métodos: la algorítmica, la cual ha sido relevante para la mayor parte de las ciencias y las humanidades. Se ha hecho hincapié en el diseño de algoritmos correctos y eficientes. Se explicó la noción de algoritmo, cómo analizar su eficiencia y cómo convencerse de que resuelve correctamente un problema. Se mostraron, además, algunas de las técnicas para resolver problemas, como "divide y vencerás", recursividad y exploración exhaustiva, al igual que ejemplos de algoritmos clásicos.

PROGRAMACIÓN

TEMA

3

© Sergi Larripa.

Un enfoque científico se caracteriza comúnmente por las palabras lógico, sistemático, impersonal, calmado, racional; mientras que un enfoque artístico se caracteriza por las palabras estético, creativo, humanitario, inquieto, irracional. Me parece que ambos enfoques, aparentemente contradictorios, tienen un enorme valor en lo que respecta a programación de computadoras.

DONALD E. KNUTH, 1974.

3.1 INTRODUCCIÓN

3.1.1 La programación y su importancia

En la actualidad, "programación" es un término utilizado en muchas situaciones. Mientras que un contador programa una hoja de cálculo, un diseñador gráfico programa un editor de imágenes, un químico analiza la estructura tridimensional de una molécula

Alan J. Perlis (1922-1990). Computólogo estadunidense pionero en lenguajes de programación. Fue el primero en recibir el premio ACM Turing por su contribución en el área de compiladores y su participación en la familia de lenguajes de programación Algol.

Curiosidades

Con la programación se construyen grandes sistemas de cómputo, enormes y detalladas catedrales virtuales. El símil es en ocasiones literal, como en el caso de la Sagrada Familia, en Barcelona. Esta obra excelsa del arquitecto Antoni Gaudí (1852-1926), todavía inconclusa, que se empezó a construir el 3 de noviembre de 1883, fue digitalizada por Toni Meca, quien trabajó durante siete años en este proyecto. El resultado fue un modelo virtual increíblemente detallado, conformado por más de 35 millones de polígonos, es decir, 10 veces el tamaño del modelo que se creó para el barco de la película Titanic. Gracias al trabajo de Toni Meca podemos ver un modelo tridimensional en una pantalla de la versión terminada de la catedral.

y un físico realiza complicados cálculos numéricos. Aunque la manera en que se realiza la programación varía de un área a otra, los conceptos y fundamentos son los mismos.

Al igual que las matemáticas básicas, la programación se ha convertido en una habilidad necesaria. Sin embargo, a diferencia de las primeras, al programar se construyen objetos dinámicos, máquinas virtuales que realizan cosas para nosotros, con los cuales podemos jugar y experimentar. Diseñar programas es divertido y constituye un medio excelente para desarrollar habilidades básicas en muchos aspectos: pensamiento analítico, síntesis creativa, abstracción y atención al detalle. En resumen, diseñar programas obliga a observar con un enfoque activo.

Para aprender a programar es necesario interactuar directamente con la computadora y, como se verá en este tema, el ambiente de programación propuesto es una gran ayuda para el usuario, ya que indica los errores inmediatamente y permite que se exploren alternativas, se experimente con las construcciones y se evalúen los avances.

Ya se mencionó que los datos son las creaturas que habitan el mundo de la computación y que los algoritmos indican cómo manipularlas. Un algoritmo se puede escribir en español, en inglés o en cualquier idioma. Un programa es la escritura formal y precisa de un algoritmo en un lenguaje de programación, que es el lenguaje que una computadora entiende.

Es importante resaltar que la programación es divertida y abre las puertas a mundos nuevos, como cuando aprendemos un idioma. Programar implica usar símbolos, reglas y construcciones caprichosas mediante las cuales se pueden expresar ideas, conocimientos y algoritmos. Al igual que cuando se aprende un idioma, se requiere práctica y perseverancia para aprender a programar.

Los programadores y diseñadores de sistemas de cómputo cargan con grandes responsabilidades, y para que esta tarea no sea abrumadora, se crearon estrategias y formas de trabajo para asegurar, con relativa confianza, que un sistema de cómputo funcione como se espera. Por ejemplo, la mayoría de los programas modernos se escriben en capas o **módulos**; cada uno se prueba y se verifica por separado, de modo que su desarrollo, adecuación y mantenimiento se faciliten.

Aunque los programadores y diseñadores de sistemas toman muy en serio su tarea, en la historia se encuentran muchos ejemplos de errores de programación catastróficos. Por ejemplo, el primer viaje interplanetario que intentó realizar la NASA tenía como destino Venus; la nave se llamó *Mariner 1* y tuvo que ser destruida poco después del despegue, debido a dos errores. Primero, su recepción de rumbo por antena tenía demasiado ruido y, segundo, su programa de guía interno tenía un error: ¡le faltaba un guión!

A lo largo de la historia han sucedido grandes catástrofes provocadas por errores de software en diversas áreas. Aquí se mencionan algunas:[1]

Misiones espaciales:

- El *Mariner 1* de la NASA se salió de curso durante el despegue por un error en su software, escrito en lenguaje Fortran (22 de julio de 1962).
- El *Apollo 11* de la NASA tuvo un problema de aterrizaje (20 de julio de 1969).
- El *Voyager 2* de la NASA tuvo un contratiempo (25 de enero de 1986).
- El *Phobos 1* se extravió (10 de septiembre de 1988).
- El *Ariane 5* de la ESA se autodestruyó 40 segundos después del despegue (4 de junio de 1996).

[1] Wikipedia: errores de computación <http://en.wikipedia.org/wiki/Computer_bug>.

- El *Mars Climate Orbiter* de la NASA se destruyó porque utilizaba unidades imperiales en lugar de métricas (23 de septiembre de 1999).
- El *Mars Polar Lander* se extravió (3 de diciembre de 1999).
- El *Mars Rover* de la NASA se congeló debido a que tenía muchos archivos abiertos en su memoria *flash* (21 de enero de 2004).

Medicina:
- Los accidentes por radiación de la Therac-25 causaron al menos cinco muertes (1985-1987).
- La errónea utilización de un software de diagnóstico culminó con la muerte de ocho pacientes por exceso de radiación en Panamá (en 2000).

Computación:
- El problema del año 2000 o Y2K que causó enormes pérdidas financieras y una ola de terror por el inminente colapso económico fue ocasionado por almacenar sólo los últimos dos dígitos del año en los programas.
- Un error en el algoritmo de división usado en el procesador Pentium ocasionó a la empresa Intel pérdidas millonarias.

Milicia:
- El error de software en un misil Patriot mató a 28 estadunidenses en Dhahran, Arabia Saudita (25 de febrero de 1991).

Edsger W. Dijkstra (1930-2002) Reconocido computólogo holandés, recibió el premio ACM Turing en 1972 por sus contribuciones al área de lenguajes de programación. Poco antes de morir recibió el premio PODC por un artículo que dio inicio al estudio de la autoestabilización.

Edsger W. Dijkstra | *© Hamilton Richards.*

3.1.2 Lenguajes de programación

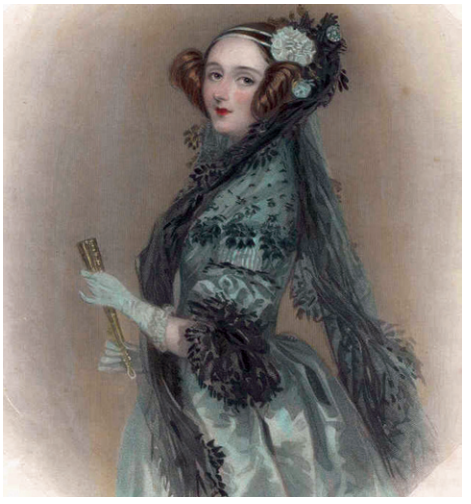
Programar representa un reto, en parte porque no es fácil describir una idea de manera precisa y detallada. Al "hablarle" a la computadora en un lenguaje de programación, cada palabra y cada oración tienen un sentido único, definido con precisión. No basta con decirle: "ordéname la lista de números", hay que proporcionarle más detalles: de cuántos números consiste, dónde están guardados, si deben ordenarse de menor a mayor o viceversa, etcétera. Además, el lenguaje de programación consta de un número fijo de instrucciones. Si la computadora no tiene una instrucción para ordenar números, hay que programarla a partir de las instrucciones que ya tenga o de otras que hayamos programado con anterioridad.

En la actualidad, algunos ingenieros que construyen computadoras se empeñan en que éstas sepan ejecutar pocas y muy sencillas instrucciones, de modo que faciliten lo más posible su construcción y favorezcan su velocidad. Los programadores generalmente tiran en el sentido opuesto: mientras más instrucciones

Concepto
Un módulo es un componente autocontenido de un sistema que, entre otras cosas, tiene una interfaz bien definida acerca de cómo interactúa con otros módulos. En programación, diseñar y programar o construir un módulo puede ser una tarea compleja. Sin embargo, una vez que un módulo existe puede fácilmente conectarse o desconectarse del sistema.

tengan y más poderosas sean éstas, más fácil resultará "escribir" un programa. El programador requiere de una sola instrucción que pueda usarse para ordenar listas de números, así como de otros elementos; por ejemplo, nombres de personas. En ingeniería se sabe que estas labores se pueden programar a través de instrucciones simples; sin embargo, a veces se muestra reticencia a implementar estas instrucciones en la electrónica. Por lo tanto, en la actualidad, el *hardware* de la máquina sólo es capaz de ejecutar instrucciones simples, las cuales constituyen el *lenguaje de máquina.* Programar un sistema complejo, como los que se usan comúnmente (una hoja de cálculo, un procesador de texto, un videojuego), por medio de ese lenguaje, sería poco menos que imposible. Por eso se han inventado los *lenguajes de alto nivel*. Se llaman así porque el nivel de abstracción en el que se encuentran hace que una sola de sus instrucciones equivalga a varias instrucciones del lenguaje de máquina. Por ejemplo, en algunos lenguajes de alto nivel sí existe una instrucción para ordenar listas.

Como la máquina sólo entiende su limitado repertorio de instrucciones de lenguaje de máquina, al programar en un lenguaje de alto nivel se necesitan *compiladores*: programas dedicados a traducir programas escritos en lenguajes de alto nivel a su equivalente en lenguaje de máquina.



Ada Lovelace.

Existen muchos lenguajes de programación. Cualquier cosa que se pueda programar en un lenguaje se puede programar en cualquier otro. Distintos programas escritos en diferentes lenguajes pueden ser ejecutados en una misma computadora gracias a los compiladores que los traducen a un programa escrito en lenguaje de máquina. Al igual que otros lenguajes (lenguajes hablados, lenguas), cada lenguaje de programación tiene un estilo peculiar y una facilidad especial para expresar cierto tipo de ideas.

Curiosidades
Se considera que Ada Augusta Lovelace (hija del célebre poeta británico Lord Byron) fue la primera programadora de la historia. Ella se concentró en la tarea de elaborar programas para una máquina que nunca se concretó: la máquina analítica diseñada por Charles Babbage. Para programarla se proporcionaban las instrucciones codificadas en trozos de cartón perforados, de acuerdo con un patrón que indicaba a la máquina las operaciones que debía realizar y en qué orden.

3.1.3 Notas acerca de programación y el lenguaje presentado

Como se mencionó anteriormente, aprender a programar tiene algunas similitudes con aprender un idioma o un deporte. Se pueden explicar las reglas, estrategias y el lenguaje propio del juego, pero la única manera de aprenderlo bien es jugarlo, practicar mucho, desarrollar la creatividad y un estilo propio. A lo largo de este tema se ofrecerá una visión de los conceptos fundamentales de la programación, que se espera sirvan de base para llevar el juego a la práctica con éxito.

Siempre que un autor presenta una introducción a la programación, se encuentra con que tiene que elegir un lenguaje de programación entre las decenas y decenas que se usan actualmente. Cada uno cuenta con ventajas y desventajas, pero usualmente el elegido es exaltado por sus adeptos como si fuera el rey de los lenguajes de programación. Esto se debe en parte a la fuerza de la costumbre (mientras más tiempo se ha usado un lenguaje, éste resulta más cómodo), y en parte a que cada uno es especialmente adecuado para programar problemas de cierto tipo. Sin embargo, para presentar conceptos básicos de programación que sirvan como introducción a esta fascinante disciplina, en realidad no importa mucho cuál sea el elegido. Las ideas básicas trascienden cualquier lenguaje y son de aplicación universal.

Es importante, sin embargo, elegir un lenguaje simple y conciso para no distraer al lector con detalles innecesarios o herramientas sofisticadas que nunca va a emplear. Por otro lado, es conveniente seguir un enfoque similar al que utilizan muchas de las principales universidades en el mundo para enseñar programación, y así aprovechar la literatura y las herramientas de soporte para el aprendizaje ya existentes. Es así como se eligió Scheme como el lenguaje de programación. Vale la pena aclarar que, de cualquier modo, no es difícil traducir la siguiente exposición a cualquier otro lenguaje moderno de programación. En el DVD anexo se ofrece información sobre una propuesta actual que ha tenido mucho éxito en años recientes para enseñar programación y que, por supuesto, explica con mayor detalle los principios fundamentales de la programación que se presentan aquí.

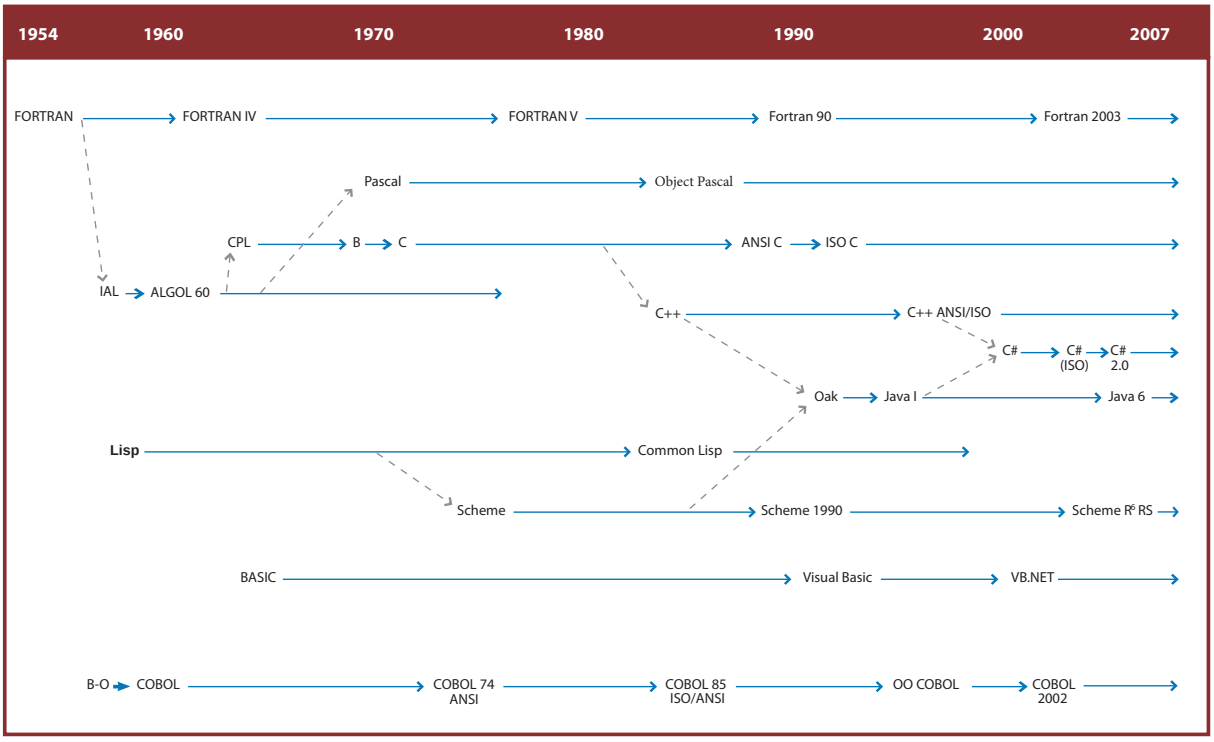
3.1.4 Un primer ejemplo: las torres de Hanoi

Pues bien, la magia debe comenzar. Para ello se recurrirá al algoritmo que se desarrolló en el tema anterior:

```
Resuelve TH (N, Fuente, Libre, Destino)
  Si N es 0
       Ningún movimiento
   De otro modo
       Resuelve(N - 1, Fuente, Destino, Libre)
       Mueve Fuente = > Destino
       Resuelve(N - 1, Libre, Fuente, Destino)
```

Curiosidades
Desde 1822, con la máquina diferencial de Charles Babbage, ha existido la necesidad de comunicar a las computadoras lo que queremos que hagan mediante una forma de lenguaje de programación. Con esta máquina era necesario cambiar sus engranes manualmente, así como más adelante, para otras computadoras, abrir y cerrar interruptores o "realambrar" tableros. La historia moderna de los lenguajes de programación comienza con Fortran en 1954, diseñado en IBM para cómputo científico. En 1958 se inventó Lisp para desarrollos de inteligencia artificial y, en 1959, se diseñó Cobol para uso en negocios.

Figura 1. Historia de los lenguajes de programación.



Curiosidades
Los premios A.M. Turing se entregan anualmente por la ACM (Association for Computing Machinery), una de las asociaciones de computación más importantes del mundo. Estos premios se entregan a personas cuyas contribuciones técnicas sean perdurables y de importancia para la comunidad de computólogos. Algunos los equiparan a los premios Nobel, pero en computación. Deben su nombre a Alan Mathison Turing, considerado uno de los padres de la computación moderna.

Sorprende lo fácil que es programarlo en Scheme. Las siguientes líneas constituyen un programa que resuelve el problema de las torres de Hanoi:

```
(define (resuelve-hanoi N fuente libre destino)
  (cond ((> N 0)
    (resuelve-hanoi (- N 1) fuente destino libre)
    (display "Moviendo disco ") (display (- N 1))
    (display " de: ") (display fuente)
    (display " -> ") (display destino) (newline)
    (resuelve-hanoi (- N 1) libre fuente destino))))
```

La primera línea es casi idéntica, salvo por detalles como el orden y los paréntesis. Las siguientes tres son la parte condicional que pregunta si ya no hay discos que mover, eso mismo hace el **cond** en el programa (**cond** es abreviatura de *conditional*). Obsérvese la parte fundamental del algoritmo, esto es, las últimas cuatro líneas: en primer lugar está `Resuelve(N - 1, Fuente, Destino, Libre)`, la primera llamada recursiva que resuelve las torres de Hanoi para los primeros N – 1 discos y en nuestro programa casi idéntica a `(resuelve-hanoi (– N 1) fuente destino libre)`, salvo por los paréntesis y detalles como comillas o el orden N – 1 en vez de – N 1. Lo mismo sucede con la última línea, que es la segunda llamada recursiva. Esto plantea el problema de descifrar qué hace:

```
(display "Moviendo disco") (display (- N 1))
(display " de: ") (display fuente)
(display " -> ") (display destino) (newline)
```

Que es esencialmente lo mismo que: `Mueve Fuente => Destino` del algoritmo original. Lo único que hacen esas tres líneas es desplegar en la pantalla el movimiento que se realiza, que es también la tarea de `Mueve` en el algoritmo. Aquí va una ejecución del programa en acción. Primero, se teclea el siguiente comando al intérprete de Scheme (más adelante se describirá en detalle):

```
> (resuelve-hanoi 3 'A 'B 'C)
```

Entonces la computadora responde con el resultado de ejecutar el programa:

```
Moviendo disco 0 de:     A -> C
Moviendo disco 1 de:     A -> B
Moviendo disco 0 de:     C -> B
Moviendo disco 2 de:     A -> C
Moviendo disco 0 de:     B -> A
Moviendo disco 1 de:     B -> C
Moviendo disco 0 de:     A -> C
```

Nótese que se han marcado en negritas las líneas que indican movimientos finales de discos a la torre de destino, es decir, C. Ya se había dicho que bastaba decir A –> C para indicar que se debe mover el disco más pequeño de la torre A a la C. Aquí, sin embargo, se imprime el número de disco para facilitar la lectura, cero (0) es el más chico y dos (2) en este caso el más grande.

Fue muy fácil, ¿no es así? Sin embargo, se dieron por sentado algunos detalles de programación que se explicarán posteriormente. Vamos a ubicar y referir algunos de esos detalles. Por principio de cuentas, todas las palabras en negritas son palabras del lenguaje Scheme con un significado fijo, llamadas palabras reservadas; algunas son procedimientos primitivos como – (resta) o `display` (que imprime algo en la pantalla); otras son expresiones de control o formas especiales, como `define` (definición de variables o procedimientos) o `cond` (forma condicional).

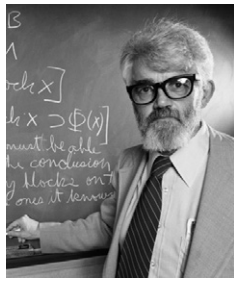
Ahora se desarmará el programa para señalar los detalles del lenguaje: `(define (resuelve-hanoi N fuente libre destino))`, *define* un procedimiento (o programa) que se llama `resuelve-hanoi`. Este procedimiento recibe cuatro argumentos: el número de discos, `N`, así como tres torres `(fuente, libre y destino)`. Una vez definido un procedimiento, se puede ejecutar; esto se hizo después con: `(resuelve-hanoi 3 'A 'B 'C)`. Nótese que antes había un símbolo >, que se conoce como *prompt* y que es la forma del intérprete del lenguaje para indicar que está listo para recibir instrucciones y ejecutarlas.

Todo lo que sigue después de la primera línea se conoce como el *cuerpo* del procedimiento. Usualmente se utiliza una combinación de instrucciones de control y expresiones en estas definiciones. En el programa de las torres tenemos una expresión única formada por la estructura de control, `cond`. En el procedimiento se puede traducir como: "Si el disco que manejamos no es el más pequeño, `(> N 0)`, entonces hay que: *a*] mover todos menos el mayor a la torre de apoyo o libre; *b*] mover el disco más grande a la torre de destino y *c*] mover todos los discos que se pusieron en la torre de apoyo a la torre destino". Los pasos (*a*) y (*b*) ya mencionados son llamadas *recursivas*. Cabe recordar que en el tema anterior hay una explicación sobre la recursividad.

Cuando se ejecuta el procedimiento, la computadora (o mejor dicho, el intérprete de Scheme que lo traduce a lenguaje de máquina y luego lo ejecuta en la computadora) asocia el nombre `N` con `3`. También asocia `fuente` con `'A`, que es un símbolo, etcétera (posteriormente se verá más de esto). Cuando se define un procedimiento, los nombres de los argumentos se llaman *parámetros*, aunque es usual llamarlos *argumentos*.

¡Y eso es todo!, el programa recursivo es prácticamente igual a la descripción recursiva del algoritmo del tema 2. Por supuesto, lleva un poco de tiempo acostumbrarse a la sintaxis del lenguaje y a los nombres de cada parte.

Curiosidades
Lisp por list processing, o procesamiento de listas, fue creado por John McCarthy (1927), quien recibió el premio Turing en 1971 por sus contribuciones a la inteligencia artificial. Pero aunque el Lisp se ha utilizado con mucho éxito en aplicaciones de inteligencia artificial, es de uso general. Originalmente servía para manipular símbolos, no sólo números y otros datos simples, como con Fortran, el único lenguaje de programación en uso más viejo que Lisp.



John McCarthy | *© Latin Stock México.*

3.2 NOCIONES BÁSICAS DE SCHEME

Para que la experiencia sea memorable y divertida —sobre todo si éste es el primer encuentro con la programación—, en este libro se utilizará un ambiente de programación que integra, igual que los sistemas comerciales para desarrollo, todo lo que vamos a necesitar: un editor inteligente que ayuda a escribir los programas, un sistema de ayuda y documentación, una máquina virtual o intérprete que los ejecuta y una serie de herramientas para seguir paso a paso la ejecución de tus programas.

El ambiente que se utiliza se llama DrScheme[2] y el lenguaje de programación es Scheme; sin embargo, es importante mencionar que DrScheme soporta otros lenguajes, como

[2] DrScheme es un ambiente desarrollado por el grupo de investigación en lenguajes de programación: *Programming Languages Team* (PLT), que está formado por media docena de investigadores de varias universidades de Estados Unidos. Para mayor información y obtener una copia de DrScheme, que es software libre y de distribución gratuita, visite: <http://www.drscheme.org>.

Curiosidades
DrScheme es un ambiente de programación gráfico, interactivo e integrado para distintos lenguajes de programación (Scheme, MzScheme y MrEd). DrScheme puede ser ejecutado en distintas plataformas como Windows (versiones 95 y superiores), Mac OS X (versiones 10.3 y superiores) y prácticamente en todos los sistemas operativos Unix con el sistema de ventanas X. Sólo hay que visitar la página www.drscheme.org y seguir las sencillas instrucciones para instalarlo. Una vez instalado, debe seleccionarse el lenguaje de programación que se desea usar. Una buena opción para los objetivos de este módulo es PLT –> Full Swindle. Puedes seleccionar idioma español desde el menú de ayuda, si prefieres interactuar con DrScheme en nuestra lengua.

Java y Algol. Una gran ventaja de DrScheme es que empieza como un ambiente de programación para principiantes y "crece" con el usuario, conforme aprende y se convierte en un programador.

3.2.1 El lenguaje de programación Scheme

Se empleará una de las principales vertientes de Lisp, el lenguaje Scheme. De hecho, Scheme es una familia de lenguajes de programación, donde cada uno de los miembros es similar a los demás, pero también cuentan con diferentes módulos para facilitar o soportar diferentes metodologías, paradigmas o enfoques de programación. Aquí se usará PLT Scheme, un lenguaje muy socorrido en el mundo de la enseñanza de la programación y la investigación de temas relacionados con los lenguajes.

Scheme tiene características que lo convierten en el vehículo perfecto para estudiar construcciones y conceptos fundamentales de la programación. Por ejemplo —y ésta es una característica que comparten Lisp y todos sus descendientes—, las descripciones de procesos que en Scheme se conocen como procedimientos,[3] pueden ser tratados como datos.

3.2.2 El ambiente de programación DrScheme

DrScheme provee resaltado del texto del programa fuente para mostrar errores de sintaxis o de ejecución. Además, soporta, como ya se mencionó, varios lenguajes con distinto nivel de soporte para el usuario, desde principiante hasta nivel de Scheme extendido, con las bibliotecas para desarrollo de aplicaciones gráficas, un sistema de clases y objetos, TCP/IP para acceso a redes de computadoras, etcétera.

La ventana principal de DrScheme luce como se muestra en la figura 2. Se observa que hay dos partes principales: la superior se llama *ventana de edición*, donde se escriben los programas; y la inferior, *de interacciones*, donde la máquina virtual o intérprete de Scheme corre en todo momento.

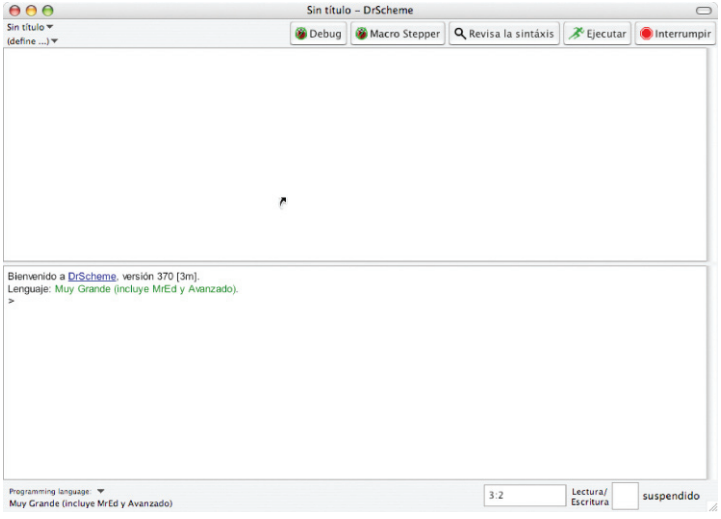


Figura 2. Ventana de DrScheme cuando inicia.

[3] Para quienes estén acostumbrados a otros lenguajes de programación como Pascal, C, C++, C# o Java, lo que aquí se llamarán procedimientos o programas, en esos lenguajes se conoce como funciones o métodos.

Los dos botones en la parte superior derecha son muy importantes: *ejecutar* permite probar todo lo que está escrito en la ventana de interacciones; es decir, el intérprete de Scheme evalúa las expresiones en la ventana de edición y posteriormente pasa a la ventana de interacciones para que se prueben las definiciones o se muestren los errores generados. El segundo botón, *interrumpir*, permite suspender la ejecución actual; es muy útil cuando algo se sale de control y no se detiene.

3.2.3 Metodología de diseño

Ya casi está todo listo para empezar a programar, pero hace falta ponerse de acuerdo con algo vital: la metodología de diseño que se seguirá. Como se dijo anteriormente, programar es el *arte de diseñar procesos.* Al igual que en muchas otras áreas relacionadas con el diseño, aquí se harán descripciones en el lenguaje común para describir los procesos y explicar cómo evolucionarán, para después traducirlos, como los buenos hechiceros, al arcano lenguaje de programación.

Se seguirá una serie de guías de diseño para desarrollar los procedimientos. La intención es que estas recetas lleven de manera natural desde la descripción del problema hasta la solución computacional, paso a paso, y generando resultados intermedios que faciliten tu comprensión de la programación. También, se espera apoyar con esto el desarrollo de las habilidades de lectura, análisis, organización, experimentación y pensamiento, que son útiles en muchos otros aspectos. En la tabla 1 se muestran los pasos de la receta de diseño con los resultados intermedios esperados al final de cada uno:

Tabla 1. Pasos básicos de la receta de diseño de procedimientos.

1	Análisis del problema y definición de datos.	Determinación de cuáles son los datos de entrada del problema y qué características poseen.
2	Contrato, propósito y enunciados de efecto, encabezado.	Descripción informal del comportamiento del procedimiento.
3	Ejemplos.	Ilustran el funcionamiento deseado del procedimiento.
4	Definición del procedimiento.	La transformación del modelo en una definición completa, refinando los detalles omitidos de la versión anterior.
5	Pruebas.	Se prueba el procedimiento con datos de entrada para los que ya se conoce el resultado. El objetivo es detectar posibles errores.

3.2.4 Expresiones primitivas y datos simples

Armados con las herramientas necesarias para esta travesía y recorrer los fundamentos de la programación, ¡manos a la obra! Para dar una idea clara de qué se espera, se verá cómo funciona el ambiente de trabajo, comenzando por el intérprete de Scheme que se encuentra en la ventana de interacciones de DrScheme.

Expresiones

Cada vez que se escribe algo como una expresión y se ingresa al intérprete con *entrar*, éste responde imprimiendo en pantalla el resultado de su evaluación. Existen muchos tipos de expresiones, de las cuales las más sencillas son los números. En la figura 3 se muestran algunas interacciones muy sencillas con números.

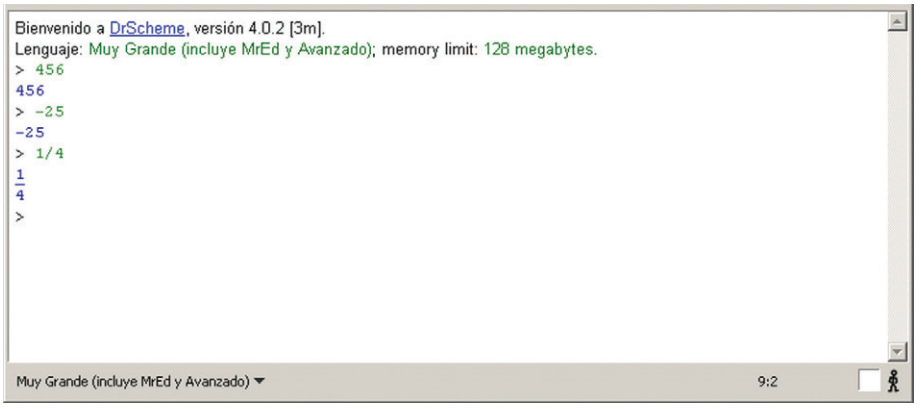


Figura 3. Ejemplo de evaluaciones en DrScheme.

Nótese que el intérprete regresa (marcado en azul) exactamente el mismo valor; esto se debe a que los números y algunos otros elementos son *autoevaluados*, es decir, que su valor de regreso es el mismo que el de su representación.

Por convención y espacio, a partir de este momento no se mostrarán las imágenes de DrScheme, sino que se indicarán directamente las expresiones, procedimientos o programas, y se marcarán con azul las evaluaciones o los mensajes que imprima DrScheme.

Hay varios procedimientos primitivos, o sea que ya están integrados al sistema para operar con números: + (suma), - (resta), / (división), * (multiplicación), entre otros. Para realizar operaciones con estas primitivas se utiliza la notación prefija:

```
(+ 2 3)   (- 5 1)   (* 3 4)   (/ 25 3)
```

Todas estas formas también son expresiones, por lo que se puede utilizar el intérprete como cualquier calculadora de bolsillo y, por supuesto, hará lo correcto, por ejemplo:

```
> (+ 2 3)
5
```

Un número es una expresión compuesta. Las expresiones compuestas serán las más comunes y no hay límite en las combinaciones de expresiones que se pueden realizar. Por ejemplo, las siguientes son expresiones válidas:

```
(+ (- 5 1) 4)        (/ (+ 1 2) (* 3 4))
```

La primera regresa 8 y la segunda ¼. ¿Cómo funciona esta evaluación? Para explicarlo se usará un sencillo modelo de sustitución de expresiones. Así, la primera expresión indica:

```
> (+ (- 5 1) 4)
> (+ 4 4)
8
```

Lo que sucedió es que primero se evalúan las expresiones internas y se sustituyen en la versión original. Las sustituciones se realizan de adentro hacia afuera; primero, las más internas y así sucesivamente hasta obtener un resultado. En general, todas las expresiones en Scheme tienen la forma

```
(operación A … B)
```

lo que facilita saber qué tiene que evaluarse primero. Si todos los argumentos A…B son números, se puede realizar la operación y obtener un valor. Si no es así, entonces primero se evalúa A…B (en cualquier orden) y se sustituyen sus resultados en la expresión original. Con una expresión escrita de la manera familiar, como: "5 + 3 * 2" no sucede lo mismo. ¿Cómo saber si primero se debe sumar y luego multiplicar para obtener 16 de resultado? O hacerlo al revés, y obtener 11 como resultado.

Se sabe debido a la convención de dar siempre preferencia al * y ejecutarlo antes del +. La versión en Scheme para obtener el mismo resultado es: `(+ 5 (* 3 2))`. En la manera familiar de escribir sería impráctico tener que definir una precedencia entre las operaciones y, además, a veces se requiere que las operaciones se evalúen en un orden diferente al convencional, así que de todos modos se termina por usar paréntesis para indicar el orden deseado de evaluación, como en (5 + 3) * 2.

Además de las operaciones básicas, como suma y multiplicación, Scheme también soporta varias operaciones matemáticas avanzadas: `(sqrt n)`, raíz cuadrada de n; `(exp n m)`, exponencial n^m; `(log n)`, el logaritmo natural de n; `(sin n)`, que calcula el seno de n radianes. Existen varias operaciones aritméticas más y, en general, una buena forma de saber si DrScheme las soporta o no es escribir un ejemplo en la ventana de interacciones.

Una característica que se utilizará frecuentemente es la habilidad del intérprete para ignorar espacios blancos; así, por ejemplo, la expresión:

```
> (- (+ 2 (* 12 (/ (expt 2 5) 2))) 3)
```

que se evalúa 191, puede escribirse como:

```
> (- (+ 2
       (* 12
            (/ (expt 2 5)
         2)
   )
   )
3)
```

lo que facilita la lectura y comprensión de la expresión.

Variables y procedimientos

Al igual que en álgebra, en programación se utilizan variables como *marcadores* para cantidades que no se conocen. Por ejemplo, el área de un círculo de radio *r* está dada por: $3.14 \times r^2$. En esta fórmula, *r* es un marcador para cualquier número positivo. Estos marcadores se conocen como variables.

Al programar, una expresión que contiene variables se convierte en una regla que indica cómo calcular una operación determinada cuando obtenemos un valor para las variables. Un procedimiento o programa es una de esas reglas. Así, se define un procedimiento para calcular el área de un círculo:

```
(define (área-círculo r)
 (* 3.14 (* r r)))
```

Más aún, se puede definir también la variable *pi* con su valor usual:

```
(define pi 3.14)
```

y entonces reescribir el método original para que utilice esta variable:

```
(define (área-círculo r)
 (* pi (* r r)))
```

Ahora se puede *aplicar* o utilizar esta nueva definición de la misma manera en que se usan las primitivas. Nótese que la definición misma del procedimiento indica cómo se debe utilizar: `(área-círculo r)`. Por ejemplo:

```
> (área-círculo 3)
28.26
```

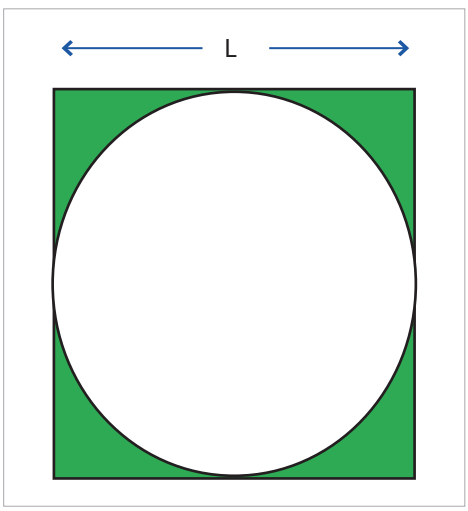


Figura 4. Círculo dentro de un cuadrado.

En este ejemplo, al número 3 se le denomina el argumento del procedimiento. En contraste, a la *r* se le llama el parámetro.

Se puede utilizar el procedimiento `área-círculo` para construir procedimientos más complicados, como ocurrió con la primitiva `*`. Un buen intento es el siguiente: hacer un programa que calcule el área sombreada en la figura 4.

El procedimiento para determinar el área del círculo fue muy sencillo. Se ve más complicado, pero sólo se tiene que calcular el área del cuadrado y restarle el área del círculo. Es momento de utilizar la receta de diseño; son seis sencillos pasos:

```
;; Contrato: área-cuadrado-círculo: L[número] -> número
   ;; Propósito: calcular el área que se obtiene de restar
```

```
   el área del
;; círculo a la del cuadrado. Recibe el tamaño L de un lado
;; del cuadrado.
;; Ejemplo: (área-cuadrado-círculo 5) debe producir 5.375
;; Definición:
(define (área-cuadrado-círculo L
 (- (área-cuadrado L)
   (área-círculo (/ L 2))))
;; Pruebas:
(área-cuadrado-círculo 5)
```

Hay muchos detalles que deben explicarse aquí:

1] Todo se escribe en la ventana de definiciones de DrScheme.
2] Las líneas que empiezan con ";" son comentarios que ignora el intérprete de Scheme.
3] Ya se conoce el `área-círculo`, pero se utiliza además `área-cuadrado`. ¿Cuál es el procedimiento? Todavía no se ha definido. Se podría dejar como ejercicio, pero es muy sencillo: el área de un cuadrado de lado L es L^2. Adelante se encuentra la definición.
4] ¿Por qué utilizar `(/ L 2)` como radio del círculo?
5] Algunos pasos de la receta no son explícitos, por ejemplo el primero, el análisis del problema. No se hizo mayor énfasis, porque ya se indicó que la resta de las áreas de las figuras determina el área sombreada y eso funciona como análisis. Se agregó, sin embargo, una primera sección llamada *contrato*, cuyo nombre se debe a que todos los procedimientos consumen y "regresan" algo, en este caso números. La L entrada representa la longitud del lado del cuadrado (y ya se sabe que la mitad de eso es el radio del círculo) y la salida representa el área sombreada.

Ésta es la definición que faltó para determinar el área de un cuadrado. Como se señaló en el tema 2, se conoce como diseño de arriba abajo.

```
(define (área-cuadrado L)
 (* L L))
```

Hay otras formas de resolver el mismo problema; por ejemplo, en lugar de utilizar `área-círculo` se puede insertar `(* pi (* (/ L 2) (/ L 2))`, que calcula el área del círculo. Sin embargo, como es fácil imaginar, luciría más complejo el procedimiento completo. El uso de `área-círculo` y `área-cuadrado` agrega claridad, hace más fácil el mantenimiento y corrige errores en el procedimiento. Estas funciones que calculan una parte de la solución final reciben el nombre de procedimientos auxiliares.

Como regla, trata de formular definiciones o procedimientos auxiliares para todas las dependencias entre las cantidades que se mencionan en la definición del problema. En el ejemplo anterior, las relaciones obvias eran el área del círculo interior y el área del cuadrado. El nombre estándar de estos procedimientos auxiliares es *subrutina* y se presentó en el tema 2.

Expresiones condicionales

Por ejemplo, ¿cómo se le podría hacer para calcular el valor absoluto de un entero? En matemáticas se diría (y esto sirve como análisis del problema):

$$|x| = \begin{cases} x \text{ si } x \geq 0 \\ -x \text{ si } x < 0 \end{cases}$$

Por fortuna, Scheme incluye muchas primitivas para realizar este tipo de pruebas o comparaciones y también diversas formas de llevar a cabo una u otra alternativa. A continuación, algunas interacciones básicas:

```
> (= 5 5)
#t
> (> 5 6)
#f
> (> 5 3)
#t
> (>= 5 12)
#f
> (and (> 5 3) (> 5 2))
#t
> (or (> 5 7) (> 5 2))
#t
> (not (> 5 7))
#t
```

Los valores `#t` y `#f` son utilizados por Scheme para representar los valores de verdad, *verdadero y falso* respectivamente. Los procedimientos para hacer comparaciones numéricas, `<`, `>`, `=`, `<=` y `>=`, funcionan como lo esperado y los otros procedimientos son de lógica y representan la conjunción: `and` ("y" en español), la disyunción `or` ("o" en español), la negación `not` ("no" en español).

Aún no se puede construir el procedimiento para calcular el valor absoluto de un número. Falta un mecanismo para decidir entre distintas opciones. Hay varias alternativas en Scheme para realizar esta labor, las cuales se conocen como expresiones condicionales; la más general se da a través de la forma primitiva `cond`, cuya forma general es:

```
(cond
 [pregunta respuesta ]
 ...
 [pregunta respuesta ]
 )
```

```
(cond
 [pregunta respuesta ]
 ...
 [else respuesta ]
 )
```

¡Ah!, por cierto, en DrScheme se pueden utilizar paréntesis o corchetes de manera indistinta. En general, se utilizarán para facilitar la lectura cuando hay varios paréntesis juntos.

Regresando a las expresiones `cond`; el valor total de la expresión es la primera respuesta cuya pregunta sea verdadera. Considérense los siguientes ejemplos:

```
> (cond
        [(> 5 6) 1]
        [(< 5 6) 2]
        [(= 5 6) 3]
        )
2
> (cond
   [(and (> 5 4) (= 5 6)) 1]
   [(= 5 6) 2]
   [else 3]
)
3
```

Obsérvese que tanto las preguntas como las respuestas son expresiones cualesquiera; sin embargo, en el caso de las preguntas, Scheme espera recibir un valor de verdad, ya sea falso o verdadero. Algo muy práctico es que en Scheme y en muchos otros lenguajes de programación, cualquier expresión que no es falsa `(#f)` es verdadera. Por ejemplo, `5` o `(+ 3 5)` son expresiones que se consideran verdaderas, aunque no correspondan a un valor de verdad, propiamente hablando. La expresión `else` funciona como último recurso y puede interpretarse como: "Si no hay otra opción, utilizar esta respuesta".

Ahora todo está listo para escribir un procedimiento que calcule el valor absoluto de un entero; ya se ha hecho el análisis y se puede proceder directamente a la receta de diseño:

```
;; Contrato: absoluto: n[entero] -> entero no-negativo
;; Propósito: calcular el valor absoluto de un entero
;; Ejemplo: (absoluto -3) debe producir 3
;; Definición:
(define (absoluto n)
 (cond [(>= n 0) n]
    [(< n 0) (- n)]))
;; pruebas:
(absoluto 5)
(absoluto 0)
(absoluto -5)
```

¿Qué nuevas cosas hay en este procedimiento absoluto? ¿Qué significa `(- n)`? Se supone que la operación – (resta) recibe dos argumentos, ¿o no? La respuesta es sí y no. La resta como se conoce es una operación binaria. Sin embargo, en Scheme, la mayoría de las operaciones matemáticas que son asociativas, como la resta, reciben desde uno hasta cualquier cantidad de argumentos. Analícense los siguientes ejemplos:

```
> (- 1)
-1
> (- 5 4 3)
-2
```

```
> (- 5 4 3 2 1)
-5
```

El último ejemplo sería equivalente a la operación:

```
> (- (- (- (- 5 4) 3) 2) 1)
-5
```

que es la versión que muestra la *asociación* para la resta de manera explícita. Es decir, la resta de cinco números se lee naturalmente de izquierda a derecha: a la resta de 5 menos 4, restarle 3; al resultado obtenido restarle 2 y al siguiente resultado restarle 1. Cuando la resta recibe un argumento único, como (`- 5`), funciona como la negación y por ello el resultado es `-5`.

La forma `cond` es muy general y resulta una práctica común en todos los lengua-jes de programación. Para los casos más típicos existen formas especiales; por ejemplo, cuando se quiere plantear una sola pregunta y decidir entre un camino u otro, izquierda o derecha, arriba o abajo, existe otra forma especial que se llama `if`. La sintaxis de `if` es más sencilla porque utiliza menos paréntesis:

```
(if pregunta respuesta-verdadera respuesta-falsa)
```

Es decir, la expresión `respuesta-verdadera` se evalúa cuando la pregunta regresa verdadero (`#t`); cuando regresa falso `#f`, se evalúa la expresión `respuesta-falsa`.

Estructuras de control

Reflexionando un poco sobre "la forma" de los procedimientos que se escribieron arriba, por ejemplo: `absoluto`. En este procedimiento se utilizó una expresión condicional con tres alternativas. ¿Qué sucede cuando se evalúa (`absoluto -5`)? Ocurren las siguientes acciones en orden consecutivo:

Curiosidades
La programación funcional es uno de los tipos fundamentales de programación o paradigma de programación, que trata los cómputos como evaluación de funciones matemáticas. Los lenguajes puramente funcionales evitan la mutación o modificación de datos. Scheme facilita el estilo funcional de programación, pero también soporta operaciones de mutación.

1] El intérprete nota que `-5` es un número; entonces, evalúa el procedimiento absoluto y pasa como argumento `-5`.
2] El cuerpo del procedimiento es una expresión única condicional, así que se evalúa la primera pregunta: (`> = n 0`), donde `n = -5`; es decir, se evalúa por nuestro modelo de sustitución: (`> = -5 0`), que regresa falso, `#f`.
3] Finalmente se evalúa (`< -5 0`), que es verdadero, `#t`, y posteriormente la respuesta de esta alternativa: (`- n`); es decir (`- -5`), que regresa `5`.

Es una ejecución lineal, pero recuérdese que los algoritmos realmente interesantes requieren repetir una tarea una y otra vez.

3.2.5 Recursividad

En el capítulo anterior se analizó la recursividad con detalle. Es interesante notar que en Scheme la recursividad es una forma natural de repetir tareas por el tipo de programación funcional que soporta. Al detallar la versión de factorial, utilizando recursividad y siguiendo la receta de diseño se obtiene:

```
;; Contrato: factorial: n[entero positivo] -> entero
;; Propósito: calcular el factorial del número de entrada.
;; Ejemplo: (factorial 5) debe producir 120
;; Definición:
(define (factorial n)
 (if (<= n 1)
   1
   (* n (factorial (- n 1)))))
;; Pruebas:
(factorial 5)
(factorial 30)
```

El cuerpo del modelo de procedimiento es una condicional sencilla, de las que sólo tienen dos caminos, así que es posible utilizar la forma `if`. Se puede ver claramente que la siguiente parte se trata de una multiplicación de *n* por el factorial de `n – 1`. Al examinar las pruebas se observa que sí funciona:

```
> (factorial 5)
120
> (factorial 30)
265252859812191058636308480000000
```

Hay un procedimiento de Scheme que calcula el resultado de un número elevado a una potencia, `exp`., pero un reto interesante es programar un procedimiento propio, de modo que se anima al lector a crear un procedimiento llamado potencia.

Primero se analiza el problema: ¿cuál es el caso más sencillo de las exponenciales? Cuando el exponente es 0, porque cualquier base elevada a 0 es 1. Éste será el caso base.[4] ¿Cómo luce en la receta de diseño, tomando en cuenta que se tiene que hacer una pregunta con dos posibles respuestas para saber si el exponente es cero o no?

```
;; Contrato: potencia: base[número] exp[entero] -> número
;; Propósito: calcular el valor de la función exponencial,
tomando
;; como base el primer número y como exponente, el segundo.
;; Ejemplo: (potencia 2 3) debe producir 8
;; Definición:
(define (potencia base exp)
 (if (= exp 0)
   1
   (* base (potencia base (- exp 1)))))
;; Pruebas:
(potencia 2 3)
(potencia 3 5)
```

[4] Todos los procedimientos recursivos necesitan un caso base, aunque existen algunos con más de uno. También se conocen como ruta de escape, porque *terminan* la recursividad. Debe verificarse que cada llamada recursiva debe acercar a uno de los casos base; de otra forma, nunca terminará el programa.

3.2.6 Ciclos

Recursividad es un concepto muy poderoso en computación y es equivalente a herramientas para armar ciclos, comunes en muchos lenguajes de programación. Si alguien ha programado en Fortran, Pascal, C, C#, C++, Java, etc., seguramente estará familiarizado con construcciones como `for` o `while`.

Aunque no es lo más común en Scheme, ni en los lenguajes funcionales en general, una de las maneras de repetir una actividad en un programa se logra a través de ciclos. En muchos lenguajes existen construcciones conocidas como `for`, `while`, `repeat` o `do`, que son distintas formas de configurar ciclos o secuencias de expresiones que se repiten un cierto número de veces o hasta que una condición se cumpla.

Más adelante se verán otras variantes para realizar ciclos sobre estructuras particulares, por ejemplo, sobre una lista. Por el momento, sin embargo, la explicación se enfocará en el uso del `while`, el recurso más utilizado para generar ciclos. La construcción general es:

```
(while condición
    expresiones)
```

Con `while`, la idea es que mientras se cumpla la `condición`, es decir, mientras la condición se evalúe como verdadera, se ejecutarán las `expresiones`. Veamos `while` en acción con un programa para calcular el factorial n! = $n \times n-1 \times \ldots \times 1$. Es claro que se quiere repetir la multiplicación varias veces para poder recorrer de la n al 1, o al revés. Será necesario utilizar una variable `contador` para contar de cero a n y una segunda variable para almacenar los resultados de las multiplicaciones parciales, `resultado-parcial`. ¿Cuál debe ser el valor inicial de las variables `contador` y `resultado-parcial`? Y, más aún, ¿cómo almacenar el valor de esa multiplicación como el *nuevo* resultado parcial? Se hará un pequeño paréntesis para introducir la construcción que hace falta.

3.2.7 Asignación

Es usual querer conservar valores intermedios durante un cómputo; para logralo, Scheme ofrece un mecanismo que permite realizar *asignaciones* de un valor a una variable.

Variables globales

La primera parte que consiste en definir una variable ya es conocida y se realiza igual que con las definiciones de procedimientos, es decir, con `define`:

```
(define variable valor-inicial)
```

Una vez que la variable existe, se puede modificar su valor con la forma especial `set!`; aquí, al igual que en los predicados, el signo de admiración se utiliza por convención para señalar que en adelante *modifica* el valor que se le había dado al argumento. La asignación se debe usar con cuidado porque en su presencia el modelo de sustitución para las evaluaciones no funciona. Así se utiliza `set!`:

```
(set! variable nuevo-valor)
```

Ahora sí, todo está listo. Utilizando `set!` se va a escribir el procedimiento `factorial` para que modifique la variable que guarda el resultado intermedio:

```
;; Definición:
(define contador 1)
(define resultado-parcial 1)
(define (factorial n)
 (while (<= contador n)
     (set! resultado-parcial (* contador resultado-parcial))
     (set! contador (+ contador 1))
     )
 ;; regresamos el resultado
 resultado-parcial)
```

¿Por qué no se selecciona cero como el valor inicial del `resultado-parcial`? ¿Cuánto vale el factorial de 0?

Como se puede notar, se modificó también el contador al incrementar un 1 y con esto, efectivamente, se logra avanzar de uno en uno el `contador` desde 1 hasta *n*. El programa calcula de manera correcta el factorial de un número entero positivo si se ejecuta una sola vez. Este diseño no es bueno debido a la utilización de las variables globales `contador` y `resultado-parcial` que no están asociadas únicamente al procedimiento factorial para el cual están pensadas. Por eso no siempre funciona bien si se ejecuta repetidamente.

Curiosidades
El paradigma de programación, como el que se usa en el caso del factorial con contadores, recibe el nombre de programación imperativa; en ella se hace un uso extenso de asignaciones.

Variables locales

Se ha visto cómo hacer asignaciones y a partir de esto crear variables globales con `define` y `set!`; sin embargo, la idea de las variables globales es permitir definiciones que tengan sentido por sí mismas. En el ejemplo de `factorial` citado arriba, había una variable `contador` y otra `resultado-parcial`, que sólo tenían sentido en el procedimiento, por ello se aclara que era una mala práctica de programación. Ahora se utiliza una forma local de asignación llamada `let`, cuya forma general es:

```
(let [(var1 valor1)
   (var2 valor2)
   ...
   (varN valorN)]
  expresiones)
```

Ahora se puede reescribir el procedimiento `factorial`, en estilo imperativo, pero con variables locales:

```
(define (factorial n)
 (let ((contador 1)
    (resultado-parcial 1))
  (while (<= contador n)
     (set! resultado-parcial (* contador resultado-parcial))
      (set! contador (+ contador 1))
      )
```

```
  ;; regresamos el resultado
  resultado-parcial))
```

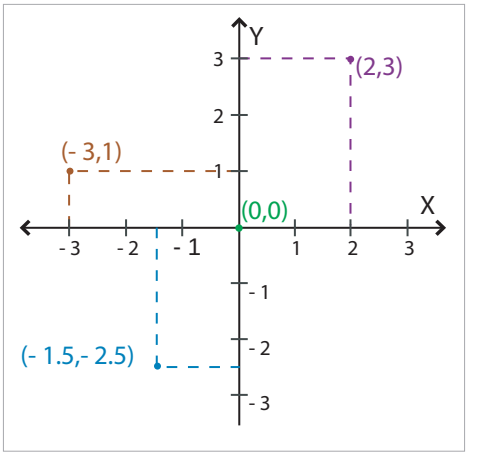


Figura 5. Ejemplo de plano cartesiano con algunos puntos.

lo cual, además de calcular el factorial correctamente siempre que se ejecute, es más claro, y los contadores que sólo tienen sentido para el procedimiento son *invisibles* desde fuera.

A continuación se presentará otro ejemplo que aclara el uso de `let`. Recuérdese la manera de calcular la distancia entre dos puntos en el plano. Por ejemplo, la distancia entre los puntos (-3,1) y (2,3), los puntos café y verde, respectivamente, en la figura 5, se calcula de la siguiente forma:

$$d=\sqrt{(-3-2)^2 + (1-3)^2}$$

Es decir, se restan las coordenadas *X*, luego las *Y*; se elevan al cuadrado esas cantidades, se suman y se saca raíz cuadrada de todo. Es muy sencillo, así que aquí se muestra la versión con receta de diseño:

```
;; Contrato: distancia número número número número -> número
;; Propósito: Calcular la distancia entre dos puntos
cualesquiera
  en el plano.
;; Ejemplo: (distancia -3 1 2 3) debe producir 5.38
;; Definición
(define (distancia x1 y1 x2 y2)
 (sqrt (+ (* (- x1 x2) (- x1 x2))
      (* (- y1 y2) (- y1 y2)))))
;; Pruebas:
(distancia -3 1 2 3)
```

Como puede notarse, se repiten dos veces las restas: `(- x1 x2)` y `(- y1 y2)` y éste es un caso típico para utilizar `let`, como una especie de abreviatura:

```
(define (distancia x1 y1 x2 y2)
 (let ((dif-x (- x1 x2))
    (dif-y (- y1 y2)))
  (sqrt (+ (* dif-x dif-x)
       (* dif-y dif-y)))))
```

En estos casos, además de mejorar la visibilidad del procedimiento, también se reduce la posibilidad de cometer errores al escribir menos veces cada término, así el intérprete puede ejecutar más rápido el procedimiento. Por ejemplo, en lugar de calcular cuatro restas, DrScheme sólo calcula dos restas y utiliza los resultados de éstas las cuatro veces que son requeridas.

En el capítulo anterior se analizó el problema de encontrar los puntos más cercanos en un conjunto de *n* puntos, y el algoritmo de fuerza bruta que se definió es:

```
Para cada pareja de puntos x, y:
   Sea d la distancia de x a y
   Si d es menor a dmin, la menor distancia vista hasta ahora,
  Sea dmin igual a d
```

¿Cómo escribir eso en Scheme? Ya se sabe calcular la distancia entre dos puntos, pero ¿cómo representar un conjunto de puntos?

3.3 ABSTRACCIÓN CON DATOS

La sección anterior se centró en procesos computacionales y en el diseño de procedimientos: datos básicos (números) y operaciones primitivas para manejarlos; cómo combinar procedimientos, condicionales, parámetros y cómo utilizar `define` para definir nuevos procedimientos o abstraer el comportamiento deseado de procesos. Ahora se verá cómo representar otros datos y cómo combinar distintas representaciones. Esto permite elevar el nivel conceptual con el que se diseñan problemas.

3.3.1 Definición de estructuras

La definición de estructuras es un mecanismo importante de Scheme para representar todo tipo de objetos con un número fijo de propiedades. Por ejemplo, los puntos en un plano cartesiano, los datos de un estudiante, etcétera.

```
(define-struct nombre (propiedad1 … propiedadn))
```

Cuando DrScheme evalúa una definición de estructura, crea n + 1 operaciones por el usuario, que puede utilizar para programar y crear datos complejos:

1] `make-nombre`, el constructor de estructuras nombre.
2] Para extraer cada una de las propiedades, como la propiedad *i* de una estructura de tipo nombre `nombre-propiedadi`.

Por ejemplo, supóngase que se debe diseñar un proceso que consuma una estructura tiempo y que deba producir el número de segundos que transcurren desde la medianoche hasta la hora que representa la estructura. A continuación se da la definición de la estructura tiempo:

```
(define-struct tiempo (horas minutos segundos))
```

Esto dará origen a cuatro procedimientos: `make-tiempo`, `tiempo-horas`, `tiempo-minutos` y `tiempo-segundos`. Por ejemplo, las 08:25:30 se representan así: (`make-tiempo 8 25 30`). Ahora, pasando al análisis del problema, que no es muy complicado, el número de segundos es una cuestión de multiplicaciones. Una hora tiene 60 minutos, un minuto tiene 60 segundos. La receta de diseño, considerando la estructura tiempo, es:

```
;; Contrato: estructura-tiempo -> número
;; Propósito: Calcular el número de segundos desde la media
;; noche, hasta la hora que representa el argumento.
;; Ejemplo: (tiempo->segundos (make-tiempo 8 25 30))
;; debe producir 30330
;; Definición:
(define (tiempo->segundos t)
 (+ (* (+ (* (tiempo-horas t) 60) ;; # horas en minutos
      (tiempo-minutos t))   ;; minutos
    60)
   (tiempo-segundos t)))
;; Pruebas:
(tiempo->segundos (make-tiempo 8 25 30))
```

Las estructuras permiten ordenar los datos de un problema en forma compacta y útil; se seguirán utilizando más adelante.

Además de los procedimientos para construir y extraer los datos de una estructura, DrScheme crea un predicado[5] para preguntar si un objeto es o no miembro de esa estructura. Éste permite distinguir un ejemplo de una estructura particular:

```
> (tiempo? (make-tiempo 8 25 30))
#t
```

Aprovechando esta nueva herramienta; se van a representar los puntos del plano cartesiano y reescribir el procedimiento de distancia con estos puntos.

```
(define-struct punto (x y))
;; Contrato: distancia2 punto punto -> número
;; Propósito: Calcular la distancia entre dos puntos
;; cualesquiera en el plano
;; Ejemplo: (distancia2 (make-punto -3 1) (make-punto 2 3))
;;     debe producir 5.38
(define (distancia2 punto1 punto2)
   (let ((dif-x (- (punto-x punto1) (punto-x
  punto2)))
       (dif-y (- (punto-y punto1) (punto-y punto2))))
  (sqrt (+ (* dif-x dif-x)
      (* dif-y dif-y)))))
;; Pruebas:
(distancia2 (make-punto -3 1) (make-punto 2 3))
```

Esto ofrece un mayor *encapsulamiento* de la información involucrada, en este caso sobre puntos en el plano, que permite a los programadores ver el programa y entender claramente qué hace, sin necesidad de imaginar que `x1` y `x2` son coordenadas *X* de dos

[5] Predicado es el nombre estándar que reciben los procedimientos cuyos valores de regreso son falso (#f) o verdadero (#t). Scheme ofrece varias primitivas que son predicados: `zero?`, `equal?`, `number?`, entre otros. Por convención, los nombres de los predicados terminan en símbolo de interrogación.

puntos en el plano. Es más fácil dar mantenimiento a este programa o hacer modificaciones que en la versión anterior.

3.3.2 Constructores y selectores

Con la definición de cada estructura se obtiene un *constructor*, un procedimiento que, dados los argumentos adecuados, genera una instancia o ejemplo de la estructura. Esto es útil para todas las estructuras que definimos; sin embargo, existen otras estructuras que se podrían llamar primitivas y que son la forma estándar de almacenar datos complejos en Scheme. Una *lista* en Scheme es la implementación de un tipo de **dato abstracto**; los elementos de la lista, puestos entre paréntesis, constituyen los datos. De esta manera, tanto las listas, que son datos, como los programas lucen igual. Algunos ejemplos de listas son:

Concepto
Un tipo de dato abstracto o ADT (por sus siglas en inglés) especifica un conjunto de datos y las operaciones para manipularlos. Estos tipos de datos son abstractos por su independencia de una implementación particular: no importa cómo se lleven a cabo las operaciones, sólo qué operaciones pueden hacerse y cuáles son sus efectos.

```
(1 2 3 4 5)
(a b c 1 2 3)
("México" "Estados Unidos" "Francia" "España")
```

¿Qué hará DrScheme si se trata de evaluar una de las expresiones descritas arriba? Marca un error. Algo como `"Aplicación de función: se esperaba un procedimiento, pero recibí: 1; los argumentos fueron: 2 3 4 5"`. DrScheme hizo lo que siempre hace para evaluar expresiones: tomar el primer elemento en la lista, que debe ser un procedimiento; evaluar los argumentos y realizar la aplicación.

Pero en este caso no se pretende que evalúe la expresión, sino que la considere en una lista. Se puede realizar lo anterior con un procedimiento que se llama `quote`, que emplea como abreviatura un apóstrofo (`'`). Por ejemplo:

```
> '(1 2 3 4 5)
(1 2 3 4 5)
```

Lo que hace (`'`) es indicarle a DrScheme que no evalúe lo que sigue, sino que lo tome *literal*. Ya se conoce la primera forma de representar listas, pero, ¿cómo crearlas desde un programa? La forma más sencilla es utilizar alguno de los siguientes constructores de listas:

```
> (list 1 2 3 4 5)
(1 2 3 4 5)
> (cons 1 '(2 3 4 5))
(1 2 3 4 5)
> (append '(1 2) '(3 4 5))
(1 2 3 4 5)
```

El constructor más empleado se llama `list` y reúne en una lista todos los elementos que recibe como argumentos. Por otro lado, uno de los constructores más poderosos que se utilizará con frecuencia es `cons` (de *constructor*, en inglés), ya que permite agregar un elemento al inicio de la lista. Recibe dos argumentos: un elemento y una lista. Finalmente, `append` posibilita unir dos listas en una.

Una lista vacía se representa con `'()` y DrScheme ofrece un predicado llamado `empty?` para que regrese `#t` si el argumento que recibe es la lista vacía. Para "desarmar" u obtener elementos particulares de la lista se tienen varias opciones:

```
> (define l (list 1 2 3 4 5))
> (first l)
1
> (second l)
2
> (rest l)
(2 3 4 5)
> (third l)
3
> (nth l 3)
4
```

Para obtener elementos particulares del primero al octavo, DrScheme ofrece procedimientos particulares: `first`, `second`, `third`, `fourth`, `fifth`, `sixth`, `seventh` y `eighth`. En general, el que resulta más útil en los programas es `first`, junto con `rest`, que regresa una lista con todos los elementos de la lista original, excepto el primero. Finalmente, `nth`, devuelve el enésimo elemento en la lista; sin embargo, hay que tomar en cuenta que este procedimiento numera los elementos de la lista desde cero, así `(nth lista 0)` es igual a `(first lista)`.

Ya se tiene suficiente información sobre el ADT, lista en Scheme, ahora se harán algunos programas para probar todo lo analizado. Aunque existe en Scheme un método primitivo llamado `length`, ¿cómo obtener la longitud de una lista? Una alternativa es la siguiente receta de diseño:

```
;; Contrato: longitud lista -> número
;; Propósito: Calcular la longitud (número de elementos)
;; en una lista.
;; Ejemplo: (longitud '(1 2 3 4 5)) debe producir 5
;; Definición:
(define (longitud lista)
  (if (empty? lista)
  0 ;; longitud lista vacía
  (+ 1 (longitud (rest lista)))))
;; Pruebas:
(longitud '(1 2 3 4 5))
(longitud '(a b c 1 2 3))
```

Es importante que primero se pregunte si la lista está vacía para no intentar desarmar con `rest` una lista vacía, lo que produciría que DrScheme marcara un error.

¿Cómo calcular la suma de todos los números en una lista? Se parece mucho al cálculo anterior, sólo que en lugar de sumar uno por uno, se suman todos los valores de los elementos de la lista. Aquí va con receta:

```
;; Contrato: suma lista -> número
;; Propósito: calcular la suma de todos los números
;; en una lista.
;; Ejemplo: (suma '(1 2 3 4 5)) debe producir 15
;; Procedimiento:
(define (suma lista)
```

```
(if (empty? lista)
  0 ;; La suma de cero números
  (+ (first lista) (suma (rest lista)))))
```

3.3.3 Operaciones con listas

Se han revisado ya varios procedimientos para construir y desarmar listas, pero qué sucede si se quiere modificar el contenido de una lista. Con lo que se conoce hasta ahora, es posible recorrer una lista para ordenarla y *crear* una nueva, lo cual es una buena opción para listas pequeñas. En caso de listas grandes se desperdicia mucha memoria al duplicarlas, así que se mostrará algo equivalente a la asignación general, `set-first!` y `set-rest!` que actualizan, de manera destructiva, el primer elemento y el resto (todos menos el primero) de una lista, respectivamente. Existe un procedimiento más que se llama `list-set!` que asigna un valor a una posición particular, por ejemplo:

```
> (define lista '(50 20 5 12 35))
> (list-set! lista 2 1000)
> lista
(50 20 1000 12 35)
```

Se recordará que en el tema 2 se plantearon varios algoritmos de ordenamiento que funcionaban con listas de números. ¿Qué tal si se programa uno de ellos? Por ejemplo, el ordenamiento por combinación, `mergesort`.

En el tema anterior se definió el siguiente algoritmo:

```
OrdenaM (L)
  Si longitud de L es 1
    Regresa L
  De otro modo
     Divide L en 2 del mismo tamaño (aproximadamente), L1 y L2
OrdenaM(L1)
OrdenaM(L2)
Mezcla(L1,L2)
```

Y a continuación se muestra la versión en Scheme, que es idéntica al algoritmo descrito arriba:

```
;; contrato: L[lista de números] -> lista ordenada
;; propósito: ordenar una lista de números, L, de menor a mayor.
;; ejemplo: (mergesort '(5 4 3 2 1)) debe producir (1 2 3 4 5)
;; definición:
(define (mergesort L)
 (if (<= (length L) 1)
   L
   (let ((mitad (quotient (length L) 2)))
    (merge
     (mergesort (list-head L mitad))
     (mergesort (list-tail L mitad))))))
```

Hay algunos detalles que aclarar; por ejemplo, ¿cuál es la mitad de una lista? Eso es fácil, es la longitud de la lista dividida entre 2. Sin embargo, si la lista es de longitud impar esto produciría un número racional; por ejemplo, si la lista es de longitud 5, la mitad es 2.5. Por este motivo se utiliza `quotient` en lugar de `/`, sólo se mantiene el cociente de la división: `(quotient 5 2)` regresa `2`.

Hay otros procedimientos que no se habían visto antes, pero son muy sencillos: `(list-head L N)` regresa los primeros `N` elementos de la lista `L` y `(list-tail L N)` se salta los primeros `N` elementos y regresa el resto de la lista `L`. La primera, `list-head`, no es una primitiva de Scheme, por eso se define aquí, pero la segunda sí es primitiva.

```
(define (list-head L hasta)
 (if (zero? hasta)
   '()
   (cons (car L) (list-head (rest L) (- hasta 1)))))
```

Finalmente, se obtendrá el procedimiento que mezcla las dos sublistas en el algoritmo. La idea esencial es la siguiente: se coloca un dedo índice en la primera posición de cada lista y se comparan; el menor de ellos se pone en el resultado y se avanza el dedo (recortando el problema) en esa lista y se vuelve a intentar. En la figura 6 se muestra un ejemplo con dos listas ordenadas (1, 3, 4) y (2, 5, 6); los dedos índices están marcados por las flechas al inicio de cada lista. Entonces se comparan 1 con 2 y el menor se lleva al resultado, como se muestra en la figura 7 y avanza el índice izquierdo.

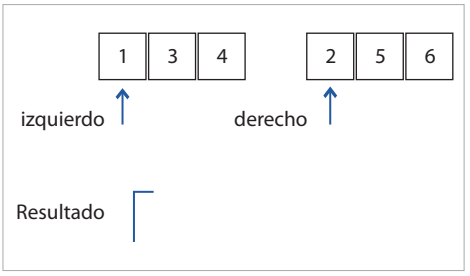


Figura 6. Algoritmo de dos dedos. Estado inicial.

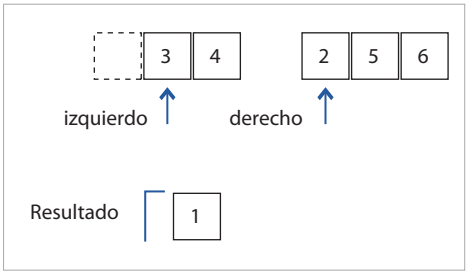


Figura 7. Después de la primera comparación, se avanza el menor al resultado.

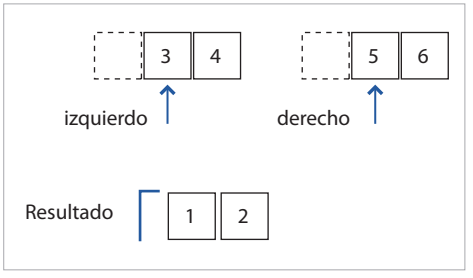


Figura 8. Movemos el menor al resultado y se avanza el índice derecho.

Después de avanzar el índice izquierdo, se compara 4 con 5 y se mueve el 4 al resultado. Cuando avanza el índice izquierdo se puede notar que la lista ya se acabó. En ese momento ya no hay más comparaciones que hacer; el resto de los elementos en la otra lista, la derecha, se agregan al final del resultado. El código es:

```
(define (merge L1 L2)
 (cond ((empty? L1) L2)
    ((empty? L2) L1)
    ((< (car L1) (car L2))
     (cons (car L1) (merge (rest L1) L2)))
    (else
     (cons (car L2) (merge L1 (rest L2))))))
```

A continuación se presentan algunas pruebas del procedimiento:

```
> (mergesort '(5 4 3 2 1))
(1 2 3 4 5)
> (mergesort '(20 3 17 4 5 8 1 65 19 9))
(1 3 4 5 8 9 17 19 20 65)
```

3.3.4 Recorriendo listas

Con procedimientos, como los que se han visto hasta el momento, es muy sencillo recorrer listas; sin embargo, el patrón de recorrido es tan común que Scheme ofrece un procedimiento muy efectivo para *procesar* listas: `map`. Lo que hace `map` es aprovechar que en Scheme los procedimientos son de alto nivel y entonces pueden utilizarse como argumentos. La forma general de `map` es la siguiente:

```
(map procedimiento lista)
```

que aplica el procedimiento a cada uno de los elementos de la lista, de manera consecutiva, y almacena los resultados en una nueva lista, que es su valor de regreso. Los siguientes son algunos ejemplos con `map`; el primero suma 5 a cada uno de los números en la lista:

```
(define (suma-5 n)
 (+ 5 n))
> (map suma-5 '(1 2 3 4 5))
(6 7 8 9 10)
```

El siguiente ejemplo multiplica por dos cada número:

```
(define (x-2 n)
 (* n 2))
> (map x-2 '(1 2 3 4 5))
(2 4 6 8 10)
```

Una tarea interesante es *filtrar* el contenido de las listas. Por ejemplo, ¿cómo hacer para quedarse con todos los números pares o los impares? Hay que recordar que Scheme tiene las primitivas `even?` y `odd?`, así que manos a la obra: en este caso se va a programar un procedimiento que se llama *filtro*, que recibe un predicado y una lista y regresa otra lista que tiene a todos los elementos para los que el predicado regresa `#t`.

```
;; Contrato: predicado lista -> lista
;; Propósito: una lista con todos los elementos
;; para los que el predicado regresa #t.
;; Ejemplos: (filtro odd? '(1 2 3 4 5))
;; debe producir (1 3 5)
;; Procedimiento:
(define (filtro pred lista)
 (cond ((empty? lista) '())
    ((pred (first lista))
     (cons (first lista)
        (filtro pred (rest lista))))
    (else (filtro pred (rest lista)))))
```

Éste es el tercer ejemplo de un procedimiento doblemente recursivo porque hay dos líneas distintas en las que se utiliza el mismo procedimiento. El primer ejemplo fue el de las torres de Hanoi y el tercero el algoritmo de ordenamiento por combinación, `mergesort`. Es importante notar que en ambos usos del procedimiento se *recorta* el problema. La primera vez que se usa es porque se encuentra un elemento que pasó el filtro, en el segundo caso se emplea cuando el elemento actual no pasó. Algunos ejemplos son:

```
> (filtro odd? '(1 2 3 4 5))
(1 3 5)
> (filtro even? '(1 2 3 4 5))
(2 4)
```

Entrada y salida

En la introducción se mostró un procedimiento que resuelve el problema de las torres de Hanoi. En ese programa se aprecia que para desplegar cosas en la pantalla se utiliza `display`, que imprime la representación de cualquier cosa que se le dé como argumento. También se presenta un procedimiento que imprime una nueva línea en la pantalla, se llama `newline`. Algunos ejemplos son:

```
> (display "Hola mundo")
Hola mundo
> (display '(1 2 3 4 5))
(1 2 3 4 5)
> (newline)
>
```

El procedimiento para hacer lo contrario, es decir, leer cosas que entren por el teclado, se llama `read`, pero de éste se hablará más adelante.

3.3.5 Datos simbólicos

Hasta ahora se ha hablado de datos simples: números o datos compuestos como las listas de números. Sin embargo, hay muchos ejemplos de programación que se relacionan con

datos simbólicos; por ejemplo, editores de texto, software para realizar cálculos algebraicos o incluso interpretar y ejecutar programas, como DrScheme.

De hecho ya se mostró una forma de manejar cualquier tipo de datos, con `quote` (`'`); cuando lo utilizas con una lista, DrScheme no interpreta de manera usual la lista, la toma de manera literal. ¿Qué sucedería si se evalúa la siguiente expresión: `'hola`? DrScheme, igual que con las listas, tomará literalmente *hola* y lo presentará en la pantalla; este tipo de elementos se llaman símbolos y Scheme ofrece un predicado para saber si algo es o no un símbolo: `symbol?`

Con esta información se distingue entre símbolos y valores, por ejemplo:

```
> (define uno 1)
> (define dos 2)
> (list uno dos)
(1 2)
> (list ‘uno ‘dos)
(uno dos)
> (list ‘uno uno)
(uno 1)
```

Ahora se mostrarán algunas primitivas más para trabajar con símbolos. Por ejemplo, con `eq?` Se puede saber si dos símbolos son iguales, es como `=` para números. Para mostrar su uso se hará un pequeño programa para buscar un símbolo en una lista; se procederá directo a la receta, porque es un procedimiento similar a los que ya se han revisado con números:

```
;; Contrato: símbolo lista -> booleano
;; Propósito: saber si un símbolo se encuentra en la lista
;; Ejemplos: (está? ‘luis ‘(sergio josé ernesto luis jesús
paco))
;;  debe producir #t
;; Procedimiento:
(define (está? elem lista)
 (cond ((empty? lista) #f)
    ((eq? elem (first lista)))
    (else (está? elem (rest lista)))))
;; Pruebas:
(está? ‘luis ‘(sergio josé ernesto luis jesús paco))
```

Todos los procedimientos que se han realizado hasta el momento son provistos por primitivas de Scheme. Hay un procedimiento llamado `memq`, parecido a `está?` que determina si un símbolo aparece en una lista. Observa, sin embargo, que no utiliza un signo de interrogación al final; esto se debe a que regresa la sublista a partir del elemento que se busca, si es que éste aparece, `#f` en caso contrario. Scheme provee un procedimiento llamado `equal?`, el cual se parece a `eq?` o a `=`, pero funciona con listas y, en general, con todo tipo de elementos del lenguaje.

Puntos más cercanos

Ahora regresando al problema de encontrar los puntos más cercanos en un conjunto de puntos. Para seguir el algoritmo definido en el módulo anterior, se intentará resolverlo por partes: primero, se atenderá la tarea de generar todas las parejas de puntos y, segundo, seguir el algoritmo para comparar las distancias, manteniendo siempre la menor (y los puntos con esa distancia). ¿Cómo encontrar todas las parejas de puntos posibles? Si se piensa que el conjunto es una lista, una alternativa es seleccionar el primer punto y armar parejas de éste con todos los siguientes, después tomar el segundo y armar parejas con todos los que le siguen. ¿Por qué ya no es necesario combinar el segundo elemento con el primero? Se explica con la receta:

```
(define (crea-parejas lista)
 (if (empty? lista)
   '()
   (append (parejas (first lista) (rest lista))
       (crea-parejas (rest lista)))))
```

Se puede notar que el procedimiento auxiliar `parejas` separa el primero de la lista y lo asocia con todos los que le siguen. Como la distancia entre los puntos de una pareja es un dato que se utilizará frecuentemente, es preferible calcularla al momento de crear la pareja y almacenar este valor. Conviene utilizar una nueva estructura:

```
(define-struct pareja (p1 p2 d))

(define (parejas elem lista)
 (if (empty? lista)
   '()
   (cons (make-pareja elem (first lista)
      (distancia2 elem (first lista)))
   (parejas elem (rest lista)))))
```

Nótese cómo al crear una pareja con `make-pareja`, se calcula la distancia entre ambos puntos de la pareja y se almacena. Sólo resta convertir en código el algoritmo que ya se ha explicado en el capítulo anterior:

```
;; Contrato: lista -> imprime el resultado
;; Propósito: regresar los dos puntos más cercanos entre sí.
;; Ejemplos: (puntos-más-cercanos '(p1 p2 ... pN))
;; donde cada punto es una estructura punto.
;; Procedimiento:
(define (puntos-más-cercanos lista)
 (let ((lista-puntos (crea-parejas lista)))
  (if (empty? lista)
    (display "No hay suficientes puntos en la lista")
    (menor-distancia (first lista-puntos) (rest lista-
puntos)))))
```

```
(define (menor-distancia menor lista)
 (if (empty? lista)
   (imprime menor)
   (let ((siguiente (first lista)))
    (menor-distancia
     (if (< (pareja-d siguiente) (pareja-d menor))
       ;; Tenemos un nueva pareja más cercana
       siguiente
       ;; La pareja "menor" sigue siendo la más pequeña
       menor)
     (rest lista)))))
```

En este caso, el procedimiento auxiliar, `menor-distancia`, es el encargado de comparar las distancias ya calculadas, entre todas las parejas de puntos. La idea es muy sencilla: se tienen N parejas en la lista; entonces, se asume que la primera es la de menor distancia y se pasa con el argumento menor de menor-distancia y el resto de las parejas. El procedimiento auxiliar compara la distancia de este supuesto menor con la distancia de la siguiente pareja y, si es menor esta última, se hace la llamada recursiva, pero ahora la segunda pareja es menor, y así sucesivamente. Aquí están unos ejemplos de su ejecución:

```
> (define p1 (make-punto 1 1))
> (define p2 (make-punto 2 2))
> (define p3 (make-punto 5 5))
> (puntos-más-cercanos (list p1 p2 p3))
Distancia más pequeña: P1(1,1) - P2(2,2) = 1.4142135623730951
> (define p4 (make-punto 5.5 6))
> (puntos-más-cercanos (list p4 p2 p3 p1))
Distancia más pequeña: P1(5.5,6) - P2(5,5) = 1.118033988749895
```

Este programa es un ejemplo típico de *construcción modular* de soluciones, similar a la construcción de un edificio, por ejemplo la catedral de la Sagrada Familia, que se ilustra en el inicio de este capítulo. Se trata de un problema complicado, dividido en varios subproblemas más sencillos, cada uno con un procedimiento (como encontrar la menor distancia o crear todas las parejas de puntos) que, al combinarlos, ofrecen la solución final.

3.3.6 Vectores, gráficas y laberintos

Las listas son estructuras abstractas de datos muy útiles y, como se puede apreciar, son sumamente versátiles: han sido utilizadas para almacenar números, símbolos, etcétera. Ahora se presentarán otras estructuras de datos, mismas que se usarán para programar algunos ejemplos.

Vectores y matrices

Un vector es una estructura de datos en la cual se puede acceder a cualquier elemento individual a través de un número entero o índice y, lo más importante, en tiempo cons-

tante. En una lista, acceder a un elemento particular implica *recorrer* la lista hasta encontrarlo, lo que hace atractivos a los vectores para ciertas aplicaciones. En la siguiente tabla se muestran las operaciones principales de esta estructura abstracta de datos.

`(make-vector n)`
`(make-vector n)`
`inicio)`
	Constructor: crea un vector de tamaño `n` y, si se incluye, todas las posiciones contienen `inicio` o bien cero (0).

`(vector-ref vector` `pos)`	Obtener el valor en la posición `pos` del vector.
`(vector-set! vector` `pos obj)`	Almacena el objeto `obj` en la posición `pos` del vector.
`(vector-length vector)`	Regresa el tamaño del vector.
`(vector->list` `vector)`	Convierte el vector en una lista.

¿Con las primitivas para vectores se puede definir una matriz bidimensional, por ejemplo de 5 × 5 elementos? Sí, se puede crear un vector de tamaño cinco y después utilizar `vector-set!` y asignar otros vectores en cada posición, ésta es una tarea fácil para un ciclo:

```
(define i 0)
(define v (make-vector 5))
(while (< i 5)
        (vector-set! v i (make-vector 5))
        (set! i (+ i 1)))
```

Al final, `v` es una matriz de 5 x 5 elementos, así que es posible almacenar y recuperar objetos de cualquiera de sus 25 posiciones, por ejemplo:

```
> v
#5(#5(0) #5(0) #5(0) #5(0) #5(0))
> (vector-set! (vector-ref v 2) 0 “Una cadena”)
> (vector-set! (vector-ref v 2) 1 2500)
> v
#5(#5(0) #5(0) #5(“Una cadena” 2500 0) #5(0) #5(0))
> (vector-ref (vector-ref v 2) 0)
“Una cadena”
> (vector-ref v 2)
#5(“Una cadena” 2500 0)
```

Como se observa, la notación `#N` y luego una lista son utilizadas por DrScheme para indicar que se trata de un arreglo de tamaño `N`. La lista que sigue está conformada por los elementos del arreglo y si todas las posiciones tienen el mismo valor, sólo repite dicho valor. Así, `#5(0)`se traduce en un arreglo de cinco posiciones, cada una con un cero. En los ejemplos, `v` es un vector de tamaño cinco pero cada elemento del arreglo es a su vez otro arreglo distinto de cinco posiciones cada uno. Un ejemplo interesante con vectores es el ordenamiento por selección que se definió en el capítulo anterior:

1] Encontrar el menor elemento en la lista.
2] Intercambiarlo con el primero de la lista.
3] Repetir 1 y 2 para el resto de la lista empezando en la segunda posición.

Sólo que, en lugar de una lista, se va a utilizar un vector. En el DVD que acompaña este libro se encuentra una versión con listas; sin embargo, es un poco más complicada por diferencias fundamentales entre las estructuras lista y vector. Siguiendo el principio de construcción en bloques, es claro que el primer paso constituye en sí mismo un procedimiento y que se pueden juntar 2 y 3 en un segundo procedimiento que resuelve el problema. Aquí está la versión en código:

```
;; contrato: v[vector de números] -> vector ordenado
;; propósito: ordenar el vector de manera ascendente,
;; utilizando el algoritmo por selección.
;; ejemplo: (ordenar #(5 4 3 2 1)) debe producir #(1 2 3 4 5)
;; definición:
(define (ordenar v)
 (ordenar-por-inserción v 0 (vector-length v)))

(define (ordenar-por-inserción v desde hasta)
 (if (= desde hasta)
   v ;; Hemos terminado, v está ordenado
   (let ((menor (posición-del-menor v desde (+ desde 1) hasta))
      (primero (vector-ref v desde)))
    ;; Primero, ponemos el menor en su posición:
    (vector-set! v desde (vector-ref v menor))
    ;; Y al número que estaba ahí, en la posición del menor:
    (vector-set! v menor primero)
    ;; Y ordenamos el resto del vector:
     (ordenar-por-inserción v (+ desde 1) hasta))))
```

La primera definición, `ordenar`, facilita llamar al procedimiento que lleva a cabo el ordenamiento, `ordenar-por-inserción`. Los pasos 2 y 3 del algoritmo de ordenamiento arriba, se definen en el procedimiento auxiliar, `ordenar-por-inserción`, que recibe el vector y algunos datos fijos acerca del vector: `desde` y `hasta`. En el primer paso `desde` vale 0 y se aumenta en 1 conforme avanza el programa, mientras que `hasta`[7] nunca cambia y es el tamaño del vector. El paso 1 del algoritmo lo realiza el procedimiento `posición-del-menor`:

```
(define (posición-del-menor v menor sig v-len)
 (cond ((= sig v-len) menor) ;; última posición
    ((< (vector-ref v sig) (vector-ref v menor))
     ;; nuevo menor en la posición "sig"
     (posición-del-menor v sig (+ sig 1) v-len))
     (else
     ;; el número en "menor", sigue siendo
        el más pequeño
     (posición-del-menor v menor (+ sig 1) v-len))))
```

Curiosidades
En la película *El cubo* (1997), un hombre despierta completamente solo dentro de un cuarto en forma de cubo. No tiene idea de dónde se encuentra ni cómo llegó ahí. El cubo es perfecto; en cada uno de sus seis lados hay una compuerta para salir de éste. Al abrir una de las compuertas, el hombre se da cuenta de que ésta lo lleva a otro cubo idéntico al primero. Paulatinamente, se percata de que está encerrado en un laberinto gigantesco en forma de cubos interconectados, que además se mueven. El hombre debe encontrar la salida o morir, ya que los cubos no tienen agua ni comida. Éste es un ejemplo de un laberinto tridimensional móvil, entre los tantos laberintos dificilísimos de resolver que se han inventado. Hay más ejemplos en <http://www.clickmazes.com/mazes/ixmaze.htm>.

Este procedimiento encuentra al menor en el vector, entre las posiciones `menor` y `v-len`, que es la longitud del vector. Por la recursividad en `ordenar-por-inserción`, el menor va creciendo conforme se ordena el vector. A continuación unas pruebas:

```
> (ordenar #(1 2 3 4 5 4 3 2 1))
#(1 1 2 2 3 3 4 4 5)
> (ordenar #(100 80 50 60 10))
#(10 50 60 80 100)
```

Gráficas y laberintos

Existen muchos tipos de laberinto. En computación se estudia la creación de laberintos y cómo encontrar la salida. La relación con gráficas es íntima debido a que la forma de los caminos en sí, dentro de un laberinto, no es importante para encontrar la salida; lo que importa es qué caminos existen y de qué manera están interconectados. Algunos son más difíciles de resolver que otros, pero aquí la discusión se centrará en un tipo particular conocido como *laberinto perfecto*. Este tipo de laberintos debe su nombre a que dos posiciones en el laberinto (`entrada-salida`) están conectadas entre sí por un camino único (figura 9).

Un laberinto perfecto no tiene secciones inaccesibles, no tiene rutas circulares y no tiene áreas abiertas.

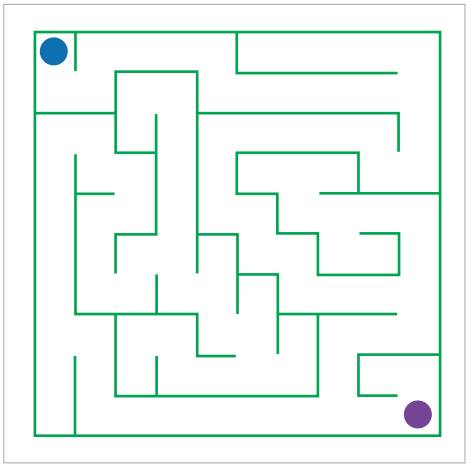


Figura 9. Ejemplo de laberinto perfecto.

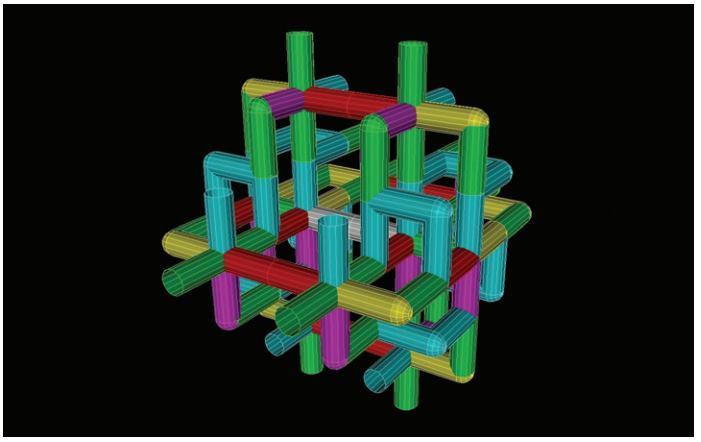


Figura 10. Ejemplo de un laberinto en tres dimensiones.

3.3.7 Construcción de laberintos perfectos

Ahora se utilizará una estructura gráfica para generar un laberinto perfecto y, más adelante, se usará la misma estructura y un algoritmo importante para encontrar la salida. ¿Qué hace fáciles a los laberintos perfectos? Que todas las posiciones o celdas se pueden ver como un cuadrado con dos posibles paredes: derecha y abajo. El marco del laberinto se convierte en paredes para todas las celdas en las orillas. Entonces, la estrategia de construcción es muy sencilla:

1] Seleccionar una celda como entrada y otra como salida del laberinto. Por convención, la entrada será la celda en la esquina superior izquierda y la salida en la esquina inferior derecha.
2] Al inicio, todas las paredes del laberinto están puestas. Es decir, no existe ningún camino, cada celda del laberinto tiene cuatro paredes.
3] Ubicarse en la entrada (funciona también si elegimos una celda aleatoria).
4] Localizar de manera aleatoria una celda vecina no visitada aún.
5] Si existe tal celda vecina, moverse hacia ella, tirando la pared entre las celdas. Si no hay celda vecina, regresar a la celda anterior.
6] Repetir los pasos 4 y 5 hasta que se hayan visitado todas las celdas del laberinto.

Los pasos 3 a 6 constituyen un algoritmo muy famoso de exploración general de gráficas, conocido como búsqueda a profundidad (`dfs`, por sus siglas en inglés). Una vez más, se va a descomponer el problema completo en varios problemas menores. Se comenzará por el procedimiento principal: `crea-laberinto` y las variables globales para almacenar el laberinto, el número de renglones y el número de columnas.

```
(define renglones 0)
(define columnas 0)
(define laberinto #f)
(define-struct celda (x y derecha abajo visitada))

(define (crea-laberinto rens cols)
 (let ((lab (make-vector rens))
    (i 0))
  (while (< i rens)
      (let ((columna (make-vector cols))
         (j 0))
       (while (< j cols)
          (vector-set! columna j (make-celda i j #t #t #f))
          (set! j (+ j 1)))
       (vector-set! lab i columna))
      (set! i (+ i 1)))
  ;; Se almacena la información en las variables globales:
  (set! renglones rens)
  (set! columnas cols)
  (set! laberinto lab)
  ;; Ahora generan las rutas:
  (crea-rutas (vector-ref (vector-ref laberinto 0) 0) #f)))
```

¿Qué puntos de la estrategia se han definido? 1, 2 y 3, por supuesto. Los dos ciclos `while` anidados permiten recorrer todas las celdas del laberinto e inicializar cada una, de modo que tengan ambas paredes y estén marcadas como *no visitadas*. La estructura celda almacena varias cosas: su posición en el laberinto (cuadrícula), si las paredes derecha e inferior aún están de pie y, finalmente, un campo que se llamará *visitada*, que se utilizará en el algoritmo de búsqueda a profundidad para saber si el algoritmo ya pasó por esta celda. La siguiente parte importante es el procedimiento `crea-rutas`, implementado por los puntos 3 al 6:

```
(define (crea-rutas celda dir)
 ;; Primero, se marca la celda actual como visitada.
 (set-celda-visitada! celda #t)
 ;; Se verifica la dirección y se tira la pared en esa
dirección.
 (if dir
   (tira-pared celda dir))
 ;; Después, se recorren las posibles direcciones, tirando
paredes.
 (dfs celda (posibles-direcciones (celda-x celda) (celda-y
celda))))

(define (dfs celda direcciones)
 (if (null? direcciones)
   '()
   (let ((vecino (obtener-vecino celda (car direcciones))))
    (if (not (celda-visitada vecino))
      ;; Vecino no visitado, se recorre primero:
      (crea-rutas vecino (car direcciones)))
    ;; Y luego se recorren las demás direcciones (backtrack).
    (dfs celda (rest direcciones)))))
```

Los comentarios explican la función de cada parte de los procedimientos. Sin embargo, un concepto nuevo e importante aquí es que `crea-rutas` y `dfs` son procedimientos mutuamente recursivos; es decir, hacemos llamadas entre ellos conforme avanzamos. Aquí se explica la lógica de esto:

- `Crea-rutas` se encarga de marcar una celda, que nunca había sido visitada antes, como visitada. Después, localiza todas las posibles direcciones en las que nos podemos mover desde esta celda y llama a `dfs` para recorrer tales rutas.
- `dfs`, como ya se dijo, recorre todas las direcciones posibles desde la celda actual. Sin embargo, cuando selecciona de manera aleatoria una celda vecina que no ha sido visitada… ¿qué hacer con una celda nunca antes visitada? ¡Exacto!, se envía la celda a `crea-rutas`.

Las definiciones auxiliares que se verán a continuación cumplen funciones importantes y se puede pensar en ellas como pequeños bloques de construcción: `posibles-direcciones`, `tira-paredes` y `obtener-vecino`. La segunda y tercera son muy sencillas. Continuando con la tercera: ¿cómo obtener el vecino de una celda en una dirección determinada? Si es el vecino a la izquierda, se resta 1 a la columna de la celda actual;

por ejemplo, si se está en la celda (5,3), el vecino de la izquierda está en el mismo renglón, 5, pero en la columna anterior, o sea, 2. Y algo similar para las demás direcciones. Obsérvese con detalle:

```
(define (obtener-vecino celda dir)
 (cond ((eq? dir 'izquierda)
     (vector-ref
      (vector-ref laberinto (celda-x celda))
      (- (celda-y celda) 1)))
    ((eq? dir 'derecha)
     (vector-ref
      (vector-ref laberinto (celda-x celda))
      (+ (celda-y celda) 1)))
    ((eq? dir 'arriba)
     (vector-ref
      (vector-ref laberinto (- (celda-x celda) 1))
      (celda-y celda)))
    (else
     (vector-ref
     (vector-ref laberinto (+ (celda-x celda) 1))
      (celda-y celda)))))
```

¿Cómo tirar una pared? Se asigna `#f` a la pared en cuestión. Hay un par de puntos que considerar: primero, recuérdese que cada celda se encarga de su pared derecha y su pared inferior; segundo, se mueve a la celda vecina y desde ahí se tira la pared. Por ejemplo, supóngase que se está en la posición (3,3) y se hace un movimiento hacia la derecha, a la celda ubicada en (3,4). ¿Qué pared se debe tirar? La pared entre las columnas 3 y 4 en el renglón 3. A continuación se presenta la definición:

```
(define (tira-pared celda dir)
 (cond ((eq? dir 'izquierda)
     (set-celda-derecha! celda #f))
    ((eq? dir 'derecha)
     (set-celda-derecha! (obtener-vecino celda
        'izquierda) #f))
    ((eq? dir 'arriba)
     (set-celda-abajo! celda #f))
    (else
     (set-celda-abajo! (obtener-vecino celda
        'arriba) #f))))
```

Finalmente, `posibles-direcciones` es un procedimiento con un pequeño truco, ya que se necesita seleccionar de manera aleatoria a los vecinos. Entonces, lo que se hace primero es generar las direcciones válidas desde la posición actual; por ejemplo, en la posición (0,0), las únicas direcciones válidas son hacia la derecha y abajo, pero en (1,1), es posible moverse hacia las cuatro direcciones. Aquí está la primera parte:

```
(define (posibles-direcciones x y)
 (random-list
```

```
(filter symbol?
    (list (and (> y 0) 'izquierda)
       (and (< y (- columnas 1)) 'derecha)
      (and (> x 0) 'arriba)
       (and (< x (- renglones 1)) 'abajo)))))
```

Y ahora, como se quieren *reordenar* esas direcciones de manera aleatoria, se llama a `random-list`, que genera un número aleatorio entre 0 y el número de direcciones posibles y con ese valor *rota* la lista, para darle un orden distinto. Por ejemplo, si la lista de direcciones posibles es `'(izquierda derecha abajo)` y el número aleatorio es 2, es decir, la última posición de la lista, entonces `random-list` regresa `'(abajo izquierda derecha)`. Este procedimiento da variedad a los laberintos, lucen más bellos y complicados.

```
(define (random-list L)
 (let* ((len (length L))
    (primera (random len))
    (i 0)
    (resultado '()))
  (while (< i len)
     (set! resultado
        (append (list
            (list-ref L (módulo (+ primera i) len)))
            resultado))
       (set! i (+ i 1)))
  resultado))
```

A continuación se muestran algunos ejemplos del programa. Estas versiones gráficas, así como el código que las genera vienen en el DVD que acompaña este libro. Al abrir el archivo "laberinto" en DrScheme y hacer clic en Ejecutar, aparecerá un diálogo que solicita el número de renglones y columnas del laberinto. Aquí se muestran uno de 10 × 20 y otro de 20 × 10, donde el punto azul indica la entrada y el morado la salida del laberinto.

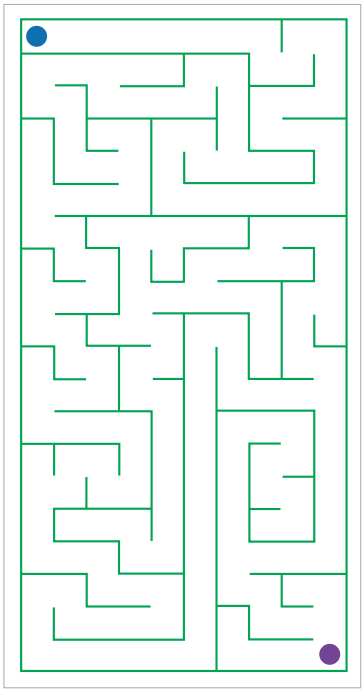


Figura 12. Laberinto de 20 × 10.

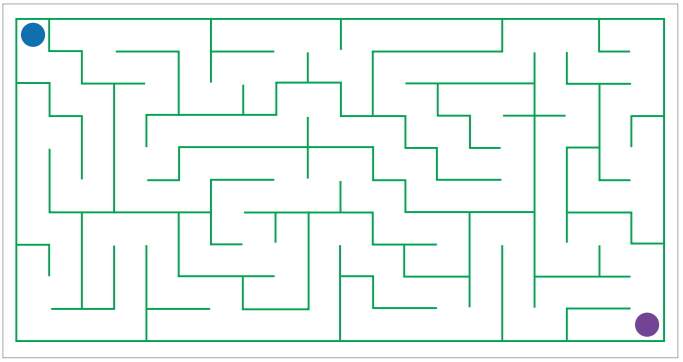


Figura 11. Laberinto de 10 × 20.

3.4 RESUMEN

La computadora ejecuta programas que manipulan datos siguiendo los pasos de uno o más algoritmos que interactúan entre sí. Estos programas pueden tener desde unas cuantas instrucciones hasta millones de ellas, formando así enormes catedrales virtuales. Para aprender a programar hay que conocer un lenguaje de programación, así como técnicas para armar los bloques que componen el programa. De esta manera se logra una construcción sólida, correcta y eficiente.

TEMA

4

© *Mountain.*

El cambio de átomos a bits es irrevocable e imparable.
NICHOLAS NEGROPONTE, 1995.

Omnibus ex nihil ducendis sufficit unum.
Uno basta para derivar el todo de la nada.
GOTTFRIED WILHELM LEIBNIZ, SIGLO XVII.

Lo principal es la sabiduría; adquiere sabiduría, y con todo lo que obtengas adquiere entendimiento.
PROVERBIOS 4:7.

4.1 LOS MIEDOS DE LA FUTURA SUEGRA DE ARCADIO

Después de casi una hora de que su suegra se había encerrado en el despacho, Arcadio escucha sus llamados desesperados.

—¡Esta máquina no sirve para nada!

—¡Voy, suegrita! Espéreme que estoy hablando con la tía Carmelita de Colombia, que también tiene un problema técnico.

—¿Cuál problema? El problema lo tengo yo: llevo una hora tratando de enviarle un fax, meto y meto la hoja, y ¡esta máquina endemoniada me la regresa!

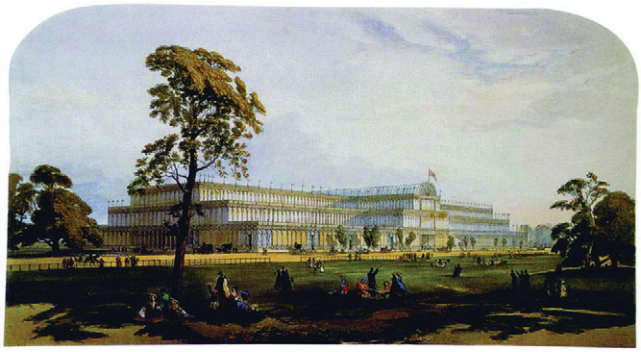
—Pues precisamente me dice la tía que para qué le estamos enviando la misma hoja tantas veces.

—¿¡Cómo que tantas veces!? No se la he logrado enviar ni una sola vez. ¿No ves que acá tengo en mi mano la hoja?

Esta historia es ilustrativa de la elusiva noción de *información*. ¿Qué es exactamente lo que se hace al enviar un documento por fax? En principio, una máquina de fax podría lograr tal precisión, que el documento recibido y el enviado fueran indistinguibles. Además, es muy diferente a enviar una carta; como bien dice la suegra de Arcadio, después de todo ella sigue con el documento en la mano, ¿no? Parecería que lo que se envió es algo intangible, sin olor ni sabor, que además tiene una maleabilidad sin fin, ya que puede ir cambiando de forma y naturaleza suavemente, sin perder su esencia, al pasar de un medio a otro, como es el papel, la luz (ya que la máquina del fax de alguna manera lo lee mediante luz, sin necesidad de tocarlo), la electricidad de la línea telefónica, las ondas electromagnéticas de la transmisión vía satélite, y de regreso hasta llegar a las manos de Carmelita, en forma de tinta sobre papel.

Eso que se transmitió a enormes velocidades y que el ser humano no se imaginó que fuera posible, por miles de años hasta el siglo XIX, es *información*. Una cosa extraña que, a la vez que está hecha de nada, puede ser valiosísima: un empresario teme mucho menos que le roben su *laptop*, a que le roben la información que contiene ésta acerca de sus patentes y carteras de clientes. Y, como veremos en este módulo, esta "nada" se puede medir y cambiar de forma, se puede comprimir, expandir, esconder, compartir, fortalecer y procesar. Esta "nada" puede serlo "todo", desde números y listas de nombres, hasta diccionarios, sofisticadas estrategias de ajedrez, códigos genéticos gigantescos o enormes sinfonías musicales. La información es la base del conocimiento y está íntimamente relacionada con el aprendizaje. No es casualidad que los computólogos la estudien desde diversos ángulos ni que existan importantes aplicaciones de cómputo que deben aprender para resolver mejor sus tareas.

Curiosidades
La invención del fax se le atribuye a Alexander Bain, quien la presentó en la Gran Exibición de 1851 en Londres, la primera de la afamada serie de Ferias Mundiales donde se presentaban los inventos y novedades culturales del año. El Palacio de Cristal fue creado especialmente para este evento, donde entre muchas otras cosas se presentaron los primeros baños públicos y un barómetro hecho de sanguijuelas.



Palacio de Cristal | *© Anónimo.*

4.2 SÍMBOLOS

Nunca deja de sorprender el hecho de que toda la amplia variedad de elementos almacenados en una computadora —canciones, un video de una cantante favorita, un programa para jugar ajedrez, un diccionario de español, las fotos de los viajes, correos electrónicos, dibujos y programas que ayudan a hacerlos, listas de clientes, la contabilidad de una empresa y tantas otras cosas— no sea más que secuencias de 0 y 1 llamados *bits*. No es que la computadora esté llena de ceros y unos, ¿cómo podría ser? No se trata de un circulito y un palito; son sólo los símbolos que se usan para representar cada uno de los dos estados en los que puede encontrarse un bit. Además, de alguna manera, en el interior del cerebro también se tienen todas esas cosas: poemas, canciones, reglas de multiplicación y todo lo demás. Es decir, todos esos datos están representados de alguna manera en la computadora y en el cerebro. Se verá cómo se pueden representar diversos tipos de datos,

y cómo hacerlo de manera *eficiente*. Pero también es importante hacerlo de manera *robusta* si no se quiere que una pequeña falla, un error en un solo cambio de 0 a 1, cambie el significado del dato. Y a veces se desea representar datos de manera *segura*, es decir, que nadie además de las personas designadas para ello los puedan ver.

4.2.1 Símbolos, palabras, mensajes

Antiguas civilizaciones han usado campanas en ritos religiosos, aun antes del desarrollo del lenguaje escrito. En su forma básica, una campana es un mecanismo que permite producir un solo sonido al ser golpeada. Se puede representar el hecho de golpear una campana mediante un símbolo, por ejemplo "∩" . Así, desde la Edad Media se tocaba varias veces una campana para indicar la hora. Por ejemplo, "∩∩∩∩" podría significar que "son las cuatro de la tarde". En efecto, en inglés, la palabra "reloj", o sea *clock*, deriva del holandés *klok* que significa "campana". Para no usar un símbolo raro, que no sepamos cómo pronunciar, cambiemos "∩" por "1", pues así es más fácil escribir "∩∩∩∩" como "1111" y pronunciar "cuatro unos" sin que esto implique que se está haciendo referencia a un número, sino solamente a un símbolo.

Faro | © *Emad Victor Shenada.*

Curiosidades
Ésta es una reconstrucción detallada en tres dimensiones por computadora basada en un estudio reciente (2006) del Faro de Alejandría, que fue construido en el siglo III antes de nuestra era en la isla de Faros en Alejandría, Egipto. El faro, que algunos estiman de más de 180 metros de altura —lo que la habría hecho la construcción más alta del mundo hasta el siglo XIV—, tenía un espejo en la cúspide para reflejar la luz del sol durante el día, y de noche, una fogata. Se cree que la señal de luz emitida desde el faro para ayudar a los barcos en la navegación se podía ver hasta 56 kilómetros desde la costa. Hoy en día siguen en operación menos de 1 500 faros en el mundo, que asisten en la navegación para identificar costas peligrosas y entradas seguras a puertos.

El sonido de la campana es una forma de transmitir símbolos. Evidentemente, lo esencial es cuántas veces se toca la campana, y no la campana misma, ni su tono o intensidad. Se puede lograr el mismo efecto de otras maneras. Desde los tiempos de la *Ilíada* de Homero se usaban fogatas para anunciar el arribo de barcos. Para indicar que un enemigo se aproxima, por ejemplo, se podría hacer aparecer una luz tres veces, o sea "111".

Si a una secuencia de símbolos se le llama palabra, entonces cada palabra puede estar representando algún mensaje. La palabra "111" indica que un enemigo se aproxima. Así, aparentemente basta con un solo símbolo para representar cualquier cosa. Leibniz dijo: "Con uno, todo puede ser obtenido de nada." Es decir, dado un conjunto de mensajes que se quieren representar, y posibles símbolos a usar, se elije un conjunto de palabras con esos símbolos como alfabeto, y una función de codificación que indica qué mensaje corresponde a qué palabra.

Los símbolos son abstractos; para comunicarlos o almacenarlos es necesaria alguna implementación física. Por ejemplo, el "111" se puede implementar usando sonido, pulsos eléctricos, luz o escribiendo tres palitos en forma de "uno" en un papel, como en la siguiente figura:

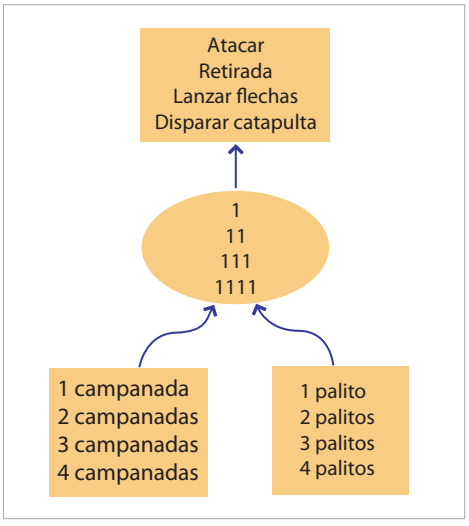


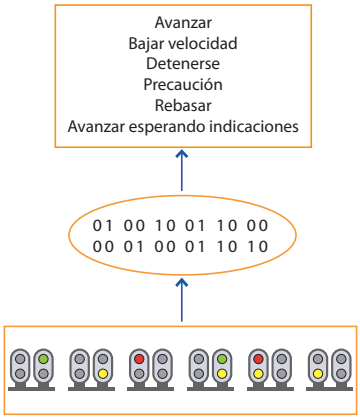
Figura 1. Ejemplos de uso de diferentes símbolos para representar un conjunto de mensajes.

También se pueden representar números usando un solo símbolo. Para representar el 246 se podría simplemente utilizar la palabra de 246 unos, y una manera de transmitirla sería tocando la campana 246 veces. ¿Existe la posibilidad de representar todos los números, negativos y positivos, inclusive el cero, mediante un solo símbolo? Claro, sólo es cuestión de tener un acuerdo acerca de qué codificación usar. Como ejemplo véase la tabla 1:

Palabra	Número representado
1	0
11	1
111	-1
1111	2
11111	-2
111111	3
Etc.	

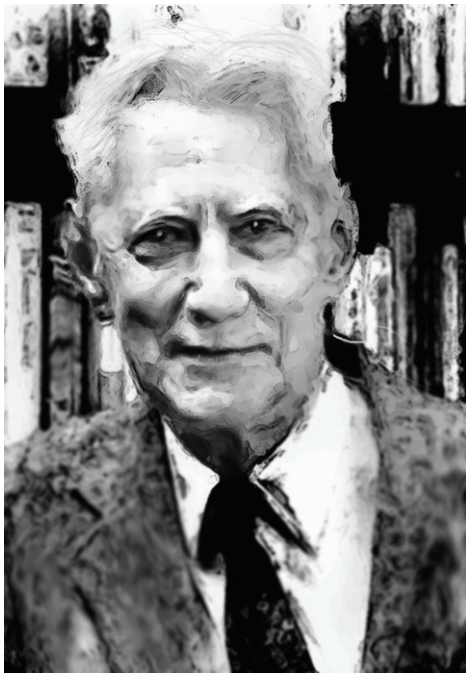
Tabla 1. Lista de palabras representadas por números.

Curiosidades
Un tipo de semáforo de ferrocarril tiene cuatro luces, con las cuales se envían seis mensajes con los siguientes significados:



Si con un solo símbolo se puede representar un conjunto de mensajes cualesquiera, con más de uno es posible hacerlo aún mejor.

El problema con todos los mecanismos usados antes del siglo XIX —como señales de humo, fogatas, sonidos de cuernos, etc.— es que estaban limitados a un conjunto predeterminado de mensajes, como "viene el enemigo". Parece muy fácil representar conjuntos arbitrarios de mensajes inclusive con un solo símbolo, pero no es muy eficiente. Por ejemplo, si se utiliza una campana, tomaría nueve veces el tiempo para representar el número 9 que para representar el 1. En cambio, con sólo cuatro luces es posible producir 16 mensajes diferentes, usando todas las combinaciones posibles, prendiéndolas y apagándolas.



Claude E. Shannon (1916-2001) En un célebre artículo publicado en 1948, Shannon inauguró el área de la computación hoy denominada Teoría de la Información. Acuñó los términos y los conceptos precisos de cantidad de información y de entropía. Shannon fue el primero en usar la palabra bit. Aunque se le atribuye su origen a John Tukey, quien, en un trabajo en los Laboratorios Bell el año anterior, contrajo binary digit a simplemente bit. Estos laboratorios dieron lugar a inventos tan importantes como el transistor (1947), el láser (1958), la celda solar (1954) y el lenguaje C junto con el sistema Unix (1969-1972).

4.2.2 Bits

Cuando se piensa en usar dos símbolos, es común usar como convención 0 y 1. Entonces las palabras obtenidas son secuencias como "1011001", y a cada uno de sus elementos se les conoce como *bit*. Como dice Nicholas Negroponte en *Being Digital*, "un bit no tiene color, tamaño o peso, y puede viajar a la velocidad de la luz. Es el elemento más pequeño

en el ADN de la información. Es un estado de ser: prendido o apagado, verdadero o falso, arriba o abajo, adentro o afuera, blanco o negro. El significado del 1 o el 0 es un asunto independiente".

¿Para qué usar dos símbolos si con uno basta? ¿Para qué construir una campana china Zhong, como la de la imagen al inicio de este tema, que puede producir dos sonidos diferentes, quizás el bit más antiguo? ¿Por qué no usar tres símbolos o más? Por ejemplo, el alfabeto español consta de 29 símbolos, y el sistema numérico utiliza 10, los dígitos del 0 al 9. Posteriormente se verá que es más eficiente usar más de un símbolo que sólo uno; sin embargo, cuando se usa más de un símbolo, la diferencia en la eficiencia entre usar un número mayor o menor de símbolos ya no es tan grande como en el primer caso.

El papel fundamental del bit se debe al hecho de que es la manera más simple, entre las eficientes, de codificar información. Es decir, al usar un solo símbolo se requerirán palabras de longitud n para representar n mensajes posibles, mientras que si se utilizan bits sólo se requieren palabras de longitud igual al logaritmo en base 2 de n para representarlas, ya que si $x = \log_2 n$ entonces las palabras de longitud n serán las 2^x combinaciones posibles de 0 y 1. Esto implica una ganancia exponencial en la economía de representación, debido a que por cada bit más que se emplea se duplica el número posible de mensajes: todos los que ya se tenían, seguidos de un 0 y también seguidos de un 1. Téngase en cuenta que el logaritmo con base b de x es n, es decir, $\log_b x = n$ significa que $b^n = x$. Y se tiene la siguiente relación:

$$\frac{}{\log_b x} = \log_a b$$
$$\log_b x$$

Por otro lado, si se pasa de bits a un alfabeto de b símbolos, $b > 2$ y de palabras de longitud log con base 2 a palabras de longitud log con base b, la ganancia sería de log con base 2 de b. Lo cual no hace mucha diferencia, ya que log con base 2 de b solamente es una constante. Por ejemplo, sería posible representar las 29 letras del alfabeto con cinco bits, ya que con éstos se obtienen 32 combinaciones posibles de 0 y 1. De manera que un texto en español se expandiría cinco veces al pasarlo a un alfabeto binario.

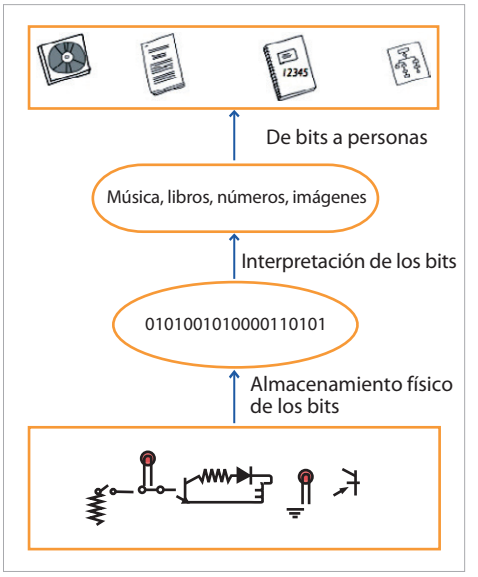


Figura 2. Secuencia para almacenar o transmitir datos.

A continuación se verá cómo representar diversos tipos de datos mediante secuencias de bits, como datos numéricos, letras, imágenes. Cabe señalar que esto es independiente de la manera de almacenar o transmitir estos datos físicamente. El computólogo estudia formas de representar datos mediante bits de manera eficiente, robusta y segura. Se encarga de las dos capas intermedias de la siguiente figura. Es asunto de ingenieros electrónicos diseñar circuitos para almacenar los datos, lo cual se representa en la capa inferior de la figura, o para comunicarlos a personas u otras computadoras, diseñar altavoces, pantallas, antenas y transmisores. Típicamente, en una computadora se representan los 0 y 1 en circuitos electrónicos, usando dos niveles de voltaje diferentes.

Curiosidades

En China, ya para el siglo IV antes de nuestra era se había desarrollado la filosofía bipolar del mundo basada en la interacción de dos energías: la femenina, *yin*, representada por dos rayas (- -); y la masculina, *yang*, representada por una raya continua (—). Los elementos de la naturaleza se representan mediante combinaciones de estos dos tipos de líneas. Por ejemplo, un yin encima de un yang y hasta abajo un yin representan al viento, el suroeste y el final del verano. El sistema se describe en el antiguo libro chino *I Ching*. Cuando Leibniz se enteró de las 64 figuras formadas por los hexagramas del *I Ching*, de las cuales las primeras cinco (A, B, C, D, E) se ilustran abajo, se dio cuenta inmediatamente de que no son más que bits: podríamos decir que el 0 representa - - y el 1 representa — y a los ojos de Leibniz representaban (erróneamente) números, lo cual lo llevó a exclamar la cita del inicio de este módulo: *omnibus ex nihil ducendis sufficit unum* (uno basta para derivar el todo de la nada).

A = 0 0 0 0 0 0 = 0
B = 0 0 0 0 0 1 = 1
C = 0 0 0 0 1 0 = 2
D = 0 0 0 0 1 1 = 3
E = 0 0 0 1 0 0 = 4

4.3 REPRESENTANDO EL MUNDO MEDIANTE BITS

¿Cómo representar números, letras, imágenes y sonidos mediante bits? En las siguientes líneas se tratará de argumentar que cualquier cosa es en principio representable mediante bits.

4.3.1 Representando números

Los números se podrían representar de muchas maneras diferentes usando bits. Como ejemplo, podemos tomar el conjunto de todas las palabras de 0 y 1, y arbitrariamente asignarle a cada número entero una de ellas. A continuación se verá la representación binaria que permite que se puedan manipular y hacer operaciones aritméticas eficientemente.

Midiendo ingredientes

Vale la pena recordar la historia de Úrsula y su madre cocinando galletas. Quedó pendiente ver cómo había hecho Úrsula para medir los ingredientes usando sólo las pocas cucharas medidoras de que disponía: de una, dos, cuatro y ocho onzas. Para recordar las cantidades basta echar una mirada a las dos columnas de la izquierda de la tabla 2.

Ingrediente	Cantidad	8 Oz	4 Oz	2 Oz	1 Oz
Avena	15 Oz	✓	✓	✓	✓
Harina	12 Oz	✓	✓	✗	✗
Chispas de chocolate	10 Oz	✓	✗	✓	✗
Nueces trituradas	9 Oz	✓	✗	✗	✓
Mantequilla	7 Oz	✗	✓	✓	✓
Azúcar mascabado	6 Oz	✗	✓	✓	✗
Azúcar	5 Oz	✗	✓	✗	✓
Chocolate amargo	3 Oz	✗	✗	✓	✓

Tabla 2. Lista de ingredientes y las cucharas usadas para medirlos.

Lo primero que midió Úrsula fue la avena: 15 onzas. Primero llenó la cuchara de ocho onzas y luego la de cuatro, tenía entonces 8 + 4 = 12 onzas; sólo faltaban tres, que se podrían medir exactamente usando las dos cucharas restantes. Es decir, usando una vez cada cuchara se podían medir las 15 onzas, lo que sin duda era mejor que usar 15 veces la cuchara de una onza, o tres veces la pareja de cuatro y de una. Para medir las 12 onzas de harina sólo tenía que hacer parcialmente lo que había hecho antes, usar la de ocho y la de cuatro. Las 10 onzas de chispas de chocolate se podían medir usando la de ocho y la de dos. En la tabla 2 también se muestran las combinaciones de cucharas que usó Úrsula para medir todas las cantidades en onzas. De hecho, Úrsula se dio cuenta de que cualquier cantidad entera de onzas desde cero hasta 15 podía ser medida usando sólo las cuatro cucharas disponibles a lo más una vez cada una.

En general, la estrategia de Úrsula consistió en usar la cuchara más grande posible, cuya capacidad fuera menor o igual a la cantidad que necesitaba medir. Si era igual, el asunto estaba resuelto, si no, entonces quedaba algo por medir y procedía igual, eligiendo la cuchara con la mayor capacidad posible que no excediera la cantidad necesaria.

Curiosidades
No todos los sistemas numéricos que ha usado la humanidad a lo largo de la historia han sido posicionales. Los sistemas, como el romano o el egipcio, en los que un símbolo vale siempre lo mismo sin importar su posición dentro del número, se denominan aditivos. Algunos pueblos, como los antiguos celtas, los mayas y los indios, usaron sistemas posicionales; los primeros dos tomaron al 20 como base del sistema, mientras que los últimos prefirieron el 10. Nuestro actual sistema proviene del que usaban los indios y, posteriormente, los árabes, por lo que solemos llamarlo indoarábigo.

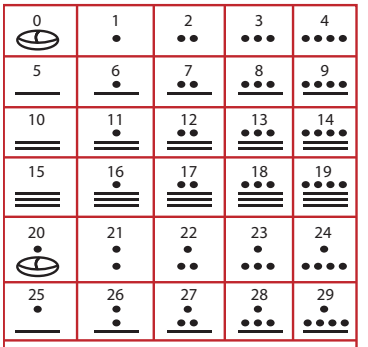
El sistema numérico binario

Para resolver el problema de medir las cantidades necesarias de cada ingrediente usando sólo las cucharas disponibles, Úrsula recurrió a representar cada una de las cantidades, de manera no muy distinta a la que usamos cotidianamente. Desde la enseñanza elemental se aprende que el número 4649, por ejemplo, está compuesto por: cuatro unidades de millar, seis centenas, cuatro decenas y nueve unidades o, lo que es lo mismo:

$$4649 = 4 \times 1000 + 6 \times 100 + 4 \times 10 + 9 = 4 \times 10^3 + 6 \times 10^2 + 4 \times 10^1 + 9 \times 10^0$$

Recordemos que cualquier número elevado a la potencia cero vale 1; es decir, el valor de cada dígito del número se multiplica por una potencia de 10. Esta característica del sistema numérico usual es compartida por todos los sistemas llamados *posicionales*. En ellos, el valor de un número está siempre vinculado con lo que se denomina la base del sistema. En la vida cotidiana usamos el 10 como base, por eso se dice que es un sistema *decimal* o de base 10.

Sin saberlo, Úrsula se dio cuenta de esto. Cualquier cantidad entre 0 y 15 onzas puede ser medida con las cucharas que tenía porque a éstas les cabía exactamente una potencia de 2. De hecho, Úrsula entró en contacto con el sistema *binario* o con base 2. En este sistema, los únicos dígitos posibles son el 0 y el 1, los *dígitos binarios*, o sea el ideal para usarse en una computadora, con bits. En la tabla 3 se puede ver cómo se expresa cualquier entero en el rango mencionado, reemplazando las "palomas" por unos y los "taches" por ceros.



Sistema numérico maya

Así pues, el 11 en el sistema decimal (al que denotaremos como 11_{10}) se escribe en binario como 1011 (o para ser consistentes con nuestra notación 1011_2) porque:

$$1011_2 = 1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 1 \times 8 + 1 \times 4 + 1 \times 2 + 1 \times 1 = 8 + 0 + 2 + 1 = 11_{10}$$

Número	$8 = 2^3$	$4 = 2^2$	$2 = 2^1$	$1 = 2^0$
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
5	0	1	0	1
6	0	1	1	0
7	0	1	1	1
8	1	0	0	0
9	1	0	0	1
10	1	0	1	0
11	1	0	1	1
12	1	1	0	0
13	1	1	0	1
14	1	1	1	0
15	1	1	1	1

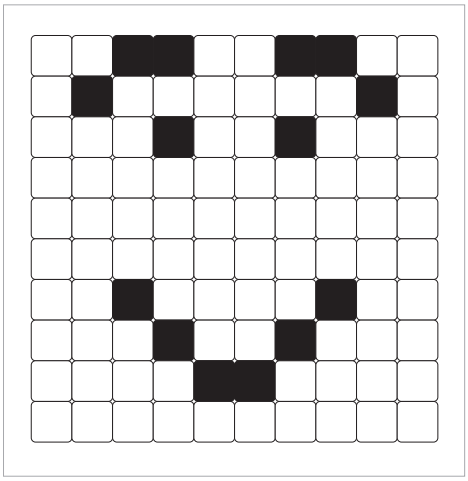
Tabla 3. Números que se pueden representar con cuatro bits.

4.3.2 Representando imágenes

La pantalla de una computadora consiste en pequeños elementos que se pueden controlar independientemente para que emitan luz o no, de distintos colores, y éstos son llamados *pixeles*. En su forma más simple, un pixel puede emitir luz blanca o no, en cuyo caso se ve negro. En pantallas a colores, al iluminar elementos de distintos colores, el ojo los combina y ve la gama cromática que produce la pantalla. Pero considérese únicamente el

caso de pantallas en blanco y negro, ya que para ilustrar las ideas es suficiente. En éstas, mediante un bit, podemos definir si el pixel está o no prendido. Si la pantalla fuera en blanco y negro, simplemente consistiría de un arreglo rectangular de pixeles.

Una manera muy simple de representar la imagen anterior podría ser mediante una palabra de 100 bits, 10 bits para cada uno de los 10 renglones, usando la convención en la que escribimos en el primer renglón de izquierda a derecha, seguido del segundo, y así sucesivamente. Los primeros cinco renglones se verían así:



0011001100010000001000010010000000
0000000000000000

Mediante este método se puede representar cualquier imagen en blanco y negro. En particular, es posible representar letras y números. Pero con frecuencia únicamente se representa algún tipo particular de imagen, como letras y números, o curvas y líneas, por ejemplo. En este caso existen representaciones mucho más eficientes. Este tema se verá con más detalle en el tema sobre multimedia.

Curiosidades
Asociado con todo sistema posicional, existe un conjunto de dígitos válidos y sus correspondientes valores. El número de dígitos siempre es igual a la base; así, en el sistema decimal existen 10 dígitos, a saber: {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9"} (se han puesto comillas para distinguir al símbolo del dígito de su valor). El valor de cada dígito es, respectivamente: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. En el sistema binario hay, por supuesto, sólo dos dígitos {"0", "1"} con valores 0 y 1, respectivamente.

4.3.3 Codificación y el mundo

El lector podría pensar que si usa un procesador de texto que manipula letras y otros símbolos en muy diversas **fuentes**, y también ve imágenes y películas, y escucha música, no es posible que todo eso sean sólo números binarios.

En efecto, para los humanos no lo son, nosotros vemos claramente en la pantalla una letra "A" en arial a 12 puntos o una imagen hecha de millones de puntitos de colores o escuchamos el sonido de una guitarra. Todo esto, como ya se ha mencionado, se representa dentro de la computadora como bits. El truco está en *discretizar y codificar.*

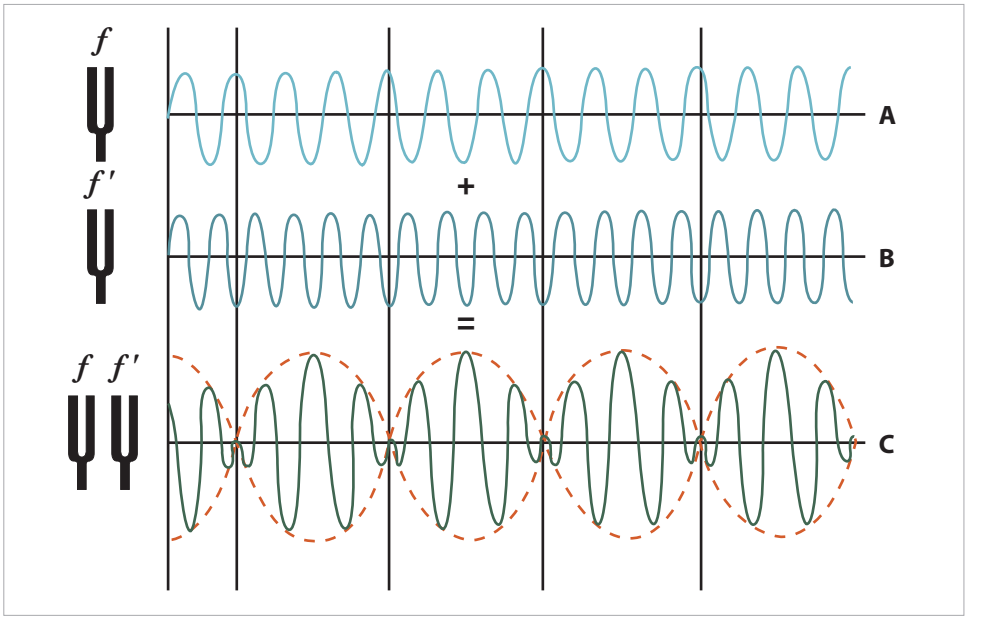
Curiosidades
Las computadoras modernas despliegan letras en la pantalla coloreando sus pixeles con colores adicionales al blanco y el negro para mejorar su apariencia. Ésta es una ampliación de cómo despliega en la pantalla una "a" una computadora Mac (una PC lo haría un poco diferente). Si la miras de lejos se ve sólo negra, pero más bonita.

El mundo discreto contra el mundo continuo

Como se ha explicado antes, cualquier cosa que pueda estar en un estado dentro de un conjunto finito de estados, se puede representar mediante bits. Un número entero en cierto rango, por ejemplo entre -1000 y +1000, se puede representar mediante bits, asignándole una secuencia de bits a cada valor en el rango. Un número real en ese rango no, ya que hay una infinidad de números reales en el rango. Así que lo que se hace cuando se requiere trabajar con números reales en una computadora es *discretizarlos*; es decir, aproximarlos mediante un conjunto finito y representar este conjunto mediante bits.

Usando esta misma idea es como se representan imágenes o sonidos en la computadora. En el ámbito auditivo, se puede pensar que un sonido es una onda que se propaga en el aire, y se representa mediante una función continua que nos dice cómo va variando su intensidad en el tiempo.

Figura 3. Combinación de dos tonos para producir un tercero.



Curiosidades

El filósofo griego Zenón de Elea (*circa* 450 antes de nuestra era) propuso varias paradojas para cuestionar las nociones de tiempo y espacio de su época que aún hoy son relevantes. La más famosa es la de Aquiles y la Tortuga, que nos hace pensar en lo siguiente: para recorrer 100 metros primero hay que recorrer 50, luego la mitad de los 50 restantes, es decir, 25. Pero para recorrer esos 25 metros, primero hay que recorrer la mitad, y luego el resto. La consecuencia es que nunca se llega a la meta, ya que cada vez que se está a punto de llegar, aún hay que recorrer primero la mitad de lo que queda y luego la otra mitad. Algo similar sucede en cuanto al tiempo: lo que se tarda en recorrer los 100 metros es la suma de una infinidad de intervalos de tiempo, cada uno de la mitad de tamaño que el anterior.

En la figura 3 se observan las ondas producidas por un diapasón que emite un cierto tono, luego otro que emite un tono más agudo y, en tercer lugar, cómo se vería la onda de los dos diapasones sonando simultáneamente. En la siguiente figura se ve cómo muestreando —es decir, midiendo la señal a intervalos de tiempo regulares— se obtiene una aproximación a la amplitud de la onda en cada instante de tiempo de la muestra. Las muestras se representan mediante bits, discretizando la amplitud continua. El resultado es una secuencia de números expresados en bits, que representan el sonido con mayor o menor calidad, en función de qué tan precisa sea la aproximación a la onda elegida, en sus dos dimensiones: qué tan frecuentemente se muestree la onda (distancia entre las flechas rojas) y cuántos bits se le asignen a cada muestra (longitud posible de cada flecha).

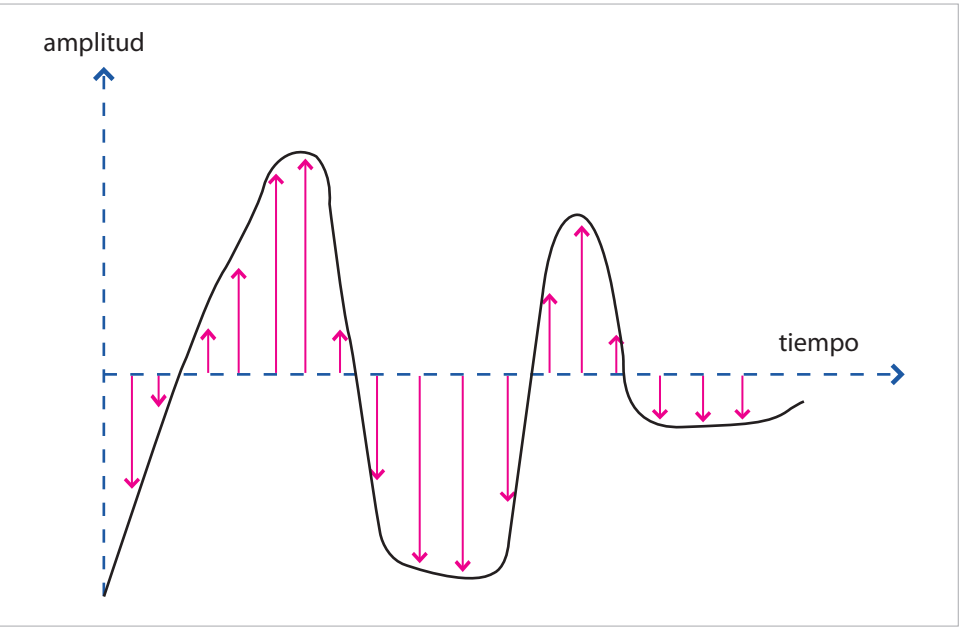


Figura 4. Muestreo de una señal.

Códigos en la vida cotidiana

Codificar es representar una cosa usando otra. Levantar el dedo pulgar de la mano hacia arriba es una manera de codificar el mensaje "todo salió bien"; SOS es el código universal para pedir auxilio; 10,4 es el código que suele usar la policía para decir "de acuerdo", y *tango* es el código que se usa en toda transmisión militar radial de la OTAN para representar la letra "t". En la vida cotidiana usamos códigos todo el tiempo: la CURP (clave única del registro de población) es un código que identifica individualmente a cada ciudadano mexicano. El número de cuenta que asigna la UNAM a cada alumno es lo que la institución usa para representarlo e identificar unívocamente su historial académico.

Representando texto

Para representar texto en la computadora se requiere de codificaciones eficientes y universales que permitan intercambiar eficientemente textos en diversos idiomas. En un procesador de texto hay un código que sirve para especificar el tipo de letra por usar, para indicar que una letra está en **fuente** Arial, uno para indicar su tamaño y otro para indicar de qué letra se trata. Con base en éstos, se determina qué dibujar en la pantalla y cómo hacerlo.

Para representar textos se pretende usar un código capaz de representar cualquier símbolo de los que se utilizan en cierto conjunto de idiomas; lo que suele hacerse en ese caso es definir un cierto tamaño de representación fijo, digamos 16 bits, y luego definir qué símbolo alfabético es representado por cada uno de los números binarios posibles de 16 bits.

A esto se le llama un *código de bloque*, porque todos los códigos miden lo mismo (16 bits en nuestro ejemplo). El código ASCII, que se solía utilizar hasta fines del siglo pasado, era de ocho bits. El código Unicode usado hoy día pretende representar todo carácter de cualquier idioma.

Compresión

En la actualidad es de fundamental importancia almacenar datos en la computadora de una manera eficiente, comprimiéndolos lo más posible, ya que la cantidad de datos que se manejan crece constantemente. Por otro lado, también en la naturaleza la compresión es muy importante, por las mismas razones de eficiencia y, más aún, podría decirse que todo aprendizaje consiste en comprimir información.

4.3.4 Compresión en la computadora

Si se piensa en el tipo de imagen que con frecuencia se envía por fax —una hoja blanca con algo de texto—, inmediatamente llama la atención el enorme desperdicio de bits incurrido por la representación, en la que cada pixel se representa por un bit, pues es de esperarse que una imagen de este tipo tenga muchos espacios en blanco. Por supuesto, esto no sucede con cualquier imagen, pero con muchas sí.

De manera que el computólogo se pregunta cómo encontrar una representación más eficiente o, dicho de otro modo, cómo comprimir una secuencia de bits como la anterior en otra más pequeña sin perder información. De hecho, muchos sistemas de compresión permiten inclusive perder algo de información, con el objetivo de comprimir aún más la entrada. Ya se ha visto cómo representar los números naturales 0, 1, 2, 3… mediante bits. Así se puede codificar una imagen escribiendo, renglón por renglón, el número consecutivo de pixeles blancos, seguido por el de negros, luego blancos, etc. O sea, que se podrían

representar los primeros cinco renglones de la cara de una imagen mediante la palabra (donde las comas indican que son diez pixeles consecutivos del mismo color):

222221161131213,10,10

Algo similar se hace en las máquinas de fax, que con frecuencia utilizan 100 pixeles por pulgada, por lo que una fila de pixeles blancos en la parte superior de una hoja de siete pulgadas de ancho podría ocupar 700 bits de almacenamiento. Usando el código que acabamos de describir, representamos con sólo 10 bits la fila en blanco, comprimiendo de esta manera 70 veces la información. Éste es un ejemplo extremo, en el que toda la fila es del mismo color, pero en promedio una máquina de fax logra una compresión de unas siete veces, lo que implica una reducción en la transmisión del fax en la misma proporción.

La reducción del tamaño de una secuencia de bits que representa alguna información es lo que llamamos *compresión*. La capacidad de almacenamiento de las computadoras ha crecido a una velocidad increíble en los últimos años: de 25 años a la fecha, un millón de veces aproximadamente. Sin embargo, las necesidades de almacenamiento han crecido aún más rápido, lo cual ha hecho que las técnicas de compresión sigan siendo importantísimas en los sistemas de cómputo, que deben manejar archivos enormes de imágenes, audio y video. Una película fácilmente puede requerir de 25 000 000 000 de bits para almacenarse, estando ya comprimida. Y como un video resulta susceptible de comprimirse en gran medida, ya que un cuadro tiende a no cambiar demasiado con respecto al siguiente, y se pueden usar 25 imágenes cada segundo, se pueden alcanzar compresiones de hasta 300 veces. Entonces, ¿qué es preferible, comprar más memoria o comprimir más? Cambiar espacio por tiempo, ya que comprimir y descomprimir toma tiempo. En el recuadro acerca de Samuel Morse se puede ver un antiguo ejemplo de compresión de datos.

Curiosidades
Un código de barras es una manera de representar información de manera que facilite su lectura mecánica. El primer código de barras fue desarrollado en 1948 por dos estudiantes en el Instituto Drexel de Tecnología, Bernard Silver y Norman Joseph Woodland, pero no se usó comercialmente con éxito sino hasta los años ochenta. Al inicio, solamente se utilizaba para representar números, pero ahora puede representar el código ASCII completo.

4.3.5 Compresión en la naturaleza

Curiosidades
Muchos sistemas de compresión de datos pierden información, con el objetivo de lograr comprimir aún más la entrada dada. Esto es común en compresión de audio o video, ya que la mente humana no percibe pequeñas distorsiones en este tipo de medios. Por ejemplo, es posible comprimir audio 10 veces sin que el oído lo note, y comprimir video 300 veces con muy poca pérdida visual. Para imágenes fotográficas es común usar el estándar jpeg y para audio y video el estándar mpeg.

El aprendizaje, en general, está íntimamente relacionado con la compresión de datos observados, de resultados de experimentos o de observación de comportamientos. Se puede afirmar que si un estudiante simplemente memoriza una serie de ejemplos, uno tras otro, no aprende. Si después de leer un libro nos preguntan de qué se trató, podemos elegir con cuánto detalle contar la historia, es decir, presentar el libro con distintos grados de compresión.

Compresión y ciencia

Se podría decir que la ciencia es el arte de la compresión de datos. El científico observa múltiples experimentos e intenta representar alguna propiedad de manera breve, un patrón que se repite, eliminando así el ruido de alguna esencia de interés en los datos. Por ejemplo, al observar una y otra vez el lanzamiento de balas de cañón de distintas formas y tamaños, el científico no recuerda las decenas de experimentos con sus pequeñas variaciones, sino que comprime toda esta información en una ley que describe el movimiento como si fuera el de un punto que se mueve siguiendo la forma de una parábola, y que está descrita mediante una ecuación matemática. De forma similar, un niño no aprende a hablar un idioma memorizando cientos de oraciones, sino que de alguna manera las va com-

primiendo todas en reglas y patrones que le permiten entender el lenguaje e inclusive hablarlo, generando nuevas oraciones que nunca había escuchado.

Compresión en animales

De igual manera, la naturaleza, siempre sabia, utiliza la compresión. Experimentos con hormigas, como los descritos por Li y Vitanyi, muestran que éstas comprimen los datos de los caminos que van recorriendo, es decir, aprenden qué caminos llevan a lugares con comida, y comunican esta información a otras hormigas. En un experimento, se coloca comida en algunas de las hojas de un árbol binario y en otras no.

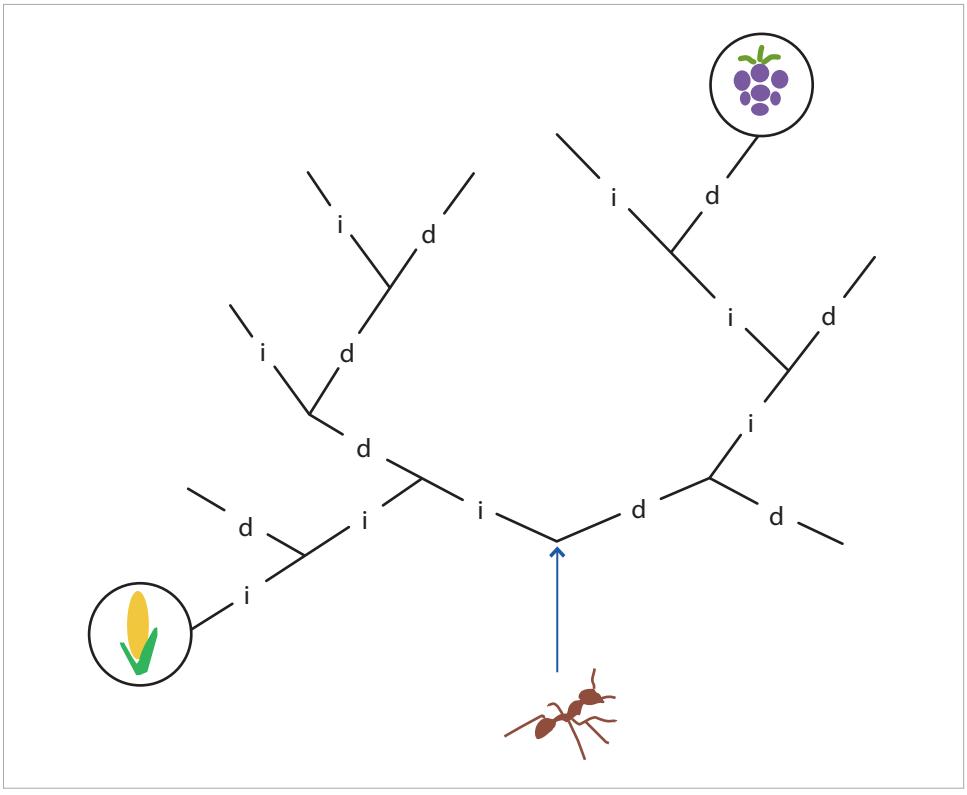


Figura 5. Qué camino aprenden más fácilmente las hormigas.

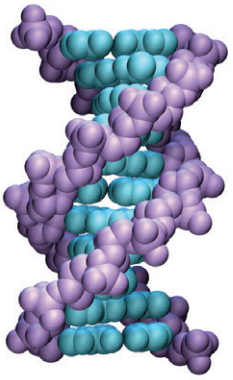
Se observó que a las hormigas les toma más tiempo comunicar patrones de caminos más "aleatorios" que más "regulares". El camino que lleva al elote es fácil de aprender, ya que consiste solamente en tomar bifurcaciones a la izquierda, mientras que el camino hacia las uvas "diid" es menos regular.

Compresión y aprendizaje

Cada camino que recorre una hormiga se puede representar mediante una secuencia de bits. Algunos caminos son muy fáciles de aprender, por más largos que sean. Por ejemplo, el camino "0101010101010101010101010101" se aprende fácilmente, ya que simplemente se describe como "cada vez que encuentres una bifurcación, toma el camino de la izquierda o el de la derecha, alternadamente". Mientras que en el caso de otros caminos, como "0100100011110100011101010000010", que no siguen ningún patrón aparente,

Curiosidades
A lo largo de la historia el hombre se ha preguntado cómo se pasa la información hereditaria de una generación a otra. En 1953, James Watson y Francis Crick descubrieron la estructrura y propiedades del ADN, la molécula que lleva esta información. Descubrieron que las características de un ser humano están codificadas en una larga cadena de ácidos nucleicos, arreglada en una doble hélice, como si fuera una escalera de cuerda trenzada con tres mil millones de peldaños, que se pueden representar por cadenas de cuatro letras, A, T, C y G, lo que les mereció el Premio Nobel en 1962. En 1968, Nirenberg, Khorana y Holley recibieron el Premio Nobel (con la contribución de otros científicos) por descifrar el código de la vida: el ADN representa un vocabulario de palabras de 20 letras.

ADN | © Latin Stock México.



debe aprenderse de memoria una secuencia entera de bits. En realidad, la dificultad no tiene que ver con la aleatoriedad del camino tal cual, ya que si se decidiera con una moneda cada uno de los bits de la secuencia, ambas secuencias tienen exactamente la misma probabilidad de ocurrir. La dificultad está relacionada, de alguna manera, con la cantidad de información en el camino. Se verá más acerca de la medición de la información más adelante.

Una manera de precisar qué quiere decir exactamente que un camino sea más fácil o difícil de aprender es usando la llamada Complejidad de Kolmogorov, que indica cuál sería el programa de computadora más pequeño que podría producir la secuencia. Es decir, esta teoría permite estudiar qué tanto es posible comprimir una cadena de bits.

En general, los computólogos han estudiado mucho el aprendizaje, desde diversos puntos de vista, y el área de aprendizaje en computación tiene muchas aplicaciones: los sistemas de compras en internet se adaptan a los gustos de los usuarios, los programas para jugar ajedrez aprenden las estrategias del oponente, los programas que reconocen la voz (y la escritura) se adaptan al estilo de una persona y cometen menos errores mientras más la escuchan, los robots aprenden acerca del ambiente donde se mueven, los programas para diagnósticos médicos aprenden de la experiencia, los algoritmos para finanzas y la bolsa de valores, etcétera.

Expansión. Representaciones para garantizar integridad

En ocasiones es útil expandir los datos, en lugar de comprimirlos, con el fin de introducir redundancias que permitan detectar y hasta corregir errores. Esto se hace tanto en computadoras como en la naturaleza.

4.3.6 Expansión en computadoras

Dos tipos de expansión son muy comunes en los sistemas de cómputo. El primero es el de la replicación de la información. Si se quiere estar seguro de no perder un dato, simplemente se guarda dos veces. A continuación se mostrará que hay formas mucho más eficientes de hacer esto, cuando se trata de almacenar información en el mismo sitio. La replicación es importante más bien en el contexto de transmisión de información y cuando es conveniente guardar la información replicada en lugares diferentes. Las bases de datos distribuidas y las aplicaciones de internet almacenan los mismos datos en distintos lugares no sólo por razones de tolerancia a fallas, sino también para que estén más cerca geográficamente del usuario, reduciendo así el problema de cuellos de botella al accesar los datos. Se darán más detalles sobre este tipo de problemas en el tema 7.

El otro tipo de mecanismo de expansión sobre el cual se profundizará es el que se realiza mediante codificación. Se trata de representar los datos con el objetivo de cuidar la integridad de los que se almacenan en el mismo lugar o para su transmisión; consiste en usar representaciones capaces de ser utilizadas para garantizar que lo que se recupera de los datos, luego de enviarlos o almacenarlos, sea lo mismo que se envió o almacenó originalmente, o al menos que se puedan detectar errores.

4.3.7 Códigos detectores y correctores de errores

Arcadio abrió la puerta de su recámara sumido en algunas cavilaciones, pero apenas dio un paso fue devuelto violentamente a la realidad por un extraño sonido: había pisado uno de sus CD. Revisó la superficie plateada y aparentemente no se había dañado de manera relevante, al menos no se veía más rayada de lo que ya estaba. Cuando lo puso, las primeras tres canciones se oyeron bien pero, a 2 minutos y 17 segundos de empezada la cuarta, el disco se trabó. Éste probablemente no sea el término adecuado, pero es el que resulta más acertado intuitivamente. El aparato reproducía la misma sílaba de la canción una y otra vez, como si tuviera hipo, hasta que Arcadio presionaba el botón de adelantar para forzarlo a "pasar el bache", por decirlo así, y luego continuaba sin contratiempos durante un rato.

En realidad, Arcadio fue testigo de lo que ocurre cuando un aparato reproductor de CD se topa con un error. En los discos compactos se almacenan datos con cadenas de bits grabados sobre su superficie. Con el fin de que el disco no sea tan delicado en su manejo y resulte posible reproducirlo con fidelidad a pesar de que tenga partículas de polvo, huellas digitales, manchas o pequeñas rayaduras inevitables, cada una de estas cadenas de bits va codificada usando una representación que permite al aparato reproductor "darse cuenta" de ciertos errores en los bits leídos y corregirlos. Cada partícula de polvo o huella digital puede afectar potencialmente algunos de los bits leídos y cambiarlos; reproducir los datos leídos suponiéndolos inmaculados resultaría un desastre ininteligible. Por tanto, los datos se codifican de tal forma que, luego de ser leídos, el aparato reproductor pueda verificarlos para determinar: *a*] si son correctos o *b*] si no lo son y pueden ser corregidos, es decir, qué hay que cambiarles para corregirlos.

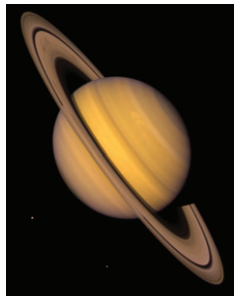
El algoritmo estándar de los reproductores de discos compactos de audio indica que si se detectan errores en los datos y éstos no pueden ser corregidos, entonces se debe proceder a enviar al dispositivo de audio la última señal correcta leída. De allí que parezca que el disco se "atora" o "tiene hipo". Esta misma idea se utiliza en casi todos los contextos en los que es necesario almacenar o transmitir información de manera confiable. En el disco duro de cualquier computadora los datos se almacenan en bloques de ceros y unos, y éstos van sucedidos de un código que permite verificar lo que se leyó y corregir un cierto número de errores. Prácticamente todas las transmisiones de televisión digital, de telefonía, satelitales, de internet y los datos almacenados en DVD, por ejemplo, se *codifican usando códigos detectores y correctores de error.*

El objetivo es, por supuesto, garantizar, en la medida de lo posible, que la información que se recupera del medio de transmisión o almacenamiento sea exactamente la misma que se envió o se almacenó originalmente. Se da por sentado que el canal de transmisión o el medio de almacenamiento no es infalible, es decir, que se presupone la existencia de errores que pueden alterar lo que se almacena o transmite por un canal. A esto se le llama, genéricamente, *ruido*.

En la teoría de la información se estudian diversos tipos de ruido. Sin embargo, la característica más relevante del ruido es que es aleatorio, porque no se puede saber de antemano qué bits serán alterados. En el modelo de ruido más sencillo se supone que un bit tiene la misma probabilidad de ser alterado que cualquier otro, y que la alteración de uno de ellos no tiene nada que ver con la posible alteración del siguiente. A esto suele llamársele *ruido blanco*. Un ejemplo de ruido blanco es el "hormigueo gris" que queda en la pantalla cuando termina la programación de las transmisiones por televisión.

La clave para lograr "darse cuenta" de que han ocurrido errores y el primer paso para saber cuáles son los datos incorrectos es añadirle a la información cierta redundancia. La

Curiosidades
Las sondas espaciales que el hombre ha enviado a explorar los confines del Sistema Solar envían a la Tierra las imágenes fotográficas que capturan, codificadas con mecanismos que permiten detectar y corregir errores. Las transmisiones de estas sondas están sometidas a perturbaciones ocasionadas por las emisiones electromagnéticas del Sol, lo que suele llamarse "viento solar". Ésta es una célebre fotografía de Saturno tomada por la sonda Voyager 1 en 1979.



Saturno | © *NASA/JPL Caltech.*

idea general para hacer esto es poseer un catálogo de palabras suficientemente amplio para decir lo que sea necesario, pero que al mismo tiempo sean palabras suficientemente diferentes entre sí. Será más fácil comprender esta situación con un ejemplo.

Supongamos que se necesita comunicar 16 cosas diferentes. Si hay que decir 16 cosas diferentes se pueden usar cuatro bits para codificarlas, dado que $2^4 = 16$. Ahora, si se desea transmitir datos a un compañero, cada uno de éstos será una de esas 16 cosas que se han representado con cuatro bits. Lo único que debe hacerse es enviar el código —es decir, cuatro bits— por cada cosa que se desea comunicar al receptor. De manera que aunque se tengan varias cosas que decir, sólo se envían cuatro bits por cada una. Complicando las cosas un poco, asumiendo que el canal de comunicación que se usa para transmitirle al receptor tiene ruido y que éste puede echar a perder algunos de los bits enviados, bien pudiera ser que se envíe la palabra 0010 y un compañero reciba 0011 porque se alteró el último bit. En la figura 6 se observa cómo mientras más errores se introducen, más se aleja el mensaje de la palabra original. El primer círculo contiene palabras con un error, el segundo con dos, etcétera.

¿Puede el receptor de la palabra 0011 percatarse de que ha ocurrido un error? Él no sabe qué se le envió, la única manera de comunicarse con el remitente es a través del canal que se está usando. La respuesta es *no*. El código que el receptor recibe es tan válido como el que se le envió; el receptor puede, con toda confianza, suponer que lo que recibió es lo que efectivamente se le mandó porque no tiene modo de distinguir las palabras del código correctas de las que no lo son. El problema es que se utilizó todo el poder de expresividad posible de cuatro bits para crear las palabras del código que se quería decir, y ninguna palabra de cuatro bits es inválida. *No hay espacio entre las palabras.*

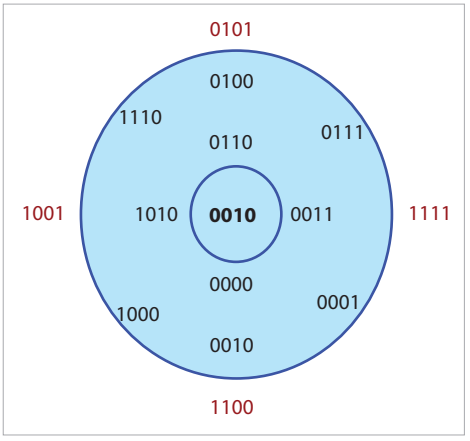


Figura 6. Posibles cambios de la palabra original.

Para poder distinguir cuando algo está mal no se debe usar todo el poder expresivo del código; es necesario que algunas de las posibles secuencias, de la longitud que se usen, sean inválidas. Éstas forman "huecos" y logran que entre las palabras del código haya suficiente distancia. En la figura 7 se observa que si se incluyen en el código las palabras 0100 y 0010, pero no las de alrededor de éstas, es posible detectar si hubo un (sólo uno) error. Las palabras 0101, 1110, 0110, 0000, 1010 o 0011 no serían válidas. Pero aunque así se puede detectar un error, *no es posible corregirlo*; si se recibe 0000, no hay manera de saber si la palabra enviada fue 0100 o 0010, ya que no hay manera de saber si fue el segundo o el tercer bit el que cambió.

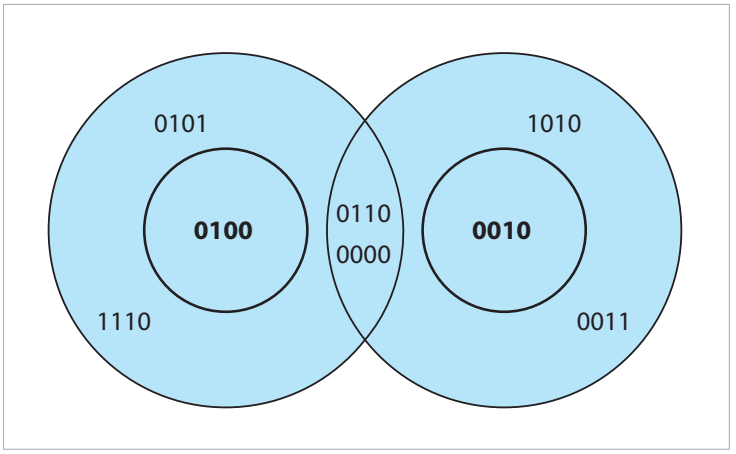


Figura 7. Errores detectables si algunas secuencias se consideran inválidas.

Si sólo se necesita expresar ocho cosas diferentes y se pueden usar los mismos cuatro bits, hay más espacio. Suponiendo que se usan como catálogo de palabras de

código válidas todas las que terminan con 0. Si ahora se envía 0110 y del otro lado se recibe 0111, el receptor sí puede darse cuenta de que algo se estropeó en el canal, porque sabe que no es posible enviar algo que termine con "1" pues no está en el catálogo de palabras válidas. A los tres bits estrictamente necesarios para expresar ocho cosas se les ha añadido uno más; entonces ya se puede hacer algo con respecto a los errores. Desgraciadamente, si se echa a perder el segundo bit, como se mencionó antes, y se recibe 0010, el receptor no podrá darse cuenta del error. No basta entonces con añadir datos, también hay que cuidar que las distancias entre las palabras del código estén bien distribuidas; que los huecos rodeen bien a todas las palabras.

Un esquema similar al que se usó en el párrafo anterior ha sido muy usado a lo largo de la historia. Se trata de añadir un bit a cada palabra del código, de tal forma que siempre se complete un número par o impar. Según se elija un número par o impar, el sistema adquiere su nombre. Por ejemplo, si se desean codificar 16 datos posibles, se emplearán cuatro bits para cada uno de ellos, pero esta vez cuando la palabra de cuatro bits tenga un número impar de unos se le agregará otro más al final y cuando no sea así se le agregará un cero. Este esquema, claro está, es el de *paridad* par. Así, la palabra 0110 se convierte en 01100, el último cero significa que la palabra completa de cinco bits tiene un número par de bits "prendidos", es decir, con valor 1, en los primeros cuatro. En cambio, si deseamos enviar 0010, deberíamos enviar 00101, dado que el quinto bit indica que se tiene un número impar de unos en los primeros cuatro bits. Ahora sí, el receptor se percatará de que algún bit fue cambiado accidentalmente en el canal durante la transmisión.

Con un esquema de verificación de paridad es posible *darse cuenta* de que ha ocurrido un error —o, mejor dicho, un número impar de errores— en los bits transmitidos, pero cuando el número de errores es par, pasarán desapercibidos. Para poder detectar un mayor número de errores o, mejor aún, corregirlos, se debe añadir más redundancia.

La clave para poder corregir errores es hacer que las palabras del código que se usan para enviar los datos sean muy diferentes entre sí. Si dos palabras se parecen mucho, por ejemplo: 0110101 y 0110001, basta con que ocurra un error en el quinto bit para que una se convierta en la otra y el receptor del mensaje la dé por buena. Cuanto mayor sea la distancia entre las palabras del código, mayor será el número de errores que deben ocurrir en el canal para que se convierta en otra palabra válida. Por esto, una de las cualidades más importantes de los códigos detectores y correctores de errores es la medida de la dife-rencia mínima entre sus palabras, a lo que suele llamársele *distancia mínima del código*.

4.4 MEDIR INFORMACIÓN

Hemos hablado acerca de comprimir y expandir datos sin cambiar la información representada. Pero surgen las siguientes preguntas: ¿qué tanto se puede comprimir?, ¿cómo se sabe si una representación es óptima o, más en general, cómo se puede medir la cantidad de información representada? ¿Cuál es la unidad de medida de la información? Ciertamente, no son los litros ni los gramos. Son los bits. Shannon bien merece el nombre de "padre de la teoría de la información" por haber contestado a estas y muchas otras preguntas relacionadas.

Adivinar el regalo

Arcadio presionó el último dígito del número de Úrsula y cuando escuchó su melodiosa voz le informó que le había comprado un regalo, un disco compacto. Luego le pidió que adivinara de cuál se trataba haciendo preguntas de "sí o no" únicamente.

—A ver... ¿es música popular? —comenzó Úrsula.
—Sí.
—¿Banda?
—No.
—¿Pop?
—Sí.
—¿Nacional?
—No.
—¿Mujer?
—No.
—¿Es de Robbie Williams?
—¡Sí! Eres buena para esto, te tomó sólo seis preguntas.

El juego practicado por Arcadio y Úrsula resulta muy interesante desde el punto de vista computacional, porque nos da la clave acerca de cómo cuantificar la información. Para comprender este concepto pensemos en la siguiente situación: A le pide a U que adivine el número entero en que está pensando en ese momento —puede ser cualquiera entre 0 y 15— haciendo preguntas cuya respuesta sea "sí" o "no".

Es posible preguntarse ahora qué sabe U acerca del número que eligió A. Salvo el hecho de que es algún elemento del conjunto {0, 1, 2, ..., 14, 15}, U no sabe nada. El número elegido por A es completamente arbitrario, elegido de acuerdo con lo que en lenguaje matemático se conoce como *probabilidad uniforme*. ¿Cómo procedería U a hacer las preguntas para determinar el número que piensa A? Por supuesto, una opción es comenzar preguntando si se trata del 0 y si la respuesta es "sí" se termina el juego, pero si la respuesta es "no" se procedería a preguntar por el 1 y así sucesivamente hasta que la respuesta sea "sí" o hasta que se llegue al 14 (entonces la respuesta obvia sería el 15). Con este procedimiento seguro se resuelve el problema, pero no es eficiente porque en el peor de los casos hay que hacer 15 preguntas. Inclusive en el caso promedio se tendrían que hacer siete preguntas. De hecho, preguntar "¿el número que escogiste es *x*?" no es eficiente porque el número de preguntas potenciales es siempre el mismo: el tamaño del rango menos uno. Cada respuesta proporciona muy poca *información* acerca del número elegido; en la mayoría de los casos, sólo indica cuál no es, lo que nos deja con un conjunto de posibilidades igual al que teníamos antes, menos uno.

En cambio, si preguntamos "¿el número es menor que 8?", la respuesta proporciona mucho más información, pues indica en qué mitad de la lista de 16 números está el que se quiere adivinar. Después se puede preguntar por el que se encuentre en la mitad, de la mitad donde se espera que esté el número buscado; eso nos ubica, con sólo dos preguntas, en un conjunto de posibilidades con una cuarta parte del tamaño del conjunto total inicial. Una pregunta más y se estará en una octava parte; luego de la cuarta pregunta se sabe exactamente qué número es el elegido.

En la figura 8 de la siguiente página se observa un despliegue de todas las opciones en función de las respuestas previas. A una estructura como ésta se le denomina *árbol de decisión*. Cada vez que se recibe una respuesta, se desciende por una de las ramas, la etiquetada con la respuesta recibida. Conforme se avanza, el conjunto de posibilidades es de la mitad del tamaño del conjunto previo.

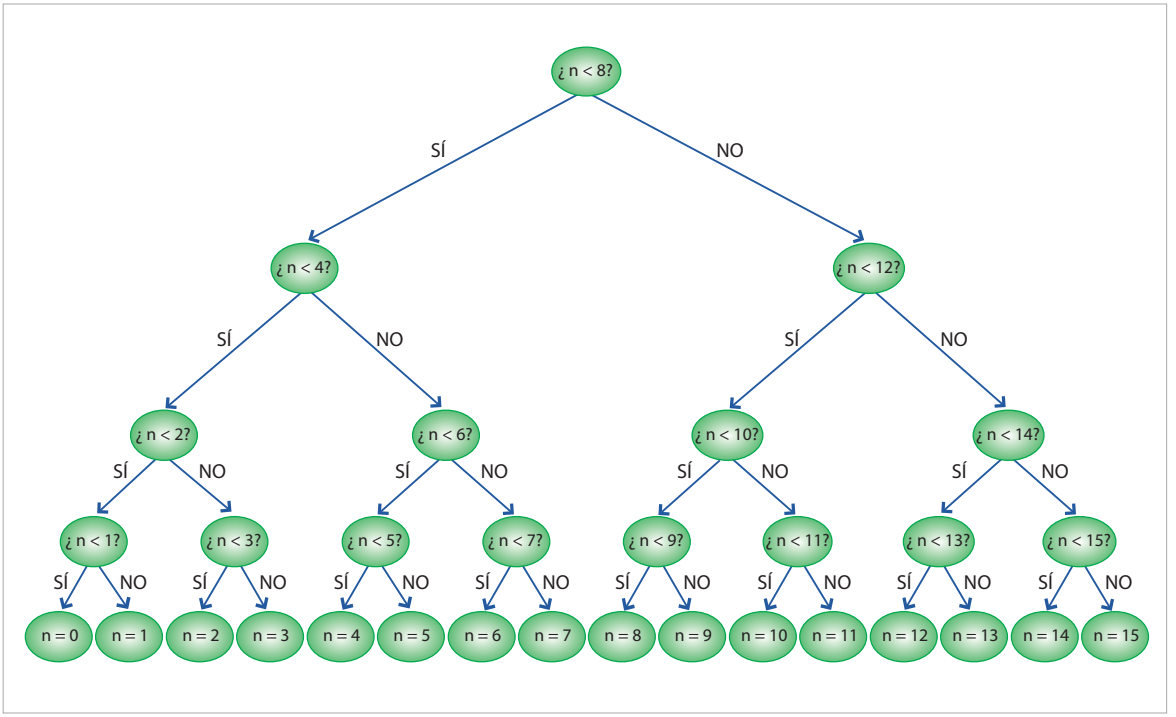


Figura 8. Árbol de decisión para determinar un número aleatorio entre 0 y 15.

Del análisis del ejemplo se concluye que el número óptimo de preguntas es cuatro. Vale la pena observar que, si se sigue una ruta descendente desde la raíz del árbol hasta una hoja cualquiera, cada arista está etiquetada como "sí" o "no"; de manera que se puede especificar el camino mediante bits, con la secuencia de "sí" y "no" necesaria para llegar a una hoja desde la raíz. Por ejemplo, la ruta para llegar a la hoja "n = 10" es: no, sí, no, sí; la del 6 es sí, no, no, sí. Si se reemplazan los "no" por 1 y los "sí" por 0, se obtiene algo que resulta familiar: la ruta del 10 es 1010 y la del 6 es 0110, que son justamente las representaciones binarias de esos números, como se vio cuando se expresaron los ingredientes de la receta de galletas.

Resumiendo lo anterior, si no hay pistas útiles para poder restringir la búsqueda de un elemento de un conjunto, entonces el número mínimo de preguntas "sí o no" que hay que hacer coincide con la longitud de la expresión binaria mínima necesaria para poder escribir todos los números del conjunto. Si se reflexiona acerca de que cada vez que alguien responde una pregunta está aportando información, entonces se puede decir que, de hecho, la longitud de la expresión binaria mide la cantidad de información necesaria para determinar cada número del conjunto, y es básicamente igual al logaritmo con base 2 del tamaño del conjunto. El tamaño del código usado para representar a los elementos del conjunto indica la cantidad de información contenida en ellos.

4.4.1 Cantidad de información y entropía

Hay una diferencia notable con el caso que enfrenta Úrsula al adivinar el disco que Arcadio le compró. Allí sí se sabe algo, no se parte de que fue elegido de manera completamente aleatoria. Poniéndolo en otros términos, las listas de popularidad, los conocidos "top 10" o "top 40" de discos, se forman contando cuántos discos se venden de cada intérprete o artista; es decir, con base en estadísticas. Así que lo que la lista dice es algo como:

si *x* está arriba de *y* es porque es más frecuente que un cliente cualquiera compre el disco *x* y no el *y*; podría ser, por ejemplo, que de cada 100 discos comprados, 37 sean *x* y 23 sean *y*. Un árbol de decisión eficiente para determinar qué disco compró Arcadio debe tomar en consideración ese hecho. En cuanto Úrsula supo que Arcadio le había comprado un disco de regalo, intentó preguntar por cosas que fueran más probables de ocurrir. Cuesta menos trabajo deducir algo muy común que algo raro. Si Arcadio le hubiera comprado un disco de música medieval letona, Úrsula hubiera tenido que hacer más preguntas para adivinar.

En cualquier representación de datos que pretenda ser eficiente se debe utilizar un criterio análogo al de la "popularidad": decir que ocurrió algo que todo mundo sabe que ocurre muy frecuentemente no es noticia; en otras palabras, aporta poca información o *se requieren pocas preguntas para deducir qué ha ocurrido.* En cambio, cuando lo que ocurre es un evento raro, fuera de lo común, sí se necesita mucha información para deducirlo y, por tanto, para representarlo.

La estrategia de la representación eficiente entonces es considerar qué tan probables son los datos que se pretende representar. Si un dato aparece 70 veces en un conjunto de 100, lo ideal sería que la longitud de su representación fuera la mitad de la usada para otro que ocurre 35 veces. Datos más frecuentes corresponden a longitudes de representación menores, porque hay que decirlos más seguido y conviene decir menos las más de las veces. Datos menos frecuentes corresponden con longitudes mayores porque, como se dicen pocas veces, se necesita más información para que puedan ser determinados.

Un ejemplo concreto: una agencia de venta de automóviles desea determinar una codificación eficiente para almacenar los colores de las unidades en su inventario, pero usar los nombres de los colores es poco práctico porque siempre son los mismos y son largos, digamos que se trata de los mostrados en la columna izquierda de la tabla 4. En la segunda columna de la izquierda se muestra el número total de vehículos vendidos de cada color durante los últimos seis meses y constituye un indicador de qué tan frecuente o raro es cada color en el inventario.

En la siguiente columna aparece un código eficiente para representar cada color. Queda fuera del alcance de este texto explicar cómo se obtuvo dicho código, pero lo que hay que notar es que la longitud de las palabras, el número de bits de cada una, está en proporción inversa a la frecuencia del color correspondiente. Si se tuvieran 26 automóviles en el patio de la agencia: 10 rojos, 5 azules, 3 verdes, 3 amarillos, 2 plata, 2 oro, 1 gris y 0 rosas (cantidades que guardan entre sí aproximadamente las proporciones de la tabla), se requerirían 67 bits para codificar los colores de todos. En cambio, si no se supiera nada, si todos los colores fueran igualmente usuales, entonces se requerirían tres bits para cada color, dado que con tres bits se pueden decir exactamente $2^3 = 8$ cosas, igual al número de colores que se tienen. En este caso, para guardar los colores de todos los vehículos del patio se necesitarían 78 bits, puesto que $26 \times 3 = 78$. Si en vez de los 26 vehículos de la agencia se consideraran los 2 540 que están en el patio de la planta de producción y cuyos números relativos guardan la misma relación de proporcionalidad que se muestra en la tabla, entonces se usarían 6 520 bits en sus colores, mientras que la codificación de tres bits por color requeriría de 7 620.

En computación, a la longitud mínima de bits necesarios para codificar (representar) un dato se le denomina *cantidad de información* y la cantidad promedio de bits necesarios para representar cada uno de los datos de un conjunto se conoce como *entropía de información*. Éstas son cantidades que se calculan solamente con base en qué tan frecuente o probable es cada uno de los datos de un conjunto, así que pueden ser números fraccionarios. En el ejemplo de los colores de automóviles, la entropía de información

Curiosidades
Samuel Morse (1791-1872). Inventor del telégrafo, ideó también el código que lleva su nombre con la misma premisa de eficiencia que hemos mencionado aquí: el código Morse para la letra "E" es sólo un punto; el de la "T" es sólo una raya; el de la "Z", dos rayas y dos puntos. Esto corresponde con la frecuencia con la que esas letras se usan en idioma inglés. Morse quería que su código fuera eficiente para transmitir mensajes en inglés, así que asoció códigos más largos a letras más improbables y más cortos a las más frecuentes: en promedio, en un texto en inglés de 1 000 letras, 127 serán "E", 91 serán "T" y sólo una será "Z".

Samuel Morse | *© Anónimo.*

del conjunto de colores, con base en su probabilidad (estimada a través de las frecuencias que se muestran), resulta ser 2.49, lo que significa que cada color requiere un código binario que mide en promedio 2.49 bits de longitud. Los datos usados en este cálculo están en la tabla 5.

4.4.2 Codificación eficiente

La intención de los esquemas de codificación que asocian códigos más largos a datos más raros y más cortos a los más comunes, es minimizar la longitud promedio de los códigos usados para representar los mensajes construidos con los datos de esas características. Es decir, pretenden que los códigos que se usen para representar cada uno de los datos del conjunto tengan una longitud, entera en bits, lo más parecida posible a la cantidad de información del dato que se codifica, para que en promedio se use un número de bits que se aproxime lo más posible a la entropía del conjunto. En el ejemplo, el código mostrado en la tabla 4 tiene una longitud promedio de palabra de 2.57 bits aproximadamente y la entropía, como ya se dijo, es de 2.49 bits, así que el código está bastante cerca de la entropía.

Siempre que se codifica eficientemente un conjunto de datos, el código usado tiene una *longitud promedio* al menos tan grande como la entropía del conjunto; a lo más que se puede aspirar, entonces, es a tener una longitud promedio que sea exactamente el valor de la entropía. Éste es un límite de qué tan eficiente se puede ser para almacenar datos: ningún código puede expresar un conjunto de datos en menos bits que la cantidad de información inherente al conjunto, a menos que pierda información.

Nombre del color	Frecuencia	Código binario eficiente	Longitud de palabra de código
Rojo bravío	50	1	1
Azul oceánico	23	011	3
Verde pacífico	15	001	3
Amarillo flama	12	0100	4
Plata	12	0101	4
Oro	9	0001	4
Gris acero	5	00001	5
Rosa pasional	1	00000	5

Tabla 4. Lista de colores de automóviles, frecuencia de venta y código para almacenar los colores eficientemente.

Nombre del color	Probabilidad estimada	Cantidad de información	Prob x Info
Rojo bravío	0.393	1.34	0.53
Azul oceánico	0.181	2.46	0.446
Verde pacífico	0.12	3.08	0.364
Amarillo flama	0.094	3.4	0.322
Plata	0.094	3.4	0.322
Oro	0.07	3.81	0.271
Gris acero	0.039	4.66	0.184
Rosa pasional	0.007	6.98	0.055
Entropía			**2.492**

Tabla 5. Probabilidades estimadas, cantidad de información de cada color y entropía en el ejemplo de los colores de automóviles.

Al igual que su concepto homónimo de la física, la entropía en computación se refiere al grado de desorden en el conjunto de datos; cuanto mayor es éste, más aleatorios son los datos, más impredecibles, y por tanto se requiere en promedio mayor información para deducir cualquiera de ellos.

Curiosidades
Desde el año 1900 antes de nuestra era los egipcios usaron la criptografía. Se sabe que Khnumhotep II, un arquitecto de Amenemhet II, la usó para cifrar documentos acerca de las construcciones que hizo para el faraón. La escítala era un dispositivo utilizado por los antiguos griegos para cifrar mensajes mediante el método de transposición.

Escítala | © *Anónimo.*

Escondiendo información: criptografía

Los datos se pueden comprimir, expandir y también esconder. El objetivo básico de la criptografía es enviar un mensaje de manera que solamente el destinatario pueda entender su significado. Comunicarse en secreto es esencial en una guerra, así que la historia de la criptografía se remonta a la historia antigua.

En los tiempos modernos se usa la criptografía para comunicarse en secreto, con fines tanto militares como civiles. En todas las transacciones bancarias y comerciales, la criptografía es indispensable para mantener la confidencialidad de la información y permitir identificar a las partes que se comunican mediante firmas electrónicas y otros mecanismos. Por ejemplo, cada vez que tecleamos un *password* —o una clave de acceso o contraseña— en algún sistema (ya sea una computadora, un cajero automático u otros), se usa criptografía para verificar su validez de manera que la computadora no necesite tenerlo almacenado. Es decir, la criptografía le permite a esa computadora verificar que el *password* es el correcto sin conocerlo. Gracias a la criptografía se pueden hacer esta y muchas otras cosas que parecen imposibles, como monederos electrónicos, firmas digitales, protecciones de material con derechos de autor, jugar cartas por internet, enviar mensajes al futuro que no se puedan abrir antes de tiempo, etc. Veremos cómo es posible calcular la edad promedio de un grupo de personas sin necesidad de que nadie le diga a nadie cuántos años tiene o que cualquiera pueda encriptar mensajes para que sólo una persona los pueda descifrar.

La criptografía moderna es una de las ramas más sofisticadas de la computación, y en la cual se usan matemáticas más avanzadas; es un área en la que participan, además, la ingeniería, la computación e inclusive los abogados, quienes diseñan el contexto legal de los sistemas que usan criptografía, como las firmas digitales o las votaciones electrónicas. La teoría de la información y las ideas en torno a ella que hemos descrito están íntimamente relacionadas con la criptografía, pues se requiere evitar que la información sea visible para personas no autorizadas. También intervienen ideas de análisis de algoritmos y complejidad, debido a que el enfoque en la criptografía moderna no es mantener un secreto a toda costa. No interesa si una persona no autorizada puede o no descifrar el mensaje, lo importante es garantizar que le tomaría *demasiado tiempo* hacerlo.

La criptografía es una de las ramas de la computación más importantes, ya que la integridad y buen funcionamiento de todos los sistemas dependen de ésta, en un mundo en que prácticamente ya no existe ninguna computadora que no esté conectada a internet y en el cual la mayoría de las transacciones comerciales pasan en algún momento por la red.

Previniendo el escarnio

—Perdón, no se me había ocurrido, respecto a la comida del viernes en tu casa, ¿qué llevo?
—fueron las palabras de Arcadio.
—Pues no sé... nada —respondió Úrsula—. Pero qué bueno que hablas —dijo murmurando extrañamente y continuó—: yofo crefeofo quefe defebefes quifitafartefe efel afadofornofo defe lafa cefejafa.
—¿Qué?
—¿No sabes hablar en efe?
—¿Qué es eso?
—¡Uy!, olvídalo, te mando un mensaje a tu cel.

Momentos después, Arcadio recibió un mensaje de texto en su teléfono celular: "No podía hablar normal porque aquí estaba mi papá, te dije que te quitaras el adorno de la ceja, no le gustan, o sea: nofo lefe gufustafan."

4.4.3 Criptografía básica

La intención primaria (ya se han mencionado muchas otras, como firmas o identificación) de la criptografía es lograr que dos personas o, para ser más generales, dos entidades diferentes (A y B) logren comunicarse de modo que sólo ellos puedan entender el significado de los mensajes. Se presupone entonces la existencia de una tercera entidad, el enemigo, que está permanentemente al acecho procurando comprender lo que A y B se dicen. Para que el enemigo no tenga éxito, A y B deben enviarse mensajes *cifrados*.

Existe otra área dedicada al ocultamiento de los mensajes, se denomina *esteganografía*, pero no será posible tratarla en el espacio de este libro. A su vez, la criptografía es parte de un área más general llamada *criptología*. Otra parte de la criptología es el criptoanálisis, que se encarga de intentar descifrar las comunicaciones secretas.

Cifrar datos es entonces representarlos de tal forma que resulten ininteligibles para quien "no está autorizado". El proceso criptográfico básico es un algoritmo que recibe como entrada los *datos claros*, comprensibles para cualquiera, y un elemento adicional llamado *clave de cifrado*; la salida del algoritmo es el conjunto de datos cifrados. El mensaje cifrado es entonces transmitido o almacenado, a través de o en un medio, respectivamente, que es susceptible de ser intervenido por el enemigo. El receptor del mensaje, por su parte, debe recuperar los datos originales, para lo cual ejecuta también un algoritmo. El algoritmo de descifrado recibe como entrada los datos cifrados, una clave de descifrado, que puede o no ser la misma que se usó para cifrar, y produce como salida los datos claros originales que fueron cifrados por el emisor, solamente en el caso de que la clave de descifrado sea la correspondiente a la clave de cifrado. Así que la "persona autorizada" es aquella que conoce esta clave.

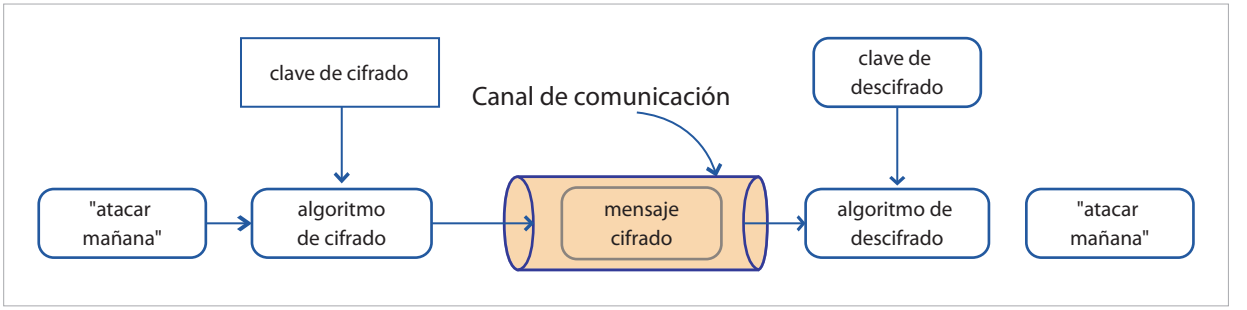


Figura 9. Esquema básico de cifrado.

En la criptografía clásica, los algoritmos son simétricos; es decir, la clave de cifrado y descifrado son iguales. En este caso, se trata de diseñar los algoritmos de cifrado y descifrado de manera tal que únicamente las personas que conozcan la clave puedan entender los mensajes. Ésta es la criptografía que permite cifrar y descifrar más rápidamente, y existen algoritmos que se usan mucho para hacer esto, como los basados en el estándar DES (Data Encryption Standard), pero tiene el inconveniente de que cada pareja de personas que se desean comunicar, primero deben ponerse de acuerdo en una clave secreta que sólo ellos conozcan. Sería mucho más eficiente tener dos claves: una pública y otra

secreta; de tal forma que Arcadio pudiera publicar en un directorio su clave pública y todo aquel que la viera pudiera enviarle mensajes que solamente él pueda leer con la correspondiente llave secreta. Hasta antes de los años setenta se creía que esto era imposible, pero en un trabajo publicado en 1976 por Diffie y Hellman —donde muestran cómo se puede lograr que dos personas que solamente se pueden comunicar a través de un canal inseguro, al que cualquiera tiene acceso, se pongan de acuerdo en una clave secreta— nace la *criptografía de llave pública*. Ahora existen algoritmos de este tipo, llamados *asimétricos*, usados en todas las transacciones seguras de internet. Uno muy usado es el RSA (por las iniciales de los apellidos de sus creadores: Ron Rivest, Adi Shamir y Len Adleman). Sin embargo, los que se conocen son mucho más lentos que los simétricos usados.

En el lenguaje de la efe usado por Úrsula, la introducción de las letras "f" es lo que ocasiona que el mensaje cifrado ya no sea reconocible de inmediato cuando se le escucha. Luego de introducir cada letra "f" se duplica la vocal que la precedió, así que por cada vocal del mensaje original, en el mensaje cifrado existen dos, de hecho la misma: una está antes y otra después de la "f". Existen mucho otros algoritmos simples de cifrado, los métodos clásicos son:

- *Transposición*: reordena las letras de un mensaje; por ejemplo, "necesito ayuda" se podría convertir en "enecisotyadua".
- *Sustitución*: cada letra se sustituye por otra; por ejemplo, "necesito ayuda" se podría convertir en "ofdftjupbzveb" al cambiar cada letra por la que le sigue en el alfabeto.

El lenguaje efe sólo ha servido para ilustrar los conceptos, y en realidad no es más útil que cualquiera de los métodos de transposición o sustitución. Cualquiera de estos métodos es fácil de descifrar sin necesidad de tener la clave secreta, usando una computadora y viendo suficientes mensajes cifrados. La idea básica para hacer esto es tomar estadísticas acerca de la frecuencia de las letras o combinaciones de éstas en un conjunto suficientemente grande de mensajes cifrados, y compararlas con las estadísticas conocidas del idioma español (o en el que se estén escribiendo los mensajes a descifrar). En efecto, con el advenimiento de veloces computadoras que pueden analizar miles y miles de posibilidades por segundo, la criptografía cambió.

La intensa interacción entre el criptoanálisis y la criptografía se volvió indispensable. Es muy fácil diseñar algoritmos que aparentemente "transforman mucho" un mensaje, de manera que parezca imposible descifrarlo, hasta que se hace un sofisticado criptoanálisis y se logra. Un esquema criptográfico realmente útil es mucho más complicado, y dado que en realidad no se sabe cómo probar con toda certeza cuándo un esquema práctico es seguro, se publica el esquema y la comunidad internacional lo criptoanaliza, y si no logra violarlo durante mucho tiempo se considera que el esquema es seguro. Hay esquemas criptográficos actuales con los cuales se logra que, por ejemplo, el cambio de uno solo de los cientos de bits que posee el mensaje de entrada genere la alteración total del mensaje cifrado.

4.4.4 Protocolos criptográficos

El comercio electrónico, tan común hoy en día, sería imposible sin el uso de esquemas criptográficos confiables. A través de internet se envía información confidencial que, de ser capturada por delincuentes, podría ser utilizada para cometer algún fraude. Si el número de tarjeta de crédito de un cliente de una tienda virtual pudiera ser capturado por

un intruso que hubiera intervenido el canal de comunicación, las consecuencias serían terribles para el usuario, para el banco emisor de la tarjeta y para la tienda virtual.

Funciones de un solo sentido

Para garantizar la confidencialidad de los datos en un contexto como el de comercio electrónico descrito arriba, se utilizan dos esquemas criptográficos. El primero se encarga de poner de acuerdo con una clave a la máquina cliente con la máquina servidor (la tienda virtual). El segundo usa esta clave para cifrar y descifrar los datos intercambiados durante la transacción.

Para poner de acuerdo a las dos máquinas con una clave común para ambas, sin que ésta viaje a través del inseguro canal de comunicación, se utiliza un concepto fundamental de la criptografía moderna: el de *función de un solo sentido*. El objetivo es tener una función a la que se la da un argumento de entrada y se obtiene fácilmente un resultado, pero es muy difícil tratar de deducir cuál fue la entrada a partir de la salida. Una manera de construir una función de un solo sentido es como sigue: es mucho más fácil multiplicar dos números para obtener otro, n, que factorizar n en los dos números originales. Los detalles de cómo hacer esto correctamente requieren del uso de la rama de las matemáticas llamada *teoría de números.*

Curiosidades
G. H. Hardy (1877-1947) fue un gran matemático inglés que hizo importantes contribuciones a la teoría de números, famoso por preferir que las matemáticas se mantengan puras y por haber dicho: "Nunca he hecho nada 'útil'. Ninguno de mis descubrimientos ha hecho o puede llegar a hacer, directa o indirectamente, para bien o para mal, la más mínima diferencia en el servicio del mundo." Es precisamente la teoría de números la que ha hecho posible la criptografía moderna.

El objetivo práctico de todo sistema criptográfico no es lograr la seguridad perfecta, que como se verá más adelante resulta impráctica, sino lograr la seguridad adecuada: es decir, que el enemigo tarde lo suficiente en descifrar el mensaje como para que ya no le sea útil. Que sepa el plan de batalla después de que ésta ha terminado, para decirlo en términos simples.

Para que el cliente y el servidor de una transacción comercial se pongan de acuerdo en una clave sin que ésta pueda ser capturada por el enemigo, cada uno de ellos aplica una función de un solo sentido a un argumento diferente e intercambian los resultados. Luego cada uno vuelve a aplicar la misma función, pero ahora sobre el argumento que recibió de su contraparte; así, ambos llegan al mismo resultado y el enemigo potencial que interviene el canal no puede saber cuáles eran los argumentos iniciales porque es difícil determinarlos a partir de los resultados que viajaron por el canal.

Un ejemplo más concreto sería elegir un catálogo de pinturas para muros. En las muestras de pintura hay tonos con diferencias muy sutiles, apenas perceptibles. Tratar de determinar cuál es la mezcla precisa de colores básicos que dan como resultado exacto uno de los colores del catálogo es difícil. Saber cuáles y en qué proporción mezclarlos para obtener un color particular es lo que implica una función de un solo sentido: es fácil mezclar colores, pero no es fácil, dado un color en particular, determinar cuál es la mezcla correcta que lo produce.

Supóngase que A y B desean ponerse de acuerdo en un color en particular sin que nadie más pueda deducir cuál es aun cuando esté observando lo que se intercambian A y B.

1] Tanto A como B se ponen de acuerdo en un color inicial del catálogo de colores que poseen en común, por ejemplo X. Tanto A como B toman una lata con capacidad para tres litros y ponen en ella un litro exacto de X.
2] A elige un color aleatorio del catálogo: Y, y agrega a su lata un litro exacto de Y. Ahora la lata contiene dos litros de la mezcla XY.
3] B elige un color aleatorio del catálogo: Z, y agrega a su lata un litro exacto de Z. Ahora la lata contiene dos litros de la mezcla XZ.

4] A y B intercambian sus latas. El posible enemigo sabe cuál color es *X*, pero ahora lo ve pasar mezclado con otro y no puede determinar con cuál.
5] A recibe la lata de B y le añade un litro de su propio color *Y*. A termina con tres litros de la mezcla *XYZ*.
6] B recibe la lata de A y le añade un litro de su propio color *Z*. B termina con tres litros de la misma mezcla *XYZ*.

Ahora A y B pueden usar el secreto con el que se pusieron de acuerdo como clave para cifrar sus futuras comunicaciones cifradas. Por supuesto, en la vida real el secreto es un número.

Datos, información, conocimiento y sabiduría

Hasta ahora hemos hablado de diferentes tipos de representaciones en función de su uso. Sin embargo, sólo se han mencionado representaciones de *bajo nivel*, por decirlo así. En esencia, se ha hecho referencia a cadenas de bits y de cómo estas cadenas pueden servir para representar datos de modo eficiente, seguro o preservando la integridad. ¿Cómo se llega a representar, usando sólo bits, reglas sofisticadas de diagnósticos médicos, estrategias para resolver problemas, programas de traducción de un lenguaje a otro, por ejemplo? En la convivencia cotidiana con los dispositivos de cómputo se suelen usar representaciones más próximas a las necesidades diarias. Se manipulan datos en una hoja de cálculo y no hay necesidad de percatarse de que los números que aparecen en ella se representan en binario al interior de la máquina. Las necesidades humanas de manipular, manejar e intercambiar datos demandan representaciones mucho más abstractas, de *alto nivel*.

Tómese una hoja de cálculo, de tipo Excel. ¿De dónde surge su enorme valor? Una hoja de cálculo está constituida por un arreglo bidimensional de celdas indexadas por filas y columnas. Cada una de estas celdas contiene un dato de un tipo: cadena de caracteres, número con cierta precisión, valor de verdad (falso o verdadero), valor monetario, hora, fecha, etc. Estas celdas pueden ser manipuladas individualmente o agrupadas para efectuar cálculos con ellas. Es posible procesar subconjuntos de ellas aplicándoles una fórmula (es decir, un algoritmo), ordenarlas de acuerdo con cierto criterio o filtrarlas separándolas en subconjuntos más pequeños. La hoja de cálculo responde a nuestra necesidad de manejar conjuntos de datos complejos, constituidos por elementos más simples de tipos diferentes, a los que se les asocian distintos significados, pero que poseen cierta unidad determinada por la estructura tabular.

Las tablas sirven para representar *relaciones* entre distintos datos. Cada columna o cada fila de una tabla se relaciona con las otras de alguna manera. Lo realmente importante en una hoja de cálculo es que, más allá de los datos concretos individuales que aparecen en cada celda, permite representar algo mucho más abstracto: una relación, un vínculo entre ellos. Manipular la hoja de cálculo y aplicar operaciones sobre su contenido permite transformar los datos en información útil que, a pesar de haber estado siempre allí en estado latente, sólo es posible extraer explotando la relación que existe entre los datos.

Este mismo concepto, el de representar vínculos abstractos entre series de datos, es el que se generaliza en lo que se conoce como *bases de datos*. Esencialmente, una base de datos consiste en almacenar muchos datos, manteniendo relaciones entre éstos. En las bases de datos clásicas, estas relaciones se mantienen mediante un conjunto de tablas diferentes, cada una de las cuales establece un vínculo directo entre los datos contenidos en ella, y luego se establecen vínculos indirectos entre los datos de distintas tablas. En una base de datos de control escolar de una institución educativa, por ejemplo, podría existir una tabla

en la que estén contenidos los datos personales de los alumnos: nombre, número de identificación, dirección, número telefónico, asignaturas en las que está inscrito, etc., y otra para las asignaturas mismas: nombre de la materia, clave, nombre del profesor y lista de alumnos inscritos. Ambas tablas están, por supuesto, vinculadas: a través de la primera se hace referencia a la segunda y viceversa. Una base de datos permite establecer varios niveles de relación y, por tanto, se potencia el poder de extraer información útil de ella, respecto al que se posee en una simple hoja de cálculo.

¿Dónde está la vida que hemos perdido en vivir?
¿Dónde está la sabiduría que hemos perdido en conocimiento?
¿Dónde el conocimiento que hemos perdido en información?
El primer coro de la roca, de T. S. Eliot.

Probablemente, hoy más que nunca es evidente el poder que se puede obtener mediante representaciones adecuadas de datos. Internet existía mucho antes de que se inventara el concepto de hipertexto y la red –o *web*–, la telaraña mundial (*world wide web*); sin embargo, su increíble magnitud actual, la versatilidad de sus aplicaciones y el volumen de información que contiene sólo pueden explicarse a partir de éstos. El cambio esencial lo constituyó la creación de una manera de representar datos de tal forma que fueran accesibles para un gran público y que permitieran establecer relaciones entre ellos. En la red, una página tiene ligas a otras páginas relacionadas.

En HTML (Hypertext Markup Language) y sus parientes cercanos, los lenguajes de marcado en general (como SGML o XML), el concepto es simple: establecer una manera de asignar a los datos un significado, asociarles cierto comportamiento dependiendo del tipo que representan. Poder decir que una cadena de texto es realmente una referencia a un documento que se encuentra en otro lugar, que es el título del documento o que posee un formato específico al desplegarse, es lo que concede utilidad a los lenguajes de marcado. Hacen posible que lo que de otro modo sería sólo un texto lineal posea cualidades que lo elevan a un nivel de abstracción mucho más cercano a lo que necesitamos los seres humanos para extraer información.

Ilustración 1. Diferencia entre información, conocimiento y sabiduría.

4.5 RESUMEN Y CONCLUSIONES

Los datos, que son la materia prima del cómputo, pueden ser de cualquier tipo: números, imágenes, música, videos, libros, estadísticas, mapas, etc. A partir de los datos se obtiene información y, con ésta, conocimiento y finalmente sabiduría. Pero en el fondo, todo son bits, es decir, secuencias de 0 y 1. Los sistemas de cómputo no hacen más que procesar y comunicar información representada mediante secuencias de bits.

4.5.1 Tiempo contra espacio

Un rasgo notable de los mecanismos de representación es el hecho de que el tiempo y el espacio son indistinguibles; es decir, para las ramas de la computación que hemos abordado en este capítulo, es completamente intrascendente si la representación de los datos será usada para almacenarlos y poder recuperarlos en el futuro, o para enviarlos a través de algún medio de transmisión a un destino remoto. En el primer caso los datos viajan a través

del tiempo, en el segundo a través del espacio, y para la teoría de la información o la criptografía es irrelevante cuál de las dos opciones se utiliza, lo importante es que el medio, ya sea de almacenamiento o de transmisión, es: *a*] el recurso que debemos optimizar, en el caso de las representaciones eficientes; *b*] lo que es intervenido por el enemigo, en el caso de la criptografía, y *c*] lo que es susceptible de introducir errores en los datos, en el caso de las representaciones orientadas a preservar la integridad de los datos.

4.5.2 Limitaciones de codificación

Otro rasgo notable es el hecho de que en todos los mecanismos de representación existe lo que se podría llamar *la perfección*, que es inalcanzable desde el punto de vista práctico. Vale la pena aclarar este punto: Shannon demostró que es posible formular un código tan eficiente como se quiera, es decir, cuya longitud promedio de palabra sea tan próxima a la entropía como se desee, si se está dispuesto a pagar el precio de tener un modelo estadístico muy preciso de los datos que se desea representar. En este caso sólo se usó lo que se denomina *estadística de primer orden*, utilizando sólo la probabilidad o frecuencia de cada objeto por representar, sin relación alguna con los demás. Si se pensara en la probabilidad de que algo ocurra *luego de que ha ocurrido alguna otra cosa,* se estaría hablando de estadísticas de segundo orden, que proporcionan una idea mucho más clara del comportamiento de los datos. Ya que en cada idioma existe una probabilidad característica para cada letra del alfabeto, con base en esto se puede crear un código eficiente para representar textos en español, codificando cada letra. Pero si además se toma en consideración la probabilidad de que ocurra cada una de las letras, suponiendo que la inmediata anterior fue alguna otra, entonces se tendrá una idea más clara de las características del español. En este idioma la aparición de una "s" después de una "e", por ejemplo, ocurre 2.4% de las veces, "es" es la pareja de letras más frecuente, en inglés en cambio la pareja más frecuente es "th", que ocurre 2.7% de las veces, mientras que "es" ocurre sólo 1%. Se puede continuar con este proceso, considerando cuál es la probabilidad de que aparezca una "r" luego de una secuencia "pe", por ejemplo, y cuanto mayor sea el número de letras en la secuencia "histórica" que se considere, mayor será la precisión y, como demostró Shannon, la eficiencia del código para representar los datos. Por supuesto, si se trata de guardar la tabla de probabilidades del alfabeto de 29 letras del español, el modelo de orden 1 es factible, pero para un modelo de orden 4, con 29^4 probabilidades por cada una de las 29 letras, ya no es económico. Se termina gastando más al guardar el modelo que al guardar los datos con una representación trivial de bloque de cinco bits por letra.

4.5.3 Limitaciones de la criptografía

En el caso de la criptografía ocurre algo similar. Está demostrado que la seguridad perfecta existiría, pues se puede hacer algo completamente indescifrable si se logra que:

a] la clave sea generada de manera completamente aleatoria,
b] la clave sea de la misma longitud que el mensaje,
c] la clave se use una, y sólo una, vez.

La tarea de generar números perfectamente aleatorios en sí ya es imposible. Pero en términos prácticos es suficiente generar números *seudoaleatorios seguros*, lo cual quiere

decir que a cualquier algoritmo razonablemente eficiente le es imposible adivinar un bit de la clave con probabilidad significativamente mayor a un medio. Pero inclusive contando con un algoritmo para generar números seudoaleatorios seguros (para lo cual existen varias propuestas), un sistema como éste sería de poca utilidad. Que la clave sea tan larga como el mensaje y se use una sola vez implica que tanto quien cifra como quien descifra deban guardar, en principio, una clave demasiado grande que sólo les servirá para un mensaje, y punto. Esto, en su conjunto, resulta bastante impráctico, y se usa en muy pocas situaciones.

Se ha demostrado también que, dado un canal cuya tasa de error se conoce —es decir, se sabe la frecuencia con la que echa a perder un bit de la transmisión—, es posible diseñar un código que tenga la capacidad de detectar y corregir los errores que se desee. Es decir, se puede construir una manera de codificar los datos de tal forma que el número de errores que pasan inadvertidos sea tan bajo como se quiera. Pero esto tiene un costo en la capacidad de expresividad del código: se puede añadir tanta redundancia como sea necesario, pero entonces el número de bits empleados en cada palabra excede rápidamente el que sería estrictamente necesario para decir el número de cosas que se desea expresar. En el extremo podría ocurrir que sólo 0.0001% de los bits equivocados pasen inadvertidos, en un canal muy malo, pero usando 32 bits para decir sólo cuatro cosas, por ejemplo.

4.5.4 Derechos de autor

Las ramas de la computación que se han tratado hasta ahora son algunas de las que han recibido mayor atención y han resultado ser más útiles. En buena medida, el comercio por internet es posible porque existe la criptografía que impide que alguien capture fácilmente el número de tarjeta de crédito de otra persona y lo utilice para su beneficio. Las comunicaciones modernas, todas ellas, son impensables sin códigos detectores y correctores de errores, y la enorme cantidad de información contenida en una canción o en una imagen no podría ser manejable sin representaciones eficientes.

La criptografía también es la base de las soluciones de protección de derechos de autor de información. Tradicionalmente, hablando por ejemplo de libros, un autor pone una advertencia en su libro para evitar que se fotocopie y se vendan copias más baratas sin pagarle regalías. En el mundo moderno de la computación se ha visto cómo todo es información, y el libro en sí no es lo que vale, sino el texto, es decir, la información que contiene. Y cualquiera puede reproducirla tantas veces como quiera y enviarla por internet a miles de personas con unos simples movimientos de sus dedos sobre el teclado. Lo mismo se puede hacer con canciones o películas. Parecería que evitar esto es imposible. Es decir, hágase lo que se haga, una vez que una persona está escuchando una canción, puede perfectamente grabarla y reenviarla a quien quiera. Podría pensarse, por ejemplo, en encriptar una canción antes de venderla; algo similar se ha hecho para protección de películas en DVD, con las restricciones para reproducir solamente en una región dada del mundo. Como bien se sabe, estos esfuerzos han fracasado, siempre es posible eliminar esas protecciones. Sin embargo, existen técnicas criptográficas creativas que intentan resolver el problema de otro modo, mediante las llamadas *marcas de agua*.

La idea consiste en tomar la información que se desea proteger, como puede ser una imagen, e introducir cambios imperceptibles al ojo humano, una especie de ruido con cierto patrón, que permiten identificar al creador de la imagen.

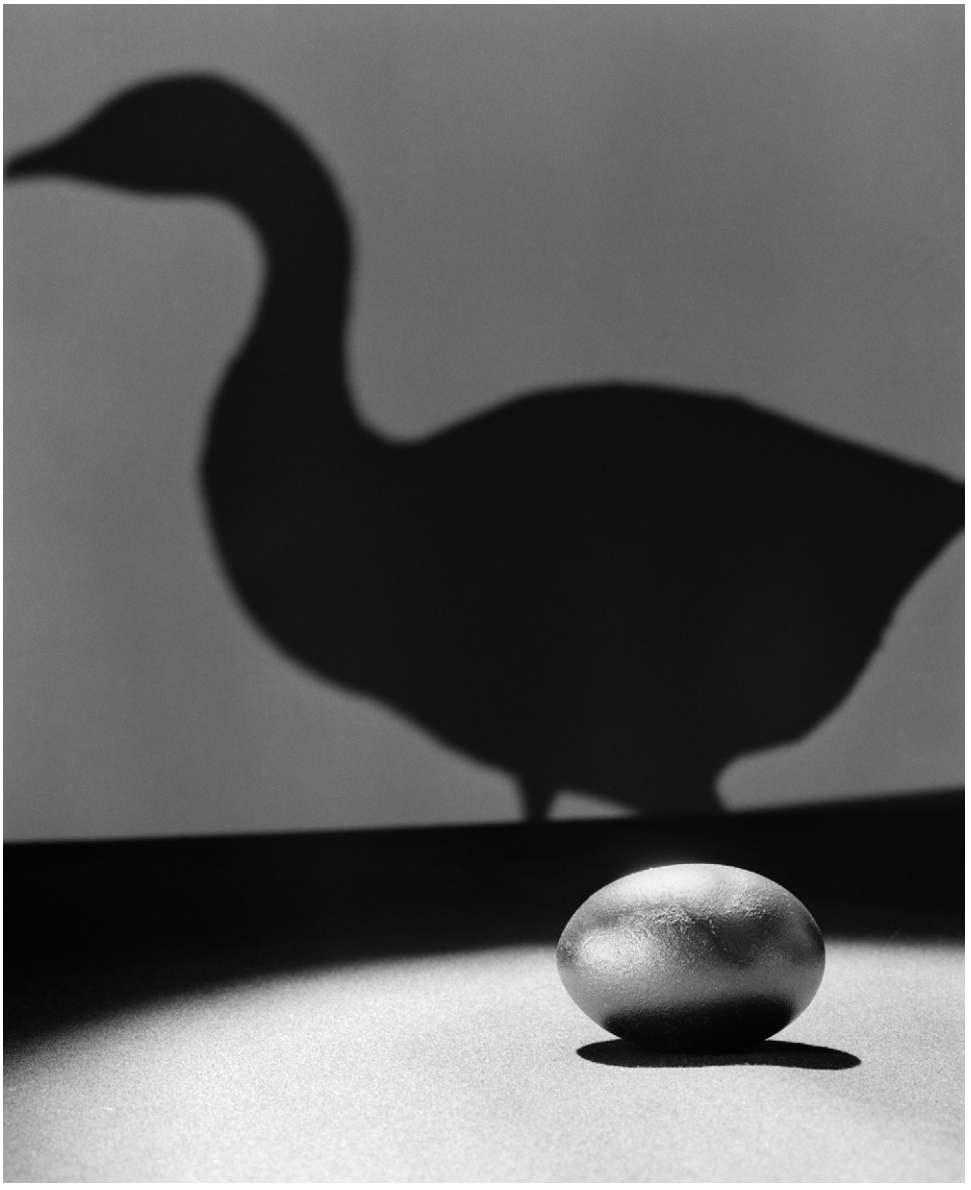
De cualquier modo, el mundo de la información y la computación ha reavivado una intensa polémica acerca de la noción de derechos de autor, y ha hecho que la industria

del cine y la música empiece a pensar en modelos diferentes de negocios para ganar dinero, no directamente por medio de la venta de discos, sino a partir de, por ejemplo, servicios. Lo mismo sucede con los creadores: cada vez más músicos y escritores ponen sus obras en internet a disposición de todos, sin costo. Está abierta la discusión: ¿es buena o mala la piratería? ¿Qué significa piratería en el contexto del mundo moderno de la información?

ABSTRACCIÓN

TEMA

5



© *Latin Stock México.*

Arcadio estaba muy emocionado porque su tío Aureliano le prometió que cuando lo fuera a visitar a Estados Unidos le permitiría manejar su coche convertible. Arcadio llegó ya muy noche del aeropuerto a casa de su tío, pero al día siguiente lo primero que hizo al despertarse fue convencerlo de salir a pasear en el convertible.

Abstracción es la actividad por excelencia en la ciencia de la computación, la herramienta intelectual que permite a los científicos de la computación expresar su entendimiento de un problema, mantener la complejidad manejable, y seleccionar el nivel de detalle y grado de generalidad que necesitan en el momento.

COMPUTER SCIENCE: REFLECTIONS ON THE FIELD, 2004.

—Qué bien se maneja este coche, tío.

—Sí, pero no quites los ojos de la calle, que está nuevecito.

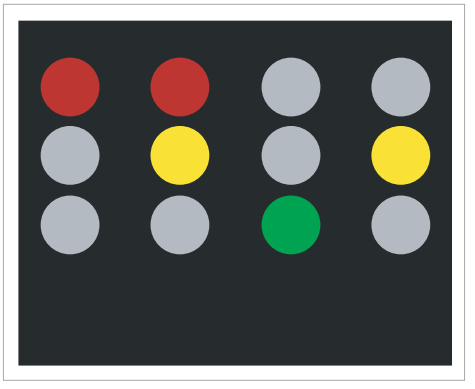
—Claro tío, soy muy cuidadoso al manejar.

—¡¿Qué haces?! ¡Aaaaaaah! —gritó Aureliano cuando Arcadio reaccionó al ver que el semáforo cambiaba de verde a amarillo, y el coche que venía atrás casi se estrella contra ellos, dando un fuerte enfrenón y rechinando las llantas.

Curiosidades
La mayoría de los semáforos del mundo consisten de tres luces, que cambian en el orden rojo, verde, amarillo. Pero en algunos países, como el Reino Unido, Polonia y Alemania, hay semáforos que van de rojo, a rojo con amarillo (para indicar que ya viene el verde), a verde y luego a amarillo.

Arcadio no sabía que en Estados Unidos se permite pasar un semáforo en amarillo, pensó que era como en México, en donde podría ser infracción y motivo de multa. De manera que tomó por sorpresa al coche de atrás, que no se esperaba que Arcadio fuera a frenar en ese momento.

A un nivel de abstracción, un semáforo consiste en tres estados diferentes, cada uno con un significado particular. Es irrelevante cómo se implementa cada estado a este nivel de abstracción. No sólo es irrelevante, es útil no especificar la implementación de los estados, ya que permite evitar distracciones y concentrarnos en la esencia de interés: tres señales diferentes para control de tráfico. Además, permite la independencia de distintas implementaciones. Por ejemplo, una implementación puede ser en luces roja, verde y amarilla, pero no se limita a éstas. Se puede construir un semáforo para ciegos, usando sonidos, o uno para personas con problemas para ver colores con luces de distintas formas, que satisfagan las mismas propiedades de control de tráfico. Cada una de estas implementaciones tiene sus propios detalles y dificultades, que son independientes del significado en sí de cada estado del semáforo.



Pero, ¿cómo explicar la semántica del semáforo? Es decir, ¿qué significa cada uno de los tres estados? Se necesitan dos modelos: por un lado, el de los tres estados y cómo cambian de uno a otro; por otro, el reglamento de tránsito, que no es más que otro modelo que asigna significado al primer modelo que, en este caso, es un crucero, avanzar o no, etcétera.

En esta historia de Arcadio se presenta la problemática de la abstracción. Se tienen dos modelos y se quiere razonar acerca de la relación entre ellos. Ya sea que uno sirva para definir la semántica del otro o para implementarlo, obteniendo beneficios como facilitar su construcción (por ejemplo, usando piezas como focos, disponibles en cualquier tlapalería) o posibilitar distintas implementaciones (para ciegos y daltónicos), y permitiendo la verificación de una implementación de manera no ambigua: un policía debe poder decidir si un auto cometió una infracción o no analizando la relación entre dos modelos: el reglamento de tránsito y una situación real de un auto cruzando un semáforo. Se posibilita así la construcción de sistemas complejos, pero el fabricante de focos rojos puede concentrarse en su propia labor sin preocuparse de cómo se usará su producto.

La abstracción posibilita el razonamiento acerca de las propie-dades de los sistemas. ¿No se facilitaría la construcción del semáforo dejándolo siempre en verde? Claro, el semáforo en sí es parte de una abstracción de más alto nivel, de la cual surge el requerimiento de que tenga tres luces que vayan rotando: un cruce de calles. El sistema en cuestión no contiene uno sino varios semáforos para controlar el tráfico en el crucero. Y los semáforos son una sola entre las posibles implementaciones del control de tráfico del crucero. De ahí proviene una preo-cupación a la que el computólogo le ha dedicado libros enteros: la justicia. Sin ésta sería muy fácil resolver el problema: simplemente se deja en verde uno —y sólo uno— de los semáforos del crucero, siempre. Otro tipo de razonamiento que se puede hacer gracias a la abstracción permite analizar si un modelo es implementable. ¿Qué tal que no haya manera de cumplir con el reglamento de tránsito? ¿Que algunas de sus reglas

sean demasiado estrictas, ambiguas o contradictorias? Para razonar acerca de las relaciones entre modelos, la herramienta adecuada es la lógica matemática, como siempre que es necesario hacer razonamientos complejos de manera cuidadosa. Se verá que es posible usar distintas lógicas, con distintos poderes, diferentes capacidades de expresar propiedades, según el modelo en cuestión y los razonamientos requeridos.

5.1 LA ABSTRACCIÓN

5.1.1 Los inicios: sentido de número y contar

B. Mazur, T. Dantzig y J. Mazur[1] afirman que si se ha de juzgar el desarrollo de nuestros ancestros remotos por el estado mental de las tribus contemporáneas no se puede evadir la conclusión de que el comienzo fue modesto. Fue a partir de un rudimentario sentido aritmético, no mucho más visionario del que poseen los pájaros, que se desarrolló el concepto de número. Y no hay muchas dudas con respecto a que, basándose exclusivamente en esta percepción directa de lo que es un número, el hombre no hubiera podido avanzar en el arte del cómputo mucho más allá de lo que los pájaros lo han hecho.
Sin embargo, a través de una serie de circunstancias sobresalientes, el hombre ha aprendido a superar su excesivamente limitada percepción de número a través de un artificio que estaría destinado a ejercer una tremenda influencia en su vida futura. Este artificio es la abstracción y, en el caso particular de los números, el conteo.

Hay lenguas primitivas que cuentan con palabras para cada color del arco iris pero no existe una para color; hay otras en las que existen palabras para todos los números pero no una para número. Lo mismo ocurre con otros conceptos. La lengua inglesa es muy pródiga en expresiones nativas para tipos particulares de colecciones: flock (manada), herd (hato, rebaño), set (conjunto), lot (mucho) y bunch (manojo) se aplican en casos especiales; sin embargo, las palabras *collection* (colección) y *aggregate* (agregado, total) son de origen extranjero.

Lo concreto precede a lo abstracto. "Requiere muchas eras descubrir —dice Bertrand Russell— que una pareja de faisanes y un par de días son, ambos, ejemplares del número 2." Hoy en día se tienen múltiples maneras de expresar la idea del 2: pareja, par, apareamiento, matrimonio, etcétera.

Un ejemplo ilustrativo de la extrema concreción del concepto antiguo de número se puede ver en la lengua de la tribu thimshi de Columbia Británica, Canadá. En ella hay siete diferentes conjuntos de palabras para los números: uno para los objetos planos y los animales; una para los objetos redondos y el tiempo; uno para contar hombres; uno para árboles y objetos largos; uno para canoas;

Curiosidades
Probablemente nunca se sepa lo que indujo al hombre hace 30 000 años a penetrar con una antorcha en las entrañas de la profunda cueva de Lascaux, en la región de la Dordogne francesa, y cubrir de bellísimas pinturas sus paredes. Pero existe la seguridad de que alguna noción y necesidad de contar tenía, al ver las pinturas donde aparecen puntos en diversas configuraciones. En algunas aparecen 13 o 29 puntos, indicando su probable interés por el ciclo lunar de 29 días.

Imagen prehistórica | *© Latin Stock México.*

Curiosidades
La Columna de la Independencia de la ciudad de México fue inaugurada el 16 de septiembre de 1910. En enero de 1906 se habían terminado la base, el zócalo y el pedestal. Una vez iniciada la construcción, y cuando ya se habían colocado más de 2 400 piedras de cantera, la cimentación original no pudo soportar el gran peso del monumento y la columna empezó a perder verticalidad por el hundimiento. Se decidió demoler lo construido y se diseñó una nueva cimentación. La abstracción en la que se basó el diseño inicial de los cimientos no fue la adecuada. Este tipo de problema suele ocurrir en computación, en la construcción de sistemas de software, por ejemplo. Siempre es mejor desechar un sistema cuando su diseño fundamental, la abstracción que constituye su "cimentación", no es la adecuada. Es más barato tirar el trabajo hecho que tratar de "enderezar" el sistema corrigiendo los problemas que surgirán durante su elaboración.

Ángel de la Independencia | *© Ulises00.*

[1] Barry Mazur, Tobías Dantzig y Joseph Mazur, "Number Sense in Humans", *Pi Press*, 2005.

uno para medidas y otro para contar objetos no definidos en los grupos anteriores. Este último probablemente es de desarrollo ulterior; los primeros deben de ser reliquias de tiempos remotos, cuando los miembros de la tribu aún no habían aprendido a contar.

Es mediante contar que se consolida la hasta entonces concreta y, por tanto, heterogénea noción de pluralidad, característica del hombre primitivo; es el concepto abstracto y homogéneo de número lo que hace posible las matemáticas y, más en general, el pensamiento simbólico y la computación como procesamiento de información.

5.1.2 Abstracción: el camino del conocimiento

El mundo es un lugar complicado, hasta el objeto más trivial o el fenómeno más cotidiano resulta ser todo un reto si se pretende analizar detenidamente. Para cada cosa hay una multiplicidad de lecturas posibles que dependen de lo que se quiera leer, de las preguntas que se deseen responder a propósito de ella. Esta complejidad es inherente al mundo, en cada hecho y en cada objeto concurren un sinfín de otros hechos u objetos que les dan origen; y los primeros, a su vez, influyen en otra multitud de hechos u objetos que devienen de ellos. Cada objeto, por insignificante que sea, cada inocuo acontecimiento, resulta ser el punto de cruce de una infinidad de caminos. El mundo es, como querría Jorge Luis Borges, el jardín de los senderos que se bifurcan.

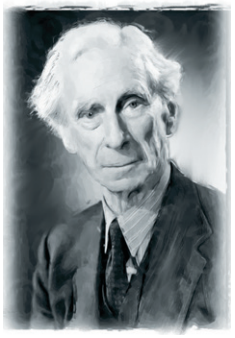
Enfrentar el mundo tal cual es, con el objeto de conocer algo acerca de él, sería impensable sin la abstracción. Nada se podría saber acerca del movimiento de una canica que rueda por la superficie de una mesa si, en la obsesión por los detalles, se centrase la atención primero de las rugosidades de la mesa, las fibras del cedro de que está hecha, la densidad del aire, la temperatura del ambiente, las cualidades de refracción del vidrio de la canica y los fotones que chocan con ella. En estricto sentido, todo lo mencionado determina el movimiento real de la canica, pero descubrir la influencia que cada cosa tiene sobre el fenómeno observado es muy complicado y, además, en muchos casos intrascendente. De haber considerado relevante la temperatura de Pisa en una tarde de verano, probablemente Galileo hubiera muerto sin entrever la gravitación.

La abstracción determina las lecturas posibles de cada hecho y de cada objeto. En función de lo que se pretenda descubrir, la abstracción indica qué debe ser considerado relevante y qué debe despreciarse. Luego de comer con un amigo el otro día, él tomó un palillo mondadientes. Difícilmente se puede pensar en un objeto más humilde, una pequeña astilla de madera. El amigo no lo usó para lo que fue elaborado, jugueteó con él entre los dedos, de hecho habló brevemente acerca de la longitud del palillo y el punto en el que hay que colocarlo para que se equilibre sobre un dedo. Analizado desde el punto de vista de un ambientalista, lo relevante será, probablemente, que está hecho de madera y puede discutirse acerca del impacto ambiental que tiene la fabricación de palillos. El mercadólogo discurrirá acerca de la preferencia del público por los palillos de madera en contraste con los de plástico.

En el lenguaje cotidiano, a veces se usa el término abstracción en el sentido opuesto al que posee en realidad. Cuando algo parece complicado, ininteligible, se suele decir que "es muy abstracto". Pero en realidad la abstracción es el filtro usado para quedarse con lo que se considera esencial de un fenómeno, quitando toda complicación innecesaria.

El conocimiento humano se divide en diversas disciplinas porque cada una de ellas se enfoca en un aspecto particular de la realidad, cada una posee sus propias abstracciones. Aun dentro de una misma disciplina existen diversas abstracciones: en la teoría de la gravitación, a un físico le interesa la masa de los cuerpos presentes en el fenómeno observado;

Bertrand Russell (1872-1970) Uno de los filósofos y matemáticos más distinguidos del siglo XX, Premio Nobel de Literatura en 1950. Escribió sobre una amplia gama de temas, desde los fundamentos de las matemáticas y la teoría de la relatividad hasta el matrimonio, los derechos de las mujeres y el pacifismo. Su gran contribución a las matemáticas es la indudablemente importante *Principia Mathematica*, obra escrita en tres volúmenes con Alfred North Whitehead, donde a partir de ciertas nociones básicas de la lógica y la teoría de conjuntos se deduce la totalidad de las matemáticas, mostrando así el poder de los lenguajes formales, la posibilidad de modelar las matemáticas y la fertilidad de la lógica.



Curiosidades
Se supone que el máximo número de objetos que una persona puede contar al verlos es siete. De ahí que para algunas culturas decir "siete" es decir "muchos", como "los siete mares".

en electromagnetismo le interesarán las cargas de los mismos. Ciertamente todas esas propiedades coexisten simultáneamente en los cuerpos, pero dependiendo de qué aspecto interese investigar se mantienen unas y se ignoran otras. De hecho, ¡cada una de estas propiedades en sí es una abstracción!

We always did feel the same, we just saw it from a different point of view.
BOB DYLAN, TANGLED UP IN BLUE.

Una buena abstracción es como una buena caricatura. En el periódico se pueden ver caricaturizados los personajes de la vida pública del país. No son retratos perfectos ni pretenden serlo, los dibujos acentúan características de los personajes reales. Es posible reconocer a cada uno de ellos sin mayor dificultad porque el caricaturista ha sabido captar alguna cosa interesante, que resulta esencial de la persona.

5.1.3 Abstracción en computación

En computación, aun más que en las demás áreas del conocimiento, la abstracción es fundamental; juega un papel muy especial debido, entre otras cosas, a su naturaleza dual. Por un lado, la computación tiene como meta construir, al igual que la ingeniería: construir máquinas, programas y sistemas que realicen cómputo. Por otro lado, al igual que la física, pretende observar la naturaleza (real o imaginaria) y entender los procesos de cómputo que ahí se realizan. Por lo tanto, la abstracción se lleva a cabo en dos direcciones opuestas.

El pintor puede y debe abstraer muchos detalles al crear su pintura. Toda buena composición es sobre todo un trabajo de abstracción.
DIEGO RIVERA.

Abstracción para construcción

Para la construcción de objetos de cómputo, la abstracción se usa de diversas maneras:

Especificación de problemas | Al resolver un problema, lo primero que hay que hacer es especificarlo. Esto es, dar un modelo general del comportamiento de cualquier solución a éste. Una solución al problema podría ser un algoritmo, cuyo comportamiento satisface la especificación dada. El algoritmo representa entonces otro modelo, que incluye más detalles que los del modelo de la especificación del problema. Se requiere de técnicas para verificar que el algoritmo satisface la especificación.

Lenguajes de programación | Al programar el algoritmo en algún lenguaje, por ejemplo Scheme, se tendrá un tercer modelo, con más detalles aún. Un lenguaje es muy preciso, tanto en su sintaxis como en su semántica. Por ejemplo, si el lenguaje requiere que cada vez que se abra un paréntesis se deba cerrar, y se olvida cerrar uno, el programa ya no corre. Mientras que un algoritmo es más cercano al lenguaje natural, con la correspondiente flexibilidad y posibles ambigüedades.

Semántica | Pero, ¿cómo saber cuál es la semántica de cada instrucción de Scheme? Es decir, una instrucción no es más que una secuencia de símbolos, como "(+ 3 *x*)". Es necesario explicar el significado de esta secuencia. En palabras se puede explicar que el efecto de la instrucción es sumarle al valor de *x* el número 3. Se ve entonces que se está pensando en una asociación de la secuencia de símbolos a un efecto, en algún modelo de cómputo. Así que se requiere de abstracción para definir el significado de un lenguaje de programación, relacionando dos modelos, el del lenguaje con el de su ejecución. Toda una rama

de la computación se dedica al estudio de cómo especificar el significado de un lenguaje de programación.

Lenguajes de distintos niveles | Y ahí no termina la historia. Las instrucciones que puede ejecutar una computadora no son directamente las de Scheme; obviamente, existen muchos otros lenguajes de programación. El programa en Scheme se debe implementar en las instrucciones de máquina, que pertenecen al lenguaje que el chip procesador de la computadora sabe ejecutar en hardware. Pero estas instrucciones son demasiado simples; por ejemplo, se refieren directamente a localidades de la memoria, y no a variables cuyos nombres puede elegir el programador. Así que se usa un lenguaje intermedio, llamado ensamblador, que facilita la tarea de escribir un programa en lenguaje de máquina.

Independencia de plataformas | Es deseable que una vez escrito un programa en Scheme, se pueda ejecutar en cualquier computadora, y cada una puede tener un procesador de otra marca, con diferentes instrucciones de máquina. El lenguaje de programación Java ganó popularidad muy rápidamente porque proveía de un modelo intermedio de computadora, llamado máquina virtual, de manera que un programa escrito en Java se tradujera en uno más detallado que corriera en esta máquina virtual, y luego se encargara uno de traducir la máquina virtual a cada plataforma de cómputo diferente. El costo es una pérdida en eficiencia, ya que no se puede traducir directamente el programa en Java a una máquina específica y aprovechar sus peculiaridades, pero la ganancia es que se pueden desarrollar programas que corran en cualquier computadora.

Reusar componentes | Y se podría seguir y seguir con esta historia, ya que los programas y computadoras modernas han llegado a niveles de complejidad tan enormes, que se usan más y más capas, tanto hacia arriba como hacia abajo. Hacia arriba se usan construcciones útiles a muchos programas: por ejemplo, ya que todos usan elementos de la interfaz de usuario similares —como ventanas, ratón, menús...—, se utilizan herramientas que evitan tener que programarlas cada vez, o bibliotecas de soluciones a problemas comunes a diversas situaciones, como podría ser un módulo para ordenar objetos.

Domar la complejidad | Hacia abajo se requiere abstracción para manejar la enorme complejidad del hardware moderno de una computadora. Por ejemplo, en realidad la memoria está organizada en capas; las memorias más grandes son las más baratas, pero más lentas; las más rápidas son más caras, así que son más pequeñas. Se requieren complejos sistemas de administración de la memoria, que se encargan de mover los datos de una capa a la otra, de la manera más transparente posible para el programador. El sistema debe proveer un modelo que le permita abstraerse de los detalles sobre cómo se hace esto.

Abstracción para análisis

Al igual que en la física, durante la observación de un fenómeno en la naturaleza se abstraen detalles y se produce un modelo que incluye las características esenciales de interés. Las dos maneras más comunes de definir modelos de cómputo es mediante máquinas

y con lenguajes. Con frecuencia se desea diseñar ambos tipos para modelar el mismo fenómeno, ya que cada uno tiene sus propias ventajas. Entonces se usa la herramienta para razonar acerca de abstracciones a fin de que ambos modelos sean equivalentes. Más adelante se verá cómo es posible definir distintos modelos de cómputo, unos más poderosos que otros, unos equivalentes a otros en algún sentido, y en otro no. Lo que conduce una vez más al problema central de la abstracción: con dos modelos dados, en este caso de máquinas de cómputo, cómo razonar acerca de su relación.

Desde los inicios de la computación moderna se ha usado abstracción para analizar fenómenos relacionados con el cómputo. La "prueba de Turing" es el nombre con el que se conoce la propuesta que hizo Alan Turing en 1950 para estudiar la pregunta de si las máquinas pueden pensar. En su artículo "Computing Machinery and Intelligence" (disponible en internet), propone considerar la pregunta: "¿pueden las máquinas pensar?".

Se debe comenzar con definiciones de los términos "máquina" y "pensar", para lo cual Turing toma una posición similar a la de Descartes, en la que en lugar de adentrarse en discusiones acerca de lo que es la inteligencia, simplemente considera que algo es inteligente cuando se comporta como si lo fuera. Es decir, usa las nociones de abstracción que se han mencionado para proponer un experimento como el "juego de la imitación": una persona, el interrogador, se encuentra en un cuarto completamente aislado, sólo con un teclado típico de letras y números para escribir preguntas dirigidas a los otros dos participantes del juego, y una pantalla donde observa las respuestas a sus preguntas. Uno de esos participantes es una persona, y el otro es una computadora. El objetivo del interrogador es adivinar cuál de los otros dos participantes es la persona y cuál la computadora. La única interacción entre el interrogador y los participantes es mediante las preguntas tecleadas y las respuestas escritas en la pantalla. Si la computadora logra engañar al interrogador, se decide que es inteligente.

La prueba de Turing ha sido el centro de muchas discusiones en filosofía, ciencias cognitivas e inteligencia artificial, en muchos libros y artículos escritos durante los últimos 50 años. Independientemente de las discusiones al respecto, de la validez e implicaciones de la prueba, provee un bello ejemplo de cómo se usa la abstracción en computación. Se considera una "caja negra" con entradas y salidas, y se define un comportamiento válido para la caja, en términos de la relación que las entradas y las salidas deben guardar: si la entrada es tal, la salida debe ser tal o cual, etc. Una implementación de la caja se considera *correcta* si, y sólo si, a cualquier entrada corresponde una salida válida. Cómo está construida, cómo lo logra, lo que hay adentro de la caja se ignora.

Esto es similar a las discusiones acerca de si las computadoras saben jugar ajedrez, el rey de los juegos. Desde siempre, las personas que presumían de grandes capacidades mentales se medían por su habilidad para jugarlo. Existen interesantes discusiones filosóficas al respecto, pero el hecho es que hoy en día ya hay programas a los que ninguna persona "normal" les puede ganar un solo partido (únicamente un profesional del ajedrez puede hacerlo).

5.2 MODELOS DE CÓMPUTO

Preguntarse si las máquinas pueden pensar es como preguntarse si los submarinos saben nadar.
EDSGER W. DIJKSTRA.

Mucho de lo referente a computación es acerca de máquinas de estado. Las máquinas de estado son para los computólogos lo que las ecuaciones son para los físicos. Casi cualquier fenómeno que tenga que ver con computación puede ser descrito mediante una

máquina de estado. De hecho, como dice Leslie Lamport, dado que las máquinas de estado pueden ser descritas y manipuladas mediante matemáticas ordinarias y cotidianas, es decir, conjuntos, funciones y lógica sencilla, éstas proveen un marco de trabajo uniforme para estudiar el cómputo de manera formal, con matemáticas sencillas.

5.2.1 Máquinas de estados finitos

—Hola, Úrsula, mira lo que compré en el mercado de antigüedades.

—¿Para qué sirve? Es sólo una caja con una luz verde, un botón de reset y un micrófono —le contesta a Arcadio.

—Pues ni idea. Sólo sé que cuando le dices palabras, a veces se prende la luz verde y a veces no.

—Pues ábrela y mira qué tiene adentro.

—Traté pero no pude. No tiene tornillos, está muy bien sellada.

—A ver —dice Úrsula—, le voy a decir algo: hola.

La luz verde no se prende. Después de varios intentos, como "perro", "loca" y "Arcadio", para los cuales la luz no se prendió, Úrsula dice:

—Creo que tu caja no sirve… a ver, haz otro intento: "Francia". ¡Se prendió la luz!

—París. Uy, no se prendió. España. ¡Se prendió otra vez!

—México. No se prendió. Canadá. Tampoco. China. Tampoco.

—Italia. ¡Sí, otra vez se prendió!

Después de jugar un buen rato con la caja, deciden que sirve para indicar los países de la Unión Europea, ya que cuando le dicen a la caja un país de otro continente la luz no se prende.

Los computólogos le llaman a esto "caja negra", ya que su apariencia no da ninguna pista acerca de su función, de su comportamiento, ni de cómo está construida, o de qué materiales. Pero ya sea que su mecanismo esté hecho de engranes de metal, de circuitos electrónicos o de un ratón muy listo, el hecho es que se debe suponer que entre palabra y palabra, cuando no hay actividad, sus partes internas se estabilizan configurando un estado, sin saber qué signifique exactamente esto (podría ser niveles de voltaje o posiciones de los engranes, por ejemplo). Lo cierto es que se puede suponer que el número de estados posibles de la caja es finito; un cierto número, quizá grande, pero finito, ya que la caja en sí es finita. Al inicio, la caja se encuentra en uno de esos estados, considerado el inicial. El botón de *reset* sirve precisamente para regresar a la máquina a este estado. Ahora bien, cuando le dicen una palabra a la caja, lo único que se puede saber es que produce una salida (prende o no la luz) y cambia a otro estado (que podría ser el mismo en el que estaba). Así que, aparentemente, sin necesidad de saber nada acerca de la caja negra, se puede describir su comportamiento mediante un autómata finito, es decir, una gráfica cuyos vértices son los estados, y arcos dirigidos de un vértice a otro etiquetados con las entradas y salidas posibles.

Como ya se ha discutido en el tema 4, se puede suponer que tanto las entradas como las salidas son bits que podrían estar codificando palabras, sonidos o cualquier otra cosa. Ignorando de momento qué representan; a final de cuentas, es sólo una caja negra, y no hay manera de saber su "intención". Solamente se obtienen secuencias de salidas al darle entradas. En la figura 1 aparece un autómata de siete estados; sobre cada arco está indicada la entrada que ocasiona que transite al estado al final del arco.

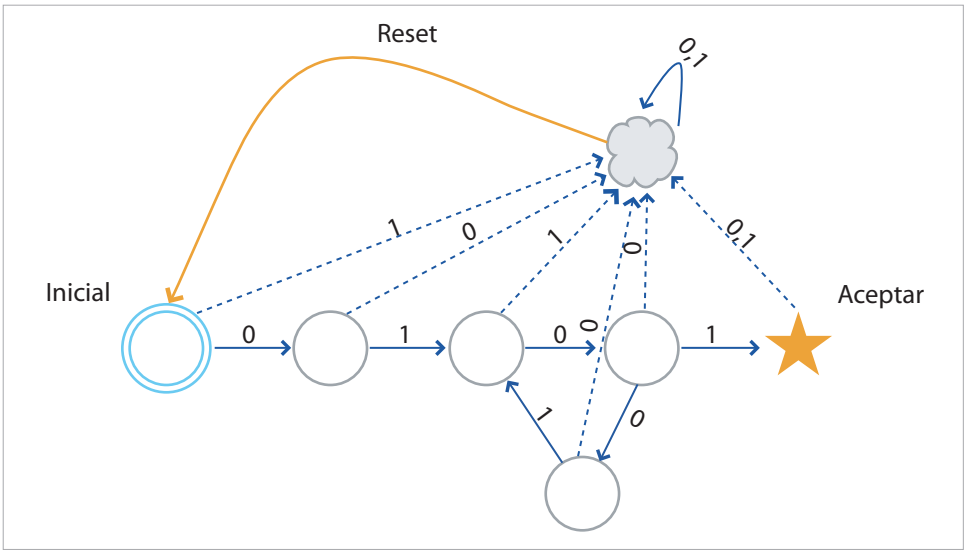


Figura 1. Ejemplo de un autómata de siete estados.

En este autómata hay tres entradas posibles: 0, 1 y reset, y de cualquier estado, al darle reset, se llega al estado inicial (aunque no se pintaron todas las líneas de reset para no abarrotar la figura). Se puede pensar en que al llegar al estado aceptar, se prende la luz verde.

¿Qué hace este autómata? Acepta la cadena 0101, la 0100101 y, en general, la cadena que empieza con 010, seguida de 010 tantas veces como se quiera (quizá cero veces), y terminar con 1. Es decir, acepta todas las cadenas de la forma 010(010)*1. Que haya otros autómatas que aceptan el mismo conjunto de cadenas es una indicación de lo que ya se suponía. Hay muchas maneras de implementar la misma caja negra, así que para un observador externo es imposible distinguir una de otra.

Cajas negras y limitaciones de la ciencia

Ahora bien, ¿hay manera de que Arcadio y Úrsula descubran cuál es el autómata que representa el comportamiento de la caja negra? Esto es, como observadores externos, pueden probar más y más cadenas de entradas y observar las salidas producidas, pero nunca pueden probar *todas* las cadenas de entrada, porque hay un número infinito de éstas. Saben que la caja tiene un número finito de estados, pero no saben cuántos tiene. Si descubrieran un letrerito pequeño debajo de la caja que dijera "Caja negra modelo X42901 de 5 estados" podrían saber con toda exactitud cuál es el conjunto de cadenas que acepta (es demasiado difícil explicar cómo hacer esto en un texto introductorio a la computación). El hecho es que sin esta información es imposible determinarlo: por más y más entradas que le den a la caja negra, siempre puede haber una para la cual se comporte de manera inesperada. Recuérdese el fenómeno del Cisne Negro descrito en el tema 2. No porque sólo se conozcan cisnes blancos quiere decir que todos los cisnes son blancos. Esto nos habla de una limitante inherente a la ciencia en general. Que una ley física se cumpla no quiere decir que sea "verdadera", sólo que se observa su cumplimiento en todos los experimentos realizados.

Durante cientos de años se pensó que las leyes de Newton eran verdaderas hasta que llegó Einstein y demostró que no, al menos no del todo. Que se haya observado durante siglos que se cumplían significa que no estaban tan mal. En efecto, casi siempre se cumplen. En algunas situaciones de extrema velocidad no. Así que en realidad las teorías de Einstein depuraron las leyes de Newton. Así funciona la ciencia, como una sucesión de aproximaciones que describen cada vez mejor los fenómenos de la naturaleza. Al igual que las cajas negras, mientras más entradas se prueben, mejor se les podrá describir.

Los autómatas finitos no lo pueden todo

En cualquier procesador de palabras o en cualquier programa editor se tiene, por ejemplo, la función de búsqueda. En casi todos ellos existe además la opción de buscar todas las cadenas de texto que satisfagan ciertas características, por ejemplo, todas la cadenas que terminan en "ando" o las que empiezan con una "a", luego tienen alguna otra letra y después otra "a". Generalmente, para denotar estos conjuntos de cadenas se utiliza lo que se denominan expresiones regulares: el primer ejemplo sería *ando, el segundo a?a*. Para realizar este **tipo de búsquedas** se usan autómatas finitos.

Cada expresión regular define por sí misma un conjunto de cadenas de símbolos, llamadas *palabras*. En el primer ejemplo están consideradas las cadenas "silbando" y "caminando", pero no "ayer" o "silbar"; en el segundo ejemplo están "ana", "alas", "anazasi" y "amante", pero no "amor" o "banana".

Un resultado importante, y quizás sorprendente, de la teoría de la computación es que los autómatas finitos definen conjuntos de palabras que provienen de expresiones regulares y nada más esto. Son muy útiles, como se puede comprobar cuando se usa un procesador de texto, pero ciertamente su utilidad es limitada. Por ejemplo, no existe ningún autómata finito que reconozca los palíndromos marcados; es decir, todas las cadenas de la forma zmz', donde z' es la misma cadena que z, pero escrita al revés, y m es una letra especial que marca el centro del palíndromo. Hay muchas cosas, inclusive sencillas, que no son computables por medio de un autómata finito.

Cajas negras en la vida real

Resultan insospechadamente usuales las aplicaciones de los autómatas finitos: en estacionamientos y edificios públicos es cada vez más frecuente encontrar puertas automáticas, cuyo funcionamiento está regulado por un autómata finito. El siguiente diagrama ilustra la operación de la pluma de un estacionamiento automático.

La pluma puede estar en uno de dos estados posibles: abierto o cerrado. Si no se está procesando ningún boleto de estacionamiento —es decir, no hay nadie en la pluma—, entonces permanece en ese estado; lo mismo ocurre si el boleto que se introdujo no ha sido pagado. Si el boleto ha sido pagado entonces se pasa al estado abierto, donde se permanece mientras no haya pasado el cliente con su automóvil, es decir, mientras esté en tránsito. Una vez que el cliente ha pasado, se regresa al estado cerrado.

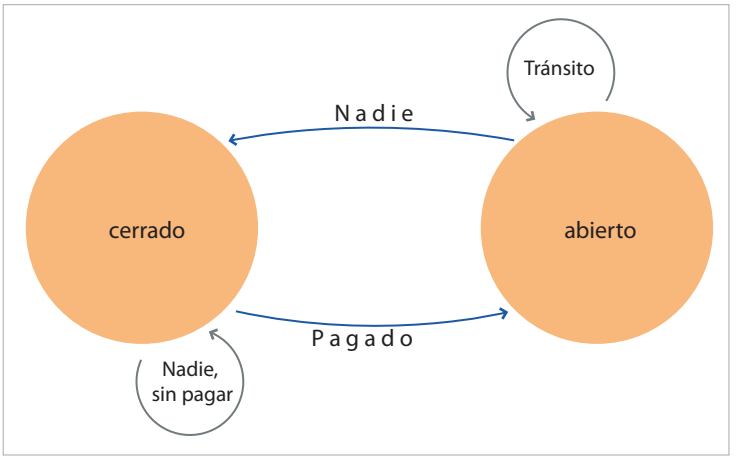


Figura 2. Autómata que modela la operación de una pluma en un estacionamiento automático.

Otro caso popular es el de las máquinas expendedoras de golosinas. Una vez que se ha depositado el dinero entra en operación una máquina de estados finitos. Cada estado está asociado a una de las posibles golosinas disponibles, se gira el resorte que empuja la seleccionada hasta que cae y la que sigue ocupa su lugar. Luego se pasa al estado (mejor dicho, conjunto de estados) en que se entrega el cambio.

Máquinas de estados y lenguajes

Como ya se mencionó en el tema de Algoritmos, una computadora entiende un lenguaje en el que se escriben los algoritmos que es capaz de ejecutar. Ese lenguaje es una manera de describir un modelo de la máquina que representa esa abstracción, se puede decir que *lo esencial* de la máquina es que puede hacer sumas y multiplicaciones, por ejemplo. Los algoritmos que debe ser capaz de ejecutar están dados en términos de esas operaciones y no otras. De hecho, el lenguaje de la computadora abstracta define lo que ésta puede hacer. Dado el lenguaje de la máquina habrá cosas que no puedan expresarse como algoritmo, porque no podrían escribirse en términos del lenguaje propio de la máquina. Máquinas de estados y lenguajes son dos caras de una misma moneda. A continuación se presentará el ejemplo de un modelo de cómputo descrito a partir de un lenguaje.

5.2.2 La geometría plana, un modelo de cómputo restringido

Una de las primeras ramas de la ciencia que cultivó la humanidad es la geometría. Euclides, en el siglo II antes de nuestra era, formuló los principios fundamentales de lo que hoy se denomina geometría sintética. En este tipo de geometría, las únicas operaciones válidas son aquellas que pueden llevarse a cabo usando solamente una regla (sin graduar, por lo que no se puede "medir" con ella y solamente se pueden trazar líneas rectas) y un compás, y existen algoritmos con estas operaciones para obtener innumerables construcciones geométricas. Aunque no todas.

Para construir un triángulo equilátero, por ejemplo, se tiene el siguiente algoritmo:

1] Se traza un segmento de recta de longitud arbitraria.
2] Se apoya el compás en un extremo del segmento y se abre hasta el otro. Se traza el arco de circunferencia usando ese punto de apoyo y esa apertura.
3] Se apoya el compás en el otro extremo del segmento y se traza otro arco de circunferencia sin cambiar la apertura.
4] Se elige como vértice del triángulo uno de los puntos en donde se cruzan los arcos y se trazan los segmentos que van desde él hasta los extremos del segmento.

Curiosidades
La teoría de Galois ha logrado responder a la pregunta de qué problemas se pueden resolver utilizando sólo regla y compás. Pero no se sabe mucho acerca de la complejidad de éstos. Es decir, acerca de cuántos pasos son necesarios para resolver un problema dado.

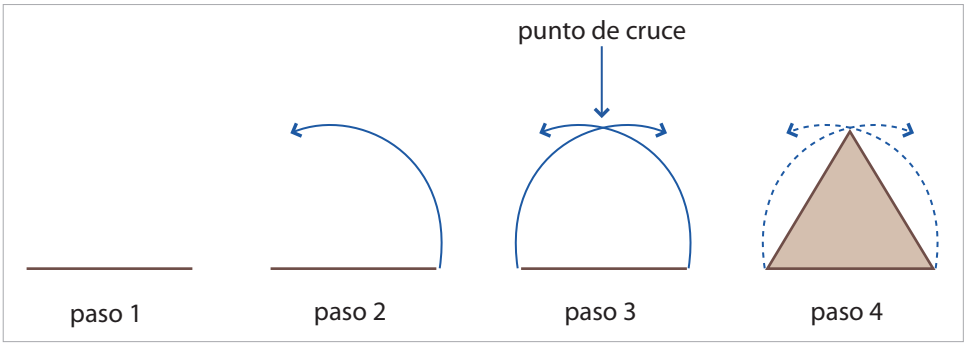


Figura 3. Construcción de un triángulo equilátero.

Euclides (300 a.C.). Su tratado de geometría Los elementos influyó el desarrollo de las matemáticas por más de 2 000 años | © Latin Stock México.

También hay algoritmos que permiten trazar un pentágono, un hexágono o dividir un ángulo en dos partes iguales, por ejemplo. Sin embargo, los geómetras griegos buscaron sin éxito un algoritmo para llevar a cabo las siguientes construcciones:

- La trisección del ángulo: dividir un ángulo en tres partes iguales.
- La duplicación del cubo: dado un cubo, construir otro cuyo volumen sea el doble del primero.
- La cuadratura del círculo: dado un círculo, construir un cuadrado de la misma área.

Todas éstas, por supuesto, han de hacerse con sólo regla y compás. Es aquí donde reside el problema. Pasaron siglos antes de que los matemáticos lograran demostrar, en el siglo XIX, que no existe algoritmo que resuelva estos problemas usando sólo regla y compás. Es decir, resultan *no computables* en la máquina abstracta de la geometría sintética, cuyo lenguaje está hecho sólo de operaciones con esos dos instrumentos.

Hoy en día, en cambio, se suele estudiar la geometría analítica creada por Descartes en el siglo XVII; en ella se introduce el concepto de *sistema coordenado* y se formulan expresiones algebraicas que corresponden con las construcciones geométricas. Usando las herramientas, o el lenguaje de la geometría analítica, los tres problemas mencionados sí tienen solución y, de hecho, no es complicada.

5.2.3 Modelos de computadoras

Lo que se conoce como computadora es: la caja que contiene el procesador junto al monitor, el teclado y el ratón. A final de cuentas, no es más que una caja negra. ¿Tiene ésta el mismo poder de cómputo que un autómata finito? Ya se discutió anteriormente que un autómata finito sólo puede computar expresiones regulares; cosas tan sencillas como un palíndromo están más allá de su alcance. Asimismo, se argumentó que cualquier caja negra no es más que un autómata finito. ¿Cuál es la solución a esta paradoja? El cerebro mismo, ¿no es una caja negra con la que interactuamos mediante entradas y salidas y, a final de cuentas, de tamaño finito? Una computadora no es más que la realización concreta de una abstracción. Pero, ¿de cuál?

La máquina de Turing

En 1936 el matemático inglés Alan Turing formuló el concepto de la máquina que lleva su nombre. La máquina de Turing es una abstracción fundamental en computación. Con ella se logró capturar lo verdaderamente esencial del concepto de algoritmo. Antes de Turing el concepto era vago, propenso a la ambigüedad; algunos, como el matemático David Hilbert, lo describían como "un procedimiento efectivo". Kurt Gödel se aproximó mejor, describiéndolo como una secuencia finita de deducciones en el terreno de la lógica. Pero la abstracción de Turing logra, por una parte, captar lo mínimo indispensable sin ambigüedades de interpretación y, por otra, formular el concepto de una forma general y simple. Existen otros modelos de lo que significa algoritmo, sorprendentemente varios de los más fundamentales fueron establecidos de manera simultánea, en 1936, pero todos equivalentes al de Turing que es un tipo de máquina de estados.

Por supuesto, la máquina de estados propuesta por Turing debe modelar lo que hoy se conoce como una computadora, y por lo tanto poder computar más cosas que expresiones regulares, así que no puede tener un número finito de estados. Pero ya se dijo varias veces que cualquier caja negra sólo puede tener un número finito de estados. Las computadoras modernas se distinguen por dos cosas:

- Tienen un número de estados inimaginablemente grande: muchísimos millones de millones.
- Se les pueden agregar estados, añadiéndoles memoria.

Alan Turing, en uno de los artículos más importantes del siglo xx, pudo intuir un modelo de la esencia de una computadora antes que ninguna se hubiera construido, a partir de un ejercicio mental, en lo que se considera una de las exposiciones más brillantes de un argumento informal e intuitivo que lleva a una noción precisa y formal. Turing pide imaginar a una persona haciendo matemáticas, sentada en un escritorio. Tiene lápiz y papel donde escribir símbolos. Un matemático va escribiendo símbolos, uno tras otro, quizá borrando o cambiando alguno de vez en cuando. Cuando se le acaba la hoja, toma otra. En principio, cuántas hojas toma, cuántos símbolos escribe, qué tan largo es el cálculo que está haciendo, no están limitados más que por su propio tiempo de vida. Quizá ni eso, ya que otros matemáticos pueden continuar con su trabajo. Eso sí, trabaja con un número de símbolos finito. Cualquier descubrimiento de matemáticas posible se podrá hacer de esta forma, cualquier cómputo, cualquier procesamiento de información. Es así como Turing deriva su modelo de máquina de estados, muy similar a un autómata finito, con la única diferencia de que debe tener "hojas de papel" disponibles, ilimitadamente.

Con más detalle, una máquina de Turing puede visualizarse como una caja negra, que es un autómata finito, llamada unidad de control, que supuestamente representa la mente del matemático. Ésta posee una cabeza de lectura y escritura que recorre una cinta sin fin, representando las hojas de papel. La cinta se puede pensar dividida en celdas de tamaño fijo, en cada una está escrito un símbolo de un conjunto finito, llamado alfabeto. La cabeza recorre la cinta en cualquiera de sus dos direcciones y puede leer de cada celda o escribir en ella un símbolo (del alfabeto de salida). La máquina de Turing inicia con la cabeza colocada en un sitio particular de la cinta llamado *celda inicial* y la unidad de control en un estado inicial particular; la unidad de control, de acuerdo con las transiciones de su autómata finito, indica, en función del símbolo de entrada sobre el que está la cabeza, cuál debe ser el estado siguiente de la máquina, cuál debe ser el símbolo que debe escribirse en la celda actual y la dirección en la que debe moverse la cabeza. En algún momento, si los datos de entrada fueron "correctos", es decir, concuerdan con lo que la máquina esperaba encontrar en la cinta a partir de la celda inicial, ésta termina llegando a un estado final, y en la cinta se encuentra el resultado de la ejecución del algoritmo. Si los datos de entrada en la cinta no coinciden con lo esperado por la máquina, entonces ésta puede llegar a un estado de error; se puede pensar que llega a un estado en el que se detiene y en un panel lateral despliega el mensaje "la entrada no es aceptada". Una tercera posibilidad es que la máquina no termine, que la entrada haga que la máquina entre en un ciclo infinito en el cual se repiten indefinidamente algunos de los estados, sin llegar nunca al estado final.

La idea es que una máquina de Turing es un modelo de una computadora, tal y como las usadas hoy en día. Este modelo logra captar todos los elementos mencionados en el módulo dedicado a los algoritmos, a saber:

- Se poseen datos de entrada. El contenido original de la cinta.
- Se cuenta con un conjunto de estados por los que transita la máquina abstracta durante la ejecución del algoritmo. Un estado está determinado por el contenido de la cinta y el estado del control finito.
- Se producen datos de salida. El contenido de la cinta luego de la ejecución del algoritmo.
- Existe un conjunto finito de instrucciones que le especifican a la máquina, en todo momento, cuáles deben ser las acciones a realizar dado el estado actual y el dato de entrada que recibe, codificadas en el control finito de la máquina.
- El *problema* que resuelve la máquina es representado por una función *f*, tal que si la entrada a la máquina es *x*, cuando la máquina se detiene, en la cinta se encuentra $f(x)$.

En el caso de la máquina de Turing, las instrucciones del último punto sólo dicen "si estando en el estado *a*, en la celda de la cinta ves el símbolo *x*, pasar al estado *e*, escribe el símbolo *y*, y mueve la cabeza en dirección *d*". Se tiene un conjunto finito muy bien especificado: *a* y *e* están en el conjunto de estados, al que llamaremos Q y que contiene al estado inicial q_0 y al estado final q_F *x y* y en el conjunto de símbolos posibles o alfabeto, al que se denominará A; *d* es una de dos opciones, los elementos del conjunto D = {Izq, Der, Alto}; así que cualquier ambigüedad queda eliminada. Se podría decir que el lenguaje en el que se especifican los algoritmos a una máquina de Turing está hecho de quintetas de la forma (*a*, *x*, *e*, *y*, *d*).

Llama la atención que la cinta de la máquina sea infinita. La intención de esto es que el poder de cómputo no se vea limitado *a priori*, como se explicó con el ejemplo del matemático y sus hojas de papel. Lo importante es que, por supuesto, en cualquier momento durante la ejecución de una máquina de Turing sólo una sección finita de la cinta se utiliza.

Un ejemplo de máquina de Turing

Considérese un ejemplo, aprovechando lo visto en el tema de representación de datos: se codificarán numéricamente los elementos de los conjuntos Q, A y D y se codificarán también la entrada y la salida escritas en la cinta. Se hará una máquina que incremente en uno un número dado como entrada, este número será dado como una secuencia (posiblemente vacía) de unos consecutivos en la cinta de entrada (comenzando en la celda inicial). Así, para dar como entrada el número 5, se debe proporcionar una secuencia de cinco 1 consecutivos en la cinta. Se asumirá que las celdas desocupadas de la cinta contienen un cero. La máquina comienza en la celda inicial al comienzo de la secuencia de entrada. En ese estado se permanecerá recorriendo la cinta (es decir, cada vez que se encuentre un 0 se reemplaza por el mismo 0) hasta que se encuentre el primer 1 de la secuencia de entrada, se cambia entonces al estado 2 recorriendo la entrada completa (ahora se hace lo complementario, sobrescribiendo los 1 que se encuentran) hasta el primer 0 a la derecha de ésta; en ese momento se le reemplaza por un 1 y se pasa al estado que indica que se ha terminado la ejecu-

Estado actual	Símbolo de entrada					
	0			1		
	Estado siguiente	Salida	Mov.	Estado siguiente	Salida	Mov.
q_0	q_0	0	Der	q_1	1	Der
q_1	q_F	1	Der	q_1	1	Der
q_F	q_F	0	Alto	q_F	1	Alto

ción (el estado q_F). La cabeza termina en la primera celda a la derecha de la salida. En la siguiente tabla se muestra la especificación completa de esta máquina de Turing.

Limitaciones de la máquina de Turing

Una vez que se tiene un modelo de una computadora, cabe preguntarse cuál es su poder, qué cosas puede computar y cuáles no. Turing pudo así probar que existen problemas no computables, que no se pueden resolver mediante una máquina de Turing. Una manera relativamente simple de probar esto es notando lo siguiente:

Hay más problemas que máquinas de Turing

¿Cuántas posibles máquinas de Turing existen? A fin de cuentas, una máquina de Turing cualquiera es una lista de quintetas (a, x, e, y, d), donde cada una de las variables está tomada de un conjunto finito. Si de momento se considera que las quintetas no forman parte de una lista y se eliminan las comas que separan los elementos de cada una, queda una secuencia de números, aquellos con los que se codifican los elementos. Todo esto no es, de hecho, más que un número, posiblemente muy grande, pero un número a fin de cuentas. En el ejemplo, leyendo los números de la tabla de arriba abajo y de izquierda a derecha se tiene: 110121123112113302312. Podría usarse esto como el nombre de la máquina y de hecho identificar a cualquier máquina por el número entero que se obtiene de leer su tabla completa. Puede pensarse ahora que cualquier número entero no negativo puede verse, en efecto, como la tabla de una máquina de Turing, a lo mejor una muy inútil, por ejemplo una que, sin importar la entrada, siempre escriba un 1 y termine. Habrá otra que recorra tres lugares hacia la derecha y termine sin hacer nada, y también varias iguales que sólo avancen cinco lugares a la derecha, tres a la izquierda y escriban un 8, en tanto que otra calcule el trigésimo séptimo dígito de π. En fin, habrá muchas máquinas de Turing, válidas sin duda, aunque algunas sin sentido. Pero lo importante aquí es que puede concluirse que hay tantas máquinas de Turing como números naturales, es decir, el número de máquinas de Turing es infinito y numerable (al igual que los números naturales, pueden recorrerse de uno en uno). Mientras que no es difícil mostrar que el número de problemas es infinito no numerable —esto es, hay tantos como números reales existen—, un número estrictamente mayor que el de números naturales.

Otros modelos de computadoras

Las máquinas de Turing resultan de poca utilidad cuando se trata de determinar la complejidad de un algoritmo, o para planear la construcción de una computadora. El modelo es demasiado simple, pero de muy alto nivel de abstracción. Por ejemplo, no representa explícitamente el programa que se está ejecutando, su lista de instrucciones.

Esto da lugar a otra abstracción, como la denominada máquina de acceso aleatorio (RAM, por sus siglas en inglés), del estilo de la arquitectura de von Neumann, que son más cercanos a la forma en la que está construida una computadora moderna pero con exactamente el mismo poder de cómputo que una máquina de Turing. En el capítulo de computadoras se verá esto con más detalle. Ésta es, esencialmente, una abstracción de la máquina real en la que el algoritmo, luego de ser traducido a un programa, será ejecutado. En el modelo RAM se supone que las instrucciones o pasos que constituyen el algoritmo son ejecutados secuencialmente, uno tras otro; ningún paso se ejecuta al mismo tiempo o con un traslape temporal con otro. Los recursos consumidos durante la ejecución son contabilizados, eliminando detalles de manera uniforme, así que, en general, cada paso del algoritmo tarda lo mismo cada vez que se ejecuta.

Todos los algoritmos que se han tratado en este texto han sido inicialmente mostrados en una especie de lenguaje adecuado para que el procedimiento pueda comprenderse sin malos entendidos, a los que se les suele llamar seudocódigos. No han sido escritos en un lenguaje de programación específico. Éstos se pueden representar mediante una máquina de estados, como el modelo RAM. Sin entrar en detalles, este modelo define una memoria donde están almacenadas una tras otra las instrucciones del programa, y un lugar especial de la memoria con un número que indica cuál es la siguiente instrucción a ejecutar.

5.2.4 Tesis de Church-Turing

El modelo de Turing es el más general que se tiene para decidir si algo es computable o no. Los computólogos trabajan todos los días suponiendo el modelo de cómputo. De hecho hasta se ha bautizado esta hipótesis como la tesis de Church-Turing, que se puede enunciar como: todo problema computable lo es porque hay una máquina de Turing que lo resuelve. Hay muchos modelos alternativos de lo que es computable, diferentes del propuesto por Turing, pero son equivalentes a éste, o más débiles.

En la sección anterior se concluyó que el número posible de máquinas de Turing es infinito, pero numerable: aunque uno nunca termina de contarlas, se pueden recorrer una a una contándolas. En cambio, el número de posibles funciones es también infinito, pero de un tamaño diferente: no numerable. Excede el nivel de este libro una discusión mayor al respecto, pero para dar una idea aproximada de lo que se habló, las funciones computables son sólo minúsculas salpicaduras dispersas en el inmenso océano de todas las funciones.

Existen muchos problemas (funciones) que no son computables y que resultan de gran importancia práctica. Por ejemplo, sería muy provechoso tener un programa que revise si otros programas son correctos. Se puede demostrar que no existe tal programa. Así que se dedica muchísimo esfuerzo a tratar de que los programas y sistemas de cómputo que se construyen no tengan errores, pero sólo es posible aproximarse a esta meta.

La tesis de Church-Turing habla de cómputo en un sentido muy preciso: dada una entrada, el modelo computa una salida. En realidad esto no abarca todas las posibilidades de lo que puede ser el cómputo. En el mundo de internet y la web, hay sistemas de cómputo que reciben entradas continuamente, una tras otra, en distintos lugares del sistema, y deben computar salidas continuamente, a pesar de fallas e información incompleta. Todo esto da lugar a una enorme variedad de modelos y dificultades de otro tipo, que el computólogo estudia. Algo de esto se verá en el tema de redes.

5.3 LÓGICA

5.3.1 El sueño de Leibniz

¿Cómo piensa el ser humano?, ¿cuál es el misterioso mecanismo que permite deducir algo a partir de lo que se sabe? De hecho, ¿es un mecanismo? Estas preguntas desembocaron en el surgimiento de la ciencia de la computación en el siglo XX. El pensamiento es el medio por el que se formulan las abstracciones que permiten explicar los fenómenos observados en el mundo. Así, las preguntas anteriores surgen de la inquietud por saber si

los posibles "mecanismos" del pensamiento son adecuados para explicar los "mecanismos" que determinan el funcionamiento del universo.

El modelo de las mónadas de Leibniz fue un intento por establecer la definición última de los misteriosos hilos que la naturaleza o un dios, como el filósofo prefiera, utiliza para gobernarlo todo. Vivimos en el mejor de los mundos posibles, tal es la conclusión que Leibniz desprende del hecho de suponer un dios perfecto como arquitecto del mundo. De ello deduce la existencia de entidades precisas, abstractas, inmateriales, que lo regulan todo y cuya secreta interacción debe obedecer reglas, como las del álgebra. Explicar el mundo consiste entonces en encontrar el lenguaje de las mónadas y los elementos constitutivos de él, su alfabeto.

Leibniz recibió inspiración de Ramón Lull, catalán del siglo XIII, quien había supuesto que la verdad es única y dictada por la omnisciencia de Dios y que era una pena que, para que los seres humanos pudieran descubrirla, las premisas de las que se podía deducir tuvieran que pasar por el intelecto. Esto sometía el proceso a la falible habilidad humana para razonar correctamente y, peor aún, ocasionaba disputas violentas y guerras. A partir de esto, Lull decidió crear una especie de rudimentaria máquina que fuera capaz de automatizar el proceso de deducción de la verdad: el Ars Magna.

Leibniz, 500 años más tarde, prefirió aventurar el lenguaje de las mónadas. El objetivo es descubrir las reglas que permiten, literalmente, calcular verdades mediante la manipulación simbólica y garantizar que sólo verdades pueden desprenderse como consecuencia de la aplicación de las reglas. El lenguaje debe ser perfecto. Es en esta corriente de pensamiento en la que la lógica matemática actual tiene su germen. Establecer las reglas mediante las que se debe razonar para deducir nuevas verdades a partir de premisas previamente establecidas como verdaderas. Como se verá posteriormente, existen diversos tipos de lógica, tanto más poderosas cuanto más complejas; las más simples no proporcionan todos los elementos necesarios en el lenguaje para poder hacer deducciones sofisticadas, resultan ser una abstracción excesiva. Las más generales, con mayores elementos, son también más poderosas. Todas ellas son útiles en computación; cada una constituye la abstracción fundamental en su propio ámbito.

Leibniz | © *Latin Stock México.*

Gottfried Wilhelm von Leibniz
(1646-1716)
Anticipó nociones que aparecieron mucho más tarde en biología, medicina, geología, teoría de la probabilidad, psicología, ingeniería y ciencias de la información. Descubrió el cálculo infinitesimal, independientemente de Newton, y su notación es la que se halla desde entonces en uso general. También inventó el sistema binario, en que se basan casi todas las arquitecturas de computación actuales. Se le adjudica haber utilizado por primera vez la palabra función, que proviene del latín *functo*, que significa "el acto de realizar". Dominaba el latín, el griego, el francés, el inglés y el alemán, e incluso llegó a interesarse por la escritura china y el I Ching.

5.3.2 Un problema fundamental

En el tema de Información se habló de las preguntas más simples que permiten buscar un dato. Preguntas *binarias*, cuya respuesta sólo puede ser sí o no. Éste es el problema más simple que puede formularse: saber si algo ocurre o no, si es o no es, si es falso o verdadero. Y se explicó ahí que da lugar a la noción de *bit*. Un problema está definido por un conjunto de entradas, y para cada una, las salidas que deben ser producidas por un algoritmo que lo resuelve. A los problemas que tienen sólo dos salidas posibles se les llama *problemas de decisión*.

El matemático David Hilbert propuso el *Entscheidungsproblem* (problema de decidibilidad en alemán) en 1928. Éste consiste esencialmente en diseñar un algoritmo para resolver un problema de decisión en un sistema formal. De especial interés es el sistema formal de la aritmética. Es decir, diseñar un algoritmo que tome como entrada un enunciado matemático acerca de los números enteros, y conteste sí o no, según si el enunciado es verdadero o falso. Por ejemplo, la Conjetura de Goldbach dice "cualquier número entero par mayor a 2 puede ser escrito como la suma de dos números primos": $4 = 2 + 2$, $6 = 3 + 3$, $8 = 3 + 5$, etc. Con la ayuda de computadoras se ha logrado verificar que el enunciado es cierto para números enormes, hasta de 17 cifras, pero aún no se ha podido res-

ponder si este enunciado es cierto o no en general. La pregunta de Hilbert se remonta a Leibniz, quien en el siglo XVII soñaba con diseñar tal algoritmo, al construir exitosamente una máquina mecánica de cálculo, que pudiera manipular símbolos para determinar si una frase en matemáticas es un teorema. Tanto los resultados de Turing que se han mencionado como simultáneamente los de Alonzo Church en 1936, de manera independiente, demostraron que no existe tal algoritmo.

La geometría plana que se vio en la sección anterior es otro ejemplo de un sistema formal, y ahí un enunciado podría ser, por ejemplo, que dados dos puntos por ellos pasa una y sólo una recta. En general, se puede definir un sistema formal mediante un conjunto de postulados que se asumen ciertos, llamados *axiomas*, y a partir de ellos se deducen nuevos enunciados verdaderos en el sistema: se les suele llamar teoremas; los axiomas junto con todos los teoremas que se pueden deducir a partir de ellos constituyen lo que se denomina un *sistema formal*. En términos generales el problema de decidibilidad consiste en lo siguiente: dados un conjunto de axiomas y una proposición, diseñar un algoritmo que responda si la proposición es verdadera o no en el sistema formal. O sea decidir si la proposición es o no un teorema del sistema formal.

Las máquinas de Turing son justamente una formalización de la noción de algoritmo, y el Teorema de Church-Turing (no confundir con la Tesis de Church-Turing) dice que existen proposiciones de aritmética que ninguna máquina de Turing puede decidir. No se puede saber si son verdaderas o no, mecánicamente. Existen sistemas en los que, en efecto, toda proposición es decidible, pero son más simples que la aritmética. Church y Turing demostraron que el *Entscheidungsproblem* en general no tiene solución.

5.3.3 La limitación inherente de las matemáticas

A fines del siglo XIX y principios del XX había un gran interés por la lógica. Allí surgió el problema de la decidibilidad que se ha mencionado y, junto con éste, algunos otros. Este interés no era fortuito, la intención última era establecer una estructura sólida, perfecta para las matemáticas, sobre la lógica, la herramienta para razonar de manera impecable que permite deducir conocimiento nuevo a partir de lo ya establecido. De allí que cobrara auge también el método axiomático en las matemáticas: si cualquier área de las matemáticas podía ser construida a partir de una colección finita de postulados elementales, usando las reglas de la lógica para derivar todo lo demás, entonces todo era un sistema formal y la esperanza era que esto deviniera en la *perfección*.

La perfección habla de sistemas formales cuyos axiomas se pueden enumerar mediante una máquina de Turing, y se refiere a que:

- El sistema es completo. Lo que significa que cualquier proposición verdadera en el sistema puede deducirse a partir de los axiomas.
- El sistema es consistente. Lo que significa que no existe ninguna proposición que pueda ser demostrada como verdadera en el sistema y que ocurra lo mismo para su negación.

Los últimos dos incisos son incompatibles para sistemas formales suficientemente complejos, según lo demostró el lógico Kurt Gödel en 1931. Ningún sistema formal que contenga, al menos, los axiomas de la aritmética, puede ser al mismo tiempo completo y consistente. Es decir, para la aritmética misma, que es consistente, hay enunciados que a pesar de ser verdad no pueden demostrarse en el sistema a partir de los axiomas y usando las reglas de la lógica.

5.3.4 Álgebra booleana

Al poseer un sistema formal basado en axiomas, la intención era que pudiera automatizarse la deducción de cualquier verdad matemática, que una máquina dedujera las cosas sin necesidad de que un ser humano, propenso al error, interviniera. La verdad debía poder deducirse por pura manipulación simbólica llevada a cabo con las reglas de la lógica.

Uno de los lógicos matemáticos que contribuyeron relevantemente en ese sentido fue George Boole —él se dio a la tarea de formular las reglas de manipulación que permitían deducir la verdad o falsedad de expresiones construidas con símbolos que representan enunciados, sin importar lo que los símbolos representaran, se pudieran manipular algebraicamente las expresiones construidas con ellos para deducir nuevas expresiones válidas—. El resultado más útil para la computación fue lo que se conoce como el álgebra de las funciones de conmutación, un caso simple de álgebra de Boole que es, de hecho, un sistema formal. En el álgebra de las funciones de conmutación se poseen al menos dos elementos (típicamente, verdadero y falso, o 1 y 0) y tres operaciones elementales: una operación unaria (que posee un solo operando) llamada negación: si A denota una variable, entonces $\tilde{A}$ denota la negación de A, si A es verdadero o 1, entonces $\tilde{A}$ es falso o 0, respectivamente. Las otras dos operaciones se llaman disyunción y conjunción, y son las que ya se conocen de la lógica simbólica. En la tabla que sigue se muestran las tablas de verdad de estas operaciones en la lógica proposicional.

Cualquier otra función de conmutación puede formularse con base en las tres funciones elementales mencionadas, es decir, la tabla de verdad de cualquier función cuyas variables sean booleanas (falso o verdadero) puede obtenerse combinando estas tres operaciones. Esto resulta sumamente útil, como se verá en el capítulo dedicado a sistemas de cómputo, pero no es todo lo que se necesita.

A	B	A ∧ B	A ∨ B
F	F	F	F
F	V	F	V
V	F	F	V
V	V	V	V

5.3.5 Lógica de primer orden

En efecto, en la lógica proposicional o equivalentemente, en el álgebra de las funciones de conmutación, no se pueden hacer deducciones muy complejas. Por ejemplo, las siguientes afirmaciones:

- Las plantas verdes tienen clorofila.
- El cilantro es una planta verde.
- Entonces, el cilantro tiene clorofila.

En la lógica proposicional quedarían representadas por:

- P
- Q
- P => Q

Lo que no sirve de mucho porque no dice nada acerca de la validez de la última expresión a partir de las anteriores. La lógica proposicional considera independientes las afirmaciones anteriores, no puede expresar la tercera como conclusión de las anteriores, se necesita mayor complejidad en la abstracción para expresar propiedades que son satisfechas por ciertos objetos y además para decir cuántos de esos objetos las satisfacen.

En la lógica de primer orden, si se usa el símbolo $V(x)$ para decir que x es una planta verde y $C(x)$ para decir que x tiene clorofila, entonces se tiene:

- Para toda x, si $V(x)$ entonces $C(x)$.
- V (cilantro).
- Por lo tanto, C (cilantro).

Ahora puede verse la tercera afirmación como una deducción derivada de las dos primeras; es posible apreciar mayores detalles de las afirmaciones al considerarlas constituidas elementos menores como la cuantificación de "para toda" y los predicados V y C, que ahora son cosas susceptibles de ser evaluadas como falsas o verdaderas dependiendo de a quién le son aplicadas.

Ésta es una herramienta esencial, no sólo en computación, porque toda la matemática está basada en ella. La lógica de primer orden se utiliza para demostrar cualquier afirmación hecha en matemáticas, es su herramienta deductiva. Es, por cierto, una abstracción muy poderosa; en la última representación del ejemplo ya no importa realmente qué significa V o C o cilantro. Muy bien se podría decir que cilantro es el nombre de un amigo, que $V(x)$ significa que x es un holgazán y que $C(x)$ significa que x reprueba el curso y la deducción seguiría siendo válida. Más aún, podrían ser cosas que ni siquiera tengan sentido: cilantro es el nombre de un habitante del planeta Hoth, $V(x)$ significa que es un "trupiciante" y $C(x)$ significa "algerobio". La primera afirmación dice que todo trupiciante es algerobio, no se sabe qué significa eso pero no importa, el caso es que si luego nos dicen que cilantro es un trupiciante entonces se deduce que es algerobio. La validez de la afirmación proviene de la forma que tiene, de lo que se escribe, no de lo que significa lo que se escribe (si es que significa algo).

Con base en esta herramienta es posible demostrar la corrección de los algoritmos y verificar, en casos particulares, cuándo terminan y cuál es el estado en el que lo hacen.

Verificación de corrección de programas

El siguiente es un algoritmo para calcular la media aritmética (promedio) de una lista de números. Con `lista[i]` se denota el i-ésimo número de la lista.

```
1. Promedio(lista)
2. suma = 0
3. for i∈{1, 2, ..., tamaño(lista)} do
4.   suma = suma + lista[i]
5. endfor
6. regresa suma / tamaño(lista)
7. end.
```

A manera de ejemplo, se analizará este algoritmo. Lo que se pretende calcular es tan simple que intuitivamente se sabe que es correcto; el código del algoritmo expresa exactamente el proceso que se realiza para calcular el promedio manualmente, a saber, calcular:

$$\frac{\sum_{k=0}^{tamaño\ (lista)} lista\ [k]}{tamaño\ (lista)}$$

No hay motivos para pensar que el algoritmo no es correcto. Pero, en general, los algoritmos que suelen utilizarse en muchas situaciones reales son bastante más complejos; asegurarse de que entregan el resultado correcto siempre puede ser muy difícil. Cabría preguntarse, ¿por qué se sabe que el algoritmo es correcto?, ¿cómo asegurarse de que realmente lleva a cabo el cálculo para el que fue pensado?

El algoritmo supuestamente calcula el promedio de los números guardados en una lista. En la línea 2 se asigna un cero a la variable suma, se puede decir entonces que cuando se llega por primera vez a la línea 3, la variable suma contiene la suma de los primeros cero elementos de la lista. La siguiente vez que se llega a la línea 3, la variable `i` tiene un 1 y la variable `suma` contiene lo que tenía (0), más el valor guardado en la posición 1 de la lista; es posible afirmar que la variable `suma` contiene entonces la suma del primer elemento de la lista. En la siguiente llegada a la línea 3, `i` vale 2 y la variable `suma` contiene la suma de los primeros dos elementos de la lista y así sucesivamente. Siempre que se llega a la línea 3, la variable `suma` contiene la suma de los primeros `i` elementos de la lista. En notación de predicados se diría:

a] Si P = {0, ..., *tamaño(lista)*} entonces, para toda i P,

$$\text{Suma} = \sum_{k=0}^{i} \text{lista [k]}$$

luego de la *i*-ésima repetición del ciclo de las líneas 3 a 5.
También se tiene que:

b] Al llegar a la línea 6, *i* = *tamaño(lista)*.
Así que fácilmente se deduce que:

c] Al llegar a la línea 6, `suma` contiene la suma de todos los elementos de la lista. Es decir:

$$suma = \sum_{k=0}^{tamaño\ (lista)} lista\ [k]$$

Finalmente, en la línea 6 se calcula el resultado que es, como se ve en el código, el contenido de la variable `suma` dividido por el tamaño de la lista. De donde se deduce:

d] El resultado final del algoritmo es:

$$\frac{suma}{tamaño\ (lista)} = \frac{\sum_{k=0}^{tamaño\ (lista)} lista\ [k]}{tamaño\ (lista)}$$

Lo que coincide con la definición de promedio, por lo que ahora se puede tener la seguridad de que el algoritmo es correcto.

A una expresión como la del predicado A se le denomina *invariante de ciclo*, es un predicado verdadero siempre que se ejecuta un ciclo (no necesariamente durante la ejecución del mismo). Usando la lógica de primer orden y el invariante de ciclo fue posible demostrar que el algoritmo es correcto. Éste es un procedimiento sólido, y todo buen programador recurre a él con frecuencia, al menos implícitamente.

Razonar para resolver problemas-acertijos

Con la lógica de primer orden como herramienta deductiva es posible resolver problemas que resultan insolubles a simple vista. Considérese el ejemplo de la siguiente historia: luego de una desigual batalla en la que resultó vencedor, Harald, hijo de Hafni, se entregó al sueño, agotado por la lucha. Sleipnir, hechicero enemigo de Harald, aprovechó la condición del héroe para secuestrarlo y así, en medio de su profundo sueño, éste fue llevado por medio de artes mágicas a un amplio salón en el castillo de Sleipnir. Así permaneció Harald, dormido, ignorante de su situación, durante varios días. Cuando despertó se halló encerrado en el enorme salón sin otra compañía que una lechuza y un oso, con la habilidad de hablar el lenguaje humano, y que en realidad eran otras víctimas de Sleipnir. Se oyó de pronto la estentórea voz de Sleipnir retumbando en el salón: "Sólo podrás salir si adivinas qué día de la semana es hoy". El mago confiaba en mantener atrapado al héroe a sabiendas de que éste había perdido la noción del tiempo por haber estado dormido varios días. Acto seguido anunció: "Seré benévolo contigo, Harald, hijo de Hafni. El oso que te acompaña está obligado a decir sólo mentiras los lunes, martes y miércoles; la lechuza no puede sino mentir los jueves, viernes y sábados; el resto de los días ambos sólo dicen verdades. Demuestra que tu mente es tan poderosa como tu espada." El oso le dijo entonces a Harald: "Ayer fue uno de los días en que me toca mentir"; la lechuza declaró por su parte: "¡Qué coincidencia! Ayer también fue uno de los días en que debo mentir". Harald se sentó abatido en el suelo; luego de unos momentos de reflexión, se levantó y gritó: "Sleipnir, permíteme liberar del hechizo a estos desgraciados junto conmigo". Al cabo de unos minutos se oyó la voz del hechicero: "Harald, hijo de Hafni, he de reconocer tanto la valía de tu generoso corazón como tu soberbia, ¿te crees capaz de vencerme? No seré yo quien te reproche tu fracaso, sino los infelices que te acompañan. Te concederé su libertad y la ruptura del hechizo que pesa sobre ellos si logras saber qué día de la semana es hoy. Mi honor queda empeñado en ello." Acto seguido Harald gritó: "Hoy es el día consagrado al poder de Thor,[2] que su martillo te aplaste si no nos liberas en este jueves." Cayeron las puertas del recinto, Harald pudo sentir el viento de las montañas otra vez y en sus futuras aventuras le acompañaron siempre Alder y Elder, dos robustos guerreros que alguna vez fueron un oso y una lechuza.

¿Cómo logró Harald saber el día de la semana?

1] Supóngase que L es la lechuza. Los días en que la lechuza miente son $M(L) = \{ju, vi, sa\}$. Luego, los días en que dice sólo la verdad son $V(L) = \{do, lu, ma, mi\}$.
2] Si O es el oso. Los días en que miente son $M(O) = \{lu, ma, mi\}$. Los días en que sólo dice la verdad son $V(O) = \{ju, vi, sa, do\}$.
3] Se ve claramente que no hay días en que la lechuza y el oso mientan simultáneamente, en notación de conjuntos: $M(L) \cap M(O) = \emptyset$.
4] Así que para cualquier día x, uno de los dos animales necesariamente dice la verdad.
5] Con $A(x)$ se denotará el día anterior a x. Según el oso $A(\text{hoy}) \in M(O)$. Según la lechuza $A(\text{hoy}) \in M(L)$. Pero esto significaría que $A(\text{hoy}) \in M(L) \cap M(O)$. Esto, por lo expresado en el inciso 3, no puede ser; entonces alguno de los dos miente (el inciso 4 impide que ambos mientan).
6] Supóngase que quien miente es el oso. Eso significa que hoy es uno de los días en que le toca mentir a él y a la lechuza no. Como la lechuza no miente entonces lo que dice es verdad, ayer fue uno de sus días de mentir. Eso significaría que $A(\text{hoy}) = sa$ y entonces hoy sería domingo. Pero si fuera domingo el oso tampoco estaría mintiendo (los domingos nadie miente) y eso contradice lo que se está suponiendo.

[2] Esto es, *Thursday* en Inglés, que es el día dedicado a Zeus (en griego), Júpiter o Jove en la mitología latina, es decir, nuestro jueves.

7] Supóngase entonces que es la lechuza quien miente. Entonces hoy es uno de los días en que le toca mentir a la lechuza y al oso no. Éste hoy dice la verdad y ayer le tocó mentir, eso significa que A(hoy) = mi y, por tanto, hoy es jueves, como bien dedujo Harald, lo que confirma que es día que la lechuza miente.

5.3.6 Lógica y conocimiento

Una lógica provee de un lenguaje con la expresividad adecuada para resolver un problema. Los acertijos planteados resultan más o menos simples de resolver usando la lógica de primer orden. Pero para resolver el siguiente, aun sin darse cuenta de ello, se necesita otro tipo de lenguaje.

Tres alpinistas

En una fría madrugada de invierno, en medio de la oscuridad total, tres alpinistas que pernoctan en una cabaña, a los que se llamará *A*, *B* y *C*, se levantan al mismo tiempo y toman sus gorras de lana de una mochila en la que hay cinco de ellas: tres rojas y dos negras. Al salir de la cabaña cada uno de ellos puede ver la gorra de sus compañeros, pero no la propia. Todos saben cuántas gorras hay en total de cada color. *A* le pregunta a *B*, estando presente *C*, si sabe el color de la gorra que el propio *B* tiene, *B* dice que no; de igual forma, *A* le pregunta a *C*, en presencia de *B*, si sabe el color de su gorra, *C* dice ignorar el color de su propia gorra. *A* declara entonces que tiene una gorra roja, ¿cómo lo determinó?

Para resolver el acertijo es necesario poner en claro lo que sabe cada uno de los personajes y lo que saben todos ellos. Es decir, el conocimiento individual y el común. Como sólo hay dos negras y no alcanzan para todos, necesariamente hay al menos una gorra roja en alguno de los alpinistas y eso es algo que todos saben. Se puede expresar esto en términos de lógica como: A tiene gorra roja Ó B tiene gorra roja Ó C tiene gorra roja. Se ha usado el conectivo Ó en el sentido que suele usarse en lógica, para denotar una disyunción de proposiciones, esto es, al menos una de las tres proposiciones es verdadera. En la lista que aparece a continuación y que contiene lo que todos saben, esta disyunción aparece modificada en el primer inciso.

1] *A* o *B* o *C* tiene gorra roja.
2] *A* sabe el color de la gorra de *B*.
3] *A* sabe el color de la gorra de *C*.
4] *B* sabe el color de la gorra de *A*.
5] *B* sabe el color de la gorra de *C*.
6] *C* sabe el color de la gorra de *A*.
7] *C* sabe el color de la gorra de *B*.

Después de la primera pregunta de *A*, el acervo de conocimiento común se incrementa, *C* responde que no sabe el color de su propia gorra, es decir:

8] *C* no sabe el color de la gorra de *C*.

Dados 6 y 7, si *C* viera dos gorras negras en *A* y *B*, sabría que, como son todas las negras que hay, él debe tener una roja. Sin embargo, dado que *C* declara no saber el color de su gorra se deduce que no ve dos negras, es decir, al menos ve una roja. Es decir:

9] *A* tiene gorra roja o *B* tiene gorra roja.

Se integra ahora también al conocimiento común. Luego de la segunda pregunta de *A*, todos saben que:

10] *B* no sabe el color de la gorra de *B*.

Dados 4 y 5, si *B* viera dos negras, por el mismo razonamiento anterior sabría que él posee una roja; dado que declara no saber, significa que al menos ve una gorra roja. Es decir:

11] *A* tiene gorra roja o *C* tiene gorra roja.

Con este conocimiento acumulado, *A* puede deducir que tiene una gorra roja, porque si fuera negra, dada la respuesta de *C* a la primera pregunta, *B* hubiera podido saber, por el inciso 9, que él tenía gorra roja. Pero *B* no supo (10), así que la gorra de *A* no puede ser negra. Por lo tanto tiene que ser roja. ¡Ajá! Además es posible darse cuenta de que nunca se usa el hecho de que *A* puede ver las gorras de sus compañeros. De hecho *A* podría ser invidente y aún así deducir correctamente el color de su gorra, porque la deducción se basa sólo en el conocimiento que se acumula a lo largo del proceso.

Al tipo de lógica que se utiliza para hacer deducciones basadas en conocimiento se le denomina lógica epistémica, y cuando se utiliza específicamente la noción de saber adquiere el sobrenombre de lógica modal. Éste es también un sistema formal basado en axiomas, de hecho los mismos de la lógica de primer orden con algunos adicionales que le dan poder de deducción basado en el conocimiento. Se da por sentado, por ejemplo, que algo que se sabe es verdadero, y que si algo se sabe, entonces también se sabe que eso se sabe y que si no se sabe, entonces se sabe que no se sabe. Parece trabalenguas, pero es lógica.

El acontecimiento que marcó el fin de la segunda guerra mundial, la invasión aliada de Europa, se puede deducir de la lógica modal:

- Los alemanes no saben que la invasión de Europa será por Normandía.
- Los ingleses (mediante una eficaz red de espionaje) saben que los alemanes no saben que la invasión será por Normandía, y además.
- Los alemanes no saben que los ingleses saben que los alemanes no saben que la invasión será por Normandía.

Así se pudo fraguar el engaño del día D. Los ingleses tomaron de la morgue el cadáver de un indigente, le inventaron una identidad falsa vinculada con el Estado mayor, le esposaron un portafolios con documentos falsos que indicaban que la invasión de Europa tendría lugar por el paso de Calais, donde Inglaterra y el continente están más cerca, y dejaron el cuerpo a la deriva cerca de la costa española (Francisco Franco era afín al eje). Después, el servicio de inteligencia británico corroboró que el engaño había tenido éxito y Adolfo Hitler estaba convencido de que, de realizarse una invasión, ésta tendría lugar por el paso de Calais, lugar en el que se reforzó la defensa alemana (el "Muro del Atlántico"), mientras que en Normandía, sin ser despreciable, la defensa era menor.

La lógica modal y su poder de deducción la hacen una de las herramientas más útiles en el análisis de problemas que involucran conocimiento. Tiene por tanto aplicaciones evidentes en diversas áreas de la inteligencia artificial, pero también es útil para formular las abstracciones necesarias para resolver problemas que involucran la participación de diversas entidades computacionales, como se verá en el capítulo de redes, en términos llanos, de diversas computadoras que colaboran para lograr un objetivo común. De hecho,

hace posible minimizar la comunicación entre procesos distantes de tal forma que se intercambie sólo lo indispensable para tener el conocimiento común que les permita lograr su objetivo o bien demostrar que hay problemas que no pueden resolverse a menos que se garantice un mínimo conocimiento común.

Como se pudo percibir en los acertijos anteriores, el objetivo es que el conocimiento nuevo que se incorpora en cada paso, en cada deducción, reduzca el número de posibilidades entre las que se encuentra la solución. Al principio se sabía poco acerca de cuál es realmente el estado de las cosas (los colores de las gorras de todos); existen muchos mundos posibles y en cada uno de ellos los colores de las gorras de *A*, *B* y *C* son diferentes, todas las combinaciones posibles. Luego, conforme se aprende algo nuevo, el número de mundos posibles se reduce paulatinamente hasta llegar a uno solo, o bien, a un conjunto de ellos en los que hay alguna característica común.

5.4 ANÁLISIS DE PROBLEMAS

5.4.1 En el banco

Al cruzar el umbral de entrada del banco, Arcadio se despide con resignación de su deseo por concluir rápidamente su trámite bancario: cobrar el cheque que amablemente le dio su madre para comprar el regalo de Úrsula. Arcadio se forma en la fila marcada para cuentahabientes; para no variar, esta fila es la más larga. Hay otras dos filas, una para clientes preferentes y otra para empresas.

Arcadio calcula con fastidio que han transcurrido 10 minutos y aún le faltan siete lugares para llegar al inicio de la fila. Así, Arcadio se mueve a su pasatiempo favorito, observar. Antes de ser atendido y mientras su fila avanza lentamente, analiza el comportamiento de las filas a su izquierda (clientes preferentes) y derecha (empresas).

La fila de clientes preferentes siempre es muy corta y avanza a mayor velocidad que las otras dos. En particular, avanza tres veces más rápido que la suya. Además de injusto para su noble causa de comprarle un regalo a Úrsula, suena como una mala inversión para el banco contar con el mismo número de cajeros para clientes preferentes que para las otras dos filas juntas. Sobre todo cuando el banco dispone de una caja única para trámites de empresas, donde los mensajeros tardan 20 minutos en ser atendidos en la caja y, suponiendo un promedio de cinco personas en la fila en todo momento, al menos dos horas desde que llegan al banco hasta que salen.

El timbre que anuncia que una caja para cuentahabientes está disponible lo saca de su análisis y Arcadio revisa el tablero digital con el número de la caja donde podrá, 27 minutos después de haber llegado, cambiar por efectivo el cheque que su madre le dio. Exactamente treinta minutos después de entrar al banco, Arcadio se encuentra de nuevo en la parada del autobús y ansioso por llegar a su destino para iniciar la búsqueda del regalo perfecto.

5.4.2 La visión del computólogo

Mientras que para los usuarios comunes de un banco la prioridad es realizar las operaciones eficientemente, para el banco la meta es maximizar sus ganancias. Ambos intereses se mezclan y operan de manera simultánea en las sucursales del banco. Haciendo uso de la abstracción, se puede apreciar el equilibrio entre ambos puntos de vista.

El banco podría considerar preferentes a aquellos clientes que depositan y mantienen importantes sumas de dinero en sus cuentas. Atender con mayor celeridad a estos clientes es prioritario. Esto no significa que el resto de los usuarios no sean importantes para el banco, por el contrario. El grupo más numeroso de clientes representa una parte importante de las utilidades y una fuente vital de prestigio, ya que un cliente satisfecho usualmente promueve el servicio entre sus familiares y conocidos.

El último grupo, los mensajeros de las empresas, también es importante. Sin embargo, las empresas los contratan para realizar decenas o cientos de operaciones bancarias diariamente.

Una solución aceptable para este *análisis* de los requerimientos del banco es, por supuesto, ofrecer distintas filas para los clientes. Este tipo de análisis es común en computación y se deriva directamente de la abstracción y la especificación de problemas.

Al tener distintas filas para los tipos de clientes y números variables de cajas para atender cada fila, el banco está administrando sus recursos humanos y técnicos para:

1] Maximizar su ganancia económica.
2] Ofrecer el servicio más eficiente posible a los distintos tipos de clientes.

Si el banco contrata más cajeros, atiende más rápido a los clientes, pero gasta más en sueldos y en infraestructura requerida para soportar más transacciones concurrentes. Por otro lado, si el banco recorta su personal para ahorrar en sueldos y tecnología, corre el riesgo de perder clientes porque sus servicios en sucursal serán deficientes. Así, el banco intenta encontrar un punto intermedio entre mejorar la calidad del servicio y maximizar su utilidad neta.

Resolver este tipo de situaciones, donde se intenta optimizar varias funciones, posiblemente contrarias entre sí al mismo tiempo, es muy común en computación. La o las abstracciones necesarias en este tipo de problemas deben ser capaces de capturar lo esencial de dos puntos de vista radicalmente diferentes, para proceder a buscar el mecanismo mediante el cual sea posible lograr un equilibro entre ambos.

5.4.3 Abstracción en programación

En la programación se utilizan dos tipos esenciales de abstracción: de control o de datos. La primera es la abstracción de acciones, mientras que la de datos se refiere a *estructuras de datos,* es decir, al modo en que éstos se organizan en la memoria de la computadora, la manera en que se relacionan y los mecanismos que se usarán para acceder a ellos.

Por ejemplo, la abstracción de control en el contexto de programación estructurada es el uso de subprogramas y flujos de control con formatos específicos. La abstracción de datos permite manejar detalles de éstos de manera significativa, y esto es la motivación principal detrás de los *tipos de datos*.

La computación ofrece independencia del mundo concreto; mientras el hardware implementa un modelo de computación que es intercambiable con otros, el software está estructurado en arquitecturas que permiten crear sistemas enormes concentrándose en unos cuantos asuntos a la vez. Estas arquitecturas utilizan diversos tipos de abstracción.

Una forma central de abstracción en computación es la del lenguaje: nuevos lenguajes son desarrollados para expresar aspectos específicos de un sistema —por ejemplo, lenguajes de modelado para apoyar la planeación. Los lenguajes de programación pueden ser procesados con una computadora. Algunas características de los lenguajes de progra-

mación permiten al programador crear nuevas abstracciones, que incluyen subrutinas, módulos y los componentes de software.

A principios del decenio de los ochenta del siglo pasado comenzó a cobrar relevancia un nuevo estilo de programación, rodeado de toda una serie de conceptos que le daban coherencia: la programación orientada a objetos. En este paradigma (marco conceptual) el software se modela mediante abstracciones, los objetos del mundo real involucrados en el problema que se pretende resolver con el programa. Las cualidades que se consideran relevantes de los objetos reales son mapeadas entonces en una serie de atributos de los objetos que "habitan" en el mundo abstracto del programa. En un programa de manejo del personal de una empresa, por ejemplo, una entidad de tipo persona poseería probablemente características como éstas: nombre, cargo, departamento al que pertenece en la empresa, monto de salario mensual que percibe, etc. En un programa de manejo de datos de un censo de población, en cambio, una persona podría poseer atributos como: edad, número de hijos, tipo de casa que habita, ingreso medio mensual, sexo, lugar de nacimiento; en un programa de estudio de variación genética podría poseer atributos como: color de ojos o de piel, tipo sanguíneo. En fin, los atributos del objeto en un programa pretenden representar las características que poseen los objetos reales que modelan y que influyen determinantemente en el problema que se pretende resolver.

Esta manera de concebir la programación, aunada a la tecnología, ha dado lugar a un desarrollo sustancial de lo que se puede hacer con los programas. En la década de los setenta hubiera sido impensable tener aplicaciones como las que conocemos hoy en día, y el desarrollo de aplicaciones con una fracción de la complejidad que poseen las actuales hubiera requerido muchos años más de diseño y programación de lo que requieren hoy.

Una de las ventajas evidentes de la abstracción usada en la programación orientada a objetos es que el hecho de abstraer procesos o comportamientos comunes a una colección de objetos diferentes da lugar a programar esas cosas comunes como parte de una plataforma sobre la que se montan luego todos los elementos que las comparten. Este concepto, que los ingenieros de software llaman *reusabilidad de código*, es impensable sin la abstracción que permite identificar lo común. El resultado obvio es que los programadores desarrollan cada vez menos cosas redundantes y, en la medida de lo posible, reutilizan códigos que ya han sido probados exhaustivamente en el pasado, lo que reduce los errores potenciales inherentes a rehacer las cosas una y otra vez con pequeñas variantes. En síntesis, se desarrollan sistemas de software en menor tiempo y con mejores garantías de calidad y confiabilidad.

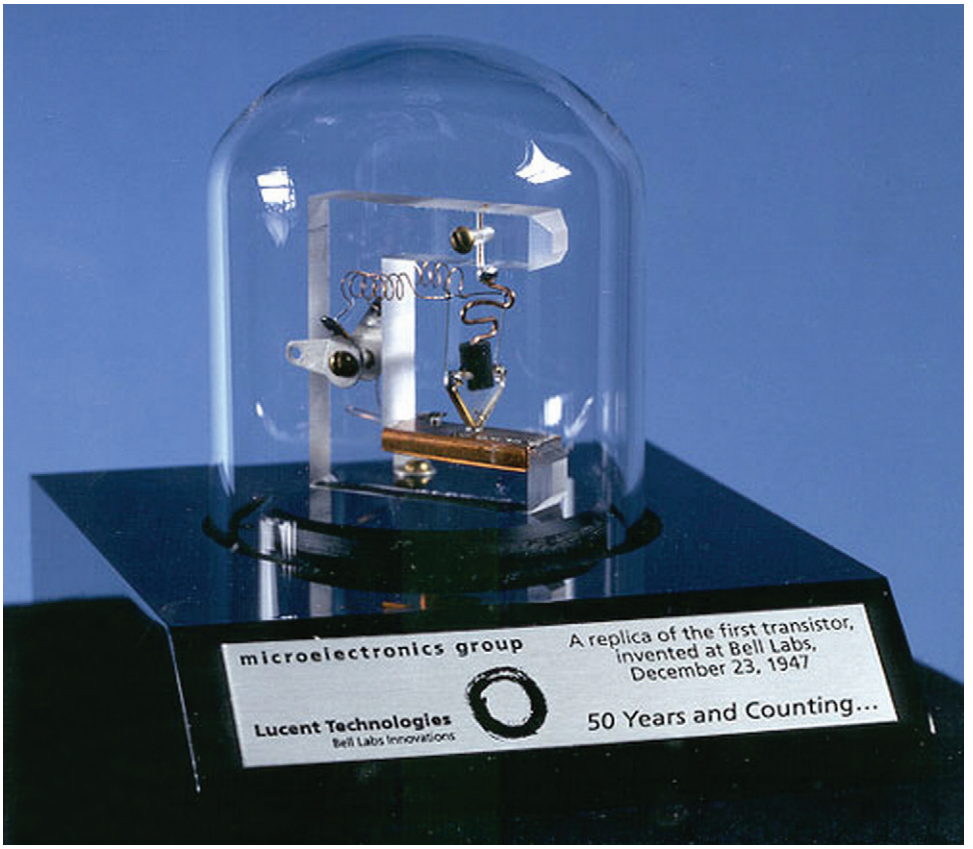
5.5 RESUMEN

Para capturar la esencia de los fenómenos, ya sean del mundo o de la imaginación, se usan modelos. Cuando estos fenómenos son complejos, con demasiados detalles y relaciones entre sus elementos, es difícil entenderlos. El computólogo crea abstracciones que representan los modelos, usando la notación apropiada para el fenómeno. Una actividad central de la computación es diseñar estos modelos, lenguajes para describirlos, analizarlos y ejecutarlos. Es así como logra construir sistemas que a la vez son enormes, flexibles, eficientes y confiables.

COMPUTADORAS

TEMA

6



© *Lucent Technologies.*

Y bien, Babbage, ¿con qué estás soñando?", a lo que respondí: "Pienso que todas estas tablas (señalando a las tablas de logaritmos) pueden ser calculadas por máquinas.

CHARLES BABBAGE, 1812.

La tecnología suficientemente avanzada es indistinguible de la magia.

ARTHUR C. CLARKE, 1973.

6.1 PROBLEMAS DE ELECTRICIDAD

—No puede ser, Arcadio, cuéntamelo desde el principio.

—Así sucedió: el problema es que hay dos lámparas en la estancia, una sobre la sala, otra sobre el comedor, controladas por un solo interruptor. El papá de Úrsula quería poner un interruptor doble para

poder independizarlas y así iluminar independientemente la sala y el comedor; él ya había pasado el cable extra por el ducto y sólo había que conectar los interruptores.

—¿Y cuál es el problema?

—No sé qué ocurrió, pero si prendíamos el interruptor de abajo solamente no ocurría nada, si prendíamos el de arriba tampoco, pero si dejábamos el de arriba prendido y prendíamos el de abajo, entonces se encendían ambos focos.

—¡Ay Arcadio, si serás burro! Hicieron una conjunción —dijo Remedios, que suele ser considerada la pequeña genio del salón.

—¿Una qué?

—Conjunción, ¿te acuerdas de eso? Lo vimos en mate el semestre pasado; lógica simbólica.

—¡Ah! Claro, la operación "Y".

—Claro, el foco se enciende sólo si ambos interruptores están en posición de encendido.

Para entender el argumento de Remedios, se considera que A y B (véase la tabla de verdad de la siguiente página) son los interruptores eléctricos y V significa encendido, la columna A y B es el efecto en la lámpara. De manera gráfica, significa que conectaron los interruptores en serie, la salida del primero como entrada del segundo, como se muestra a continuación:

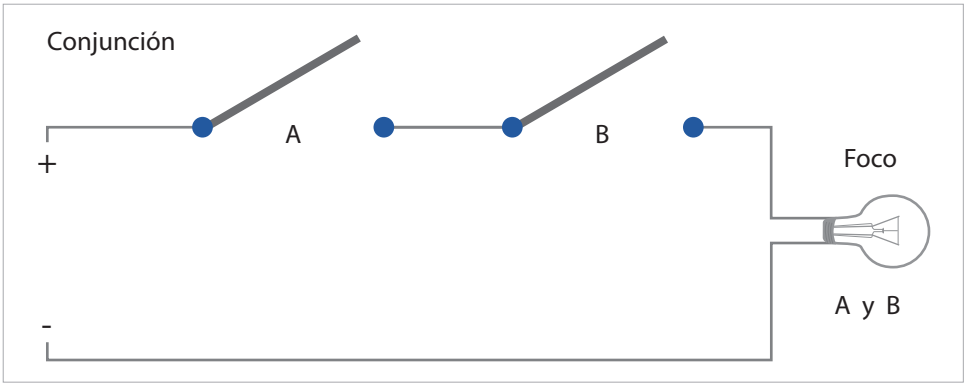


Figura 1. Conjunción.

Por lo que sólo si ambos interruptores están cerrados dejan pasar corriente al foco. El foco calcula la conjunción de A y B, que son los interruptores.

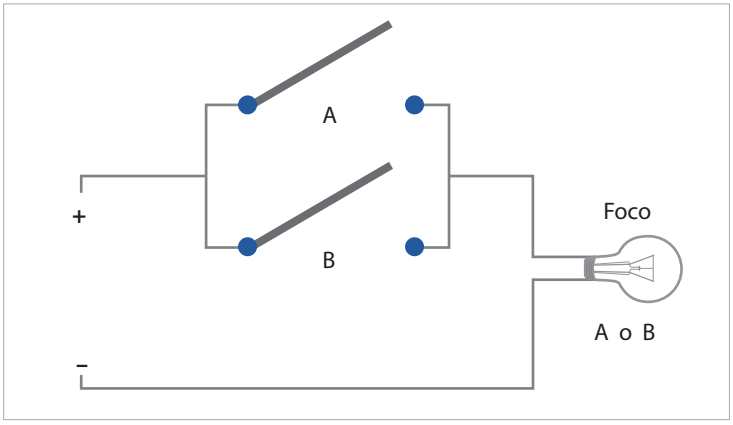


Figura 2. Disyunción.

6.2 SÓTANO: TRANSISTORES Y FUNCIONES DE CONMUTACIÓN

Curiosidades
Una tabla de verdad es una tabla matemática que se utiliza para calcular valores de expresiones lógicas en cada uno de sus argumentos funcionales. La tabla de verdad de la conjunción que menciona Remedios es:

A	B	A y B
F	F	F
F	V	F
V	F	F
V	V	V

Como Arcadio se dio cuenta, es posible calcular funciones lógicas, como la conjunción y la disyunción (figuras 1 y 2 respectivamente), con base en interruptores. Éste, por cierto, resulta ser el principio fundamental en que se basa la construcción de los modernos equipos de cómputo. Las funciones que se pueden calcular con base en interruptores, a las que se denomina *funciones de conmutación*, constituyen un conjunto con una estructura interesante: un álgebra booleana, y es esto lo que le confiere a nuestras computadoras todo su poder: tanto las operaciones elementales de la lógica como las de la aritmética, por ejemplo, se pueden expresar como combinaciones de unas cuantas funciones de conmutación.

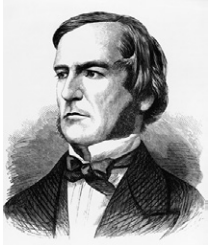
Las funciones de conmutación son importantes porque en ellas todo se expresa en términos binarios: 0 y 1, verdadero y falso, blanco y negro; no importa, el caso es que sólo hay dos valores posibles. Esto, como se vio en el tema sobre información, responde al hecho de que las computadoras trabajan en binario; todo lo que procesan y almacenan puede considerarse expresado como cadenas de ceros y unos.

Como se aclaró en aquel tema, esto no es literal, no hay ceros y unos escritos en la memoria de las computadoras, lo que ocurre es que tanto la memoria como el resto de los circuitos de la computadora están capacitados para distinguir entre dos niveles de voltaje, cero y cinco volts; típicamente, a uno de ellos se le asocia entonces el valor 0 y al otro el valor 1, de allí que se diga que se trabaja en binario: hay o no hay voltaje, hay o no hay corriente, hay o no hay carga eléctrica; así es realmente el maniqueo mundo de las computadoras digitales.

De allí la importancia de los interruptores. Cada interruptor, como su nombre lo indica, interrumpe o no el flujo de corriente; tiene, pues, una respuesta binaria. A lo largo de la historia de la tecnología de las computadoras se han usado diversos dispositivos como interruptores, relevadores, bulbos (formalmente llamados tubos de vacío) y, finalmente, transistores, los que se siguen usando hasta la fecha. La imagen de la portada de este tema es del primer transistor que se ensambló en los Laboratorios Bell en 1947. De hecho, típicamente se utiliza este indicador tecnológico como medio de clasificación histórica:

1] *Primera generación*: 1940 a 1956. Máquinas hechas con base en tubos de vacío.
2] *Segunda generación*: 1957 a 1963. Construidas con base en transistores individuales.
3] *Tercera generación*: 1964 a 1970. Hechas con base en circuitos integrados de baja densidad.
4] *Cuarta generación*: 1971 al presente. Computadoras basadas en circuitos integrados de muy alta densidad.

En la lista previa, el término *densidad* se refiere al número de transistores que es posible empacar juntos en una sola cajita para llevar a cabo una labor o conjunto de labores específicas. Estas cajitas, también llamadas pastillas, circuitos integrados o chips, son los



George Boole | © Latin Stock México.

Curiosidades
El matemático británico George Boole (1815-1864) formuló las reglas que gobiernan las operaciones del álgebra que ahora lleva su nombre: el álgebra booleana. Su labor era, sin embargo, más ambiciosa: pretendía encontrar las reglas que gobiernan el pensamiento lógico en general; podríamos decir que quería descubrir los algoritmos que gobiernan el pensamiento humano. Más tarde, el trabajo de Boole fue retomado por Edward Vermilye Huntington (1874-1952), el matemático estadunidense que estableció las reglas básicas a partir de las cuales se pueden deducir las formuladas por Boole. En términos matemáticos, Huntington realizó la formulación axiomática del álgebra de Boole en 1933.

Edward V. Huntington.

pequeños dispositivos negros rodeados de patas de metal que se pueden ver en las tablas de circuitos de prácticamente todos los aparatos electrónicos. Baja densidad significa que en un solo chip, con un área de unos cuantos centímetros cuadrados, es posible poner sólo unas decenas o cientos de transistores. A partir de los años setenta del siglo XX fue posible poner miles de transistores, y ya en la década de los ochenta se pasó de los cientos de miles al millón de transistores. El número ha seguido creciendo, y a principios del siglo XXI era ya de miles de millones de transistores.

Bulbo | © *Tvezymer.*

6.2.1 Transistores

Los transistores son entonces el componente más elemental de las modernas computadoras, porque se comportan como interruptores: un transistor deja o no pasar la corriente eléctrica a través de él, pero a diferencia de los interruptores de Arcadio, y de los que se usaban en las primeras redes de teléfonos y computadoras, el proceso no es mecánico, sino eléctrico.

Transistores | © *Dary.*

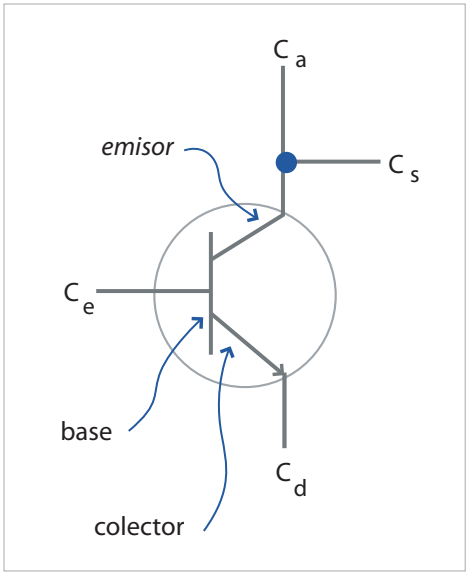


Figura 3. Representación de un transistor.

Curiosidades

John Bardeen, William Shockley y Walter Brattain fueron el equipo de AT&T que inventó el transistor, por lo que recibieron el Premio Nobel de Física en 1956. Bardeen obtuvo luego otro Premio Nobel por sus trabajos en superconductores. Shockley fundó su propia compañía para fabricar transistores, la Shockley Semiconductor, para lo cual eligió un lugar agradable cerca de donde vivía su madre y de la Universidad de Stanford, de donde pensaba reclutar jóvenes talentos. El lugar se llama hoy Silicon Valley y es, sin duda, el parque industrial más importante en el mundo de la tecnología.

Esquemáticamente, así se representa un transistor:

C_a es la corriente de alimentación, siempre está allí, proviene del suministro de corriente del aparato. C_e es la corriente de entrada: si al transistor se le suministra una corriente de entrada que exceda cierto umbral entonces se dice que la entrada está en alto o, para ser consistentes, está en 1. C_s es la corriente de salida que, como se verá, depende de la de entrada. Por último, C_d es la corriente de drenaje del transistor.

En abril de 1965, Gordon E. Moore creó la llamada Ley de Moore, que establece, mediante la observación empírica del fenómeno, cómo evoluciona el poder de cómputo con el tiempo. La densidad —esto es, el número de transistores que se pueden poner por unidad de área— crece a una tasa aproximada de 50% al año; es decir, se duplica cada dos años. Este crecimiento de densidad trae como consecuencia uno equiparable con el poder de cómputo contenido en cada chip y significa que éste crece exponencialmente a lo largo del tiempo. La gráfica 1 muestra los datos de número de transistores en los microprocesadores de la compañía Intel (precisamente la compañía que Moore contribuyó a fundar en 1968), desde el 4004 de 1971, a la sazón el primero de la historia, hasta el

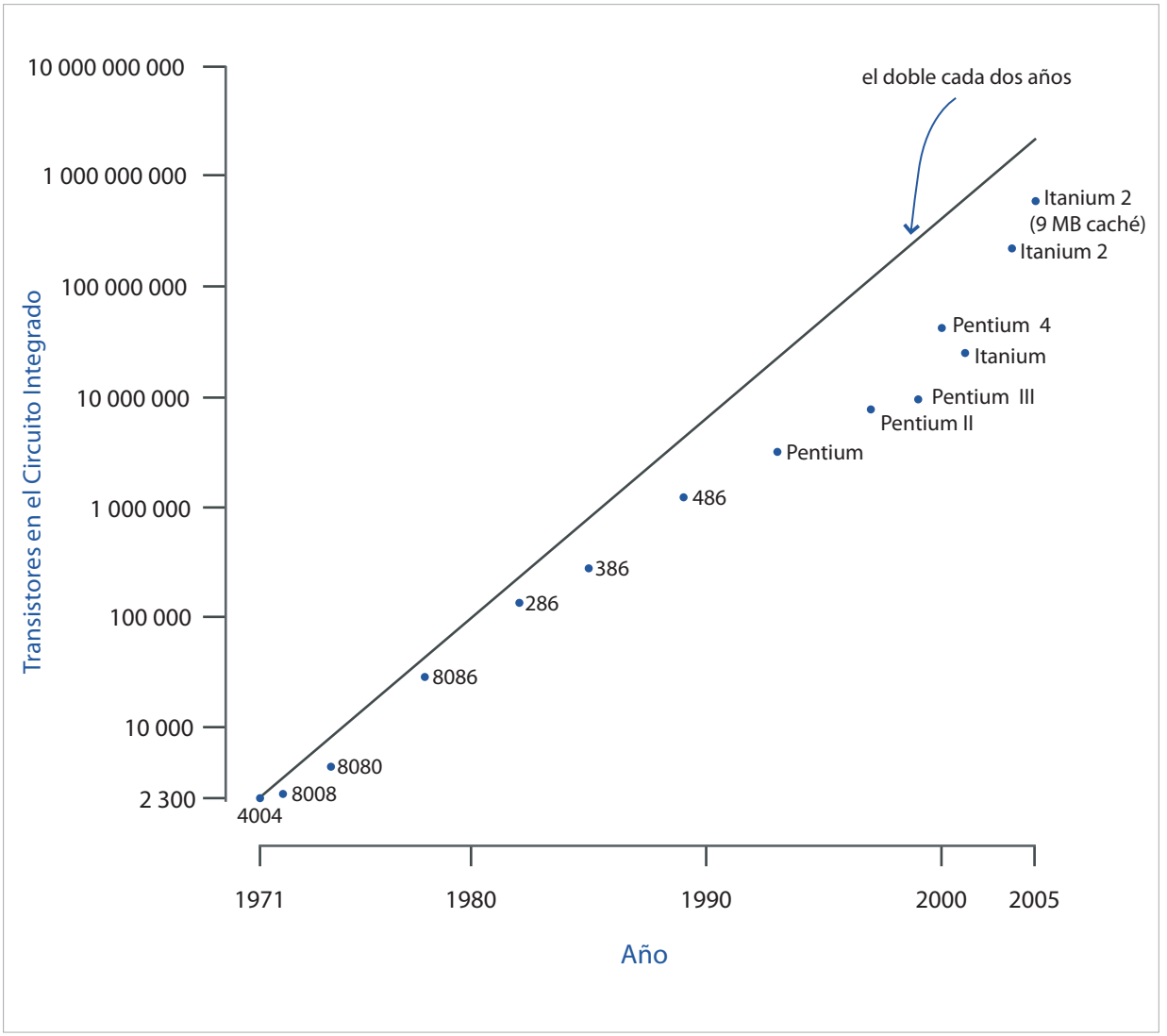
Curiosidades

Con el paso de los años, el tamaño de los transistores se ha ido reduciendo cada vez más; compárese el tamaño de un bulbo de 1947 con otros transistores más recientes. En 1989, el procesador 486 DX rebasó el millón de transistores y el tamaño de cada uno fue de una micra; diez años después, los procesadores contenían más de 28 millones de transistores y el tamaño de cada transistor se redujo a 0.18 micras.

Itanium 2 de 2005; nótese que en el eje vertical la escala es logarítmica (en potencias de 10) mientras que el horizontal es lineal, lo que evidencia el comportamiento exponencial mencionado. En abril de 2005, sin embargo, el propio Moore declaró que la ley que había descubierto estaba por caducar: subir la densidad significa, claro está, hacer cada vez más pequeños los transistores, que en la actualidad tienen partes de apenas unos seis átomos de grosor y, aun cuando se pudiera continuar miniaturizando, al mismo ritmo se generan problemas colaterales muy difíciles de resolver. La densidad de los microprocesadores actuales hace que éstos generen mucho calor, energía desperdiciada que por añadidura daña los circuitos. Hoy en día es imposible ver el chip del procesador en la computadora, suele estar oculto bajo una enorme masa de aluminio, cuyo único propósito es disipar el calor para que el chip no se funda; desde hace unos años se ha añadido, además, un pequeño ventilador sobre el disipador.

Las cosas funcionan así: los transistores están hechos de un material denominado semiconductor, una cosa que a veces se comporta como conductor (deja fluir la corriente eléctrica a través de él) y a veces como aislante (no deja fluir la corriente). Su comportamiento depende de si le es suministrada o no corriente eléctrica por otro lado. En un transistor, la corriente de alimentación C_a entra por una línea llamada emisor, la corrien-

Gráfica 1. Ley de Moore.



te de entrada C_e por otra llamada base y la de drenaje C_d sale por el colector. La corriente que entra por la base es la que determina si el transistor se comportará como conductor o como aislante: si C_e excede cierto umbral conduce, si no se comporta como aislante. Un transistor se comporta de manera parecida a una tubería de agua con una válvula de escape: ésta hace las veces de la corriente de entrada C_e; si la válvula de escape está cerrada, entonces toda el agua fluye por la salida, si no se va al drenaje.

El lector atento probablemente habrá notado que la corriente de salida es contraria a la de entrada: cuando ésta es 1 aquélla es 0 y viceversa. En efecto, un transistor puesto en esta situación es un inversor.

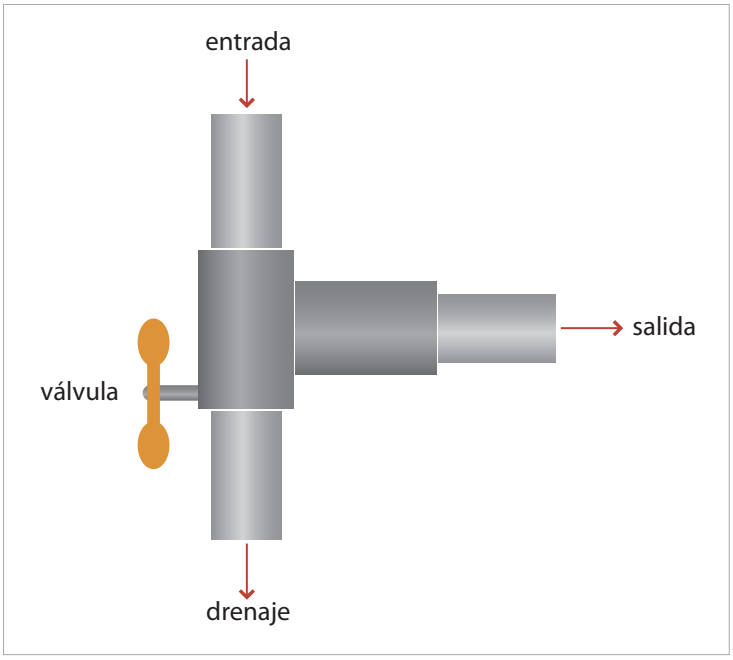


Figura 4. Una tubería hidráulica como símil de un transistor.

6.3 PLANTA BAJA: COMPUERTAS Y CIRCUITOS INTEGRADOS

6.3.1 Compuertas elementales

Remedios le mostró a Arcadio algunas de las funciones de conmutación más simples: la conjunción ("y" o *and* en inglés) y la disyunción ("o" u *or* en inglés). Éstas son exactamente las mismas funciones que suelen usarse en el cálculo proposicional, como se puede ver en el tema de abstracción. Si se añade a este par de funciones una más, como la negación (el "no", o *not* en inglés), se obtiene un conjunto con el que es posible construir cualquier otra función de conmutación.

La negación es aún más simple que la disyunción o la conjunción: recibe un único argumento y entrega el complemento binario del mismo como resultado. En la siguiente tabla de verdad se muestra la negación.

A	No A
0	1
1	0

Tabla 1. Tabla de verdad de la negación.

Conviene aclarar algunas cosas a propósito de esta manera de representar funciones usando una tabla de verdad. Probablemente, el lector está acostumbrado a representar funciones mediante lo que se denomina una regla de correspondencia: una expresión algebraica que dice cómo calcular al valor que adquiere la variable dependiente, dado el valor de la independiente; algo del estilo de $f(x) = 2x + 4$, si se da un valor para la variable independiente, $x = 3$, entonces la regla dice qué valor de la variable dependiente le corresponde: $f(3) = 2 \times 3 + 4 = 10$. Pero hay que recordar el sentido original del concepto de función: *una relación que vincula a cada elemento de un conjunto (dominio) con uno y sólo uno de otro (contradominio)*. La regla de correspondencia es particularmente útil en el caso de las funciones en las que el dominio es infinito, porque de manera sintética permite definir qué elemento del contradominio le corresponde a ¡todos y cada uno de los elementos del dominio infinito!

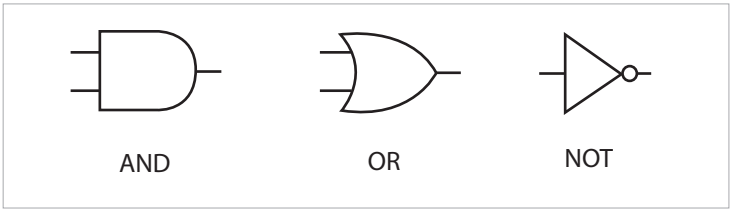
En el caso de las funciones de conmutación, sin embargo, la situación es bastante más sencilla, el dominio es finito; se puede hacer una lista de todos sus elementos y, por tanto, una tabla —a la que se denomina *tabla de verdad*—, en la que se especifique qué elemen-

to del contradominio le toca a cada quien. Por supuesto, también se podría usar una regla de correspondencia (y se hará posteriormente).

Las tres funciones elementales mencionadas: disyunción, conjunción y negación, constituyen una base (por cierto, no la única) sobre la que se puede construir cualquier otra función de conmutación y, por tanto, cualquier computadora. Debido a esto, los ingenieros suelen usar símbolos especiales para denotarlas al diseñar circuitos.

Es entonces importante poder expresar cualquier función de conmutación en términos de AND, OR y NOT. En adelante, se usará el símbolo "+" para denotar el OR o disyunción; el AND o conjunción se de-notará sin usar un símbolo particular, simplemente poniendo juntas las variables que intervienen en la operación: por ejemplo, AB denota la conjunción de A y B, variables booleanas. El NOT o negación se denotará poniendo una barra sobre la variable que se niega. Por ejemplo, la expresión:

Figura 5. Símbolos de las compuertas AND, OR y NOT.



$$f(A, B) = AB + A\overline{B}$$

corresponde a una función cuya tabla de verdad es la tabla 2. Se muestra el cálculo de las variables negadas (columnas 4 y 5) y el de los términos parciales (columnas 6 y 7) para facilitar el de la función en la tercera columna:

A	B	f(A, B)	$\overline{A}$	$\overline{B}$	$A\overline{B}$	$\overline{A}B$
0	0	0	1	1	0	0
0	1	1	1	0	0	1
1	0	1	0	1	1	0
1	1	0	0	0	0	0

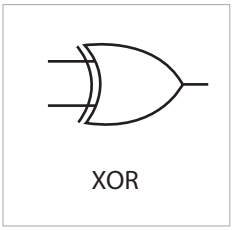
Tabla 2. Tabla de verdad de la función f (A, B).

Esta función es también muy simple y bastante útil. Se le suele llamar *disyunción exclusiva* (XOR, en inglés). Es muy similar a la disyunción común, salvo que no es verdadera cuando ambos argumentos son verdaderos (de allí que se la llame exclusiva: si la disyunción es "A o B", la disyunción exclusiva es "A o B pero no ambos"). Si se piensa en interruptores, el XOR puede ser calculado por un interruptor de los llamados "de escalera": cuando ambos interruptores, el que está al pie de la escalera y el que está arriba, se encuentran en la misma posición, entonces el foco está apagado; si alguno de ellos cambia de posición, el foco se enciende.

La disyunción exclusiva es tan útil por sí misma que suele usarse el operador "$\oplus$" para representarla en las expresiones de funciones de conmutación, y los diseñadores de circuitos también poseen un símbolo para ella:

A los símbolos usados para denotar las funciones AND, OR, NOT y XOR en los circuitos de electrónica digital se les denomina compuertas lógicas.

Figura 6. Símbolo de compuerta XOR.



6.3.2 Diseño lógico

Es posible diseñar circuitos que calculen cualquier función de conmutación usando compuertas. Por ejemplo, la función:

$$S(A, B, C) = A \oplus B \oplus C$$

es calculada por el circuito:

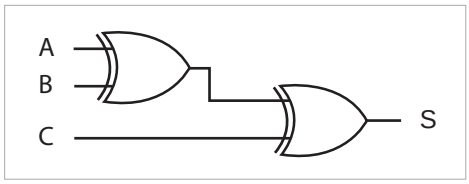


Figura 7. Circuito para calcular la función S.

Por otra parte, la función:

$$T(A, B, C) = C(A \oplus B) + AB$$

se calcula con el circuito:

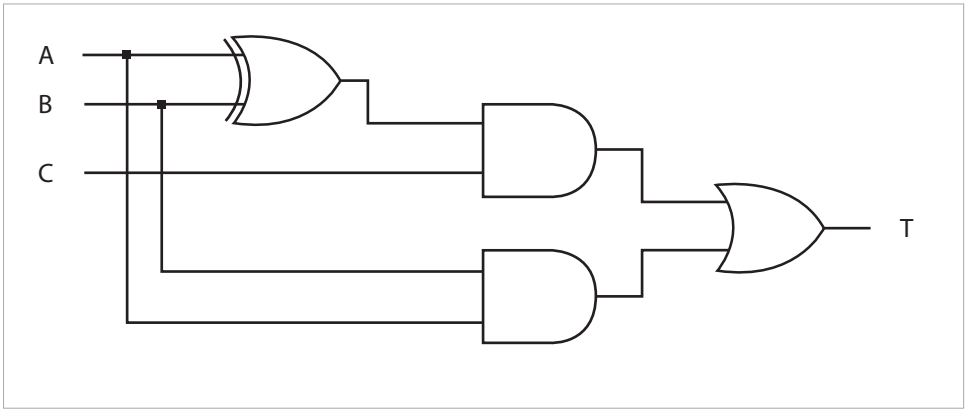


Figura 8. Circuito para calcular la función T.

Estas dos funciones tienen, juntas, un significado peculiar. Para hacerlo evidente se debe recordar lo que se sabe acerca de la representación binaria posicional de los números: usando dos bits, el 0 se representa como 00, el 1 como 01, el 2 como 10 y el 3 como 11. Si se ponen las funciones en una tabla de verdad (tabla 3) y se toman en cuenta las columnas T y S juntas, como la expresión posicional binaria de un número, resulta que son la suma de los valores de las otras tres columnas: A, B y C. Es decir, si se presta atención a un renglón cualquiera de la tabla y se observa el número binario escrito en las columnas T y S, resulta ser el número de unos contenidos en ese mismo renglón en las columnas A, B y C. ¡Bien! Se ha construido la tabla de verdad de un circuito sumador. Normalmente, a la columna T se le denomina el acarreo de la suma, y a la columna S la suma propiamente dicha. En el sistema decimal, cuando se suman 8 y 5 el resultado es 13, es decir, tres unidades y una decena; en lenguaje coloquial "3 y llevamos 1"; ése 1 es el acarreo que significa que sumando los dígitos en una posición determinada se ha completado una unidad de la siguiente posición. En binario, si se suma 1 + 1 se obtiene 2, que en binario se escribe 10, es decir "0 y llevamos 1", o sea 0 unidades y una "biena".

A	B	C	T	S
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

Tabla 3. Tabla de verdad de las funciones S y T.

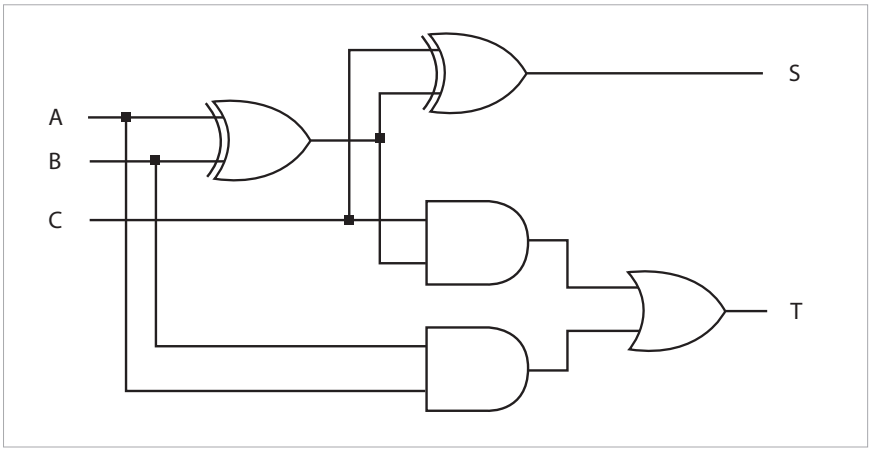


Figura 9. Diseño de sumador completo.

El circuito que se construyó se denomina *sumador completo de un bit* y se podría fabricar y ponerlo en una cajita cuyo interior se vería así:

Se denomina sumador de un bit y no de tres, porque realmente está sumando una sola posición de un par de números binarios, considerando que ya se tenía un acarreo previo, proveniente de la posición anterior, al que se designa C. Los bits que se suman son A y B. Así que ahora se puede ser más ambicioso y pensar en un sumador de números de ocho bits, lo que se vería así:

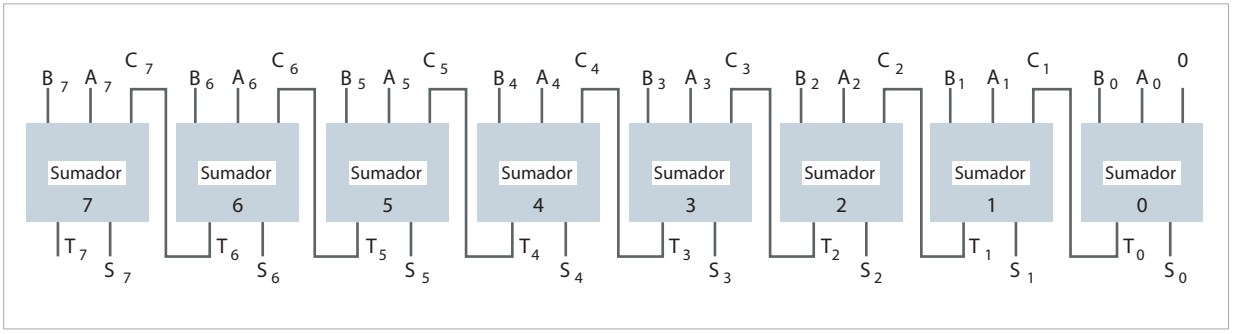


Figura 10. Diseño de un sumador de operandos de ocho bits, encadenando sumadores de un bit.

Los números que se suman son expresados en binario como: $A_7\ A_6\ A_5\ A_4\ A_3\ A_2\ A_1\ A_0$ y $B_7\ B_6\ B_5\ B_4\ B_3\ B_2\ B_1\ B_0$, el acarreo de entrada a la suma de A y B es, por supuesto, 0, por eso la línea de acarreo de entrada en el extremo derecho tiene escrito un 0.

Podría hacerse un sumador para números de cualquier tamaño, sólo hay que encadenar tantos sumadores de un bit como sea necesario. Del mismo modo, se puede pensar en hacer un circuito que multiplique o que haga cualquier operación que se ocurra. Éste es el principio de construcción de circuitos integrados: en cada chip se empacan unidades funcionales que permiten hacer algo particular. En este caso, se podría pensar en un primer chip con un sumador de un bit, luego intentar ser más ambiciosos y poner en un chip un sumador de ocho bits en cuyo interior se tendrían, de hecho, ocho sumadores de un bit; después podrían agregarse circuitos para hacer otras operaciones y añadir uno más para poder elegir cuál operación hacer.

Esto constituiría lo que se denomina una unidad aritmético-lógica (ALU, por sus siglas en inglés): el corazón de cualquier procesador central de una computadora o, en algún sentido, su cerebro.

6.3.3 Transistores y compuertas NAND

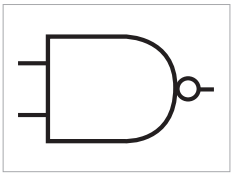


Figura 11.Símbolo de la compuerta NAND.

Ya que un solo transistor se comporta, en principio, como una compuerta NOT o inversor, si se conectan en serie dos de éstos se obtendrá, no una compuerta AND, como se esperaría al conectar de esa forma dos interruptores, sino una cosa llamada compuerta NAND, la negación de un AND. Lo que puede representarse como en la figura 11.

Si se conectan en paralelo, de igual modo no se obtiene un OR, sino un NOR, como se muestra en la tabla:

A	B	A NAND B	A NOR B
F	F	V	V
F	V	V	F
V	F	V	F
V	V	F	F

Tabla 4. Tabla de verdad de las compuertas NAND y NOR.

Ahora puede plantearse la pregunta: ¿se pueden construir circuitos usando estas funciones? La respuesta es sí. De hecho resulta que podría escogerse alguna de las dos y con eso bastaría. Para probar esto se requiere expresar las funciones previas, el AND, el OR y el NOT en términos de NAND o NOR. Se utilizará la primera. Si se usa el símbolo $\odot$ para representar al NAND, resulta:

$$\overline{A} = A \odot A$$
$$A\ B = (A \odot B) \odot (A \odot B)$$
$$A+B = (A \odot A) \odot (B \odot B)$$

O bien, esquemáticamente:

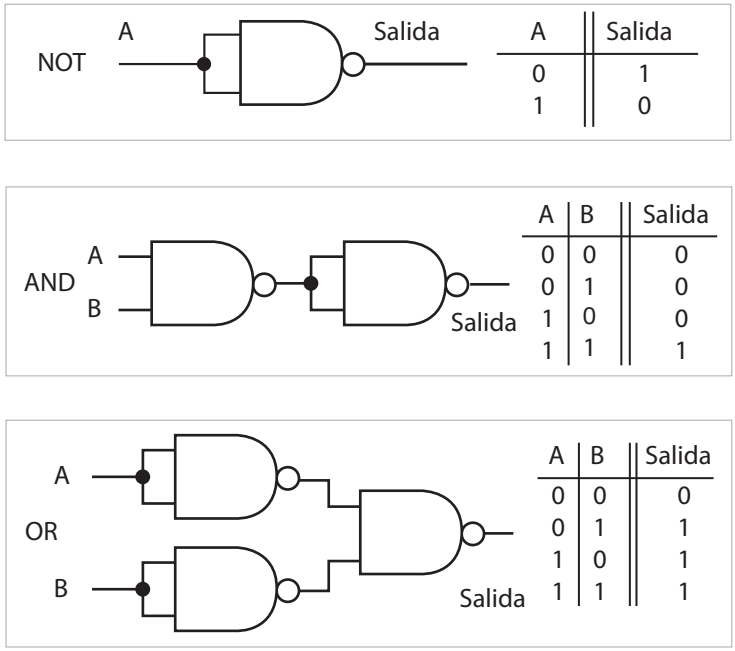
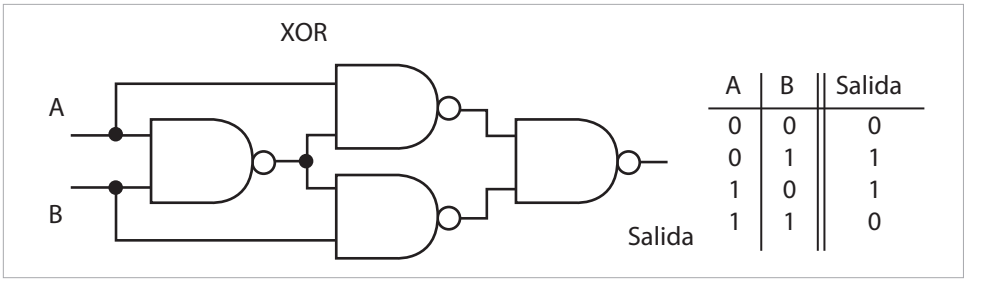


Figura 12. Circuitos para las funciones NOT, AND y OR en términos de NAND.

Curiosidades
Luego de colaborar en el proyecto de construcción de ENIAC (Computador e Integrador Numérico Electrónico), la primera computadora electrónica digital de propósito general de la historia, el equipo formado por Eckert, Mauchly, Goldstine y Von Neumann, se dio a la tarea de diseñar una nueva máquina, a la que denominaron EDVAC (Electronic Discrete Variable Automatic Computer). Una de las cosas que más molestaba al equipo era la dificultad de programar ENIAC, pues esto requería ir a los paneles de conexión de la computadora, desconectar cientos de cables de su lugar actual y reconectarlos en posiciones diferentes, labor tan delicada que sólo era encomendada a mujeres (ciertamente más cuidadosas en este tipo de labores). De una de las discusiones acerca de esto surgió la idea de una computadora de programa almacenado o de arquitectura de Von Neumann. En el nombre se omite a los demás por una casualidad; aparentemente fue John von Neumann quien puso en blanco y negro los resultados del trabajo conjunto en un memorándum que nunca debió haber salido del ámbito restringido del equipo de colaboradores: "First draft of a report on the EDVAC" (Primer borrador de un reporte sobre la EDVAC).

Es decir, cualquiera de las tres operaciones fundamentales: el OR, el AND y el NOT se pueden expresar usando sólo el NAND. Cualquier función de conmutación, entonces, se puede escribir usando únicamente el NAND. Así que no hay problema con que eso sea lo único que los transistores sepan hacer.

Por cierto, sólo para que nada falte:



A	B	Salida
0	0	0
0	1	1
1	0	1
1	1	0

Figura 13. Circuito de la función XOR.

6.4 PRIMER PISO: ARQUITECTURA DE COMPUTADORAS

En el tema sobre abstracción se vio que toda computadora se puede modelar mediante una máquina de Turing, o cualquiera de los otros modelos equivalentes. Ahora se mostrará el modelo propuesto por Von Neumann, que sigue siendo la base de la mayoría de las computadoras modernas.

6.4.1 Arquitectura de Von Neumann

La unidad aritmético-lógica de una computadora puede ser su corazón, pero de muy poco serviría por sí misma. No podría hacer cosas mucho más complejas que las que hace una simple calculadora de bolsillo. Algo fundamental de una computadora es que es *programable*, aun las más simples, como la de un reproductor de MP3, la de un microondas, la de un aparato para grabar programas de televisión, la de un cajero automático o la de una caja registradora. Esencialmente, una computadora es un *ejecutor de algoritmos*; cada tarea llevada a cabo por una computadora, por simple que sea, está programada de antemano, es un algoritmo que ha sido expresado en el lenguaje que entiende la computadora, por lo que ésta se limita a ejecutarlo paso por paso.

En la práctica, las computadoras actuales requieren más cosas que una ALU:

- La unidad central de proceso (CPU, por sus siglas en inglés),[1] donde se ejecutan los programas, constituida por unidades menores;
- la unidad aritmético-lógica, que se encarga de llevar a cabo las operaciones;
- y la unidad de control, que determina qué operación se debe hacer y con qué operandos.

Puede haber más de una CPU, en cuyo caso se dice que la computadora es un sistema multiprocesador.

[1] Coloquialmente, en ocasiones se le llama CPU a la caja que contiene la computadora propiamente dicha, en la que se conectan el teclado, el ratón y el monitor. En realidad, el CPU está dentro de ese gabinete y no es lo único que hay allí, también están la memoria y buena parte de los dispositivos de entrada y salida.

- Memoria. En ella se almacenan dos cosas: programas y datos. A la memoria que se borra cuando la computadora se apaga se le denomina memoria primaria, a la que permanece se le denomina secundaria. La memoria principal de una máquina, a la que se le suele llamar RAM,[2] es primaria; un disco duro o un CD son memoria secunda-ria. Para que un programa pueda ejecutarse es indispensable que esté en memoria primaria.
- Dispositivos de entrada y salida. Son todos aquellos dispositivos que permiten la interacción de la computadora con el medio exterior: teclado, ratón, tarjeta de red, impresora, tarjeta de video, puertos USB, paralelos, seriales, etcétera.

Esquemáticamente, se pueden representar los componentes genéricos de una computadora como se muestra en la figura 14:

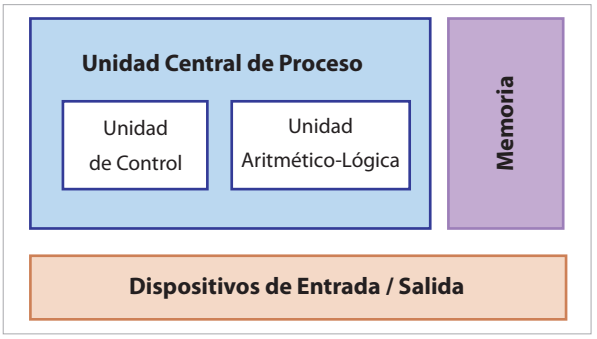


Figura 14. Representación esquemática de la arquitectura de Von Neumann.

A la estructura mostrada se le denomina arquitectura de Von Neumann; su característica fundamental es que posee una memoria para almacenar no sólo los datos sobre los que se ejecuta el algoritmo, sino también el programa que codifica dicho algoritmo. Hasta antes de la invención de esta cualidad fundamental, las computadoras eran programadas reconfigurando el cableado de los componentes; es decir, se "reconstruía" la computadora cada vez que se cambiaba el algoritmo que ejecutaba. No es necesario decir que esto era muy complicado y propenso al error, así que realmente fue una gran idea que la computadora pudiera guardar el programa en la memoria, para así poder cambiarlo fácilmente.

Se vio que un programa consiste en una secuencia de instrucciones que manipula datos. De la ejecución de éstas se encarga la unidad aritmético-lógica del procesador. Se deben guardar en la memoria de la computadora tanto las instrucciones del programa como los datos que manipula. A lo largo de la ejecución, la máquina que ejecuta el programa transita por una secuencia de estados, determinados por lo que se ha hecho sobre los datos de entrada y las características de los resultados parciales obtenidos hasta el momento.

En el caso del procesador, es necesario mantener el estado de alguna manera, recordar qué se ha obtenido hasta el momento. La manera de hacerlo es almacenando datos en la memoria, pero como la memoria externa al procesador suele ser muy lenta y estos datos serán usados asiduamente, los procesadores poseen pequeñas cantidades de memoria de

Curiosidades

John von Neumann (28 de diciembre de 1903-8 de febrero de 1957) nació en Hungría y recibió su doctorado en matemáticas a los 23 años, después de haber estudiado ingeniería química; fue uno de los pensadores más influyentes del siglo XX. Trabajó en el desarrollo de las primeras armas nucleares. Es considerado el padre de la teoría de juegos; también fue muy importante en política, concibió el concepto de MAD (Mutually Assured Destruction, o destrucción mutua asegurada), que dominó la estrategia nuclear estadunidense durante los tiempos de posguerra. Fue pionero de la computadora digital moderna y de la aplicación de la teoría de operadores a la mecánica cuántica. Nunca olvidaba nada de lo que leía y su habilidad para realizar cálculos mentales era legendaria. Todos los que lo conocían estaban de acuerdo en dos cosas acerca de Von Neumann: lo carismático y amable de su personalidad, y lo mucho más inteligente que era.

John von Neumann | *© Latin Stock México.*

[2] El término RAM proviene de las siglas en inglés de Random Access Memory, memoria de acceso aleatorio. El calificativo "aleatorio" aquí significa que se puede acceder a cualquier lugar de la memoria arbitrariamente en cualquier momento. Se le dio ese nombre para distinguirla del tipo de memoria en la que el acceso debía llevarse a cabo en cierto orden.

acceso expedito dentro de ellos. Esta memoria está directamente conectada a la unidad aritmético-lógica y allí se ponen los datos traídos de la memoria primaria externa para usarlos como operandos y se almacenan los resultados antes de llevarlos de regreso a la memoria externa. Esta pequeña cantidad de memoria interna está dividida en unidades manejables por la unidad aritmético-lógica en celdas llamadas registros del procesador. Es el tamaño de éstos lo que da su nombre a la arquitectura; un procesador de 32 bits tiene registros de ese tamaño y uno de 64 bits tiene registros del doble.

6.4.2 Frecuencia de operación

Cuando se anuncia un nuevo procesador, por lo general se dice a cuántos gigahertz "corre"; a esto se le llama *frecuencia de operación* y se refiere al número de pasos por segundo que ejecuta el procesador. Todos los procesadores están diseñados con base en un reloj que, a la manera de los barcos de remo antiguos, tiene la función de que todos sus transistores y demás componentes trabajen sincrónicamente, al ritmo del reloj. Esto para el diseño del procesador y la estabilidad de su operación.

Ejemplo de un sistema en que el tambor marca la frecuencia de operación.

De esta manera, todo el procesador divide el tiempo total de ejecución de cada instrucción en lo que se denomina ciclos de reloj y que no son otra cosa sino pulsos regulares que indican cuándo hacer un cambio o transmitir un dato de un lugar a otro a través de señales eléctricas, y cuándo se debe estar en reposo esperando que todo se estabilice.

Cuando la adrenalina produce sus efectos en una persona y ésta se encuentra "acelerada", no sólo en sentido coloquial sino también en lo referente al ritmo cardiaco, se dice que está apurada y haciendo más cosas mucho más rápido de lo usual. En un procesador pasa lo mismo: cuanto mayor sea su frecuencia de operación (los gigahertz a los que corre) mayor será el número de cosas que hace por unidad de tiempo. Una frecuencia de un hertz significa que algo ocurre una vez por segundo, un kilohertz son mil ciclos por segundo, un megahertz es un millón y un gigahertz es mil millones de ciclos por segundo. Eso significa que... ¡maravilla de maravillas!, una computadora hace ¡miles de millones de operaciones en un segundo!

6.4.3 La jerarquía de memoria: la idea del caché

Hoy en día, en cada computadora existe una muy variada gama de dispositivos de memoria. Ya se habló de la distinción entre la primaria (volátil) y la secundaria, pero dentro de cada una de estas categorías existe una gran diversidad. La memoria fundamental para

todo sistema de cómputo, de la que depende esencialmente la ejecución de los algoritmos, es la primaria, dado que las arquitecturas de Von Neumann requieren que el programa se almacene en ésta para poder ser ejecutado.

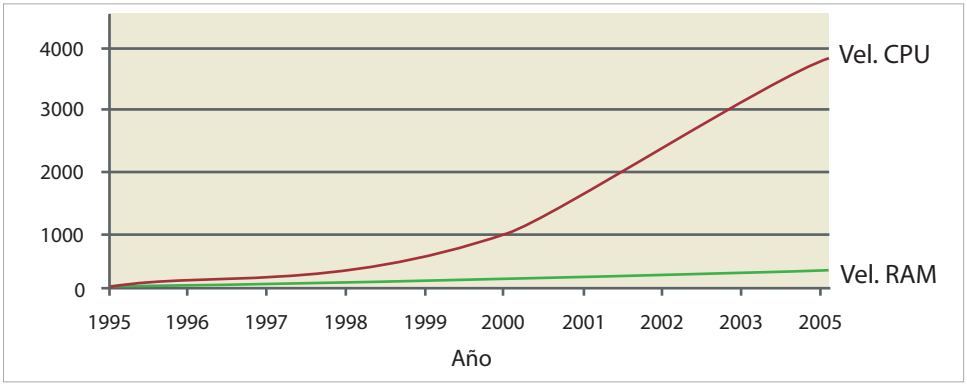
Existen diversos tipos de memoria primaria dependiendo de la tecnología usada para su fabricación, lo que conlleva características que influyen notablemente en el desempeño. Existen memorias a las que, cuando se les solicita un dato, pueden entregarlo a la misma velocidad a la que trabaja el procesador; es decir, su frecuencia de operación es la misma que la de éste. Pero esto no es lo común, normalmente las memorias tardan mucho más en responder, son cientos de veces más lentas que los procesadores, lo que significa que cuando se requiere de un dato en el procesador, éste tendría que esperar una eternidad (en su escala de tiempo, cientos o miles de ciclos de reloj) para poder continuar su labor.

Pero la solución real no consiste en ponerle a la computadora sólo la memoria más rápida posible, la que no se tarda nada, porque es muy cara. En efecto, la memoria es tanto más cara cuanto más rápida. Construir un sistema usando sólo la memoria más rápida lo haría increíblemente costoso, prácticamente nadie podría comprarlo. Por supuesto, el otro extremo tampoco es deseable: poner sólo memoria de bajo desempeño haría un sistema muy barato, pero también inútil.

Para lograr el equilibro deseable se construye lo que se denomina una jerarquía de memoria: el sistema posee varios tipos de memoria, desde la muy cara hasta la más barata; más de la barata y menos de la más cara. La más veloz, de la que hay poca, se pone cerca del procesador para que lo atienda preferentemente. Pero como, debido a su tamaño, a lo mejor no cabe todo lo que el procesador necesita en un momento dado, se pone otro nivel de memoria, un poco más lejos del procesador, un poco más lenta, pero con mayor capacidad de almacenamiento. Y así sucesivamente hasta llegar a la memoria principal del sistema, que normalmente se denomina RAM.

El nivel de memoria más cercano al procesador forma parte de la estructura misma de éste. Está constituido por los ya mencionados registros del procesador en donde se almacenan los operandos que se requieren en el corto plazo. Los procesadores, a lo largo de la historia, han tenido cada vez más registros: se ha pasado de tener uno solo en el decenio de los setenta del siglo XX hasta cientos en las máquinas actuales.

A los niveles de memoria que se encuentran entre la memoria principal y los registros del procesador se les denomina *memoria caché*. Al caché más veloz, más caro, más pequeño y, por supuesto, más cercano al procesador se le denomina *caché de nivel L1*, el siguiente es el de L2 y así sucesivamente hasta el caché inmediato anterior a la memoria principal. El número de cachés es decisión del diseñador del sistema, y cuanto mayor sea mejor será el desempeño porque siempre, o casi siempre, el que responde a una petición de un dato de memoria es el mejor caché de entre los que poseen el dato.



Gráfica 2. Brecha entre las velocidades de los procesadores y las de la memoria a lo largo del tiempo (velocidad en mHz).

La idea del caché es un gran invento que se usa en diversas situaciones de computación. Por ejemplo, el funcionamiento de internet y la web depende de esta idea. En el caso de las computadoras, durante toda la historia de las computadoras electrónicas, la memoria siempre ha sido mucho más lenta que el procesador, por lo que el desempeño total del sistema, sin importar qué tan bueno sea el procesador, depende esencialmente de la memoria. Más aún, esta brecha de velocidades entre el procesador y la memoria ha ido creciendo con el tiempo (véase la gráfica anterior). La existencia de uno o varios niveles de caché amortigua el impacto negativo que tiene la velocidad de la memoria en el desempeño: a mayor número de niveles, más suave será la caída entre la velocidad de un nivel y el siguiente.

El funcionamiento de la memoria caché se basa en lo que se denomina *principio de localidad*. Este principio tiene dos variantes:

1] *Temporal*. Una vez que se utiliza un dato de memoria es altamente probable que este mismo dato y su lugar de almacenamiento sigan siendo utilizados por un tiempo relativamente corto.
2] *Espacial*. Una vez que se utiliza un dato de memoria es altamente probable que los lugares de almacenamiento alrededor de él y los datos allí guardados sean utilizados en un futuro cercano.

Caché en la vida cotidiana

Cuando una persona acude a la biblioteca para llevar a cabo un par de tareas escolares se utiliza el principio de localidad. Supóngase que tiene una tarea de historia: la Revolución industrial. Se dirige a la estantería (memoria principal) y saca tres ejemplares que hablan sobre el tema, todos los libros sobre la Revolución industrial están cerca en la estantería porque los libros se acomodan de acuerdo a un código (colocación) que depende del tema del libro. Toma los libros y los lleva a su mesa (caché), donde usa cada uno de ellos por un tiempo (localidad temporal), luego se da cuenta de que necesita algo que no está en ellos y regresa a la estantería, al lugar de donde sacó los tres primeros, y saca otros dos que están por allí (localidad espacial) porque tratan de lo mismo y los lleva a su mesa.

Luego de un par de horas, las cosas cambian porque la persona necesita también hacer tarea de biología: fotosíntesis, por ejemplo. Se dirige a la estantería, ahora a un lugar completamente diferente, y saca dos libros de plantas, regresa a su mesa y ¡uf!, de pronto se da cuenta de un problema, la mesa está repleta de libros de la Revolución industrial. Los quita, los deja en el lugar de donde el personal los toma para volverlos a colocar en su sitio y pone entonces los de plantas. Ahora su caché posee libros que corresponden a lugares en la estantería lejanos de los anteriores, pero cercanos entre sí, y como es muy pequeño comparado con toda la estantería, tiene que reemplazar los libros que tenía allí por los nuevos.

Esto es lo que ocurre también en los cachés de las computadoras. Una memoria caché de cierto nivel posee siempre un subconjunto, bastante pequeño, de las cosas guardadas en el siguiente nivel más grande de memoria; cuando se requiere de cosas que no están en un caché, hay que traerlas del siguiente nivel (en general de alguno de los siguientes) para ponerlas en éste; si ya no caben habrá que quitar algo de lo que se tiene almacenado actualmente para poder colocar lo nuevo. Después de todo, de acuerdo con el principio de localidad, lo más reciente será usado frecuentemente en el futuro cercano, de manera que más vale tenerlo ahí cuando sea solicitado otra vez.

6.5 SEGUNDO PISO: LENGUAJES DE BAJO NIVEL

El uso de la abstracción en la construcción del hardware de la computadora es muy importante. Por ejemplo, las compuertas lógicas, la jerarquía de memoria, etc. En la parte del software también, como ya se empezó a discutir en el tema sobre abstracción, y se describe con más detalle a continuación.

6.5.1 Lenguaje de máquina

Se ha dicho que una computadora no es otra cosa que un ejecutor de algoritmos, y que éstos consisten en una secuencia de instrucciones. Por otra parte, también se ha comentado que absolutamente todo lo que se almacena en la memoria de una computadora, todos los datos con los que opera, son cadenas de ceros y unos (en realidad, presencias o ausencias de corriente). Se deduce entonces que las instrucciones que debe ejecutar un procesador están representadas por cadenas binarias. En efecto, en el mundo real de las computadoras electrónicas digitales, este lenguaje está hecho de ceros y unos. Las instrucciones que la computadora "entiende" son cadenas de bits. Al lenguaje que puede ejecutar el procesador directamente se le denomina *lenguaje de máquina*.

La primera labor del diseñador de computadoras es determinar el catálogo de cosas que el procesador central sabrá hacer por sí mismo, usando únicamente sus componentes físicos, su hardware. Debe decidir qué instrucciones implementar en hardware, y cuáles dejar para ser programadas en software utilizando las del hardware. Sería ideal implementar muchísimas en hardware, ya que se ejecutarían más rápido, pero éste se complicaría mucho. Conforme se va logrando miniaturizar más y más los circuitos electrónicos se pueden incluir más instrucciones en hardware.

Se puede decidir, por ejemplo, que la computadora sabrá sumar, restar y hacer operaciones como el OR o el AND con operandos de 32 bits de longitud, o también que no sabrá multiplicar. Esto, por supuesto, no significa que la máquina no pueda multiplicar, sólo que para hacerlo deberá ejecutar un algoritmo basado en las operaciones que sí sabe hacer por sí misma, como la suma. Al catálogo de instrucciones que la computadora sabe hacer sin ayuda de un programa se le llama *conjunto de instrucciones*.

Este conjunto de instrucciones se numera, con lo que ahora cada operación diferente tendrá un código asociado que la identifica y que se utilizará después, en lenguaje de máquina, para indicarle al procesador que debe llevar a cabo esa operación. Así, por ejemplo, la suma puede ser la número 23 (0010111 en binario en siete bits) de un catálogo de 128 instrucciones diferentes, de manera que el código 0010111, al ser leído por la unidad de control del procesador, le indica que debe ordenarle a la ALU que realice una suma.

Con lo que se ha visto acerca del procesador y sus registros, se puede tener ya una idea clara de cómo se ven las instrucciones de lenguaje de máquina. Una instrucción de suma (a la que se le había asignado el código de operación 23 o 0010111 en binario) entre dos registros del procesador, por ejemplo, el número 10 (001010 en binario en seis bits) y el 14 (001110 en binario) podría comenzar así: 0010111,001010,001110 (se han puesto comas para separar el código de operación y los números de los registros, pero por supuesto no estarían en la memoria de la máquina). Esta suma genera un resultado que habrá que guardar en algún lado, por ejemplo, en el registro 12 (001100 en binario). Con esto en mente y quitando las comas, la instrucción completa se vería así:

00101110010100011100011100

Curiosidades
En la práctica son comunes dos tipos de conjuntos de instrucciones: CISC y RISC (por sus siglas en inglés), de computadora de conjunto de instrucciones completo y reducido, respectivamente. Para cada uno de estos conjuntos existen diversos procesadores que implementan o entienden dicho conjunto de instrucciones. Hoy en día, muchas arquitecturas de dispositivos tan variados como las computadoras de mano (Palm, PocketPC), las consolas de videojuegos (Nintendo), los reproductores de música, los teléfonos (iPod, iPhone, varios modelos de Nokia, etc.) y, por supuesto, una plétora de computadoras están basadas en RISC.

Cuando el procesador recibe esta cadena, su unidad de control ya sabe que los primeros siete bits son el código que indica qué operación se debe realizar, sabe que los siguientes dos segmentos de seis bits son los nombres de los registros cuyo contenido será usado como operando y que los últimos bits son el nombre del registro en el que se debe guardar el resultado de la operación. Éste es el estilo de cosas que el procesador ejecuta, las instrucciones en el lenguaje nativo de la computadora.

Por supuesto, cada procesador particular tiene su propio conjunto de instrucciones y sus propios códigos de operación, entre otras cosas. Un programa escrito en el lenguaje de un procesador particular no es ejecutable por otro. Por poner un caso concreto, un programa en lenguaje de máquina para un procesador fabricado por una compañía no es entendible para un procesador fabricado por otra; ambas compañías fabrican procesadores, pero con conjuntos de instrucciones y códigos de operación radicalmente diferentes.

Pero a pesar de que un programa en lenguaje de máquina para un procesador no es, en principio, ejecutable por otro, sabemos que ambos procesadores no son más que un par de máquinas de Turing, como se vio en el tema de abstracción. Ambas igualmente capaces, ambas igualmente poderosas en cuanto a lo que saben hacer, ambas esencialmente la misma máquina de Turing universal: una máquina muy general, capaz de simular cualquier máquina de Turing particular, es decir, capaz de ejecutar cualquier algoritmo.

6.5.2 Ejecución con cauce segmentado

La ejecución de cada instrucción de lenguaje de máquina conlleva casi siempre varias tareas menores:

1] Traer la instrucción de la memoria. A esto en la jerga de los arquitectos de computadoras se le denomina *fetch*, por la palabra en inglés equivalente a *traer*.
2] Determinar los operandos y disponerlos como entrada a la unidad aritmético-lógica. Esta etapa recibe el nombre de *decodificación*.
3] Ejecutar la operación. Lo que suele llamarse *ejecución* propiamente dicha.
4] Alterar el estado del procesador de acuerdo con el resultado obtenido, en particular escribirlo en alguno de los registros. A lo que, nuevamente en la jerga de los arquitectos, se le denomina *write-back* y que en esencia significa *escribir resultados*.

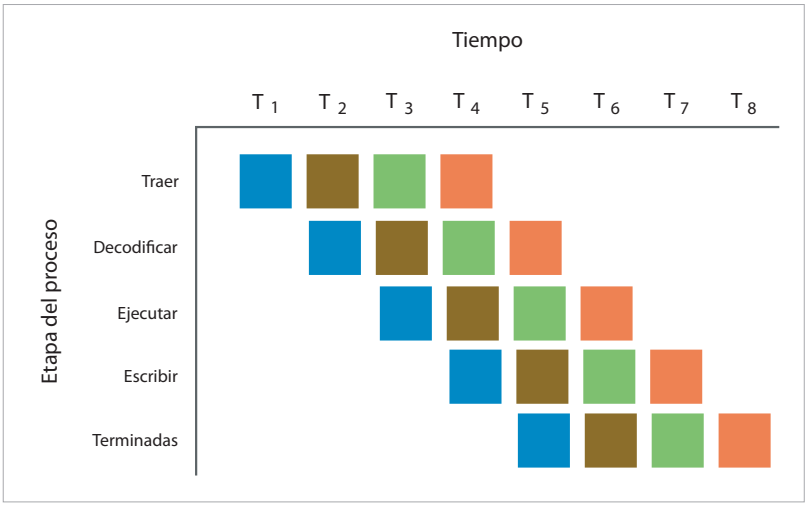
A partir del decenio de los ochenta del siglo pasado, los diseñadores de procesadores se dieron cuenta de que, si lograban hacer que todas las etapas de todas las instrucciones tardaran el mismo tiempo, podían lograr que la ejecución de programas en el procesador fuera mucho más eficiente. El truco es usar la misma estrategia que usó Henry Ford para producir en serie el famoso Modelo T, el primer automóvil hecho en una línea de ensamblaje. Ford dividió el proceso de elaboración de un Modelo T en varias etapas; normalmente todas las etapas serían ejecutadas por un equipo de obreros sobre un automóvil, una tras otra hasta terminarlo y sólo entonces podrían comenzar con uno nuevo. En este esquema, si cada etapa tarda, digamos, 12 minutos, y hay un total de cinco etapas, entonces cada automóvil tardará una hora en ser armado; un observador colocado fuera de la planta de producción vería entonces salir un automóvil por hora.

Ahora cámbiese el esquema, en vez de que todos los obreros estén haciéndose bolas sobre un solo coche a la vez, hay que ponerlos a lo largo de una línea junto a una banda móvil. Sobre la banda se coloca un chasis y se mueve hasta el primer obrero que se encarga *solamente* de llevar a cabo las tareas de la primera etapa, luego de 12 minutos se le pone un

nuevo chasis en la banda y se desplaza un lugar hacia adelante. Ahora, el primer obrero recibe un nuevo chasis sobre el que trabaja y el segundo obrero de la fila puede hacer las labores propias de la segunda etapa sobre el automóvil que acaba de ser trabajado por el primer obrero. Luego de 12 minutos, se repite la operación hasta que al cabo de una hora ya todos los obreros están trabajando: cada uno haciendo sólo una de las etapas y nada más, cada uno sobre un automóvil diferente, pero todos al mismo tiempo. Por supuesto, el tiempo de elaboración de cada automóvil no cambia, sigue siendo de una hora, pero ahora el observador externo ve salir un automóvil completo cada 12 minutos y no cada hora como antes.

Cada una de las etapas listadas arriba para la ejecución de una instrucción tarda una unidad de tiempo pequeña (en las computadoras actuales es realmente pequeña, posiblemente una fracción de nanosegundo) y, por cierto, es justo el tamaño de esa unidad lo que determina la duración de los mencionados ciclos de reloj. Si todas tardan lo mismo se podría pensar en hacer lo análogo a una línea de ensamble: un trozo del procesador trae la *n*-ésima instrucción del programa, al mismo tiempo otro trozo del procesador hace la decodificación de la instrucción inmediata anterior, la $n - 1$; al mismo tiempo, otro trozo del procesador ejecuta en la ALU la operación de la instrucción $n - 2$; simultáneamente, otro fragmento más realiza la escritura de resultados de la instrucción $n - 3$. ¡Brillante! ¿No?

A esta estrategia se le denomina *ejecución de cauce segmentado*. Suena complicado. El término en inglés es *pipeline*, que literalmente significa tubería, pero no dice mucho, así que es mejor usar el complicado término en español en vez de la traducción literal.



En la figura 15, cada instrucción se representa con un color diferente. A partir de T_1, en cada paso temporal (ciclo de reloj) se alimenta una nueva instrucción al procesador. A partir de T_5, se termina con una instrucción en cada paso temporal.

La ejecución de cauce segmentado parece más fácil de lo que es en realidad, pues hay algunos problemas no triviales que resolver para implementarla. Por ejemplo: ¿qué pasa

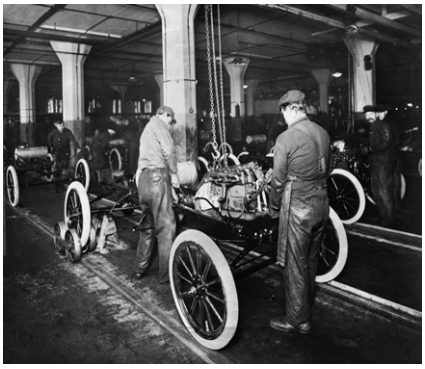


Figura 15. Ejecución de cauce segmentado | © Latin Stock México.

Curiosidades

En 1780, un estadunidense de nombre Oliver Evans dividió el proceso de elaboración de harina de trigo en etapas y luego hizo que cada uno de los empleados del molino se dedicara a una y sólo una de las etapas de producción de harina. Éste es el primer ejemplo de lo que se conoce como línea de producción. Más tarde, en 1913, Henry Ford implantó un sistema similar para producir en serie su famoso Modelo T. En la línea de ensamble de Ford había estaciones, en cada una de las cuales se llevaba a cabo una labor particular: en una se colocaba el motor, en otra las puertas, en otra el radiador, etc. Por un lado de la línea entraba sólo el chasis de los automóviles, y a lo largo de ella se iba completando paulatinamente hasta que al final de la línea salía completo. Éste es precisamente el principio del concepto de cauce segmentado o *pipeline* en inglés.

si la instrucción $n-1$, en su fase de decodificación, necesita un dato almacenado en un registro que aún no ha sido escrito en él por la instrucción $n-3$? Esto es perfectamente plausible, porque después de todo la instrucción $n-3$ precede a la $n-1$, y ésta puede basar su operación en los datos calculados por aquélla. A un problema como éste se le denomina *conflicto de datos* (en inglés *data hazard*) y su solución eficiente excede con mucho el alcance de este libro. La solución fácil es detener toda la "línea de producción" salvo aquellas etapas que están calculando lo que se necesita para continuar; esto introduce intervalos de espera indeseables en los que el procesador no está ocupado al cien por ciento.

Los procesadores actuales llevan el concepto de cauce segmentado aún más allá. Actualmente poseen no sólo una, sino varias "líneas de ensamble", varios cauces dedicados a tareas específicas: uno para operaciones aritmético-lógicas con números enteros, otro para instrucciones de acceso a memoria y otro para operaciones que tienen que ver con el manejo de gráficos en la pantalla, por ejemplo. Así, cada instrucción se coloca en el cauce que le corresponde una vez que se determina de qué tipo es. A esto se le denomina ejecución superescalar. Prácticamente todos los procesadores de este siglo son superescalares.

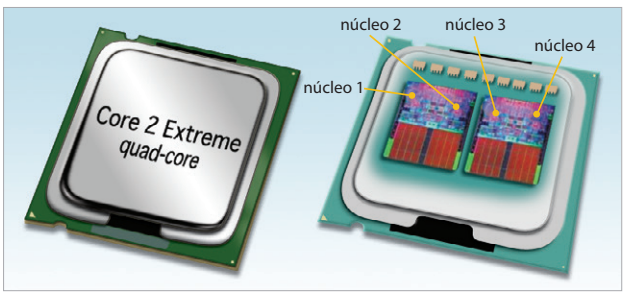
Curiosidades
El hecho mencionado, de que ya es prácticamente imposible continuar con un crecimiento de densidad como el establecido por la ley de Moore, traería como consecuencia un estancamiento en la velocidad de los procesadores. Para paliar, no resolver, el problema, desde hace unos años se diseñan arquitecturas de procesadores en los que realmente hay más de una unidad de procesamiento en cada chip. A esto se le ha llamado arquitectura multinúcleo (*multicore* en inglés). El desempeño de una de estas arquitecturas no es comparable al que se obtendría incrementando la densidad al ritmo señalado por la ley de Moore, pero al menos permite no dejarlo igual. En una arquitectura multinúcleo, cada núcleo se encarga de su propia secuencia de instrucciones, se podría decir, sin ser del todo precisos, que se encarga de su propio subprograma, a lo que se le suele llamar hilo de ejecución. Por lo que realmente es una pequeña computadora en la que se ejecutan en paralelo (simultáneamente) varios hilos de ejecución.

6.5.3 Lenguaje ensamblador

Por supuesto, sería insufrible programar en lenguaje de máquina, habría que recordar todos esos códigos de operación y varias cosas más que no vienen al caso, además de que resultaría muy fácil cometer errores: un solo bit equivocado y el significado puede ser radicalmente distinto del que se pretendía. Por eso se crearon, casi desde el inicio de la computación electrónica, los lenguajes ensambladores. El lenguaje ensamblador, al igual que el de máquina, es particular de cada procesador, de hecho es una simple transcripción de éste. En vez de poner 0010111, se utiliza una palabra mnemotécnica que indique que el código corresponde a una suma, por ejemplo `suma`, en vez de usar 001010, se pone explícitamente `R10`, se reemplaza 001110 por `R14` y finalmente se pone `R12` en vez de 001100, con lo que nuestra instrucción anterior 0010111 001010 001110 001100 quedaría expresada como `suma R10, R14, R12`. Aunque ya casi no se utiliza en la actualidad, la programación en lenguaje ensamblador fue muy usual en los años sesenta y setenta del siglo pasado, cuando era indispensable hacer programas breves y la única manera de lograrlo era programando en lenguaje ensamblador.

Por supuesto, la computadora debe recibir sólo lenguaje de máquina a la hora de ejecutar un programa, así que un programa escrito en lenguaje ensamblador debe ser traducido a lenguaje de máquina. Una labor más o menos sencilla, dado que hay una equivalencia uno a uno entre las instrucciones en uno y en otro. Esta tarea de traducción la lleva a cabo un programa especial llamado, por cierto, *ensamblador*.

Dado que todas las computadoras son esencialmente iguales, y que programar en lenguaje de máquina o en ensambladores es una labor altamente especializada y difícil, se han inventado lenguajes con los que es mucho más fácil programar olvidándose de las peculiaridades de cada máquina. En el tema sobre programación se tuvo contacto con uno de ellos, pero hay literalmente cientos de lenguajes de programación



diferentes. A estos lenguajes se les denomina *lenguajes de alto nivel*. El término se refiere al hecho de que cada instrucción en uno de estos lenguajes equivale, generalmente, a varias instrucciones de lenguaje de máquina. Por supuesto, dado que la equivalencia es uno a uno entre el lenguaje ensamblador y el de máquina, ambos son de bajo nivel.

6.6 TERCER PISO: SISTEMAS OPERATIVOS

—Úrsula, córrele porque si no, no nos va a dar tiempo.
—Estoy lista, vamos al súper, mamá.
—¿Traes la lista de lo que vamos a comprar?
—Sí, aquí la tengo.

Así, Úrsula y su madre se dirigen al supermercado para comprar todo lo necesario para preparar la cena tradicional anual. En esta ocasión habrá más invitados que de costumbre porque vienen algunos familiares de Guadalajara y es la primera vez que Arcadio la pasará con ellos.

Después de recorrer varias veces la tienda, adquieren todo lo necesario y regresan a casa, donde la mamá de Úrsula inicia los complejos rituales de preparación, en *pipeline*, de la ensalada, el pastel de carne y los higos con relleno de queso con naranja, que son el menú para la cena. En esta ocasión, Úrsula sirve como ayudante de cocina y no deja de preguntar cada pequeño detalle sobre la preparación: ¿cómo sabes cuánta ensalada preparar?, ¿será suficiente el pastel de carne para 15 personas?, ¿qué utensilios utilizaremos para servir?, ¿servimos los platos en la cocina?, ¿cómo sabes cuánto quiere cada persona?

6.6.1 Manejo de procesos: planificadores

Éstas son dudas válidas para una novata en la cocina, en *administrar* un hogar o un restaurante. Con las capacidades tan grandes que tienen los sistemas de cómputo modernos surgen el mismo tipo de complicaciones: ¿se puede ejecutar más de un programa a la vez?, ¿cuántos recursos (entendiéndolos como la cantidad de memoria, el tiempo del procesador que se puede utilizar, etc.) pueden asignarse a una aplicación particular? Es claro que se requiere de un *administrador* para la computadora.

Dado que el administrador de la computadora debe funcionar eficientemente, tiene que ser un programa que controle y ordene el acceso a los recursos, que ofrezca mecanismos para que los otros programas puedan solicitar y acceder a esos recursos y, aunque no es tan obvio, también debe proveer seguridad y privacidad para los programas que se ejecutan en el sistema. Este programa administrador recibe el nombre de *sistema operativo*.

En un sistema de cómputo en cualquier casa u oficina moderna, una persona o varias pueden estar ejecutando de manera simultánea distintas aplicaciones: un navegador de web, una hoja de cálculo, un editor de textos, compiladores y ambientes de desarrollo, etc. Como ya se ha visto, el avance de este tipo de computadoras hasta hace un par de años estaba dominado por sistemas con un solo procesador. Entonces, ¿cómo pueden estarse ejecutando de manera simultánea tantos procesos?

La manera de resolver esto en los sistemas operativos modernos es a través de una pequeña porción del sistema operativo, uno de los muchos programas que *conforman* el sistema operativo, que se conoce como proceso planificador de tareas (o *scheduler*, en inglés) y cuya tarea consiste en administrar la interacción de todos los procesos en el

Curiosidades
Los primeros sistemas operativos que se desarrollaron para equipos *mainframe* y microcomputadoras, tenían un planificador muy primitivo y soportaban la ejecución de un solo proceso. El programa entonces tenía el control total del equipo. La habilidad multitarea se introdujo en los *mainframes* en los años sesenta y en las microcomputadoras en los ochenta, pero las técnicas se consolidaron y utilizaron ampliamente varios años después.

equipo: cuántos hay en ejecución, qué recursos tienen, exactamente en qué línea de su programa va el procesador, si están o no esperando enviar o recibir información de los dispositivos de entrada y salida, cómo almacenar información en un archivo o recibir lo que escribió el usuario por medio del teclado.

Tómese como ejemplo la sección de carnes frías de un supermercado típico: cuando llega una persona, debe tomar una ficha con un número, y cuando uno de los dependientes se desocupa y activa un control, se enciende el número que indica qué cliente será atendido a continuación. Cuando ve aparecer su número en la pantalla, el cliente dice: "Sigo yo", y entonces el dependiente le pregunta qué desea. En esta situación, el dependiente junto con el esquema de números, controles y el tablero, funcionan como un planificador para la sección de carnes frías. ¿Qué recursos administra *este planificador*? Antes que nada, la atención del dependiente, que no podría atender todas las solicitudes de las compradoras al mismo tiempo; después la báscula para pesar y cotizar los productos, las carnes frías, etcétera.

En la práctica, las computadoras con esos millones de operaciones que pueden realizar por segundo dan la impresión de que hacen *muchas cosas a la vez*, aunque en realidad funcionan de una manera similar al ejemplo de las carnes frías: le asignan un tiempo determinado a un proceso —por ejemplo, al editor de textos—, y después de ese tiempo o cuando el procesador indique que terminó o que debe esperar a que el usuario escriba o haga algo, el planificador "saca" de ejecución al proceso, guarda todos los datos que están en los registros y que corresponden al editor, los almacena en un lugar seguro y va por otro proceso, lo instala en los lugares necesarios e inicia su ejecución, otra vez por un tiempo determinado. ¿Cuánto tiempo está un proceso en ejecución? Depende del sistema operativo particular, el proceso y muchos otros factores, pero usualmente se mantiene entre unas decenas de milisegundos y unos cuantos segundos, tiempo suficiente para ejecutar desde varias centenas de operaciones hasta varios millones del programa en ejecución. Esto se conoce como *multitarea* y prácticamente todos los sistemas operativos actuales lo utilizan.

6.6.2 Manejo de memoria

El manejo de procesos es una de las características más importantes y visibles en un sistema operativo actual, es el control y enlace con los usuarios del sistema. Sin embargo, como ya se ha visto por las diferencias abismales entre las velocidades del procesador y la jerarquía de memoria, uno de los principales retos para el sistema es administrar efectivamente la jerarquía de memoria. Para todo fin práctico, para una persona que utiliza un programa como el editor de textos o para el programador de ese editor no es relevante si un dato que requiere el programa o el usuario está en el disco, en la memoria RAM o en algún otro lugar de la jerarquía de memoria. Pero para el sistema operativo tal información sí es relevante, ya que puede impactar de manera importante en el rendimiento total del sistema.

Otro factor importante en los sistemas operativos modernos es el vínculo entre el administrador de disco y la jerarquía de memoria, ya que el último nivel de ésta generalmente reside en el disco duro. Esta actividad tiene que ver con lo que se conoce como memoria virtual, que provee al programador una abstracción sencilla de la memoria, ya que ofrece a los procesos la *visión* de que tienen a su disposición mucha más memoria de la que físicamente está instalada en la computadora, a través de un concepto relativamente sencillo: partir la memoria que requiere el programa y sus datos en páginas, que son de un tamaño

fijo, usualmente cuatro KB o 16 KB. Así, sólo una fracción de las páginas del proceso están presentes en la memoria física, las demás se mantienen en el siguiente nivel de la jerarquía, usualmente el disco duro de la computadora.

Cuando un dato o segmento de código es solicitado por el usuario, el sistema operativo suspende temporalmente su ejecución y va al disco por las páginas de memoria requeridas, mientras el planificador selecciona otro proceso para aprovechar el tiempo. ¿Cuántos procesos puede ejecutar un sistema operativo dando esta *ilusión* de que se ejecutan de manera simultánea? En general, depende del tamaño de los procesos y el número de éstos que el sistema pueda manejar en la memoria sin tener que ir *demasiadas* veces al disco duro. En todo caso, se tiene un ejemplo más de abstracción. Al programador se le presenta un modelo de una computadora para él solo y la implementación permite la ejecución de varias tareas.

6.6.3 Sistemas de archivos (otra abstracción)

Otro aspecto importante de los sistemas operativos es ofrecer soporte para almacenar la información de los procesos, así como los procesos mismos, mediante distintos sistemas de archivos. El concepto de archivo es una abstracción común a todos los sistemas operativos, aunque en la práctica su implantación física varía considerablemente de un sistema de archivos al siguiente. ¿Qué es un archivo? En efecto, el nombre viene de los archiveros de papel, donde un archivo sirve para almacenar un conjunto de información arbitraria. Sólo que en un sistema de cómputo, en lugar de archiveros, se utiliza algún mecanismo de almacenamiento durable, como un disco duro, un disco compacto o una unidad de cinta magnética. El modelo para organizar su información en archivos, contenidos unos dentro de otros, le presenta al programador un esquema que abstrae detalles de cómo y dónde exactamente se almacenan sus datos.

Entonces, el sistema de archivos tiene la enorme responsabilidad de administrar y manejar los archivos de todos los usuarios y programas del sistema. Es usual que los sistemas de archivos estén organizados por medio de estructuras de datos similares a un árbol, es decir, de manera jerárquica, donde se tienen archivos y directorios. Los directorios funcionan como los archiveros físicos, adentro pueden contener muchos archivos e incluso otros directorios. Con esto se pueden almacenar cientos de millones de archivos distintos o archivos enormes, por ejemplo, toda la información del censo de población del país.

Los sistemas de archivos hacen muchas cosas más, como preocuparse por la seguridad del sistema para que un usuario no pueda ver o modificar los archivos de otro usuario del sistema. Algunos incluyen además soporte para fallas graves, como caídas del sistema, y recuperan de manera automática todos los cambios no salvados antes de la falla. Otros son soportados de manera universal por distintos sistemas operativos y sirven para almacenar los archivos en un dispositivo de respaldo externo, como una cinta o CD-ROM.

6.6.4 Interfaz de texto o gráfica

Hasta arriba de la jerarquía de abstracciones sobre las cuales está implementada una computadora, se encuentran las de la interfaz con el usuario. Ésta debe proveer modelos, abstracciones lo más cercanas posibles a las vivencias del usuario, a su sentido común, a fin de despreocuparlo, en la medida de lo posible, de las peculiaridades de la construcción de la computadora.

"Quiero medio kilo de jamón de pavo y un cuarto de queso", es una instrucción estándar que el dependiente de la sección de carnes frías puede entender y realizar, pero ¿cómo interactuar con el sistema operativo y darle instrucciones? A lo largo de la historia de las computadoras se ha pasado desde la complicada interacción de cambiar los cables y reconfigurarla hasta las elaboradas y atractivas interfaces gráficas de usuario (GUI, por sus siglas en inglés), que ofrecen un ambiente integrado virtual: con un escritorio o lugar para instalar y encontrar cosas fácilmente (archivos, programas, herramientas para administrar el sistema, etc.), menúes para organizar aplicaciones, conexiones de red y, por supuesto, una gama de programas que, en algunos casos, se consideran parte del sistema operativo y en otros parte de la GUI.

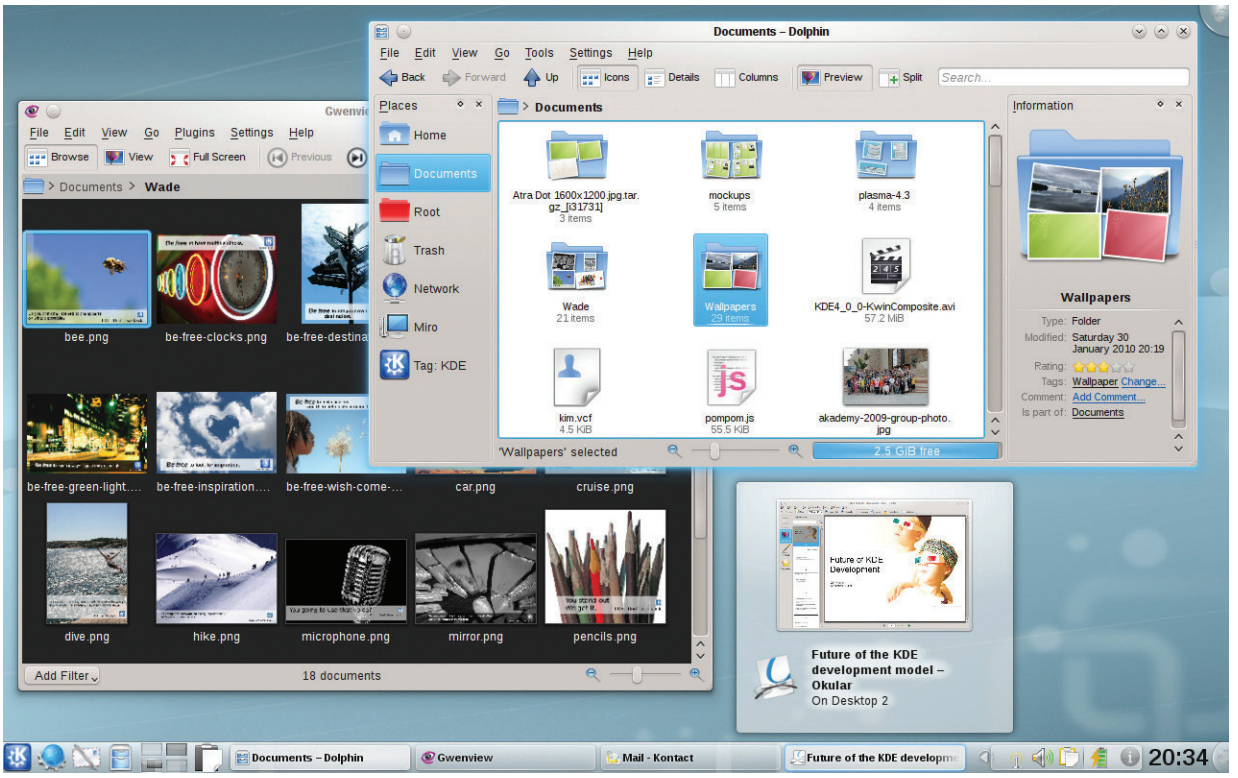


Figura 16. Ejemplo de escritorio, utilizando el ambiente manejador de ventanas X con el escritorio KDE en un sistema operativo Linux | © KDE.org.

Si la GUI es parte integral del sistema operativo, obliga al usuario del sistema a trabajar exclusivamente con esta interfaz, mientras que en otros sistemas se utiliza una mezcla de aplicaciones que dan mayores libertades y permiten seleccionar entre varias GUI alternativas. Por ejemplo, en la imagen de arriba se aprecia el ambiente de escritorio KDE, corriendo sobre el manejador de ventanas *X*, usualmente conocido como X11, y que es la base de la mayoría de las GUI para los sistemas operativos basados en Unix.

Antes de las GUI, los sistemas operativos utilizaban una interfaz de texto para ingresar comandos. En su forma más usual, se conocen como terminales, que lucen como la ventana principal de un editor de texto, pero hay un *prompt* que indica al usuario que el sistema operativo está listo para recibir (y ejecutar) un comando. Hoy, aunque no son indispensables, la mayoría de los sistemas operativos ofrecen un programa que se llama terminal, símbolo de sistema o algo similar, y que presenta una ventana, dentro del GUI, donde se pueden ingresar comandos.

6.7 PENTHOUSE: SOFTWARE DE APLICACIÓN

Hasta el momento se ha explicado cómo funciona, y cómo, mediante instrucciones, programas, lenguajes de bajo y alto nivel, un usuario puede comunicarle sus instrucciones a una computadora. Sin embargo, para la mayoría de las personas, ésas no son las maneras estándar de operar un sistema de cómputo. En realidad, la mayoría de los usuarios utilizan la computadora para acceder a otras herramientas especiales, como editores de texto, un navegador para visitar distintos sitios web o leer y enviar correos electrónicos.

Estas herramientas son usualmente piezas complicadas de software escritas con un propósito particular en uno o varios lenguajes de programación, generalmente de alto nivel, y que ofrecen mecanismos de operación avanzados, probablemente a través de iconos que se pueden presionar con el puntero (la flechita) por medio del ratón de la computadora, o a través de la voz o el teclado. Este software es comúnmente llamado *software de aplicación*, aplicación o simplemente software.

Existen, literalmente, miles de aplicaciones para hacer todo tipo de cosas, por lo que no es sensato pensar en mostrar en un solo libro —o en cien para el caso— qué hacen o cómo se utilizan. Lo que sí se desea dejar en claro al lector es el mensaje de que, detrás de las interfaces y formas de interactuar con una computadora, con un sistema de cómputo, con una aplicación, están todos los conocimientos que se han mostrado en este libro y algunos más que, por razones de espacio y por las propias limitaciones de los autores, no fue posible incluir, y deben ser guías en este viaje que implica la tecnología y la computación. En el último tema, sobre aplicaciones, no se muestra una herramienta particular o un programa, sino algunos ejemplos donde la utilización de la computación ha marcado una diferencia fundamental, donde alcanzar los mismos resultados sin la ayuda, sin el músculo de cálculo que aporta la computación, sería imposible.

6.8 RESUMEN

Los dispositivos de cómputo, que han permeado en casi todos los ámbitos de la vida cotidiana, constituyen uno de los avances tecnológicos más sorprendentes de la humanidad. En este capítulo se analizaron los sistemas de cómputo, se presentaron los fundamentos de su operación y también se mostró que su utilidad proviene del indisoluble vínculo entre el hardware y el software. Se ha logrado construir dispositivos de cómputo enormemente complejos debido al uso de la abstracción.

TEMA

7

La función primordial de la web no es reemplazar átomos por bits para que podamos, por ejemplo, comprar en línea. La web ni siquiera está para darle poder a grupos, como los consumidores. En realidad, la web está cambiando nuestro entendimiento de cómo desde un principio se conectan las cosas. Y lo más importante, la web está pegando no sólo páginas sino a nosotros mismos, los seres humanos, de nuevas maneras. Nos estamos conectando de nuevas formas que aún estamos inventando.

DAVID WEINBERGER, 2002.

Un sistema distribuido es aquel en el que la falla de una computadora, que tú ni siquiera sabías que existía, puede inutilizar la tuya.

LESLIE LAMPORT, 1987.

© Chris Harrison, Carnegie Mellon University, USA.

7.1 CÓMPUTO DISTRIBUIDO

7.1.1 Introducción

Desde la popularización de la computadora personal en el decenio de los ochenta del siglo pasado, la humanidad ha presenciado enormes avances en el poder de cómputo de estos sistemas, principalmente debido a la creciente capacidad de integración en los microprocesadores: de unos miles de transistores en un chip a miles de millones en el mismo espacio. El desarrollo de este tipo de computadoras las ha llevado a ámbitos en los que antes eran impensables: casas, escuelas y oficinas.

En los últimos años, sin embargo, los principales fabricantes de microprocesadores han dejado de incrementar su velocidad debido al sobrecalentamiento en los chips. En su lugar, la nueva tendencia es agregar varios núcleos o procesadores en el mismo chip, que se comunican entre sí por medio de memorias electrónicas de gran velocidad. Esto agrega poderío a las computadoras, porque se explota el paralelismo: lograr que varios procesadores trabajen en la misma tarea. Estos sistemas son conocidos como *multicore*.

Los retos para programar y aprovechar el paralelismo de estos nuevos sistemas personales, *multicore*, así como el de grandes cúmulos de computadoras, llamados *grids* o clústers y, por supuesto, de las supercomputadoras, son muy variados e involucran una serie de conocimientos y principios que son la base del cómputo distribuido. Éste es, sin lugar a dudas, uno de los retos más interesantes de la computación moderna.

¿Qué tipo de problemas se presentan en el cómputo distribuido? Por ejemplo, los procesadores dentro de un mismo chip tienen que coordinar su acceso a una posición de memoria compartida; en un cúmulo de servidores el reto está en balancear la carga para ofrecer un mejor servicio, etc. En los siguientes apartados veremos estos temas y sus principios básicos, con la ayuda de situaciones cotidianas.

7.1.2 Exclusión mutua

Arcadio y Pilar son vecinos y comparten un patio. Arcadio tiene un perro y Pilar un gato. A ambos animales les gusta correr en el patio pero, por supuesto, no se toleran mutuamente. Después de varios eventos desafortunados, los dueños deciden que tienen que coordinarse para que sus mascotas nunca coincidan en el patio. Obviamente, quedan descartadas las opciones que impidan que una mascota pueda salir al patio cuando éste está vacío. ¿Qué deben hacer entonces? Tienen que ponerse de acuerdo en cómo decidir si dejan o no salir a sus mascotas al patio. Este tipo de procedimientos, para ponerse de acuerdo, son lo que llamamos un *protocolo*, de lo cual se hablara más adelante.

El patio es demasiado grande, por lo que Pilar o Arcadio no pueden simplemente asomarse y verificar si está libre o no. Si Arcadio quiere sacar al perro, podría caminar hasta la casa de Pilar y preguntarle, pero esto toma mucho tiempo y tal vez esté lloviendo. ¿Qué tal si le llama por teléfono? Esto podría no funcionar, porque tal vez Pilar se está bañando o su teléfono no tiene batería y no lo escucha.

Después de varios intentos de coordinación fallidos, se les ocurre lo siguiente: instalan dos postes fuera de sus respectivas casas e instalan en cada uno una bandera. Cuando Pilar quiere sacar a su gato hace lo siguiente:

1] Eleva su bandera en el mástil.
2] Cuando la bandera de Arcadio está abajo, saca al gato.
3] Cuando el gato regresa a casa, baja su bandera.

El comportamiento de Arcadio es un poco más complicado:

1] Eleva su bandera en el mástil.
2] Mientras la bandera de Pilar está arriba:
   a] Arcadio baja su bandera.
   b] Arcadio espera hasta que Pilar baje la bandera.
   c] Arcadio eleva su bandera.
3] Tan pronto como su bandera está arriba y la de ella abajo, saca al perro.
4] Cuando su perro regresa, baja la bandera.

Esta solución funciona porque de manera intuitiva cada uno eleva su propia bandera y luego observa la bandera del otro, por lo que al menos uno de los dos verá la bandera del otro arriba y no sacará su mascota al patio.

7.1.3 Propiedades de la exclusión mutua

Con el protocolo en uso ha sido posible que las mascotas no estén en el patio al mismo tiempo; esta propiedad se conoce como *exclusión mutua*.

La exclusión mutua es muy importante, pero no es la única propiedad que interesa. También se quiere maximizar la utilización del patio, es decir, si una mascota quiere salir al patio, eventualmente debe lograrlo y, si los dos quieren salir, uno de los dos debería poder hacerlo. Esta propiedad se conoce como *libre de abrazo mortal* (*deadlock* en inglés). Sería terrible que por un error en el protocolo ninguna de las dos mascotas pudiera salir y el patio estuviera vacío.

Otra propiedad de interés es prevenir la hambruna (*starvation* en inglés); es decir, que una de las mascotas quiera entrar al patio pero tenga que esperar indefinidamente. Con respecto a esta propiedad, el protocolo mencionado arriba se comporta de manera injusta, porque si ambas banderas suben, entonces Arcadio y su perro deben esperar. También podría suceder que, mientras espera, Arcadio se distrajera viendo la televisión o leyendo y Pilar bajara la bandera sin que él lo note; y que el gato quisiera volver a salir y Pilar volviera a subir la bandera, de manera que éste salga dos veces al patio y el perro ninguna. Esto podría suceder más veces.

Otra propiedad de interés es la espera. Si inmediatamente después de subir la bandera, Pilar saliera de la ciudad de emergencia porque su madre está enferma y no regresara en una semana, Arcadio vería siempre la bandera arriba y, por lo tanto, no dejaría salir a su perro en una semana. En la práctica, todos los protocolos para lograr exclusión mutua involucran espera, aunque algunos la manejan mejor que Arcadio y Pilar y, en esos casos, se conoce como espera acotada.

Curiosidades
SMS (siglas en inglés de "servicio de mensajes cortos") es un protocolo de comunicación que permite el intercambio de mensajes entre teléfonos celulares. La tecnología SMS ha generado una revolución en las comunicaciones e impulsó el desarrollo de los mensajes de texto a tal grado que, en buena parte del mundo, se utiliza para indicar que se está enviando un mensaje de texto, aun cuando la tecnología de transporte es distinta (por ejemplo, MMS, siglas en inglés de otro protocolo muy común hoy día: el "servicio de mensajes multimedia").

7.2 COMUNICACIÓN

7.2.1 Otro problema de la vida real

Úrsula toma su teléfono celular y escribe un mensaje de texto o SMS, como coloquialmente se le conoce, dirigido a Arcadio: "Nos vemos a la entrada del cine a las 6 pm". Antes de enviar el mensaje, Úrsula se pregunta si llegará con éxito. Pueden ocurrir varias cosas, como que Arcadio no reciba el mensaje por una falla en la red celular o con su teléfono. Entonces, Úrsula se pregunta cómo puede resolver esto, y al final de su mensaje añade: "confirma por favor".

¿Es suficiente para quedar de acuerdo? Pues no, porque cuando Arcadio recibe el mensaje contesta con otro diciendo "De acuerdo, a las 6", pero entonces, igual que Úrsula, también se pregunta qué pasaría si no le llega, así que decide jugar seguro y extiende el mensaje a "De acuerdo, a las 6. Por favor confirma mi confirmación". Y cuando Úrsula recibe la confirmación, ella sabe que Arcadio sabe, pero Arcadio no sabe que Úrsula también sabe.

Así, la cosa se hace interminable. Cada uno, Úrsula y Arcadio, ven posibles mundos distintos. Al inicio, Úrsula ve dos posibilidades, que Arcadio lea su SMS o que no lo haga. Después de la primera confirmación, todo se aclara y sólo hay un mundo posible: Arcadio leyó su SMS. Sin embargo, ahora para Arcadio hay dos mundos: Úrsula recibió su SMS o no.

Con este tipo de comunicación resulta imposible llegar a un acuerdo.

7.2.2 Un mismo lenguaje

¿Cómo comunicarse con otra persona? Existen muy diversas formas, una de las más simples es mediante el habla o las señas. Al que habla o hace las señas se le denomina *emisor* y al que escucha o ve las señas, *receptor*. ¿Qué sucede cuando no se entiende el idioma del que está hablando? Es posible escucharlo, pero no comprender lo que dice. ¿Qué sucede, por ejemplo, cuando se habla por teléfono y no se escuchan bien algunas palabras de la conversación? Si se escucha parte de la frase pero no se entiende, se puede pedir que repitan la palabra o el mensaje; pero si no se oye nada por algún problema en la línea, es posible entender otra cosa o captar la idea correcta del mensaje por el contexto.

Todas estas preguntas son válidas para otros tipos de comunicación, como la que ocurre con un aparato receptor al captar las señales de radio o entre distintas computadoras conectadas a través de una red. Ambos extremos de la comunicación tienen que *hablar el mismo lenguaje*. Además, para que la comunicación sea exitosa, las partes deben ponerse de acuerdo en la forma en que ésta se llevará a cabo. Por ejemplo, considérese el siguiente fragmento de comunicación:

—Sí, ¿bueno?

—Hola. Úrsula, ¿eres tú? Habla Arcadio.

—¡Arcadio, qué gusto! ¿Cómo estás?

—Bien, gracias. ¿Te llegó mi regalo?

—Sí, me llegó, muchas gracias, está increíble... Pasó algo chistosísimo, cuando…

—Qué bueno que te gustó, no estaba... perdón, continúa.

—Ah, gracias, es que cuando llegó el regalo que me enviaste, Carlos me avisó, pero yo entendí que mi mamá había hecho paella y no "hay un regalo para ella".

—Ja ja, sí, qué chistoso.

7.2.3 Protocolo

Aun sin etiquetar las líneas con la persona que las dice, es fácil determinar quién dijo qué, ¿por qué sucede esto? En este ejemplo particular, porque existe la costumbre, al menos en México, de iniciar las conversaciones telefónicas con un saludo y después realizar intervenciones intercaladas, usualmente en forma de preguntas y respuestas, entre las personas al teléfono. En computación se denomina *protocolo* a esta manera de regular la comunicación.

Por supuesto, muchos protocolos en la comunicación humana varían de una ciudad o familia a otra, y se aprenden a través de la experiencia y la repetición. El protocolo se convierte en un acuerdo de comportamiento y es común, además, que especifique qué hacer en situaciones extraordinarias, por ejemplo, cuando se cruzan las líneas o no se escucha cierta parte de la conversación. A continuación se detallan algunas de las partes del protocolo inmerso en una conversación telefónica típica:

1] Al inicio cuando suena el teléfono y Úrsula descuelga la bocina, lo que se llama *contestar* o *atender* la llamada. "Hola", "bueno", "diga", "residencia de la familia Gómez", "departamento de compras, buenos días", son todas formas típicas de iniciar una llamada.
2] Después del saludo, Úrsula hace una pausa que le indica al interlocutor que le *corresponde* hacer su intervención. ¿Cuánto debe durar esta pausa? Usualmente un par de

segundos. ¿Qué contesta Arcadio en el ejemplo? "Hola. Úrsula, ¿eres tú? Habla Arcadio", es decir, contesta al saludo de manera similar y se identifica diciendo su nombre. Acto seguido hace una pausa, otra vez es turno de Úrsula.

3] En algún momento de este ir y venir de intervenciones, tienen que ocurrir dos cosas: primero, tocar el tema central y el motivo de la llamada; y segundo, decidir que es tiempo de concluirla.

3.1] ¿Es indispensable que exista un tema central? Siempre existe un tema. En conversaciones familiares o de amigos, puede ser tan sencillo como simplemente saludar y mantenerse en contacto. En conversaciones de negocios o profesionales, es típico aclarar que inicia el tema central con frases como "además de saludarle, el motivo de esta llamada es avisarle que..." o "¿recibió nuestra propuesta...?", etcétera.

3.2] Una vez que se han satisfecho los requerimientos de información mediante la discusión del o de los temas centrales, se espera que una de las partes avise que la conversación está por concluir y dé oportunidad a que la otra parte se dé por enterada y se despida. Por ejemplo, con algo como "bien, me dio gusto hablar contigo" o "tengo que correr, porque ya voy tarde a una cita".

La interacción que ocurre en una conversación telefónica es un buen ejemplo de *causalidad*, que representa la relación entre un evento (llamado causa, en el ejemplo "decir algo y luego hacer una pausa") y otro evento llamado *efecto*, que es una consecuencia del primero. La causalidad es un tema filosófico profundo y supone lo que puede parecer un detalle obvio: la causa debe ocurrir antes que el efecto. En el ejemplo, uno debe escuchar el saludo del interlocutor y esperar una pausa antes de responder. ¿Qué otra cosa es importante en la causalidad? En el ejemplo, se habla *y se escucha*, es decir, las personas que intervienen en la llamada están conectadas de alguna manera en ambos extremos de la línea telefónica, y la causa y el efecto deben estar conectados espacialmente o mediante una cadena de causas y efectos. En el caso del teléfono se trata de una cadena de cosas no obvia: un aparato telefónico conectado a una roseta en la pared, que a su vez está conectada a un registro, a un equipo de conmutación o central telefónica, a una red regional, nacional o mundial, y luego algo similar hasta llegar al aparato telefónico de la otra persona.

7.2.4 Consenso

Aunque se han tocado aspectos de la comunicación en general, se aprovechará la agradable distracción que proporciona la causalidad para hablar de otros tópicos relacionados y que, desde la vista del computólogo, involucran conceptos fundamentales. Para comenzar, una historia:

Hace mucho tiempo, el sultán turco inició la invasión del Imperio Bizantino. El emperador, al enterarse de tan terrible noticia, ordenó a sus ejércitos que salieran al paso de los invasores desde distintos puntos. Cada ejército estaba dirigido por un general.

Los ejércitos de Constantinopla que habrían de repeler la invasión turca eran suficientemente poderosos, pero tenían que coordinar sus acciones: todos atacaban o todos se retiraban simultáneamente. Avanzaron entonces con paso firme hasta rodear al ejército turco y los generales pasaron la noche en vela, considerando si debían ordenar el ataque al amanecer. Como había sido acordado previamente, los generales debían tomar una decisión de *consenso* sobre si atacar o no a los turcos, de manera que por la noche cada uno de ellos envió mensajeros a todos los demás.

La riqueza de los turcos no tenía límites y los bizantinos eran famosos por traidores, así que el problema con el plan de los generales es simple: algunos generales podían haber sido sobornados por el sultán turco. ¿Cómo pueden saber los generales leales si todos atacarán o se retirarán, a pesar de que algunos generales traidores mientan? Si unos atacan y otros se retiran, las consecuencias serían desastrosas.

Este problema se analizará por casos, ¿es posible ponerse de acuerdo con sólo tres generales? Suponiendo que sólo hay un traidor, en la figura 1 sería el que está pintado de verde, el 1. Para simplificar, considérese que el general 1 siempre es el general supremo de los bizantinos y a él corresponde dar la orden inicial. ¿Pueden los generales 2 y 3 saber sin lugar a dudas lo que deben hacer? No, no es posible, y sin importar cuál de los tres generales sea el traidor, siempre sucedería algo similar.

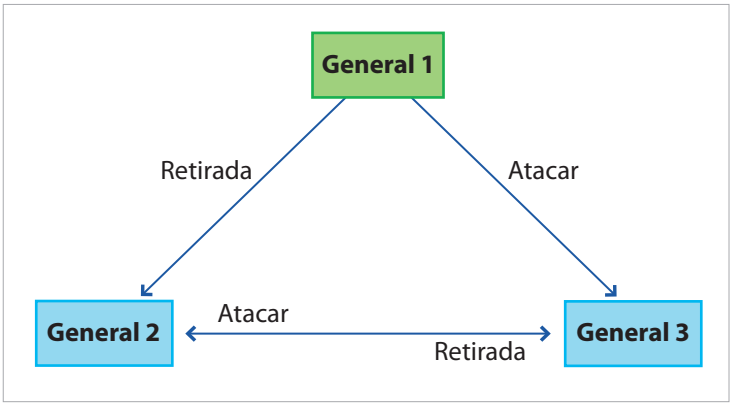


Figura 1. El general 1 es traidor.

Cada general recibe exactamente dos órdenes, pero como una proviene del traidor ambas serán contrarias entre sí, una es ataque y la otra retirada, de tal suerte que no hay manera de saber cuál está mal. ¿Qué sucede si hay cuatro generales? En este caso, sí sería posible alcanzar una decisión de consenso, ya que los generales leales reciben más mensajes correctos que espurios, como puede apreciarse en la figura 2, donde se supone que el traidor es el general 4, pues envía órdenes contrarias a los demás, es decir, de retirada. Pero no es grave, porque los demás genera-les reciben dos órdenes de atacar y solamente una de retirarse, por lo que se suman a la mayoría.

En general, se puede demostrar que puede lograrse consenso cuando hay t traidores si existen G número de generales, tal que $G > 3t$. En el ejemplo se cumple cuando $G = 4$ generales y $t = 1$ traidor, ya que $4 > 3 \times 1$.

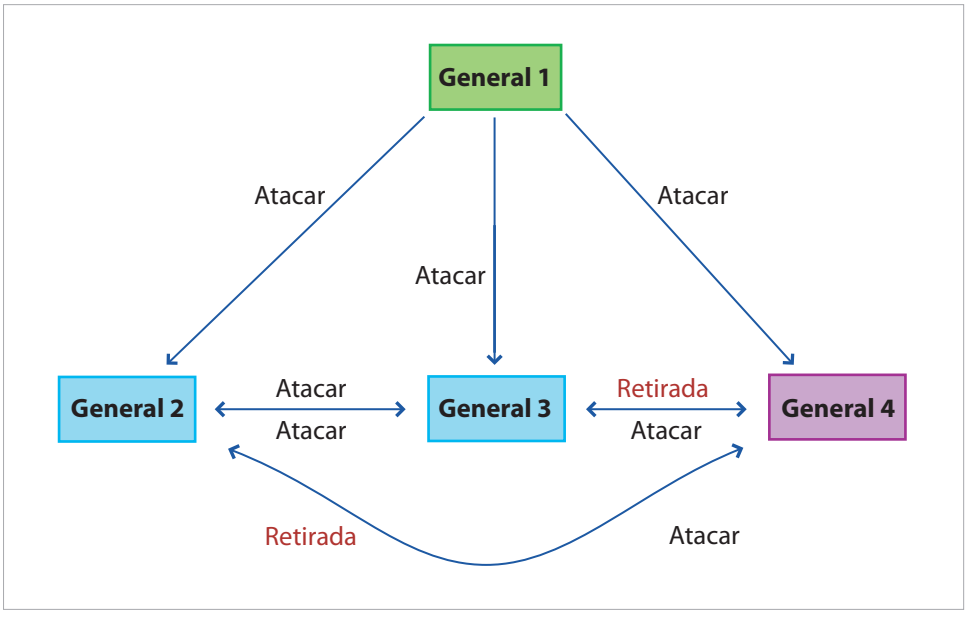


Figura 2. El general 4 es el traidor.

¿Cuántos generales se requieren al menos para lograr consenso en la decisión si hubiera tres generales traidores, es decir si $t = 3$? Se requieren mínimo "$3t + 1$" o lo que es lo mismo: $3 \times 3 + 1 = 10$ generales. Y, en este caso con tres traidores, ¿cuántos mensajes se tienen que enviar? Se tienen que realizar $t + 1$ "rondas" de mensajes. En la figura 2 se muestra la primera ronda, que *intuitivamente* significa, por ejemplo para el caso del general 2: "el general 1 (el supremo) me ordena…", pero una segunda ronda obliga al general 2 a decir "el general 3 me ha dicho que el general 1 ordenó…" y así para todos la segunda ronda involucra repetir lo que otro de los generales dice: "el general 3 me ha dicho que el general 4 dice que el general 1 ordenó…", y así sucesivamente.

7.2.5 Comunicación uno a uno

Regresando al ejemplo de comunicación vía telefónica, es importante resaltar que la intención es intercambiar información, en el caso más común, con otra persona. Conforme avanza la conversación mediante el intercambio de mensajes de manera alternada, como se mostró en el protocolo arriba, se modifica la cantidad y, como se vio en el ejemplo de los generales bizantinos, la calidad del conocimiento de las personas involucradas. La primera parte, la relacionada con la *cantidad* de información en un mensaje, tiene varios componentes: compresión de datos y detección de errores.

Como se recordará, en computación interesan los códigos con detección y corrección de errores. Los idiomas utilizan palabras frecuentes con pocas letras, como las preposiciones *a*, *ante*, *de*, etc., o los pronombres *yo*, *ella*, *él*; sin embargo, algunas otras palabras comunes como *universidad*, *abuelita* o *televisión* no lo son tanto. Esto es importante porque para comunicarse —por ejemplo, por teléfono— se puede enviar más información con menos sílabas de esta manera.

Otra característica importante del lenguaje hablado es que aunque uno de los interlocutores, por ejemplo Arcadio, no logre escuchar parte de una oración que Úrsula le dice porque está pasando un avión en ese instante, de todas formas puede comprender el significado del mensaje.

Estas dos características que son comunes a la mayoría de los lenguajes hablados por los seres humanos, la compresión de datos y la corrección de errores, son los conceptos principales en la teoría de la información. Es importante notar que en esta teoría interesa sólo la parte cuantitativa de la comunicación y no la cualitativa o el *significado* de la misma. Por ejemplo, para decir "buenos días, con permiso" se utiliza más o menos el mismo tiempo que "¡no respira!, inicio RCP", pero claramente el segundo mensaje conlleva más información.

La comunicación es un proceso que permite intercambiar información por varios métodos y requiere que todas las partes entiendan un lenguaje común en el que se expresan los mensajes que se intercambian. La comunicación ocurre en muchos niveles y de distintas formas, por ello, al igual que en otras áreas de la ciencia, en computación se le dedica gran atención e interesan las siguientes dimensiones:

- Contenido
- Fuente
- Destino
- Canal/medio
- Destino/receptor
- Propósito

Hasta el momento se han discutido distintos aspectos de la comunicación entre dos personas a través de una llamada telefónica. Ahora, ¿cómo dotar a las comunicaciones electrónicas del mismo nivel de robustez que tiene el intercambio de mensajes entre seres humanos? Nótese que si bien al utilizar teléfonos se involucra *tecnología*, la interpretación y análisis de los mensajes lo realizan las personas involucradas. Lograr que dos sistemas de cómputo puedan intercambiar mensajes y tengan la capacidad de detectar e incluso corregir errores será el siguiente reto.

7.2.6 Comunicación transitoria y persistente

Regresando al protocolo de Arcadio y Pilar para coordinar la salida de sus respectivas mascotas al patio mediante banderas, ahora se analizará desde el punto de vista del computólogo y la comunicación en sistemas concurrentes, donde se encuentran dos tipos de comunicación naturales:

- *Comunicación transitoria*: requiere que ambas partes participen al mismo tiempo. Hablar en persona y las conversaciones telefónicas son ejemplos de este tipo de comunicación.
- *Comunicación persistente*: los participantes pueden interactuar en distintos tiempos. Enviar mensajes por correo y subir o bajar las banderas son ejemplos de comunicación persistente.

La exclusión mutua requiere de comunicación persistente. En computación, este concepto se utiliza en diversas partes, por ejemplo, un sistema operativo moderno utiliza exclusión mutua para permitir que un programa pueda utilizar los recursos compartidos del sistema, como la pantalla o el teclado, y la manera en que los distintos programas pueden acceder a estos recursos es utilizando interrupciones, que son similares a las banderas de Arcadio y Pilar, sólo que en este caso el sistema operativo mismo revisa el estado de las interrupciones y actúa en consecuencia, ya sea permitiendo que se ejecuten y accedan a los recursos o no.

Curiosidades
Durante el siglo pasado se vivió una revolución en las telecomunicaciones, con nuevos medios para comunicación a distancia. La primera transmisión transatlántica de radio se realizó en 1906 y abrió la puerta a nuevas formas de transmisión de información:
- Telecomunicaciones analógicas, que incluyen telefonía, radio y televisión convencionales.
- Telecomunicaciones digitales, que incluyen la telegrafía y redes de computadoras.

Los medios de comunicación modernos permiten intercambios intensos de mensajes a larga distancia, y favorecen la comunicación entre varios participantes o muchos-a-muchos, por ejemplo, a través del correo electrónico y los foros de internet. Por otro lado, los medios estándar de comunicación masiva favorecen la comunicación uno-a-muchos, por ejemplo, la televisión, la radio, el cine, los periódicos y las revistas.

7.3 REDES

7.3.1 Correo terrestre

Se van a introducir nuevos retos y conceptos relacionados con la comunicación, a partir de otro ejemplo común: el envío de cartas por correo terrestre. Cuando se envía una carta por correo a una dirección en otra ciudad, ¿qué camino sigue, cuántas oficinas de correo reciben y envían la carta? Más aún, ¿cómo saber si el destinatario recibió o no la carta?

Una posibilidad para responder a la primera pregunta es depositar la carta en el buzón más cercano: el cartero recoge la carta y la lleva a la oficina de correos local. Una vez en la oficina, un empleado (o una máquina) revisa la dirección del destinatario y selecciona la oficina de correos que entrega en esa zona. Si la oficina seleccionada es cercana, por ejemplo en otro lugar de la misma ciudad, entonces agrega la carta al paquete que se envía a esa oficina. Si la oficina no se encuentra en la misma ciudad, entonces la carta se agrega al paquete que va a la oficina de correo más adecuada.

Figura 3. Sobre con propaganda de 1905 | © Anónimo.

Al igual que con la conversación telefónica, en una carta se encuentran varios elementos comunes que generan un protocolo. El contenido de la carta misma es libre, aunque se parece a la transcripción de la mitad de una conversación telefónica: usualmente inicia con un saludo, seguido del contenido o tema principal y una despedida. Sin embargo, ahí no termina el proceso. De hecho, se podría decir que en este momento se inicia un nuevo protocolo, el de *entrega* de la carta:

1] La carta se dobla e introduce en un sobre o un paquete para protegerla y agregar un cierto nivel de confidencialidad.
2] En el sobre se escriben en lugares claramente distinguibles y estándar los datos del destinatario y el remitente.
3] Se compra un cierto número de estampillas, que se pegan en el sobre en la esquina superior derecha.
4] Se deposita la carta en un buzón oficial (aquí termina la interacción del remitente con el envío de la carta).
5] El servicio postal nacional se hace cargo de recoger las cartas y procesar el envío y entrega o, en caso de algún error como no encontrar al destinatario porque cambió de domicilio, se encarga de regresar la carta al remitente.

Curiosidades
Una de las primeras rutas de correo terrestre cumplió 150 años en 2007: la ruta entre San Antonio y San Diego en Estados Unidos, y había cinco estaciones entre ambas ciudades. La ruta de correo tenía unas 1 450 millas, el recorrido era peligroso y llevaba varios días a caballo.

La oficina de correo *más adecuada* ¿cuál sería? Posiblemente aquella que se encuentre más cerca de la ciudad destino. Por ejemplo, si el destinatario de la carta está en Veracruz y la carta sale de una oficina en el sur de la Ciudad de México, tiene sentido enviar la carta a la oficina de correos por la salida a Puebla. Esta oficina puede enviar paquetes de correo a una oficina en Puebla y así hasta llegar a los límites entre Puebla y Veracruz.

En computación, el problema de encontrar la mejor ruta para enviar paquetes se conoce como *enrutamiento*, y funciona más o menos como se comentó en el caso de las cartas. Las oficinas de correo que pueden atravesar fronteras y están conectadas con distintas regiones postales reciben el nombre de *puertas de enlace*.

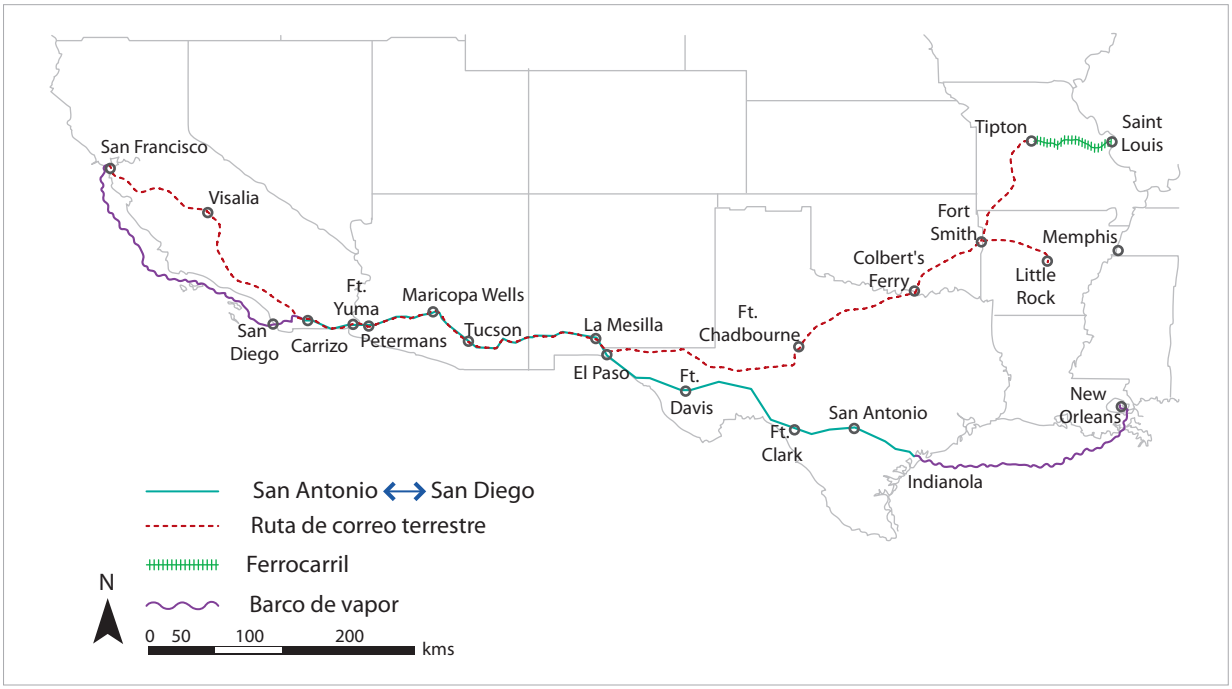


Figura 4. Ruta SA & SD.

Ahora un segundo problema, ¿cómo saber si la carta que se envió fue recibida? Hay varias opciones: esperar que el destinatario conteste con otra carta, preguntarle por teléfono o algún otro medio, o enviar la carta por correo certificado o paquetería para poder verificar en la oficina de correos que fue recibida y cuándo.

Si se piensa de nuevo en los generales de la historia anterior, conforme reciben a los mensajeros de los demás generales, con cada mensaje se modifica lo que saben y lo que no. Por ejemplo, supongamos que el mensajero del general 2 llegó con el general 3 y le dijo "recibí la orden de atacar", en ese momento el general 3 sabe que el general 2 envió a su mensajero, pero también sabe que el general 3 no sabe que él ya recibió el mensaje. Podría suceder que uno, varios o todos los mensajeros de un general fueran capturados por los turcos invasores. El general 3, por su parte, también tiene un cierto número de conocimientos y dudas: sabe que su mensajero ya salió para avisar del ataque al general 2, pero no sabe que el general 2 ya recibió su mensaje.

Este problema de certidumbre o incertidumbre, dependiendo de la óptica, se estudia en muy diversos contextos: en física, en teoría de juegos, en teoría de la información, en mercados financieros, en la predicción del clima, etc. Como se observó en el ejemplo de los generales bizantinos, la incertidumbre se propaga a través de la red de generales conforme avanzan los mensajeros.

Curiosidades
El Servicio Postal Mexicano (Sepomex) cuenta con más de 35 000 puestos de servicio, alrededor de 3 000 rutas y circuitos terrestres, y anualmente maneja más de 700 000 000 de piezas (cartas, paquetes, etcétera).

7.3.2 Redes de computadoras

Una red de computadoras se compone de múltiples computadoras conectadas que se comunican por medio de cables o del aire. Una red casera, por ejemplo, se compone de un par de computadoras que comparten una impresora y archivos. El tamaño y la capacidad de crecer de una red de computadoras están determinados por el medio físico de comunicación y por el software o protocolo que controla dicha comunicación.

Curiosidades
Las redes de telecomunicaciones pueden clasificarse por el estándar que implementan, el modelo de referencia OSI (Open System Interconection; modelo de interconexión de sistemas abiertos) o la suite de protocolos de internet. En la práctica, la mayoría de las redes utiliza el Protocolo de Internet (IP). También clasificamos una red por su escala, y las más comunes son las redes de área local (LAN, por sus siglas en inglés) y las redes de área amplia (WAN, por sus siglas en inglés). Por su método de conexión, los más comunes son ethernet e inalámbrico. También podemos clasificar las redes por su topología: de estrella, de anillo y, finalmente, de árbol o jerárquicas.

La topología es importante porque permite a los dispositivos de hardware que conectan las computadoras y los componentes de la red que vean y utilicen sus relaciones lógicas entre ellos. Esto significa que la topología de una red es independiente de la disposición física de la misma.

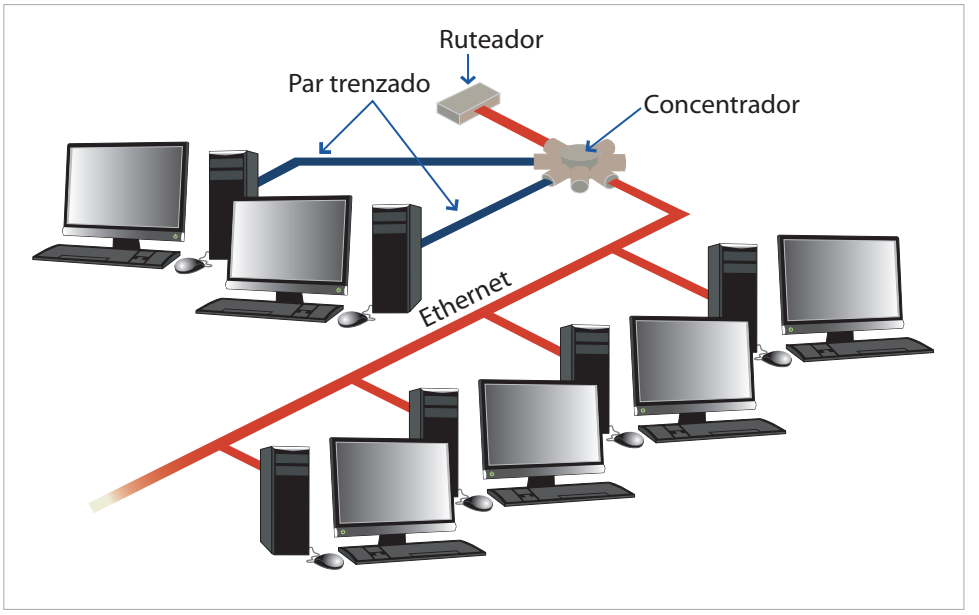


Figura 5. Red típica de ethernet.

7.3.3 Conmutación de paquetes

Concepto
Ethernet es una familia de tecnologías de redes de computadoras basada en paquetes de datos de tamaño fijo, utilizada para conectar redes de área local. El nombre se tomó del concepto físico de "éter", que se utiliza para describir un medio para transmisión de luz. **Ethernet** se encuentra en uso desde 1990 y es el método de conexión más común. Las redes inalámbricas Wi-Fi, muy comunes en casas, aeropuertos y cientos de lugares públicos hoy día, son una extensión de ethernet.

Uno de los conceptos más importantes de la era de las redes de computadoras es la *conmutación de paquetes*, en la cual los paquetes (unidades de información) son dirigidos entre nodos a través de líneas de datos que incluyen más tráfico de otros paquetes. En cada nodo, los paquetes son encolados o almacenados temporalmente, lo que resulta en un retraso variable. Se puede pensar en cada paquete como una carta y en cada nodo como una oficina (o destino) para el correo.

La conmutación de paquetes se encuentra en el corazón de los protocolos básicos de red; provee, por plantearlo de una manera sencilla, la base, la primera capa, para las redes de computadoras. Encima de IP se han construido otros protocolos para solucionar diversos problemas, como el protocolo de control de transmisión (TCP, por sus siglas en inglés), que permite recibir paquetes que llegan en desorden o que no llegan completos.

Las redes telefónicas funcionan con un principio muy diferente al de conmutación de paquetes. Cuando Arcadio descuelga el teléfono y marca el número de casa de Úrsula, la red telefónica cierra varios interruptores de acuerdo con la ruta establecida para llegar de su casa a la de ella, lo cual forma un circuito cerrado por el que se puede llevar a cabo la conversación entre ambos, un circuito temporal dedicado que se rompe cuando alguno de ellos cuelga el teléfono. Por ello, la red telefónica es llamada de conmutación de circuitos. Las redes de conmutación de circuitos son anteriores a las de conmutación de paquetes; de hecho, éstas surgieron como una alternativa confiable para robustecer el sistema de comunicaciones estadunidense.

A su vez, sobre TCP/IP existe una gama de protocolos con fines especializados, o protocolos de aplicación, como el protocolo de transferencia de hipertexto (HTTP, por sus siglas en inglés), que permite que ciertos agentes accedan a documentos, archivos y otros recursos en la red mundial web (WWW, por sus siglas en inglés). Más adelante retomaremos estos temas.

7.3.4 Comunicación entre homólogos y abstracción

El señor secretario de comercio exterior del país X, luego de un concienzudo análisis y reuniones con su equipo de asesores, decide comunicarse con su homólogo del país Y. Llama a su secretario particular y le dice en correcto español: "Alejandro, necesito enviarle una carta al secretario de comercio exterior de Y, quiero que le digas que le solicito que nos reunamos a la brevedad". X necesita establecer una política arancelaria de importaciones de manufacturas de acero que afectará a varias compañías de Y pero, a cambio, está dispuesto a establecer unas cuotas temporales a sus exportaciones de productos primarios que beneficiarán a los productores de Y.

Alejandro no es economista, no entiende casi nada de lo que habla el señor secretario, pero no importa, su inglés es mucho mejor que el que balbucea su jefe. Así que redacta una carta muy pulcra que le hace llegar al secretario particular del secretario de comercio exterior de Y, quien a su vez, aunque no entiende muchos de los términos que se usan en la carta, le comunica a su jefe el contenido de la misma en un fluido francés, lengua oficial de Y.

Analizando este ejemplo, los secretarios de comercio de X y de Y hablan un lenguaje propio de su investidura, los términos de ese lenguaje son bien entendidos por ambos; cuando uno de ellos dice algo da por sentado que el otro lo entiende. La labor de sus secretarios particulares es, justamente, que cada uno de los secretarios de comercio pueda hablar como si estuviera frente a su homólogo del otro país. La comunicación física, el envío y la recepción de mensajes no se da al nivel de los secretarios de economía, sino más abajo, entre sus secretarios particulares, y posiblemente aún más abajo, entre los mensajeros. Para comunicarse entre sí, los secretarios de comercio usan el nivel de abstracción propio de ellos, pero los mensajes van de uno a otro a través de otros niveles de abstracción más básicos que se encargan de las traducciones e interpretaciones necesarias.

En las redes de computadoras opera el mismo principio. Los protocolos de comunicación establecen siempre las reglas a un cierto nivel de abstracción, suponiendo que los interlocutores se encuentran justo a ese nivel. Y luego, para llevar a cabo la realización efectiva de la comunicación, se montan en los servicios ofrecidos por capas de abstracción inferiores, cuyas reglas de operación también están dadas a su propio nivel y que, a su vez, recurren a capas inferiores para hacer efectiva la comunicación.

El número de capas de abstracción es variable y depende de la complejidad de la tarea que se pretende llevar a cabo. Tareas más complejas requieren de mayor nivel de abstracción para hacerse y, por tanto, de mayor infraestructura de comunicaciones.

Todas las redes de computadoras se construyen por capas. Las únicas entidades que realmente "se hablan" son las de la capa de abstracción más baja, es decir, los bits (bueno, realmente señales eléctricas u ópticas que los representan). Estas cadenas de bits se pasan o provienen de la capa inmediata superior que probablemente hace ciertas verificaciones y corrección de errores elementales, y a su vez se comunica con la capa inmediata superior, para recibir de ella lo que hay que transmitir o para pasarle lo que se le debe entregar.

Curiosidades

En la década de los sesenta del siglo pasado, en plena guerra fría, Paul Baran (1926), ingeniero eléctrico polaco que trabajaba para la Corporación RAND, se ocupaba del siguiente problema: imaginemos que Estados Unidos recibe de la Unión Soviética un ataque nuclear que probablemente destruye gran parte de la infraestructura de comunicaciones en ciertas regiones estratégicas. Si se confía la capacidad de contraataque a las comunicaciones telefónicas, lo más probable es que no se pueda responder. La clave para que sea posible mantener comunicaciones en una situación como ésta es que existan diferentes rutas entre las entidades que se comunican y que una sola "conversación" pueda ser enrutada de diferentes maneras, dependiendo de los caminos disponibles, dando origen a lo que hoy llamamos una red de conmutación de paquetes.

Paul Baran.

Al igual que con los secretarios de comercio exterior, puede ser que los homólogos hablen distinta lengua, tengan diferente manera de decir algo, pero el protocolo se debe encargar de que ambos entiendan lo mismo. Pueden usar diferentes modalidades de representación de los datos que se dicen, a lo mejor los números enteros de una máquina son de diferente tamaño que los de otra o usan diferentes conjuntos de caracteres, pero la infraestructura completa debe garantizar que se entiendan bien. Todas las traducciones necesarias se hacen en el camino.

7.3.5 Enrutamiento

El crecimiento y la constante integración de redes para formar otras cada vez más grandes presentan complicaciones para la selección de rutas en la red para el envío de datos y la dirección del tráfico. El enrutamiento tiene la tarea de encontrar la mejor ruta posible, y diversos tipos de redes utilizan esta técnica, incluyendo las telefónicas, de computadoras o de transporte, como en las oficinas de correo ya mencionadas.

7.3.6 Congestión

Curiosidades
Este problema de congestión es bien conocido para la mayoría de las personas que viven en grandes ciudades, pues lo sufren todos los días en la forma de tráfico vehicular. En la ciudad de México, por ejemplo, diariamente circulan más de 3 000 000 de vehículos, lo cual, además de generar congestión o tráfico vehicular, es la principal causa de contaminación.

¿Qué sucede si en un momento dado llegan dos o tres veces más cartas que las usuales en una oficina?, ¿o si dos de los tres empleados de una oficina están enfermos y no pueden llegar a trabajar? Lo más probable es que muchas de las cartas recibidas no puedan ser procesadas y salgan con diversos retrasos a sus destinos finales.

Dada la manera en que se realiza la transmisión de paquetes en una red, es posible que un nodo particular o una línea de transmisión se sature de manera tal que su calidad de servicio se deteriore. Una vez que esto ocurre, algunos de los efectos típicos incluyen retraso o pérdida de paquetes, o la imposibilidad de aceptar nuevas conexiones.

El problema de congestión se estudia desde diversos ángulos, ya que es común a prácticamente todos los medios de comunicación conocidos —desde un punto de vista psicológico y social, desde el de la física e incluso desde el punto de vista económico—. ¿Cuánta incomodidad provocada por el tráfico se puede soportar o qué tanta inversión se requiere para mejorar los medios de transporte? En el caso de las redes modernas, se utilizan técnicas para controlar o evitar la congestión de la red, por ejemplo *exponentail backoff* (802.11 y ethernet), *window reduction* (TCP) y *fair queueing* (enrutadores).

7.3.7 Transmisión

En el siguiente experimento, un grupo de alumnos es llevado a un salón donde no hay luz, por lo que no pueden ver nada. Entonces se les pide que, uno por uno, todos digan su nombre. ¿Cuál sería el mejor algoritmo para lograrlo en el menor tiempo posible?

¿Quién comienza, quién continúa? Una posibilidad es utilizar el siguiente algoritmo:

1] Intentar decir el nombre.
2] Si al iniciar a hablar, se escucha un ruido (otro alumno diciendo su nombre), callarse y pasar a 4.
3] Si no hay ruido, decir el nombre y terminar.
4] Esperar el doble de tiempo que antes y regresar a 1.

Es decir, se intenta decir el nombre; si hay ruido, se debe esperar dos segundos y se vuelve a intentar; si hay ruido, ahora la espera es de $2^2 = 4$ segundos y se vuelve a intentar y así hasta lograr decir el nombre. Éste es un algoritmo utilizado en redes ethernet estándar para evitar la congestión y se conoce como *exponential backoff*.

¿Si tienen que hacer lo mismo, pero en un salón con luz, como procederían? Esto es típico, por ejemplo, en el primer día de clases, cuando el profesor les pide a todos que se presenten diciendo su nombre. Hay varias opciones: la primera sería definir un orden, por ejemplo recorriendo las filas de adelante hacia atrás, de izquierda a derecha. Otra opción sería que cada alumno levante la mano si aún no se ha presentado y esperar a que el profesor seleccione a uno. Este último algoritmo, donde existe un moderador, es muy socorrido en computación y deben considerarse dos conceptos importantes: imparcialidad y hambruna.

En este contexto, *imparcialidad* es la idea de que a todos los estudiantes —o a los paquetes en una red de computadoras— se les dé una oportunidad de transmitir. El problema de que un moderador elija al siguiente es que puede seleccionar a los mismos para transmitir, ignorando a algunos, los cuales estarían entonces en un estado de hambruna, pues no pueden avanzar en su tarea.

En las redes de computadoras existen distintos tipos de transmisión de datos, entre ellos unicast, que va de un nodo a otro de manera directa y, en el otro extremo, broadcast, donde la transmisión se realiza de manera simultánea a todos los nodos en la red.

7.4 INTERNET: RED DE REDES

Internet es una red de computadoras mundial y de acceso público que utiliza IP para transmitir datos; es una red de redes formada por millones de redes gubernamentales, de negocios, académicas e incluso domésticas, que juntas dan soporte a una gama muy amplia de servicios, como correo electrónico, charla en línea y varios tipos de mensajería instantánea, transferencia de archivos y, probablemente el más socorrido de todos, las páginas y documentos web interligados que conforman la *world wide web*, conocida simplemente como *la web*.

7.4.1 Infraestructura

Como ya se mencionó, internet está formada por muchas redes más pequeñas, por lo que su infraestructura es heterogénea y diversa. Sin embargo, lo que sirve como pegamento para mantenerla unida son los protocolos en las diversas capas de la red: el nivel más bajo lo atiende IP, en la actualidad la versión 4 o IPV4 es el protocolo dominante, aunque ya está lista y muchas redes locales utilizan la sexta versión, IPV6. El siguiente nivel en la estructura es TCP, para conexiones virtuales y para agregar un cierto nivel de garantía y confianza a la transmisión de datos. Finalmente, en la capa superior de internet, se encuentran diversos protocolos de aplicación, cuya finalidad es definir mensajes y formatos de datos específicos para enviar y recibir entre aplicaciones en cada extremo de las líneas de comunicación.

Se ha dicho que internet es una red de conmutación de paquetes, y esta característica es lo que permite atacar una serie de problemas implicados en ella: por ejemplo, ¿cómo se pueden dar garantías de que el ancho de banda, es decir, de que el canal de comunicación, sea adecuado para llevar a cabo una videoconferencia por internet?, ¿qué hacer si no han

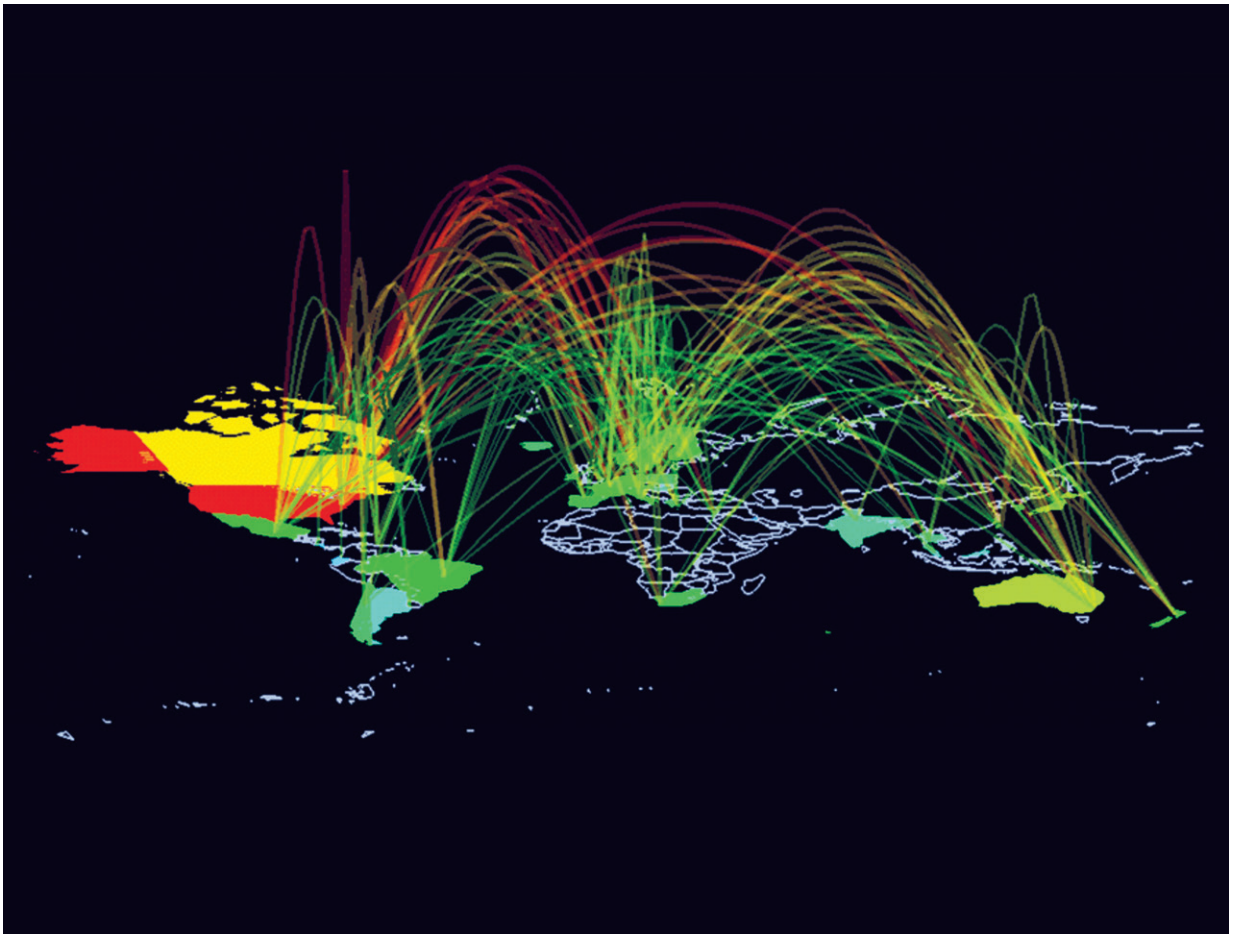


Figura 6. Arcos que muestran el tráfico en internet | © Stephen G. Eick.

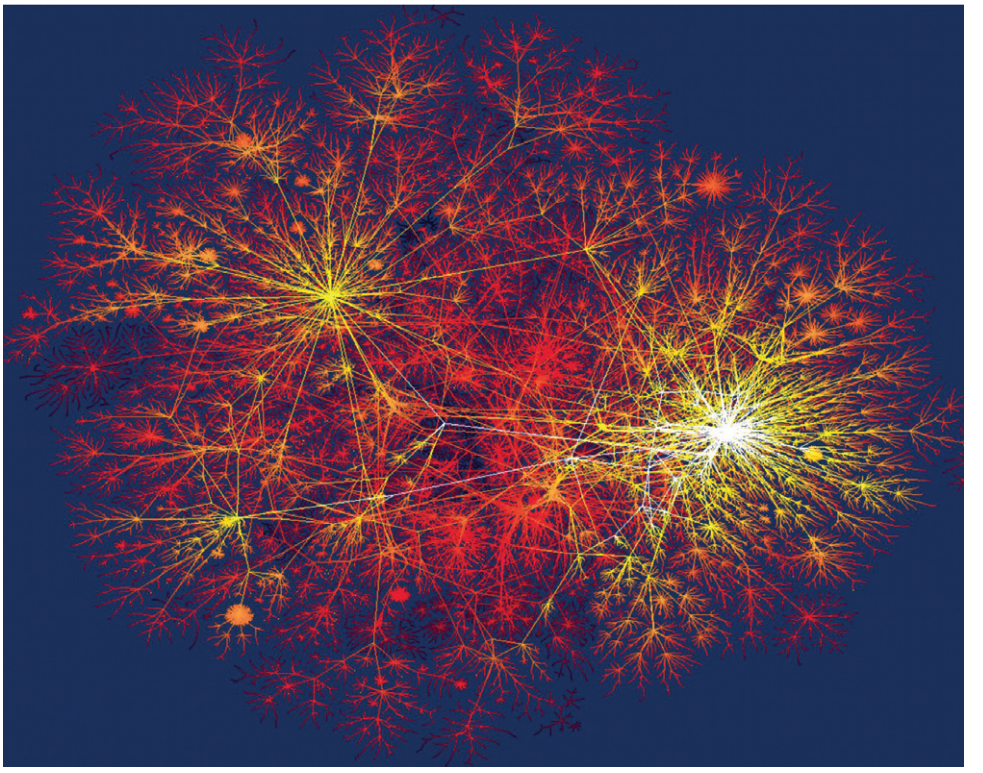


Figura 7. Instantánea de conectividad en internet (datos de algunos proveedores de internet comerciales) | © K. C. Claffy.

llegado todos los paquetes de un mensaje?, ¿cuánto tiempo debe esperar una computadora aquellos paquetes que no han llegado?

Si alguna vez se ha armado un mueble de esos que ahora venden en cualquier tienda y que vienen desarmados o que se pueden comprar por partes —primero un librero con tres divisiones y luego, si se necesita, nuevas divisiones para expandirlo—, es probable que se pueda apreciar la simplicidad e importancia de la conmutación de paquetes: en la fábrica diseñaron y construyeron el mueble (por ejemplo, el escritorio para una computadora), después lo dividieron en varias piezas, las numeraron de alguna forma e hicieron un manual de ensamblado. Al adquirir el mueble, éste se lleva en una caja con las piezas y el manual para ensamblarlo (conectar las piezas 1 y 3 con un tornillo de número 10, a continuación conectar las 5 y 7, etcétera).

Para cualquier mensaje que se transmita por internet sucede algo similar, la computadora que envía el mensaje lo parte en muchos pedazos de tamaño fijo, los identifica de una manera estándar y los envía. A diferencia del mueble, sin embargo, aquí otro protocolo se encarga de decidir cuál es la mejor ruta para cada uno de los paquetes, por lo que pueden llegar en desorden y a distintos tiempos a la computadora destino. ¿Qué tan grande puede ser cada paquete? No mucho, a lo más de 1 500 caracteres de extensión, que se meten en un sobre del protocolo IP para su viaje por la red. El número de paquetes requerido para enviar un correo electrónico es, por supuesto, mucho menor que el empleado para enviar la película de la boda de tus padres a tu hermana que vive en otro país.

Una vez que se reciben todos los paquetes, TCP arma nuevamente el mensaje original, pero ahora del lado del receptor. Si uno de los paquetes se dañó o corrompió durante la transición, el protocolo vuelve a solicitar el paquete al emisor. Lo mismo ocurre si después de cierto tiempo no ha llegado alguno de los paquetes. Juntos, estos protocolos, y teniendo una conexión con cierta velocidad y capacidad en la computadora local a la red, ¿se podría garantizar la transmisión de audio y video en una videoconferencia con alumnos y profesores de otra escuela para tener una clase conjunta? La respuesta es no, no es suficiente para garantizarlo, porque muchas otras computadoras y redes son utilizadas durante la transmisión de los paquetes, entre el sitio local y el remoto.

En la práctica, por supuesto, se logran configuraciones de las distintas redes locales involucradas en transmisiones especiales, como una videoconferencia o un programa de radio por internet, que permiten *reservar* un espacio de la capacidad de la red local y de su conexión a internet, así como otorgar mayor prioridad a los paquetes que llevan pedazos de audio y video, para *garantizar* que se podrá realizar la videoconferencia o transmisión de audio. Ésta es una parte de lo que se conoce como calidad de servicio en redes (QOS, por sus siglas en inglés) que, en efecto, está relacionado con el concepto de congestión mencionado anteriormente. *Es relevante comprender cómo se enlazan distintas redes para formar internet, lo que la convierte en una red sin límite de escalamiento; esto es, su estructura y dinámica son independientes del tamaño (número de nodos) en la red.*

7.4.2 Direccionamiento

La autoridad que coordina la asignación de identificadores únicos en internet es la Internet Corporation for Assigned Names and Numbers (ICANN), lo que incluye nombres de dominios, direcciones IP y los números de puerto y parámetros para los distintos protocolos. Es vital para el buen funcionamiento de internet contar con un espacio de nombres global, en el cual exista uno y sólo un dueño de cada nombre.

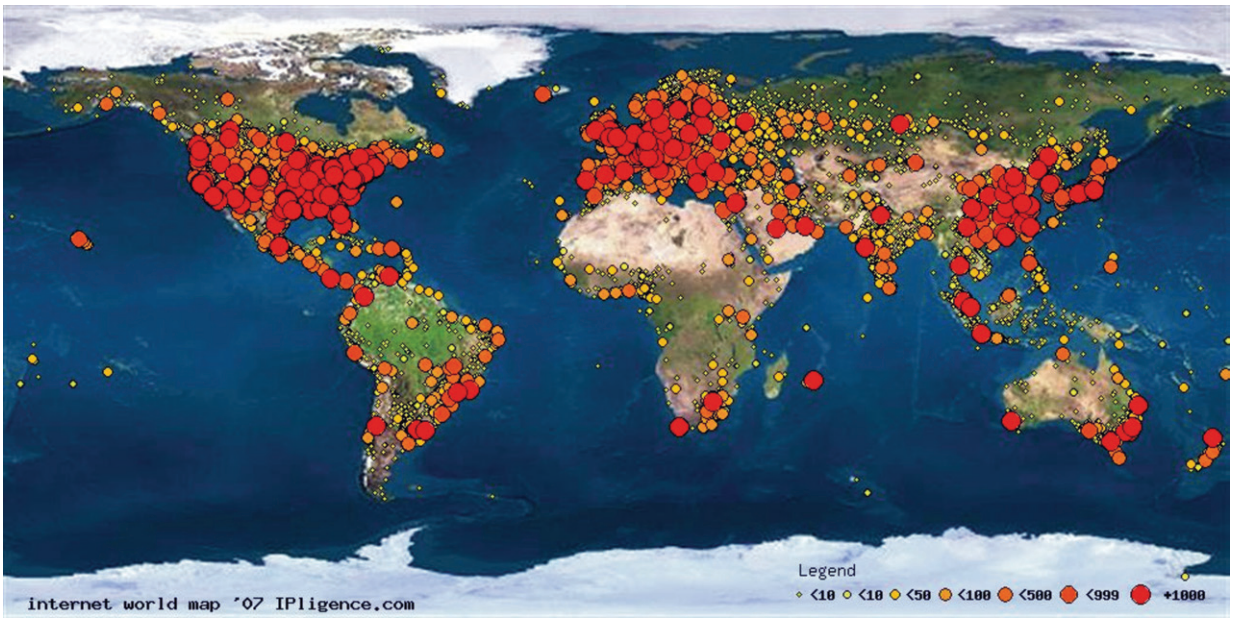


Figura 8. Distribución de internet en el mundo | © IPligence.

Área geográfica	Número de direcciones	Porcentaje
África	40 241 664	1.519
Antártica	15 620	0.001
Asia	371 297 015	14.015
Caribe	1 681 866	0.063
América Central	2 557 340	0.097
Europa	569 838 903	21.510
Oriente Medio	12 011 131	0.453
América del Norte	1 481 754 661	55.932
Oceanía	76 417 711	2.885
América del Sur	93 409 304	3.525

Curiosidades
Las oficinas centrales de ICANN se encuentran en Marina del Rey, en California, Estados Unidos. No obstante, están dirigidas por una junta de directores conformada por expertos de las comunidades técnicas, de negocio, académicas y no comerciales de internet. Debido a que internet es una red global, formada por muchas redes de manera voluntaria, no existe un cuerpo que la gobierne. El rol de ICANN, como único organismo de organización centralizado, se limita al manejo de nombres, direcciones IP y números de puerto y parámetros.

Las direcciones IP en el mundo están distribuidas como se muestra en la siguiente tabla. En la figura 8 se puede ver gráficamente la distribución.

7.4.3 Idioma

Una problemática de internet hoy en día es el manejo de distintos idiomas, sobre todo para los protocolos de aplicación. Esto se debe a la combinación de dos situaciones: en primer lugar, el origen de internet se dio en Estados Unidos y, por tanto, el inglés fue su lengua franca; y, en segundo, las primeras computadoras tenían capacidades limitadas y no era sencillo manejar caracteres acentuados o que no podían representarse en alfabeto latín base.

Las tecnologías de internet han evolucionado a pasos agigantados en años recientes, particularmente en el uso de Unicode, lo que permite el desarrollo y comunicación en la mayoría de los idiomas comunes.

7.5 WEB

Internet, a pesar de ser una red de proporciones descomunales, sigue siendo un canal para soportar comunicaciones. Las aplicaciones actuales que aprovechan o extienden internet son muy grandes y están en constante evolución, por lo que explicaremos sólo las más relevantes y aquellas que han ayudado a convertirlo en uno de los avances tecnológicos más importantes en la historia de la humanidad.

7.5.1 Qué es la web

Muchas personas se refieren a internet y la web como si fueran la misma cosa; sin embargo, como se aclaró antes, estos dos términos representan cosas distintas. La web utiliza *http*, un protocolo de aplicación de internet, para entregar lo que sus usuarios solicitan.

La web es un conjunto enorme de documentos entrelazados, imágenes y otros recursos, todo vinculado por hiperligas y URL. El concepto URL es un término técnico utilizado en la web para hacer referencia a un recurso, corresponde a las sigla en inglés de Uniform Resource Locator y se utiliza, para todo fin práctico, como URI (Uniform Resource Identifier), que es un nombre más adecuado. En resumen, un URL o URI sirve para identificar de manera única un recurso en la web, ya sea una página, una imagen, un video o cualquier otro recurso que pueda transmitirse por este medio.

Los productos de software que pueden acceder a los distintos contenidos que se ofrecen en la web son llamados *agentes usuarios*. Normalmente, un navegador, que es un agente usuario, como Firefox o Internet Explorer, permiten al usuario acceder a páginas web y navegar de una a otra, utilizando las hiperligas. Una página web puede contener prácticamente cualquier combinación de datos de computadora, incluyendo fotografías, gráficas, sonido, video, texto, multimedia y contenido interactivo; por ejemplo, juegos, aplicaciones de oficina y científicas, etcétera.

Es muy sencillo para las organizaciones y las personas publicar información e ideas accesibles a una gran audiencia a través de la web. Es fácil encontrar formas de publicar información en la web, sin necesidad de realizar una inversión o esfuerzo iniciales considerables. Sin embargo, mantener sitios profesionales requiere un gran trabajo de administración, diseño, edición y mantenimiento de información, que para muchas empresas e individuos representa un reto muy costoso.

Cómo funciona

Ver o visitar una página web inicia al ingresar el URL de la página en un navegador, o siguiendo una hiperliga a esa página o recurso. Después de esto, el navegador inicia una serie de comunicaciones, tras bambalinas, para solicitar la página y mostrarla:

- Primero, la porción que incluye el nombre del servidor debe *resolverse* para obtener una dirección IP. Esta resolución se realiza mediante otro protocolo de internet, que mantiene una base de datos distribuida de nombres y se conoce como *sistema de nombres de dominios* o DNS, por sus siglas en inglés.
- Una vez que se conoce la dirección IP del servidor, el navegador envía un paquete de solicitud http para obtener el recurso deseado.

Curiosidades
Unicode es un estándar que permite a las computadoras representar y manipular un texto expresado en cualquiera de los sistemas de escritura modernos. El repertorio actual de Unicode asciende a más de 100 000 caracteres, un conjunto de tablas de códigos para referencia visual, una metodología de codificación, un conjunto de codificación de caracteres y una lista precisa de propiedades de caracteres, tales como mayúsculas y minúsculas. El estándar Unicode ha sido implementado en varias tecnologías recientes, incluyendo XML (Extensible Markup Language; lenguaje de marcas extensible), varios lenguajes de programación y sistemas operativos.

Curiosidades
La superficie de la web, es decir, el contenido que puede ser indexado por los motores de búsqueda convencionales, de acuerdo con un estudio de enero de 2005, contenía 11 500 000 000 de páginas web. En contraste, según estudios de 2001, se cree que la web profunda, es decir, todo el contenido en la web, incluyendo aquello que no es público, es varios órdenes de magnitud mayor. En el año 2000 se calculó que la web profunda contenía 550 000 000 000 de páginas individuales, que ocupaban unos 7 500 terabytes.

- El servidor procesa la solicitud y, en el caso usual de una página web, envía en paquetes la página solicitada.
- El navegador analiza rápidamente la página recibida y, de ser necesario, hace más solicitudes al servidor para obtener las imágenes y demás elementos necesarios antes de desplegar la página en la pantalla del usuario.

Muchas páginas web incluyen hiperligas a otras páginas relacionadas y otros documentos o recursos. Este tipo de colecciones de recursos relacionados, interconectados por medio de hiperligas, es lo que da su nombre a la web, como una *red* de información.

7.5.2 Memoria inmediata o caché

Si un usuario visita nuevamente una página web después de un intervalo breve de tiempo, la página no es solicitada otra vez al servidor. La mayoría de los navegadores utilizan un espacio de almacenamiento local, llamado *caché*, que sirve como memoria inmediata, por lo que sólo se solicitarán al servidor web, vía http, aquellas secciones de la página que hayan cambiado y lo demás saldrá de su memoria inmediata.

Utilizar este tipo de memorias inmediatas o cachés ayuda a reducir el tráfico web en internet y acelera el despliegue del contenido. Puesto que esto es una buena idea, existen diversas técnicas y tecnologías para llevar los beneficios a otros niveles. Por ejemplo, en muchas organizaciones se utilizan servidores caché que almacenan las páginas web cada vez que uno de los usuarios de la organización las visita por primera vez. De esta manera, si alguien más en la organización solicita el mismo recurso, ya se encuentra en el servidor local de la organización. Algunos motores de búsqueda o buscadores, como Google o Yahoo!, también almacenan el contenido de los sitios web que visitan.

Curiosidades
Akamai Technologies es una compañía que ofrece una plataforma de cómputo distribuida para servir contenido y aplicaciones en internet. Akamai hace, de manera transparente, una copia espejo del contenido en los servidores del cliente y lo replica en sus servidores por todo el mundo. A continuación, cuando un navegador solicita una página, aunque el nombre del servidor es el mismo, la dirección IP a la que resuelve ahora es de un servidor Akamai, que se espera siempre encontrar uno disponible y cercano al cliente. Buena parte del contenido web se distribuye, sin que lo sepamos, a través de los servidores de Akamai.

7.5.3 Estándares

Dos organizaciones, World Wide Web Consortium (W3C) e Internet Engineering Task Force (IETF), generan y mantienen varios estándares formales y otras especificaciones técnicas que definen la operación de diferentes aspectos de la web, internet y del intercambio de información. Es común, cuando se discuten estándares para la web, ver las siguientes publicaciones:

- Recomendaciones para lenguajes de marcado, especialmente HTML y XHTML, de W3C, que definen la estructura e interpretación de documentos hipertexto.
- Recomendaciones para hojas de estilo, específicamente CSS, de W3C.
- Estándares para ECMAscript, mejor conocido como JavaScript, un lenguaje de programación para extender las páginas web. El intérprete, encargado de ejecutar los programas escritos en JavaScript, está incluido en el navegador web en la máquina del cliente.
- Uniform Resource Identifier (URI), que es el sistema universal para hacer referencia a recursos en internet, tales como documentos hipertexto o imágenes. URI está definido por el RFC 3986 de IETF.

7.5.4 La revolución web

La web ha sido un terremoto para nuestra cultura, está revolucionando nuestras economías, nuestras ideas acerca de compartir trabajo creativo e incluso la forma de concebir la religión, la educación, el gobierno, la democracia, etc. ¿Por qué? Según David Weinberger, la web es un mundo nuevo que estamos colonizando. Somos como los europeos cuando viajaron a América por primera vez: no sabemos qué vamos a encontrar o qué hacer para encontrarlo, ¿empacamos equipo para escalar montañas, canoas para ríos y lagunas, nos preparamos para el desierto o las tres opciones? Pero los europeos tenían una ventaja, porque aunque no conocían la geografía del nuevo continente al menos sabían que existía una. La web no tiene geografía, no existe un paisaje, no hay nada natural.

El concepto de distancia en la web no existe, podría pensarse en el número de brincos que un paquete de datos tiene que hacer para moverse entre dos computadoras como la distancia, pero entonces es probable que una persona en Alaska esté más cerca de usted, que su hermana que escucha la misma transmisión de datos en otra escuela, en la misma Ciudad de México.

La web tiene muy pocas reglas y no existen autoridades. El sentido común no funciona aquí y aún no emerge entre los seres humanos otro sentido que lo reemplace. No es una sorpresa, entonces, que existan dificultades para entender cómo se realizan los negocios en este nuevo mundo. No se sabe cómo referirse a un lugar sin fronteras, sin cercanías o lejanías, a un lugar donde una misma persona puede tener muchas identidades virtuales, completamente desconectadas de su aspecto físico. Nuevos mundos crean nuevas personas, y la web está transformando las relaciones humanas paulatinamente en formas que no se alcanzan a comprender y cuyo resultado es difícil imaginar.

7.5.5 La web semántica: llevar la web a nuevos niveles

La visión principal de la web semántica es extender los principios de la web de documentos a datos. Esto permitirá incrementar su potencial, ya que se podrán compartir datos de manera efectiva con comunidades cada vez más grandes, además de que dichos datos sean procesados en forma automática por medio de herramientas o manualmente. Habrá ligas entre distintas páginas web que darán significado a sus conexiones.

Por ejemplo, hoy se puede poner una liga que te lleve de una página web personal a la de un libro leído y, por supuesto, también se podría tener una liga a otro libro que no se ha leído, pero del que se tiene una buena opinión y sería deseable leer. Esto es interesante, pero sólo es de utilidad para otro individuo que pueda entender los significados involucrados en la página. La intención de la web semántica es que, en el futuro, las páginas web formalicen sus significados de manera tal que puedan ser procesadas automáticamente mediante programas. Se está impulsando un nuevo lenguaje, llamado RDF, inventado por los creadores de la web semántica, que permite agregar significado a las páginas, de tal manera que tu página "libro" no haga referencia sólo a la palabra libro, que es una secuencia de caracteres, sino al concepto "libro" y que la liga *cree* la noción de recomendación. Así, dos páginas web estarán relacionadas entre sí por la noción recomendación.

Como se discutió en el tema de información, el conocimiento surge de relacionar elementos de distintas formas, y la web semántica pretende que exista un lenguaje universal que ayude a describir relaciones entre los elementos que forman la web. Con esto se logrará algo que hoy es muy difícil: escribir programas que permitan extraer información

de un sitio y la compaginen con información proveniente de otro. Hoy se pueden hacer programas para atacar este problema, pero están basados en coincidencias de caracteres y palabras, no en su significado. Y, por si fuera poco, esta tarea es complicada porque depende de la forma y estilo de las páginas, que pueden ser cambiadas en cualquier momento por sus dueños o creadores.

Con la web semántica, el problema de la presentación ya no es tan relevante, porque sin importar cómo luce una página web en un navegador, el significado de la información no cambia.

7.5.6 Motores de búsqueda o buscadores

Un motor de búsqueda o buscador, como coloquialmente se le conoce, es un sistema para obtener información, diseñado para facilitar la localización de información almacenada en un sistema de cómputo. Un objetivo primordial de los buscadores es minimizar el tiempo requerido para encontrar información y, cuando se habla de la que se encuentra en internet, también se quiere minimizar la cantidad que se debe revisar.

Los buscadores más comunes son aquellos especializados en la web, o buscadores web, aunque los hay de muchos tipos, entre ellos los personales, que sirven para localizar rápidamente información en un equipo de cómputo personal, o los institucionales, que buscan solamente en servidores y máquinas locales de la organización.

Cómo funcionan

Existen muchos buscadores, cada uno con sus propias técnicas y tecnología. A manera de ejemplo, aquí se hablará sobre la tecnología de uno de los buscadores más importantes que concentra más de 50% de las búsquedas en la web: Google. Antes que nada, ¿cómo sabe el buscador qué palabras aparecen en un sitio o página web dada? Se logra generando un índice, a través de programas que navegan de manera autónoma en la web. Estos programas reciben el nombre de arañas, robots o rastreadores. El robot de Google se llama GoogleBot y recorre las páginas web como si fueran una autopista; cada vez que encuentra una página nueva, analiza el código, lo rastrea y lo envía a su centro de datos.

Otro robot, en este caso el FreshBot, se encarga de visitar sitios previamente indexados para mantener su información al día, buscar cambios e integrarlos a su índice. La base de datos de Google mantiene miles de millones de páginas y utiliza un algoritmo propietario para calificar la relevancia de los sitios. Cuando un usuario ingresa una *cadena de búsqueda*, utilizan estos valores de relevancia para entregar los resultados ordenados.

Algunos datos curiosos de Google:

- Su nombre nació de la palabra *googol*, que significa 1 seguido de cien ceros.
- Su índice incluye miles de millones de páginas y su capacidad es tal que puede buscar en todas ellas en menos de medio segundo.
- Sus grupos de discusión contienen más de 845 000 000 de mensajes de Usenet, lo que representa la colección más grande de mensajes: más de un terabyte de conversación humana.

7.6 APLICACIONES

7.6.1 E-mail

E-mail o correo electrónico es un método para componer, enviar, almacenar y recibir mensajes a través de sistemas de comunicaciones electrónicos. Cuando se hable de *E-mail*, *email*, *mail* o simplemente *correo*, se hará referencia al protocolo de internet llamado Simple Mail Transfer Protocol (SMTP). La idea de correo electrónico antecede a la de internet y, de hecho, fue una herramienta vital en la creación de ésta. El correo electrónico se ideó como una manera para que los usuarios de una computadora de tiempo compartido (*mainframe*), en 1965, pudieran comunicarse. El correo electrónico representa, hoy en día, uno de los medios más utilizados para comunicarse a distancia.

Cómo funciona

En la figura 9 se muestra una secuencia típica de los pasos involucrados en la creación, transmisión y recepción de un correo electrónico. Aunque luce como una tarea complicada, para los usuarios que redactan y envían un mensaje, como para el que lo recibe, en realidad es muy sencillo.

Úrsula redacta un correo utilizando su *cliente de correo* o Mail User Agent (MUA). Ingresa o busca en su agenda la dirección de correo electrónico de Arcadio y presiona el botón *enviar*. Una dirección de correo electrónico es una cadena de la forma: partelocal@dominio.org, lo que se conoce como una dirección de dominio plenamente calificada. La parte que antecede la @ es la parte local de la dirección, usualmente el nombre de usuario del receptor, y la parte que está después, el nombre del dominio.

Después de que Úrsula presiona el botón de enviar, se realizan las siguientes actividades:

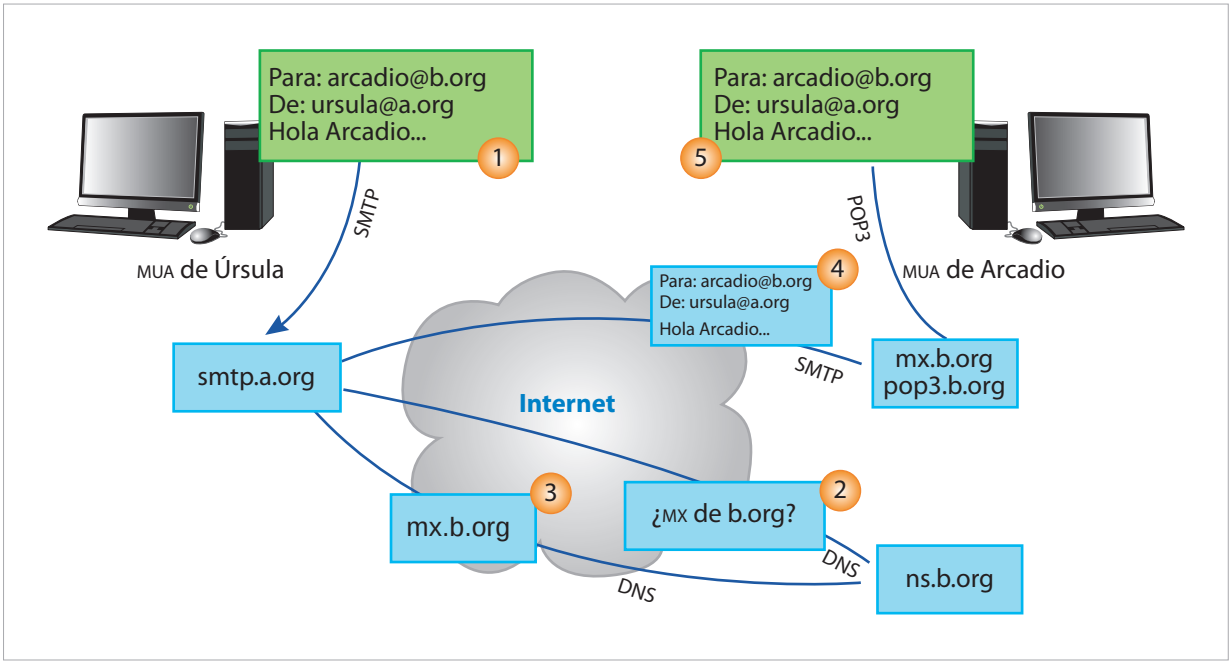


Figura 9. Secuencia de mensajes durante el envío de un correo electrónico.

- El cliente de correo da formato al mensaje y utiliza el SMTP para enviar el mensaje al agente de transferencia de correo (MTA, por sus siglas en inglés). En este ejemplo, smtp.a.org, que es manejado por el proveedor de internet de Úrsula.
- El MTA revisa la dirección de correo que provee SMTP, en este caso arcadio@b.org. El MTA localiza el dominio en el servidor de nombres, DNS, para determinar quién acepta el correo para dicho dominio.
- El servidor de nombres para el dominio b.org contesta con una entrada MX que indica los servidores de intercambio de correo para el dominio; en el ejemplo este servidor es mx.b.org.
- Entonces smtp.a.org envía el mensaje a mx.b.org, utilizando SMTP. En el servidor de correo receptor se almacena el mensaje recibido en el buzón de entrada de Arcadio.
- Finalmente, Arcadio presiona el botón para obtener nuevos mensajes en su MUA, que recoge, en este ejemplo, el mensaje utilizando otro protocolo de internet conocido como POP3 (Post Office Protocol).

Ésta es una secuencia típica, aunque puede cambiar de un lugar o usuario a otro, ya que existe una variedad de protocolos distintos para acceder a los mensajes de correo; pueden incluirse opciones para cifrar o agregar una firma electrónica a los mensajes y su correspondiente validación, etcétera.

Formato

Aunque definir el mensaje que se envía por correo electrónico, de la misma manera que cuando se redacta una carta con pluma y papel, es una tarea intelectual, el formato de dicho mensaje se puede representar mediante una serie de estándares de internet. Algunos de estos estándares son llamados, de manera colectiva, MIME (Multipurpose Internet Mail Extensions) y permiten enviar mensajes que incluyen más que sólo texto, por ejemplo, imágenes, archivos de audio y video, y casi cualquier otro formato soportado por un sistema de cómputo.

Es tarea del cliente de correo o MUA del que envía el mensaje codificar el mensaje y los archivos anexos a éste, de manera tal que se apeguen a los estándares de formato. Un correo electrónico consta de dos partes esenciales:

- Encabezado: una lista de campos, tales como De, Para, Asunto y otra información acerca del correo.
- Cuerpo del mensaje: el mensaje que desea enviarse por correo electrónico, usualmente texto libre que puede contener una sección especial para la firma del usuario. Esta firma no es como una rúbrica, sino algún texto fijo que siempre se agrega a los correos que se envían, por ejemplo: el nombre y puesto en la empresa, o una cita textual interesante.

Spam, phishing y virus

Aunque el correo electrónico es una herramienta útil y, en el caso de muchas personas, el medio principal para comunicarse con colegas de estudio, compañeros de trabajo o la familia, hay ciertos fenómenos que atentan contra este medio de comunicación en internet: spam, phishing y virus o gusanos de correo.

Se conoce con el nombre de *spam* al correo comercial no solicitado y, hoy en día, se ha convertido en causa de incomodidad para los usuarios que lo reciben de manera constante. Como se trata de un medio relativamente barato, existen muchas personas y organizaciones que se dedican a enviar cientos de millones de correos basura o spam de manera cotidiana, por lo que resulta usual que una persona reciba en su correo electrónico desde un puñado hasta varios cientos de correos basura todos los días.

Phishing es una actividad criminal mediante la cual alguien intenta adquirir información sensible o confidencial —como nombres de usuario y contraseñas, detalles de tarjetas de crédito, etc.— de manera fraudulenta. Quienes realizan esta actividad lo hacen suplantando, en un mensaje electrónico o correo, la identidad de alguna entidad o institución, como un banco o un sitio de venta de bienes en línea. Para engañar a usuarios de tecnologías de la información, *phishing* aprovecha técnicas de una nueva área conocida como ingeniería social.

Los virus y gusanos de correo son programas malignos que se transmiten usualmente como archivos adjuntos a un correo electrónico. Por lo general, una vez que uno de estos gusanos logra infectar un equipo de cómputo, se replica de manera automática y se autoenvía utilizando correos electrónicos a todos los contactos de la persona cuya máquina fue infectada.

Curiosidades
La ingeniería social es una colección de técnicas para manipular a las personas para que realicen acciones o divulguen información confidencial. Aunque estas tareas son similares a las de un simple fraude, el término ingeniería social se utiliza exclusivamente cuando la trampa involucra almacenar información u obtener acceso para un sistema de cómputo; además, en muchos casos el atacante nunca se enfrenta cara a cara con la víctima.

7.6.2 Mensajería instantánea

La mensajería instantánea es una forma de comunicación en tiempo real, entre dos o más personas, basada en texto escrito. El texto es enviado vía computadoras conectadas en una red, como internet, pero a diferencia del correo electrónico, la mensajería instantánea ofrece una experiencia cercana a *hablar* con la otra persona. De hecho, existen diversas extensiones para la mensajería instantánea que además del texto permiten integrar audio y video. La mensajería instantánea cubre un nicho importante en distintas organizaciones para permitir que las personas colaboren eficientemente, sin tener que desplazarse de su lugar de trabajo a otra oficina o sin necesidad de utilizar el teléfono. Por supuesto, este tipo de comunicación es de las preferidas por familiares y amigos, porque resulta barata y eficiente.

Cómo funciona

Existen distintas implementaciones de mensajería instantánea, lo que significa que no hay un estándar de comunicación común y, por lo tanto, el funcionamiento preciso depende de cuál de las distintas implementaciones de mensajería se usa. Sin embargo, éstos son los componentes comunes a todas las implementaciones:

1] Para comunicarse con otra persona utilizando mensajería instantánea se necesita un cliente. En la actualidad hay clientes que soportan múltiples protocolos, por lo que es posible hablar con amigos en distintas redes.
2] Una vez que alguien se conecta a la red, se almacena en uno de los servidores de la red en los que se está en línea.
3] El mismo servidor le avisa al cliente cuáles de sus contactos están en línea y viceversa, a los que están en línea les avisa que el cliente se ha conectado.
4] A continuación se puede iniciar una conversación con alguno de los contactos.

La comunicación entre el cliente y los servidores de la red varían y cada uno tiene su propio protocolo. Las redes más comunes en la actualidad son: AIM (America Online), Skype, ICQ y Jabber. Este último, Jabber, está basado en el protocolo XMPP (Extensible Message and Presence Protocol), que permite que se agreguen funciones fácilmente, como la transmisión de archivos o audio, y es un protocolo abierto, como el del correo electrónico, lo que significa que cualquiera en internet puede instalar un servidor de Jabber. La mayoría de estas redes, así como los distintos clientes, son softwares libres (véase el apartado "Colaboración y software libre", más adelante).

7.6.3 Acceso remoto

Internet permite que los usuarios de computadoras puedan conectarse a otras computadoras a través de la red. Esta comunicación se puede llevar a cabo en una variedad de formas y puede incluir distintos mecanismos de seguridad, autenticación y permisos o autorización para realizar ciertas tareas. Esta facilidad está transformando la manera en que se organizan las empresas y organizaciones. Así, hoy en día es común que muchas personas trabajen desde casa y se conecten a las redes internas de sus empresas, con un ambiente de trabajo idéntico al que se obtiene en sus oficinas, pero con muchas ventajas: se reducen los costos para el empleado al no tener gastos de transporte, y para la organización, pues no tiene que contemplar un ambiente de trabajo adecuado para todos los miembros de la organización, los empleados no pierden tiempo en desplazarse a las oficinas, se pueden tener grupos de trabajo distribuidos en distintas ciudades en todo el mundo, ahorra en insumos, etcétera.

Por ejemplo, cuando un trabajador está lejos de su oficina, probablemente en un viaje de negocios o placer, puede ejecutar una sesión remota en su computadora de escritorio usual utilizando la red virtual privada (VPN, por sus siglas en inglés) de la empresa, a través de Internet. Para los administradores de estas VPN se incrementan los retos, pues ahora deben defender la red de la organización, incluyendo los ambientes remotos utilizados por los miembros de la organización en casa o sitios remotos.

7.6.4 Colaboración y software libre

Colaboración, en el estricto sentido de la palabra, es un proceso recursivo (sí, exacto, relativo a la recursividad que se revisó en el tema 2), en el cual dos o más personas trabajan juntas construyendo consensos y compartiendo conocimiento para alcanzar un fin que es, en la mayoría de las ocasiones, de naturaleza creativa. Los proyectos de colaboración no siempre requieren de un líder y pueden ofrecer mejores resultados que otras técnicas de organización, como la descentralización.

Los modelos estructurados de colaboración impulsan la introspección del comportamiento y comunicación, permitiendo que las posibilidades de éxito de los equipos o grupos que los utilizan se eleven considerablemente. En el contexto de internet que nos interesa aquí, se encuentran diversos ejemplos de modelos de colaboración bien estructurados.

Con la proliferación de aplicaciones en internet se han acortado y abaratado las distancias, por lo que hoy es posible compartir ideas, conocimiento y habilidades entre miembros de grupos distribuidos a lo largo del planeta. Esto no sólo ha facilitado, sino que ha promovido la formación de grupos para resolver problemas específicos o promover ideas.

7.6.5 Producción por comunes

A raíz de algunos proyectos exitosos de colaboración se han acuñado nuevos términos, como la *producción por comunes*, que describe el modelo de producción económica en el cual la energía creativa de muchos se coordina, usualmente con ayuda de aplicaciones en internet, para formar proyectos trascendentes, sin la jerarquía de una organización típica y sin compensación económica. Algunos ejemplos de productos creados utilizando producción por comunes son:

- Linux, un sistema operativo para computadoras. Se calcula que 12.7% de los servidores en el mundo utilizan Linux y, gracias a su adopción de muchas empresas y usuarios en el mundo, quienes comercializan Linux esperan que sus ganancias para 2008 superen los 35 000 000 000 de dólares.
- Slashdot, un sitio de noticias y anuncios. Este sitio recibe 5 500 000 visitantes por mes, y para muchos sitios tiene una consecuencia desafortunada que se conoce como *efecto slashdot*, ya que si una noticia en Slashdot hace referencia a una página web en un sitio determinado, en unas cuantas horas varios miles y algunas veces hasta millones de personas intentan acceder al sitio, lo que puede saturar el servidor.
- Wikipedia, una enciclopedia en línea. Desde su creación en 2001, ha crecido para convertirse en uno de los principales sitios de referencia en la web. Colaboran en la edición de sus más de 5 300 000 artículos alrededor de 75 000 voluntarios en todo el mundo.

Curiosidades
Uno de los ejemplos más significativos de proyectos colaborativos es el movimiento del software libre, que ha permitido crear GNU y Linux, partiendo de cero. El software libre hace referencia a productos de software que pueden ser utilizados, estudiados y modificados sin restricción. GNU es un proyecto masivo de colaboración para producir software libre. Junto con el núcleo Linux, otro producto software libre, han logrado ofrecer un sistema operativo completo, moderno y que ha impulsado una gran cantidad de cambios en la industria del cómputo mundial.

7.7 RESUMEN

Uno de los retos más importantes de la computación moderna es el cómputo distribuido, ya sea para programar supercomputadoras o las recientes computadoras personales con varios procesadores, o multicore. Se revisaron los principios esenciales del cómputo distribuido, el paralelismo, la exclusión mutua como un mecanismo para sincronizar el acceso a un recurso compartido. En el fondo de estos temas, se encuentran la comunicación, los protocolos, las redes de computadoras, internet, la web y las principales aplicaciones para comunicación basadas en internet.

MULTIMEDIA

TEMA

8

© *Latin Stock México.*

La esperanza es que, dentro de no muchos años, cerebros humanos y computadoras estén íntimamente acoplados, y que la sociedad resultante piense como ningún cerebro humano ha pensado.

J. C. R. LICKLIDER, 1960.

La música ha mostrado ser una de las más poderosas fuerzas que han forjado a la computación.

NICHOLAS NEGROPONTE, 1995.

8.1 INTRODUCCIÓN

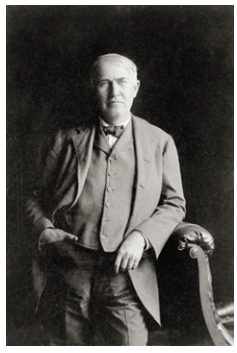
Hace unos 20 000 años, la primera experiencia multimedia fue creada a partir de las cuevas subterráneas, como la de Lascaux en la región de la Dordogne en el sur de Francia. El hombre de Cro-Magnon decoró con hermosísimos murales estas cavernas resonantes, iluminadas con tintineantes velas con olor a grasa animal, donde se realizaban rituales en un teatro mágico de los sentidos.

En efecto, la comunicación y la transmisión de información son fundamentales para garantizar la supervivencia de las especies. Por ejemplo, especies como las abejas intercambian información mediante movimientos, para indicar el lugar donde se encuentra el polen. Otras especies usan un lenguaje químico y muchas otras, durante el cortejo, usan variados medios de comunicación, en una suerte de danza en la que exhiben sus habili-

dades o por diferentes aromas. Las luciérnagas, por ejemplo, usan bioluminiscencia para proporcionar información sobre su disponibilidad al apareamiento. En fin, la transferencia de información entre seres vivos es fundamental para su supervivencia y reproducción. En la especie humana, la transmisión del conocimiento ha sido un elemento básico para garantizar su supervivencia y desarrollo. Existen muchos medios para representar y transmitir el conocimiento; entre los más importantes se encuentran el lenguaje y la escritura. También existen muchos medios para comunicarse basados en los sentidos: el texto, el sonido, el video, la animación y otros más específicos como la lectura de códigos por el tacto que usan las personas invidentes. El texto es un medio básico de transmisión de información y es innegable el poder de la palabra escrita. Por medio de escritos se puede conocer la historia, la forma de pensar de otras personas, diferentes culturas y costumbres. La transmisión de la información y del conocimiento a través de la escritura ha sido una piedra angular en el desarrollo de la humanidad. La electrónica y la computación facilitan este medio de comunicación.

A la combinación simultánea de diferentes medios de comunicación y transferencia de información se le conoce como *multimedia*. La multimedia es una manera interactiva y eficaz de combinar simultáneamente, por ejemplo, texto, sonido, video y animación.

Thomas Edison (1847-1931). Inventor y empresario estadunidense que desarrolló dispositivos de gran influencia en la vida mundial, como el fonógrafo y el foco de larga duración.



© Latin Stock México.

8.2 TEXTO

El texto es una secuencia de signos de un alfabeto codificado en un sistema de escritura; como se mencionó anteriormente, la escritura ha sido un medio fundamental de comunicación. En la actualidad, con el uso de las computadoras, los procesos de generación, edición y producción de texto se han agilizado.

Cada símbolo o letra en la computadora tiene una representación en bits, por medio de un código, como el ASCII (American Standard Code for Information Interchange). De esta forma, cada letra o símbolo dentro de la computadora es una secuencia de bits.

Concepto
Una grabación digital convierte la señal analógica del sonido en un flujo de números discretos que representan los cambios en la presión del aire a través del tiempo.

8.2.1 Diseño de tipos

Si en algo se ha insistido a lo largo de este libro es en aclarar que existen problemas que no se pueden resolver con sólo incrementar la capacidad de cómputo, como se explicó en el primer tema. El lector no debe quedarse con la impresión de que para mejorar la resolución en las pantallas de las computadoras sólo hace falta un procesador de gráficos más poderoso. Por el contrario, en años recientes han surgido muchos avances en este frente gracias al diseño de tipos. Por desgracia, la experiencia de siglos de los diseñadores de tipos para imprentas no puede aplicarse directamente a los tipos de computadora. Desde los años ochenta, una **fuente** digital está constituida por miles de matrices de 0 y 1 para cada letra, número y símbolo, así como para cada tamaño, estilo y resolución.

Esto, por supuesto, ha evolucionado, por lo que en la actualidad las fuentes de computadora utilizan contornos escalables para cada letra, en lugar de una matriz para cada posibilidad de letra. Para desplegar una letra en pantalla, el software que maneja las fuentes toma el contorno de la letra deseada, lo adapta al tamaño adecuado y, casi al instante, crea una matriz de la letra. Todo esto sucede en un tiempo de 10 a 20 milisegundos, desde que se presiona la tecla hasta que se ve en pantalla.

Concepto
Una fuente es un conjunto de imágenes que representan los caracteres de un estilo tipográfico particular. La tipografía es el arte y la técnica del diseño, manejo y selección de tipos.

Hinting contra antialiasing

Debido a que es un programa el que genera la imagen apropiada para cada letra, pueden suscitarse inconsistencias. Por ejemplo, las serifas (pequeños trazos ubicados generalmente en los extremos de las astas de los caracteres tipográficos, también llamados patines) pueden mostrarse en pantalla con un ancho diferente para distintas letras, unas con un elemento de resolución de ancho y otras con dos. Para resolver este tipo de problemas, se han desarrollado distintas técnicas. Una de las primeras se conoce como *hinting*, que sirve para distorsionar levemente la imagen, de modo que luzca mejor en pantalla.

Aplicar un *antialias* a la imagen de una letra tiene la intención de hacer que ésta luzca muy suave en la pantalla, que tenga una apariencia pareja, similar al texto impreso en la página de un libro. Pero, por supuesto, una desventaja importante en este caso es que la letra resulta menos clara.

8.3 SONIDO

El sonido es una perturbación de energía mecánica que se propaga como una onda a través de la materia, por lo que tiene las siguientes propiedades: frecuencia, amplitud de onda, periodo, amplitud y velocidad. La ciencia que estudia el sonido se conoce como acústica. Desde el punto de vista de la computación, y en particular de la multimedia, el reto se encuentra en la grabación y reproducción del sonido.

Curiosidades
La frecuencia de muestreo es uno de los factores más importantes en las grabaciones digitales, ya que si esta frecuencia es muy baja, el sonido original no puede reconstruirse a partir de la señal de muestreo. Para evitar problemas, se utiliza el doble de la frecuencia más alta en la secuencia de sonido original, como frecuencia de muestreo.

Durante más de cien años, desde la aparición del fonógrafo inventado por Edison en 1877, se ha grabado el sonido; sin embargo, para fines prácticos, el tipo de grabación que interesa en computación es la **grabación *digital*** del sonido. El procedimiento para llevar a cabo la grabación y reproducción digital del sonido se detalla a continuación:

Grabación

- La señal analógica se transmite desde el dispositivo de entrada a un convertidor analógico digital (ADC, por sus siglas en inglés).
- El ADC convierte esta señal en una serie de números binarios. La cantidad de números que se producen por segundo se conoce como frecuencia de muestreo (*sample rate*).
- Los números se transmiten a través de cables a un dispositivo de almacenamiento digital en la computadora, por ejemplo al disco duro o a un quemador de discos compactos.

Reproducción

- La secuencia de números se transmite desde el dispositivo de almacenamiento al convertidor digital analógico (DAC), que convierte los números en sonido.
- Se transmite el sonido a las bocinas.

8.3.1 Almacenar bits

Tanto para audio como para video, el principal reto consiste en almacenar los datos, pues aun si se convierte el sonido en números, se requiere un esquema lo suficientemente rápido como para no perder información. Por ejemplo, para grabar una canción en sonido estéreo (dos canales) con una frecuencia de muestreo de 44.1 kHz y un **tamaño de palabra**

de 16 bits, el software y el hardware de grabación deben ser capaces de operar 1 411 200 bits por segundo.

Una de las principales ventajas de la grabación digital sobre la analógica es que se pueden corregir errores en la secuencia de audio. Por ejemplo, si se modifica uno de los números en la secuencia, por un error físico en el medio de almacenamiento, el software de reproducción puede detectar que lo que leyó debía ser un 1 y corregirlo. En las grabaciones analógicas, los errores se hacen más graves con el tiempo. Existen muchas técnicas para corrección de errores que no se abordarán aquí por razones de espacio.

Concepto
El tamaño de palabra es el número de bits utilizado para representar una onda de audio única. El tamaño de palabra afecta directamente la distorsión. En la actualidad se utiliza como límite práctico un tamaño de palabra de 24 bits, ya que su relación señal-a-ruido excede el de la mayoría de los circuitos analógicos, que son utilizados en (al menos) dos puntos en el proceso de grabación y reproducción del sonido.

8.4 IMÁGENES Y VIDEO

El video, un medio muy importante de comunicación, consiste en la grabación de imágenes. En este tema, primero se presentará una introducción sobre las imágenes digitales y, posteriormente, en la parte de animación, se enriquecerán los conceptos relacionados con este tema de estudio.

Un antiguo proverbio afirma que una imagen dice más que mil palabras. En verdad se cree que así es: una imagen posee un gran contenido de información. En ella se pueden ver la forma, colores y texturas de los objetos, así como las relaciones entre ellos. La visión es el sentido más avanzado que poseen los seres humanos. Por ende, no es de extrañar que las imágenes sean el objeto de estudio más importante de la percepción humana.

8.4.1 Imágenes digitales

Una **imagen** digital está compuesta por un número finito de elementos, cada uno con un lugar y valor específicos. Estos elementos de forma cuadrada son llamados pixeles (*picture elements*). Por ejemplo, una imagen de diferentes tonos de gris queda representada por una matriz; o sea, un arreglo rectangular de números, donde cada elemento o pixel de la imagen mantiene un valor numérico que indica el valor de intensidad o tono de gris en ese punto. Cuando se desea representar texto, solamente existen dos valores de intensidad de gris: el blanco y el negro. A este tipo de imágenes se les llama imágenes binarias. Una imagen binaria es aquella que mantiene dos valores de intensidad: 0 y 1, es decir que sus intensidades sólo mantienen dos estados. Este tipo de representación es muy común en textos o símbolos donde sólo se desea indicar su silueta ausente de color. La figura 1 muestra un ejemplo de una imagen binaria, el escudo de la UNAM.

Figura 1. Ejemplo de una imagen binaria: el escudo de la UNAM.

Una vez representada una escena por medio de una imagen digital, es muy simple realizar otro tipo de representaciones. Por ejemplo, la mostrada en la figura 2 (que utiliza el tipo de información de la figura 1), donde a cada pixel "negro" se le confirió una altura previamente definida; posteriormente, se colocó una malla sobre la imagen y se proyectó en perspectiva.

Concepto
En términos formales se puede definir una imagen como una función bidimensional $f(x,y)$, donde x y y son las coordenadas cartesianas y f la amplitud de la intensidad o nivel de gris de la imagen en ese punto de coordenadas. Cuando los valores de las coordenadas y el valor de la amplitud son todos finitos, es decir, poseen cantidades discretas, se trata de una imagen digital.

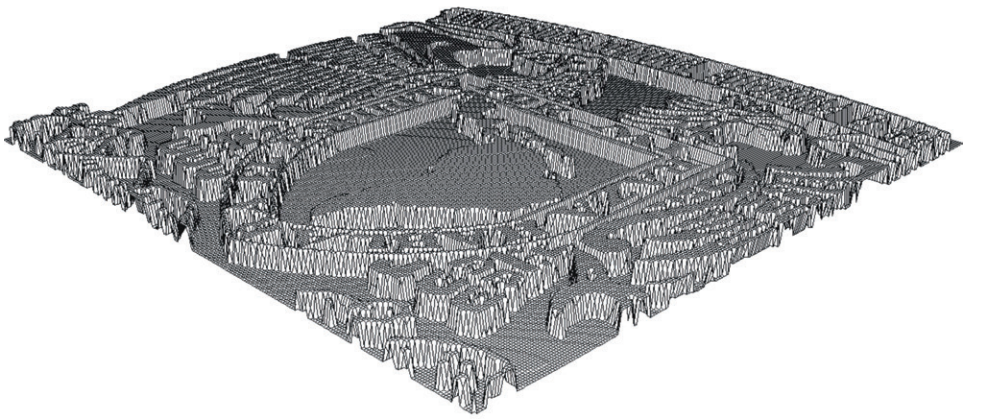


Figura 2. Otra representación de una parte de la imagen binaria mostrada en la figura 1 | © Ernesto Bribiesca.

Figura 3. Ejemplo de una imagen digital | © Ernesto Bribiesca.

Curiosidades

La habilidad del ojo humano para distinguir colores se basa en la sensibilidad de las células retinianas a la luz. La retina tiene tres tipos de células receptoras de color o conos. El primer cono responde a lo que conocemos como violeta (o longitudes de onda de 420 nm), el segundo amarillo-verdoso (564 nm) y el tercero a verde (534 nm). Incorrectamente, estos conos son conocidos como: azul, rojo y verde, respectivamente. Se estima que el ojo humano puede distinguir 10 millones de colores. Curiosamente, el modelo de color más común es: RGB, siglas en inglés de rojo, verde y azul, para televisiones de rayos catódicos, de plasma y de cristal líquido.

La figura 3 muestra un ejemplo de imagen digital a colores, donde la combinación de los colores básicos —rojo, verde y azul— permite obtener una gran variedad. Esta imagen digital está formada por 640 × 480 pixeles, es decir, 640 elementos en cada una de las 480 líneas.

Si se observa con detenimiento la figura 3, se puede extraer una gran cantidad de información. Entre las principales características se tienen: el color, es posible distinguir claramente los diferentes colores en la imagen; la forma de cada uno de los elementos y su textura (es posible distinguir los diferentes tipos de texturas presentes); el contexto de información, esto es, las relaciones que mantienen los objetos dentro de la imagen; y, finalmente, la experiencia permite identificar la mayor parte de estos elementos. La cantidad de información que se puede extraer de una imagen es impresionante.

La figura 3 representa una planta conocida como buganvilia. A primera vista, se distinguen con facilidad las flores y las hojas de la planta; se pueden ver los diferentes colores y tonalidades del verde de las hojas; al fondo se observan algunas ramas de la planta; en las flores se distinguen con facilidad las partes de las mismas, como pistilos y polen, además de los distintos matices de color rosa de los pétalos.

En resumen, el color y la forma son fundamentales en el reconocimiento. Sin embargo, también es posible ver fácilmente las texturas de los elementos de la imagen, por ejemplo, la suave textura de las hojas. Asimismo, se pueden apreciar la iluminación y las sombras.

Las relaciones entre los diferentes elementos de una imagen son muy notorias. Una parte muy importante en todo reconocimiento es la experiencia, es decir, el conocimiento previamente acumulado, producto de experiencias anteriores. ¿Qué otras cosas se pueden decir de la planta de la figura 3?

Se puede afirmar con seguridad que se trata de una planta viva, no parece deshidratada o disecada; se ve saludable, ya que posee hojas simétricas y sus formas son armónicas; su iluminación es natural, producto de la luz solar (incluso se puede intentar adivinar a qué hora fue fotografiada); en fin, se puede apreciar su belleza y reconocer algunas de sus

características, pero si esta descripción la hiciera un botánico, seguramente tendría mucho más que comentar, ya que su experiencia es distinta. Como dato adicional, la buganvilia es originaria de América del Sur y, lo que comúnmente se llaman flores, en realidad son hojas modificadas que se denominan *brácteas*. Las verdaderas flores son diminutas, blancas y están rodeadas por las brácteas. ¿Será posible encontrar una flor en la figura 3?

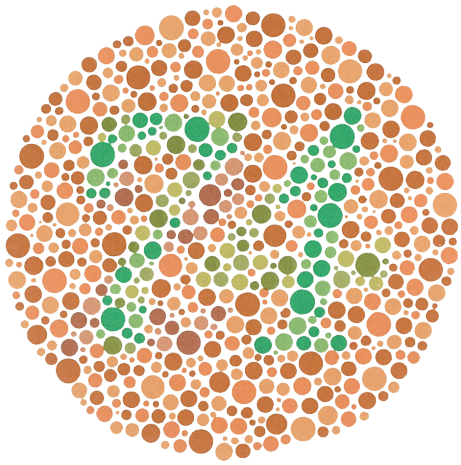


Figura 4. Ejemplo de plato con el número 74 para la prueba de Ishihara | © Anónimo.

Como se mencionó, un elemento de resolución es el pixel, el cual tiene forma cuadrada. Sin embargo, no es el único; existen también otros elementos de resolución: los polígonos, los hexágonos y los triángulos, por ejemplo. Los polígonos deben llenar el plano, es decir, no deben existir espacios entre ellos. La figura 5(a) muestra un objeto; en la (b) se aprecia el objeto representado por pixeles; en la (c), por hexágonos y, finalmente, en la (d), por triángulos.

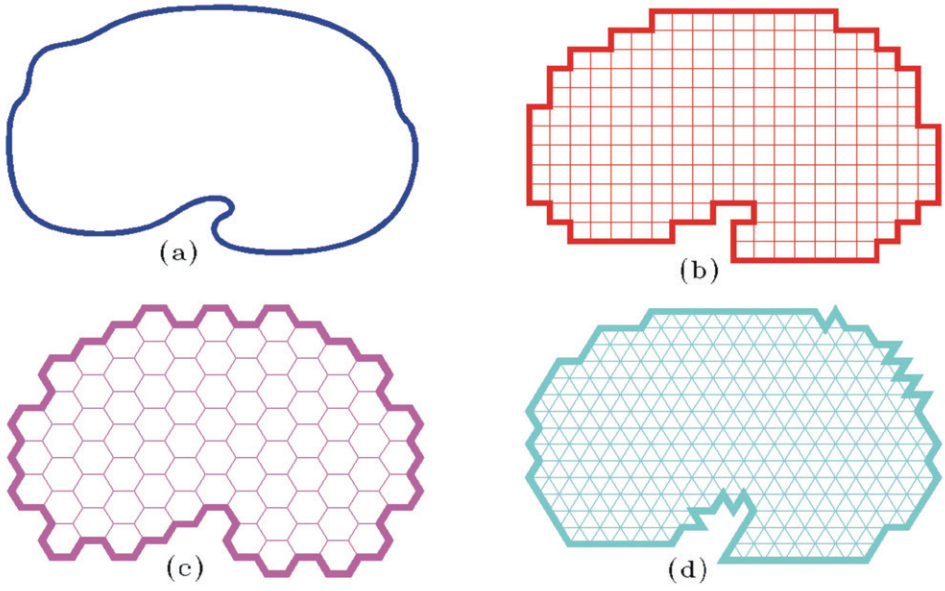


Figura 5. Un mismo objeto representado por diferentes polígonos regulares.

Cada una de estas representaciones posee ventajas y desventajas, aunque la representación que se usa más, por lo general, es la que se realiza por medio de pixeles. No obstante, de vez en cuando aparecen en la literatura científica artículos que muestran imágenes representadas por hexágonos. Cabe señalar que este tipo de celosías es común en la naturaleza: un panal de abejas tiene una estructura formada por hexágonos para optimizar el espacio; las células retinianas del ojo humano con conos y bastones forman estructuras hexagonales.

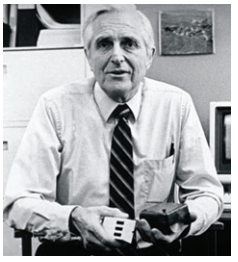
8.4.2 Adquisición de imágenes

Curiosidades
La percepción o visión es el término que utiliza la psicología para definir la habilidad de interpretar la información, en forma de luz visible, que llega al ojo.

La adquisición de imágenes se logra a través de diferentes medios. Generalmente, se tiene una escena del universo físico y un sensor que la captura. Existe una gran variedad de sensores que trabajan en diferentes rangos, a esto se debe la diversidad y riqueza de las imágenes. Una ventaja de la adquisición de imágenes es que no es necesario hacer contacto físico con el objeto de estudio, es decir, que se puede capturar la información a distancia.

Lo que captura el sensor son las manifestaciones de energía de la escena, la energía que refleja un objeto. Entre los principales tipos de sensores se encuentran los que capturan la luz reflejada por el objeto en diferentes rangos del espectro electromagnético. Por ejemplo, es necesario que la escena esté iluminada para que las cámaras fotográficas y digitales de uso común puedan captar las imágenes. Sin embargo, existe una variedad considerable de manifestaciones de energía de los objetos, como los sensores térmicos, capaces de detectar el calor emitido por un cuerpo. En este sentido, la naturaleza también posee esta tecnología, ya que las víboras son capaces de detectar a sus víctimas por el calor que emiten, independientemente de si es de día o de noche.

Douglas Engelbart
(1925). Inventor estadunidense de origen sueco y noruego, creador del ratón de computadora, un dispositivo apuntador en un sistema de cómputo estándar. Engelbart fue uno de los pioneros en el área de interacción humano-máquina y trabajó en el equipo que desarrolló el hipertexto, las redes de computadoras y una de las primeras interfaces gráficas de usuario. En el año 2000, fue condecorado con la Medalla Nacional de Tecnología, que es la máxima presea en materia de tecnología que entrega el gobierno de Estados Unidos.



© Doug Engelbart Institute.

Otro tipo de sensor capta los campos magnéticos o la radiactividad que emiten los objetos, lo que da como resultado, por ejemplo, las imágenes médicas, donde se utilizan técnicas de tomografías PET (por emisión de positrones). Hay una gran variedad de sensores y una abundante cantidad de imágenes, pero finalmente el objetivo es generar modelos en dos y tres dimensiones para que posteriormente sean clasificados y analizados por los expertos.

8.4.3 Algunas características de las imágenes digitales

Sin importar el tipo de sensor utilizado para obtener la imagen, es necesario almacenarla y representarla de alguna manera; para esto, existen diferentes maneras de describir dicha imagen. A continuación se detallan las más comunes y sus principales usos.

Imágenes en formato vector

Una de las formas más comunes para describir imágenes es utilizar números: posición y tamaño de formas geométricas, ya sean líneas, rectángulos, círculos y curvas. A este tipo de imágenes se les denomina *imágenes de vector*; en ellas está almacenada toda la información geométrica del objeto de estudio: sus coordenadas. Esto permite cambios de escala fácilmente, de ahí que sea el formato preferido por los diseñadores y arquitectos que usan software asistido por computadora como apoyo para sus diseños. Para representar curvas se utiliza un sistema coordenado o espacio de usuario, llamado así porque es el lugar donde los usuarios pueden colocar y relacionar entre sí los elementos de la imagen.

Imágenes en formato raster o bitmap

Una imagen en formato *raster* es una rejilla rectangular de pixeles. Este formato presenta varias ventajas y se usa bastante. Pero es importante resaltar que no se trata de un forma-

to sólido para cambios de escala, ya que los cambios que se realicen deben ser en múltiplos de pixeles, pues los valores intermedios ocasionan pérdida de información.

Existe una gran variedad de formatos de archivos para almacenar imágenes *raster* o *bitmap*. En la tabla 1 se muestran los más comunes con una breve descripción:

Nombre	Extensión	Descripción
Join Photographic Experts Group	.jpg	Formato muy utilizado para fotografías, tiene pérdida de información.
Portable Network Graphics	.png	Imagen sin pérdida de información, soporta 16 bits de profundidad.
Graphics Interchange Format	.gif	Formato indexado de 8 bits, PNG es un superconjunto de éste. Soporta animación.

Tabla 1. Formatos de archivos para bitmaps.

Megapixeles

¿De qué tamaño es una imagen digital? Si se piensa en una imagen como una matriz de pixeles, entonces una imagen digital de 2 048 × 1 535 tiene 3 143 680 pixeles o aproximadamente tres megapixeles. Por ejemplo, la buganvilia de la figura 3 es una imagen de 640 × 480 = 307 200 pixeles o 0.3 megapixeles. En la tabla 2 se muestran resoluciones típicas; algunas son utilizadas en televisiones, en monitores o en cámaras fotográficas y de video caseras.

Dimensiones	Megapixeles	Nombre	Comentario
640 x 480	0.3	VGA	VGA
800 x 600	0.4	SVGA	
1024 x 768	0.8	XGA	Dimensión más común de las pantallas de computadora en la actualidad.
1280 x 960	1.2		
1600 x 1200	2.1	UXGA	Videoproyectores.
1920 x 1080	2.1	1080i HDTV	Formato para televisión digital de alta definición.

Tabla 2. Resoluciones típicas.

Curiosidades
La mayoría de las imágenes digitales se comprimen. Existen varios tipos de compresión, con o sin pérdida de información. ¿Quién en su sano juicio querría perder información (o sea, los detalles) en una imagen? Resulta que el ojo humano puede esconder o ignorar imperfecciones y, por eso, aunque se quite detalle a las imágenes, el cerebro prácticamente no lo detecta. JPG, uno de los formatos más utilizados, tiene pérdidas de información.

Profundidad de color y compresión de imágenes

El valor de los pixeles se almacena en la memoria de la computadora; por lo tanto, la información se guarda de manera binaria y la continuidad espacial de la imagen se aproxima por los espacios en la malla de color (*sample depth*). Esto se conoce como profundidad de color, y existen varias profundidades muy utilizadas:

- *8 bits*. Esta profundidad utiliza enteros de ocho bits, que pueden representar hasta 256 valores: $2^8 = 256$. Representan los niveles de brillo.
- *12 bits*. Para imágenes muy detalladas, tanto en brillo como en sombras. Algunas cámaras digitales, sobre todo las profesionales, utilizan el formato *raw*, que es de 12

Concepto
En matemáticas, el concepto de curva intenta capturar la idea intuitiva de un objeto continuo y unidimensional. En el estricto sentido, aunque esto ya no parezca obvio, una línea recta también es una curva.

bits y soporta $2^{12} = 4\,096$. Permiten gran precisión para almacenar la imagen sin perder detalle.

- *16 bits*. La mayoría de los programas para manipulación de imágenes y los formatos comunes PNG y TIF realizan operaciones a 16 bits en imágenes de ocho bits para evitar perder información.

Existen otras profundidades que poseen mayor precisión, pero que se utilizan en aplicaciones muy especiales, como el cine. Por supuesto, la precisión y forma de representación de estas profundidades implica un costo. Las imágenes bitmap utilizan mucho espacio en memoria. Por ejemplo, una imagen de 2.1 megapixeles a colores —con el modelo RGB en ocho bits— utiliza: $1\,600 \times 1\,200 \times 3$ bytes = 5 760 000 bytes = 5.7 megabytes.

8.4.4 Segmentación de imágenes

La segmentación de imágenes es la partición de una imagen en regiones homogéneas u objetos. La forma de detectar regiones homogéneas que representan un objeto de la imagen es la clasificación. Existen varios tipos de clasificación: de color, de contorno, etcétera.

Probablemente, la forma más obvia de clasificación en una imagen se realiza basándose en su color, a través de la identificación de regiones que pertenecen a un color o a un rango de color determinados. Dado el color de un objeto previamente identificado, ¿qué otros objetos de la imagen tienen el mismo color, es decir, son de la misma clase?

Otra técnica de detección de regiones homogéneas u objetos se realiza por medio de la detección de sus contornos. Una forma de detectar un contorno dentro de una imagen digital consiste en calcular la diferencia entre pixeles vecinos. Se define un umbral o valor de diferencia entre pixeles y se calculan para toda la imagen.

En una estructura cuadriculada hay dos formas de vecindad y conectividad entre los pixeles: la conectividad 4, que sólo considera a los pixeles vecinos de la izquierda, derecha, arriba y abajo; y la conectividad 8, donde se consideran los vecinos antes mencionados, más los cuatro vecinos de las esquinas. La figura 3 fue segmentada por medio del cómputo de la diferencia entre pixeles vecinos usando conectividad 8 y definiendo un umbral de diferencia. Como resultado de esta segmentación se obtuvo la imagen segmentada mostrada en la figura 6.

Figura 6. Detección de contornos en la imagen digital mostrada en la figura 3 | © Ernesto Bribiesca.

Esta imagen permite observar los contornos como límites de regiones homogéneas o de objetos; por ejemplo, se pueden identificar claramente los contornos de las hojas de la planta, tanto las verdes como las rosas, así como los contornos de los pistilos y otros elementos. En este ejemplo de segmentación se intentó conservar los colores originales.

8.4.5 Representaciones y descripciones de objetos

Una vez segmentada la imagen es posible extraer las regiones u objetos de la misma, los cuales corresponden a objetos en dos y tres dimensiones. Además, dependiendo de la imagen de procedencia, pueden ser curvas, superficies o sólidos.

Representaciones de curvas

Hay diferentes sistemas de coordenadas: cartesianas (rectangulares), polares, esféricas, cilíndricas, entre otras. Cada una de estas representaciones tiene ventajas y desventajas, por lo que deben seleccionarse de acuerdo con las necesidades del problema de estudio.

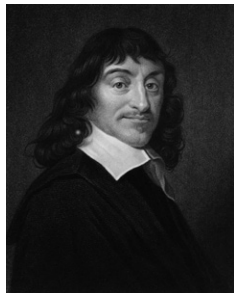
En la práctica, el sistema de coordenadas cartesianas es muy útil para determinar, por medio de dos números, cualquier punto en el plano. Para definir estas coordenadas se utilizan dos líneas perpendiculares dirigidas: el eje de las *x* y el eje de las *y*, así como una unidad de longitud, usualmente enteros. Pero, ¿para qué sirven los sistemas de coordenadas en este contexto? A través de un sistema coordenado, podemos describir figuras geométricas, utilizando ecuaciones algebraicas, es decir, ecuaciones que son satisfechas por los puntos en la figura.

Es importante considerar que cuando se diseñaron las representaciones de curvas por medio de coordenadas no se contaba con computadoras. En la actualidad se cuenta con representaciones nuevas, basadas en las anteriores, pero con un enfoque computarizado. Por ejemplo, para las imágenes digitales todo se mide en múltiplos de pixeles, en cantidades enteras.

Cualquier **curva** puede digitalizarse y representarse por medio de sus coordenadas. La figura 7 muestra un ejemplo de una curva continua y su digitalización. Sus coordenadas son introducidas a la computadora y después desplegadas en un dispositivo de salida.

Es importante que las representaciones sean compactas, que permitan la transferencia de información con facilidad y que sean lo más estándar posible para su incorporación a diferentes sistemas. Las representaciones interesantes de curvas, si se considera lo antes mencionado, corresponden a los llamados códigos de cadenas.

René Descartes (1596-1650). También conocido como Renatus Cartesius, fue un filósofo, matemático, científico y escritor francés. Se le considera el "padre de la filosofía moderna" e, incluso, "padre de las matemáticas modernas".



© Latin Stock México.

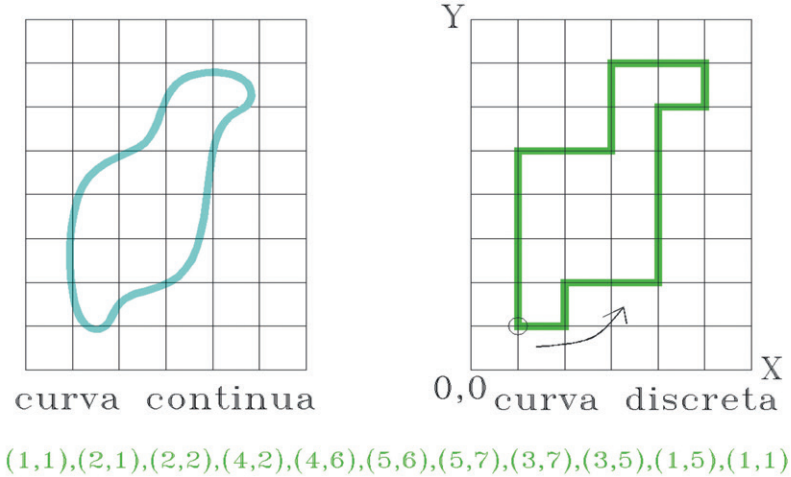


Figura 7. Cualquier curva puede ser digitalizada.

Curiosidades
Si se calculan todas las combinaciones de los elementos de las cadenas de Freeman, digamos de 10 elementos, y sólo se consideran las formas cerradas sin cruces internos e invariantes a la rotación, se obtendrán todas las formas cerradas a esa resolución.

Código de cadenas de Freeman

En 1961, el profesor Herbert Freeman[1] de la Universidad Rutgers propuso un método para representar curvas o contornos de regiones en el plano a través de un código de cadenas. En la figura 8(a) se aprecian las diferentes direcciones de los elementos de las cadenas de Freeman. Los elementos van del 0 al 7: los elementos 0, 2, 4 y 6 tienen una longitud de uno; y los elementos 1, 3, 5 y 7, una longitud de $\sqrt{2}$. Todas sus posiciones son discretas, es decir, inician y terminan en un vértice de la cuadrícula. De ahí que cualquier curva en el plano, ya sea abierta o cerrada, se pueda representar usando este código.

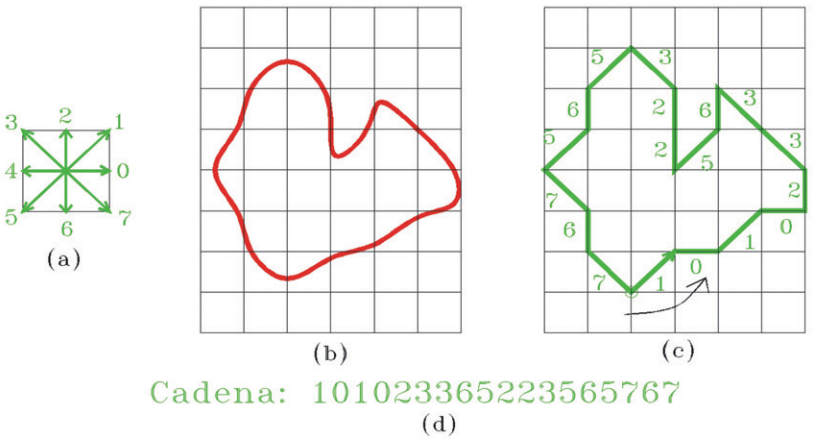


Figura 8. Cadenas de Freeman: (a) elementos; (b) ejemplo de curva cerrada simple; (c) elementos de la cadena de la curva ya "normalizada" en (b); (d) cadena de la curva.

La figura 8(b) presenta un ejemplo de una curva continua en el plano; en (c) se muestra su versión digital en una resolución previamente definida, es decir, al tamaño del cuadro considerado. Es fácil notar que los vectores de las cadenas de Freeman se aproximan lo más posible al contorno debido a su cercanía con los vértices de la cuadrícula. Como se puede observar, la curva discreta mostrada en la figura 8(c) está compuesta por 18 segmentos de línea recta y representada por la cadena: 101023365223565767.

El origen de la curva en la figura 8(c) está representado por una pequeña circunferencia. Con la información de la cadena, las coordenadas (x,y) del origen de la curva y el tamaño de la cuadrícula es posible reconstruir la curva discreta sin perder información.

Una característica importante de este código, que se puede observar directamente, es su capacidad de compresión sin pérdida de información. Si se representa cada elemento de la cadena por tres bits, los ocho elementos de la cadena en forma binaria quedarían como: 000, 001, 010, 011, 100, 101, 110 y 111, respectivamente. Lo anterior proporciona una forma compacta de representar formas de objetos o curvas.

Código de cadenas de vértices

En 1999, en la Universidad Nacional Autónoma de México se desarrolló el código de cadenas de vértices (Vertex Chain Code, VCC).[2] Este código de cadenas se basa en el princi-

[1] Herbert Freeman, "On the Encoding of Arbitrary Geometric Configurations", *IRE Transactions on Electronic Computers, EC-10*, 1961, pp. 260-268.
[2] Ernesto Bribiesca, "A New Chain Code", *Pattern Recognition*, 32, 1999, pp. 235-251.

pio de imágenes digitales. Cada objeto o región está compuesto por un número finito de pixeles conectados unos a otros. Si la imagen usa pixeles cuadrados, el VCC utiliza conectividad 4 (aunque también es posible utilizar celdas triangulares o hexagonales), como se muestra en la figura 9.

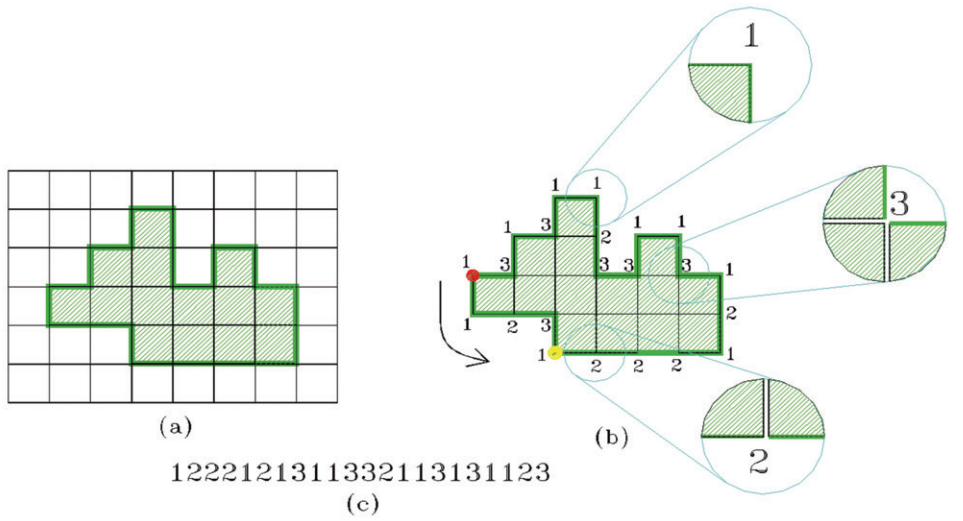


Figura 9. El código de cadena de vértices: (a) ejemplo de una región compuesta por pixeles; (b) generación del VCC; (c) cadena de elementos de la región presentada en (a).

El VCC representa el número de vértices de los pixeles en la forma en que tocan el contorno de la región. Observa la figura 9: en (a) se tiene un ejemplo de región compuesta de 14 pixeles. Los elementos del VCC aparecen en (b), donde es fácil apreciar que el elemento 1 corresponde a un vértice que toca un pixel del contorno; el 2 corresponde a dos vértices de dos pixeles, y el 3 a tres vértices de tres pixeles de la región. En esta forma, se generan los elementos del VCC para cualquier región.

La figura 9(c) muestra la cadena considerando el origen (representado por un punto) y sentido indicado por la flecha, lo que produce la cadena: 122212131133211313123. Es importante resaltar que esta cadena no se basa en símbolos, como en el código de cadenas de Freeman, que bien podrían ser letras o números. En el VCC los elementos de la cadena indican propiedades intrínsecas de la forma que representan; nos señalan el número de vértices que toca el contorno: ¿cuántos son rectos?, cantidad de los 2; ¿cuántos son cóncavos?, cantidad de los 3; ¿cuántos son convexos?, cantidad de los 1. Así, el VCC permite tener un descriptor de forma que indica las propiedades geométricas de ésta.

Es conveniente tener siempre una misma cadena para cada forma o región. Arbitrariamente se seleccionó un punto como origen. Sin embargo, ¿qué hubiera pasado de haber seleccionado otro origen? Se produce otra cadena. Para generar una cadena independiente del origen se recomienda seleccionar el origen que corresponda a la cadena mínima. Cada origen diferente alrededor de la curva genera una cadena distinta. Entonces, se considera a cada cadena como un número entero y se selecciona aquella que representa el número menor. En caso de que existan dos cadenas mínimas iguales, se deberá agregar una definición más para describir un origen único.

La cadena en la figura 9(c) está compuesta por 22 elementos. De esta manera, se generan 22 cadenas, las cuales se presentan a continuación:

12221213113321131311231
22212131133211313112312
22121311332113131123122
21213113321131311231222
12131133211313112312221
21311332113131123122212
13113321131311231222121
31133211313112312221213
11332113131123122121311
13321131311231222121311
33211313112312221213113
32113131123122121311331
21131311231222121311332
11313112312221213113321
13131123122121311332113
31311231222121311332113
13112312221213113321131
31123122212131133211313
11231222121311332113131
12312221213113321131311
23122212131133211313112
31222121311332113131231

De todas estas cadenas se selecciona la de menor valor y ésa es la que representa la región invariante al origen. Esta cadena se indica con números en negritas. Igual que en el código de cadenas de Freeman, en el VCC se puede obtener una importante capacidad de compresión sin perder información. Si se representa cada elemento de la cadena para una celosía de cuadros en forma binaria, se obtiene para sus tres elementos: 01, 10 y 11, respectivamente. Sólo se emplean dos bits para representar cada uno de los elementos de la cadena.

Representaciones de curvas tridimensionales

Curiosidades
A las curvas 3D representadas en este libro se les ha dado un grosor, es decir, son mostradas como cuerdas. Esto proporciona una mejor visualización y comprensión: se puede ver su profundidad y otras características.

Ahora se tratará la representación de las curvas en el espacio o tridimensionales (curvas 3D). La representación de curvas 3D es un tópico de gran importancia en diferentes áreas del conocimiento, ya que se presentan de manera natural en muchas situaciones, como por ejemplo en la trayectoria que describe un avión o un pájaro al volar, la doble hélice de la cadena del ADN, la trayectoria que describe una planta como la enredadera, etcétera.

Como se mencionó anteriormente, una **curva** es una configuración unidimensional, y una forma natural de representarla es por medio de sus coordenadas cartesianas, sólo que ahora se utilizarán tres ejes de referencia: x, y, z. Cualquier curva 3D puede ser digitalizada y representada en forma discreta. La figura 10 ilustra un ejemplo de una curva cerrada en 3D.

Para digitalizar curvas 3D es necesario intercalarlas en una cuadrícula tridimensional. Si se toma la secuencia de los puntos de la curva y se codifican los vértices de la cuadrícula más cercanos a la curva, se obtiene su digitalización. El tamaño del segmento de la cuadrícula define la resolución que se desea para la curva. Una vez digitalizada la curva se obtienen los valores discretos de la misma, es decir, si está en $\mathbb{Z}^3$ ($\mathbb{Z}$ es el conjunto de los números enteros). Dentro de la cuadrícula 3D los pasos siempre caen en los vértices de la misma, no en posiciones intermedias. La figura 11 muestra la versión discreta de la curva representada en la figura 10.

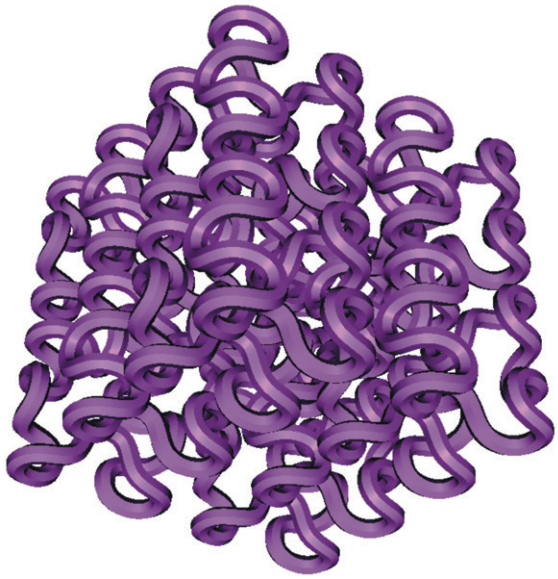


Figura 10. Representación continua de una curva 3D.

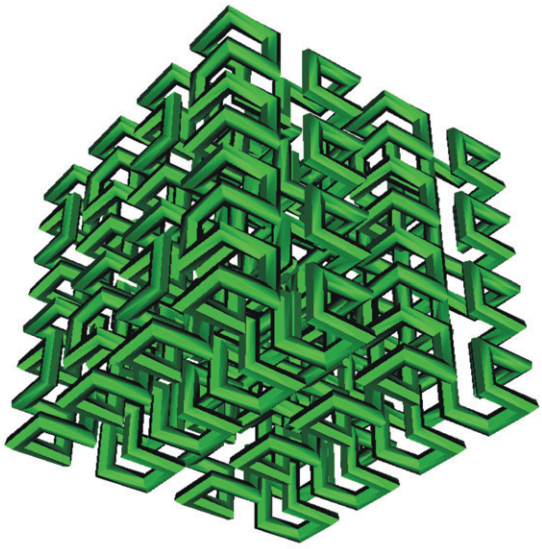


Figura 11. Representación discreta de la curva presentada en la figura 10.

Al igual que en la representación de dos dimensiones, se han encontrado importantes ventajas en la representación de curvas 3D por medio de códigos de cadenas, puesto que son una forma simple de almacenamiento de información y de compresión. Los códigos de cadena se han vuelto un estándar para muchos algoritmos de análisis de forma de objetos. Aquí se muestra el código de cadenas de cambios de dirección ortogonal,[3] llamado código de cadenas 3D.

Este código se presenta en la figura 12: (a) ejemplo de curva 3D continua; (b) versión digitalizada de curva 3D, constituida por segmentos de línea recta, todos ellos en posiciones ortogonales. Esta curva digitalizada está compuesta de 10 segmentos. Las figuras (c)-(g) muestran los diferentes elementos de la cadena para cada cambio de dirección.

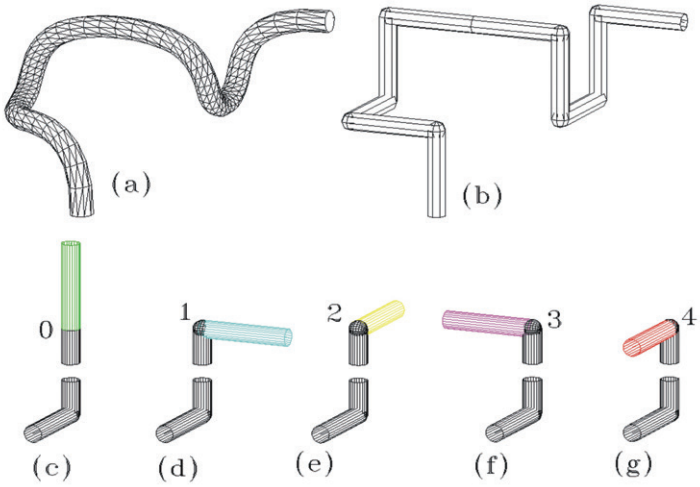


Figura 12. Definición de los elementos de código de cadenas 3D: (a) ejemplo de una curva continua; (b) versión discreta de la curva presentada en (a); (c) el elemento de cadena 0; (d) el elemento 1; (e) el elemento 2; (f) el elemento 3; (g) el elemento 4, respectivamente.

Con sólo cinco elementos[4] de cadena podemos representar cualquier curva discreta. No está permitido ir y regresar por el mismo segmento de línea recta. Cada cambio de dirección está formado por dos segmentos de línea recta contiguos, y cada elemento de cadena por dos cambios de dirección; es decir, siempre se necesitan las dos direcciones anteriores de los segmentos de recta para definir el actual. Con el fin de facilitar la comprensión de los diferentes cambios de dirección, se han asignado colores a cada uno de ellos. Así, los elementos de la cadena para representar cualquier curva discreta 3D están definidos como sigue:

[3] Ernesto Bribiesca, "A Chain Code for Representing 3D Curves", *Pattern Recognition*, 33, 2000, pp. 755-765.

[4] A. Guzmán, "Canonical Shape Description for 3-D Stick Bodies", en *MCC Technical Report*, núm. ACA-254-87, Austin, Texas, 1987.

1] El elemento de cadena 0 indica que no hay cambio de dirección del último segmento, es decir, que sigue en *línea recta*. Está representado en la figura 12(c) con el color verde.
2] El elemento 1 indica un cambio de dirección hacia la *derecha*, a partir del cambio de dirección anterior. Se muestra en la figura 12(d) con el color cian.
3] El elemento 2(e) indica un cambio de dirección tipo *escalera* y se muestra con el color amarillo.
4] El elemento 3(f) indica un cambio de dirección hacia la *izquierda*, a partir del cambio de dirección anterior, con el color magenta.
5] El elemento 4 describe un cambio de dirección de *regreso*, de *vuelta en U,* como se ilustra en la figura 12(g) con el color rojo.

La figura 13 muestra cómo asignar los elementos a una curva.

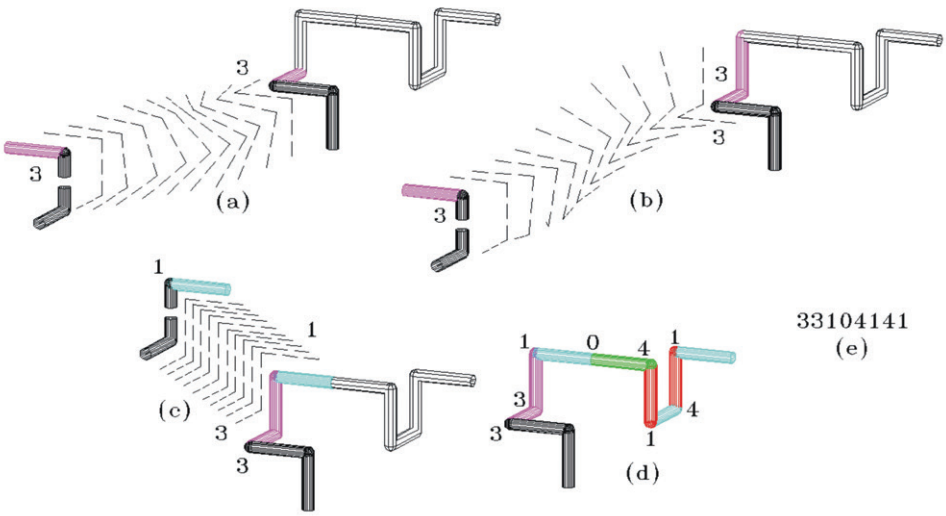


Figura 13. Asignación de los elementos a la curva: (a) cálculo del primer elemento de la curva; (b) cálculo del segundo; (c) cálculo del tercero; (d) todos los elementos de la curva con sus respectivos colores; (e) cadena de la curva.

El la figura 13(a), el origen de la curva se localiza en la parte baja. Ahora supóngase que se camina sobre la curva, al estar frente a un ángulo recto ¿cuál es el siguiente paso? Hacia la izquierda, por lo tanto le corresponde un 3, o sea, el color magenta. Los dos primeros segmentos al iniciar la curva no son etiquetados, debido a que sirven de referencia para definir el primer elemento, es decir, el 3. En la figura 13(b), para calcular el siguiente elemento de la cadena se toman como referencia, de nuevo, los dos elementos anteriores; se camina sobre la curva otra vez, se va hacia la izquierda, por lo que le corresponde un 3 (magenta). La figura 13(c) muestra el siguiente elemento de la curva, sólo que ahora se gira hacia la derecha y le corresponde un 1 (cian).

Finalmente, en la figura 13(d) se muestran todos los elementos de la cadena. Es importante notar que cuando existe un 0 es necesario ver cuál fue el último cambio de dirección diferente de este número para definir el elemento actual. Cuando se encuentra un 0 como en el ejemplo propuesto, se busca el anterior para definir el actual; en este caso se forma un 4, o sea el color rojo. La cadena está compuesta de los siguientes elementos: 33104141.

Para el caso de curvas abiertas, el número de segmentos de la curva siempre es el número de elementos de la cadena más dos, ya que los dos primeros se utilizan como referencia. La secuencia en colores de la cadena mencionada es la siguiente: magenta; magenta, cian, verde, rojo, cian, rojo y cian.

La figura 14 muestra dos ejemplos de curvas, utilizando los colores del código de cadenas 3D. Una esfera indica el origen de la curva abierta, mostrada en el lado izquierdo de la figura. Hay que recordar que los primeros dos segmentos siempre se usan como referencia. Esta curva abierta está compuesta de 40 segmentos de línea recta, o sea, de una cadena de 38 elementos.

La curva cerrada está representada por 24 segmentos de línea recta y por 24 elementos de cadena. En el caso de las curvas cerradas, tanto el número de segmentos de línea recta

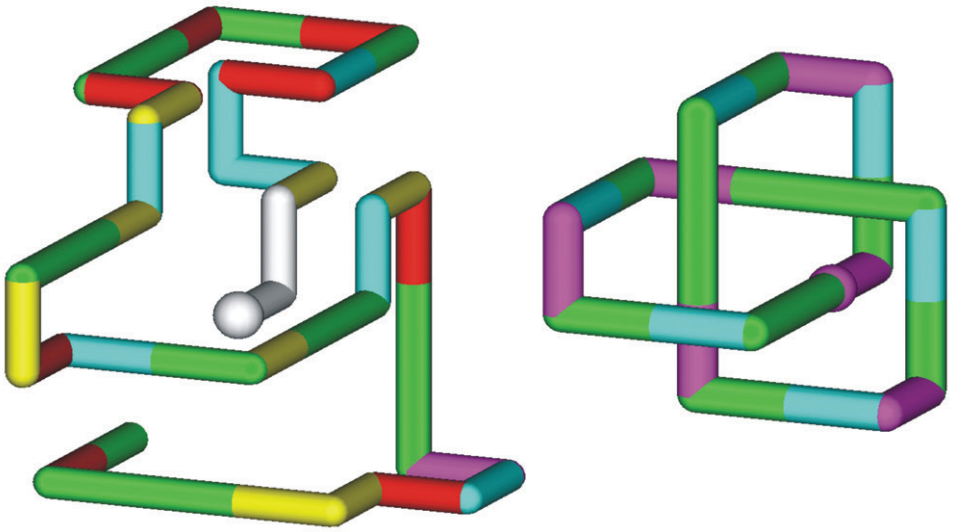


Figura 14. Representación del código de cadenas 3D usando colores. Ejemplos de una curva abierta (izquierda) y una cerrada (derecha).

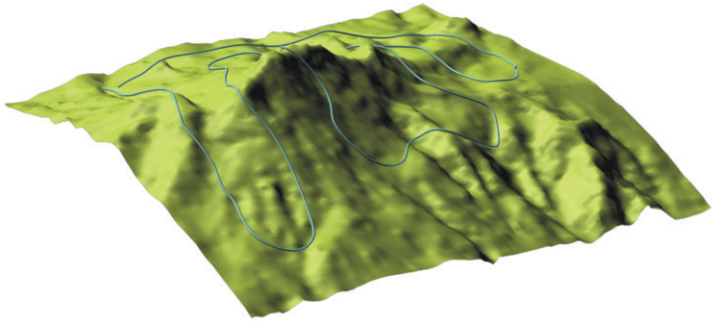
como el número de elementos de la cadena son iguales. Esto se debe a que siempre se pueden apreciar los dos cambios de dirección anteriores (diferentes de 0) para definir el cambio presente. El origen de la curva abierta se representa también por una esfera, y la dirección de la codificación se da en dirección al segmento verde contiguo.

Al igual que en el VCC, esta notación puede ser invariante al origen de construcción, al transformar las cadenas a números enteros, recorrer todos los posibles orígenes alrededor de la curva y seleccionar la cadena de menor valor. Para curvas abiertas, sólo hay dos orígenes y se debe seleccionar el que produzca el entero de menor valor.

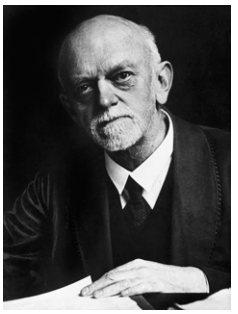
Un ejemplo donde se aplicó el código de cadenas 3D es la representación de la trayectoria de un avión volando sobre el volcán Iztaccíhuatl (véase la figura 15).

Figura 15. Trayectoria de un avión volando sobre el volcán Iztaccíhuatl.

Otra aplicación importante de este código ha sido la descripción de algunas curvas matemáticas conocidas, como la curva de Hilbert.[5] Esta curva es famosa porque posee un patrón, parecido a una silla de montar, que se repite y va cubriendo el espacio que la contiene en sus diferentes etapas. La figura 11 muestra una de las etapas de la curva de Hilbert; en este caso, para no



[5] W. Gilbert, "A Cube-Filling Hilbert Curve", en *The Mathematical Intelligencer*; vol. 6, núm. 3, 1984, pp. 78-79.



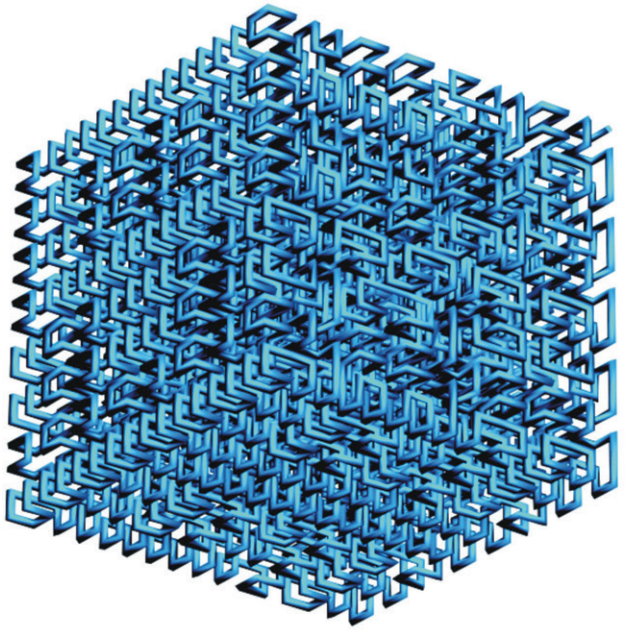
David Hilbert (1862-1943) | © Latin Stock México.

saturar la imagen con colores, sólo se trazó con uno. Está compuesta de 512 elementos de cadena, los cuales son:

433411434114334143242143411433 4
143341143411433414324214341143 33
411143341432421434114334143341 14
341143341432421434114334143341 12
423341143411433414324214341143 34
143341143411433414324214341143 33
411143341432421434114334143341 14
341143341432421434114334143341 11
433411434114334143242143411433 34
143341143411433414324214341143 33
411143341432421434114334143341 14
341143341432421434114334143341 12
423341143411433414324214341143 34
143341143411433414324214341143 33
411143341432421434114334143341 14
341143341432421434114334143341 11

Para ilustrar el potencial de este código, también se generó el siguiente estado de la curva de Hilbert, el cual se muestra en la figura 16. En cada estado, la curva se incrementa en un factor de 8; en este estado la curva contiene 4 096 elementos. Al igual que en el ejemplo anterior, para no saturar la imagen con colores, sólo se utilizó uno. Se observa que esta curva es abierta y la anterior es cerrada. Ambas versiones de curvas se pueden usar.

Figura 16. Curva de Hilbert representada por el código de cadenas 3D.



Una ventaja muy importante de este código es su nivel de compresión de información, ya que sólo requiere 3 bits para representar cada elemento de la cadena, es decir: 000, 001, 010, 011 y 100 para representar los elementos 0, 1, 2, 3 y 4, respectivamente.

Finalmente, este código se ha usado para generar familias de curvas con distintas resoluciones, es decir, con un número de elementos diferente. Por ejemplo, si se desea conocer todas las curvas de 3 elementos, se calculan todas sus combinaciones: 000, 001, 002, 003, 004, 010, 011, 012, 013, 014, 020, 021, 022, 023, 024, 030, 031, 032… y así sucesivamente.

Actualmente, la UNAM posee el récord mundial en la generación de la familia completa de todas las curvas 3D compuestas de 24 segmentos: 282,429,536,481 curvas.[6] El récord anterior fue para la familia completa de curvas de 20 segmentos, citado por Brian Hayes en su artículo "Square Knots" en la revista *American Scientist*.[7]

[6] Ernesto Bribiesca, "A Method for Computing Families of Discrete Knots Using Knot Numbers", *Journal of Knot Theory and Its Ramifications*, 14, 2005, pp. 405-424.
[7] B. Hayes, "Square Knots", *American Scientist*, 85, 1997, pp. 506-510.

Representaciones de superficies

Cuando una imagen se segmenta es posible extraer los objetos de la misma. Una forma común de representar un objeto es por medio de su superficie. La figura 17 muestra el modelo del volcán Popocatépetl. Aunque se trata de un sólido, una forma de representarlo es a través de la superficie que lo envuelve.

Concepto
Rendering es el proceso mediante el cual se genera una imagen a partir de un modelo, utilizando programas de cómputo. El modelo es, como se ha mostrado aquí, la descripción de un objeto tridimensional.

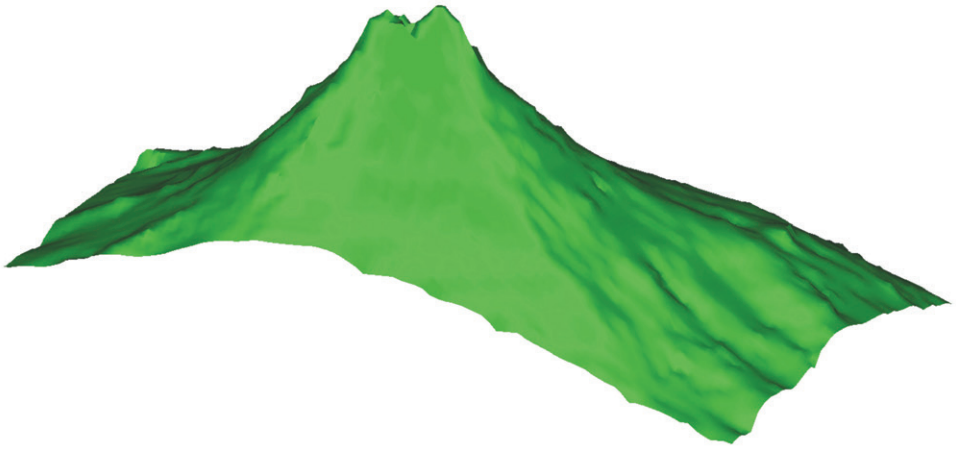


Figura 17. Volcán Popocatépetl, realzado en Z.

Una forma matemática para representar superficies es por medio de matrices; una matriz es un arreglo rectangular de números. Se coloca una red o malla (la resolución de la red se define previamente) sobre el modelo, como se ilustra en la figura 18. Esta malla tiene un espaciamiento previamente definido; está compuesta de 100×70 elementos y cada nodo de la red representa una altura en ese punto del modelo. Por supuesto, a mayor número de elementos, mayor precisión.

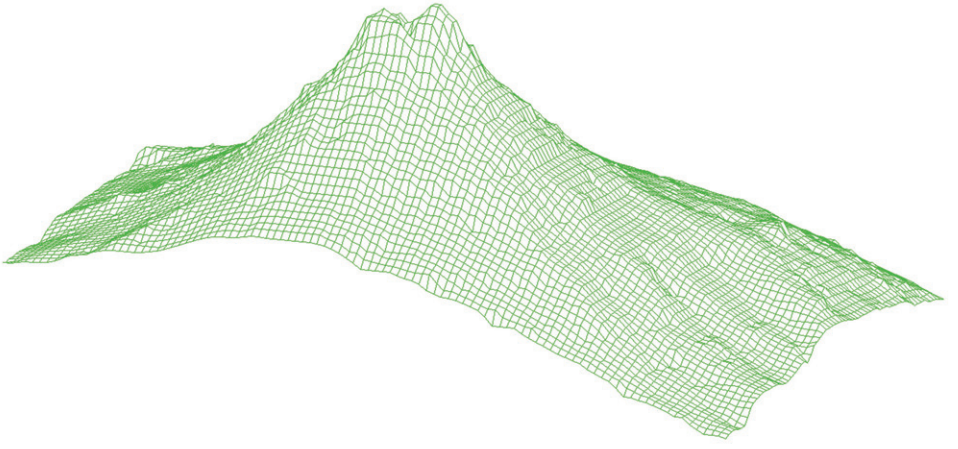


Figura 18. Volcán Popocatépetl representado por una malla de 100×70 elementos.

Representaciones de sólidos

Además de la superficie envolvente que se mencionó arriba, otra manera común de representar sólidos es por medio de "rebanadas"; es decir, a través de la obtención de cortes

de imágenes del sólido, tomando en cuenta un incremento constante en cada uno de los cortes. Al *unir* los cortes se obtiene un modelo tridimensional del objeto de estudio.

Cada uno de los cortes representa el objeto de estudio por medio de pixeles. Si se considera el incremento del siguiente corte igual a la longitud de uno de los lados del pixel, entonces se obtiene un voxel (*volume element* en inglés) y tiene una forma cúbica. Finalmente, el sólido queda representado por un conjunto de voxeles conectados entre sí. La figura 19 muestra el volcán Popocatépetl representado por voxeles.

Curiosidades
La tomografía es la obtención de imágenes, por secciones, de algún objeto. Este método se utiliza en varias disciplinas: medicina, arqueología, biología, geofísica y ciencia de materiales, entre otras. La palabra tomografía viene del latín *tomos*, que significa "sección" o "corte", y *grafía*, que quiere decir "representación gráfica".

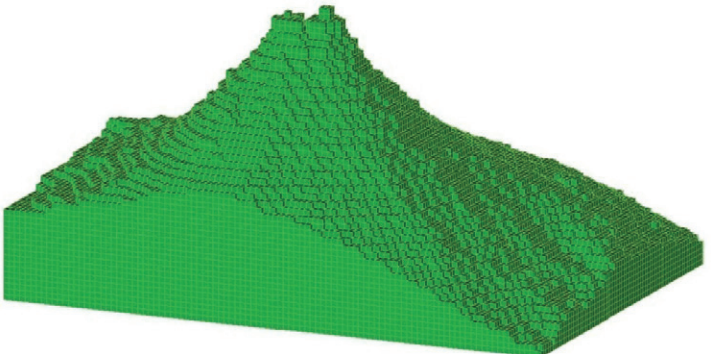


Figura 19. Volcán Popocatépetl representado por voxeles.

La forma de almacenar este modelo es por medio de matrices 3D. En este caso, sólo existen dos valores para los elementos de la matriz: el 0 y el 1; el 0 indica ausencia de materia y el 1 indica presencia de materia. Esta forma de almacenar la información tridimensional tiene ventajas, como la fácil obtención de cortes del modelo, así como su volumen y algunas otras características.

Una desventaja de esta representación es la cantidad de información redundante que contiene el modelo, en contraste con la notación de la superficie envolvente. En la notación de la superficie envolvente no existe redundancia en la información, sólo se almacenan las coordenadas de la superficie. Sin embargo, la notación de las superficies se complica cuando existen concavidades en el modelo, debido a que para cada una hay que generar una nueva superficie y, si se desea calcular el volumen total del sólido, hay que restar las superficies de las con-cavidades a la superficie envolvente del sólido.

Curiosidades
Las medidas de las estructuras cerebrales han probado ser de gran utilidad para la determinación de cambios relacionados con las patologías cerebrales, tales como: esquizofrenia, efecto de fármacos en desórdenes bipolares y cambios relacionados con la vejez.

La figura 20 muestra una importante aplicación de la representación por medio de voxeles; ilustra la materia gris de un cerebro humano compuesta por 876 224 voxeles. En este caso, cada voxel equivale a un milímetro cúbico de materia gris. Esta notación facilita la cuantificación del volumen de la materia gris y el cálculo de su superficie.

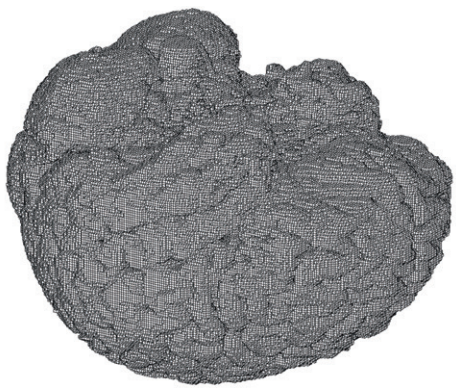


Figura 20. Materia gris de un cerebro humano representado por 876 224 voxeles (imagen derivada de un proyecto de la Universidad Autónoma Metropolitana y la UNAM).

8.4.6 Reconocimiento de objetos

Existen diversas técnicas para identificar objetos, basadas en patrones específicos, entre los cuales destacan tres patrones: sintácticos, estadísticos y estructurales. Los primeros se basan en técnicas sintácticas y presentan una gran variedad. Los patrones estadísticos, como su nombre lo indica, se basan principalmente en métodos estadísticos y, finalmente, los estructurales, utilizan métodos geométricos. Sus aplicaciones son casi tan variadas como las mismas disciplinas.

Para un médico es muy importante reconocer en una radiografía una fractura, fisura o tumor en un hueso. Los geólogos necesitan identificar fallas y fracturas en estructuras geológicas, por medio de fotografías aéreas o imágenes digitales de recursos de la Tierra provenientes de satélites. Es muy importante en aspectos legales y financieros poder reconocer la firma de una persona o reconocer su huella dactilar. El reconocimiento automático de rostros humanos es de gran interés para muchas disciplinas; en la actualidad es un área de investigación muy activa. El código de barras que se encuentra en prácticamente todos los productos de tiendas de autoservicio es un ejemplo clásico de cómo funciona el reconocimiento de patrones. En fin, se podrían citar infinidad de ejemplos sobre el reconocimiento.

Un método básico para el reconocimiento es la llamada máxima correlación. En ésta, se selecciona una *subimagen* de la imagen original; después, empleando la *subimagen* a manera de ventana, se recorre la imagen original a incrementos de un pixel, de arriba hacia abajo y de izquierda a derecha. En cada paso se calcula la suma de las diferencias entre los pixeles de la ventana y de la imagen. Cuando la suma da cero como resultado, se ha encontrado la máxima correlación. La figura 21 muestra la imagen y la ventana en la parte inferior; el cuadro enmarcado por líneas negras representa el máximo empate.

Curiosidades
La extracción de objetos de una imagen abre una gama de posibilidades. Una de las más interesantes es el reconocimiento de patrones, es decir, la identificación de objetos. El reconocimiento de patrones es un término importante en computación y se utiliza en diversas áreas, como análisis de imágenes, robótica e inteligencia artificial.

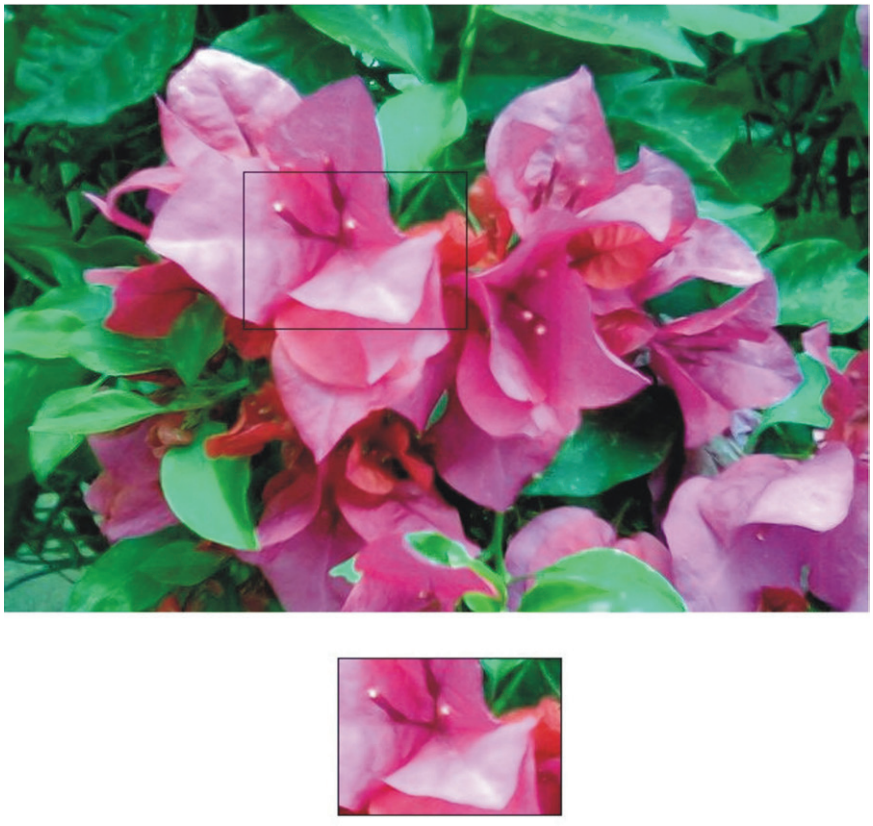


Figura 21. Reconocimiento de objetos por máxima correlación | © Ernesto Bribiesca.

8.5 ANIMACIÓN

Curiosidades
Para producir una hora de una película se requieren 86 400 imágenes. Recuérdese que se proyectan a razón de 24 imágenes por segundo.

Gregory MacNicol, especialista en tecnología 3D, define la animación como una simple verdad: "La animación es una ilusión". Así es, la animación es una simulación del movimiento. Los filmes están compuestos por una serie de imágenes estáticas. Generalmente, las películas son proyectadas en una pantalla a razón de 24 imágenes por segundo. Una producción en video usa el mismo principio y se muestra a 30 veces por segundo. El éxito de esta simulación se basa en el comportamiento del ojo humano. Éste retiene por un instante la última imagen percibida y cuando aparecen las siguientes hace lo mismo. Todo lo anterior más la compilación de información que realiza el cerebro producen la sensación de un movimiento continuo.

Definitivamente, con el advenimiento de las computadoras, el cine y la producción de video se han visto muy favorecidos. De hecho, las computadoras se han convertido en poderosos asistentes en los procesos de video y producción de películas, y han contribuido con nuevas formas, técnicas y estrategias. La computadora puede crear automáticamente una secuencia de animación, computando los intervalos entre cada una de las imágenes, y colocar las posiciones de las luces y el tipo e intensidad de las mismas. Puede computar los diferentes puntos de vista de las cámaras y su enfoque, los acercamientos, así como los distanciamientos de los objetivos. Una opción muy interesante para el uso de las computadoras en la animación es la extrapolación e interpolación entre imágenes, es decir, la transformación de una imagen en otra, un aspecto muy común en los videos.

Con el objetivo de explicar los pasos para realizar una animación, a continuación se explica la simulación del vuelo de un avión sobre el volcán Iztaccíhuatl. Los pasos fundamentales para crear esta animación son:

1] Formar un modelo. Se toman las coordenadas del modelo digital de elevaciones (como se mencionó anteriormente, una elevación es una matriz de alturas). En los modelos digitales de elevación, la altura está dada en metros sobre el nivel del mar. Asimismo, cada elemento de la matriz representa una altura. El tamaño de la matriz es de 140 × 140. Parte de ésta se muestra a continuación.

2913 2900 2886 2874 2869 2871 2872 2870 2868 2860 2863…
2916 2908 2892 2880 2874 2870 2868 2860 2864 2862 2861…
2912 2910 2890 2889 2883 2876 2871 2866 2865 2860 2854…
2909 2894 2888 2883 2878 2874 2868 2860 2859 2855 2840…
2904 2900 2890 2880 2877 2870 2871 2865 2850 2852 2847…
2898 2890 2892 2886 2888 2876 2872 2860 2855 2850 2840…

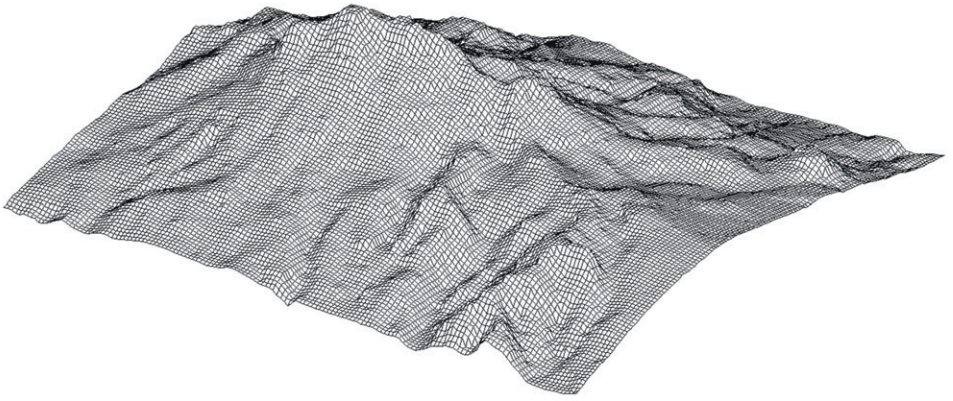


Figura 22. Modelo digital de elevaciones del volcán Iztaccíhuatl, representado por una malla 3D de 140 × 140 elementos.

2] Una vez consideradas todas las alturas de la matriz, se asigna a cada uno de los vértices de la malla 3D una malla con una resolución de 140 × 140 elementos, como se muestra en la figura 22.

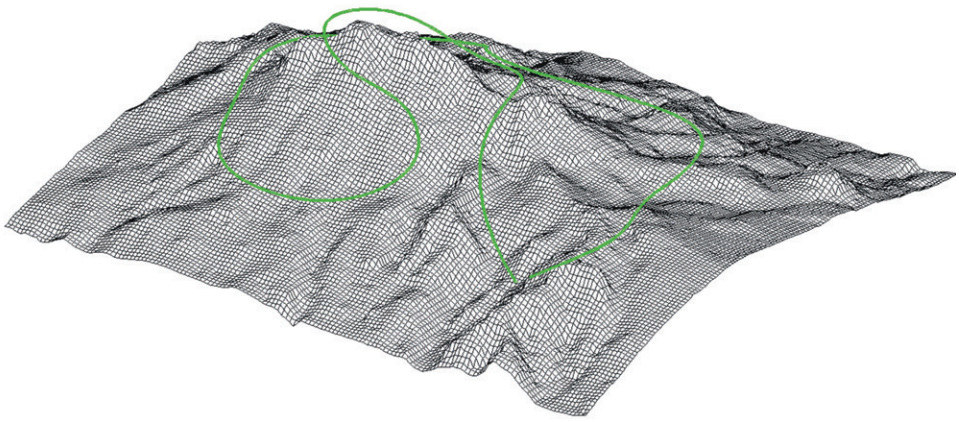


Figura 23. Trazo de la trayectoria de un avión volando sobre el volcán Iztaccíhuatl.

3] Se diseña la ruta de vuelo, misma que se puede realizar por medio del código de cadenas 3D. La figura 23 muestra la selección de esta ruta.

4] Se divide la ruta de vuelo en el número de intervalos en los que se desea proyectar la animación; en este caso se dividió en 2 000 puntos. Posteriormente, se posiciona la cámara en cada uno de los puntos y se toma una imagen desde cada punto de vista, siguiendo la dirección de la trayectoria del avión. Cada imagen es procesada (***rendering***)[8] y almacenada. A continuación se presenta una secuencia de 10 imágenes (de las 2 000 generadas) en la figura 24.

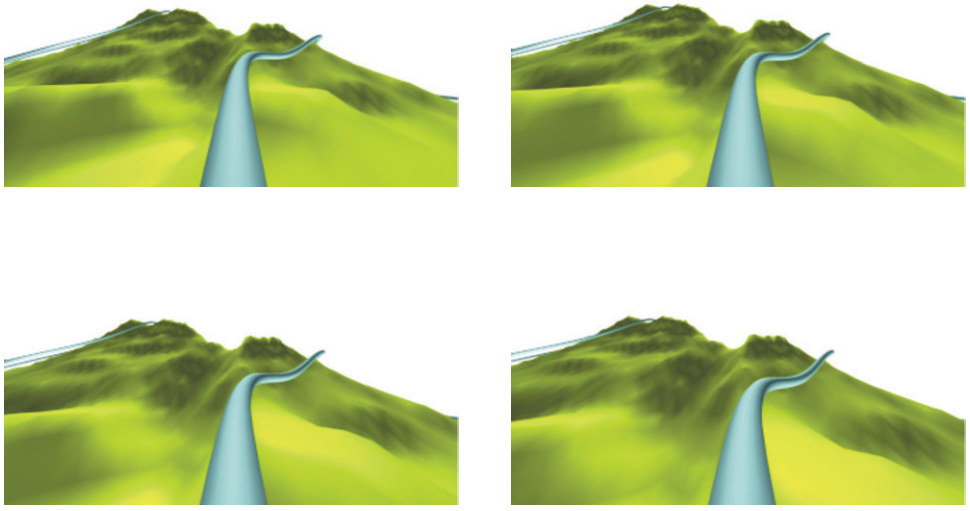


Figura 24. Secuencia de 10 imágenes de la simulación del vuelo de un avión sobre el volcán Iztaccíhuatl.

[8] *Rendering* puede traducirse como "representación"; sin embargo, en el medio de la computación es un anglicismo muy utilizado y por ello se optó por no utilizar su equivalente en español.

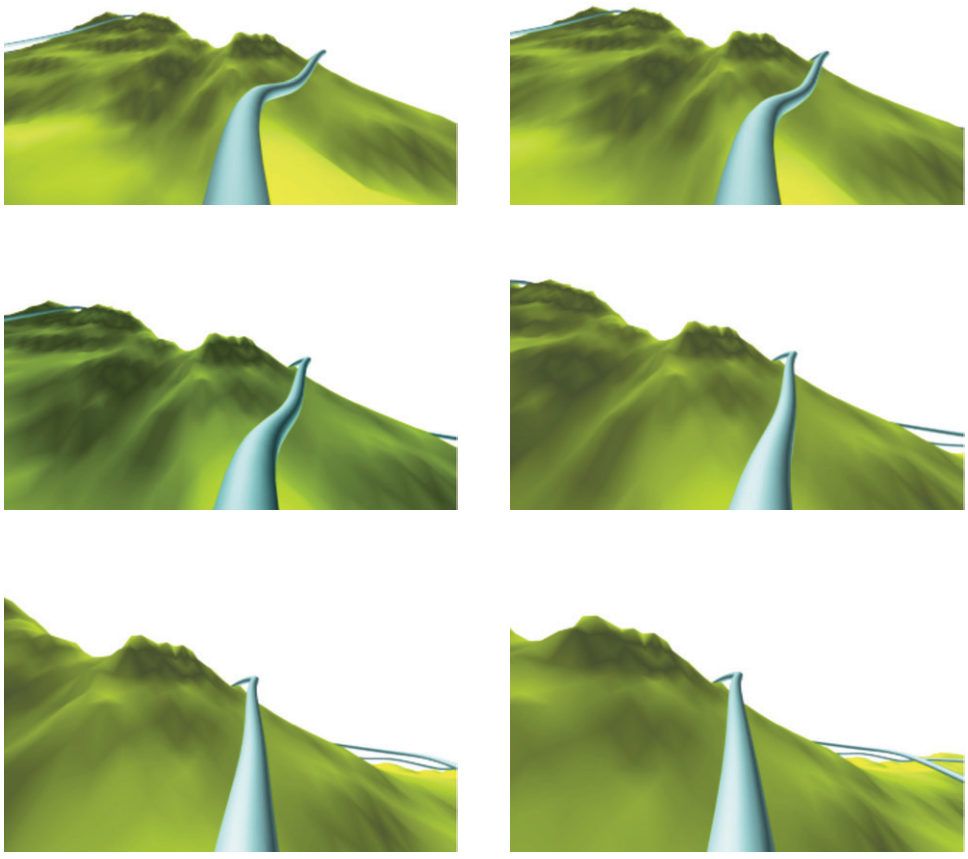


Figura 24. Secuencia de 10 imágenes… (continuación).

5] Finalmente, se despliega cada una de las 2 000 imágenes en secuencia. Para lograr una mejor animación, con un programa se sincroniza la proyección a 30 imágenes por segundo. De esta forma se obtiene una animación de más de un minuto.

En el DVD que acompaña este libro se presentan dos animaciones o simulaciones de dos vuelos: uno, sobrevolando el volcán Iztaccíhuatl, y otro, el Valle de México. El primer clip de video contiene un total de 2 000 imágenes a una velocidad de 30 imágenes por segundo, con una duración de un minuto con seis segundos. El segundo clip está compuesto por 3 000 imágenes a una velocidad de 30 imágenes por segundo, con una duración total de 1 minuto con 40 segundos. Ambos videos muestran una simulación de vuelo, donde las cámaras se mantienen sobre las trayectorias de las rutas establecidas. Ambas trayectorias de curvas fueron almacenadas en el código de cadenas 3D; sin embargo, para su despliegue y animación se conservaron sus versiones continuas.

Curiosidades
Los modelos digitales de elevaciones provienen de información cartográfica del Instituto Nacional de Estadística y Geografía (INEGI), con una escala 1:250 000. La programación se realizó en el lenguaje C. La altura de los modelos de elevación fue realzada por un factor de 3 para mejorar la identificación de la topografía.

8.5.1 Realidad virtual inmersiva

Ixtli,[9] el Observatorio de Visualización de la UNAM, es una sala de alta tecnología diseñada para visualizar y simular objetos complejos e imágenes en tercera dimensión mediante un sistema de realidad virtual inmersiva. Las aplicaciones que pueden hacer uso de la inmer-

[9] <www.ixtli.unam.mx>.

sión son aquellas que obtienen beneficios de la libertad de interacción y de la sensación de presencia dentro del mundo tridimensional. Algo también muy interesante es que gran parte del software utilizado en la sala Ixtli es libre.

Este lugar de encuentro multidisciplinario, en el cual las nuevas tecnologías computacionales y de electrónica dan vida al trabajo docente y de investigación de los universitarios, apoderándose de nuestros sentidos y percepciones para crear una ilusión total de tridimensionalidad, posee las más avanzadas técnicas de realidad virtual para disposición de los académicos en la enseñanza y la investigación en todas las áreas del conocimiento humano.

La tecnología y el diseño de esta herramienta de trabajo permiten múltiples usos, lo que la hace única en México y en toda América Latina. Además, es la sala con mayor capacidad de cómputo intensivo en operación en una institución de educación superior en el país. En Ixtli se puede ver, escuchar y tener una experiencia realmente innovadora a través de una pantalla curva, especialmente diseñada para realzar y mejorar las representaciones de los diferentes proyectos de investigación en el quehacer universitario y, sobre todo, para comprender mejor la realidad y los resultados de las investigaciones.

Dentro de esta sala es posible crear una sensación de inmersión dentro de un modelo computarizado, donde también se puede interactuar en este mundo virtual. El número de aplicaciones de estas técnicas es grande y variado, por mencionar algunas: en medicina, los médicos las usan para visualizar órganos en tres dimensiones, enseñar su funcionamiento y explicar técnicas de cirugía; en matemáticas, los topólogos pueden visualizar y clasificar diferentes tipos de nudos por medio de la navegación en los mismos; en la arquitectura, la realidad virtual inmersiva se usa para navegar por los espacios que se diseñan, con el objetivo de sentir sus dimensiones y adaptarlos antes de construirlos; en las geociencias, para analizar regiones de la República Mexicana mediante vuelos en modelos digitales de terreno e imágenes satelitales, y para poder realizar clasificaciones de especies o estudios socioeconómicos, usando sistemas de información geográfica.

En fin, el número de aplicaciones es muy variado en las áreas científicas, sociales y artísticas. La figura 25 muestra la pantalla curva de la sala Ixtli. La figura 26 representa el modelo de un nudo complejo. La figura 27 ilustra la familia completa de todas las curvas 3D discretas compuestas de 12 segmentos (1 403 en total). Si se tuvieran seis palillos y se acomodaran solamente en posiciones ortogonales en el espacio, es decir, en múltiplos de 90 grados, formando curvas cerradas, ¿cuántas curvas 3D discretas diferentes se formarían?

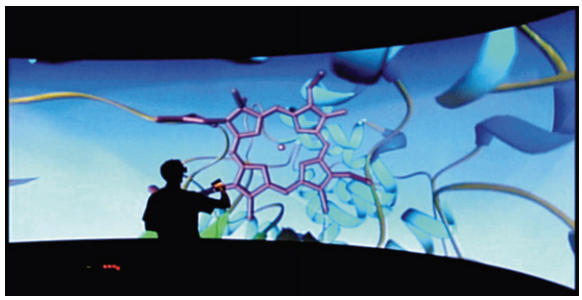


Figura 25. La pantalla curva de la sala Ixtli de la UNAM | © Ixtli-DGSCA-UNAM.

Figura 26. Modelo tridimensional de un nudo | © Ernesto Bribiesca.

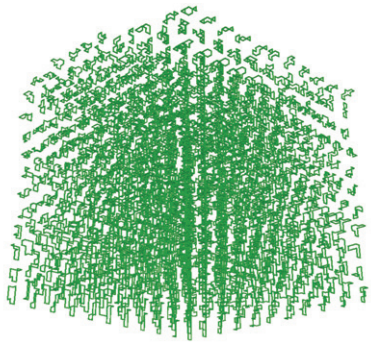


Figura 27. Familia completa de todas las curvas 3D discretas, compuestas de 12 segmentos (1 403 en total) | © Ernesto Bribiesca.

La figura 28 muestra la estructura química del virus del dengue, aislado en el estado de Guerrero. La figura 29 ilustra un embrión de ratón con implantes de células fosforescentes para destacar su estructura fisiológica. Finalmente, la figura 30 muestra un modelo de murales y edificios del sitio arqueológico maya de Bonampak, reconstruidos digitalmente para realizar recorridos virtuales.

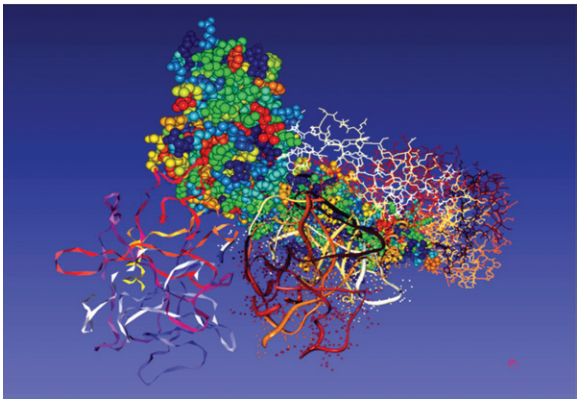


Figura 28. Estructura química del virus del dengue, aislado en el estado de Guerrero | © Ixtli-DGTIC-UNAM.

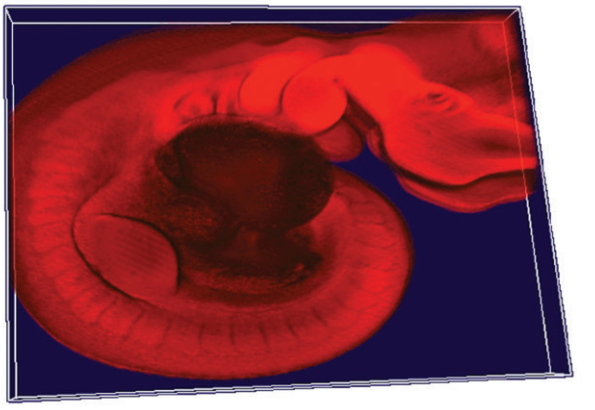


Figura 29. Embrión de ratón con implantes de células fosforescentes para destacar su estructura fisiológica | © Ixtli-DGTIC-UNAM.

Figura 30. Modelo de murales y edificios arquitectónicos de las ruinas mayas de Bonampak, reconstruidos digitalmente para realizar recorridos virtuales | © Ixtli-DGTIC-UNAM.

8.5.2. La animación en el cine

Una de la áreas de la computación con mayor impacto en las últimas tres décadas es, sin lugar a duda, la animación por computadora aplicada al cine. Al inicio de la década de los años ochenta, las computadoras se utilizaban en el cine sólo con la finalidad de crear fondos basados en vectores (por ejemplo, la película *Tron*) o imágenes generadas por computadora o imágenes CGI (Computer Generated Imagery), como en *Star Trek II: The Wrath of Khan*.

Para 2003, la industria cinematográfica ya contaba con personajes totalmente animados por computadora, cuyas actuaciones son tan buenas que incluso algunos han recibido reconocimientos; por ejemplo, Gollum en *Lord of the Rings: The Two Towers* obtuvo un Oscar por el mejor actor de reparto.

En la década de los años noventa, la utilización de computadoras para animaciones en cine alcanzó un desarrollo inusitado. Los directores de cine y los estudios cinematográficos fueron más ambiciosos y mezclaron todo tipo de técnicas. Por ejemplo, *The Beauty and the Beast* (junio de 1991) fue la primera película que combinó la animación tradicional con CGI. *Terminator 2: Judgement Day* (mayo de 1992) incursionó en efectos "líquidos" y de "transformación" y, en noviembre de 1995, se estrenó *Toy Story*, que fue la primera película totalmente CGI. A esta última siguieron muchas películas de dibujos animados, aunque probablemente las películas más recordadas del decenio en este sentido son *The Matrix* (marzo de 1999) y el primer capítulo del éxito de los setenta *Star Wars* (mayo de 1977).

Durante los primeros años del siglo XXI, el cine ha continuado explorando (y empujando) los límites de la animación por computadora. Para la película *Transformers* (2007), un grupo de aproximadamente 350 ingenieros, armados con todo tipo de fotografías de partes mecánicas, crearon un personaje de nombre Bumblebee de carne y hueso o, mejor dicho, de mecánica y animación, a partir de las 750 partes mecánicas de un Camaro. Esta película se convirtió en un hito en cuanto a animación se refiere, no precisamente por la manera en que las armadoras de autos funcionan, sino por la magia y la técnica de los efectos especiales que empleó la empresa Industrial Light & Magic (ILM).

Curiosidades
Industrial Light Magic (ILM) forma parte de una de las empresas más importantes de entretenimiento en el mundo: Lucasfilm, que fue fundada por George Lucas, creador de *Star Wars*, en 1971. ILM es la parte encargada de la generación de animaciones y efectos especiales y, junto con Skywalker Sound, satisface buena parte de las necesidades digitales de la industria del entretenimiento.

8.5.3 Generación de imágenes para la pantalla azul

Se puede afirmar que en la actualidad casi todas las películas utilizan computadoras para llevar a cabo sus efectos especiales. Esta técnica mezcla personajes y objetos reales (autos, motocicletas, aviones), o parte de ellos y un fondo de color azul. Las pantallas azules, que no son siempre azules, pues también hay verdes y rojas, son pantallas brillantes, iluminadas de manera uniforme. Se filma una secuencia con los personajes y después se *sustituye* el fondo azul con una animación, otro color o con otras imágenes.

Muchas películas recientes han llevado esta técnica al extremo y, en lugar de utilizarla para filmar sólo las secuencias de acción intensa y agregar después los efectos especiales por medio de computadoras, han filmado la película entera con una pantalla azul de fondo. Esto, por supuesto, complica la actuación de las personas, porque no "ven" lo que estará a su alrededor en la escena final. Algunos ejemplos de películas filmadas en su totalidad con pantalla azul son: *Sin City* (marzo de 2005) y *300* (diciembre de 2006), ambas adaptaciones de cómics y novelas del notable escritor Frank Miller.

Para concluir, es innegable que las ciencias de la computación han desempeñado un papel fundamental en las técnicas modernas de los sistemas de multimedia.

8.6 RESUMEN

Los datos, aquellas criaturas que habitan el mundo de la computación, pueden representar texto, voz, música, imágenes y videos. Los sistemas de cómputo modernos almacenan y procesan los datos, y éstos aprovechan la variedad de medios para comunicarse con los humanos.

APLICACIONES

TEMA

9

© *José Galaviz.*

9.1 INTRODUCCIÓN

La diferencia entre la teoría y la práctica en teoría es mucho menor que la diferencia entre la teoría y la práctica en la práctica.

PROVERBIO POPULAR.

La perspectiva que dan las ciencias de la computación es diferente. Tanto en una **imagen** digital en dos dimensiones compuesta por pixeles como en una en tres dimensiones compuesta por voxeles, las posiciones de los pixeles o voxeles están perfectamente definidas, es decir, uno se desplaza en posiciones fijas a pasos discretos a múltiplos de pixeles o voxeles; no existen posiciones intermedias. Matemáticamente, esto significa que solamente andamos en el conjunto de los números enteros, o sea, en $\mathbb{Z}^2$ y $\mathbb{Z}^3$ para dos y tres dimensiones, respectivamente. Por ejemplo, si se desea saber el número de caminos posibles entre dos pixeles de una imagen, éste siempre será finito, pues siempre se podrá calcular.

Cuando se vieron los códigos de cadenas en el tema sobre multimedia se mencionó que realizando todas las combinaciones posibles de los elementos de cadena, los cuales representan vectores unitarios a diferentes direcciones discretas, se podían conocer todas las formas cerradas posibles en un determinado número de segmentos (el número de segmentos define la resolución de la forma). Así, en este universo de formas estarían, por ejemplo, las siluetas de todos los peces que existen actualmente, los que existieron y los que existirán. Todo lo anterior genera un ambiente amistoso de certidumbre y seguridad.

No es lo mismo estar observando cómo se desplaza un insecto y analizando cómo evade los objetos que se encuentra en su camino que hacer un robot que imita o emula los movimientos del insecto para desplazarse con "autonomía" y "tomar decisiones" para evadir los objetos que encuentra en su camino.

Tampoco es lo mismo estar en una granja clasificando los diferentes sonidos que producen las gallinas ante diferentes depredadores que hacer un robot que "entienda" el lenguaje natural y que dada una orden en ese lenguaje realice una acción.

Definitivamente, la computación ha retado al hombre a tener otra perspectiva del universo físico que le rodea a la hora de tratar de modelar y simular los fenómenos de la naturaleza y de intentar imitar las actividades de la inteligencia humana. Esta perspectiva ha ampliado nuestros horizontes y nos ha ayudado a sentir más respeto y admiración por el universo físico que nos rodea y por todo lo que en él existe.

Las computadoras agilizan la mayoría de los procesos en casi todas las ramas del conocimiento humano. Por mencionar un ejemplo: en 1596, el matemático alemán Ludolf van Ceulen calculó el número π con 35 cifras decimales, número con el que trabajó casi hasta el día de su muerte a los 70 años de edad; pidió que los 35 dígitos de π se inscribieran, sobre la lápida de su sepulcro como un digno epitafio. Su deseo fue cumplido. El valor que dio para π es, en parte 3.14159265358979323846… En recuerdo de su hazaña, los alemanes llaman todavía a π el número ludolfiano. Ahora ese mismo cálculo se hace rápidamente en una computadora. ¿Qué pensaría el mismo Ludolf van Ceulen de saber que la obra de su vida ahora se resume en unos instantes con el uso de las computadoras?

9.2 CIENCIAS DE LA TIERRA

La computación ha penetrado profundamente en las disciplinas correspondientes a las ciencias de la Tierra o geociencias, que son las disciplinas que estudian la morfología, la evolución, las estructuras internas y superficiales de nuestro planeta. La evolución en las geociencias ha sido muy activa; la geografía se ha ido haciendo más compleja, y han surgido nuevas áreas, como la geomorfología, la geofísica, la geoquímica, la geología, la hidrología, la meteorología, la edafología, la climatología, la tectónica, la petrografía, la paleontología, la oceanografía, la sismología, la mineralogía, la vulcanología, la petrología y otras. Todas estas disciplinas se encuentran entrelazadas y conforman las geociencias o ciencias de la Tierra. La computación ha jugado un papel muy importante en cada una de ellas y ha sido un catalizador para su rápido desarrollo. A continuación se mencionarán algunas aplicaciones.

9.2.1 Cartografía automatizada

En la cartografía, la producción automática de mapas derivados de información primaria es fundamental para estudios de planeación y desarrollo. La figura 1 muestra el modelo

Curiosidades
Un antecedente importante de los países desarrollados fue el inventario exhaustivo de sus recursos naturales y potenciales. Lo anterior se hizo por medio de cartografía a detalle. ¿Cómo puede un país hacer una buena planificación y tomar decisiones adecuadas si no conoce los recursos con que cuenta y de los que carece?

digital[1] del Valle de México desde una vista superior. La orientación de modelo presentado en la figura 1 es la siguiente: el Norte corresponde a la parte superior y fácilmente se puede distinguir la Sierra de Guadalupe; del lado izquierdo se puede observar el valle de Toluca; en la parte derecha empieza la topografía de los volcanes Iztaccíhuatl y Popocatépetl; y la parte baja corresponde al Sur y se pueden distinguir el Ajusco y el cerro Pico del Águila.

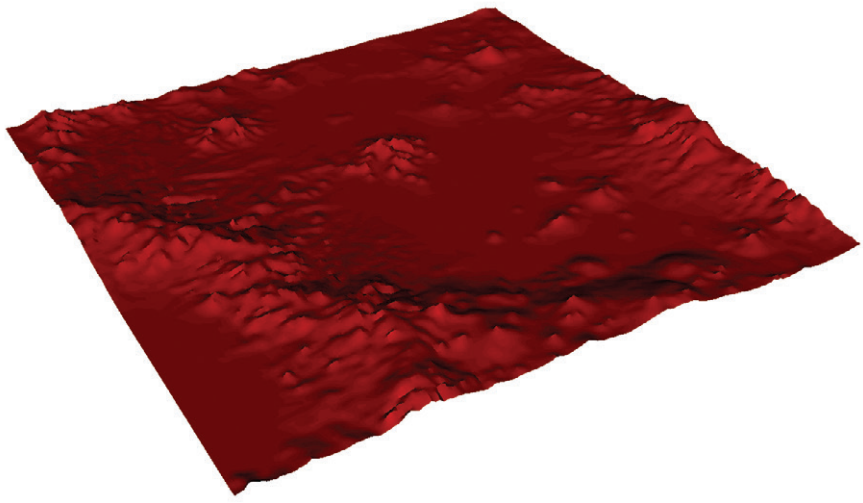
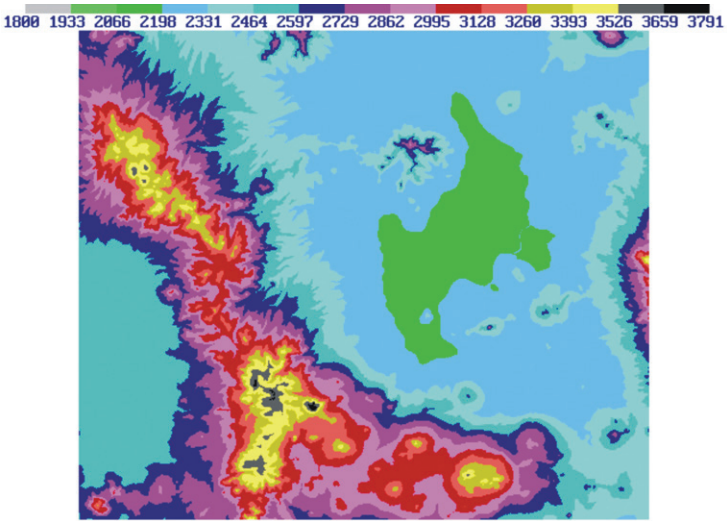


Figura 1. Modelo digital de elevaciones del Valle de México, visto desde arriba | © Ernesto Bribiesca.

La figura 2 muestra un mapa derivado de rangos de altura de la información anterior asignando un color a cada rango de altura previamente definido. Aquí es donde la computación juega un papel fundamental en la cartografía computarizada. Los rangos de altura mostrados están en metros sobre el nivel del mar y definidos por diferentes colores, los cuales se muestran en la parte superior de la imagen.

Figura 2. Mapa de rangos de alturas en metros sobre el nivel del mar derivado del modelo digital de elevaciones del Valle de México | © Ernesto Bribiesca.



9.2.2 Geomorfología

La geomorfología es el estudio de las formas de la superficie terrestre. El análisis de este tipo de estructuras es de suma importancia para diferentes estudios. Usando técnicas de reconocimiento de patrones, se pueden detectar fallas geológicas o analizar diferentes modelos de drenaje sobre el terreno, y también es posible clasificar diferentes tipos de volcanes. La figura 3 muestra el volcán La Maliche o Malintzin ("venerable señora hierba", en náhuatl) haciendo la comparación con una estructura cónica. En esta parte del estudio es muy importante la aplicación de las técnicas de medidas de similitud o semejanza entre objetos 3D, como los mostrados en la figura 3.

[1] Los modelos digitales de elevaciones fueron obtenidos de información cartográfica proveniente del INEGI de escala 1:250000. Toda la programación fue hecha en lenguaje C. La altura de los modelos digitales de elevación fue realizada por un factor de 3 para una mejor identificación de la topografía digital.

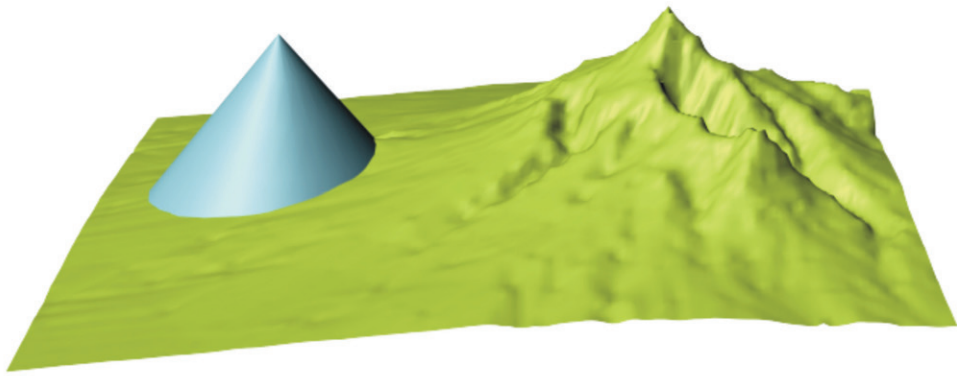


Figura 3. Análisis de estructuras geológicas: volcán la Malinche realzado en Z.

Concepto
Los satélites artificiales dedicados a la captura de información climatológica generan una gran cantidad de imágenes que forman grandes bases de datos, que no sería posible procesar sin ayuda de computadoras y complejos sistemas de información.

9.2.3 Climatología

La climatología es una rama de la geografía física dedicada al estudio del clima y el tiempo. El clima está totalmente ligado a las actividades humanas y por eso es de gran importancia su conocimiento y predicción. Las ciencias de la computación han ayudado enormemente a esta disciplina. Toda la información que se captura y la cantidad de variables que se manejan resultan impresionantes. No sería posible procesar todas estas variables manualmente. Los satélites artificiales dedicados a la captura de este tipo de información generan una gran cantidad de imágenes y forman grandes bases de datos. La figura 4 muestra dos imágenes del estado de Tabasco en dos fechas diferentes, tomadas por un satélite de la NASA, que muestran las áreas afectadas por inundaciones a finales de 2007.

En las dos imágenes mostradas es posible ver las diferencias de una manera notoria. En la imagen superior se muestran las inundaciones causadas por las lluvias, a diferencia de la imagen inferior. En ambas imágenes, el agua se presenta en colores azul oscuro y negro. Así, usando técnicas de análisis de imágenes digitales es factible clasificar los cuerpos de agua con base en la luz reflejada y hacer un rápido inventario de las zonas afectadas.

La figura 5 muestra una imagen tomada del satélite americano-francés altimétrico *Jason* y corresponde a las condiciones generadas por los fenómenos denominados El Niño y La Niña.

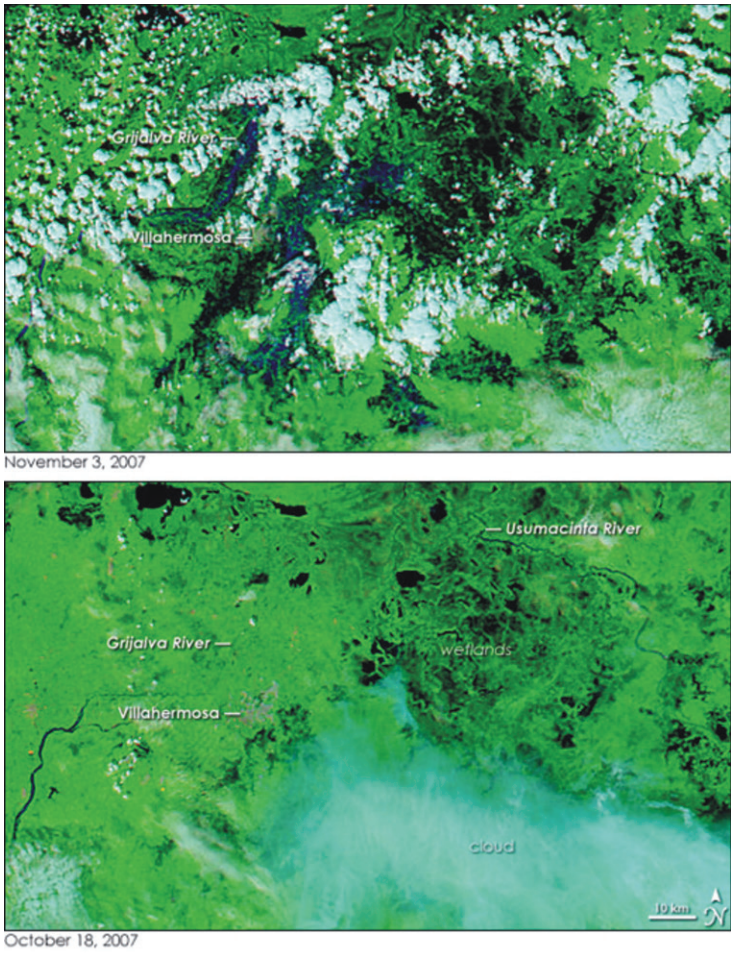


Figura 4. Imágenes del estado de Tabasco en dos fechas diferentes, tomadas por un satélite de la NASA, que muestran las áreas afectadas por inundaciones | © NASA Earth Observatory.

Concepto
Usando técnicas de segmentación de imágenes digitales, como se mencionó en el tema sobre multimedia, es factible clasificar los cuerpos de agua con base en su luz reflejada.

Curiosidades
Durante La Niña, los vientos son más fuertes de los que aparecen normalmente, y las aguas frías que regularmente existen en las costas del sur de América ahora se extienden al Pacífico central.

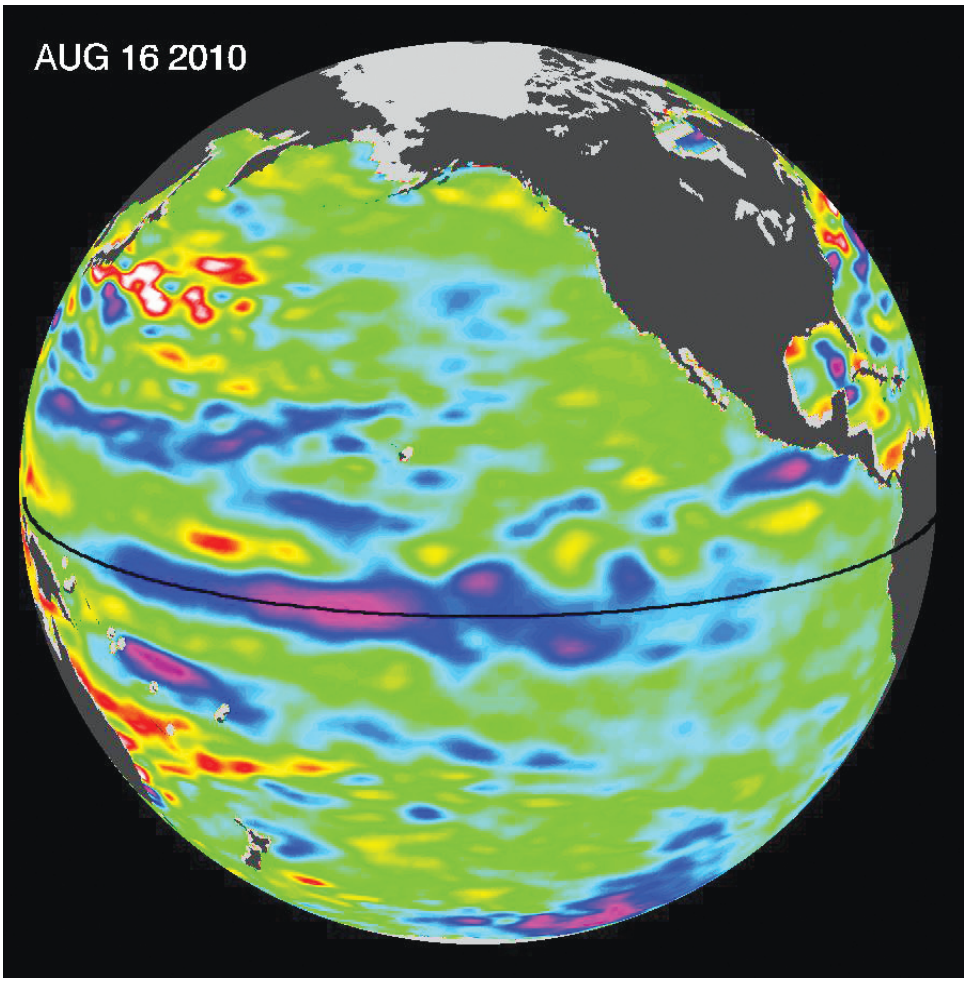


Figura 5. Imagen del Océano Pacífico tomada por el satélite americano-francés altimétrico Jason | © Courtesy NASA/JPL-Caltech.

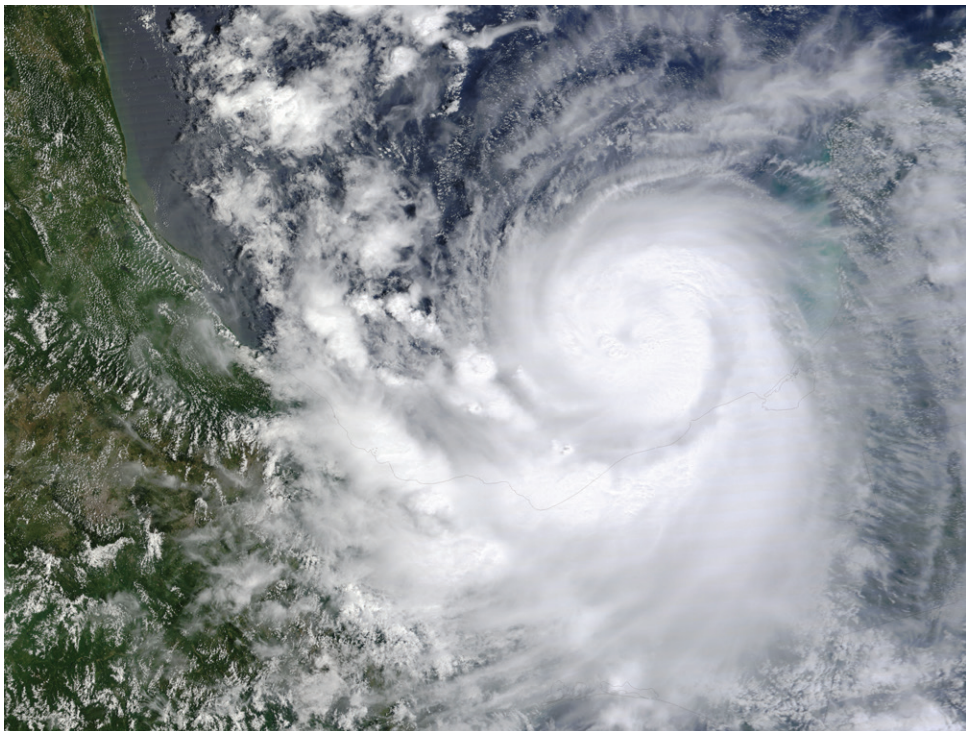
Concepto
La cantidad de información producida y el número de imágenes generadas por medio de los satélites meteorológicos son impresionantes; el número de variables que influyen en este tipo de fenómenos es muy grande. Una vez consideradas todas las variables se procede a la modelación y simulación del fenómeno para hacer predicciones por medio de sistemas de cómputo.

Los colores amarillo y rojo de la imagen en la figura 5 corresponden a temperaturas más altas de lo normal sobre la superficie del Océano Pacífico. Los colores azul y púrpura, a temperaturas más bajas de lo normal sobre la superficie del océano. Finalmente, los de color verde corresponden a las condiciones normales.

9.2.4 Meteorología

La meteorología es la ciencia que estudia el estado del tiempo, los fenómenos que se producen en la atmósfera y las leyes que los rigen. La meteorología fue muy favorecida a partir del lanzamiento del primer satélite meteorológico TIROS-1 en 1960 y con la llegada de computadoras más poderosas. Lo anterior significó el inicio de una era de difusión y procesamiento global de la información climática. Todos los fenómenos atmosféricos son de gran interés para la humanidad debido a implicaciones; por ejemplo, el conocimiento y la predicción sobre los huracanes es fundamental para evitar desastres.

La figura 6 muestra una imagen del huracán *Karl* que causó graves daños en varios estados de la República Mexicana.



Curiosidades
La temperatura es vital en la formación de huracanes. Se requiere que la superficie del agua del mar esté a una temperatura aproximadamente igual o mayor a los 26.66 grados Celsius (información tomada de la National Aeronautics and Space Administration).

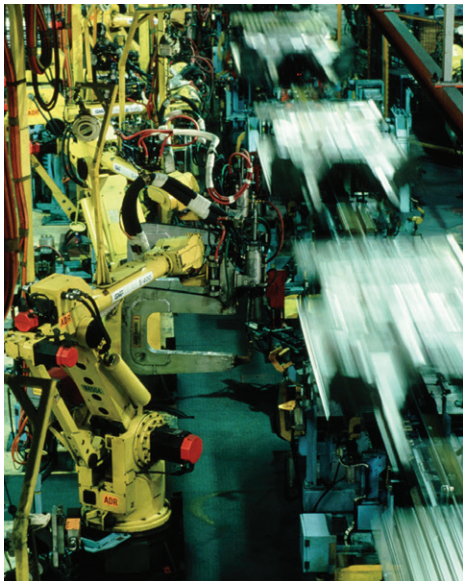
Figura 6. Imagen del huracán Karl, *tomada el 17 de septiembre de 2010* | © NASA *Earth Observatory.*

9.3 ROBÓTICA

9.3.1 Robots de servicio

Los *robots industriales*, usados principalmente en la manufactura y que en su mayoría son brazos robotizados estacionados en un solo lugar, se introdujeron en las fábricas durante el decenio de los sesenta. La figura 7 muestra estos robots en una fábrica de automóviles.

Así como se incorporaron a la vida cotidiana televisores, radios, computadoras y otras máquinas, los robots de servicio (figura 8) también lo harán en algún momento. Uno de ellos se dedicará a cortar el pasto, mientras otro realizará funciones de vigilancia. Dentro de la casa se tendrán robots pequeños que limpien, planchen y doblen la ropa, entre muchas otras funciones.



En la figura 9 se muestra el **robot de servicio** *Asimo*, fabricado por la compañía japonesa Honda, y en la figura 10 se muestra al robot TPR8, desarrollado en el laboratorio de Bio-Robótica de la Facultad de Ingeniería de la Universidad Nacional Autónoma de México.

La computación se encuentra en diversos niveles cuando se relaciona con la robótica. Por ejemplo, TPR8 utiliza el sistema ViRbot, que permite probar algoritmos para robots de servicio. Se mencionarán algunas de las partes más interesantes:

Curiosidades
Se espera que en el futuro aumente la demanda masiva de los ahora llamados robots de servicio, cuyo objetivo es simplificar el trabajo humano en casas, oficinas, tiendas, etc. Este tipo de robots son dispositivos ambulantes programables que ofrecen servicios en forma automática o semiautomática.

Figura 7. Ensamblado de automóviles usando brazos de robot | © *Latin Stock México.*

Concepto

Los robots de servicio son sistemas de software y hardware que consisten en una serie de dispositivos electrónicos y electromecánicos que se ubican en ambientes dinámicos y complejos. Todas estas características les dan una cierta autonomía, que les permite tomar decisiones a partir de una representación interna del mundo. Así, los robots deben tener dos capacidades básicas: adaptabilidad, para reaccionar en forma oportuna y apropiada a sucesos imprevistos, modificadores de su medio, y determinación para escoger las acciones apropiadas para lograr sus objetivos.

Figura 8. Robots de servicio en una casa, en un futuro cercano | © Denoir.

Curiosidades

TPR8 fue diseñado para competir en RoboCup, una competencia mundial cuyo objetivo es que en el año 2050 el campeón del mundo de futbol tenga un partido con el campeón del mundo de futbol del representativo de robots y que éstos ganen el partido. Para lograr este objetivo, cada año se realiza esta competencia con diferentes categorías. En la categoría denominada Junior, participan jóvenes entre 12 y 19 años, los cuales compiten con sus robots en pruebas de futbol y de rescate, en donde los robots realizan actividades en lugares dañados por desastres naturales, como terremotos.

Figura 9. Robot Asimo fabricado por Honda | © Xavier Caballé.

Figura 10. Robot TPR8, desarrollado en el laboratorio de Bio-Robótica, Facultad de Ingeniería, UNAM | © Jesús Savage.

El ambiente virtual

Una interfaz gráfica de tres dimensiones, la cual utiliza técnicas de *graficación* para poder representar objetos de tres dimensiones en monitores planos, permite visualizar diversos robots virtuales, que son una simulación veraz de los reales: pueden aparentar las mismas órdenes, con ligeras variaciones (véase la figura 11).

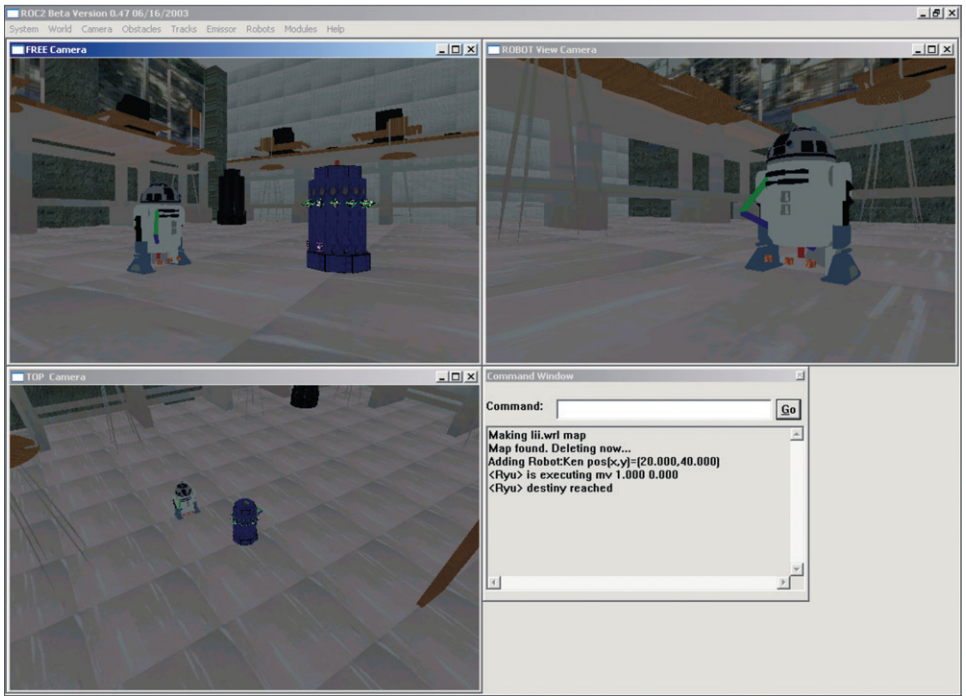


Figura 11. Robot real y virtual | © Jesús Savage.

Mientras un robot utiliza sensores internos y externos para captar su entorno, en el robot virtual se utilizan modelos matemáticos para simular el entorno.

Las tareas del robot y la interfaz hombre-robot

A lo largo del día, el robot debe realizar diversas tareas de acuerdo con la hora de su programación. En este camino, es muy importante la interacción con los humanos para que los robots sean capaces de reconocer en forma natural las órdenes dadas por una persona.

Percepción

El robot obtiene una representación simbólica de los datos que vienen de los sensores, de las tareas del robot y de la interfaz hombre-robot. Con esa representación simbólica se genera una *creencia*. En la figura 12, la representación simbólica genera dos postulados dudosos: *hay un agujero adelante del robot* o *hay una sombra enfrente de él*. ¿Qué debe hacer el robot? Si, en efecto, hay un agujero enfrente no debe avanzar o caerá en él. Es fácil ver que el robot requiere de más información; usualmente estos sistemas incluyen mapas del medio ambiente y alguna manera de localizar su posición en estos mapas.

Concepto

En temas anteriores se han evaluado distintas técnicas de búsquedas básicas en inteligencia artificial. Por ejemplo, para generar (y resolver) el problema de laberintos perfectos, en el tema sobre programación se utilizó la búsqueda a profundidad. Ésta y muchas otras desarrolladas a lo largo de los años, permiten seleccionar la ruta más adecuada entre conjuntos muy grandes de opciones. Estas búsquedas se dividen en dos: búsquedas ciegas o de fuerza bruta, como la búsqueda a profundidad o amplitud, y las búsquedas heurísticas: escalada simple, máxima pendiente, minimax, alfa-beta, etcétera.

Toma de decisiones

Una vez que el robot obtiene toda la información posible de sus sensores, de su posición, etc., entonces debe decidir un curso de acción. Para simular razonamiento, se utilizan diversas técnicas de *inteligencia artificial* y, dado un cierto conjunto de valores de entrada, localiza y realiza la acción más apropiada. En el ejemplo del robot frente a la sombra, una vez que determina que en efecto es una sombra, entonces selecciona la acción de *proseguir el camino*.

Figura 12. Sombra enfrente del robot.

Aprendizaje

Sin embargo, para poder convivir con los seres humanos, un robot debe tener dos capacidades más: corregir sus errores y aprender cosas nuevas. En la actualidad existen varios métodos para que los sistemas artificiales aprendan, como son los algoritmos genéticos, las redes bayesianas, neuronales y artificiales, así como la programación genética.

Curiosidades
Existen pequeños robots que vuelan, como si fueran moscas. En su artículo "Fly, Robot Fly", Robert Wood describe este tipo de robots miniatura que vuelan todavía a nivel de prototipos, pero que tendrán un número considerable de aplicaciones: en operaciones de búsqueda y rescate, en ambientes inhóspitos para el hombre, en exploración y monitoreo, en inspección de construcciones y otras.

9.3.2 Los robots en la literatura

Los robots han aparecido en la cultura popular desde hace muchos años, en el cine, el teatro, la televisión y la literatura. A continuación se presenta una selección de algunos de los libros más representativos en donde aparecen robots.

Frankestein o el moderno Prometeo, de Mary Shelley (1818), da inicio a una tradición en donde el ente creado se revela contra su creador.

En *RUR (Robots Universales de Rossum)*, de Karel Capek (1924), por primera vez se utiliza la palabra *robot*, que en checo significa trabajo forzado. Sus robots en realidad eran androides.

Ray Bradbury, en uno de los cuentos del libro *Crónicas marcianas* (1951), describe una casa completamente automatizada en donde robots tipo ratón se encargan de hacer la limpieza.

Stanislaw Lem publica *Ciberiada* (1965), donde introduce a los constructores Trurl y Clapaucio. En 1971 publica *Memorias de un viajero estelar* en donde Ijon Tichy encuentra diferentes tipos de robots en sus múltiples viajes. A través de cuentos cortos y viajes a diferentes planetas, estos personajes discuten cuestiones filosóficas con otros robots. Por ejemplo, en uno de los planetas visitados encuentran una secta monástica de robots que creen en Dios, y al hombre lo ven solamente como un instrumento de Dios para que ellos fueran creados.

Philip K. Dick publica varios cuentos en donde aparecen robots; *Sueñan los androides con ovejas eléctricas* (1968) es una de las novelas más notables, llevada al cine como *Blade Runner*, considerada una de las mejores cien películas de todos los tiempos y, ciertamente, un clásico en el cine de ciencia ficción.

Brian W. Aldiss, en el cuento "Supertoys Last All Summer Long", presenta robots humanoides (niños) que son utilizados por parejas infértiles; la película *Inteligencia artificial* está basada en este libro.

Odisea espacial, de Arthur C. Clarke (1968), describe una computadora que controla una nave espacial. Esta computadora tiene la capacidad de conversar en forma natural con los astronautas de la nave y debido a conflictos internos decide matar a los tripulantes.

Isaac Asimov publica por primera vez historias de robots en 1940 con el cuento corto "Robbie". En *The Complete Robot,* Asimov inventó el término robótica y también las tres leyes de la robótica que dominan el comportamiento de los robots:

1] Ningún robot dañará a un ser humano o permitirá, por inacción, que éste sufra daño.
2] Un robot obedecerá las órdenes de un ser humano siempre que éstas no contradigan la primera ley.
3] Un robot salvaguardará su propia existencia, siempre que no entre en conflicto con la primera o segunda leyes.

Desde el nacimiento de *Robbie*, el primer robot niñera, en los cuentos de Asimov, la compañía U.S. Robotics seguirá produciendo robots más sofisticados hasta llegar a los robots que tienen cerebros positrónicos.

De 1942 a 1983, Asimov escribe su serie de libros sobre imperios galácticos, la serie *Fundación*. Después de la muerte de Asimov, la serie de Robots-Fundación fue continuada por Gregory Benford.

9.4 JUEGOS

Para la mayoría, son conocidos los usos de la computación en los juegos de video, consolas de juego, juegos para computadora y, en general, aquellos diseñados para cualquier tipo de dispositivo personal, desde teléfonos celulares hasta computadoras de mano. Este tipo de juegos involucra diversos conocimientos fundamentales de cómputo, aunque su principal atractivo es la parte gráfica y de animación. Sin embargo, esta sección se centrará en juegos que representan retos computacionales *casi* imposibles de resolver, como el ajedrez o el go. ¿Por qué se dice que son "casi imposibles" de resolver? Si el lector ha jugado gato, damas inglesas, damas chinas o ajedrez, probablemente tiene una buena idea: porque en cada paso existen muchas opciones, pero eso no es todo, conforme se aprende a jugar cualquiera de éstos, se cae en la cuenta de que es necesario pensar "si yo muevo aquí, ¿qué hará mi oponente?" y entonces se evalúan algunas posibilidades de jugadas y luego las que hará el oponente y nuevamente cuál será el siguiente paso si eso sucede y así sucesivamente. ¿Cuántas posibilidades de movimiento existen, cuántos niveles *en el futuro* se evalúan para la siguiente jugada?

Este tipo de juegos, al igual que muchos ejemplos que se vieron en el primer tema, problemas difíciles, tienen un crecimiento exponencial, como el de las torres de Hanoi. Que no tengan una solución sencilla, sin embargo, no impedirá que se intente resolverlos, ¿o sí? De hecho, el ajedrez se considera, en términos computacionales, un problema resuelto. En 1997, un grupo de investigadores de IBM construyó una computadora llamada Deep Blue que le ganó a Garry Kaspárov, el ajedrecista evaluado con el más alto nivel de toda la historia.

Para comprender el tamaño del reto, se revisarán algunos números interesantes, los cuales se muestran en la figura 13.

Curiosidades
El campeón humano-máquina de damas es también un sistema de cómputo: Chinook. Este programa nació en 1989 y en 1990 ganó el derecho para competir en un campeonato mundial. En 1992 llegó a la final. Sin embargo, en 1994 ganó el campeonato y dos años después fue retirado porque estaba clara su superioridad sobre cualquier humano. Chinook ganó un campeonato mundial tres años antes que Deep Blue en el ajedrez y entró en 1996 al libro de récords Guinness como el primer programa en lograr tal hazaña. El juego de damas tiene aproximadamente 5×10^{20} posibles posiciones, lo que representa un reto muy grande. Sin embargo, Chinook, a través de una combinación de estrategias de inteligencia artificial y bases de datos especializadas para los inicios y finales de juego, es capaz de jugar de manera perfecta; es decir, que nunca pierde, y si su oponente juega también perfecto, entonces empatan. ¡Chinook calcula el resultado para todas las posibles posiciones en un juego!

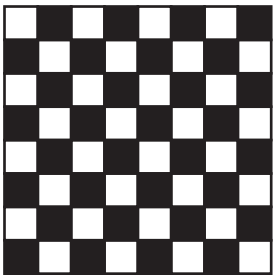
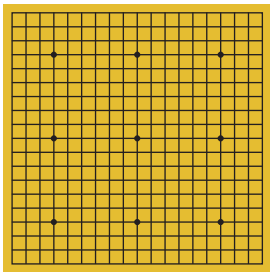
Ajedrez		Go
		
Cuadriculado		
8 x 8		19 x 19
Número promedio de opciones por turno		
35		200-300
Longitud del juego típico		
60 movimientos		200 movimientos
Número de posibles posiciones de juego		
10^{44}		10^{170}
Explosión de opciones (inicio típico)		
35	Movimiento 1	200
1 225	Movimiento 2	40 000
42 875	Movimiento 3	8 000 000
1 500 625	Movimiento 4	1 600 000 000

Figura 13. Juegos.

9.4.1 Ajedrez, un juego difícil

Curiosidades
Las tradicionales competencias de ajedrez iniciaron en el siglo XVI y el primer campeón mundial, Wilhelm Steinitz, se coronó en 1886. El campeón actual es el indio Viswanathan Anand.

El ajedrez es un juego recreativo y de competencia para dos jugadores. La variante actual del juego de ajedrez nació en Europa en la segunda mitad del siglo XV, pero está basado en versiones más viejas provenientes de Persia y la India. Hoy en día existen millones de jugadores de ajedrez alrededor del mundo.

El tablero de juego es un cuadrado de 8×8, con 64 posiciones. Al inicio del juego, cada jugador, uno a cargo de las piezas *blancas* y otro a cargo de las *negras*, controla 16 piezas: un rey, una reina, dos torres, dos caballos, dos alfiles y ocho peones. El objetivo del juego es hacerle *jaque mate* al rey del oponente, lo que significa que dicho rey está en inminente peligro de ataque o en *jaque* y no tiene posibilidad de escapar en la siguiente jugada.

Las reglas del ajedrez son complejas y la manera de *evaluar* qué tan bueno es un movimiento es una tarea complicada. Por supuesto, para un sistema de cómputo decidir su siguiente jugada, es un reto aún mayor, pues hay que simular la manera en la que una persona juega, evaluar qué tan bueno es cada movimiento y ejecutar el mejor posible, pero todo esto debe hacerse en un periodo corto de tiempo. Recuérdese que un juego típico dura entre 10 minutos y una hora e involucra unos 60 movimientos.

Al igual que en las otras aplicaciones de la computación que se han revisado, estudiar un juego con tantas posibilidades, como el ajedrez, ha impulsado el estado del arte en diversos temas relacionados con algoritmos y cómputo distribuido. Hoy en día, la computadora más poderosa para jugar ajedrez se llama *Hydra* y se estima que tiene una calificación Elo de más de 3 000, con lo cual, y éste es el propósito del proyecto *Hydra*, una computadora domina el mundo del ajedrez, con el reconocimiento de los humanos.

Hydra es un nombre tomado de la mitología griega que hace alusión a un monstruo de nueve cabezas que peleó (y fue derrotado) por Hércules. Pero en esta ocasión, no es un monstruo mitológico, sino un poderoso sistema de cómputo para jugar ajedrez patrocinado por el grupo de empresas PAL.

Tan monstruosa como su contraparte mitológica, *Hydra*, el monstruo de los tableros de ajedrez, acorralará a cualquiera que se cruce en su camino. Está configurada para envolver el mundo del ajedrez en cuatro versiones, todas con contrapartes en la mitología griega, Orthus (Ortos), Chimera (Quimera), Scylla (Escila) y, la más temible de todas, *Hydra* (Hidra).

Orthus es un horroroso perro de dos cabezas que cuidaba el ganado de Geryon y aterrorizaba a aquellos que se atrevían a cruzar sus campos. La Chimera es un monstruo con cabeza de león, cuerpo de cabra y cola de dragón, que arrasó con la ciudad en la que vivía. Scylla es un monstruo con seis cabezas de perro, vivía en una cueva y se dedicaba a atacar y aterrorizar marinos. En el caso del ajedrez, después de los monstruos ya descritos, Hydra es la forma perfecta y atacará al mundo del ajedrez con la fuerza de sus nueve cabezas de estrategia y velocidad. Una serpiente venenosa y enorme, capaz de infundir terror en cualquier persona. Cada vez que una cabeza es dañada, otras dos crecen en su lugar.

¿Podrá alguien enfrentar su fuerza sin par, este reto con múltiples cabezas, un duelo monstruoso? En la Antigüedad, los monstruos legendarios fueron derrotados por héroes como Hércules y Aquiles. Hoy, sin embargo, los monstruos han vuelto, están en el tablero de ajedrez y fueron creados por el hombre.

El primer sistema de cómputo en hacer historia en el mundo del ajedrez fue Deep Blue, que logró vencer en un torneo internacional en seis juegos a Garry Kaspárov en 1997, con dos victorias, una derrota y tres empates. Sorprendentemente, la principal fortaleza de Deep Blue era su *fuerza bruta*: era un equipo masivamente paralelo con 30 nodos basados en el equipo RS/6000 de IBM, que incluía 480 procesadores especiales para jugar ajedrez. El programa de Deep Blue para jugar ajedrez fue escrito en el lenguaje de programación C. ¿Qué tan poderosa era Deep Blue en términos de ajedrez? ¡Era capaz de evaluar 200 millones de posiciones por segundo! En 1997, Deep Blue estaba considerada como la supercomputadora número 259 en la lista de las más poderosas del mundo y era capaz de calcular 11.38 gigaflops. En computación se utilizan los *flops* como medida del rendimiento de un sistema de cómputo; son siglas en inglés de operaciones de punto flotante por segundo. Giga equivale a 10^9 flops, es decir, Deep Blue era capaz de calcular 11 380 000 000 flops.

Eso es mucho poder de cómputo, pero aun así suena insuficiente, ya que en un juego normal y siguiendo la técnica que se explicó arriba, donde se intentaron todos los movimientos posibles de nuestras piezas, se evaluaron las nuevas posiciones generadas, luego se estudiaron todos los posibles movimientos del contrincante para cada uno de los posibles movimientos que se hicieron y se evaluaron nuevamente las posiciones resultantes y así sucesivamente. ¿Qué se genera con este tipo de evaluaciones? Un árbol donde cada rama representa una posible partida; un jugador realiza el primer movimiento indicado

Curiosidades
El sistema de medición Elo fue creado por Arpad Elo, un profesor húngaro-americano de física, y permite evaluar de manera relativa los niveles de destreza de jugadores de ajedrez y go. El esquema Elo fue adoptado por la Federación Internacional de Ajedrez (FIDE, por sus siglas en francés), y aunque hoy se utiliza una variante de Elo, usualmente se conoce la calificación de un jugador como su FIDE o Elo, de manera indistinta. Solamente cuatro jugadores en la historia del ajedrez han obtenido más de 2 800 puntos: Garry Kaspárov de Rusia, Vladimir Kramnik de Rusia, Vaselin Topalov de Bulgaria y Viswanathan Anand de la India.

en esa rama, el contrincante contesta con el segundo, luego el jugador efectúa el tercero, él el cuarto, etcétera.

¿De qué tamaño es el árbol que se genera? De aproximadamente 10^{60} posiciones, que es un número tan grande que es más o menos mil veces más que el número de átomos de hidrógeno en el Sol. De manera que, aunque Deep Blue puede calcular 200 millones de operaciones por segundo, no podría calcular en un tiempo razonable las 10^{60} posiciones. Entonces, ¿qué se puede hacer? ¿Cómo logró vencer al campeón del mundo? Un punto intermedio es revisar sólo hasta cierto nivel en el árbol. Deep Blue *bajaba* y evaluaba las distintas posiciones usualmente hasta el nivel 12 y, en ciertos movimientos cruciales, llegaba hasta el nivel 40 del árbol, lo que significaba unos 170 millones de hojas en el árbol por segundo. La evaluación que hacía Deep Blue es muy básica, pero suficiente para servir de guía, y evaluaba la cantidad de *material* presente en el tablero, utilizando el valor estándar de cada pieza, por ejemplo, un peón vale un punto, un alfil o un caballo tres y así sucesivamente.

De acuerdo con cálculos posteriores al triunfo de Deep Blue, se encontró que con cada nivel extra en el árbol que pudiera explorar el sistema, subía entre 200 y 300 puntos su evaluación Elo. Este resultado permite concluir que el ajedrez puede resolverse mediante la computación. No sólo eso, muestra que aun para resolver problemas que parecen *imposibles*, como el ajedrez, algo de ingenio combinado con el poderío de las computadoras, esa fuerza bruta que se mencionó anteriormente y que, en resumen significa intercambiar el juicio por la búsqueda, produce soluciones aceptables.

9.4.2 Go, un juego imposible

Go, al igual que el ajedrez, es un juego de estrategia para dos jugadores. Es un juego muy viejo. Las primeras referencias escritas acerca del go datan del siglo V antes de nuestra era en China, pero se originó muchos años antes. Es un juego muy popular en Asia y ha cobrado relevancia en el resto del mundo en años recientes.

El go se juega en un tablero cuadriculado de 19×19 utilizando fichas de colores blancas y negras, conocidas como *piedras*. Las piedras se juegan en las intersecciones de las líneas. Y, aunque las reglas son muy sencillas, se utilizan estrategias complejas durante un juego estándar. El objetivo del juego es controlar una mayor parte del tablero que el oponente. Con este objetivo en mente, los jugadores intentan colocar sus piedras de manera tal que no puedan ser capturadas y que encierren partes del territorio donde las piedras del oponente pueden ser capturadas. Se dice que una o varias piedras están capturadas si no tienen grados de libertad, es decir, si están completamente rodeadas por piedras enemigas. Las piedras capturadas son destruidas y cuentan a favor del contrario.

El problema es pues la cantidad de opciones para posicionar piedras y encontrar un balance entre la defensiva y la ofensiva. De hecho, para un sistema de cómputo existen dos problemas principales: el árbol de análisis es mucho mayor que el del ajedrez, aproximadamente 10 veces mayor. El segundo problema es la evaluación de las posiciones finales. ¿Qué tan bueno es colocar una piedra en una cierta posición? Obviamente contar las piezas en ese momento siempre sumará uno, pero este valor no es suficiente, porque no toma en cuenta su aporte para defender otras piedras o para expandir el territorio. En el caso de Deep Blue, arriba, se vio que utilizó una estrategia de evaluación muy sencilla, simplemente contando las piezas. Para evaluar el go se debe ser más creativo.

9.4.3 Hacia una solución

A primera vista, los dos problemas que se mencionaron antes no tienen solución, pero se pueden imaginar maneras de darle la vuelta al problema. Comenzando por el árbol de análisis, si se quieren evaluar todas las posibles posiciones, bajando 12 niveles en el juego como lo hacía Deep Blue, se necesita evaluar cada posición un millón de veces más rápido, cosa que con la tecnología actual no es posible. ¿Qué opciones se tienen, qué se puede hacer para no evaluar todas las posiciones?

Por supuesto, es necesario recortar el árbol de análisis, decidir lo antes posible cuáles de las ramas en el árbol prometen un mejor futuro y cortar las demás. Los juegos de ajedrez utilizan una técnica de búsqueda conocida como alfa-beta, que hace justo eso, suspende la evaluación de un movimiento en cuanto es notorio que hay otros que son mejores. Por ejemplo, supóngase que un movimiento nos lleva a una posición en la que existe posibilidad de ganar el juego; si luego se evalúa otro movimiento en el que el enemigo puede empatar, se suspende la evaluación y se elimina el movimiento de nuestras opciones.

Alfa-beta permite recortar el espacio de búsqueda en un factor de raíz cuadrada del número de posiciones. En este caso, si se quiere evaluar hasta 12 movimientos en el futuro, entonces en lugar de las 38^{12} posiciones de fuerza bruta, sólo se deben evaluar 4×10^9 o cuatro mil millones de posiciones.

Con eso se acorta el número de posiciones, pero sigue siendo un árbol muy grande. Se puede recurrir a otra técnica de búsqueda que se llama recorte de movimiento nulo, donde lo que se hace es imaginar que se deja pasar un turno y permitir que el contrincante mueva dos veces seguidas. Si después de este tipo de ventaja, una posición sigue siendo buena para nosotros, se puede suspender la búsqueda, porque se ha encontrado un punto de quiebre, un punto donde es posible recortar la rama de evaluación.

Utilizando una mezcla entre alfa-beta y el recorte de movimiento nulo de manera recursiva, se puede lograr un ahorro estimado de la raíz cuadrada de los movimientos. Con esto, con un equipo de cómputo como Deep Blue se pueden evaluar hasta 12 niveles en el juego de go. Por supuesto, esto no incluye la evaluación compleja de cada posición, pero es un avance.

Curiosamente, el autor del algoritmo de recorte de movimiento nulo, Murray Campbell, fue uno de los miembros clave del equipo que construyó Deep Blue, pero no utilizó esta técnica para el ajedrez porque sus reglas impiden los movimientos nulos o *pasar*, pero esto no sucede en el go, donde sí es válido realizar un movimiento nulo.

Optimizar la evaluación de una posición en go también es posible utilizando una técnica muy socorrida en los sistemas de cómputo, la memoria caché, en la que se puede almacenar la evaluación de una fracción del juego y sólo cuando sea necesario, por ejemplo porque se agregaron piedras a esa región o una región contigua, se vuelve a evaluar esa sección.

Hoy día, con la tecnología actual de microprocesadores, es posible en teoría construir una computadora en un solo chip, que sea 100 veces más rápida que los 480 procesadores de Deep Blue. Más aún, si se instalaran 480 de estas poderosas computadoras en una arquitectura paralela, sería 100 veces más rápida. La ley de Moore, que dice que el número de transistores que se pueden integrar en un circuito crece exponencialmente y se duplica cada dos años, soporta la idea de que dentro de 10 años más se logrará otro incremento de 100 veces.

En resumen, la esperanza de diversos expertos es que durante el siguiente decenio será posible construir una computadora que pueda buscar más de tres billones de posiciones

Curiosidades
Existen varios métodos para generar números aleatorios en una computadora. Pero, realmente no son aleatorios, son seudoaleatorios porque responden a una función y parecen tener comportamiento de números aleatorios. La función generalmente se llama RAND.

por segundo, aproximadamente un millón de veces más rápida que Deep Blue y, por supuesto, que esta computadora podrá jugar go al más alto nivel y ganarle a cualquier humano.

9.4.4 Generación de sólidos con voxeles

Definitivamente, la computación es una ciencia que hace realidad muchas de las cosas que en un inicio sólo se imaginan, al ofrecer las herramientas necesarias para llevarlas a cabo. Por ejemplo, la imagen mostrada en la figura 14 se basa en la siguiente idea: se desea mostrar simbólicamente a todos los alumnos de bachillerato de la UNAM en una sola imagen formando el escudo universitario. Así, cada alumno es representado simbólicamente por un voxel, es decir, un cubo. Entonces, el número de alumnos de bachillerato[2] de la UNAM es de 106 913, es decir, $n = 106\,913$.

El programa que se desarrolló para llevar a cabo esta idea toma una imagen binaria del escudo universitario y le da volumen, es decir, lo convierte en una imagen 3D, con profundidad. Calcula el volumen total V del escudo y lo divide sobre n. Finalmente, para sacar la longitud l de cada segmento de cada voxel, le saca la raíz cúbica a la división mencionada anteriormente. Debido a cuestiones matemáticas de redondeo, a veces el programa no obtiene exactamente el número de voxeles solicitados; en ese caso una rutina del mismo programa "lo obliga" a obtener siempre el valor de n solicitado por medio de "agregar" voxeles o "quitar" los voxeles menos significativos de las orillas (generalmente son muy pocos). Con el objetivo de mejorar la representación en 3D del escudo universitario: a cada cara de cada voxel se le asigna un color aleatorio. Así, la figura 14 muestra el escudo universitario compuesto de 106 913 voxeles. Si se observa el escudo formado por capas de voxeles, cada capa es de 13 751 voxeles. Así, el escudo está formado por un poco más de siete capas, hasta completar el número exacto solicitado. Las caras de los voxeles están mostradas por diferentes tonos de color amarillo generados de manera aleatoria.

Figura 14. El escudo de la UNAM formado por 106 913 voxeles. Los voxeles representan simbólicamente a los alumnos de bachillerato. Las caras de los voxeles están mostradas por diferentes tonos de color amarillo generados de manera aleatoria | *© Ernesto Bribiesca.*

La figura 15 muestra también el escudo de la UNAM, sólo que ahora en diferentes colores seleccionados de manera aleatoria para cada una de las caras de los voxeles. Lo interesante de esta imagen es que da una idea del tamaño de Universidad basado en el número de alumnos de bachillerato. En lugar del escudo de la UNAM se podría generar casi cualquier tipo de imagen utilizando esta misma técnica.

[2] Dato tomado de la *Agenda Estadística 2006* de la Dirección General de Planeación de la UNAM.

9.5 BIOINFORMÁTICA

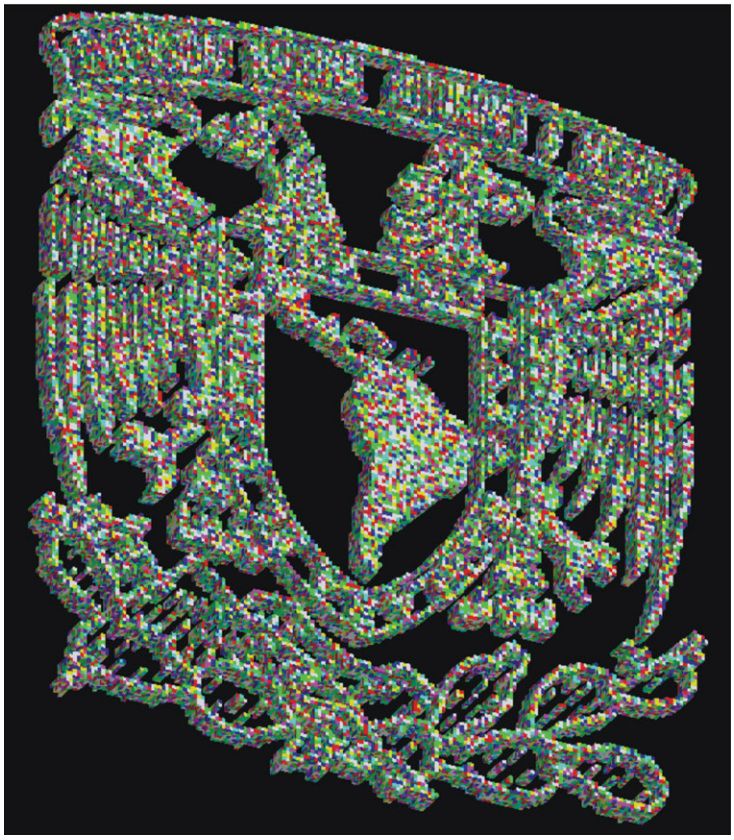


Figura 15. El escudo de la UNAM formado por 106 913 voxeles. Representación simbólica de los alumnos de bachillerato formando el escudo de la UNAM por medio de voxeles, usando diferentes colores de manera aleatoria | © Ernesto Bribiesca.

No siempre las recompensas son expeditas. En 1962, un hombre de 37 años, originario de Chicago, está parado al otro lado del mundo, en la célebre *Aula* de la Universidad de Estocolmo; la audiencia aplaude, acaba de ser presentado por el rey de Suecia como el galardonado con el Premio Nobel de Fisiología o Medicina; su nombre es James Dewey Watson y el trabajo que lo hizo acreedor al premio fue el descubrimiento de la estructura molecular del ADN, que llevó a cabo en colaboración con su colega Francis Crick en 1953, nueve años antes, a sus 28.

En el quehacer científico siempre se están haciendo lecturas interesantes de los fenómenos; claro está que la lectura profunda, la verdadera, la que permite apreciar la obra completa, no se logra a la primera, de hecho puede ser que nunca se alcance, los científicos la persiguen mediante arduas aproximaciones sucesivas: la lectura que hizo Newton de la naturaleza nos develó parte de la misteriosa trama, se creía completamente descifrada, hubo que esperar poco más de siglo y medio la relectura más exacta de Einstein. Nunca se puede estar seguro de haber hecho la lectura final.

Tanto peor es la situación cuando no hay que leer, cuando no hay libro que descifrar, cuando ni siquiera se tiene la fortuna de saber que algo reclama nuestra lectura.

En 1990, 28 años después de la ceremonia en Estocolmo, la curiosidad de Watson permanece incólume, en ese año decide encabezar uno de los mayores proyectos científicos de la historia: determinar qué se debe leer para saber qué es un ser humano, encontrar el libro de donde se habrá de extraer el significado, la esencia, biológica al menos, de lo que significa un ser humano. Con el patrocinio del Departamento de Energía y de los Institutos Nacionales de Salud de Estados Unidos, se funda el Proyecto del Genoma Humano, encabezado por el doctor Watson.

El proyecto pretende determinar la secuencia completa de nucleótidos que constituye el código genético de los seres humanos en general. Transcurrirán 13 años antes de que el trabajo esté completo, en 2003. Watson ya no encabezará los esfuerzos luego de su salida del proyecto en 1992: él cree que el libro, una vez develado, le pertenece a la humanidad, cree que no debe ser privilegio de unos cuantos poseer su propia explicación; el director de los institutos de salud piensa, en cambio, que le pertenece al que lo descubre, que se pueden patentar genes, que es el bibliotecario y no el lector el dueño de la biblioteca; el conflicto es insoluble, Watson renuncia.

A raíz del proyecto del genoma humano y haciendo uso de las técnicas de biología molecular desarrolladas durante tal proyecto, se han secuenciado muchos más genomas. El número y la variedad de los seres vivos cuyo código genético se ha develado es enorme y continúa creciendo aceleradamente, todos los días; de hecho varias veces al día se actua-

lizan los repositorios en los que se almacenan las secuencias. El acervo de datos que se posee es inmenso y se mantiene en continuo crecimiento.

Pero tener qué leer no sirve de mucho si no se lee. Es aún muy poco lo que se puede aprovechar de los datos que se tienen, comparado con el tamaño de éstos. Hacen falta herramientas de análisis. Es allí donde entra en el juego la computación.

Son muchas las cosas que se quisieran saber a propósito de los genomas disponibles y de los organismos definidos por ellos. Interesa saber para qué sirve cada segmento del código genético; en él se codifican las proteínas que determinan la función biológica de las células que lo poseen. Las células del páncreas tienen un programa detallado que les dice cómo producir insulina; las células cancerosas "olvidan" para qué fueron hechas y olvidan morir, porque hay un segmento de su código genético que ha sido alterado, se reproducen prolíficamente y producen proteínas que envían mensajes equivocados a otras células. Si se pudiese leer eficientemente el código genético, se podría percibir más fácilmente a las que han olvidado su labor y sus mensajes, sería fácil eliminarlas; de hecho, se lograría que ciertas enfermedades, cuya propensión se encuentra en los genes, nunca se desarrollen y diseñar medicamentos "a la medida" para las personas de acuerdo con su perfil genético.

Para hacer esto se requiere entender lo que se dice en las secuencias almacenadas en las enormes bases de datos internacionales en las que se guardan los genomas de mamíferos, plantas, bacterias, virus y personas. Es necesario analizar los datos allí contenidos de forma tal que permitan obtener información útil.

Una buena parte de la computación aplicada a la biología molecular, lo que suele denominarse *bioinformática*, consiste en diseñar algoritmos que permitan comparar cadenas de símbolos en un alfabeto determinado. Si las secuencias son de ADN, por ejemplo, el alfabeto está constituido por los símbolos de las cuatro bases o nucleótidos que lo forman: A (adenina), T (timina), C (citosina) y G (guanina); si en cambio las secuencias son proteínas, el alfabeto está hecho de los 20 símbolos usados para representar aminoácidos.

Comparar cadenas de símbolos no es una labor particularmente difícil en computación. De hecho, es bastante simple. Si se pretende, por ejemplo, encontrar el segmento de cadena más grande posible que resulta ser común a un conjunto de cadenas, el algoritmo para llevar a cabo esto posee una **complejidad polinomial**. Sin embargo, lo realmente interesante en bioinformática es hacer comparaciones inexactas; generalmente se pretende encontrar el segmento más largo posible que resulta ser común, salvo unas cuantas diferencias, a un conjunto de cadenas. Esto de "salvo unas cuantas diferencias" es lo que trae los problemas; los algoritmos para hacer comparaciones inexactas pueden resultar sumamente costosos en tiempo, el problema subyacente es de los que se han llamado *intratables* en este libro.

Los computólogos dedicados a la bioinformática están permanentemente buscando métodos que se aproximen a la solución de este y otros problemas intratables. Algunos de hecho recurren a técnicas de inteligencia artificial para lograrlo.

Pero… ¿por qué interesa hacer comparaciones inexactas? Resulta que, de acuerdo con lo que se sabe hasta ahora, el motor de la evolución de los organismos vivos son las mutaciones que, como sabe el lector, consisten en alteraciones fortuitas en el código genético, tales como la inserción ocasional de un símbolo que originalmente no estaba allí o la desaparición de uno que sí estaba o el cambio de uno por otro. A lo largo de millones de generaciones estas mutaciones se acumulan y entonces los organismos emparentados en la cadena evolutiva pueden tener códigos muy diferentes a simple vista, pero en realidad, salvo algunas alteraciones, son similares. Determinar este parentesco puede ser de gran

importancia porque bien pudiera ocurrir que uno tenga propiedades que lo distinguen del otro y determinar a qué se deben las diferencias.

El código genético es complejo, hay en él tramos de secuencia que realmente no codifican proteínas y que se ignora para qué puedan servir, aparentemente son vestigios inútiles de generaciones pasadas, pero nunca se sabe, sería bueno averiguarlo realmente. Otros tramos codifican proteínas que hacen que a su vez se produzcan otras proteínas, a esto se le llama *secuencias reguladoras* y de tener capacidad de detectarlas, se podría, por ejemplo, inhibir el efecto nocivo de algunos virus o detener la propagación del cáncer en un organismo. No se sabe si en algún momento se va a lograr, no se ve fácil y probablemente pasen muchos años antes de que se puedan recibir los beneficios potenciales de la investigación en bioinformática. No siempre las recompensas son expeditas.

9.6 LA COMPUTACIÓN EN LOS NEGOCIOS

Los negocios en la actualidad se pueden apreciar desde dos puntos de vista: uno tecnológico y otro social. Los administradores deben entender de cuestiones tecnológicas y manejar conceptos sociales en su negocio, para que puedan tomar decisiones encauzadas al logro de los objetivos de su organización. La tecnología tiene que ver con el cambio gerencial, ya que proporciona herramientas poderosas a los administradores. Con la computadora se han desarrollado sistemas de información que permiten obtener, analizar y comprender una gran cantidad de información.

El aspecto tecnológico es el que interesa tratar en este punto. En resumen, la importancia que tiene la computadora en los *negocios.*

El poder de las computadoras aumenta cuando son conectadas en red y más aún con internet; esto permite la distribución instantánea de información dentro y fuera del negocio. El correo electrónico y otras formas de comunicación en red permiten a los administradores ampliar su ámbito de control y dirigir trabajadores y organizaciones en cualquier lugar en el que se encuentren. Les permiten incluir más personal en el proceso de planeación, también para la autoorganización de equipos de trabajo, al mismo tiempo pueden mantener el contacto con los subordinados para supervisar su trabajo, desarrollar controles en tiempo real, crear, almacenar y diseminar los conocimientos de la organización. Las computadoras han transformado la estructura de las organizaciones, es decir, estas estructuras han reducido el número de niveles jerárquicos. Esto se debe a que sólo la información necesaria y oportuna se encuentra a disposición de los empleados de más bajo nivel jerárquico para que éstos puedan tomar decisiones sin importar que se encuentren en lugares distantes.

Como se puede observar en la figura 16, la empresa X representa un negocio compuesto de gerencias o áreas, las cuales a su vez tienen departamentos, que también tienen a su cargo subdepartamentos operativos. Esta estructura varía dependiendo del tipo, tamaño y giro del negocio de que se trate.

En cualquier organización o negocio se deben tomar decisiones sobre muchos asuntos que se presentan con regularidad (diariamente, a la semana, al mes, al trimestre, etc.), y para hacerlo es necesario que la información esté actualizada. Para lograr esto, se requiere del uso de la computadora. Con la computadora se desarrollan sistemas para obtener reportes bien estructurados que contengan información necesaria para la toma de decisiones que sirvan para el logro de los objetivos de la empresa. Algunas de las *razones más importantes que justifican el uso de las computadoras en los negocios son las siguientes:*

a] Su gran capacidad para realizar cálculos, ordenar y recuperar información.
b] Procesan grandes volúmenes de información.
c] Mayor rapidez para búsquedas de información complejas.
d] Exactitud para el manejo de operaciones aritméticas.
e] Salvaguardan la información y la mantienen disponible cuando cualquier usuario la requiera o sólo personas autorizadas.
f] Mejoran la comunicación dentro y fuera de la organización, ya que aceleran el flujo de información entre empresas remotas y personal interno.
g] Mejoran los servicios a los clientes mediante una mejor comunicación.
h] Pueden hacer que se lleven a cabo relaciones más estrechas con los proveedores, acordando mejores precios, cumpliendo con las fechas de entrega y pagos. También mejoran el procesamiento de facturas por medio del uso de procedimientos de control por lote.
i] Ayudan a dar un mayor seguimiento de los costos de mano de obra, bienes e instalaciones.
j] Introducen nuevos productos con investigación de mercados, utilizando bodegas y minería de datos.
k] Pronostican variables determinantes para el comportamiento futuro del negocio, para tomar medidas precautorias al respecto.
l] Utilizan modelos de simulación que representan situaciones críticas en el negocio, sin afectar el estado real del mismo.

Ahora se comentará cómo la computadora interviene para resolver problemas de información en cada una de las áreas de un negocio (gerencias y departamentos). En la Gerencia de Administración, en el Departamento de Personal, la computadora ayuda con el procesamiento de la información referente a: nómina, prestaciones, registro de personal, remuneraciones, relaciones sindicales, capacitación, seguimiento de empleados de la carrera en la empresa, planeación de personal, por mencionar algunas. En el Departamento de Contabilidad y Finanzas, la ayuda que brinda la computadora tiene que ver con: depósito de cheques, pago de impuestos, facturación, entrega de mercancías, presupuestos, diario, contabilidad de costos, cuentas por pagar/cobrar y manejo de fondos.

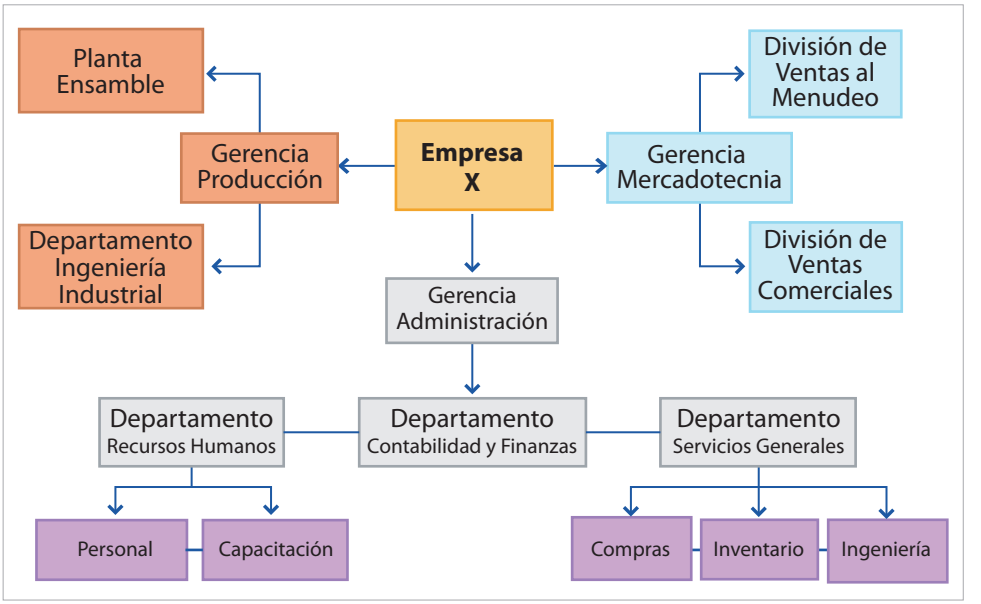


Figura 16. Estructura de las áreas y departamentos que constituyen la empresa X.

En el Departamento de Servicios Generales, con computadora se manejará la información de compras, el control de inventario para estimar el tamaño del lote y los servicios de ingeniería (véase la figura 17).

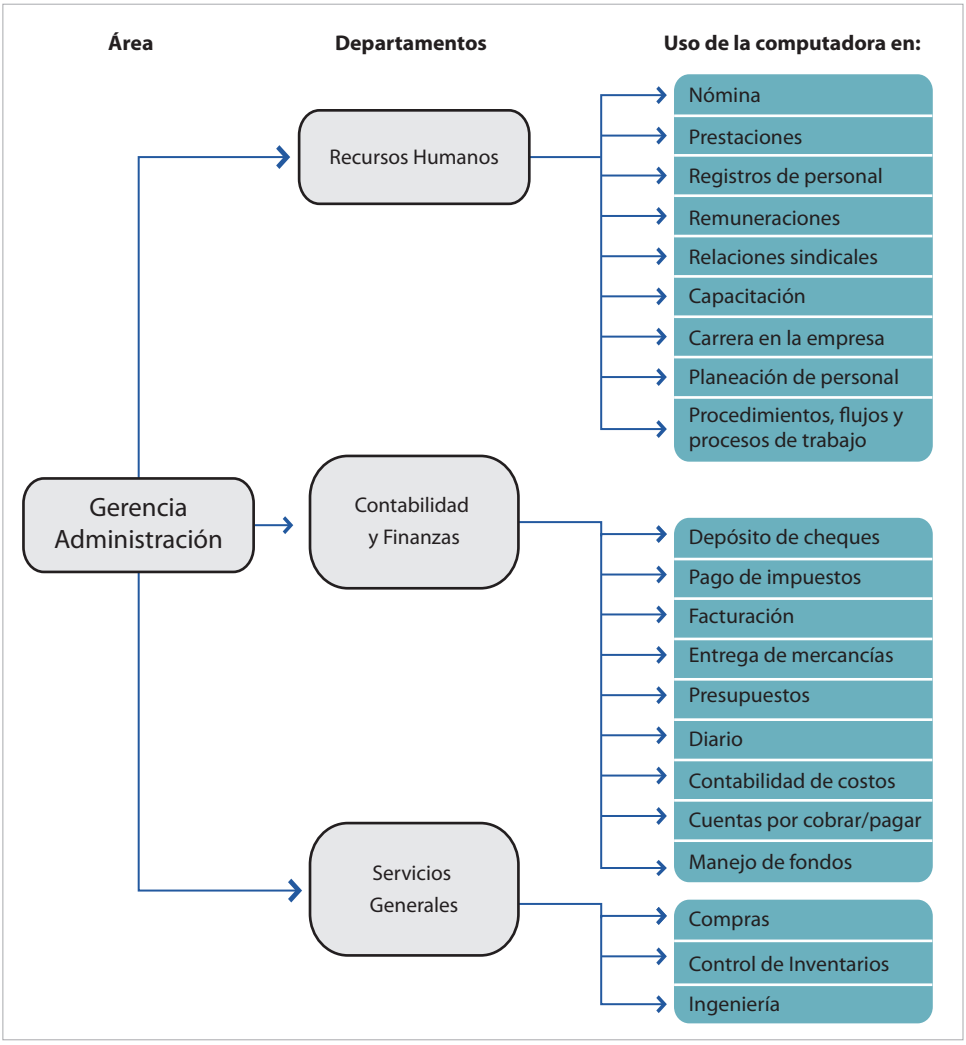


Figura 17. El uso de la computadora en la Gerencia de Administración de la empresa X.

En la Gerencia de Producción, en la Planta de Ensamble, la computadora facilita las tareas que tienen que ver con: programación de la producción, compras, recepción/embarques, control de pedidos, etc. En el Departamento de Ingeniería Industrial, en la planeación de recursos materiales, en operaciones, en el control de calidad, en la programación de tiempos y en la distribución de planta, por mencionar algunos (véase la figura 18).

En la Gerencia de Mercadotecnia, en el Departamento de Ventas, la computadora se usa en: la administración de ventas, la promoción de productos, el manejo de los precios, la introducción de nuevos productos, con un sistema de información de pedidos en tiempo real. En el Departamento de Estudios de Mercado se utiliza la computadora para llevar a cabo investigación de mercados y el estudio de productos piloto (véase la figura 19).

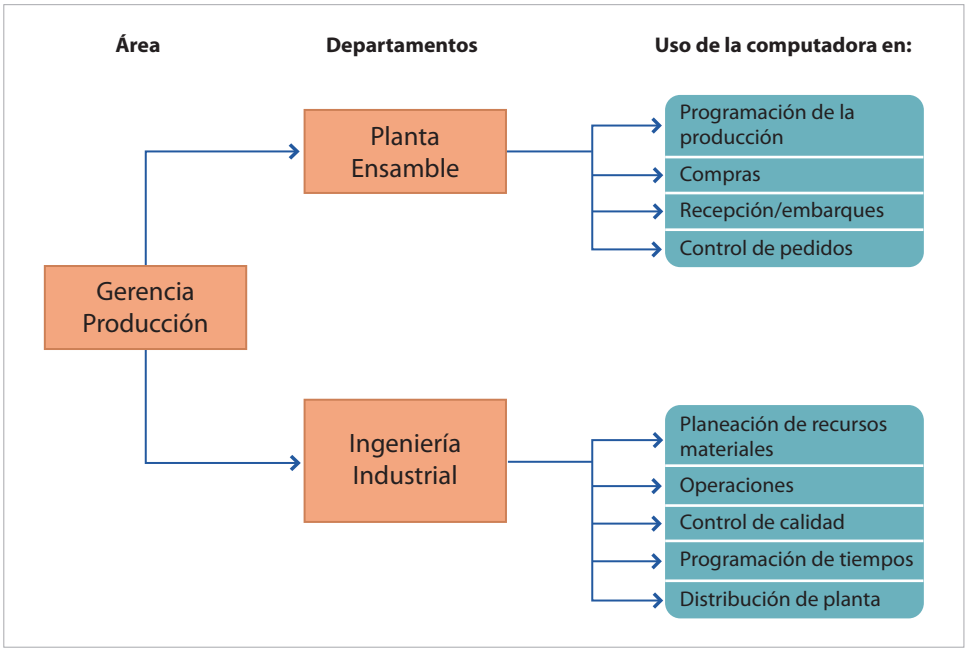


Figura 18. El uso de la computadora en la Gerencia de Producción de la empresa X.

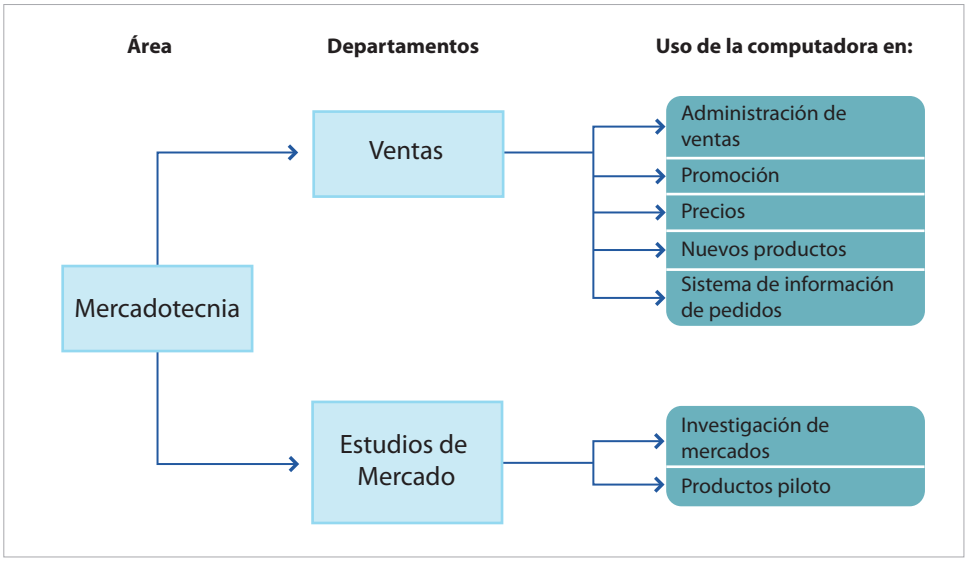


Figura 19. El uso de la computadora en la Gerencia de Mercadotecnia de la empresa X.

9.7 LA COMPUTACIÓN Y EL ARTE

La computación ha tenido una participación muy importante en las expresiones artísticas humanas. El arte digital es una disciplina creativa que abarca obras donde se hace uso de elementos digitales en las diferentes etapas, como en su producción o en su exhibición. Sin embargo, el concepto de arte digital ha sido debatido y rechazado por algunos círculos artísticos y puristas que lo clasifican más como una habilidad técnica que una expresión artística. Lo anterior puede deberse principalmente a un completo desconocimiento del arte computarizado, porque en definitiva la creación artística por medio de computadoras es un arte humano donde se expresan emociones, sensaciones e ideas con la única diferencia de que en vez de usar lienzos o paredes se usan monitores.

El arte digital tiene muchas ventajas, como la posibilidad de generar la obra por medio de parámetros variables establecidos por el artista. Así, la computadora puede producir una imagen o un sonido combinando los parámetros programados, previamente definidos con un componente de aleatoriedad. Una capacidad computarizada muy importante es la creación de mundos alternativos por medio del modelado tridimensional y de la programación de la física del entorno.

Como un ejemplo de arte con computadora se presenta un bello fractal (figura 20), generado por Adam Hauner.

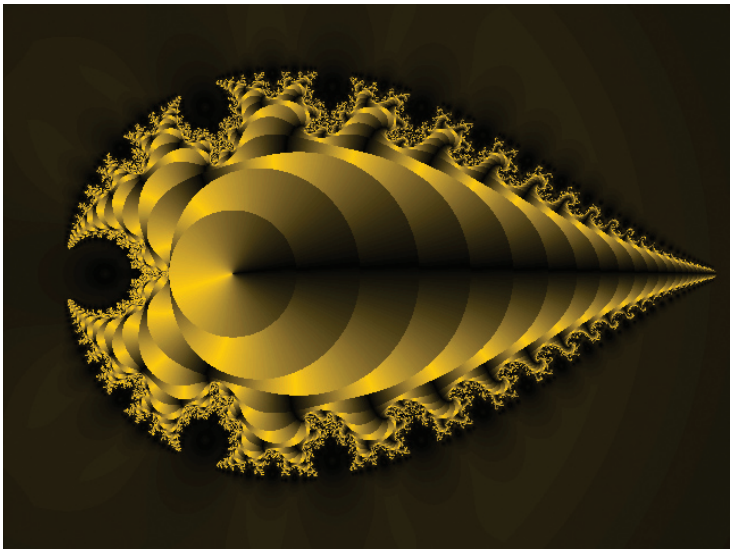


Figura 20. Fractal | © Adam Hauner.

9.8 RESUMEN

Las aplicaciones de las ciencias de la computación son muy amplias y variadas. En la actualidad es difícil encontrar un área donde no se utilice la computación. Es muy importante entender que los conocimientos fundamentales de la computación conllevan una nueva perspectiva para enfrentar diferentes problemas.

GLOSARIO

[En el texto estas palabras se indican en azul]

Algoritmo de ordenamiento por inserción: El algoritmo de ordenamiento por inserción tiene un tiempo de ejecución que crece de manera cuadrática con respecto al tamaño de su entrada, n, en el peor caso (como cuando el orden de la lista está totalmente invertido). Sin embargo, para ciertas entradas (cuando la lista está ordenada), su tiempo de ejecución es lineal. En general, al analizar un algoritmo se considera su tiempo de ejecución en el peor caso, ya que éste muestra el tiempo máximo que emplea el algoritmo en resolver el problema tomando en cuenta todas sus entradas posibles (en algunas le tomará más tiempo y en otras menos).

Algoritmo recursivo: Es un algoritmo que se define en términos de sí mismo. Dice algo así como: "para resolver el problema de tamaño n utiliza soluciones al problema de menor tamaño". De manera similar, una definición recursiva es aquella que está dada en términos de sí misma. Dice cómo obtener conceptos nuevos, usando el mismo concepto que desea describir. Al igual que con los algoritmos, para que no sea una definición circular, es necesario que esté planteada en términos de una versión más pequeña de ella misma.

Complejidad polinomial: Cuando el tiempo empleado para resolver un problema depende de alguna potencia del tamaño de la entrada, se dice que el problema es de *complejidad polinomial*. Al conjunto de todos los problemas de complejidad polinomial los computólogos lo bautizaron como **P**.

Conjunto de entradas de un algoritmo: Un algoritmo tiene un conjunto definido de entradas válidas, de distintos tamaños. Cada vez que se ejecuta el algoritmo, lo hace con un elemento de este conjunto. Cada elemento del conjunto de entradas tiene un tamaño. Mientras más grande sea la entrada, más recursos requerirá el algoritmo para producir una salida.

Curva: En matemáticas, el concepto de curva intenta capturar la idea intuitiva de un objeto continuo y unidimensional. En el estricto sentido, aunque esto ya no parezca obvio, una línea recta también es una curva.

Eficiencia de un algoritmo: Se puede evaluar qué tan bueno o eficiente es un algoritmo desde muchas perspectivas. Por ejemplo, cuánto tiempo emplea para resolver una tarea, cuánta memoria requiere, qué tan largo y difícil de entender es, etc. La complejidad de tiempo del algoritmo se define como el número de instrucciones a ejecutar, como función del tamaño de la entrada.

Ethernet: Es una familia de tecnologías de redes de computadoras basada en paquetes de datos de tamaño fijo, utilizada para conectar redes de área local. El nombre se tomó del concepto físico de éter, que se utiliza para describir un medio para transmisión de luz. Ethernet, en uso desde 1990, es el método de conexión más común. Las redes inalámbricas Wi-Fi, hoy día muy comunes en casas, aeropuertos y cientos de lugares públicos, son una extensión de Ethernet.

Fuente: Es un conjunto de imágenes que representan los caracteres de un estilo tipográfico particular. La tipografía es el arte y la técnica del diseño, manejo y selección de tipos.

Función lineal: Una función $f(n)$ es lineal si es de la forma cn, para alguna constante c, quizás más otra constante. Por ejemplo, $f(n) = 3n - 5$.

Grabación digital: Una grabación digital convierte la señal analógica del sonido en un flujo de números *discretos* que representan los cambios en la presión del aire a través del tiempo.

Gráfica: Es un conjunto de puntos llamados vértices, algunos de los cuales están conectados por líneas llamadas aristas.

Imagen: En términos formales, una imagen se puede definir como una función bidimensional $f(x,y)$, donde x y y son las coordenadas cartesianas y f la amplitud de la *intensidad* o *nivel de gris* de la imagen en ese punto de coordenadas. Cuando los valores de las coordenadas y el valor de la amplitud son todos finitos, es decir, poseen cantidades discretas, se trata de una *imagen digital.*

Modelo de cómputo: El modelo de cómputo especifica las operaciones que la máquina puede ejecutar para resolver un problema, así como el costo de ejecutar cada operación.

Módulo: Es un componente autocontenido de un sistema, que entre otras cosas tiene una interfaz bien definida acerca de cómo interactúa con otros módulos. En programación, diseñar y programar o construir un módulo puede ser una tarea compleja; sin embargo, una vez que un módulo existe, fácilmente puede conectarse o desconectarse del sistema.

Problema: Un problema consiste en la especificación de un conjunto de posibles entradas, y una relación que define las posibles salidas para cada entrada.

***Rendering*:** Es el proceso mediante el cual se genera una imagen a partir de un modelo utilizando programas de cómputo. El modelo es, como aquí se mostró, la descripción de un objeto tridimensional.

Robot de servicio: Son sistemas de software y hardware que consisten en una serie de dispositivos electrónicos y electromecánicos que se ubican en ambientes dinámicos y complejos. Todas estas características les dan una cierta autonomía, que les permite *tomar decisiones* a partir de una representación interna del mundo. Así, los robots deben tener dos capacidades básicas: *adaptabilidad*, para reaccionar en forma oportuna y apropiada a sucesos imprevistos, modificadores de su medio, y *determinación* para escoger las acciones apropiadas en el logro de sus objetivos.

Subrutina: Es la porción de un algoritmo que realiza una tarea específica y es relativamente independiente del resto del algoritmo. Dependiendo del lenguaje de programación y otras sutilezas, se puede llamar procedimiento, función o método. El uso de subrutinas dentro de un algoritmo o programa tiene muchas ventajas, como reducir la duplicación de secciones que hacen la misma tarea, reutilizar las subrutinas en otros algoritmos, descomponer algoritmos complejos en partes más simples y facilitar la comprensión y análisis del algoritmo.

Tamaño de palabra: Es el número de bits utilizado para representar una onda de audio única. El tamaño de palabra afecta directamente la *distorsión*. En la actualidad se utiliza como límite práctico un tamaño de palabra de 24 bits, ya que su relación señal-a-ruido excede el de la mayoría de los circuitos analógicos, que son utilizados en (al menos) dos puntos en el proceso de grabación y reproducción del sonido.

Tipos de búsqueda: Se han evaluado distintas técnicas de búsquedas básicas en inteligencia artificial. Por ejemplo, para generar (y resolver) el problema de laberintos perfectos, en el capítulo sobre programación se utilizó la búsqueda a profundidad. Ésta y muchas otras desarrolladas a lo largo de los años, permiten seleccionar la ruta más adecuada entre conjuntos muy grandes de opciones. Estas búsquedas se dividen en dos: *búsquedas ciegas* o de *fuerza bruta*, como la búsqueda a profundidad o amplitud, y las búsquedas *heurísticas*: escalada simple, máxima pendiente, minimax, alfa-beta, etcétera.

Tipo de dato abstracto: Un tipo de dato abstracto o ADT (por sus siglas en inglés) especifica un conjunto de datos y las operaciones para manipularlos. Estos tipos de datos son abstractos por su independencia de una implementación particular: no importa cómo se lleven a cabo las operaciones, sólo qué operaciones pueden hacerse y cuáles son sus efectos.

BIBLIOGRAFÍA

Varios de los libros que se citan a continuación fueron importantes para la elaboración de este texto.

KASNER, E. y J. NEWMAN, *Mathematics and the Imagination*, Dover Publications, 2001. Un libro para quienes gusten de las matemáticas y los retos.

PAPERT, Seymour, *The Children's Machine: Rethinking School In The Age Of The Computer*, Basic Books, 1994. Este y otros libros examinan el papel de las computadoras en la educación. Papert, un pionero de la computación, argumenta que las computadoras podrían ser usadas de manera mucho mejor para la educación infantil. Otros libros más recientes presentan también argumentos del peligro de abusar de las computadoras en las escuelas, como el de Todd Oppenheimer.

RHEINGOLD, Howard, *Tools for Thought, The History and Future of Mind-Expanding Technology,* MIT Press, 2000. Entre los libros de historia de la computación, éste destaca por su estilo ameno, contiene muchas anécdotas y descripciones de personajes fascinantes.

SHASHA, Dennis E., *Doctor Ecco's Cyberpuzzles,* Norton Press, 2004. Se ha argumentado con frecuencia que resolviendo acertijos se aprende mucho. La serie de libros con acertijos de Shasha describe muchísimos de diversos niveles, así como temas relacionados con computación y matemáticas. El que se cita es un buen ejemplo.

COMPUTACIÓN

¿Qué tan importante es el ser humano, como individuo y como especie? ¿Somos el centro del universo, o un grano insignificante en su enormidad? Una de las exploraciones más interesantes de este asunto se encuentra en los escritos de Maimónides, en la serie *Guía de los perplejos*, Fondo de Cultura Económica.

GOULD, Stephen J., *Time's Arrow, Time's Cycle: Myth and Metaphor in the Discovery of Geological Time,* Harvard University Press, 1988. En este texto, que inspiró la introducción, se abordan temas relacionados con el tiempo geológico.

HAREL, David, *Computers Ltd.: What They Really Can't Do (Popular Science)*, Paperback, Dec 22, 2003. Libro en el que se exploran los límites de las computadoras, dirigido al público en general.

ALGORÍTMICA

HAREL, David y Yishai FELDMAN, *Algorithmics, The Spirit of Computing,* 3a. ed., Addison Wesley, 2004. Otro libro de Harel, quien ha ganado premios no sólo por sus contribuciones de investigación sino también por el material de difusión para todo público. En éste, una vez más, transmite con claridad su entusiasmo por la disciplina académica de la computación y, evitando el uso de matemáticas avanzadas, logra presentar los profundos conceptos sobre los cuales se basa esta disciplina.

RIFKIN, Adam, "Teaching Parallel Programming and Software Engineering Concepts to High School Students", Technical Symposium on Computer Science Education Proceedings of the Twenty-fifth SIGCSE Symposium on Computer Science Education, SIGCSE: ACM

Special Interest Group on Computer Science Education, 1994, pp. 26-30. Este artículo propone una serie de ejercicios para estudiantes de nivel preparatoria. Los ejercicios permiten comprender las diferencias entre distintos algoritmos de ordenamiento, y entre los resultados destacados está el hecho de que resulta *natural* para el estudiante comprender algoritmos paralelos, destruyendo así el mito de que la computación paralela es *difícil*.

RINCÓN MEJÍA, Hugo Alberto, *Cuando cuentes cuántos…*, Instituto de Matemáticas, UNAM, 2002, 132 pp. Temas de matemáticas para bachillerato Núm. 1. Este libro te enseña a contar en situaciones en las que no parece trivial hacerlo. Por ejemplo, ¿de cuántas maneras se pueden sentar el mismo número de hombres y de mujeres en una mesa, de tal forma que no se sienten dos mujeres juntas o dos hombres juntos? O bien, ¿cuántas manos de póquer tienen exactamente un as? En los problemas que analizamos aquí, calculamos el tamaño del espacio de posibles soluciones, haciendo operaciones más simples, pero suele ser útil poder calcular, en casos más generales, el número de posibilidades en que algo puede ocurrir para analizar cuán complejo puede ser resolver un problema.

TAHAN, Malba, *El hombre que calculaba,* Limusa. Éste es un libro muy accesible e interesante, en particular en lo que se refiere al cálculo.

PROGRAMACIÓN

ABELSON, Harold y Gerald Jay SUSSMAN, *Structure and Interpretation of Computer Programs*, MIT Press, 2a. ed., 1996 <http://mitpress.mit.edu/sicp/>. Uno de los libros de mayor influencia en el mundo de la computación. Conocido como la Biblia de Scheme o el libro púrpura, ha servido como introducción a la computación y libro de texto para enseñar programación en el MIT y muchas otras universidades del mundo.

DYBVIG, R. Kent, *The Scheme Programming Language*, 2a. ed., Prentice Hall, 1996 <http://www.scheme.com/tspl2d/index.html>. Texto acerca de Scheme, el lenguaje de programación utilizado en este libro.

FELLEISEN, Matthias, Robert Bruce FINDLER, Matthew FLATT y Shriram KRISHNAMURTHI, *How to Design Programs, An Introduction to Computing and Programming*, The MIT Press, Cambridge, 2003 <http://www.htdp.org>. Excelente introducción a la computación y un programa completo para acercar al estudiante a la programación. Utiliza un enfoque integral para enseñar y apoyar el desarrollo de habilidades analíticas mediante el desarrollo de programas como solución a problemas. Viene acompañado de un ambiente de programación amigable que crece junto con el estudiante.

GRAHAM, Paul, *Hackers and Painters: Big Ideas from the Computer Age,* O'Reilly Media, Inc. (2004). Explora los elementos involucrados en la creación de programas de cómputo y la creatividad de los seres humanos.

INFORMACIÓN

BELL, Tim, Ian WITTEN y Mike FELLOWS, *Computer Science Unplugged… off-line activities and games for all ages,* 1988. Comenzando con temas de información, codificación, compresión y teoría de la información, presenta divertidas y a la vez muy serias actividades que un niño de primaria puede llevar a cabo, y hasta aprender acerca de los fundamentos de la computación, con explicaciones que llegan hasta el nivel de bachillerato y más. Incluye algoritmos, modelos de cómputo y lenguajes, problemas computacionalmente difíciles, criptografía e interfases humano-computadora. Es parte del esfuerzo de los autores por introducir la computación como una disciplina desde la educación básica.

DEWDNEY, A. K., *The New Turing Omnibus, 66 Excursions in Computer Science*, Freeman, 2001. Relevante para casi todos los módulos, describe en pocas páginas 66 temas centrales acerca de la teoría de la computación, la tecnología y aplicaciones.

ABSTRACCIÓN

DAVIS, Martin, *The Universal Computer, the Road from Leibniz to Turing,* Norton, 2000. Uno de los científicos más importantes en el establecimiento de los principios de computabilidad es también un excelente escritor para el público lego. La historia de la computación, enfatizando su estrecho lazo con la lógica matemática, es presentada de manera muy amena.

GARDNER, Martin, *Logic Machines, Diagrams and Boolean Algebra*, Dover, 1968. En este libro, uno de los más hábiles y populares escritores de ciencia presenta historias de dispositivos mecánicos de cómputo, algunos muy curiosos, así como bases de lógica.

COMPUTADORAS

NISAN, Noam y Shimon SCHOCKEN, *The Elements of Computing Systems: Building a Modern Computer from First Principles,* MIT Press, 2005. Un libro que reta, incluso a profesionales de la computación, a reevaluar el entendimiento de la tecnología actual. Con un enfoque basado en principios y conocimientos fundamentales, permite comprender cómo funciona un sistema de cómputo moderno y cómo se aplican sus principios.

REDES

BARABASI, Albert-Laszlo, *Linked: How Everything Is Connected to Everything Else and What It Means,* Plume, 2003. Uno de los libros fascinantes publicados recientemente con nuevas teorías que explican la importancia de las redes y los principios comunes, con sistemas que van desde redes neuronales hasta redes sociales, epidemiología, internet y la web.

BERNERS-LEE, Tim, *Weaving the Web*, Texere Publishing, 2001. Un recorrido por las ideas que dieron nacimiento a la web, por parte de su creador, así como una revisión del estado actual de la web y su futuro probable.

GRALLA, Preston, Sarah Ishida, Mina Reimer y Stephen Adams, *How the Internet Works QUE*, 8a. ed., 2006. Una excelente introducción a los protocolos y tecnologías que soportan internet. Es un libro fácil de leer y adecuado para aquellos que están descubriendo la red.

MULTIMEDIA

GONZÁLEZ, R. C. y R. E. WOODS, *Digital Image Processing,* 2a. ed., Prentice Hall, Upper Saddle River, Nueva Jersey, 07458, 2002. Un clásico en lo referente al procesamiento digital de imágenes para el nivel de licenciatura.

KOLÅS, Øyvind, *Image Processing with Gluas*: *Introduction to Pixel Molding,* 2005 <http://pippin.gimp.org/image_processing/index.html>. Una excelente introducción al arte digital.

LARSON, Kevin, "The Technology of Text", IEEE Spectrum, 2005 <http://spectrum.ieee.org/print/5049>. Interesante artículo introductorio a la caligrafía digital.

NEGROPONTE, Nicholas, *Being Digital*, Knopff, 1995. El futuro ha llegado y es digital, es lo que nos explica el autor de manera divertida y clara, discutiendo con humor la manera en que todo nuestro mundo está cambiando. Describe la evolución de los CD-ROM, multimedia, hipermedia, televisión, realidad virtual y muchas otras cosas.

APLICACIONES

BEWERSDORFF, Jörg, "Luck, Logic & White Lies: The Mathematics of Games", A. K. Peters (comp.), 2005. Este texto indaga acerca de los fundamentos matemáticos de los juegos.

KASNER, E. y J. NEWMAN, *Matemáticas e imaginación*, México, Continental, 1955. Interesante libro con datos relevantes.

<http://world.honda.com/ASIMO>. Página del famoso robot capaz de interactuar con seres humanos, creado por la compañía Honda.

APÉNDICE

PAUL STRATHERN **COMPUTACIÓN**

Turing y la computadora

TURING Y LA COMPUTADORA

PAUL STRATHERN

[Publicado en Paul Strathern, *Turing y la computadora*, traducción de Flavia Bello, Madrid, Siglo XXI de España, 1999 (adaptado por Siglo XXI Editores)]

INTRODUCCIÓN

El descubrimiento de la computadora podría ser uno de los grandes logros tecnológicos de la humanidad. Podríamos considerar incluso que está a la altura del uso del fuego, el descubrimiento de la rueda y el dominio de la electricidad. Estos avances sirvieron para domeñar las fuerzas primarias: la computadora domestica la propia inteligencia.

Más del 90% de los científicos que han existido están vivos actualmente, y las computadoras multiplican la rapidez de su trabajo a diario. (La secuenciación del genoma humano se concluirá probablemente *medio siglo* antes de lo que se predijo al descubrir su estructura, y todo gracias a estos aparatos.)

Pero no habría que dejar volar nuestras esperanzas demasiado alto. Se esperaba algo similar del desarrollo del motor de vapor, hace menos de 150 años. Y la regla de cálculo duró menos de un siglo. El avance que podría hacer que la computadora se convirtiera en un objeto inútil nos resulta inconcebible sólo porque todavía no se ha concebido.

Incluso antes de que la primera computadora se hubiera inventado, conocíamos sus límites teóricos. Sabíamos *qué* podría calcular. E incluso, durante el montaje de las primeras, se entendía la *cualidad* potencial de su capacidad: podrían desarrollar su propia inteligencia artificial. El nombre del responsable de ambas ideas es Alan Turing.

Hombre muy peculiar que llegó a verse a sí mismo como algo parecido a una computadora, Alan Turing trabajó también en la calculadora *Colossus*, que descifró los códigos del *Enigma* alemán durante la segunda guerra mundial. Al igual que Arquímedes, Turing tuvo que dejar de lado una brillante carrera para intentar salvar a su país. Arquímedes fracasó, y fue asesinado a golpe de espada por un soldado romano. Turing lo logró, y su agradecido país lo llevó a juicio por su homosexualidad.

Tras su prematura muerte, Turing fue condenado al olvido, pero actualmente cada vez más personas creen que tal vez fuera la figura más importante en la historia de la informática.

LA ERA A.C.: LAS COMPUTADORAS ANTES DE SU TIEMPO

La primera fue, por supuesto, el ábaco. Este método de cálculo fue inventado incluso antes de la rueda (evidentemente, nuestro deseo de no ser engañados es más profundo que el de viajar con comodidad). Algunos restos arqueológicos demuestran que, en torno al 4000 a. C., ya se utilizaba en China y Oriente Próximo una forma de ábaco, que parece haber evolucionado de forma independiente en las dos regiones. Algunos sugieren que esto muestra la primacía de las matemáticas: la necesidad de calcular como una función aparentemente inevitable de la condición humana.

Ábaco deriva de la palabra babilónica *abaq*, que significaba "polvo". Los estudiosos han dado explicaciones particularmente ingeniosas de esta manifiesta incongruencia. Según una de las versiones, todos los cálculos

se realizaban originalmente sobre el polvo, por lo que "polvo" acabó definiendo toda forma de cálculo. Otra opción es que el método de cálculo utilizado por el ábaco se dibujara en un principio con líneas y rayas en el polvo.

En realidad el ábaco no es en absoluto, estrictamente hablando, una computadora. El cálculo real lo hace la persona que utiliza el ábaco, que debe tener en su cabeza el programa (los pasos matemáticos necesarios).

Sea una computadora o no, lo cierto es que el ábaco y su programa humano se utilizaron en toda Europa y Asia para realizar cálculos hasta bien avanzada la Edad Media. Después se introdujo el cero en las matemáticas, lo que supuso un obstáculo para los trabajos en los que se utilizaba el ábaco. Como consecuencia de esto, los matemáticos serios desecharon rápidamente esta aportación infantil. Sin embargo, durante varios siglos el ábaco siguió utilizándose como calculadora, caja registradora, computadora y para fines similares (y no tan similares). Es más: hasta hoy el ábaco sigue desempeñando un papel fundamental en las economías locales de algunas zonas de Asia central y Rusia.

La primera calculadora conocida sigue siendo un auténtico misterio. En 1900, submarinistas cazadores de esponjas en Grecia descubrieron cerca de la diminuta isla de Antikythera restos de un antiguo naufragio del primer siglo antes de Cristo. Entre las estatuas y vasijas rotas se encontraron algunas piezas de bronce corroído, que parecían ser parte de una máquina. Cincuenta años tardaron los estudiosos en descubrir cómo encajaban estas piezas y lograr que el aparato funcionara. El resultado fue una especie de calculadora astronómica, que funcionaba igual que una computadora analógica moderna, con piezas mecánicas para hacer los cálculos. Al girar una manivela se accionaban unas palancas; éstas, a su vez, accionaban unos cuadrantes con los que se podía leer la posición del Sol y los planetas del zodiaco.

Lo que hace a este descubrimiento tan asombroso es su singularidad; nunca se ha encontrado nada de ese periodo ni remotamente semejante. En la literatura griega clásica no se menciona una máquina como ésta, ni similar. Ningún filósofo, poeta, matemático, científico o astrónomo hace referencia a un objeto así. Además, según los conocimientos actuales sobre la ciencia de la antigua Grecia, no había tradición ni conocimiento capaz de producir tal máquina. Aparentemente, la primera computadora fue una construcción estrafalaria, tal vez un juguete, de algún desconocido genio mecánico, que simplemente desapareció de la historia. Al tratarse de un objeto estrafalario sin influencias, desapareció como un cometa. Después, durante más de mil quinientos años, nada.

En general, se considera la primera calculadora "real" la que fabricó en 1623 William Schickard, catedrático de hebreo en la Universidad de Tübingen. Schickard era amigo del astrónomo Johannes Kepler, que descubrió las leyes de los movimientos planetarios. Kepler despertó el interés latente por las matemáticas del catedrático de hebreo, cuya habilidad para realizar cálculos se había apolillado un poco con el paso de los años. Así que decidió fabricar una máquina que lo ayudara con sus sumas. La máquina de Schickard se ha descrito como un "reloj de cálculo". Se pretendía que sirviera de ayuda a los astrónomos, al permitirles calcular las efemérides (las futuras posiciones del Sol, la Luna y los planetas).

Desgraciadamente, nunca sabremos si esta máquina funcionaba, o cómo se pretendía que funcionara exactamente. El primer y único prototipo aún no se había terminado cuando tanto éste como los proyectos de Schickard fueron destruidos por el fuego, durante la guerra de los Treinta Años. Schickard quedó así reducido a un mero pie de página histórico, en lugar de convertirse en el inventor del mayor avance tecnológico desde la invención del arnés.

Lo que sí sabemos es que la máquina de Schickard fue una precursora de la computadora digital, en la que los datos se introducen en forma de números. En el otro tipo de computadora, la analógica, los números de entrada y salida se sustituyen por una cantidad susceptible de ser medida, como la tensión, el peso o la longitud. Esta última se utilizó en la primera computadora analógica: la regla de cálculo, inventada en la década de 1630. La regla de cálculo más simple consta de dos reglas, ambas marcadas con escalas logarítmicas. Al deslizar las dos reglas, de forma que quede un número frente a otro, se pueden multiplicar y dividir con facilidad.

La regla de cálculo fue inventada por William Oughtred, cuyo padre había trabajado como escribiente en Eton y enseñaba a escribir a los alumnos analfabetas. Su hijo recibió las órdenes sagradas como sacerdote, pero siguió los pasos de su padre al dar algunas clases particulares aparte. En la década de 1630 creó la primera regla de cálculo rectilínea (es decir, con dos reglas rectas). Pocos años más tarde, se le ocurrió la idea de la regla de cálculo circular (que tiene un círculo móvil dentro de un anillo, en lugar de reglas deslizantes). Desgraciadamente, uno de sus alumnos se apropió la idea y la publicó primero, afirmando que el des-

cubrimiento había sido suyo. Aunque el gesto no gustó a Oughtred, se puede decir que acabó sus días feliz. Devoto monárquico, se dice que falleció en "estado de éxtasis" al oír que Carlos II había recuperado su trono.

La regla de cálculo primitiva fue evolucionando con el tiempo, hasta convertirse en un dispositivo capaz de realizar cálculos complejos. Entre los que contribuyeron a su desarrollo se encuentra James Watt, que la utilizó para calcular el diseño de sus máquinas de vapor originales, en la década de 1780. Amadeé Mannheim, un oficial de artillería francés, fue el artífice de un nuevo avance. Diseñó una forma más sofisticada de regla de cálculo, cuando aún era estudiante, lo que le permitió obtener unos resultados sobresalientes en los exámenes, que a su vez lo lanzaron a una brillante carrera dentro de la educación militar. Fue precisamente la versión de la regla de cálculo de Mannheim la que alcanzó gran popularidad durante la primera mitad del siglo xx: era el accesorio característico *de rigeur* en el bolsillo de la pechera de cualquier científico de bata blanca.

Pero volvamos a la computadora digital. El siguiente avance en este campo vino de la mano del matemático francés del siglo XVII, Blaise Pascal, que casualmente nació en 1623, el mismo año en que Schickard había inventado el "reloj de cálculo" original. El padre de Blaise Pascal era un recaudador de impuestos reales, que tenía ya bastantes dificultades para recaudar dinero como para, además, poder presentar las cuentas que necesitaba el tesorero real. Para ayudarlo, su joven y precoz hijo intentó diseñar una máquina de contabilidad. Con 19 años ya había construido un modelo que funcionaba. Los números se introducían en la máquina mediante discos graduados con números, y conectados a ejes con ruedas dentadas y engranajes. La máquina de Pascal podía sumar y restar cifras de hasta ocho dígitos. Esta máquina era extremadamente complicada, y llevaba las técnicas mecánicas del momento a sus límites, e incluso los superaba. La máquina, sin embargo, tenía muchos problemas con los engranajes. Pero Pascal era un perfeccionista, y afirmaba haber hecho "más de 50 modelos, y todos diferentes". Pascal no era sólo un gran matemático, sino que también fue el mejor filósofo religioso de su tiempo. Atormentado por una salud muy frágil, su celo religioso se incrementaba de forma inversamente proporcional a su salud. Pero siguió siendo un matemático hasta el fin de sus días, y llegó a reducir la fe a una probabilidad matemática. En su opinión, aunque se podían calcular las posibilidades de la inexistencia de Dios, era mejor apostar por su existencia, ya que no había nada que perder en caso de que no fuera cierto.

Siete de las máquinas de Pascal han llegado a nuestros días: son obras maestras del ingenio que incorporan varios principios aún utilizados en las computadoras mecánicas. Algunas de las máquinas de Pascal que han perdurado aún funcionan, aunque nadie ha descubierto todavía cómo utilizarlas para calcular las posibilidades de la existencia de Dios.

El siguiente avance significativo para la computadora digital lo logró el filósofo alemán Gottfried Leibniz, el Leonardo da Vinci de su época. Entre otras muchas cosas, Leibniz creó nada menos que dos filosofías (una optimista y otra pesimista), un plan detallado para la invasión de Egipto, una historia en 15 volúmenes de la Casa de Hannover y una calculadora muy superior a la de Pascal.

El interés de Leibniz por las calculadoras no era únicamente práctico. Cuando aún estaba en la universidad, escribió un artículo en el que explicaba la base teórica de una calculadora y sus posibilidades (un trabajo que señalaba ya el camino para las ideas básicas que Turing habría de tener sobre este tema, casi 300 años más tarde). En torno a la misma época, inventó también una matemática binaria, similar a la que sería el lenguaje de las computadoras digitales, aunque no llegó a combinar ambos elementos.

Leibniz creó su calculadora en 1673, después de haber visto una de las máquinas de Pascal en París. Desgraciadamente, Leibniz estaba arruinado en ese momento, y sus esfuerzos se vieron paralizados por la necesidad de hacer que su máquina fuera viable desde un punto de vista comercial (la máquina de Pascal era demasiado compleja para que pudiera fabricarla nadie más que él). En cuanto Leibniz hubo terminado su máquina, cruzó el Canal de la Mancha para mostrarla a la Royal Society de Londres. Sus miembros no parecieron impresionados, y abandonó el proyecto cuando aún no tenía más que un prototipo.

A pesar de estas limitaciones, la máquina de Leibniz era extraordinaria. Al igual que la de Pascal, se accionaba mediante una sucesión de ruedas dentadas, pero era capaz de hacer muchas más cosas que la de Pascal. Desde el primer momento podía multiplicar (mediante sumas repetidas), pero además Leibniz pronto añadió unos dispositivos que permitían efectuar divisiones y calcular también raíces cuadradas.

Leibniz veía un gran futuro para las calculadoras, aunque no volvió a encontrar tiempo para hacer nuevos intentos prácticos en este campo. Esto, sin embargo, no impidió que su mente, siempre activa, pensara en las calculadoras y el papel que podrían desempeñar en el futuro. Para él, algún día las calculadoras resolverían todas las disputas éticas. Bastaría con insertar los diferentes argumentos y la máquina "calcularía" cuál era superior (aunque las bases precisas para semejante cálculo quedaron en la misma categoría que el cálculo de las posibilidades de la inexistencia de Dios: es decir, siguen siendo un misterio para todos, salvo para el genio que las concibió).

Del mismo modo, Leibniz predijo también que las calculadoras quitarían trabajo a los jueces: los tribunales del futuro estarían presididos por calculadoras que emitirían tanto el fallo como la sentencia adecuada. Una presciencia tan sorprendente podría hacernos pensar en una historia de terror informático, pero Leibniz lo veía de forma muy diferente. Esencialmente, era un hombre optimista, y pensaba que "todo es para bien en el mejor de los mundos posibles". No cabe imaginar cómo habría sido el mundo si Leibniz hubiera dedicado algo más de sus excepcionales energías a la producción de calculadoras.

El siguiente avance importante en este campo se debió a un hombre que era totalmente ajeno a él: Joseph Marie Jacquard, un técnico francés dedicado al negocio de los telares. A principios del siglo XIX, creó un telar innovador, en que el patrón del tejido estaba controlado por tarjetas perforadas. Así surgió la idea de programar una máquina, aunque Jacquard no tenía ni idea de lo importante que era su invento. Afinó más su idea; sin embargo, funcionó demasiado bien: sus máquinas originaron protestas en Lyon, en la década de 1820, cuando los trabajadores de los telares que habían sido despedidos tomaron por asalto las fábricas y destruyeron muchas de las máquinas. Aún hoy se emplea el método de Jacquard para el tejido de patrones complejos.

Calculadoras de mecánica compleja, la idea de la programación, una teoría de los números racionales: los elementos básicos de la computadora moderna estaban empezando a aparecer. Pero hizo falta un genio para descubrir cómo combinar todos estos elementos dispares. En general, se reconoce a Charles Babbage como el padre de las computadoras. Al igual que muchos genios prácticos, era increíblemente poco práctico en cualquier sentido de la palabra, pero sus descubrimientos y logros estaban un siglo por delante de su tiempo.

Babbage nació en 1791, y heredó una considerable fortuna personal. Era un joven de buen carácter, que pronto demostró ser una promesa de excepción en el campo de las matemáticas. Logró que se introdujeran las notaciones matemáticas de Leibniz en Gran Bretaña (los matemáticos británicos habían insistido, patrióticamente, en utilizar la notación original —pero inferior— de Newton, con ello se aislaron a sí mismos de un siglo de avances del resto de Europa).

Después, Babbage desvió su atención hacia otro de los problemas que atenazaban a los científicos británicos: los errores que aparecían por doquier en las impresiones de las tablas astronómicas y matemáticas. Por ejemplo, se descubrió que la primera edición de *Nautical Ephemeris for Finding Latitude and Longitude at Sea (Efemérides náuticas para hallar latitudes y longitudes en el mar)* tenía más de… ¡mil errores!

Babbage decidió que había una única solución para el problema de las tablas erróneas. Había que construir una calculadora grande, infalible y multiusos. Después de solicitar y lograr ayuda gubernamental, Babbage emprendió la construcción de su aclamada "máquina diferencial núm. 1". Este proyecto era enormemente ambicioso: la máquina de Babbage tenía que ser capaz de calcular cifras de hasta 20 dígitos; también tenía que almacenar una serie de números y efectuar sumas con éstos. Los cálculos de la máquina podían reducirse a sumas porque utilizaría el método de las múltiples diferencias. Este método se basa en los polinomios (fórmulas algebraicas formadas por varios términos) y en el hecho de que éstos mantienen una diferencia constante. En su forma más simple sería como sigue:

Donde f(x) = 2x + 1			
si x = 1	2	3	4…
f (x) = 3	5	7	9…
diferencias = 2	2	2	2…

Huelga decir que la operación no resulta tan sencilla con funciones más complejas. Pero es posible hallar una diferencia constante en las diferencias entre las diferencias (o las diferencias entre las diferencias entre las diferencias). En la mayoría de los casos, si un polinomio tiene un término X^n, hay que calcular *n* diferencias para encontrar una diferencia constante. Para calcular el polinomio de una sucesión de valores de X, como cuando se calculan tablas, a una máquina le resulta más sencillo sumar la diferencia constante y volver atrás, sumando diferencias, en lugar de iniciar una serie de complejas

multiplicaciones. Además, las funciones como los logaritmos y funciones trigonométricas, que no operan de la misma forma, se pueden reducir a polinomios muy aproximados.

Como sus antecesores, la "máquina diferencial núm. 1" utilizaba ruedas dentadas y funcionaba con el sistema decimal, pero su fabricación superaba con mucho la complejidad de las máquinas anteriores, lo que hizo necesarios diversos avances dentro de la ingeniería mecánica.

Sin embargo, Babbage era muy capaz de realizar esta tarea gracias a su magistral capacidad de improvisación. A medida que su máquina crecía, se le ocurrían ideas brillantes para añadirle nuevas características que iba incorporando sobre la marcha. La "máquina diferencial núm. 1" se empezó a fabricar en 1823, pero nunca llegó a terminarse. Después de diez años de trabajo, Babbage había convertido su proyecto original en una máquina de 25 000 piezas (y sólo se habían fabricado 12 000), y el costo se elevaba a 17 470 libras (en aquellos días, esa suma era suficiente para fabricar un par de buques de guerra). Babbage había puesto grandes sumas de su propio bolsillo, pero el gobierno decidió frenar el proyecto. Era mejor invertir en una flota que en una máquina que podría acabar contribuyendo a la deuda nacional en una cifra que sólo esa misma máquina podría calcular. A pesar de todas esas dificultades, en 1827 Babbage había utilizado la única parte operativa de su máquina (formada por apenas 2 000 piezas) para calcular tablas de logaritmos de uno a 108 000. Esta parte de la "máquina diferencial núm. 1" se considera la primera calculadora automática. Había que introducir las cifras y los resultados salían en forma impresa (con lo que se reducía el margen de error humano).

Sin embargo, esto era sólo el principio para Babbage. Para la década de 1830 ya tenía el proyecto de una "máquina diferencial núm. 2". Este concepto supuso un avance significativo en las técnicas de cálculo. La "máquina diferencial núm. 2" se convirtió en la primera máquina analítica: una máquina cuyo funcionamiento estaba controlado por un programa externo. Babbage había oído hablar de la idea de Jacquard sobre las tarjetas perforadas para controlar el mecanismo de una máquina y decidió incorporarla a su propia máquina. Esto le permitiría realizar cualquier cálculo aritmético en función de unas instrucciones insertadas en tarjetas perforadas. Al igual que la primera máquina diferencial, también necesitaba una memoria para almacenar números, pero la nueva máquina debería poder realizar una secuencia de operaciones con esos números almacenados. Babbage había dado con las características esenciales de la computadora moderna.

El núcleo mecánico al que se asociarían estas características era el *plato fuerte*. Iba a contener mil ejes, con al menos 50 000 engranajes y se pretendía que calculara números de 50 dígitos en el sistema decimal.

Desgraciadamente, el gobierno no se dejó intimidar por estas increíbles posibilidades, y prefirió no hacer nuevos intentos para arruinar al erario. A estas alturas, la fatiga producida tras largos años de trabajo duro sin resultados había hecho estragos en el carácter de Babbage. El atractivo joven de Cambridge se había convertido en un vejestorio irascible que merodeaba por las calles de Londres. Le cogió manía al ruido que hacían los músicos callejeros que "no sin frecuencia hacen que los pilluelos andrajosos se pongan a bailar, y que hasta hombres casi ebrios sigan los bailes, y acompañen el ruido con sus propias voces discordes… Otro grupo que apoya con vehemencia la música callejera es el de las señoras de elástica virtud y tendencias cosmopolitas, que encuentran en ella una excusa decente para exponer sus encantos desde las ventanas abiertas". Babbage inició una campaña para que prohibieran a los músicos callejeros, afirmando que le impedían trabajar en paz. Los músicos callejeros se vengaron reuniéndose justo debajo de su ventana. Babbage dejó escrito que "en una ocasión, una banda de música estuvo tocando durante cinco horas, sin apenas pausa".

Para entonces, Babbage había invertido buena parte de su fortuna en la persecución de su sueño de las máquinas diferenciales. Durante varios años, Lady Ada Lovelace, hija del poeta Byron y una de las mujeres matemáticas más brillantes de su tiempo, le ayudó en sus esfuerzos. (El papel de esta mujer en la historia de las computadoras fue reconocido con honores cuando el Departamento de Defensa de Estados Unidos le puso su nombre, ADA, a su lenguaje de programación.) Lady Lovelace también ayudó a Babbage en un intento optimista de recuperar su fortuna. Juntos dedicaron mucho tiempo y energía a intentar crear un sistema de apuestas infalible para las carreras de caballos. Sin embargo, las demostraciones prácticas de este sistema resultaron casi tan costosas como la máquina diferencial.

A pesar de tales reveses, Babbage tuvo tiempo de inventar el quitapiedras de la locomotora y de descu-

brir que los anillos de los árboles podían leerse como registros meteorológicos. Después de su muerte, en 1871, los diseños operativos para su "máquina diferencial núm. 2" permanecieron en el olvido durante muchos años. Más tarde se construyó el núcleo de la primera máquina analítica del mundo, según unos planos modificados de la '"máquina diferencial núm. 2". Esta enorme construcción de tres toneladas puede verse hoy en día, en todo su esplendor, en el Museo de Ciencias de Londres. Y funciona (en las pruebas se configuró para que calculara 25 múltiplos del número π, de 29 dígitos decimales, tarea que sus 50 000 ruedas dentadas digirieron con insultante facilidad).

Babbage había definido las características básicas de la computadora moderna, pero sus máquinas presentaban una desventaja fundamental: funcionaban sólo dentro de las matemáticas decimales. Este problema se solucionó gracias al trabajo de George Boole, uno de sus contemporáneos. Boole nació en 1813, hijo de un zapatero de Lincoln. Aunque casi totalmente autodidacta, demostró tal agudeza intelectual que fue nombrado catedrático de matemáticas en el Queen's College en Cork, donde acabó casándose con Mary Everest, sobrina del hombre que dio nombre a la montaña.

En 1854, Boole publicó su *Investigation of the Laws of Thought* (*Investigación de las leyes del pensamiento*), que introdujo lo que actualmente se conoce como álgebra booleana. En esta obra, Boole sugería que la lógica pertenece al ámbito de las matemáticas, más que al de la filosofía. Al igual que la geometría, se basa en una serie de axiomas sencillos. Además, al igual que la aritmética tiene unas funciones primarias, como la suma, la multiplicación y la división, la lógica puede reducirse a operadores como "y", "o" o "ni". Estos operadores pueden utilizarse en un sistema binario (el sistema digital tiene diez dígitos; el sistema binario funciona igual, pero sólo con dos). El "verdadero" y "falso" de la lógica se reducen al 0 y al 1 de la matemática binaria. Así, el álgebra binaria reduce cualquier proposición lógica, independientemente del número de términos que contenga, a una simple secuencia de símbolos binarios. Eso cabría en una simple tira de papel, en la que el álgebra binaria se reduce a una secuencia de orificios (y ausencia de orificios). De este modo, se podría introducir todo un "argumento" lógico o programa en una máquina.

Con dígitos binarios, las máquinas podrían seguir instrucciones lógicas y su matemática se adaptaba perfectamente al circuito eléctrico de encendido/apagado. Así, el dígito binario (o *bit*) llegó a ser la unidad fundamental de información de los sistemas informáticos.

Sin embargo, los avances individuales de Babbage y Boole siguen sin recibir reconocimiento. En lo que al mundo respecta, el siguiente paso importante lo dio Herman Hollerith, un estadístico estadunidense. Hollerith desarrolló una "máquina de censo", que podía leer tarjetas de hasta 288 orificios y almacenar la información. Su máquina electromecánica podía leer hasta 80 tarjetas por minuto. Cuando se utilizó para el censo de estadunidenses de 1890, la máquina de Hollerith procesó todos los datos en seis semanas (la elaboración del anterior censo, de 1880, había llevado tres años). En 1896 Hollerith se metió en el mundo de los negocios, y creó la Tabulating Machine Company, que más adelante se convertiría en la International Business Machine Corporation (IBM).

Se habían descubierto los elementos necesarios para la computadora moderna (incluyendo la explotación comercial). Lo único que faltaba era que alguien descubriera qué podía hacer: las posibilidades y limitaciones teóricas. Esto fue lo que hizo Alan Turing.

VIDA Y OBRA DE UN ENIGMA

AlanTuring nació en Londres en 1912, en el seno de una familia inglesa de clase media alta. Su padre pertenecía a la administración en India y su madre era hija del ingeniero jefe del ferrocarril de Madras. En 1913 sus padres volvieron a India, dejando al recién nacido Alan y a su hermano de cinco años al cuidado de un coronel retirado y su esposa, en la ciudad de St. Leonards-on-Sea, en el condado de Sussex.

En aquellos días, los padres ingleses respetables no pensaban que fuera malo abandonar a los hijos de esta forma. Incluso los que no podían marcharse a las colonias contrataban a niñeras y mandaban a sus hijos a internados (desde los siete a los dieciocho años), para asegurarse de ver poco, y oír menos, a sus hijos. Este temprano abandono apenas afectó a John, hermano de Alan, como apenas afectó a la mayor parte de esa generación de clase media; todos acabaron convirtiéndose en típicos colegiales ingleses de colegios privados. (Sólo ahora, en que los tiempos se han vuelto mucho más exigentes, empezamos a pensar que esta especie característica de jóvenes estaba mutilada emocionalmente.) Alan Turing, sin embargo, resultó ser un niño normal, por lo que la experiencia lo afectó sobremanera. Desa-

rrolló un pronunciado tartamudeo, su autosuficiencia era rayana en la excentricidad y se sentía incapaz de participar en la charada de las costumbres sociales.

Cuando su madre volvió, para una larga visita, en 1916, Alan reaccionó con sentimientos encontrados, que habría de conservar toda su vida. Quería tiernamente a su madre, pero también pensaba que era una mujer imposible. Por su parte, parece que la señora Turing era, efectivamente, imposible, pero además difícil de querer. Su principal preocupación vital era que Alan pareciera respetable. Desgraciadamente, el "diablillo" que Alan llevaba dentro hizo que eso también fuera imposible.

En el internado, desde el principio Alan se comportó de forma especial. Era siempre el único que iba desaliñado, el que tenía tinta en los dedos, el que no encajaba en el grupo. Lo que es peor, parecía no querer encajar. Era tímido, solitario, y su tartamudeo sólo servía para empeorar las cosas. El menor de los Turing no parecía una promesa. Tuvo problemas para aprender a leer y a escribir. Sin embargo, un día decidió que quería leer y aprendió por su cuenta en tres semanas. Para cuando cumplió los once años había desarrollado una pasión por la química orgánica, pero seguía sin sentir ningún interés por otras materias y ni siquiera era capaz de hacer divisiones largas.

Cuando el matrimonio Turing regresó de la India definitivamente, decidieron quedarse en Dinard, en Bretaña, para evitar el pago de impuestos. Sin embargo, no puede decirse que se tratara estrictamente de un acto egoísta por su parte: sencillamente, no se podía considerar a alguien como miembro de la clase media si no enviaba a sus hijos a un colegio privado, que era más de lo que los Turing podían permitirse económicamente. Había que evitar este estigma social como fuera, aunque supusiera el exilio. Así que, cuando los padres volvieron por fin del extranjero, sacaron a los niños de donde estaban y los llevaron a vivir con ellos a una casa diferente. Todo por su bien, por supuesto.

De hecho, sí fue por su bien. Alan disfrutó de estas vacaciones en Francia y aprendió rápidamente el idioma. Además, su educación privilegiada acabó despertando las adormecidas cualidades que había heredado. Su abuelo paterno había sido un estudioso de las matemáticas en Cambridge, y un antecesor angloirlandés de su madre había sido el inventor del término "electrón", en 1891 (aunque lo que verdaderamente impresionaba a la familia, y compensaba el *faux pas* de ser un científico, era que había sido elegido miembro de la Royal Society).

Era estupendo que Alan jugara a hacer "tufos" con su juego de química del colegio, pero no era más que una afición. Para recibir una buena educación y salir adelante en la vida, debía aprender latín. Después de todo, para *eso* lo estaban educando. Para eso se estaba invirtiendo todo ese dinero: para que pudiera aprender algo y no se dedicara sólo a juguetear con ese espantoso juego de química. Sin el latín, nunca aprobaría el examen general para entrar en un colegio privado (en la Inglaterra de entonces, a los colegios privados se los denominaba "públicos", pero ni el examen de acceso tenía nada de "general", ni los colegios "públicos" nada de público).

Para alivio de todos, Alan consiguió entrar de panzazo en Sherborne, un colegio privado de razonable prestigio en Dorset. A los 13 años abandonó Francia, sólo para empezar su primer trimestre en el nuevo colegio. Cuando el transbordador llegó a Southampton, se encontró con que la huelga general había empezado ese mismo día. El país estaba paralizado: no había trenes, ni transportes de ningún tipo. Con el ingenio que lo caracterizaba, compró un mapa y recorrió en bicicleta los casi 90 kilómetros que lo separaban de Sherborne, a donde llegó convertido prácticamente en héroe.

Sin embargo, Turing no logró estar a la altura de lo que prometía. No tardó en ponerse de manifiesto que no iba a ser el tipo de héroe de colegio privado. Ese niño desaliñado, raro y tartamudo no parecía interesado en hacer amigos ni en ganar popularidad. Sin embargo, tampoco se rebelaba contra la sociedad. Sencillamente, adoptó un enfoque asocial: prefería hacer las cosas a su manera, en lo posible, pero siempre *dentro* del sistema. Esta actitud habría de mantenerla toda su vida. Se conformaba, pero al mismo tiempo no se conformaba.

En consecuencia, con la ética deportiva de los colegios privados, Turing se decantó por la carrera de fondo, uno de los deportes en que no había que jugar en equipo. Pese a sus pies planos, era sorprendentemente bueno en esta especialidad. Físicamente, al igual que intelectualmente, Turing siempre mostró una resistencia excepcional… siempre y cuando le apeteciera. Además, en Sherborne descubrió un profundo interés por las matemáticas. Estaba poco interesado en las clases convencionales, y empezó a estudiar el tema por su cuenta, aventajando con mu-

cho a sus compañeros, aunque al mismo tiempo seguía sin conocer ni los principios más básicos. Pronto empezó a leer acerca de la relatividad, y desarrolló un profundo interés por la criptografía. Solía hacer orificios en una hoja de papel; al colocarla sobre una página de un libro concreto se obtenía un mensaje. Sin embargo, para crear mensajes cifrados era necesario un receptor. Cuando Turing cumplió 15 años, encontró a ese alguien, un chico mayor llamado Christopher Morcom, a quien todos consideraban el mejor matemático del colegio. Su interés común por las matemáticas fue la base de su amistad. Sin embargo, para Turing esto llegó a ser algo más que amistad: se enamoró de Christopher.

Christopher era un joven de pelo claro y ojos azules, de constitución delgada. Turing encontraba en Christopher la misma inclinación a seguir unos principios (éstos, en el caso de Turing, no eran evidentes, ya que estaban eminentemente relacionados con la protección de su individualismo interior, pero en realidad eran unos principios firmes, lo suficientemente sólidos para que mantuviera las diferencias con los que lo rodeaban). Christopher era un joven de principios, sin caer en la pacatería. No se sumaba a las "conversaciones picantes" de los demás, que el propio Turing encontraba de tan mal gusto.

De vez en cuando, Christopher se ausentaba del colegio y regresaba con aspecto débil y más delgado. Turing sabía que Christopher tenía mala salud, pero desconocía la gravedad de la situación: Christopher padecía en realidad de tuberculosis bovina. A principios de 1931, cayó inesperadamente enfermo en el colegio y lo llevaron a un hospital de Londres, donde falleció a los pocos días.

Turing quedó desolado. Christopher había sido la primera persona que había logrado atravesar su coraza de soledad. Jamás olvidaría su amor adolescente por Christopher. La idea del amor secreto y casto —frecuentemente necesario en los días en que la homosexualidad era un delito— habría de ayudarlo en muchos momentos difíciles de su vida.

Turing se ganó merecidamente una beca para el King's College, en Cambridge, donde ingresó en octubre de 1931. Al principio no hablaba con mucha gente y se mantuvo aislado, disfrutando de la novedad de un cuarto individual y privado, donde podía estudiar tranquilamente. Sin embargo, su tartamudeo se hizo más evidente, y su confusión psicológica no desaparecía. Turing sufrió solo y en silencio al comprender su sexualidad, mientras otros reconocían con burla los rasgos distintivos de los homosexuales. Afortunadamente, no tardó en descubrir que uno de sus compañeros de matemáticas compartía sus inclinaciones y ambos iniciaron una relación sexual sin grandes compromisos.

A principios del decenio de 1930, Cambridge era una de las primeras instituciones matemáticas y científicas del mundo. El físico teórico anglosuizo Paul Dirac y sus colegas consideraban que la universidad sólo era superada en el campo de la física cuántica por la de Göttingen. El King's College estaba especialmente bien dotado: George Hardy, uno de los mejores matemáticos de su tiempo, y Arthur Eddington, cuyo trabajo había ratificado la teoría de la relatividad de Einstein, eran tutores residentes y dieron clases a Turing.

Sin embargo, Turing estaba particularmente interesado en la lógica matemática. En 1913, Bertrand Russell y Alfred North Whitehead, ambos tutores de Cambridge, habían publicado *Principia Mathematica.* Con ello intentaban aportar una base filosófica para las matemáticas, con el fin de establecer así su certeza. Russell y Whitehead intentaron demostrar que todo el edificio de las matemáticas podía derivarse de ciertos axiomas lógicos fundamentales (en cierto modo, esto es lo contrario de lo que Boole intentara cerca de medio siglo antes). Russell y Whitehead no lograron un éxito total, ya que en el camino encontraron algunos problemas lógicos.

Por ejemplo, tomemos una proposición como: "Lo que estoy diciendo es falso". Si la proposición es verdadera, lo que dice es falso; si es falso, lo que dice es cierto. En la jerga de los lógicos, esta proposición era formalmente indeterminable. Las matemáticas no podrían basarse en axiomas lógicos hasta que las paradojas de ese tipo se resolvieran. Sin embargo, muchos pensaban que tales dificultades eran superficiales. No llegaban al corazón del proyecto: no invalidaban el intento de basar las matemáticas en un apoyo lógico bien fundado.

En 1931 todo esto cambió, cuando el *enfant terrible* de la lógica, el austriaco Kurt Gödel, publicó su trabajo sobre las proposiciones formalmente indeterminables de *Principia Mathematica.* En ese trabajo aportó pruebas visibles de lo que, para horror de sus colegas, podría ser "el fin de las matemáticas".

Gödel tomó la proposición "esta afirmación no puede ser probada" y demostró que no podía probarse que fuera cierta (porque si lo fuera habría una contradicción), pero tampoco que fuera falsa, por la misma

razón. Gödel logró demostrar que, dentro de cualquier sistema matemático estrictamente lógico, siempre habría proposiciones cuya veracidad o falsedad no podría ser demostrada partiendo de los axiomas en los que se basara ese sistema. ¡Las matemáticas eran incompletas! Lo que es peor, parecían irremediablemente dañadas, ya que esto significaba que no podríamos estar seguros de que los axiomas básicos de la aritmética no fueran a dar resultados contradictorios. ¡Las matemáticas eran ilógicas! (y, ¡oh horror de los horrores!, ¡también la lógica era ilógica!).

Estos avances causaron un profundo efecto en Turing. Demasiado, tal vez. Porque, como de costumbre, había llegado muy lejos, pero olvidando la base. Así, sólo logró un aprobado alto en la primera parte de los *tripos* (los exámenes de Cambridge). Afortunadamente, sin embargo, en los exámenes finales estuvo a la altura y logró un sobresaliente, lo que le permitió quedarse en Cambridge y dedicarse a la investigación. Tanto Eddington como Hardy estaban convencidos de que poseía habilidades excepcionales.

Para entonces, Turing estaba ganando confianza en sí mismo. Seguía siendo un alma solitaria y un tanto extraña, pero ya no tenía motivos para ocultar su homosexualidad. En las conversaciones con sus colegas, de vez en cuando dejaba caer algún comentario casual sobre sus preferencias sexuales. Esto se integra dentro de su naturaleza: por sus principios, insistía en la franqueza, ante sí mismo y ante los demás. Había, sin embargo, excepciones. No le habló a su madre de su homosexualidad ni de su recién descubierto ateísmo.

Turing sorteaba este engaño con su particular forma de actuar. Cuando iba a casa, en Navidades, cantaba canciones propias de Pascua, y en Semana Santa cantaba villancicos (al parecer, el principio de incertidumbre de Gödel también tenía aplicaciones prácticas). Secamente, como observa el principal biógrafo de Turing, Andrew Hodges, la familia seguía siendo "el último bastión del engaño".

Mientras tanto, la madre de Alan seguía tratándolo como al patito feo de la familia, e insistía en que adoptara una apariencia más elegante, le ordenaba que se cortara el pelo, etc., en cuanto llegaba a la casa de Guilford (la familia había dejado el exilio forzado por el pago de impuestos y había vuelto a la respetabilidad de los condados locales, ahora que había dejado de pagar la educación de sus hijos).

Aunque Turing fuera elegido becario del King's College y se convirtiera en una de las mentes matemáticas más prometedoras de Gran Bretaña, su madre se seguía avergonzando de aquel niño sin remedio, siempre con la cabeza en las nubes. El aspecto aniñado de Turing contribuía a ello. Tanto física como mentalmente, mantuvo durante toda su vida un comportamiento curiosamente juvenil.

La relación que mantenía con su madre siguió siendo muy estrecha. Cuando escribía a casa, trataba incluso de mantenerla al tanto de su pensamiento matemático, y hasta mencionaba la teoría cuántica o la de la relatividad. Aunque habría que ver cuánto de todo esto entendía la señora Turing. Era una mujer inteligente, que venía de un medio culto, pero su religiosidad y su firme creencia en que había que mantener las apariencias eran lo más importante para ella. Seguía viendo a Alan como un hijo díscolo. Lo más normal, claro, era que eligiera destacar en algo tan poco *chic* como las matemáticas.

Pero, efectivamente, destacó. Su tesis de licenciatura le valió una beca, y ahora estaba dedicado a absorber los últimos avances de las mejores mentes científicas y matemáticas del momento. Después de que Hitler accediera al poder en Alemania, en 1933, muchos de los exiliados alemanes pasaron por Cambridge, y a menudo dieron conferencias. Así, Turing tuvo la oportunidad de oír a Erwin R. J. A. Schrödinger hablar de la mecánica cuántica, materia que prácticamente había inventado. También asistió a un curso completo de mecánica cuántica que impartía Max Born, recién salido de Göttingen. Otro de los exiliados de Göttingen, Richard Courant, dio un curso sobre ecuaciones diferenciales.

Tanto Born como Courant habían trabajado con David Hilbert, catedrático de la universidad de Göttingen, comúnmente considerado uno de los mejores matemáticos de la historia. Al igual que Russell y Whitehead, había intentado darle un punto de apoyo formal a las matemáticas, basándolas en unos pocos axiomas básicos. De éstos, mediante una serie de reglas bien definidas, *surgirían* todas las posibilidades matemáticas.

El "programa Hilbert", como se lo conocía, también se detuvo en seco al topar con la denominada "catástrofe Gödel", que demostraba que las matemáticas eran incoherentes desde el punto de vista de la lógica. Sin embargo, a pesar de este claro intento, la teoría de Gödel no logró acabar con las matemáticas. La gente siguió utilizándolas a pesar de él, especialmente los matemáticos. Al parecer, un triángulo seguía siendo un triángulo, los puentes no se caían, y los presupuestos nacionales aumentaban (o no aumentaban, pero

esto no era culpa de las matemáticas). De hecho, muchos entendían la demostración de Gödel como una mera interferencia sin importancia. Lo que importaba, en matemáticas, era la verdad, no la consistencia. (Pero, ¿son compatibles la verdad y la inconsistencia?)

Al margen de tales disputas, la teoría de Gödel dejaba algunas cuestiones matemáticas por resolver. Y estas cuestiones señalaban el camino para mitigar los daños. De acuerdo, un sistema axiomático, como las matemáticas, podría originar proposiciones arbitrarias (cuya veracidad o falsedad no podría ser demostrada), pero, ¿era posible determinar si tal proposición era arbitraria desde dentro del sistema? En otras palabras, ¿podría identificarse semejante proposición utilizando una serie de reglas derivadas de los axiomas básicos en los que se fundaba el sistema? ¿Podría determinarse mediante una serie de pasos concretos, de procedimientos mecánicos que cualquiera, o incluso una máquina, pudiera seguir? De ser así, estas proposiciones arbitrarias podrían ser identificadas y olvidadas sin que todo el sistema se resintiera. Sin embargo, si no pudieran ser identificadas de este modo, todo estaba perdido: las matemáticas padecerían una inconsistencia endémica.

Éste era el problema que Turing se había propuesto resolver. Era un proyecto extremadamente ambicioso: la solución era crucial para las matemáticas. Para poder resolverlo, Turing inventó un concepto cuyas consecuencias desbordarían los límites de las matemáticas.

¿Cuáles eran los procedimientos mecánicos (o reglas) que podrían utilizarse para determinar si una proposición matemática era susceptible de ser demostrada o no? Estas reglas se adentraban en el corazón mismo del cálculo. ¿Qué era un número computable y cómo se calculaba? El cálculo era un proceso estricto, como un proceso realizado por una máquina. Turing intentó definir la naturaleza teórica de una máquina semejante, ahora conocida como "la máquina de Turing".

Esta máquina sólo funcionaba de acuerdo con unas reglas, y podía calcular cualquier cosa para lo que existiera un algoritmo, es decir, una secuencia precisa de pasos que condujera a una conclusión.

Por ejemplo, tomemos el procedimiento de encontrar los factores de un número (es decir, los números primos por los que es divisible). Un ejemplo sencillo: para hallar los factores de 180, *dividir por el número primo más bajo posible, hasta llegar a un número que no sea divisible por ese número primo y repetir el proceso con el siguiente número primo ascendente, hasta completar la división* (los números primos son los que sólo son divisibles por 1 y por sí mismos, por ejemplo: 2, 3, 5, 7, 11, 13…).

$180 \div 2 = 90$
$90 \div 2 = 45$
$45 \div 3 = 15$
$15 \div 3 = 5$
$5 \div 5 = 1$
Así, $180 = 2^2 \times 3^2 \times 5$

El procedimiento, o algoritmo, es el texto que está en cursiva, y puede aplicarse a cualquier número. Se puede aplicar de forma mecánica, es decir, mediante un pensamiento mecánico o una máquina que piense.

Había que probar que una máquina de *estas* características siguiera un procedimiento determinado y realizara una tarea de acuerdo con las reglas del procedimiento. Si las reglas sirvieran para calcular números primos, calcularía números primos. Si fueran reglas de ajedrez, podría jugar al ajedrez. Cada máquina seguiría, simplemente, el procedimiento asignado.

Turing postuló después lo que él denominó una máquina "universal". En esta máquina se podría introducir un número que equivaliera a todo un procedimiento de otra máquina Turing, y la máquina seguiría el procedimiento y se comportaría del mismo modo que la máquina Turing original: jugaría al ajedrez, calcularía números primos, etcétera.

Desde este punto de partida (puramente teórico), Turing pasó a intentar demostrar su tesis. Lo que Gödel había demostrado era lógico. Lo que Turing iba a demostrar se parecía a la teoría de Gödel (en sus conclusiones), pero era *matemático.*

Turing propuso el concepto de una máquina capaz de reconocer proposiciones arbitrarias dentro de un sistema matemático. Esta máquina teórica tendría que convertirse en una máquina Turing universal. Se introduciría en ella un número que, en forma de clave, llevaría la descripción de otra máquina Turing y actuaría de la misma forma que ésta. Pero, ¿qué pasaría si en esta máquina universal hipotética se introdujera un número que indicara su *propia descripción*? ¿Cómo se comportaría como ella misma, comportándose como ella misma, comportándose como ella misma, y así sucesivamente? ¿Y cómo podría seguir el procedimiento de comportarse como ella misma, cuando ya estaba comportándose así?

Obviamente, la máquina se volvería loca. En términos teóricos, se vería enfrentada a una contradic-

ción. Dicho de otro modo: una máquina semejante no podría existir, *ni siquiera en teoría.* Lo que significaba que un cálculo semejante no era posible. Era imposible definir una serie de reglas que pudieran determinar si una proposición era susceptible de demostración (o refutación) empleando únicamente procedimientos extraídos del mismo sistema.

Así pues, las matemáticas no eran sólo incompletas desde el punto de vista de la lógica, como Gödel había demostrado, sino que también eran incompletas desde un punto de vista matemático. No había forma matemática de que las matemáticas se separasen de sus propias proposiciones arbitrarias. Turing publicó sus descubrimientos en un trabajo titulado *On Computable Numbers, with an Application to the Entscheidungsproblem* (*Sobre los números computables, con una aplicación al* Entscheidungsproblem, esta última palabra impronunciable hace referencia al problema de la determinación lógica según el planteamiento de Hilbert).

Todos los que entendieron el trabajo, aunque fuera en una medida muy limitada, admitieron que era excepcional. (Aunque pocos, en una época en que aún no había computadoras, podían darse cuenta de que marcaba un hito.) Hasta entonces, las nociones matemáticas fundamentales de computabilidad y números computables habían permanecido confusas; ahora quedaban aclaradas. El cálculo se definía en términos matemáticos precisos, tan precisos que definían el proyecto teórico de una máquina que pudiera realizar esa labor. Al mismo tiempo, Turing había definido los límites de lo que *esta máquina podría hacer.*

La máquina Turing era una computadora teórica. Actualmente está considerada el prototipo teórico de la computadora electrónica digital. Turing había dejado trazada la teoría de las computadoras antes de que la primera (tal y como la conocemos) se fabricara.

La tarea que había que acometer era obvia. Sin embargo, en 1937, cuando finalmente se publicó el trabajo de Turing, estaba aún por encima de las capacidades humanas. (Se había producido un retraso en la publicación, porque no se encontraba a nadie que tuviera los conocimientos necesarios para juzgar la originalidad del trabajo de Turing.) Para cuando se publicó *Sobre los números computables,* Turing había cruzado el Atlántico y estaba haciendo un doctorado en Princeton. Aquí, el departamento de matemáticas compartía edificio con el Instituto de Estudios Avanzados, de reciente fundación. (Este centro para la investigación científica teórica se había creado en 1933 y estaba convirtiéndose rápidamente en el mejor del mundo en su categoría, aunque, al igual que muchos de sus miembros judíos alemanes, en aquel entonces no tenía un asentamiento fijo.) Turing se encontraba entre los dioses. Einstein y Gödel eran residentes, así como Courant y Hardy. La mayoría de ellos permanecían apartados del mundanal ruido, y apenas advirtieron la presencia del joven inglés que, según confesaría atinadamente a su madre, vivía como un "solitario empedernido".

Sin embargo, sí llegó a trabar contacto con uno de los dioses del Olimpo, el matemático austrohúngaro Von Neumann. "Johnny" Von Neumann no era en absoluto un solitario desaliñado (como Einstein, Gödel, Turing, etc.), sino un elegante vienés, capaz de crear en segundos fórmulas matemáticas (y mezclas para cocteles) de increíble complejidad. Sólo Von Neumann valoró el logro de Turing en toda su dimensión. Advirtió que el joven inglés había creado, en realidad, todo un nuevo campo. (Turing lo había denominado "computabilidad", ya que no había otro término.) Von Neumann fue el que comprendió las posibilidades prácticas de la materia. Se percató de que el siguiente paso era construir una máquina Turing.

Entre tanto, Turing seguía con su doctorado, relacionado con otro de los "problemas" de Hilbert. En 1900 Hilbert había esbozado 23 problemas importantes para que los resolvieran los matemáticos del siglo XX añadiendo, de acuerdo con el positivismo típico del cambio de siglo, que "todos los problemas matemáticos son resolubles". Turing ya había demostrado que se equivocaba, pero ahora decidió hacer un esfuerzo definitivo para resolver un problema relacionado con la hipótesis de Riemann, que Hilbert había calificado como "la más importante en matemáticas". Hardy ya había luchado con este problema durante treinta años, pero sin éxito.

Por decirlo de forma sencilla, el problema de Turing estaba relacionado con la frecuencia de los números primos. A principios de los años de 1790, Karl Gauss, niño prodigio alemán de 15 años, del que muchos decían que era el único matemático capaz de equipararse a Newton, descubrió que, aparentemente, los números primos aparecían con menos frecuencia, según el patrón regular. Para el número n, el espacio entre los primos aumentaría como el logaritmo natural de n. Se descubrió que esta teoría ofrecía sólo errores marginales. Bernhard Riemann, uno de los sucesores de Gauss como catedrático de matemáticas en Göttingen, mejoró con su aportación la teoría, que incluía la hipótesis Riemann, de enorme complejidad.

Pero ni siquiera la fórmula de Riemann era absolutamente correcta.

Se descubrió que el método del logaritmo sobreestimaba ligeramente el número de primos y después de que millones de cálculos lo confirmaran, incluso para los números más altos, se aceptó que siempre era así. Entonces, uno de los colaboradores de Hardy, J. E. Littlewood, descubrió que si la hipótesis de Riemann era cierta, no podía ser así. Había un paso a partir del cual se producía la diferencia, y este paso se producía antes de llegar al número

$$10^{10^{10^{34}}}$$

Este número es inconcebiblemente alto. Como Hardy señaló, es el "mayor número utilizado con algún fin matemático". Este número, escrito en números decimales enteros llenaría una cantidad de libros con una masa superior a la del planeta Júpiter, según Hodges, biógrafo de Turing.

Los problemas que esto planteaba eran fundamentales para la teoría de los números. Turing se debatió valientemente entre estos ejercicios mentalmente agotadores, pero con escaso éxito (hasta hoy, la hipótesis Riemann, por ejemplo, sigue sin demostrarse).

Para Turing, Estados Unidos era, según el día, revitalizante y turbador. Trabajaba demasiado y pasaba demasiado tiempo solo, por lo que empezó a padecer depresiones. A esto se sumó un embarazoso incidente, relacionado con su homosexualidad y un acercamiento mal interpretado. En una de las cartas que escribió a su novio de Cambridge (que trabajaba en aquel tiempo como director de un colegio en Walsall), mencionaba de pasada, como solía, que había ideado una forma de suicidarse comiendo una manzana mortal.

Tras dos años en Estados Unidos, Turing volvió a Gran Bretaña, después de rechazar la oferta de Neumann de trabajar con él en el Instituto de Estudios Avanzados. Le fue renovada la beca en el King's College, y retomó su vida normal en Cambridge. Escribió a su madre para pedirle su osito de peluche, y asistió ansioso al estreno de *Blanca Nieves y los siete enanos* en el cine local. En especial, le chocó la escena en que la bruja malvada introduce una manzana colgada de un hilo en el puchero de veneno hirviendo y, después de ver la película, empezó a repetir su embrujo:

> Sumérjase la manzana. Que la poción de la muerte dormida la impregne bien.

Los que escuchaban sus cánticos no tenían idea de que recientemente había pensado en suicidarse con una manzana. Por su orientación sexual, estaba acostumbrado a vivir en una mentira, pero no podía evitar esa forma de franqueza oblicua (en tanto la homosexualidad fuera ilegal, era fundamental ocultarla, a pesar de que era conocida entre unos pocos de sus colegas de Cambridge). La gente seguía pensando que Turing era una persona difícil de conocer; su carácter era, incluso entonces, un enigma. Sin embargo, sí que había una clave para el que se molestara en buscarla. Desgraciadamente, a partir de entonces, empezó a resultar cada vez más difícil saber dónde buscarla.

Así pues, Turing siguió manteniendo y, de hecho, insistía en mantener su posición arbitraria dentro del sistema. Nadie podía desaprobar a Turing (era miembro del King's College y un matemático brillante), pero tampoco podían aprobarlo (su madre, por ejemplo; o a causa de su homosexualidad, entonces ilegal).

Fue en torno a esta época cuando Turing conoció al filósofo austriaco Ludwig Wittgenstein y empezó a asistir a sus clases. Wittgenstein daba clases a unos pocos elegidos, que se sentaban en unas tumbonas, en sus desnudas habitaciones, y escuchaban, mientras él "pensaba en voz alta". Ello implicaba que hubiera frecuentes periodos de largo y angustioso silencio. Después una pregunta, a la que seguía un salvaje interrogatorio, si alguien tenía la osadía de intentar responderla.

Wittgenstein daba clases sobre los fundamentos de las matemáticas, pero desde una perspectiva filosófica. Intentaba descubrir la naturaleza precisa de las matemáticas, es decir, qué eran exactamente, más que cómo funcionaban. Turing no tenía la capacidad filosófica de Wittgenstein (de hecho, ningún otro ser humano vivo poseía esa fuerza intimidatoria), pero era mejor matemático que él. En opinión de Turing, qué eran las matemáticas y para qué servían eran asuntos inextricables. Se negó a dejarse intimidar por los amenazantes ataques de Wittgenstein.

En un determinado momento, Wittgenstein propuso que un sistema como el de las matemáticas o la lógica podría seguir siendo válido, pese a contener una contradicción. El propio Turing había demostrado que las matemáticas contenían inconsistencias, pero eso no era lo mismo que "contradicciones". Le explicó a Wittgenstein que, si se intenta construir un puente basándose en unas matemáticas que contuvieran una contradicción, el puente se caería. Wittgenstein insistió en que, por el contrario, la naturaleza de las matemáticas y su aplica-

ción eran asuntos diferentes. Sin embargo, el trabajo *Sobre los números computables* de Turing había demostrado lo profundo que era el vínculo entre las matemáticas puras y las matemáticas aplicadas. Había resuelto un problema teórico fundamental de las matemáticas proponiendo una "máquina" que, pese a ser teórica, no dejaba de ser una máquina, es decir, un aparato práctico que, en principio, podría construirse.

Resulta interesante el hecho de que la demostración de Turing de que todo sistema (como el matemático o lógico) contiene proposiciones irresolubles demostraba lo contrario que la filosofía inicial de Wittgenstein. A este respecto, Wittgenstein sostenía que cualquier problema, siempre que se expresara de una forma lógica adecuada, podía ser resuelto.

Turing se adentraría ahora en el terreno de las matemáticas aplicadas con vehemencia. En 1939 estalló la guerra contra la Alemania nazi y Turing fue destinado a tareas de inteligencia. Le encargaron dirigir un equipo encargado de descifrar códigos, en el edificio de inteligencia de Bletchley Park, unos 100 km al norte de Londres. Este proyecto era alto secreto y estaba estrictamente vigilado por los militares.

El ejército no estaba preparado para lidiar con Turing. Aunque mantenía su extraño comportamiento infantil, su apariencia había evolucionado hacia una excentricidad madura propia de Cambridge (un estado muy próximo a la rareza clínica).

A primera vista, Turing parecía un hombre que hubiera pasado una mala noche. Llevaba el pelo desordenado, las uñas sucias, se sujetaba los pantalones con una vieja corbata del colegio y había dejado de afeitarse con regularidad (a menudo se cortaba al afeitarse y se desmayaba al ver la sangre). Para aquel entonces, la voz de Turing se había convertido en una especie de tartamudeo agudo y propio de la clase alta, que en ocasiones se veía interrumpido por una molesta risa nerviosa (parecida, según se cuenta, al chirriante rebuzno de un burro). Cuando se perdía en sus pensamientos —algo que le ocurría con frecuencia—, solía acompañar las intensas operaciones mentales que realizaba de unos chillidos y graznidos igualmente intensos.

La actitud social de Turing era igualmente molesta. Obviaba a toda persona cuyo intelecto no le pareciera digno de consideración; por supuesto, esto incluía a todo el personal del ejército que dirigía el establecimiento. Para empeorar las cosas, solía trabajar durante largos periodos que duraban noches y días eternos, pero después el oficial encargado de la inspección podía encontrarlo jugando al ajedrez con el chico de los recados, o durmiendo siestas de tardes enteras, con la cabeza en el escritorio.

Turing no era, sencillamente, carne de la disciplina militar. Lo que es peor, parecía que no se tomaba en serio su trabajo. Y su trabajo, o el trabajo que se *suponía* que debía hacer (como le recordaba enérgicamente el oficial al mando), era en realidad muy serio.

De hecho, era más serio incluso de lo que el ejército se imaginaba. Los esfuerzos de Turing y de los diversos equipos de personajes de inteligencia superior que trabajaban en Bletchley cambiaron, casi con certeza, el rumbo de la guerra.

La historia de Bletchley empezó en 1938, cuando Robert Lewinski, un joven ingeniero polaco, se presentó en la embajada británica de Varsovia. Afirmaba haber trabajado en Alemania, en una fábrica de máquinas de señales cifradas. Lewinski había logrado memorizar los detalles de la máquina. Rápidamente, fue conducido de forma clandestina de Polonia a París, donde se encargó de la supervisión de la fabricación de una máquina. Los británicos habían oído hablar de estas máquinas, conocidas como *Enigma*, utilizadas por el mando alemán para enviar órdenes cifradas a las fuerzas en campaña. Los comandantes de los submarinos alemanes (*U-boote*) también podían emplearlas para identificar su posición, para poder ser enviados hacia los convoyes enemigos más cercanos localizados.

El *Enigma* era sorprendentemente fácil de utilizar, aunque aparentemente su sistema de códigos era imposible de descifrar. Básicamente, el sistema consistía en dos máquinas. En la máquina emisora se configuraba una clave y el mensaje, sin cifrar, simplemente se introducía en la máquina. Tres (o más) brazos rotores eléctricos, dependiendo de la clave, desordenaban automáticamente el mensaje, y éste era finalmente transmitido. En el otro extremo, se configuraba la máquina *Enigma* receptora con la misma clave y ésta volvía a ordenar el mensaje y lo imprimía ya descodificado. Los rotores, que giraban de forma independiente, permitían literalmente miles de millones de permutaciones, por lo que cualquier enemigo que lograra interceptar una transmisión codificada se enfrentaba a una tarea aparentemente imposible, si pretendía descifrar el código. Cada día se enviaban miles de mensajes, y la clave se cambiaba tres veces al día. Los alemanes tenían razones para creer que su sistema de comunicación era indescifrable.

Ahora, gracias a Lewinski, el personal del Servicio de Inteligencia británico que trabajaba en Bletchley

sabía exactamente cómo se construía y cómo funcionaba una máquina *Enigma.* Pero eso no era suficiente. De hecho, quedaba mucho por hacer. Las dificultades generadas por el *Enigma* eran enormes. Cada vez que se pulsaba una letra, al mecanografiar un mensaje para introducirlo en la máquina, los rotores giraban. Así que, aunque se pulsara la misma letra varias veces seguidas, invariablemente produciría letras diferentes en la versión "desordenada". Para descifrar el código era necesario saber cuál era la clave definida en la máquina, ya que ésta era la que controlaba la posición inicial de los rotores. Y si el *Enigma* era de sólo tres rotores, había un millón al cubo (10^{18}) de claves posibles (los mensajes de alto secreto de la Luftwaffe se enviaban en máquinas *Enigma* de *diez* rotores).

Turing y su equipo (al que pronto se sumaron muchas de las mentes más brillantes matemáticas del país) se enfrentaban a una labor monumental. Debían buscar en la miríada de mensajes cifrados cualquier combinación, patrón o posibilidad que pudiera significar algo y después intentar reconstruir la configuración de la clave.

Turing hizo una inmediata valoración de la situación muy en su estilo. El problema era sencillo, al menos en teoría. Éste era un trabajo para una máquina Turing. La máquina descrita por Turing en su trabajo *Sobre los números computables* no era totalmente teórica. Turing había proyectado una máquina en la que las instrucciones se introducían en cinta de papel. La cinta estaba dividida en cuadrados que la máquina leería de uno en uno. En su forma más simple, cualquier problema podía reducirse a una serie de instrucciones en dígitos binarios (*bits*). Como Turing conjeturó correctamente, el problema que planteaba el *Enigma* no era un problema arbitrario. Esto quiere decir que era susceptible de ser resuelto: si se introducían las instrucciones apropiadas en una máquina Turing, ésta daría con la solución. Pero eso estaba muy bien en teoría; la práctica era otra cosa.

Turing y su equipo se pusieron a construir una máquina electromagnética que pudiera funcionar a gran velocidad, buscando en los mensajes codificados del *Enigma* algún tipo de regularidad característica recurrente o combinación que pudiera ser descifrada. (En ocasiones, como resultado de alguna acción del enemigo, encontraban la clave de algún mensaje anterior, obteniendo así más datos acerca de cómo funcionaba el *Enigma,* aunque sólo se tratara de claves ya obsoletas.) La máquina descodificadora de Turing se conocía como *Colossus;* eran tan grandes las dificultades a las que se enfrentaba el equipo de Bletchley, que se construyeron al menos diez versiones de la máquina.

El primer *Colossus* empezó a funcionar en diciembre de 1943. Los detalles de esta máquina son imprecisos, debido a la obsesión por el secreto de las esferas militares y del gobierno británico. (De hecho, hasta hace poco algunos códigos empleados en las guerras napoleónicas seguían siendo información clasificada.) Al parecer, el *Colossus* utilizaba 2 400 tubos de vacío, que realizaban cálculos en el sistema binario. No contenía un programa almacenado, aunque sí realizaba funciones semejantes a las de una computadora. Así pues, ¿se puede decir que esta máquina de Turing era efectivamente una máquina de Turing, tal como se había ideado? Sigue habiendo dudas al respecto. Sin embargo, el *Colossus* se considera en general como el antecesor de la computadora digital electromagnética.

Al margen de lo que fuera, el *Colossus* significó un enorme avance para la tecnología. La potencia combinada de sus cinco procesadores podía leer 25 000 caracteres en un segundo. Sin embargo, aún no era suficiente. Los submarinos alemanes estaban hundiendo barcos de la flota aliada en el Atlántico a un ritmo alarmante. Nada se podía hacer: todavía se necesitaban varios días para descifrar los mensajes que los *Enigma* enviaban desde y hacia los submarinos. Trabajando día y noche se consiguió reducir gradualmente este margen. Llegado un determinado momento, Gran Bretaña sólo contaba con reservas de alimentos para una semana. Por fin, el *Colossus y* el equipo de Bletchley estaban logrando descifrar los códigos en horas y luego en minutos. Finalmente, se logró determinar la posición de todos los submarinos alemanes en el Atlántico, y la pérdida de barcos de los convoyes aliados disminuyó radicalmente.

Inmediatamente, los alemanes empezaron a sospechar. Pese a sus sospechas, sin embargo, seguían convencidos de que su código *Enigma* era indescifrable. Los británicos debían de estar recibiendo información de una red de espías bien situados. No era necesario, por lo tanto, inventar una máquina codificadora más avanzada; la Gestapo pasó a la acción y empezó a realizar detenciones.

Mientras tanto, Turing seguía comportándose como siempre, arreglándose ocasionalmente para visitar a su madre (que estaba muy desilusionada, porque sus labores militares no lo obligaban a raparse el pelo). Aparentemente, el aspecto de Turing era un indicativo

de sus profundas dudas psicológicas. Seguía haciendo comentarios abiertos sobre la homosexualidad delante de sus colegas (aunque de ahí no pasaba), pero al mismo tiempo empezó a mantener una relación con una de las criptoanalistas (quien, por cierto, le enseñó a tejer guantes). Su relación duró seis meses, al cabo de los cuales Turing tuvo que admitir su futilidad.

A un periodo de relativa elegancia en el vestir le siguió un ataque de indiferencia. Sin embargo, y a pesar de su apariencia desaseada y a las largas horas de trabajo, Turing seguía estando en muy buena forma. Varias veces por semana se iba a correr por campos y bosques. Los lugareños lo miraban perplejos, cuando cogía un puñado de hierba al pasar y se iba masticándolo mientras corría. Ésta era la forma de Turing de compensar las deficiencias vitamínicas de los tiempos de guerra (antes solía comer una manzana antes de dormirse).

Esta tendencia a la autosuficiencia idiosincrásica se extendió a terrenos inesperados. Cuando estalló la guerra, Turing estaba convencido de que Gran Bretaña sería invadida. Había convertido sus ahorros en lingotes de plata y los había enterrado secretamente en los bosques cercanos a Bletchley Park. Después, había cifrado su ubicación y la había memorizado. (Desgraciadamente, este código sí logró vencer a Turing. Tras la guerra, no logró recordarlo y nunca pudo recuperar sus lingotes de plata, a pesar de realizar varias "cazas del tesoro" sistemáticas y exhaustivas, y de que llegara a inventar su propio detector de metales.)

En Bletchley ya no se dedicaban únicamente a localizar submarinos. Rápidamente, casi todas las comunicaciones alemanas se convirtieron en un libro abierto para ellos. Este trabajo tenía tal importancia que Turing llegó a cruzar el Atlántico para contactar con los estadunidenses. En el viaje se encontró con Von Neumann, que también había empezado a poner en práctica las ideas de *Sobre los números computables*. En el departamento de ingeniería de la Universidad de Pennsylvania, los estadunidenses habían empezado a trabajar en el ENIAC (Dispositivo electrónico de integración y cálculo numérico). Esta máquina era aún más colosal que el *Colossus* y contenía la asombrosa cantidad de 19 000 válvulas. Sin embargo, el ENIAC no estaría listo hasta después de la guerra. (Por otro lado, los alemanes, sin que los aliados lo supieran, estaban también trabajando en este campo. En 1943, Konrad Zuse había creado la primera calculadora multiusos controlada por un programa, que se utilizaba para el análisis en la fabricación de bombas volantes. Sin embargo, el laboratorio subterráneo de Zuse, situado en Berlín, fue bombardeado un año después.)

Cuando la guerra terminó, Turing estaba trabajando en Hanslope Park, muy cerca de Bletchley, en un proyecto de registros de voz denominado *Delilah* (por Dalila, la figura bíblica cuya engañosa voz había causado estragos en Sansón). Para entonces, el trabajo que Turing había realizado con el *Colossus* había aumentado considerablemente su comprensión de la maquinaria electrónica. Había empezado a reflexionar sobre cómo podrían las máquinas imitar el funcionamiento de la mente humana.

En 1945 Turing se unió al recién fundado Laboratorio Nacional de Física, en Teddington, a las afueras de Londres. Allí encabezaba el proyecto para la construcción de una máquina de cálculo automático (conocida como ACE). Turing intentaba diseñar una computadora electrónica digital con un programa interno. El ACE se benefició enormemente de la experiencia de Turing en la fabricación y el funcionamiento del *Colossus,* pero su punto fuerte seguía siendo meramente teórico. Al igual que la máquina Turing universal propuesta en *Sobre los números computables,* el ACE debería seguir un "diseño lógico" global, que incorporaría muchos procedimientos lógicos complejos. Desgraciadamente, estos procedimientos dieron lugar a una serie de dificultades técnicas, que no interesaban tanto a Turing. Su diseño se adelantaba mucho a su tiempo: era muy superior al ENIAC (la primera de las denominadas "máquinas Von Neumann") que los estadunidenses estaban a punto de terminar, y estaba más avanzado que cualquier otro proyecto que se estuviera realizando en Inglaterra. Sin embargo, el ACE no tuvo que enfrentarse sólo a problemas técnicos. Los problemas más importantes se debieron a la falta de fondos y a la política científica.

A diferencia de otros campos de investigación científica, la política científica florece gracias a la falta de fondos. Entre los racionamientos de la posguerra en Gran Bretaña (incluso el pan estaba racionado), las políticas científicas lograron un avance histórico: entraron en su era bizantina. Tales complejidades y sutilezas estaban totalmente al margen de los intereses de un simple mago de las matemáticas como era Turing. Ni en sus mejores momentos había sido un hombre diplomático y, así, Turing se encontró con que sus solicitudes de fondos eran siempre rechazadas.

La razón que se suele alegar para ello es que Turing tenía un carácter antipático y su aspecto descuidado e infantil hacían que la gente no lo tomara muy en serio. (Cuando asistía, por ejemplo, a alguna reunión de un departamento de Whitehall prefería cubrir los cerca de 15 kilómetros de distancia corriendo, atravesando Londres, en lugar de utilizar el transporte público. Cualquiera que haya estado en la meta de una carrera a campo traviesa, puede imaginarse el efecto que le causaba la llegada de Turing a un comité de funcionarios.) Sin embargo, la falta de don de gentes de Turing —por llamarlo de forma elegante— no era la única razón. Su proyecto era el mejor, pero otros eran mejores proyectando intrigas.

Para 1947 Turing se había percatado de que así no iba a ningún lado. La versión oficial es que dimitió de su puesto en el Laboratorio Nacional de Física y dejó que otros concluyeran el proyecto ACE. Lo que no está muy claro es si saltó él solo o si lo empujaron.

Irónicamente, esto es lo mejor que le podía haber ocurrido. Volvió a Cambridge, donde enseguida se embarcó en un trabajo revolucionario sobre la teoría informática. A pesar de su implicación en el proyecto *Colossus* y de haber sentado las bases para el ACE (que finalmente se construyó, con éxito), y su posterior participación en el desarrollo de la computadora, es por su trabajo teórico por lo que se le recuerda.

Como ya hemos visto, Turing había proyectado desde el principio que la máquina Turing realizara funciones propias de la mente humana. Pero, ¿era capaz una máquina de equipararse a una mente humana? Lo que Turing propuso y analizó entonces fue el concepto de "maquinaria inteligente". Las objeciones morales, humanas y religiosas fueron descartadas con su tacto característico: "Como son puramente emocionales, ni siquiera hace falta realmente refutarlas". Las objeciones científicas y filosóficas eran más serias. Una máquina capaz de disponer de inteligencia implicaba un enfoque mecánico de la inteligencia, que a su vez implicaba un determinismo. Sin embargo, aparentemente, la inteligencia humana incluía un elemento de libre albedrío.

La tediosa y fútil discusión filosófica entre el libre albedrío y el determinismo no ha lugar aquí. El argumento de Turing es que la mente humana *parece,* desde fuera, tener capacidad de libre albedrío. Se *comporta* como si la tuviera.

Así pues, las operaciones de la inteligencia no son meramente mecánicas, pero Turing sugería que podían ser realizadas por una máquina. En esto… ¿no hay algo ilógico? En el sentido verbal, tal vez. Sin embargo, la experiencia de Turing, adquirida en su trabajo durante la guerra en los proyectos *Colossus* y *Delilah,* indicaba lo contrario. Estas dos máquinas eran marcadamente deterministas, pero se había descubierto que también eran capaces de desarrollar un comportamiento aleatorio. (Por algo el *Colossus* había necesitado un equipo de docenas de "cuidadores" para llevarlo por el camino correcto.)

En un determinado nivel, estos ordenadores primitivos habían sido totalmente deterministas. Sin embargo, en otro nivel habían mostrado un claro comportamiento aleatorio, que parecía imitar el libre albedrío. Había una grieta en la armadura: una grieta diminuta, pero real.

El argumento fundamental de Turing es que las máquinas eran capaces de aprender. Así, podrían ampliar sus operaciones más allá de lo meramente mecánico. Se podía enseñar a una máquina a mejorar su comportamiento, hasta que llegara a mostrar "inteligencia".

En este punto, Turing superó otra de las objeciones potenciales, que podría haber limitado su tesis. Una máquina podrá mostrar inteligencia, pero sólo será el reflejo de la inteligencia de su creador. Turing no estaba de acuerdo. Utilizaba la analogía del maestro y el pupilo. El alumno podría superar en brillantez a su maestro, desarrollando una información cualitativamente superior, aunque utilice sólo la inteligencia que le ha programado su maestro. Turing llevó más lejos su argumentación. Era posible crear una máquina que jugara al ajedrez (siguiendo las reglas que se introdujeran en ella). Sin embargo "jugar contra una máquina así da una impresión muy real de estar enfrentándose intelectualmente a algo vivo". Como la computadora podía aprender, su comportamiento superaba el determinismo mecánico y mostraba un elemento de libertad que se asemejaba al de una inteligencia viva (lo cual no quería decir, necesariamente, una inteligencia humana).

Turing estaba planteando preguntas que se habían presentado desde los orígenes de la filosofía: ¿qué significa ser humano?, ¿qué es exactamente la inteligencia humana? Pero enfocaba estas preguntas desde una perspectiva original: ¿podía una máquina adquirir estas cualidades?, ¿cómo distinguimos una inteligencia humana de la inteligencia de una máquina?

Turing estaba reflexionando en un plano que trascendía las matemáticas, los números racionales e in-

cluso a las computadoras. De hecho, se involucró tanto en sus propios procesos de reflexión que acabó mirando el mundo como si él mismo fuera una computadora. A medida que exploraba las posibilidades del pensamiento, se fue adueñando del mecanismo de una computadora como su medio de pensamiento. ¿Qué es la inteligencia?

Hasta cierto punto, esta identificación con la máquina empezó a impregnar el resto de su vida. El hecho de verse a sí mismo como una máquina le proporcionó un enorme alivio psicológico para la continua confusión de su vida interior.

El regreso de Turing a Cambridge fue lo mejor que le pudo ocurrir, tanto profesional como personalmente. Durante varios periodos en la vida de Turing estas dos categorías fueron inseparables. Aquí, en Cambridge, ya no lo eran. Puede que Turing se haya identificado psicológicamente con una computadora, pero su comportamiento no lo dejaba traslucir.

Turing tenía ya 35 años, aunque seguía pareciendo unos diez años más joven. En el King's College donde vivía, su intelecto era considerado (por los pocos que le entendían) como uno de los mejores de Cambridge. Sin embargo, desgraciadamente, su comportamiento estaba muy por debajo de su intelecto. Turing adquirió la costumbre de pasearse por el patio de su facultad, en busca de jovencitos a los que invitar a tomar un té en sus habitaciones. Por las tardes solía ir a visitar inesperadamente a otros jóvenes. Solía decir: "A veces estás hablando con alguien y sabes que, en tres cuartos de hora, estarás pasando una noche maravillosa o te echarán a patadas del cuarto". El hecho de identificarse con una computadora aparentemente lo liberó de todo recelo que le impidiera manifestar su sexualidad.

Turing debió de ser una compañía peligrosa durante este tiempo. Sin embargo, afortunadamente para todos los implicados, la situación no se prolongó. Antes de que las autoridades de la universidad pudieran considerar que su comportamiento superaba los límites de la excentricidad aceptable, Turing encontró un novio estable: Neville Johnson. Este joven había obtenido una beca del instituto Sunderland y cursaba su tercer año de matemáticas. Neville ya había realizado sus dos años de servicio en el ejército, y a Turing le atrajo su actitud un tanto tosca, pero resuelta. Al parecer, Neville Johnson fue uno de los pocos que llegaron a atravesar el caparazón de Turing, su defensa frente al mundo. Un día, mientras estaban juntos en la cama, Turing confesó: "Tengo una relación más estrecha con esta cama que con otras personas." Sin embargo, pese al profundo afecto que sentía por Neville, Turing seguía rondando ocasionalmente por el patio. A estas alturas, una completa rendición al amor debía de ser ya imposible. Una computadora podía tener inteligencia, pero que tuviera emociones seguía siendo una cuestión teórica.

Entre tanto, se realizaban grandes avances en el terreno práctico. En Cambridge se estaba construyendo una máquina de computación llamada EDSAC (computadora automática de almacenamiento electrónico retardado), pero, sorprendentemente, Turing decidió evitar todo contacto con el equipo responsable. En lugar de eso, después de un año en Cambridge aceptó el puesto de director adjunto del laboratorio de computadoras en la Universidad de Manchester. En este laboratorio se estaba construyendo la Máquina Digital Automática de Manchester (conocida popularmente como MADAM).

El 21 de junio de 1948 el MADAM se convirtió en la primera computadora electrónica con un programa almacenado que funcionara, descomponiendo un número en sus factores primos (por ejemplo, $4\,620 = 2^2$, 3, 5, 7, 11).

El MADAM cumplía todas las especificaciones teóricas de una máquina Turing (tal y como se describía en *Sobre los números computables*), aunque no había sido construida según el diseño de Turing. Sin embargo, Turing se sumó con entusiasmo a la ampliación de sus capacidades originales. Diseñó unos circuitos para los equipos físicos de entrada y salida, y llegó incluso a obtener de Bletchley un teletipo de una máquina de códigos alemana. Turing se encontró pronto dedicando un gran número de horas al análisis matemático, aunque frecuentemente acababa resolviendo los problemas en breves momentos de iluminación intuitiva.

El trabajo en el MADAM no era sólo intelectual y técnico. Hacer "funcional" este monstruo en continuo crecimiento era una tarea hercúlea. El ayudante de Turing describía el proceso: "En la sala de la máquina, se avisaba al ingeniero y se utilizaban los interruptores manuales para introducir el programa de entrada; una vez hecho esto, era necesario correr escaleras arriba y poner la cinta en el lector y volver luego a la sala de la máquina". Si la máquina empezaba a leer la cinta y a seguir correctamente las instrucciones, el operario tenía que llamar al ingeniero para que conectara la corriente que activaba la función de impresión. "En cuanto el patrón del monitor indicaba que había finalizado la entrada, el ingeniero desconectaba la corriente de impresión. Solían

hacer falta varios intentos para introducir la cinta, y cada intento significaba un nuevo viaje a la sala de la cinta". Afortunadamente, Turing seguía estando en forma.

A pesar de tales dificultades atléticas, el MADAM pronto estuvo listo para acometer tareas más complejas. Sus tubos podían almacenar hasta 128 palabras (grupos de dígitos binarios que contenían instrucciones que la máquina podía utilizar) de 40 bits. Ésta no fue sólo la primera computadora operativa, sino también la primera utilizada para un objetivo constructivo a gran escala. Más adelante fue utilizada para calcular el diseño de la ruta marítima de St. Lawrence, una de las maravillas de la ingeniería del siglo XX.

Sin embargo, las tareas iniciales del MADAM fueron algo menos constructivas. Turing estaba más interesado en enseñarle a jugar al ajedrez y pasaba largas horas dedicado a ello, buscando la forma de mejorar su estrategia en el juego. Otros miembros del equipo no estaban igualmente conformes viendo al director adjunto del proyecto y al MADAM enzarzados en un combate intelectual y les hizo aún menos gracia su siguiente hazaña. Turing programó el MADAM para que escribiera una carta de amor. Ésta era la primera vez que se intentaba algo así, y la máquina escribió una misiva típica de la extraña pasión de alguien poco acostumbrado a expresar tales emociones:

> Querido cariñito:
> Eres mi ávido sentimiento amigo. Mi afecto se asocia extrañamente a tu deseo pasional. Mi deseo ansía tu corazón. Eres mi soñadora compasión: mi tierno deseo.
> Hermosamente tuyo,
> MUC

En este caso, el MADAM prefirió definirse como "MUC", la computadora de la Universidad de Manchester, elección cuyas asociaciones podrían haber resultado muy interesantes para Sigmund Freud.[1]

Las actividades de Turing eran igualmente interesantes, desde el punto de vista psicológico. No se había enamorado realmente desde su pasión escolar por Christopher Morcom. Desde luego, la prematura muerte de Christopher influyó en su incapacidad para comprometerse plenamente y de forma duradera con nadie, aunque tampoco se puede subestimar la importancia del peligro que implicaba mantener una relación amorosa homosexual en la Inglaterra de entonces. De cualquier forma, Turing se daba cuenta de su fracaso y éste lo atormentaba. (Su confesión, a Neville Johnson, de cómo recordaba su pasado — "tengo que pensar en la persona de la que estaba *enamorado* en ese momento"— no hace sino reforzar esta idea. Sus relaciones amorosas se veían frustradas, o culminaban en una breve pero insatisfactoria conflagración.)

A la luz de la identificación de Turing con una computadora, su proyecto de programar el MADAM para que redactara cartas de amor adquiere visos de patetismo. Éste no fue un acto inconsciente; Turing sabía lo que estaba haciendo, aunque los demás no se dieran cuenta. A estas alturas, se había definido abiertamente, aunque sólo de forma casual, en cuanto a sus preferencias homosexuales. Sin embargo, precisamente por eso sus colegas desconocían su secreto sufrimiento. Puede haber resuelto el problema del *Enigma,* pero seguía sin resolver el problema de su propio enigma.

Sin embargo, incluso esto seguía siendo un asunto secundario, reprimido u obviado. Turing continuaba sumergiéndose en su trabajo (el trabajo y la carrera de fondo seguía siendo su bromuro). Las partidas de ajedrez y las cartas de amor del MADAM eran fundamentales dentro del actual interés primordial de Turing: la "maquinaria inteligente" o, como se la conoce generalmente, la inteligencia artificial.

Las preguntas provocativas que Turing planteaba (y con las que frecuentemente se identificaba) sentaron las bases de este campo. Estas preguntas eran profundamente filosóficas, sin ser confusas, y al mismo tiempo seguían siendo estrictamente científicas, sin dar lugar a esos "milagros" aislados hacia los que la ciencia experimental puede degenerar con tanta facilidad. Como ocurre con la filosofía, pero no con una buena parte de la ciencia moderna, éste era un campo del conocimiento en el cual se podía vivir, ya que esclarecía la condición humana.

Turing planteó sus ideas en una serie de trabajos, de los cuales el más importante fue *Computing Machinery and Intelligence* (*Maquinaria de computación e inteligencia*), publicado en 1950. En él, Turing insistía en que se podía enseñar a las computadoras a pensar por sí mismas; eran capaces de generar un pensamiento original. De forma muy esclarecedora, tachaba de "sen-

[1] Fonéticamente, en inglés estas siglas podrían confundirse con *muck*, que equivale a *fuck*, forma popular para referirse al acto sexual. [N. de la T.]

timental" la generalizada oposición a esta noción. Para que los procesos de una computadora pudieran asemejarse a los caprichos de la inteligencia humana, Turing proponía la incorporación de un elemento aleatorio, por ejemplo, una rueda de ruleta.

Sin embargo, consideraba que muchas de las objeciones filosóficas eran tediosas y fútiles. No quería en modo alguno que la cuestión de la inteligencia de las computadoras quedara atascada en preguntas sobre el libre albedrío, la ética, la definición de la vida, etc. Así, eludió brillantemente estos problemas. Había una forma de saber si una máquina era inteligente o no: colocarla tras una pantalla y dejar que un ser humano le hiciera preguntas. La persona podría decidir entonces, basándose en respuestas escritas, si estaba tratando con un ser inteligente o una simple máquina. ¿Podría una máquina engañar a un ser humano para que pensara que era humana? Éste era el "juego de la imitación" que Turing proponía (conocido actualmente como el "test de Turing").

Turing demostró cómo un interrogador hábil podía poner a prueba la máquina, obteniendo de ella decisiones y juicios sutiles y, posiblemente, hasta respuestas emocionales. O al menos, eso parecería en las respuestas escritas. Sin embargo, Turing no esquivó todas las objeciones filosóficas (simplemente, sorteó las tediosas e improductivas). Su propio argumento filosófico era incontestable. Insistía en que el "juego de la imitación" debía aceptarse como un criterio básico. ¿Por qué? Porque así es como nosotros reaccionamos *entre nosotros.* No hay ninguna forma inmediata de saber si otra persona posee inteligencia o no. Solamente podemos inferir que son seres pensantes y conscientes comparándolos con nosotros mismos. Turing pensaba que no había ninguna razón por la que no pudiéramos actuar de la misma forma con las computadoras. Su pregunta era: "¿Por qué hay que tratarme a mí de forma distinta a una computadora?". (El hecho de que esta pregunta la planteara una persona que se veía a sí misma como una presenta una serie de cuestiones interesantes. ¿Hay alguien humano escuchándome?)

Adoptando, magnánimamente, la perspectiva humana, Turing llegó a sugerir una serie de objeciones a su argumento. La más seria de éstas se conoce como la "objeción de Lady Lovelace", por la colega de Babbage que fue la primera en plantearla. Lady Lovelace estaba convencida de que las computadoras eran incapaces de producir un pensamiento original, porque sólo pueden hacer lo que se les dice. Dicho de otro modo, sólo pueden funcionar dentro de los límites que se les hayan programado.

La respuesta de Turing era tan calculadora como una computadora: cuando programamos una computadora, sólo tenemos una vaga idea general de lo que le hemos pedido que haga. Desde luego, no hemos pensado en todas las implicaciones de la tarea.

Por analogía, hemos visto que las matemáticas fueron entendidas como una serie de números y procedimientos sencillos, como los que se podían introducir en una computadora. Sin embargo, las implicaciones de este sistema han demostrado que no son en absoluto sencillas. De hecho, no sólo han demostrado ser totalmente inagotables, sino que también han desarrollado su propia inconsistencia. Como apuntó Ehrensvard: "Hay momentos en que hasta las matemáticas parecen tener cerebro propio".

Finalmente, esta forma de pensar llevó a Turing más allá del campo de las computadoras, hacia la morfogénesis (la evolución mediante el desarrollo de patrones en los organismos). Turing se percató de que, al igual que en matemáticas, cualquier sistema simple crece en complejidad. Una estructura simétrica uniforme se desarrolla a través de la difusión de su forma en una estructura asimétrica con un patrón propio. En 1952 Turing publicó su primer trabajo sobre este tema: *The Chemical Basis of Morphogenesis* (*La base química de la morfogénesis*).

Este trabajo plantea la siguiente cuestión: ¿cómo crecen las cosas? ¿cómo *adquiere forma* la materia? (el término "morfogénesis" proviene de las palabras "forma" y "origen", en griego antiguo). Por mera coincidencia, en Cambridge, Francis Crick y James Dewey Watson estaban intentando resolver el mismo problema desde una perspectiva microbiológica. Durante el proceso acabaron descubriendo la doble hélice del ADN. Sin embargo, Turing enfocaba el problema desde una perspectiva matemática. ¿Cómo llegó a evolucionar el caldo primigenio, con su simpleza química relativa hasta convertirse en organismos de una complejidad tan enorme? Crick y Watson pretendían encontrar una explicación sobre el *cómo* de este hecho. Turing buscaba una respuesta para el *cómo y* el *porqué.* Perseguía una respuesta matemática que pudiera explicar el patrón de la vida misma en términos matemáticos. (Si Einstein podía explicar el funcionamiento último del Universo mediante fórmulas matemáticas, Turing podía describir la vida misma de

igual forma. Desde luego, a Turing no le faltaba ambición.)

¿Cómo contenía el caldo químico primigenio la información que le permitía desarrollar esa complejidad? (aquí resulta evidente el paralelismo con la pregunta de cómo una computadora podía desarrollar inteligencia). Pero, ¿qué tenían que ver estos problemas con las matemáticas? Varios ejemplos lo muestran. Tomemos, por ejemplo, una solución química inorgánica saturada, en la que se están formando cristales —o *creciendo,* porque parecen desarrollarse de una forma "orgánica" asimétrica y misteriosa. En el plano químico, no hay explicación para esta falta de simetría. Sin embargo, en el plano molecular, los movimientos y las colisiones individuales de las moléculas que hay en la solución son aleatorios. Por lo tanto, resulta poco sorprendente que los cristales adquieran formas asimétricas. En cierto modo, la complejidad se *va creando* a medida que se produce.

Un ejemplo significativo de este proceso se aprecia en la música moderna. El compositor húngaro György Ligeti ha "escrito" una pieza para 100 metrónomos, todos fijados a diferentes velocidades. Los metrónomos empiezan a funcionar al mismo tiempo, y después pierden la sincronía. Esto suena como una receta para el caos, pero lo que en realidad se desarrolla es una extraña "música virtual", que en cierto modo está *siendo creada por los propios metrónomos.*

Turing estaba convencido de que en la naturaleza se producían desarrollos matemáticos similares. Las flores, plantas y células que estudió mostraban y desarrollaban, sin excepción, unos patrones, muchos de los cuales presentaban sorprendentes secuencias matemáticas.

Por ejemplo, tanto las espirales de una piña como las semillas cubiertas de un girasol recuerdan la serie Fibonacci. Ésta es la serie 1, 1, 2, 3, 5, 8, 13, 21…, en que cada número es la suma de los dos anteriores. Las misteriosas y fascinantes propiedades de los números de Fibonacci se reflejan en las matemáticas (por ejemplo, en los triángulos de Pitágoras, los números primos y la proporción áurea) y en la naturaleza (por ejemplo, las piñas, el crecimiento de las hojas y las distancias de los planetas respecto al Sol muestran características de la serie de Fibonacci).

Los patrones de la naturaleza eran profundamente matemáticos. ¿Era posible que algo en la naturaleza de las *matemáticas* controlara el desarrollo de tal complejidad?

Ésas eran las preguntas que ocupaban a Turing a principios de los años de 1950. Siguió utilizando el MADAM en estas complejas investigaciones, aunque había sido relegado en buena medida de sus tareas sobre el desarrollo práctico del lado informático. Se esperaba que el MADAM moderno hiciera algo más que escribir cartas de amor.

Para entonces, Turing había comprado una casa en Wilmslow, un agradable barrio residencial en la periferia de Manchester y, aparentemente, era una figura de cierta eminencia. En 1951 había sido elegido miembro de la Royal Society, a la temprana edad de 39 años. Una de las personas que lo habían propuesto era el filósofo Bertrand Russell, uno de los primeros que reconoció el profundo significado *filosófico* del trabajo de Turing. (De hecho, en este sentido queda aún mucho por explorar más de medio siglo después.)

Sin embargo, en el caso de Turing la respetabilidad había sido siempre apenas una fina cubierta. Seguía trabajando durante largas horas (solía trabajar jornadas de 12 horas en el laboratorio), y "reservaba" el MADAM para utilizarlo las noches de los martes y jueves. Aun así le quedaban otras noches, largas y solitarias, en que el MADAM no podía distraerlo, por lo que de vez en cuando merodeaba en busca de jóvenes homosexuales.

Turing entabló algo parecido a una amistad con uno de sus amantes, Arnold Murray, un joven rubio y de ojos azules de Manchester. Un fin de semana Alan dejó a Arnold solo en su casa y al volver descubrió que le habían robado. Faltaban algunas cosas menudas, incluyendo una camisa y unos pares de zapatos, algunos cuchillos de plata y una brújula. Turing se sintió herido y comunicó el robo a la policía.

Esto acabó siendo un error fatal. El detective a cargo de la investigación pronto descubrió el elemento homosexual del caso y en febrero de 1952, Turing fue detenido, por los cargos de "indecencia grave".

Turing tenía un carácter firme, pero la deshonra pública le afectó, inevitablemente. Se vio obligado a ir al sur, a advertir a su madre del juicio que se avecinaba y de que era posible que se hiciera público. Hodges, su biógrafo, afirma: "La señora Turing no entendía muy bien la importancia de lo ocurrido, pero entendía lo suficiente para que tuviera con su hijo una desagradable discusión, en la que no acababan de entenderse."

Afortunadamente, no se hizo una gran publicidad del juicio en los periódicos (hubo una breve en la edición de la zona norte de *News of the World,* con el titular "El acusado tenía un gran cerebro"). El caso, casi con certeza, fue acallado por las autoridades. Tal vez

esto fuera lo mínimo que podían hacer por un hombre que había desempeñado un papel fundamental para la victoria en la segunda guerra mundial. Cabe pensar que si Turing hubiera sido un hombre más atractivo y hubiera querido participar en el juego social, todo el asunto habría quedado olvidado. Después de todo, no se puede decir que la homosexualidad no fuera común en el *establishment* británico. Pero claro, Turing no era en absoluto un miembro del *establishment*.

Al final, Turing se declaró culpable y tuvo suerte, ya que no fue condenado a prisión. En lugar de eso quedó en libertad condicional, con la condición de que se sometiera a un tratamiento hormonal para "curarlo" de su homosexualidad.

Este absurdo tratamiento con drogas tenía efectos secundarios grotescos. Para empezar, lo volvió impotente, y a un compañero al que visitó en Cambridge le confesó: "Me están creciendo pechos."

Turing intentó volcarse nuevamente en su trabajo. Ahora intentaba resolver las preguntas que había planteado en *La base química de la morfogénesis*. Sin embargo, encontró un obstáculo en las pequeñas variaciones de los sistemas de ecuaciones diferenciales de primer orden que parecían originar la asimetría. Aparentemente, éstas explicaban la teoría química de la morfogénesis cuando la complejidad se creaba a sí misma. Evidentemente, que la complejidad se creara a sí misma era un asunto complejo.

Turing pronto descubrió que esto era tan desalentador como su investigación doctoral sobre los números primos en relación con la hipótesis de Riemann. Al igual que antes, las primeras fuentes de inspiración se secaron, convirtiéndose en un desierto de cálculos. Y al igual que antes, reapareció la posibilidad del suicidio.

Esta vez, la idea resultaba más atractiva. Su trabajo se había secado y había sido excluido de toda labor creativa con las computadoras, a pesar de sus excelentes cualificaciones. Su identidad sexual estaba virtualmente anulada, y el excepcional tono físico que había mantenido como atleta de fondo había quedado convertido en una mera caricatura gracias a los medicamentos.

Así llegó la última representación de una escena que había ensayado al menos en una oportunidad. En la noche del 7 de junio de 1954, Alan Turing se tumbó y comió su habitual manzana nocturna, tratada en esta ocasión con cianuro.

EPÍLOGO

Tras su muerte, Turing fue condenado al olvido. Su trabajo en el *Colossus* durante la guerra siguió siendo secreto oficial y su exclusión final del trabajo creativo práctico en las primeras computadoras británicas hizo que los vencedores se quedaran con el botín de guerra, y sólo los *cognoscenti* de la materia supieron apreciar el brillante trabajo teórico que Turing realizó.

Y así podían haber quedado las cosas, si en 1985 Andrew Hodges no hubiera escrito una brillante y completa biografía de Turing. Ésta le brindó a Turing el reconocimiento general del público que se merecía, además de desvelar un infame escándalo sexual (en este caso, perpetrado por unas autoridades ingratas). Aquéllos que han escrito posteriormente sobre Turing tienen una gran deuda con Hodges. Sin embargo, y a pesar de las investigaciones exhaustivas, Turing siguió siendo para Hodges un misterio, como lo fuera para sus contemporáneos. A pesar de ello, los logros de Turing hablan por sí solos. Cada vez más personas lo reconocen como *el* pionero de la teoría informática, un padre fundador de la computadora moderna y, de forma casi accidental, el hombre que ganó la guerra.

Las cuestiones de la inteligencia artificial y la morfogénesis, que él fuera el primero en plantear en un sentido amplio, siguen siendo preguntas fundamentales y sin respuesta en nuestros días.

Enciclopedia de conocimientos fundamentales UNAM-Siglo XXI.
Volumen 5: Matemáticas, Física y *Computación,*
editada por la Universidad Nacional Autónoma de México y Siglo XXI Editores,
se terminó de imprimir el 23 de noviembre de 2010,
en los talleres de Compañía Editorial Ultra, S.A. de C.V.,
ubicados en Centeno 162, local 2, colonia Granjas Esmeralda,
09810, México, D. F.
El tiraje consta de 40 000 ejemplares.
Los interiores fueron impresos en papel bond de 90 g
y los forros en papel couché mate de 150 g sobre cartoné.
Para su composición se utilizaron las fuentes Minion Display 10.5/13.5, Futura 9/13.5.