

¿Por qué las ecuaciones?

Para evitar la repetición tediosa de estas palabras: es igual a: fijaré, como hago con frecuencia en el transcurso de mi trabajo, un par de paralelas o líneas gemelas de longitud uno:

=

porque no hay dos cosas que puedan ser más iguales.

*ROBERT RECORDE. The Whetstone of Witte,
1557*

Las ecuaciones son el alma de las matemáticas, la ciencia y la tecnología. Sin ellas, nuestro mundo no existiría en su forma actual.

Sin embargo, las ecuaciones tienen una reputación de ser aterradoras: los editores de Stephen Hawking le dijeron que cada ecuación sería reducir a la mitad las ventas de *Una breve historia del tiempo*, pero luego ignorando su propio consejo y se le permitió incluir $E = mc^2$ cuando supuestamente tenían vendido otros 10 millones de copias.

Yo estoy del lado de Hawking. Las ecuaciones son demasiado importantes para estar escondidas. Pero sus editores tuvieron un punto fuerte: las ecuaciones son formales y austeras, se ven complicadas, e incluso aquellos como nosotros que sentimos amor por ellas, podemos ser confundidos si somos bombardeados con ellas.

En este libro, tengo una excusa. Puesto que se trata de ecuaciones, no se puede evitar más su inclusión que como si yo escribiese un libro sobre montañismo sin usar la palabra "montaña". Quiero convencerte que las ecuaciones han desempeñado un papel vital en la creación del mundo de hoy, a partir de la cartografía a la navegación vía satélite, desde la música a la televisión, desde el descubrimiento de América a la exploración de las lunas de Júpiter.

Afortunadamente, usted no necesita ser un genio para apreciar la poesía y la belleza de una buena y significativa ecuación.

Hay dos tipos de ecuaciones en matemáticas que a primera vista tienen un aspecto muy similar. Un tipo presenta las relaciones entre las diversas cantidades matemáticas: la tarea es demostrar que la ecuación es verdadera. El otra clase proporciona información sobre una cantidad desconocida, y la tarea del matemático es *resolverla*; es hacer conocido, lo desconocido. El distinción no es clara, porque a veces la misma ecuación puede ser utilizada en ambos sentidos, pero es una guía útil. Va a encontrar los dos tipos aquí.

Las ecuaciones en las matemáticas puras son generalmente de la primera clase: revelan patrones y regularidades profundas y hermosas. Ellas son válidas porque, dado nuestros supuestos básicos acerca de la estructura lógica de las matemáticas, no hay otra alternativa. El teorema de Pitágoras, que es una ecuación expresado en el lenguaje de la geometría, es un ejemplo. Si acepta los Supuestos básicos de Euclides sobre geometría, entonces el teorema de Pitágoras es verdadero.

Las ecuaciones en matemáticas aplicadas y física matemática son por lo general de la segunda clase. Ellas codifican información sobre el verdadero mundo; expresan propiedades del universo que en principio, podría haber sido muy diferentes. La ley de la gravedad de Newton es un buen ejemplo. Nos dice cómo la fuerza de atracción entre dos cuerpos depende de sus masas, y lo lejos que están. Resolviendo las ecuaciones resultantes nos dice cómo los planetas giran alrededor del Sol, o cómo diseñar una trayectoria de una sonda espacial, pero la Ley de Newton no es un teorema matemático; es cierto por razones físicas, se encaja observaciones. La ley de la gravedad podría haber sido diferente. De hecho, es diferente: la teoría general de la relatividad de Einstein mejora la de Newton por ajustar mejor algunas observaciones, aunque sin estropear aquellos casos en los que ya se sabe que la ley de Newton hace un buen trabajo.

El curso de la historia humana se ha redirigido, una y otra vez, por una ecuación. Ecuaciones tienen poderes ocultos. Revelan los más internos secretos de la naturaleza. Esta no es la forma tradicional de los historiadores para organizar el auge y caída de las civilizaciones. Reyes, reinas, guerras y desastres naturales abundan en los libros de historia, pero las ecuaciones ocupan una capa finísima.

Esto es injusto. En la época victoriana, Michael Faraday demostró las conexiones entre el magnetismo y la electricidad en las audiencias de la Royal Institution de Londres. Al parecer, el primer ministro William Gladstone preguntó si nada de consecuencia práctica vendría de él. Se dice (sobre la base de muy poca evidencia real, pero ¿por qué arruinar una buena historia?) que Faraday respondió: "Sí, señor. Un día va a cobrar impuestos sobre ella". Si él dijo eso, él tenía razón. James Clerk Maxwell transformó observaciones experimentales iniciales y leyes empíricas sobre el magnetismo y la electricidad en un sistema de ecuaciones para el electromagnetismo. Entre las muchas consecuencias fueron la radio, radar, y la televisión.

Una ecuación deriva su poder de una fuente simple. Nos dice que dos cálculos, que parecen diferentes, tienen la misma respuesta. La clave-símbolo es el signo de igualdad, $=$. Los orígenes de la mayoría de los símbolos matemáticos se han perdido en las nieblas de la antigüedad, o son tan recientes que no hay duda de donde vinieron. El signo igual es inusual, ya que se remonta más de 450 años, sin embargo, no sólo sabemos que lo inventó, incluso sabemos por qué. El inventor fue Robert Recorde, en 1557, en *The Whetstone of Witte* (La piedra de amolar de Witte).

Utilizó dos líneas paralelas (usando la palabra obsoleta *gemowe*, que significa "gemelo") para evitar la tediosa repetición de las palabras 'es igual a'. Él eligió ese símbolo porque "*no hay dos cosas puede ser más iguales*". Recorde eligió bien. Su símbolo se ha mantenido en uso durante 450 años.

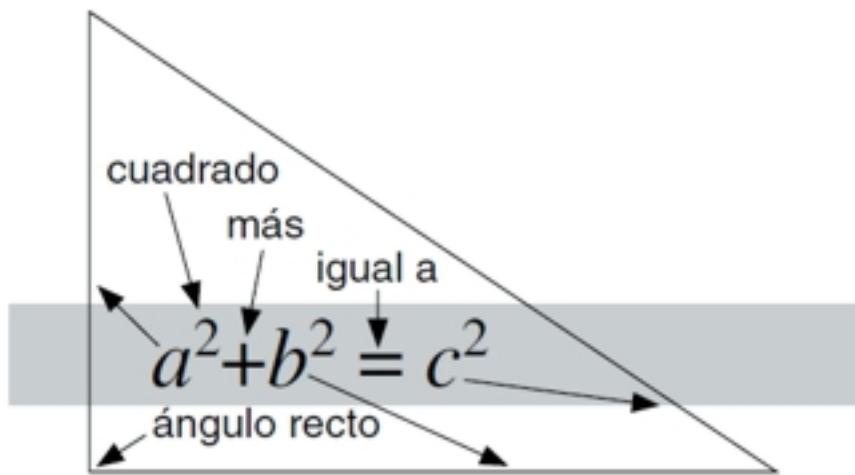
El poder de ecuaciones se encuentra en la filosóficamente difícil correspondencia entre las matemáticas, una creación colectiva de las mentes humanas, y una realidad física externa. Moldean patrones profundos en el mundo exterior.

Aprendiendo a dar valor a las ecuaciones, y a leer las historias, podemos descubrir trazos vitales del mundo a nuestro alrededor. En principio, puede haber otras formas de lograr el mismo resultado. Muchas personas prefieren palabras a los símbolos; idioma que también nos da poder sobre nuestro entorno, pero el veredicto de la ciencia y la tecnología, es que las palabras son demasiado imprecisas y demasiado limitadas, para proporcionar una vía eficaz para los aspectos más profundos de la realidad.

Están demasiado coloreadas por supuestos a nivel humano. Las palabras por sí solas no pueden proporcionar los conocimientos esenciales.

Las ecuaciones pueden. Ellas han sido una fuerza motriz en la civilización humana durante miles de años. A lo largo de la historia, las ecuaciones han estado tirando las cuerdas de la sociedad. Escondido detrás de las escenas, sin duda - pero la influencia fue allí, si se observó o no. Esta es la historia del ascenso de la humanidad, dijo a través de 17 ecuaciones.

Capítulo 1
La hipotenusa al cuadrado
Teorema de Pitágoras



¿Qué nos dice?

Como están relacionados los tres lados de un triángulo rectángulo.

¿Por qué es importante?

Nos proporciona un vínculo importante entre la geometría y el álgebra, permitiéndonos calcular distancias en términos de coordenadas. También inspiró la trigonometría.

¿Qué provocó?

Topografía, navegación y, más recientemente, relatividad general y especial, la mejor de las actuales teorías del espacio, el tiempo y la gravedad.

Pregunta a cualquier estudiante el nombre de un matemático famoso y, asumiendo que pueden pensar en uno, la mayoría de las veces optarán por Pitágoras. Si no, Arquímedes se les vendrá a la mente. Incluso el ilustre Isaac Newton desempeña el tercer papel tras estas dos superestrellas del mundo antiguo. Arquímedes fue una lumbrera y Pitágoras probablemente no lo fuera, pero se merece más crédito del

que con frecuencia se le da. No por lo que logró, sino por lo que puso en marcha. Pitágoras nació en la isla griega de Samos, en el Egeo oriental, alrededor del 570 a.C. Fue un filósofo y un geómetra. Lo poco que conocemos sobre su vida proviene de escritores bastante posteriores y su exactitud histórica es cuestionable, pero los eventos clave son probablemente correctos. Alrededor del 530 a.C. se mudó a Crotona, una colonia griega en lo que ahora es Italia. Ahí, fundó una secta filosófico-religiosa, los Pitagóricos, quienes creían que el universo estaba basado en los números. La fama actual de su fundador recae sobre el teorema que lleva su nombre. Se ha enseñado durante más de 2.000 años y ha pasado a formar parte de la cultura popular. La película de 1958 *Loco por el circo*, protagonizada por Danny Kaye, incluye una canción cuya letra en versión original dice:

*The square on the hypotenuse
of a right triangle
is equal to
the sum of the squares on the two adjacent sides.*

*«El cuadrado de la hipotenusa
de un triángulo rectángulo
es igual a
la suma de los cuadrados
de los dos catetos».*

(La versión doblada de la película hace una traducción libre y no recita el enunciado del teorema de Pitágoras. (N. de la t.))

La canción original continúa con algún doble sentido sobre no permitir a tu participio oscilar y asocia a Einstein, Newton y los hermanos Wright con el famoso teorema. Los dos primeros exclaman «¡Eureka!», pero no, ese fue Arquímedes. Deducirás que las letras no son muy buenas en lo que a rigor histórico se refiere, pero eso es Hollywood. Sin embargo, en el capítulo 13 veremos que el letrista Johnny Mercer fue muy certero con Einstein, probablemente más de lo que era consciente.

El teorema de Pitágoras aparece en un chiste muy conocido en inglés, un juego de

palabras muy tonto sobre una india (en inglés *squaw*, que al pronunciarlo suena muy parecido a *square*, cuadrado) y un hipopótamo (*hippopotamus*, que al pronunciarlo suena parecido a *hypotenuse*, hipotenusa). El chiste puede encontrarse en Internet sin problema, basta poner «*squaw on the hippopotamus*», pero es mucho más difícil descubrir de dónde proviene.¹ Hay viñetas sobre Pitágoras, camisetas y un sello griego (figura 1).

A pesar de todo este alboroto, no sabemos si realmente Pitágoras probó su teorema. Es más, no sabemos si en realidad es su teorema. Bien podría haber sido descubierto por uno de los acólitos de Pitágoras o algún escriba de Babilonia o Sumeria. Pero Pitágoras obtuvo crédito por ello y su nombre se asoció a él. Cualquiera que sea su origen, el teorema y sus consecuencias han tenido una

repercusión enorme en la historia de la humanidad. Literalmente, abrió la puerta a nuestro mundo.

Los griegos no expresaron el teorema de Pitágoras como una ecuación en el sentido simbólico moderno. Eso vino más tarde, con el desarrollo del álgebra. En la Antigüedad, el teorema se expresaba verbalmente y geométricamente. Alcanzó su forma más elegante, y su primera demostración registrada, en los escritos de Euclides de Alejandría. Alrededor del 250 a.C., Euclides se convirtió en el primer matemático moderno cuando escribió su famoso *Elementos*, el libro de texto de matemáticas más influyente de todos los tiempos. Euclides



Figura 1. Sello griego del teorema de Pitágoras.

¹ *The Penguin Book of Curious and Interesting Mathematics* de David Wells cita una forma breve del chiste. Un jefe indio tiene tres esposas que se preparan para dar a luz, una sobre la piel de un búfalo, otra sobre la piel de un oso y la otra sobre la piel de un hipopótamo. A su debido tiempo, la primera tuvo un niño, la segunda una niña y la tercera gemelos, un niño y una niña, de este modo ilustra el famoso teorema de que «*the squaw on the hippopotamus is equal to the sum of the squaws on the other two hides*» (la india del hipopótamo es igual a la suma de las indias de las otras pieles, el enunciado del teorema en inglés y el chiste suena muy parecido en inglés). El chiste se remonta al menos hasta mediados de la década de los cincuenta del siglo pasado cuando fue contado en un programa de radio de la BBC, «*My Word*», presentado por los guionistas de comedia Frank Muir y Denis Norden.

convirtió la geometría en lógica haciendo explícitos sus supuestos básicos y apelando a ellos para dar pruebas sistemáticas de todos sus teoremas. Construyó una torre conceptual cuyos fundamentos eran puntos, rectas y círculos y cuyo pináculo fue la existencia de exactamente cinco sólidos regulares.

Una de las joyas de la corona de Euclides fue lo que ahora nosotros llamamos teorema de Pitágoras, la proposición 47 del libro I de los *Elementos*. En la famosa traducción de Sir Thomas Heath esta proposición dice: «*En triángulos rectángulos, el cuadrado del lado subtendiente al ángulo recto es igual a los cuadrados de los lados adyacentes al ángulo recto*».

Por lo tanto, nada de hipopótamos. Nada de hipotenusa. Ni siquiera un explícito «suma» o «adición». Tan solo la palabra rara «*subtendiente*», que básicamente significa «ser opuesto a». Sin embargo, el teorema de Pitágoras claramente expresa una ecuación, porque contiene esa palabra fundamental: igual.

Para las matemáticas avanzadas, los griegos trabajaban con rectas y áreas en vez de con números. De modo que Pitágoras y sus sucesores griegos habrían decodificado el teorema como una igualdad de áreas: «el área de un cuadrado construido usando el lado más largo de un triángulo rectángulo es la suma de las áreas de los cuadrados construidos a partir de los otros dos lados». El lado más largo es la famosa hipotenusa, que significa «extender debajo», lo cual sucede si haces el dibujo con la orientación apropiada, como en la figura 2 (parte izquierda).

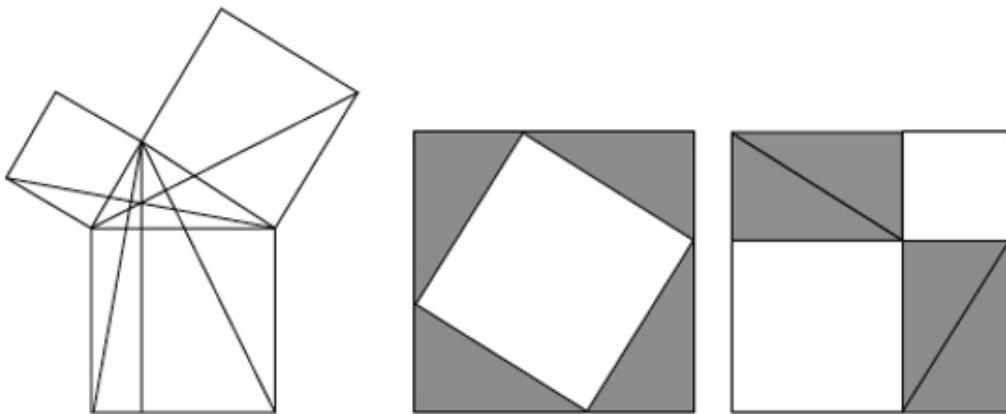


Figura 2. A la izquierda: construcción para la prueba de Euclides del teorema de Pitágoras. En el centro y a la derecha: prueba alternativa al teorema. Los cuadrados exteriores tienen áreas iguales y los triángulos sombreados tienen todos áreas

iguales. Por lo tanto, el cuadrado inclinado tiene la misma área que los otros dos cuadrados blancos juntos.

En apenas 2.000 años, el teorema de Pitágoras ha sido reformulado en la forma de la ecuación algebraica:

$$a^2 + b^2 = c^2$$

donde c es la longitud de la hipotenusa y a y b las longitudes de los otros dos lados y el pequeño 2 elevado significa «al cuadrado». Algebraicamente, el cuadrado de cualquier número es ese número multiplicado por sí mismo, y todos sabemos que el área de cualquier cuadrado es el cuadrado de la longitud de su lado. De modo que la ecuación de Pitágoras, como la renombraré, dice lo mismo que Euclides dijo, excepto por el diverso bagaje psicológico consecuencia de cómo en la Antigüedad entendían conceptos matemáticos, como números y áreas, y en lo que no voy a entrar.

La ecuación de Pitágoras tiene muchos usos e implicaciones. De manera casi inmediata, nos permite calcular la longitud de la hipotenusa dados los otros dos lados. Por ejemplo, supongamos que $a = 3$ y $b = 4$. Entonces

$$c^2 = a^2 + b^2 = 3^2 + 4^2 = 9 + 16 = 25.$$

Por lo tanto, $c = 5$. Este es el famoso triángulo 3-4-5, omnipresente en las clases de matemáticas de la escuela, y el ejemplo más simple de una terna pitagórica: un conjunto de tres números enteros que cumplen la ecuación de Pitágoras. El siguiente ejemplo más sencillo, más que versiones a escala como 6-8-10, es la terna 5-12-13. Hay infinidad de este tipo de ternas, y los griegos sabían cómo construirlas todas. Las ternas todavía conservan cierto interés en la teoría de números, incluso en la última década se han descubierto nuevas características.

En vez de usar a y b para calcular c , se puede proceder de manera indirecta y resolver la ecuación para obtener a , siempre y cuando se conozcan b y c . También se puede responder a preguntas más sutiles, como veremos a continuación.

¿Por qué es el teorema cierto? La prueba de Euclides es bastante complicada e incluye dibujar cinco líneas extra en el diagrama (figura 2, a la izquierda) y recurrir a varios teoremas anteriores probados. Los alumnos de la época victoriana (había pocas alumnas que estudiaseen geometría en aquel entonces) se referían a él con irreverencia, lo llamaban los calzones de Pitágoras. Una prueba sencilla e intuitiva, aunque no la más elegante, usa cuatro copias del triángulo para relacionar dos soluciones del mismo puzzle matemático (figura 2 a la derecha). El dibujo es de por sí convincente, pero completar los detalles lógicos requiere algo más de consideración. Por ejemplo, ¿cómo sabemos que la región blanca inclinada en el medio del dibujo es un cuadrado?

Hay evidencias tentadoras que indican que el teorema de Pitágoras era conocido mucho antes que Pitágoras. Una tabla de arcilla de Babilonia² del Museo Británico contiene, en escritura cuneiforme, un problema matemático y su respuesta, que puede ser parafraseada como:

4 es la longitud y 5 la diagonal. ¿Cuál es el ancho?

4 veces 4 es 16

5 veces 5 es 25

Quita 16 de 25 para obtener 9

¿Cuántas veces qué debo tomar para obtener 9?

3 veces 3 es 9

Por lo tanto 3 es el ancho

De modo que en Babilonia ciertamente conocían el triángulo 3-4-5, mil años antes de Pitágoras.

Otra tabla, YBC 7289, de la colección babilónica de la Universidad de Yale, es la que se muestra en la figura 3 (izquierda).

Muestra un diagrama de un cuadrado de lado 30, cuya diagonal está marcada con dos listas de números: 1, 24, 51, 10 y 42, 25, 35. En Babilonia usaban la notación de base 60 para los números, así la primera lista realmente se refiere a

$$1 + \frac{24}{60} + \frac{51}{60^2} + \frac{10}{60^3},$$

² Citado sin referencia en http://www-history.mcs.st-and.ac.uk/HistTopics/Babylonian_Pythagoras.html

que en notación decimal es 1,4142129. La raíz cuadrada de 2 es 1,4142135. La segunda lista es la primera multiplicada por 30. Por lo tanto los babilonios sabían que la diagonal de un cuadrado es su lado multiplicado por la raíz cuadrada de 2. Puesto que $1^2 + 1^2 = 2 = (\sqrt{2})^2$, esto también es un caso del teorema de Pitágoras.



Figura 3. A la izquierda: YBC 7289. A la derecha: Plimpton 322.

Es incluso más extraordinaria, aunque más enigmática, la tabla Plimpton 322 de la colección George Arthur Plimpton de la Universidad de Columbia (figura 3 a la derecha). Es una tabla de números, con cuatro columnas y quince filas. La columna final tan solo enumera el número de filas, de la 1 a la 15. En 1945, los historiadores de ciencia Otto Neugebauer y Abraham Sachs³ se dieron cuenta de que en cada fila el cuadrado del número en la tercera columna, llamémosle c , menos el cuadrado del número en la segunda columna, llamémosle b , era en sí mismo un cuadrado, llamémosle a . De esto se deduce que $a^2 + b^2 = c^2$, de modo que la tabla parece registrar temáticas pitagóricas. Al menos este es el caso, siempre y cuando cuatro errores evidentes se corrijan. Sin embargo, no está totalmente claro que Plimpton 322 tenga algo que ver con las ternas pitagóricas, e incluso si tiene que ver, podría solo haber sido una lista práctica de triángulos cuyas áreas son fáciles de calcular. Estos podrían agruparse para dar buenas aproximaciones a otros triángulos y otras

³ A. Sachs, A. Goetze e O. Neugebauer. *Mathematical Cuneiform Texts*, American Oriental Society, New Haven, 1945.

formas, quizá para la medición de tierras.

Otra icónica civilización de la Antigüedad es Egipto. Existen algunas evidencias de que Pitágoras podría haber visitado Egipto siendo joven y algunas de ellas conjeturan que fue entonces cuando aprendió su teorema. Los registros que sobreviven de las matemáticas egipcias ofrecen escaso soporte a esta idea, pero son pocos y especializados. Con frecuencia se afirma, normalmente en el contexto de las pirámides, que los egipcios diseñaron ángulos rectos usando un triángulo 3-4-5, formado por una cuerda con nudos en 12 intervalos iguales y los arqueólogos han encontrado cuerdas de ese tipo. Sin embargo, la afirmación no tiene mucho sentido. Dicha técnica no sería muy fiable, porque las cuerdas se pueden distorsionar y los nudos no tendrían una separación muy precisa. La precisión con la que se construyeron las pirámides de Guiza es superior a cualquiera que se pudiera haber logrado con una cuerda. Se han encontrado herramientas mucho más prácticas, parecidas a la escuadra de carpintero. Los egiptólogos especializados en las matemáticas del antiguo Egipto no tienen conocimiento registrado de cuerdas empleadas para formar triángulos del tipo 3-4-5 y ningún ejemplo de que dichas cuerdas existan. Así que esta historia, por muy bonita que pueda parecer, es casi con certeza un mito.

Si Pitágoras pudiese ser trasplantado a nuestro mundo actual, notaría muchas diferencias. En su época, el conocimiento médico era rudimentario, la luz provenía de velas y antorchas ardiendo, y la forma más rápida de comunicación era un mensajero a caballo o un faro encendido en la cima de una colina. El mundo conocido abarcaba la mayoría de Europa, Asia y África, pero no América, Australia, el Ártico o la Antártida. Muchas culturas consideraban que el mundo era plano, un disco circular o incluso un cuadrado alineado con los cuatro puntos cardinales. A pesar de los descubrimientos de la Grecia clásica, esta creencia estaba todavía muy extendida en la época medieval, en la forma de mapas *orbis terrae*, figura 4.

¿Quién fue el primero en darse cuenta de que la Tierra era redonda? Según Diógenes Laercio, un biógrafo griego del siglo m, fue Pitágoras. En su libro *Vidas, opiniones y sentencias de los filósofos más ilustres*, una colección de dichos y notas biográficas que es una de nuestras principales fuentes históricas para la vida privada de los filósofos de la antigua Grecia, escribió: «Pitágoras fue el primero que

dijo que la Tierra era redonda, aunque Teofrasto se lo atribuye a Parménides y Zenón a Hesíodo».

Los griegos en la Antigüedad con frecuencia reclaman que los mayores descubrimientos han sido hecho por sus famosos antepasados, con independencia del hecho histórico, de modo que no podemos tomarnos la afirmación en serio, pero lo que no se discute es que a partir del siglo V a.C. todos los filósofos y matemáticos griegos con reputación consideraban que la Tierra era redonda.



Figura 4. Mapa del mundo hecho alrededor del año 110 por el cartógrafo marroquí al-Idrisi para el rey Roger de Sicilia.

La idea sí que parece que se originó en tomo a la época de Pitágoras y quizá proviniese de uno de sus seguidores. O podría tratarse de un hecho habitual asumido, basado en la evidencia de la sombra redondeada de la Tierra en la Luna durante un eclipse, o en la analogía con la Luna, obviamente, redonda.

En cualquier caso, incluso para los griegos, la Tierra era el centro del universo y todo lo demás giraba en torno a ella. La navegación se llevaba a cabo gracias a cálculos en desuso: mirando las estrellas y siguiendo la línea de la costa. La ecuación de Pitágoras cambió todo eso. Puso a la humanidad en la senda para la comprensión actual de la geografía de nuestro planeta y su lugar en el Sistema

Solar. Fue un primer paso vital hacia las técnicas geométricas necesarias para la cartografía, la navegación y la topografía. También proporcionó la llave a una relación vital entre la geometría y el álgebra. Esta línea de desarrollo nos lleva de la Antigüedad directamente a la relatividad general y la cosmología moderna (véase el capítulo 13). La ecuación de Pitágoras abrió por completo nuevas direcciones para la exploración humana, tanto metafóricamente como literalmente. Reveló la forma de nuestro mundo y su lugar en el universo.

Muchos de los triángulos que nos encontramos en la vida real no son rectángulos, de manera que las aplicaciones directas de la ecuación podrían parecer limitadas. Sin embargo, cualquier triángulo puede dividirse en dos triángulos rectángulos, como vemos en la figura 6, y cualquier forma poligonal se puede dividir en triángulos. Así que los triángulos rectángulos son la clave, prueban que hay una relación útil entre la forma de un triángulo y la longitud de sus lados. La materia que se desarrolló a partir de esta visión es la trigonometría, que significa «medición de triángulos».

El triángulo rectángulo es fundamental en trigonometría, en particular determina las funciones básicas de la trigonometría: seno, coseno y tangente. Los nombres son de origen árabe y la historia de estas funciones y sus muchas predecesoras muestra el complicado camino por el cual surgió la versión actual del tema. No daré muchas vueltas y explicaré el resultado final. Un triángulo rectángulo tiene, obviamente, un ángulo recto, pero sus otros dos ángulos son arbitrarios, exceptuando el hecho de que suman 90° . Asociadas con cualquier ángulo hay tres funciones, esto es, reglas para calcular un número asociado a él. Para el ángulo marcado como A en la figura 5, usando la notación tradicional a, b, c para los tres lados, definimos el seno (*sen*), coseno (*cos*) y tangente (*tg*) como:

$$\text{sen } A = a/c$$

$$\cos A = b/c$$

$$\tan A = a/b$$

Estas cantidades dependen solo del ángulo A , porque todos los triángulos rectángulos con un ángulo A dado son idénticos excepto por la escala.

En consecuencia, es posible construir una tabla de valores para \sin , \cos y \tan para un rango de ángulos, y luego usarlos para averiguar características de los triángulos rectángulos. Una aplicación típica, la cual se remonta a la Antigüedad, es calcular la altura de una columna alta usando solo las medidas hechas en el suelo. Suponemos que, a una distancia de 100 metros, el ángulo que se forma al unir el punto con la cima de la columna es de 22° . Sea $A = 22^\circ$ en la figura 5, de este modo a es la altura de la columna. Entonces, la definición de la función tangente nos dice que:

$$\tan 22^\circ = a/100$$

por lo tanto

$$a = 100 \tan 22^\circ$$

Como la $\tan 22^\circ$ es 0,404, considerando tres cifras decimales, deducimos que $a = 40,4$ metros.

Una vez en posesión de las funciones trigonométricas, es sencillo extender la ecuación de Pitágoras a triángulos que no tiene un ángulo recto. La figura 6 nos muestra un triángulo con un ángulo C y lados a , b , c .

Dividimos el triángulo en dos triángulos rectángulos como se indica. Entonces aplicando dos veces el teorema de Pitágoras y algo de álgebra⁴ se prueba que:

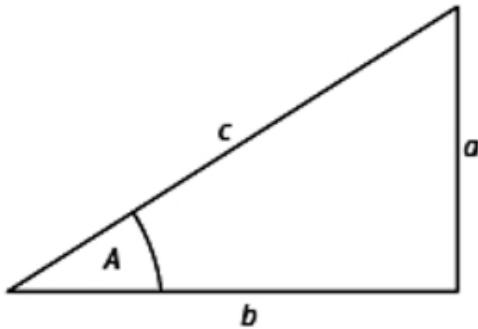


Figura 5. La trigonometría está basada en un triángulo rectángulo.

⁴ La imagen es repetida por conveniencia en la figura 60.

$$a^2 + b^2 - 2 ab \cos C = c^2$$

la cual es parecida a la ecuación de Pitágoras, excepto por el término $- 2 ab \cos C$.

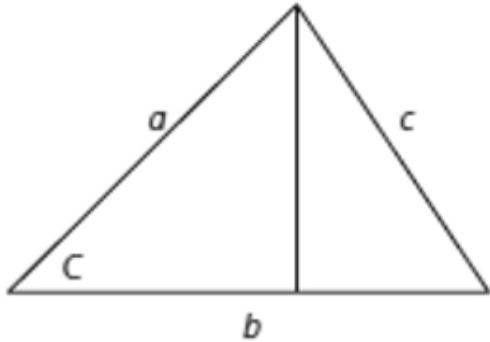


Figura 6. División de un triángulo en dos con ángulos rectos.

Este «teorema del coseno» hace el mismo trabajo que el de Pitágoras, relaciona c con a y b , pero ahora tenemos que incluir información sobre el ángulo C .

El teorema del coseno es uno de los pilares principales de la trigonometría. Si conocemos dos de los lados de un triángulo y el ángulo que se forma entre ellos, podemos usarlo para calcular el tercer lado.

Luego, con otras ecuaciones, calculamos los

ángulos restantes. En última instancia, a todas estas ecuaciones se les puede encontrar el origen en los ángulos rectángulos.

Armados con ecuaciones trigonométricas y aparatos de medida apropiados, podemos tomar mediciones y hacer mapas precisos. Esto no es una idea nueva. Aparece en el papiro de Rhind, una colección de antiguas técnicas matemáticas

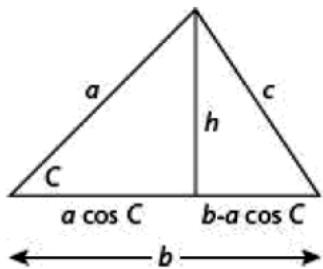


Figura 60. División de un triángulo en dos triángulos rectángulos.

La perpendicular divide el lado en dos partes. Por trigonometría, una mide $a \cos C$ y la otra, $b - a \cos C$. Se h la altura de la perpendicular. Por Pitágoras:

$$a^2 = h^2 + (a \cos C)^2$$

$$c^2 = h^2 + (b - a \cos C)^2$$

O sea,

$$a^2 - h^2 = a^2 \cos^2 C$$

$$c^2 - h^2 = (b - a \cos C)^2 = b^2 - 2ab \cos C + a^2 \cos^2 C$$

Restando la primera ecuación de la segunda, podemos despejar el término h^2 . Lo mismo ocurre con $a^2 \cos^2 C$. Resta entonces:

$$c^2 - a^2 = b^2 - 2ab \cos C$$

que lleva a la fórmula mencionada.

egipcias que datan del 1650 a.C. El filósofo griego Tales usó la geometría de triángulos para estimar la altura de las pirámides de Guiza alrededor del 600 a.C. Herón de Alejandría describió la misma técnica en el 50 d.C. Alrededor del 240 a.C., el matemático griego Eratóstenes calculó el tamaño de la Tierra observando el ángulo del Sol al mediodía en dos sitios diferentes. Alejandría y Siena (en la actualidad Asuán) en Egipto. Una serie de eruditos árabes preservaron y desarrollaron estos métodos aplicándolos, en concreto, a mediciones astronómicas tales como el tamaño de la Tierra.

La topografía empezó a despegar en 1533 cuando el cartógrafo holandés Gemma Frisius explicó cómo usar la trigonometría para elaborar mapas precisos, en *Libellus de Locorum Describendoram Ratione* (Folleto relativo al modo de describir lugares). El método se propagó por toda Europa, llegando a oídos del noble y astrónomo danés Tycho Brahe. En 1579 Tycho lo usó para trazar un mapa preciso de Hven, la isla donde estaba su observatorio. En 1615 el matemático holandés Willebrord Snellius (Snel van Royen) había transformado el método en, esencialmente, su forma moderna: la triangulación. Mediando una longitud inicial con mucho cuidado y muchos ángulos, la posición de las esquinas del triángulo y, por tanto, cualquier característica interesante en ellos, se puede calcular. Snellius calculó la distancia entre dos poblaciones holandesas, Alkmaar y Bergen op Zoom, usando una red de 33 triángulos. Escogió estas dos poblaciones porque se encontraban en el mismo meridiano y estaban separadas exactamente un grado. Sabiendo la distancia entre ellas, podía calcular el tamaño de la Tierra, que publicó en su *Eratosthenes Batavus* (El Eratóstenes holandés) en 1617. Su resultado tiene un margen de error del 4 %. También modificó las ecuaciones de trigonometría para reflejar la naturaleza esférica de la superficie terrestre, un paso importante hacia una navegación eficaz. La triangulación es un método indirecto para calcular distancias usando ángulos. Cuando se mide una franja de tierra, ya sea un edificio o un país, la principal consideración práctica es que es mucho más fácil medir ángulos que medir distancias. La triangulación nos permite medir unas pocas distancias y muchos ángulos, a partir de ahí, todo lo demás se obtiene usando las ecuaciones trigonométricas. El método empieza marcando una línea entre dos puntos, llamada

línea de base, y midiendo su longitud directamente con gran precisión. Después se escoge un punto que destaque en el terreno y sea visible desde los dos extremos de la línea de base. Ahora tenemos un triángulo y conocemos uno de sus lados y dos de sus ángulos, lo cual nos da su forma y tamaño. Podemos usar la trigonometría para calcular los otros dos lados.

A todos los efectos, ahora tenemos dos líneas de base más, los lados del triángulo que acabamos de calcular. A partir de ellas, podemos medir los ángulos a otros puntos más distantes. Continuamos con este proceso para crear una red de triángulos que cubra el área que se está midiendo. Desde cada triángulo observa los ángulos a todas las características significativas: torres de iglesia, cruces de caminos, etcétera. El mismo truco trigonométrico ubica con precisión su localización. Para un giro final, la exactitud de toda la medición puede comprobarse midiendo directamente uno de los lados.

A finales del siglo XVIII, la triangulación se empleaba de manera rutinaria en las mediciones. La Agencia Nacional para el Mapeado de Gran Bretaña empezó en 1783 y tardó setenta años en completar la tarea. El Gran Proyecto de Topografía Trigonométrica de la India, el cual entre otras cosas hizo el mapa del Himalaya y determinó la altura del monte Everest, empezó en 1801. En el siglo XXI, la mayoría de las mediciones a gran escala se hacen usando fotografías de satélites y GPS (el sistema de posicionamiento global). La triangulación explícita ya no se emplea. Pero está todavía ahí, entre bastidores, en los métodos usados para deducir ubicaciones a partir de los datos de los satélites.

El teorema de Pitágoras fue también fundamental para la invención de la geometría analítica. Este es un modo de representar figuras geométricas en términos numéricos, usando un sistema de rectas conocidas como ejes, que se etiquetan con números. La versión más popular es conocida como coordenadas cartesianas en el plano, en honor al matemático y filósofo francés René Descartes, que fue uno de los grandes pioneros en esta área, aunque no el primero. Dibuja dos líneas; una horizontal etiquetada con A y una vertical etiquetada con Y . Estas líneas son conocidas como ejes y se cruzan en un punto llamado origen. Marca puntos en estos dos ejes según su distancia al origen, como las marcas en una regla, los números positivos a la derecha y hacia arriba y los negativos a la izquierda y hacia

abajo. Ahora podemos determinar cualquier punto en el plano en términos de dos números, x e y , sus coordenadas, conectando el punto con los dos ejes como aparece en la figura 7. El par de números (x, y) especifica por completo la ubicación del punto.

Los grandes matemáticos de la Europa del siglo XVII se dieron cuenta de que, en este contexto, una recta o una curva en el plano correspondía a un conjunto de soluciones (x, y) de alguna ecuación en x e y . Por ejemplo, $y = x$ determina una línea diagonal inclinada desde la parte baja izquierda a la parte alta derecha, porque (x, y) está en esa recta si y solo si $y = x$. En general, una ecuación lineal de la forma $ax + by = c$, en la que a , b y c son constantes, se corresponde con una línea recta y viceversa.

¿Qué ecuación se corresponde con una circunferencia? Ahí es donde aparece la ecuación de Pitágoras. Esta implica que la distancia r del origen al punto (x, y) satisface:

$$r^2 = x^2 + y^2$$

y podemos resolverla para r , obteniendo:

$$r = \sqrt{x^2 + y^2}$$

Puesto que el conjunto de todos los puntos que se encuentran a una distancia r del origen es una circunferencia de radio r , cuyo centro es el origen, esa misma ecuación define una circunferencia. De modo más general, a la circunferencia de radio r con centro en (a, b) le corresponde la ecuación:

$$(x - a)^2 + (y - b)^2 = r^2$$

Y la misma ecuación determina la distancia r entre los dos puntos (a, b) y (x, y) . Por lo tanto, el teorema de Pitágoras nos dice dos cosas fundamentales: cuál es la ecuación de una circunferencia y cómo calcular la distancia entre coordenadas.

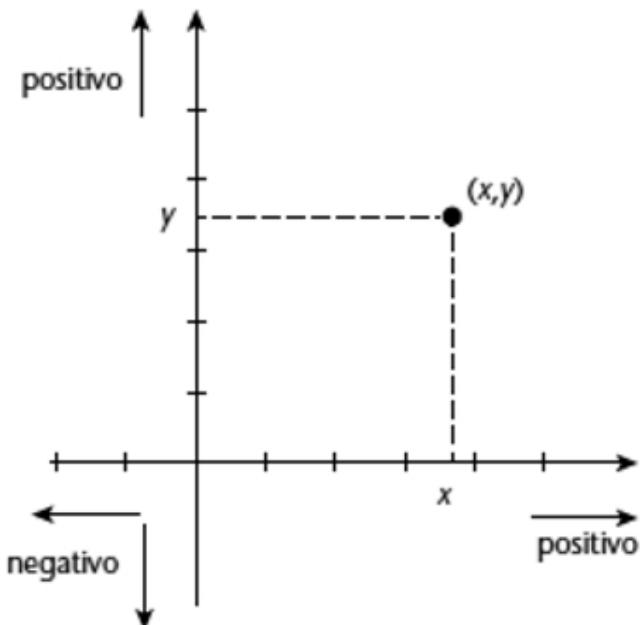


Figura 7. Los dos ejes y las coordenadas de un punto.

Entonces, el teorema de Pitágoras es importante por sí solo, pero ejerce incluso más influencia a través de sus generalizaciones. Aquí continuaré hablando solo de una ramificación de estos desarrollos posteriores para destacar la conexión con la relatividad, a la cual volveremos en el capítulo 13.

La prueba del teorema de Pitágoras en los *Elementos* de Euclides coloca el teorema firmemente en el mundo de la geometría euclidiana. Hubo un tiempo en que esa frase podría ser remplazada por «geometría» sin más, porque se asumía de modo general que la geometría euclidiana era la verdadera geometría del espacio físico. Era obvio. Como la mayoría de las cosas asumidas por ser obvias, resultó ser falsa. Euclides derivó todos sus teoremas de un pequeño número de suposiciones básicas, las cuales clasificó como definiciones, axiomas y nociones comunes. Su sistema era elegante, intuitivo y conciso, con una flagrante excepción, su quinto axioma: «si una recta que corta a otras dos rectas hace los ángulos interiores del mismo lado menores que dos ángulos rectos; las dos rectas, si se alargan indefinidamente, se cortan en el lado en que los ángulos son menores que dos ángulos rectos». Esto es un trabalenguas, la figura 8 puede ser de ayuda.

Durante más de mil años, los matemáticos trataron de arreglar lo que veían como un defecto. No estaban buscando únicamente algo más simple y más intuitivo, que llegase a la misma conclusión, aunque varios de ellos encontraron algo de este tipo. Lo que querían era librarse del axioma extraño por completo probándolo. Después de varios siglos, los matemáticos finalmente se dieron cuenta de que había geometrías alternativas no euclidianas, lo que implicaba que dicha prueba no existía. Estas nuevas geometrías eran tan consistentes lógicamente como la de Euclides, y cumplían todos sus axiomas excepto el axioma de las paralelas.

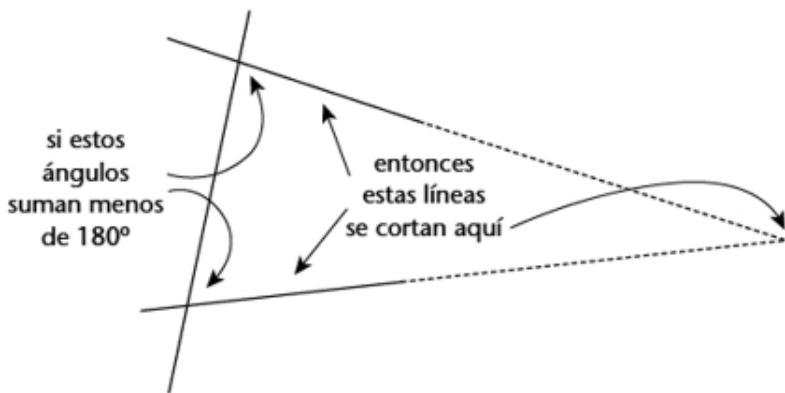


Figura 8. Axioma de las paralelas de Euclides.

Podían ser interpretadas como la geometría de las geodésicas, los caminos más cortos en superficies curvas (figura 9). Esto centró la atención en el significado de curvatura.

El plano de Euclides es plano, curvatura cero. Una esfera tiene la misma curvatura en todos los puntos y es positiva: cerca de cualquier punto parece como una cúpula. (Un sutil matiz técnico: las circunferencias máximas se cortan en dos puntos, no en uno como indica el axioma de Euclides, de modo que la geometría de la esfera se modifica mediante la identificación de puntos antipodales en la esfera, considerando que estos son idénticos. La superficie pasa a ser lo denominado plano proyectivo y la geometría se llama elíptica.)

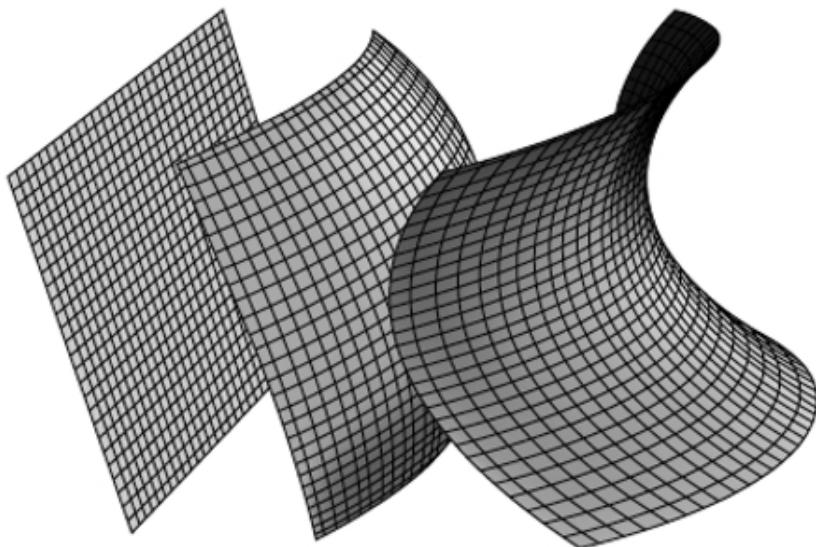


Figura 9. Curvatura de una superficie. A la izquierda: curvatura cero. En el centro: curvatura positiva. A la derecha: curvatura negativa.

También existe una superficie de curvatura constante negativa que en torno a cualquier punto de ella parece una silla de montar. Esta superficie se llama el plano hiperbólico, y puede representarse de varios modos totalmente prosaicos. Quizá el más simple es considerarlo como el interior de un círculo y definir «recta» como un arco de un círculo que se corta con la arista del círculo en un ángulo recto (figura 10).

Puede parecer que, mientras la geometría del plano podría ser no euclidiana, esto es imposible para la geometría del espacio. Puedes curvar una superficie presionándola a una tercera dimensión, pero no puedes curvar un espacio porque no hay espacio para una dimensión extra a la que empujarlo. Sin embargo, esta es una visión un poco simplista. Por ejemplo, podemos modelar un espacio hiperbólico tridimensional usando el interior de una esfera. Las rectas se definen

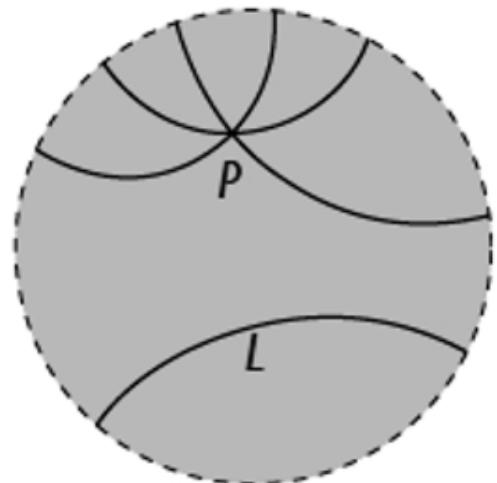


Figura 10. Círculo modelo del plano hiperbólico. Las tres líneas que pasan por P no se cruzan con L.

como arcos de circunferencia que se cortan en el borde haciendo ángulos rectos, y los planos se definen como partes de la esfera que se cortan con el borde en ángulo recto. Esta geometría es tridimensional, satisface todos los axiomas de Euclides excepto el quinto y, en un sentido que puede precisarse, define un espacio curvo tridimensional. Pero no es curvo en tomo a nada o en ninguna nueva dirección.

Tan solo es curvo.

Con todas estas nuevas geometrías disponibles, un nuevo punto de vista empezaba a ocupar el escenario central, pero como visión física, no matemática. Puesto que el espacio no tiene que ser euclíadiano, ¿qué forma tiene? Los científicos se dieron cuenta de que realmente no lo sabían. En 1813, Gauss, que sabía que en un espacio curvo los ángulos de un triángulo no suman 180° , midió los ángulos de un triángulo formado por tres montañas, Brocken, Hohehagen e Inselberg. Obtuvo una suma que era 15 segundos mayor que 180° . Si era correcto, esto indicaba que el espacio, en esa región al menos, estaba curvado positivamente. Pero necesitaríamos un triángulo mucho mayor y unas mediciones mucho más precisas para eliminar los errores de la observación. De modo que las observaciones de Gauss no eran concluyentes. El espacio podría ser euclíadiano, y también podría no serlo.

Mi comentario de que el espacio hiperbólico tridimensional es «tan solo curvo» depende de un nuevo punto de vista de la curvatura, lo cual también nos remite a Gauss. La esfera tiene una curvatura constante positiva y el plano hiperbólico tiene una curvatura constante negativa. Pero la curvatura de una superficie no tiene que ser constante. Podría ser una curva muy pronunciada en algunas zonas y menos pronunciada en otras. De hecho, podría ser positiva en algunas regiones y negativa en otras. La curvatura podría variar continuamente de una parte a otra. Si la superficie se parece a un hueso de perro, entonces los pegotes de los extremos están positivamente curvados, pero la parte que los une, está curvada negativamente.

Gauss buscó una fórmula para calificar la curvatura de una superficie en cualquiera de sus puntos. Cuando finalmente la encontró y la publicó en su *Disquisitiones Generales Circa Superficies Curva* (Investigación general sobre superficies curvas) en 1828, la llamó el «teorema egregium» (notable). ¿Qué es lo que era tan notable? Gauss había empezado con una visión naïf de la curvatura, incrustar la superficie en

el espacio tridimensional y calcular cómo de curvada estaba. Pero la respuesta le indicó que este espacio que la rodeaba no tenía importancia. No formaba parte de la fórmula. Escribió: «La fórmula ... llega por sí misma a un teorema notable: si una superficie curva se desarrolla sobre cualquier otra superficie, sea la que sea, la medida de la curvatura en cada punto no sufre ningún cambio». Con «desarrolla» quiere decir «envuelve alrededor».

Coge una hoja plana de papel, curvatura cero. Ahora envuélvela alrededor de una botella. La botella es cilíndrica y el papel se ajusta perfectamente, sin tener que doblarse, estirarse o romperse. Está curvada en lo que concierne a la apariencia visual, pero es un tipo de curvatura trivial, porque no ha cambiado la geometría del papel en ningún aspecto. Es tan solo un cambio en la manera en que el papel está colocado en el espacio que lo rodea. Pon el papel en plano y dibuja un ángulo recto, mide sus lados, compruébalos con el teorema de Pitágoras. Ahora enrolla el dibujo alrededor de la botella. La longitud de los lados medida en el papel, no cambia. Todavía se verifica el teorema de Pitágoras.

Sin embargo, la superficie de una esfera tiene curvatura distinta de cero. De modo que no es posible envolver una hoja de papel en torno a la esfera y que se ajuste bien sin tener que doblarla, estirarla o romperla. La geometría de la esfera es intrínsecamente diferente de la geometría del plano. Por ejemplo, el ecuador terrestre y las líneas de longitud de 0° a 90° al norte determinan un triángulo que tiene tres ángulos rectos y tres lados iguales (considerando que la Tierra es una esfera). Por lo tanto la ecuación de Pitágoras es falsa.

Hoy en día, llamamos curvatura a la «curvatura de Gauss» en su sentido intrínseco. Gauss explicó por qué es importante usando una analogía gráfica todavía vigente. Imagina una hormiga confinada en la superficie. ¿Cómo puede averiguar si la superficie está curvada? No puede salirse de la superficie para ver si parece curvada. Pero puede usar la fórmula de Gauss haciendo las medidas apropiadas simplemente en la superficie. Nos encontramos en la misma posición que la hormiga cuando intentamos descubrir la geometría verdadera de nuestro espacio. No podemos salimos de él. Antes podíamos emular a la hormiga tomando medidas, sin embargo, necesitamos una fórmula para la curvatura de un espacio de tres dimensiones. Gauss no dio una. Pero uno de sus estudiantes, en un arranque de

temeridad, reclamó que él la tenía.

El estudiante era Georg Bernhard Riemann, y estaba intentando lograr lo que las universidades alemanas llaman «habilitación», el paso siguiente tras el doctorado. En la época de Riemann esto significaba que podrías cobrar a los estudiantes una tasa por tus clases. Antes y ahora, obtener la habilitación requiere presentar tu investigación en una conferencia pública que es también un examen. El candidato ofrece varios temas y el examinador, que en el caso de Riemann era Gauss, escoge uno. Riemann, un brillante talento matemático, hizo una lista con varios temas ortodoxos que dominaba de sobra, pero en un ataque de locura también propuso «sobre la hipótesis en la que se funda la geometría». Gauss, que había estado interesado durante mucho tiempo en ello, lógicamente, lo escogió para el examen de Riemann.

Al instante, Riemann se arrepintió de ofrecer algo que era un gran reto. Tenía una fuerte aversión a hablar en público y no había analizado las matemáticas en detalle. Tan solo tenía algunas ideas vagas, aunque fascinantes, sobre la superficie curvada. Para cualquier dimensión. Lo que Gauss había hecho para dimensión dos con su teorema egregium, Riemann quería hacerlo para tantas dimensiones como se quisiera. Ahora tenía que ejecutarlo, y rápido. La conferencia era inminente. La presión casi le provocó una crisis nerviosa y no ayudó su trabajo para sobrevivir ayudando al colaborador de Gauss, Wilhelm Weber, experimentando con la electricidad. Bueno, quizás sí, porque mientras Riemann estaba pensando en la relación entre fuerzas eléctricas y magnéticas en su trabajo, se dio cuenta de que la fuerza puede relacionarse con la curvatura. Trabajando hacia atrás, podía usar la matemática de las fuerzas para definir la curvatura, como necesitaba en su examen. En 1854 Riemann pronunció su conferencia, la cual tuvo una calurosa acogida y con razón. Empezó definiendo lo que llamó una «variedad». Formalmente una «variedad» está definida por un sistema de muchas coordenadas, aunque con una fórmula para la distancia entre los puntos cercanos, ahora llamada métrica de Riemann. Informalmente, una variedad es un espacio multidimensional en toda su gloria. El clímax de la conferencia de Riemann fue una fórmula que generalizaba el teorema egregium de Gauss, definía la curvatura de una variedad solamente en términos de su métrica. Y es aquí donde el relato cierra el círculo por completo,

como la serpiente Uróboros se traga su propia cola, porque la métrica contiene restos visibles del teorema de Pitágoras.

Supón, por ejemplo, que la variedad tiene tres dimensiones. Sean las coordenadas de un punto (x, y, z) y sea $(x + dx, y + dy, z + dz)$ un punto cercano, donde la d significa «un poco de». Si el espacio es euclíadiano, con curvatura cero, la distancia ds entre estos dos puntos satisface la ecuación

$$ds^2 = dx^2 + dy^2 + dz^2$$

y justo esto es el teorema de Pitágoras restringido a puntos que están cerca los unos de los otros. Si el espacio es curvado, con la curvatura variable de un punto a otro, la fórmula análoga, la métrica, tiene este aspecto:

$$ds^2 = Xdx^2 + Ydy^2 + Zdz^2 + 2Udxdy + 2Vdxdz + 2Wdydz$$

Aquí X, Y, Z, U, V, W pueden depender de x, y y z . Puede parecer un poco un trabalenguas, pero, como la ecuación de Pitágoras, encierra sumas de cuadrados (y productos muy relacionados de dos cantidades como $dx dy$) más una cuantas florituras extra. Lo segundo se da porque la fórmula puede representarse como una tabla o matriz 3×3 :

$$\begin{matrix} X & U & V \\ U & Y & W \\ V & W & Z \end{matrix}$$

Donde X, Y, Z aparecen una vez, pero U, V, W aparecen dos veces. La tabla es simétrica sobre su diagonal, en el lenguaje de la geometría diferencial es un tensor simétrico. La generalización de Riemann del teorema egregium de Gauss es una fórmula para la curvatura de la variedad, en cualquier punto dado, en términos de este tensor. En el caso especial en que se puede aplicar Pitágoras, la curvatura resulta ser cero. Por lo tanto, la validez de la ecuación de Pitágoras es una prueba para la ausencia de curvatura.

Como la fórmula de Gauss, la expresión de Riemann para la curvatura depende solo

de la métrica de la variedad. Una hormiga confinada en la variedad podría observar la métrica midiendo pequeños triángulos y calculando la curvatura. La curvatura es una propiedad intrínseca de una variedad, independiente de cualquier espacio que le rodea. De hecho, la métrica ya determina la geometría, de modo que no se necesita ningún espacio que la rodee. En particular, nosotros, hormigas humanas, podemos preguntar qué forma tiene nuestro vasto y misterioso universo, y esperar responder mediante observaciones que no requieren que nos salgamos del universo. Lo cual está bien, porque no podemos hacerlo.

Riemann encontró su fórmula usando las fuerzas para definir la geometría. Cincuenta años más tarde, Einstein le dio la vuelta completamente a la idea de Riemann, usando la geometría para definir la fuerza de la gravedad en su teoría general de la relatividad e inspirando nuevas ideas sobre la forma del universo (véase el capítulo 13). Es una progresión de los eventos sorprendente. Primero surgió la ecuación de Pitágoras hace alrededor de 3.500 años para medir la tierra de un granjero. Su extensión a triángulos no rectángulos y a triángulos en la esfera nos permitió hacer mapas de nuestros continentes y medir nuestro planeta. Y una egregia generalización nos permitió medir la forma del universo. Las grandes ideas tienen comienzos pequeños.

Capítulo 2
Acortando los procesos
Logaritmos

$$\log xy = \log x + \log y$$

¿Qué nos dice?

Cómo multiplicar números sumando, en su lugar, números que están relacionados.

¿Por qué es importante?

Sumar es mucho más simple que multiplicar.

¿Qué provocó?

Métodos eficientes para calcular fenómenos astronómicos como eclipses y órbitas planetarias. Modos rápidos de realizar cálculos científicos. La compañera fiel de los ingenieros, la regla de cálculo. Descomposición radiactiva y la psicofísica de la percepción humana.

Los números se originaron en problemas prácticos: registro de la propiedad, como animales o tierras; y transacciones financieras, como impuestos y llevar las cuentas. La primera notación numérica conocida, aparte de las marcas simples de contar como IIII, se encuentra en el exterior de envolturas de arcilla. En el 8000 a.C., los contables de Mesopotamia llevaban los registros usando pequeñas piezas de formas diversas. El arqueólogo Denise Schmandt-Besserat se dio cuenta de que cada forma representaba un producto básico: una esfera para el grano, un huevo para una tinaja de aceite, etcétera. Por seguridad, las piezas se encerraban en envoltorios de arcilla. Pero era molesto romper la envoltura de arcilla para abrirla y averiguar cuántas piezas había dentro, de modo que los contables de la época grababan los símbolos en el exterior para indicar lo que había dentro. Finalmente se dieron

cuenta de que una vez que tenían estos símbolos, podían deshacerse de las piezas. El resultado fue una serie de símbolos escritos para los números, el origen de todos los símbolos numéricos posteriores y quizá, también, de la escritura.

Junto con los números llegó la aritmética: métodos para sumar, restar, multiplicar y dividir números. Instrumentos como el ábaco se usaban para hacer las sumas, luego los resultados se podían registrar con los símbolos. Con el tiempo, se encontraron formas de usar los símbolos para realizar los cálculos sin asistencia mecánica, y aunque el ábaco todavía se usa en muchas partes del mundo, las calculadoras electrónicas han suplantado los cálculos con lápiz y papel en la mayoría de los países.

La aritmética también resultó ser esencial en otros aspectos, especialmente en astronomía y topografía. Mientras los perfiles básicos de las ciencias físicas empezaban a emerger, los científicos novatos necesitaban realizar cálculos cada vez más elaborados manualmente. Con frecuencia esto consumía mucho de su tiempo, a veces meses o años, lo que se interponía en el camino de actividades más creativas. Finalmente se hizo esencial acelerar el proceso. Se inventaron innumerables instrumentos mecánicos, pero el avance más importante fue uno conceptual: pensar primero, calcular después. Usando las matemáticas de modo inteligente, se podían hacer cálculos difíciles mucho más fáciles.

Las nuevas matemáticas pronto desarrollaron una vida por sí mismas, resultando tener profundas implicaciones teóricas además de las prácticas. Hoy en día, esas ideas tempranas se han convertido en una herramienta indispensable para toda la ciencia, alcanzando incluso la psicología y las humanidades. Se usaban extensamente hasta la década de los ochenta del siglo pasado, cuando los ordenadores las volvieron obsoletas para propósitos prácticos, pero, a pesar de eso, su importancia en las matemáticas y la ciencia ha continuado creciendo.

La idea central es una técnica matemática llamada logaritmo. Su inventor fue un terrateniente escocés, pero fue un profesor de geometría con un gran interés en navegación y astronomía quien remplazó la idea brillante pero defectuosa del terrateniente por una mucho mejor.

En marzo de 1615, Henry Briggs escribió una carta a James Ussher, en la que se

registra un suceso crucial en la historia de la ciencia:

Napper, Lord de Markinston, ha puesto mi mente y manos a trabajar con sus nuevos y admirables logaritmos. Espero verlo este verano, si Dios lo permite, porque yo jamás vi un libro que me agradase e hiciese pensar más.

Briggs era el primer catedrático de geometría del Gresham College en Londres, y «Napper, Lord de Markinston» era John Napier, octavo terrateniente de Merchiston, ahora parte de la ciudad de Edimburgo, en Escocia. Napier parece haber sido un poco místico, tenía intereses teológicos fuertes, pero la mayoría se centraban en el Apocalipsis. Desde su punto de vista, su obra más importante era *Descubrimientos de todos los secretos del Apocalipsis de San Juan*, la cual le llevó a predecir que el mundo se acabaría o en 1688 o en 1700. Se creía que se había dedicado tanto a la alquimia como a la nigromancia, y sus intereses en las ciencias ocultas le crearon una reputación como mago. Según los rumores, llevaba consigo a todas partes una araña negra en una caja y poseía un «espíritu familiar» o compañía mágica: un gallito negro. Según uno de sus descendientes, Mark Napier, John empleaba a su espíritu familiar para pillar a los sirvientes que estaban robando. Encerraba al sospechoso en una habitación con el gallito y le mandaba acariciarlo, diciéndole que su pájaro mágico detectaría, de modo infalible, su culpa. Pero el misticismo de Napier tenía un corazón racional, el cual, en este ejemplo en particular, suponía cubrir al gallo con una fina capa de hollín. Un sirviente inocente tendría la confianza suficiente para acariciar al pájaro tal y como le había indicado, y se quedaría con hollín en sus manos. Uno culpable, temeroso porque se lo pillase, evitaría acariciar al pájaro. Así, irónicamente, las manos limpias probaban que se era culpable.

Napier dedicó mucho de su tiempo a las matemáticas, especialmente a métodos para acelerar los complicados cálculos aritméticos. Una invención, el ábaco neperiano, era un conjunto de diez varillas, marcadas con números, las cuales simplificaban el proceso para una multiplicación larga. La invención que creó su reputación y generó una revolución científica fue todavía mejor; no fue su libro del Apocalipsis, como él habría esperado, sino que fue su *Mirifici Logarithmorum Canonis Descriptio* (Descripción del maravilloso canon de logaritmos) de 1614. El prefacio muestra que Napier sabía exactamente lo que él había aportado y para qué

era bueno:⁵

Puesto que nada es más aburrido, compañeros matemáticos, en la práctica de las artes matemáticas, que el gran retraso sufrido en el tedio de las multiplicaciones y divisiones largas y pesadas, el hallazgo de proporciones y en la extracción de raíces cuadradas y cúbicas, y... los muchos errores escurridores que pueden surgir; yo he estado dándole vueltas a mi cabeza de cómo podría ser capaz de solventar las dificultades mencionadas para que sea un arte segura y rápida. Al final, después de pensar mucho, finalmente he encontrado un modo asombroso de acortar los procedimientos... es una tarea agradable exponer el método para el uso público de los matemáticos.

En el momento en que Briggs oyó hablar de los logaritmos se quedó encantado. Como muchos matemáticos de su época, pasaba mucho tiempo realizando cálculos astronómicos. Sabemos esto porque otra carta de Briggs a Ussher, que data de 1610, menciona los cálculos de eclipses y porque Briggs había publicado con anterioridad dos libros de tablas numéricas, uno relacionado con el Polo Norte y otro para la navegación. Todos estos trabajos habían requerido vastas cantidades de aritmética y trigonometría complicada. La invención de Napier ahorraría una gran cantidad de labor tediosa. Pero cuanto más estudiaba Briggs el libro, más convencido estaba que, aunque la estrategia de Napier era maravillosa, sus tácticas estaban equivocadas. Briggs dio con una mejora simple pero efectiva, e hizo el largo viaje a Escocia. Cuando se encontraron, «casi un cuarto de hora se pasó cada uno contemplando con admiración al otro, antes de que una palabra fuese dicha».⁶ ¿Qué era eso tan emocionante que despertaba tanta admiración? La observación vital, obvia para cualquiera que aprendiera aritmética, era que la suma de números es relativamente fácil, pero multiplicarlos no lo es. La multiplicación requiere muchas más operaciones aritméticas que una suma. Por ejemplo, sumar dos números con una longitud de diez dígitos supone diez pasos simples, pero la multiplicación necesita 200. Con los ordenadores actuales este tema es todavía importante, aunque está escondido tras los algoritmos usados para la

⁵ <http://www.17centurymaths.com/contents/napiercontents.html>

⁶ Citado de una carta escrita por John Marr a William Lilly.

multiplicación.

Pero en la época de Napier, todavía tenía que hacerse a mano. ¿No sería fantástico si hubiese algún truco matemático que convirtiese las molestas multiplicaciones en sumas rápidas y agradables? Suena demasiado bien para ser cierto, pero Napier se dio cuenta de que era posible. El truco era trabajar con potencias de un número fijo. En álgebra, las potencias de un x desconocido se indican con un número pequeño puesto algo más alto. Es decir,

$$xx = x^2$$

$$xxx = x^3$$

$$xxxx = x^4,$$

etcétera, donde, como es habitual en álgebra, una letra puesta junto a otra indica que se están multiplicando. De modo que, por ejemplo,

$$10^4 = 10 \times 10 \times 10 \times 10 = 10.000$$

No necesitas juguetear con estas expresiones durante mucho rato antes de descubrir un modo fácil de resolver, digamos, $10^4 \times 10^3$. Tan solo escríbelo:

$$\begin{aligned} 10.000 \times 1.000 &= (10 \times 10 \times 10 \times 10) \times (10 \times 10 \times 10) \\ &= 10 \times 10 \times 10 \times 10 \times 10 \times 10 \times 10 \\ &= 10.000.000 \end{aligned}$$

El número de ceros en la respuesta es 7, que es igual a $4 + 3$. El primer paso del cálculo muestra por qué es $4 + 3$, ponemos cuatro 10 y tres 10 seguidos. En resumen:

$$10^4 \times 10^3 = 10^4 + 3 = 10^7$$

Del mismo modo, cualquiera que sea el valor de x , si multiplicamos su a -ésima potencia por su b -ésima potencia, siendo a y b números enteros, entonces

obtenemos su $(a + b)$ -ésima potencia:

$$x^a x^b = x^{a+b}$$

Esta podría parecer una fórmula inocua, pero en la parte izquierda multiplicamos dos cantidades, mientras que en la derecha el paso principal es sumar a y b , lo cual es mucho más simple.

Supongamos que queremos multiplicar, por ejemplo, 2,67 y 3,51. Haciendo una multiplicación, que es larga, se obtiene 9,3717, que redondeando a dos cifras decimales es 9,37. ¿Qué pasa si intentamos usar la fórmula anterior? El truco recae en la elección de x . Si consideramos x como 1,001, entonces un poco de aritmética revela que (con el redondeo a dos decimales):

$$\begin{aligned} (1,001)^{983} &= 2,67 \\ (1,001)^{1,256} &= 3,51 \end{aligned}$$

Entonces, la fórmula nos dice que $2,67 \times 3,51$ es:

$$1,001^{(983 + 1,256)} = 1,001^{(2,239)}$$

Que, redondeando a dos decimales, es 9,37.

El núcleo del cálculo es una suma fácil: $983 + 1,256 = 2,239$. Sin embargo, si tratas de comprobar mi aritmética te darás cuenta rápidamente de que si he hecho algo, lo que he hecho ha sido el problema más difícil, no más fácil. Para resolver $1,001^{983}$ tienes que multiplicar 1,001 por sí mismo 983 veces. Y para descubrir que la potencia que hay que usar es 983, tienes que llevar a cabo más trabajo. Así que a primera vista esto parece una idea bastante poco útil.

Pero la gran perspicacia de Napier fue considerar esta objeción como errónea. Para vencerla, algún alma resistente tiene que calcular un montón de potencias de 1,001, empezando en $1,001^2$ y siguiendo hasta por ejemplo $1,001^{10,000}$. Entonces se puede publicar una tabla con estas potencias. Después de eso, la mayoría del trabajo está hecho. Tan solo tienes que arrastrar tu dedo a través de las potencias

sucesivas hasta que veas 2,67 seguido de 983, y de manera similar localizas 3,51 al lado de 1.256. Entonces sumas estos dos números para obtener 2.239. La fila correspondiente de la tabla te dice que esta potencia de 1,001 es 9,37. Trabajo hecho.

Resultados muy precisos necesitan potencias de algo mucho más próximo a 1, algo como 1,000001. Esto hace la tabla mucho más grande, con más o menos un millón de potencias. Hacer los cálculos para esa tabla es una tarea enorme. Pero solo tiene que hacerse una vez. Si algún benefactor se autosacrifica y hace el esfuerzo, generaciones futuras se ahorrarán una cantidad gigantesca de aritmética.

En el contexto de este ejemplo, podemos decir que las potencias 983 y 1.256 son los logaritmos de los números que queremos multiplicar, 2,67 y 3,51. De modo similar, 2.239 es el logaritmo de su producto, 9,38. Escribimos log como la abreviatura, y lo que hemos hecho equivale a la ecuación:

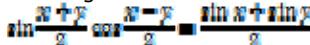
$$\log ab = \log a + \log b$$

que se cumple para cualesquiera números a y b . La elección algo arbitraria de 1,001 se llama la base. Si usamos una base diferente, los logaritmos que calculamos también son diferentes, pero para cualquiera que sea la base fijada, todo funciona del mismo modo.

Esto es lo que Napier debería haber hecho. Pero por razones que solo podemos adivinar, hizo algo ligeramente diferente. Briggs, que se acercó a la técnica desde una perspectiva fresca, descubrió dos modos de mejorar la idea de Napier.

Cuando Napier empezó a pensar en las potencias de números, a finales del siglo XVI, la idea de reducir la multiplicación a la suma circulaba ya entre los matemáticos. En Dinamarca, se usaba un método bastante más complicado conocido como «prostaféresis», basado en una fórmula con funciones trigonométricas.⁷ Napier, intrigado, fue lo suficiente listo para darse cuenta de que las potencias de un número fijo podían hacer el mismo trabajo de manera más

⁷ La prostaféresis estaba basada en una fórmula trigonométrica descubierta por François Viète, a saber:



Si tienes una tabla de senos, la fórmula te permite calcular cualquier producto usando solo sumas, restas y la división entre 2.

simple. Las tablas necesarias no existían, pero eso se remedió fácilmente. Algunas almas patrióticas debían de llevar a cabo el trabajo. El propio Napier se presentó voluntario para la tarea, pero cometió un error estratégico. En vez de usar una base que fuese ligeramente mayor que 1, usó una base ligeramente menor que 1. En consecuencia la secuencia de potencias empezaba con números grandes, que sucesivamente se iban haciendo más pequeños. Esto hizo los cálculos un poco más toscos.

Briggs descubrió este problema, y vio cómo lidiar con él: usó una base ligeramente mayor que 1. También se dio cuenta de un error más sutil, y también lo abordó. Si el método de Napier se modificase para trabajar con potencias de algo como 1,0000000001, no habría relación directa entre los logaritmos de, por ejemplo, 12,3456 y 1,23456. De modo que no estaba totalmente claro cuándo la tabla podía acabarse. La fuente del problema era el valor del log 10, porque:

$$\log 10x = \log 10 + \log x$$

Desafortunadamente log 10 no era fácil, con la base 1,0000000001 el logaritmo de 10 era 23.025.850.929. Briggs pensó que sería mucho mejor si la base se pudiese escoger para que $\log 10 = 1$. Entonces $\log 10x = 1 + \log x$, de modo que cualquiera que fuese log 1,23456, solo se necesitase sumar 1 para obtener log 12,3456. En este caso, las tablas de logaritmos solo necesitarían ir de 1 a 10. Si se tenían números más grandes, bastaba añadir el número entero apropiado.

Para conseguir $\log 10 = 1$, haces lo mismo que Napier, usando una base de 1,0000000001, pero entonces divides cada logaritmo por el curioso número de 23.025.850.929. La tabla resultante consiste en logaritmos de base 10, lo cual escribiré como $\log_{10} x$. Satisfacen:

$$\log_{10} xy = \log_{10} x + \log_{10} y$$

como antes, pero también:

$$\log_{10} 10x = \log_{10} x + 1$$

Dos años después de la muerte de Napier, Briggs empezó a trabajar en una tabla de logaritmos de base 10. En 1617 publicó *Logarithmorum Chilias Prima* (Logaritmos del primer millar), los logaritmos de enteros del 1 al 1.000 aproximados a 14 cifras decimales. En 1624, continuó con *Arithmetica Logarithmica* (Aritmética de logaritmos), una tabla de logaritmos de base 10 de números del 1 al 20.000 y del 90.000 al 100.000, con la misma precisión. Rápidamente otros siguieron el ejemplo de Briggs, llenando el gran hueco y desarrollando tablas auxiliares, tales como logaritmos de funciones trigonométricas como $\log \sin x$.

Las mismas ideas que inspiraron los logaritmos nos permiten definir las potencias x^a de una variable positiva x para valores de a que no son números enteros positivos. Todo lo que tenemos que hacer es insistir en que nuestras definiciones sean consistentes con la ecuación $x^a x^b = x^{a+b}$, y dejarse guiar por la intuición. Para evitar complicaciones molestas, es mejor asumir que x es positiva y definir x^a también como positivo. (Para x negativo, es mejor introducir números complejos, como se ve en el capítulo 5.)

Por ejemplo, ¿qué quiere decir x^0 ? Teniendo en mente que $x^1 = x$, la fórmula dice que x^0 debe satisfacer $x^0 x = x^{0+1} = x$. Dividiendo por x , tenemos que $x^0 = 1$. ¿Qué pasa ahora con x^{-1} ? Bien, la fórmula dice que $x^{-1} x = x^{-1+1} = x^0 = 1$. Dividiendo por x , tenemos que $x^{-1} = 1/x$. De manera similar, $x^{-2} = 1/x^2$, $x^{-3} = 1/x^3$, etcétera. Empieza a ponerse más interesante, y potencialmente muy útil, cuando pensamos en $x^{1/2}$. Esto tiene que satisfacer que

$$x^{1/2} x^{1/2} = x^{1/2 + 1/2} = x^1 = x$$

De modo que $x^{1/2}$ multiplicado por sí mismo es x . El único número con esta propiedad es la raíz cuadrada de x . Así que $x^{1/2} = \sqrt{x}$. De manera similar, $x^{1/3} = \sqrt[3]{x}$, la raíz cúbica. Si continuamos de este modo, podemos definir $x^{p/q}$ para cualquier fracción p/q . Entonces, usando fracciones para aproximar números reales, podemos definir x^a para cualquier número real a . Y la ecuación

$$x^a x^b = x^{a+b}$$

todavía se cumple.

También se deduce que

$$\log \sqrt{x} = \frac{1}{2} \log x$$

y que

$$\log \sqrt[3]{x} = \frac{1}{3} \log x,$$

así que podemos calcular raíces cuadradas y cúbicas fácilmente usando una tabla de logaritmos. Por ejemplo, para encontrar la raíz cuadrada de un número, consideramos el logaritmo del número y lo dividimos entre 2, y luego averiguamos qué número tiene ese resultado como su logaritmo. Para raíces cúbicas lo mismo pero dividimos entre 3. Los métodos tradicionales para estos problemas eran aburridos y complicados. Puedes ver por qué Napier saca a relucir raíces cuadradas y cúbicas en el prefacio de su libro.

Tan pronto como las tablas completas de logaritmos estuvieron disponibles, se hicieron indispensables para los científicos, ingenieros, topógrafos y navegantes. Se ahorraron tiempo, esfuerzo e incrementaron la probabilidad de que la respuesta fuese correcta. En un primer momento, la astronomía fue una beneficiaria importante, porque los astrónomos necesitaban de modo rutinario realizar cálculos largos y difíciles. El matemático y astrónomo francés Pierre Simon de Laplace dijo que la invención de logaritmos «reduce a unos pocos días la labor de muchos meses, dobla la vida del astrónomo y le ahorra errores y disgustos». A medida que el uso de la maquinaria en la industria crecía, los ingenieros empezaron a hacer más y más uso de las matemáticas: diseñar herramientas complejas, analizar la estabilidad de puentes y edificios, construir coches, camiones, barcos y aviones. Los logaritmos fueron una parte fuerte del currículum escolar de matemáticas hace unas pocas décadas. Y los ingenieros llevaban lo que en realidad era una calculadora analógica para logaritmos en sus bolsillos, una representación física de las ecuaciones básicas para logaritmos para su uso inmediato. Le llamaron una regla de cálculo, y la usaban rutinariamente en aplicaciones que iban desde la arquitectura hasta el diseño de aviones.

La primera regla de cálculo la construyó un matemático inglés, William Oughtred, en 1630, usando escalas circulares. Modificó el diseño en 1632, haciendo las dos reglas rectas. Esta fue la primera regla de cálculo. La idea es simple; cuando colocas dos varillas en fila, sus longitudes se suman. Si las varillas están marcadas usando una escala logarítmica, en la cual los números están separados según sus logaritmos, entonces los números correspondientes se multiplican. Por ejemplo, se coloca el 1 en una varilla frente al 2 de otra. Entonces, frente a cualquier número x de la primera varilla, tenemos $2x$ en la segunda. De modo que opuesto a 3 encontramos 6, etcétera (véase la figura 11). Si los números son más complicados, digamos que 2,67 y 3,51, colocamos 1 frente a 2,67 y leemos lo que haya enfrente a 3,59, a saber, 9,37. Es así de fácil.

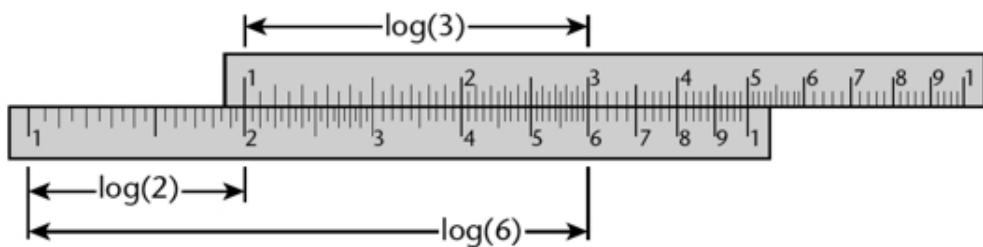


FIGURA 11. Multiplicación de 2 por 3 en una regla de cálculo.

Los ingenieros rápidamente desarrollaron reglas de cálculo elaboradas con funciones trigonométricas, raíces cuadradas, escalas log-log (logaritmos de logaritmos) para calcular potencias, etcétera. Finalmente los logaritmos se relegaron a un segundo término con los ordenadores digitales, pero incluso ahora los logaritmos todavía juegan un papel importante en la ciencia y la tecnología, junto con su inseparable compañera, la función exponencial. Para logaritmos de base 10, esta es la función 10^x ; para logaritmos neperianos, la función e^x , donde $e = 2,71828$, aproximadamente. En cada par, las dos funciones son la inversa una de la otra. Si tomas un número, formas su logaritmo y luego haces el exponencial de ese, obtienes el número con el que empezaste.

¿Por qué necesitamos logaritmos ahora que tenemos ordenadores?

En 2011 un terremoto de magnitud 9.0 en la costa este de Japón causó un tsunami gigantesco, el cual destrozó una gran área poblada y mató alrededor de 25.000

personas. En la costa había una planta de energía nuclear, Fukushima Dai-ichi (Planta de energía nuclear número 1 de Fukushima, para distinguirla de una segunda situada cerca). Constaba de seis reactores nucleares separados: tres estaban operativos cuando el tsunami la alcanzó; los otros tres se habían detenido temporalmente y su combustible había sido transferido a piscinas de agua fuera de los reactores pero dentro de los edificios de los reactores.

El tsunami arrolló las defensas de la planta, cortando el suministro de la corriente eléctrica. Los tres reactores en uso (números 1, 2 y 3) se apagaron como medida de seguridad, pero sus sistemas de refrigeración todavía se necesitaban para impedir que el combustible se fundiese. No obstante, el tsunami también destrozó los generadores de emergencia, los cuales estaban destinados a alimentar el sistema de refrigeración y otros sistemas de seguridad críticos. El siguiente nivel de seguridad, las baterías, se quedó rápidamente sin energía. El sistema de refrigeración se paró y el combustible nuclear en varios reactores empezó a sobrecalentarse. Improvisando, los operadores usaron coches de bomberos para bombear agua del mar a los tres reactores operativos, pero esta reaccionó con el revestimiento de circonio en las varillas del combustible para producir hidrógeno. La acumulación de hidrógeno causó una explosión en el edificio que albergaba el reactor 1. Los reactores 2 y 3 pronto sufrieron la misma suerte. El agua en la piscina del reactor 4 se fue por el desagüe, dejando su combustible expuesto. Para cuando los operarios recobraron alguna apariencia de control, al menos un recipiente de contención del reactor se había fracturado y la radiación se estaba filtrando al entorno local. Las autoridades japonesas evacuaron a 200.000 personas del área de los alrededores porque la radiación estaba bastante por encima de los límites de seguridad normales. Seis meses más tarde, la compañía operaria de los reactores, TEPCO, afirmó que la situación permanecía siendo crítica y que se necesitaba mucho más trabajo antes de que se pudiese considerar que los reactores estaban totalmente bajo control, pero indicaba que la fuga había sido detenida.

No quiero analizar los méritos o no de la energía nuclear aquí, pero quiero mostrar cómo el logaritmo responde una pregunta vital: si sabes cuánto material radiactivo se ha liberado y de qué tipo, ¿cuánto tiempo permanecerá en un ambiente donde podría ser peligroso?

Los elementos radiactivos se descomponen, esto es, se convierten en otros elementos a través de procesos nucleares, emitiendo partículas nucleares mientras lo hacen. Son estas partículas las que constituyen la radiación. El nivel de radiactividad disminuye con el tiempo del mismo modo que la temperatura de un cuerpo caliente disminuye cuando se enfria: exponencialmente. Así, en las unidades apropiadas, las cuales no discutiremos aquí, el nivel de radiactividad $N(t)$ con el tiempo t , cumple la ecuación:

$$N(t) = N_0 e^{-kt}$$

Donde N_0 es el nivel inicial y k es una constante que depende del elemento que nos concierne. Más exactamente, depende de la forma, o isótopo, del elemento que estemos considerando.

Una medida conveniente del tiempo que perdura la radiactividad es el período de semidesintegración, un concepto que se introdujo por primera vez en 1907. Esto es el tiempo que tarda un nivel inicial N_0 en reducirse a la mitad de su tamaño. Para calcular el período de semidesintegración, solucionamos la ecuación:

$$\frac{1}{2} N_0 = N_0 e^{-kt}$$

Tomando logaritmos en ambas partes. El resultado es:

$$t = \frac{\log 2}{k} = \frac{0.6931}{k}$$

y podemos resolver esto porque de la experimentación sabemos el valor de k .

El período de semidesintegración es un modo práctico de calcular cuánto durará la radiación. Supongamos que el período de semidesintegración es, por ejemplo, una semana. Entonces la velocidad original a la cual el material emite la radiación es la mitad después de 1 semana, habrá bajado a un cuarto después de 2 semanas, un octavo tras 3 semanas, etcétera. Tarda 10 semanas en reducirse a una milésima de su nivel original (realmente $1/1024$) y 20 semanas en bajar a una millonesima.

En accidentes con reactores nucleares convencionales, los productos radiactivos más importantes son yodo-131 (un isótopo radiactivo del yodo) y cesio-137 (un isótopo radiactivo del cesio). El primero puede causar cáncer de tiroides, porque la glándula tiroides concentra yodo. El período de semidesintegración del yodo-131 es solo 8 días, por lo tanto, causa daños pequeños si la medicación correcta está disponible, y sus peligros decrecen bastante rápidamente a menos que continúe la fuga. El tratamiento estándar es dar a la gente pastillas de yodo, las cuales reducen la cantidad de la forma radiactiva que es absorbida por el cuerpo, pero el remedio más efectivo es dejar de beber leche contaminada.

El cesio-137 es muy diferente, tiene un período de semidesintegración de 30 años. Tarda sobre 200 años en alcanzar un nivel de radiactividad que baje a una centésima el valor inicial, por lo que permanece como un peligro durante mucho tiempo. El principal asunto práctico en un accidente de un reactor es la contaminación del suelo y edificios. La descontaminación es hasta cierto punto factible, pero cara. Por ejemplo, el suelo puede quitarse, deshacerse, y almacenarse en un lugar seguro. Pero esto crea cantidades enormes de residuos radiactivos de baja actividad.

La descomposición radiactiva es tan solo una de las áreas de las muchas en las que los logaritmos de Napier y Briggs continúan sirviendo a la ciencia y a la humanidad. Si ojeas capítulos posteriores, los encontrarás en termodinámica y teoría de la información, por ejemplo. Aunque los rápidos ordenadores han hecho ahora los logaritmos redundantes para su propósito original, cálculos rápidos, siguen siendo fundamentales para la ciencia por razones más conceptuales que computacionales. Otra aplicación de los logaritmos se da en los estudios de la percepción humana: cómo sentimos el mundo alrededor nuestro. Los primeros pioneros de la psicofísica de la percepción hicieron estudios exhaustivos de la vista, el oído y el tacto, y revelaron algunas regularidades matemáticas fascinantes.

En la década de los cuarenta del siglo XIX, un doctor alemán, Ernst Weber, llevó a cabo experimentos para determinar cómo de sensible es la percepción humana. Les daba a los sujetos pesos para soportar en sus manos y les preguntaba cuándo podían percibir que uno era más pesado que otro. Quizá sorprendentemente, esta diferencia (para un sujeto experimental dado) no fuese una cantidad fija. Dependía

de cómo de pesados fuesen los pesos que se comparaban. La gente no percibía una diferencia mínima absoluta de, por ejemplo, 50 gramos. Lo que sentían era una diferencia mínima relativa, por ejemplo, un 1 % de los pesos que se comparaban. Esto es, la diferencia más pequeña que el sentido humano puede detectar es proporcional a los estímulos, la cantidad física real.

A mediados del siglo XIX, Gustav Fechner redescubrió la misma ley y la reformuló matemáticamente. Esto le llevó a una ecuación, la cual llamó ley de Weber, pero en la actualidad normalmente se hace referencia a ella como la ley de Fechner (o la ley de Weber-Fechner si eres un purista). Afirma que la sensación percibida es proporcional al logaritmo de los estímulos. Los experimentos sugieren que esta ley se aplica no solo a nuestra sensación del peso, sino también a la visión y al oído. Si miramos una luz, el brillo que percibimos varía igual que el logaritmo de la potencia de energía real. Si una fuente es diez veces más brillante que otra, entonces la diferencia que percibimos es constante, comoquiera que realmente sea el brillo de las dos fuentes. Lo mismo aplica para el volumen de los sonidos: una explosión con diez veces más energía suena una cantidad fija más fuerte.

La ley de Weber-Fechner no es totalmente exacta, pero es una buena aproximación. La evolución más o menos tuvo que definirse con algo como una escala logarítmica, porque el mundo externo plantea a nuestros sentidos unos estímulos de una gran variedad de tamaños. Un ruido podría ser tan solo un ratón escabulléndose en un seto, o podría ser un trueno; necesitamos ser capaces de oír ambos. Pero el rango de niveles de sonido es tan vasto que ningún instrumento sensorial biológico puede responder en proporción a la energía generada por el sonido. Si una oreja pudiese oír al ratón escabullirse, entonces el trueno la destrozaría. Si sintoniza a la baja los niveles de sonido, de modo que el trueno produzca una señal confortable, entonces no sería capaz de oír al ratón. La solución es comprimir los niveles de energía a un rango cómodo y los logaritmos hacen exactamente eso. Siendo sensibles a proporciones más que a valores absolutos tiene un excelente sentido y da lugar a sentidos excelentes.

Nuestra unidad estándar de ruido, el decibelio, condensa la ley de Weber-Fechner en una definición. No mide el ruido absoluto, sino el ruido relativo. Un ratón en la hierba produce unos 10 decibelios. Una conversación normal entre personas a un

metro de distancia tiene lugar a 40-60 decibelios. Una batidora dirige unos 60 decibelios a la persona que la está usando. El ruido en un coche, causado por el motor y los neumáticos, es 60-80 decibelios. Un avión reactor a una distancia de 100 metros produce 110-140 decibelios, que se incrementa a 150 al estar a treinta metros. Una vuvuzela (el molesto instrumento de plástico parecido a una trompeta que se escuchó mucho durante el Mundial de fútbol de 2010 y llevado a casa como recuerdo por fans insensatos) genera 120 decibelios a un metro; una granada detonadora militar produce hasta 180 decibelios.

Se encuentran escalas como esta sin problema porque tienen una vertiente relacionada con la seguridad. El nivel al cual el sonido puede potencialmente dañar el oído es de alrededor de 120 decibelios. Por favor, tira a la basura tu vuvuzela.

Capítulo 3
Fantasmas de cantidades difuntas
Cálculo

$$\frac{df}{dt} = \lim_{h \rightarrow 0} \frac{f(t+h) - f(t)}{h}$$

(se hace muy pequeño)

¿Qué dice?

Para encontrar la tasa de variación instantánea de una cantidad que varía con, por ejemplo, el tiempo, calcula cómo su valor cambia durante un intervalo de tiempo corto y divide por el tiempo en cuestión. Luego permite al intervalo hacerse arbitrariamente pequeño.

¿Por qué es importante?

Proporciona unas bases rigurosas para el cálculo, el principal modo con el que los científicos representan el mundo natural.

¿Qué provocó?

Cálculo de tangentes y áreas. Fórmulas para volúmenes de sólidos y longitudes de curvas. Leyes de Newton del movimiento, ecuaciones diferenciales. Las leyes de conservación de la energía y el momento. La mayoría de la física matemática.

En 1665, Carlos II era rey de Inglaterra, y su capital, Londres, era una metrópoli de medio millón de personas que se expandía descontroladamente. El arte florecía y la ciencia estaba en etapas tempranas de un desarrollo cada vez más rápido. La Royal Society, quizás la sociedad científica más antigua que todavía existe, se había

fundado 5 años antes y Carlos II le había concedido la cédula real. Los ricos vivían en casas impresionantes y el comercio era próspero, pero los pobres vivían hacinados en callejones sombríos a causa de edificios destalados que resaltaban cada vez más a medida que crecían planta a planta. Las condiciones de salubridad eran poco adecuadas; ratas y otras alimañas estaban por todas partes. A finales de 1666, un quinto de la población de Londres se había muerto víctima de la peste bubónica, propagada primero por las ratas y luego por las personas. Fue el peor desastre en la historia de la capital, y la misma tragedia azotó a toda Europa y el norte de África. El rey partió apresuradamente hacia la mucho más saludable campiña de Oxfordshire, y regresó a principios de 1666. Nadie sabía qué causaba la peste, y las autoridades locales trataron de todo: provocaban fuegos continuamente para limpiar el aire, quemaban todo lo que despidiere un olor fuerte, enterraban a los muertos rápidamente en fosas. Mataron muchos perros y gatos, lo cual irónicamente eliminó dos controles sobre la población de ratas.

Durante estos dos años, un universitario poco conocido y sin pretensiones del Trinity College, Cambridge, completaba sus estudios. Con la esperanza de evitar la peste, volvió a la casa donde había nacido, desde la cual su madre llevaba una granja. Su padre había muerto poco antes de su nacimiento, y él había sido educado por su abuela materna. Quizá inspirado por la paz y tranquilidad rural, o por no tener nada mejor que hacer con su tiempo, el joven reflexionaba sobre la ciencia y las matemáticas. Más tarde escribiría: «*En esos días estaba en la flor de la vida para la invención, y dispuesto para las matemáticas y la filosofía (natural) más que en cualquier otra época desde entonces*». Sus investigaciones le llevaron a comprender la importancia de la ley de la inversa del cuadrado de la gravedad, una idea que había estado merodeando ineficazmente durante al menos 50 años. Él dio con un método práctico para resolver problemas en el cálculo, otro concepto que estaba en el aire pero no había sido formulado en términos generales. Y descubrió que la luz del sol blanca estaba compuesta de muchos colores diferentes. Todos los colores del arco iris.

Cuando la peste se fue apaciguando, no le habló a nadie sobre los descubrimientos que había hecho. Volvió a Cambridge, hizo un máster y se convirtió en un profesor del Trinity. Elegido para la cátedra Lucasiana de Matemáticas, finalmente empezó a

publicar sus ideas y desarrollar nuevas.

El joven era Isaac Newton. Sus descubrimientos crearon una revolución en ciencias, provocando un mundo que Carlos II nunca habría creído que pudiese existir: edificios de más de cien plantas, carruajes sin caballos a más de 80 kilómetros por hora por la autopista, mientras los conductores escuchan música usando un disco mágico hecho a partir de un material extraño parecido al cristal, máquinas voladoras más pesadas que el aire que cruzan el Atlántico en seis horas, imágenes en color que se mueven y cajas que puedes llevar en tu bolsillo para hablar con el otro extremo del mundo...

Antes, Galileo Galilei, Johannes Kepler y otros habían levantado la esquina de la alfombrilla de la naturaleza y visto unas pocas de las maravillas ocultas tras ella. Ahora Newton echaba la alfombrilla a un lado. No solo revelaba que el universo seguía pautas secretas, leyes de la naturaleza, también proporcionaba herramientas matemáticas para expresar esas leyes de manera precisa y deducir sus consecuencias. El sistema del mundo era matemático; el corazón de la creación de Dios era un universo desalmado que funcionaba como un mecanismo de relojería.

La humanidad no cambió de repente la visión del mundo de religiosa a laica. Todavía no lo ha hecho por completo, y probablemente nunca lo hará. Pero después de que Newton publicase su *Philosophiæ Naturalis Principia Mathematica* (Principios matemáticos de filosofía natural) el «sistema del mundo» —el subtítulo del libro— no fue nunca más competencia exclusiva de la religión institucionalizada. Incluso así, Newton no fue el primer científico moderno, tenía un lado místico también, dedicando años de su vida a la alquimia y especulación religiosa. En los apuntes para una conferencia,⁸ el economista John Maynard Keynes, y también un erudito newtoniano, escribió:

Newton no fue el primero de la edad de la razón. Él fue el último de los magos, el último de los babilonios y sumerios, la última gran mente que miró al mundo visible e intelectual con los mismos ojos que aquellos que

⁸ Keynes nunca dio la conferencia. La Royal Society planeaba conmemorar el tricentenario en 1942, pero la Segunda Guerra Mundial se interpuso, así que las celebraciones se pospusieron a 1946. Los conferenciantes eran los físicos Edward da Costa Andrade y Niels Bohr y los matemáticos Herbert Turnbull y Jacques Hadamard. La sociedad también invitó a Keynes, cuyos intereses incluían tanto los manuscritos de Newton como la economía. Keynes había escrito una conferencia con el título «Newton, the man» (Newton, el hombre), pero murió antes de que el evento tuviese lugar. Su hermano Geoffrey leyó la conferencia en su nombre.

empezaron a construir nuestra herencia intelectual hace algo menos de 10.000 años. Isaac Newton, un niño póstumo nacido sin padre en el día de Navidad, en 1642, fue el último niño prodigo a quienes los Reyes Magos pudieron hacer un homenaje sincero y apropiado.

Hoy ignoramos en su mayoría el aspecto místico de Newton y le recordamos por sus logros científicos y matemáticos. Primordial entre ellos está su comprensión de que la naturaleza obedece leyes matemáticas y su invención del cálculo, el principal modo en el que ahora expresamos esas leyes y obtenemos sus consecuencias. El matemático y filósofo alemán Gottfried Wilhelm Leibniz también desarrolló el cálculo, más o menos independientemente, en más o menos la misma época, pero hizo poco con él. Newton usó el cálculo para comprender el universo, aunque lo mantuvo en secreto en su trabajo publicado, modelándolo de nuevo en el lenguaje geométrico clásico. Fue una figura de transición quien alejó a la humanidad de una mirada mística y medieval y la condujo a una visión moderna y racional del mundo. Después de Newton, los científicos conscientemente reconocieron que el universo tiene modelos matemáticos profundos y está equipado con técnicas potentes para explotar esa visión.

El cálculo no surgió «de la nada». Vino de preguntas tanto de la matemática pura como de la aplicada, y sus antecedentes se pueden seguir hasta Arquímedes. El propio Newton observó acertadamente «*si he visto un poco más allá es porque me he puesto a hombros de gigantes*».⁹ Primordiales entre esos gigantes eran John Wallis, Pierre de Fermat, Galileo y Kepler. Wallis desarrolló un precursor del cálculo en su *Arithmetica Infinitorum* (Aritmética del infinito) de 1656. El *De Tangentibus Linearum Curvarum* (Sobre tangentes a líneas curvas) de Fermat en 1679 presentaba un método para encontrar tangentes a curvas, un problema íntimamente relacionado con el cálculo. Kepler formula tres leyes básicas del movimiento de planetas, lo cual llevó a Newton a su ley de la gravedad, el tema del próximo capítulo. Galileo hizo grandes avances en astronomía, pero también investigó aspectos matemáticos de la naturaleza terrestre, publicando sus

⁹ Esta frase proviene de una carta que Newton escribió a Hooke en 1676. No era nueva, en 1159 John de Salisbury escribió que «Bernard de Chartres solía decir que somos como enanos a hombros de gigantes, de modo que podemos ver más que ellos». En el siglo XVII se había convertido en un cliché.

descubrimientos en *Motu* (Sobre el movimiento) en 1590. Investigó cómo se mueve un cuerpo que está cayendo, y encontró un elegante modelo matemático. Newton desarrolló este indicio en tres leyes generales de movimiento.

Para comprender el modelo de Galileo, necesitamos dos conceptos cotidianos de mecánica: velocidad y aceleración. La velocidad es cómo de rápido se mueve algo y en qué dirección. Si ignoramos la dirección, obtenemos la celeridad de un cuerpo. La aceleración es un cambio en la velocidad, lo que normalmente lleva consigo un cambio en la celeridad (surge una excepción cuando la celeridad se mantiene igual pero cambia la dirección). En la vida cotidiana usamos la aceleración para indicar que se aumenta la velocidad, y la deceleración cuando se disminuye, pero en mecánica ambos cambios son aceleración; el primero positivo, el segundo negativo. Cuando conducimos por una carretera la celeridad del coche se muestra en el velocímetro, puede ser por ejemplo 50 km/h. La dirección es la que sea que lleva el coche. Cuando pisamos el acelerador, el coche acelera y la celeridad incrementa, cuando pisamos los frenos el coche decelera, la aceleración es negativa.

Si el coche se está moviendo a celeridad fija, es fácil averiguar qué celeridad es. La abreviatura km/h lo revela: kilómetros por hora. Si el coche recorre 50 kilómetros en una hora, dividimos la distancia por el tiempo y esa es la celeridad. No necesitamos conducir durante una hora; si el coche recorre 5 kilómetros en 6 minutos, ambos, distancia y tiempo, están divididos por 10 y su relación todavía es 50 km/h. En resumen:

$$\text{celeridad} = \text{distancia recorrida dividida por el tiempo que se tarda}$$

Del mismo modo, una tasa fija de la aceleración viene dada por:

$$\text{aceleración} = \text{cambio en la celeridad dividido por el tiempo que se tarda}$$

Todo esto parece sencillo, pero surgen dificultades conceptuales cuando la celeridad o la aceleración no son fijas. Y ambas no pueden ser constantes, porque aceleración constante (distinta de cero) implica un cambio de la celeridad. Supón que conduces por una carretera comarcal, acelerando en las rectas, frenando en las curvas. Tu

celeridad no deja de cambiar, y tampoco lo hace tu aceleración. ¿Cómo podemos calcularlas en cualquier instante de tiempo dado? La respuesta pragmática es considerar un pequeño intervalo de tiempo, por ejemplo un segundo. Entonces tu celeridad instantánea a las, por ejemplo, 11:30 am es la distancia que recorres entre ese momento y un segundo después, dividida por un segundo. Funciona igual para la aceleración instantánea.

Excepto que... eso no es del todo tu celeridad instantánea. Es realmente una celeridad media, durante un intervalo de un segundo de tiempo. Hay circunstancias en las cuales un segundo es una longitud de tiempo enorme (la cuerda de una guitarra tocando Do medio vibra 440 veces cada segundo, haz un promedio de su movimiento durante un segundo completo y pensarás que está quieta). La respuesta es considerar un intervalo más corto de tiempo, una diezmillésima de segundo, quizá. Pero esto todavía no capta la celeridad instantánea. La luz visible vibra mil billones de veces (10^{15}) cada segundo, de modo que el intervalo de tiempo apropiado es menos que una milbillonésima de un segundo. E incluso entonces... bueno, siendo pedante, eso no es todavía un instante. Siguiendo esta línea de pensamiento, parece que es necesario usar un intervalo de tiempo que sea más corto que cualquier otro intervalo. Pero el único número de este tipo es 0, y esto no es útil porque en ese caso la distancia recorrida es también 0, y 0/0 no tiene sentido.

Al principio, los pioneros ignoraron estos temas y consideraron una visión pragmática. Una vez el error probable en tus medidas es mayor que la precisión que obtendrías en teoría usando intervalos de tiempo más pequeños, no sirve de nada hacerlo. Los relojes en la época de Galileo eran muy imprecisos, así que él medía el tiempo canturreando melodías para sí mismo —un músico entrenado puede subdividir una nota en intervalos muy pequeños—. Incluso entonces, cronometrar un cuerpo cayéndose es muy difícil, de modo que a Galileo se le ocurrió la trampa de ralentizar el movimiento haciendo rodar cuesta abajo bolas por una pendiente. Entonces observó la posición de la bola en los sucesivos intervalos de tiempo. Lo que encontró (estoy simplificando los números para hacer el patrón claro, pero es el mismo patrón) es que para

0, 1, 2, 3, 4, 5, 6,...

estas posiciones eran:

0, 1, 4, 9, 16, 25, 36,...

La distancia era (proporcional a) el cuadrado del tiempo. ¿Qué pasaba con la celeridad? Haciendo un promedio en los intervalos sucesivos, estas eran las diferencias:

1, 3, 5, 7, 9, 11,...

entre los sucesivos cuadrados. En cada intervalo, distinto del primero, la celeridad media incrementaba 2 unidades. Es un patrón asombroso, todavía más cuando Galileo dio con algo muy similar para docenas de mediciones con bolas de masas muy diferentes en pendientes con muchas inclinaciones diferentes.

A partir de estos experimentos y el patrón observado, Galileo dedujo algo maravilloso. La ruta de un cuerpo cayendo, o uno lanzado al aire, como una bola de cañón, es una parábola. Esto es una curva con forma de U, conocida desde la Grecia antigua. La U está al revés en este caso. Estoy ignorando la resistencia del aire, la cual cambia la forma, pues no tenía mucho efecto en las bolas rodantes de Galileo. Kepler encontró una curva relacionada, la elipse, en su análisis de las órbitas de los planetas: esto debió haberle parecido significativo a Newton también, pero esa historia debe esperar hasta el próximo capítulo.

Con solo estas series particulares de experimentos para basarse, no está claro qué principios generales subyacían en el patrón de Galileo. Newton se dio cuenta de que la fuente de los patrones eran tasas de variación. La velocidad es la tasa en la cual la posición cambia con respecto al tiempo; la aceleración es la tasa en la cual la velocidad cambia con respecto al tiempo. En las observaciones de Galileo, la posición variaba acorde al cuadrado del tiempo, la velocidad variaba linealmente y la aceleración no variaba en absoluto. Newton se dio cuenta de que con el fin de ganar una comprensión más profunda de los patrones de Galileo, y lo que

significaban para nuestra visión de la naturaleza, él tenía que asumir tasas de variación instantáneas. Cuando lo hizo, destapó el cálculo.

Podrías esperar que una idea tan importante como el cálculo se anunciase con una fanfarria de trompetas y desfiles por las calles. Sin embargo, lleva tiempo que la relevancia de ideas noveles se capte y sea apreciada, y eso pasó con el cálculo. El trabajo de Newton sobre el tema data de 1671 o antes, cuando escribió *El método de las fluxiones y las series infinitas*. No estamos seguros de la fecha porque el libro no fue publicado hasta 1736, casi una década después de su muerte. Otros cuantos manuscritos de Newton también se refieren a ideas que ahora reconocemos como cálculo diferencial e integral, las dos ramas principales del tema. Las libretas de Leibniz muestran que obtuvo sus primeros resultados importantes en cálculo en 1675, pero no publicó nada sobre el tema hasta 1684.

Después de que Newton hubiese alcanzado prominencia científica, mucho después de que ambos hombres hubiesen encontrado lo esencial del cálculo, algún amigo de Newton desató una gran controversia sin sentido, pero acalorada, sobre la prioridad, acusando a Leibniz de plagiar los manuscritos no publicados de Newton. Unos pocos matemáticos de Europa continental respondieron con contrademandas de plagio hecho por Newton. Los matemáticos ingleses y continentales estuvieron sin apenas hablarse un siglo, lo cual causó un daño enorme a los matemáticos ingleses, pero nada en absoluto a los continentales. Transformaron el cálculo en una herramienta importante de la física matemática mientras sus colegas ingleses estaban furiosos con los insultos a Newton, en vez de explotar la perspicacia de Newton. La historia es liosa y todavía es tema de discusiones en el campo académico por historiadores de la ciencia, pero en términos generales parece que Newton y Leibniz descubrieron las ideas básicas del cálculo independientemente, al menos, tan independientemente como su cultura científica y matemática común lo permitieron.

La notación de Leibniz difiere de la de Newton, pero las ideas subyacentes son más o menos idénticas. La intuición tras ellas, sin embargo, es diferente. La aproximación de Leibniz era formal, manipulando símbolos algebraicos. Newton tenía un modelo físico en el fondo de su mente, en el cual la función bajo consideración era una cantidad física que variaba con el tiempo. Esto es donde su

curioso término «fluxión» aparece, algo que fluye a medida que el tiempo pasa.

El método de Newton puede ilustrarse usando un ejemplo: una cantidad que es el cuadrado, x^2 , de otra cantidad, x . (Este es el patrón que Galileo encontró para una bola rodando: su posición es proporcional al cuadrado del tiempo que ha transcurrido. En ese caso y sería la posición y x el tiempo. El símbolo habitual para el tiempo es t , pero la coordenada estándar para el sistema de coordenadas estándar para el plano usa x e y .) Empecemos por introducir una nueva cantidad σ , que indica una pequeña cantidad en x . El correspondiente cambio en y es la diferencia:

$$(x + \sigma)^2 - x^2$$

La cual se simplifica como $2x\sigma + \sigma^2$. Por tanto, la tasa de variación (hecho un promedio en un intervalo de longitud pequeño, σ , a medida que x incrementa a $x + \sigma$) es:

$$\frac{2x\sigma + \sigma^2}{\sigma} = 2x + \sigma$$

Esto depende de σ , lo cual es lo único que podría esperarse ya que estamos haciendo el promedio de la tasa de variación en un intervalo distinto de cero. Sin embargo, si σ se va haciendo más y más pequeño, «tiende a» cero, la tasa de variación $2x + \sigma$ se acerca más y más a $2x$. Esto no depende de σ , y da la tasa de variación instantánea en x .

Leibniz realizó esencialmente los mismos cálculos, remplazando σ por dx («pequeña diferencia en x ») y definiendo dy como el correspondiente pequeño cambio en y . Cuando una variable y depende de otra variable x , la tasa de variación de y con respecto a x se llama la derivada de y . Newton escribió la derivada de y poniendo un punto sobre ella: \dot{y} . Leibniz escribió dy/dx . Para derivadas mayores, Newton usó más puntos, mientras que Leibniz escribió cosas como d^2y/dx^2 . Hoy en día, decimos que y es función de x y escribimos $y = f(x)$, pero este concepto existía de una forma rudimentaria en esa época. Usamos tanto la notación de Leibniz como una variación

de la de Newton en la cual el punto se ha remplazado con un apóstrofo, que es más fácil de imprimir: y' , y'' . También escribimos $f'(x)$ y $f''(x)$ para enfatizar que las derivadas son en sí mismas funciones. El cálculo de la derivada se llama diferenciación.

El cálculo integral —encontrar áreas— resulta ser la inversa del cálculo diferencial —encontrar pendientes—. Para ver por qué, imagina añadir una fina lámina al final del área sombreada de la figura 12.

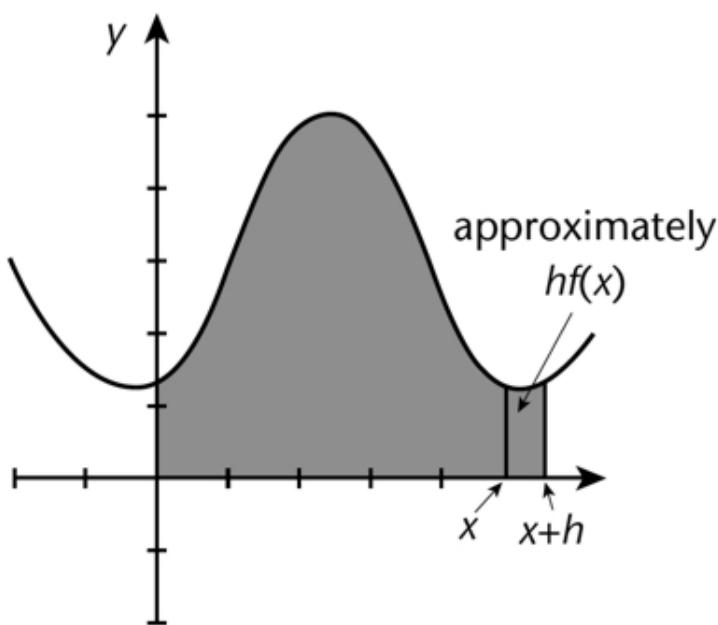


FIGURA 12. Añadiendo una pequeña lámina al área bajo la curva $y = f(x)$.

Esta lámina está muy cerca de ser un rectángulo fino y largo, de ancho σ y alto y . Su área es por tanto próxima a σy . La tasa en la cual el área cambia, con respecto a x , es la proporción $\sigma y / \sigma$, que es igual a y . Así que la derivada del área es la función original. Ambos, Newton y Leibniz, comprendieron que el modo de calcular el área, un proceso llamado integración, es lo inverso de la diferenciación en este sentido. Leibniz primero escribió la integral usando el símbolo *omn*, una abreviatura para *omnia*, suma en latín. Más tarde cambió esto a \int , una s alargada y anticuada, que también representa la «suma». Newton no tenía una notación sistemática para la integral.

No obstante, Newton sí que hizo un avance crucial. Wallis había calculado la

derivada de cualquier potencia x^a es

$$ax^a - 1$$

De modo que las derivadas de, por ejemplo, x^3 , x^4 , x^5 son $3x^2$, $4x^3$, $5x^4$. Él había ampliado este resultado para cualquier polinomio, una combinación finita de potencias, tales como $3x^7 - 25x^4 + x^2 - 3$. El truco estaba en considerar cada potencia por separado, encontrar las derivadas correspondientes y combinarlas de la misma manera. Newton se dio cuenta de que el mismo método funcionaba para series infinitas, expresiones que involucran infinidad de potencias de la variable. Esto le permitió realizar las operaciones de cálculo en muchas otras expresiones más complicadas que los polinomios.

Dada la correspondencia cercana entre las dos versiones del cálculo, difiriendo principalmente en características poco importantes de la notación, es fácil ver cómo pudo haber surgido una polémica por la prioridad. Sin embargo, la idea básica es una formulación bastante directa de la cuestión subyacente, de modo que es también fácil ver cómo Newton y Leibniz podrían haber llegado a sus versiones independientemente, a pesar de las similitudes. En cualquier caso, Fermat y Wallis se habían adelantado a ambos en muchos de sus resultados. La discusión no tiene sentido.

Una controversia más fructífera es la relativa a la estructura lógica del cálculo o, más precisamente, a la estructura ilógica del cálculo. Un crítico destacado fue el filósofo anglo-irlandés George Berkeley, obispo de Cloyne. Berkeley tuvo unas motivaciones religiosas, sentía que la visión materialista del mundo que se desarrollaba a partir del trabajo de Newton representaba a Dios como un creador distante, que se mantiene alejado de su creación una vez se pone en marcha y después la deja a su suerte, bastante diferente del Dios personal e inmanente de la creencia cristiana. De modo que atacó las inconsistencias lógicas en los fundamentos del cálculo, presumiblemente esperando desacreditar la ciencia resultante. Su ataque no tuvo un efecto apreciable en el progreso de la física matemática, por una sencilla razón: los resultados obtenidos usando el cálculo proporcionan muchísima comprensión de la naturaleza y concordaban tan bien con

los experimentos, que los fundamentos lógicos parecían poco importantes. Incluso en la actualidad, los físicos todavía toman esta visión: si funciona, ¿por qué preocuparse por nimiedades lógicas?

Berkeley defendía que no tiene lógica mantener que una pequeña cantidad (la σ de Newton y la dx de Leibniz) es distinta de cero para la mayoría de los cálculos y luego fijarla en cero, si anteriormente has dividido tanto numerador como denominador de una fracción por esa misma cantidad. La división por cero no es una operación aceptable en aritmética, porque no tiene un significado inequívoco. Por ejemplo, $0 \times 1 = 0 \times 2$, ya que ambas son 0, pero si dividimos ambos lados de la ecuación por 0, obtenemos $1 = 2$, lo cual es falso.¹⁰ Berkeley publicó sus críticas en 1734 en un folleto *The Analyst, a Discourse Addressed to an Infidel Mathematician* (El análisis, un discurso dirigido a un matemático infiel).

Newton hizo, en realidad, intentos de ordenar la lógica, apelando a una analogía física. Veía o no como una cantidad fija, sino como algo que fluía —variaba con el tiempo— acercándose más y más a cero sin realmente alcanzarlo. La derivada era también definida por una cantidad que fluía: la tasa de variación en y para ese x . Esta tasa también fluía hacia algo, pero nunca llegaba a ello, ese algo era la tasa de variación instantánea, la derivada de y respecto a x . Berkeley descartó esta idea por ser «el fantasma de una cantidad difunta».

Leibniz también tuvo un crítico persistente, el geómetra Bernard Nieuwentijt, quien puso sus críticas en papel impreso en 1694 y 1695. Leibniz no había ayudado a su caso tratando de justificar su método en términos de «infinitesimales», un término abierto a los malentendidos. Sin embargo, explicó que lo que quería decir con este término no era una cantidad fija distinta de cero que podría ser arbitrariamente pequeña (lo cual no tiene sentido lógico), sino una cantidad variable distinta de cero que puede hacerse arbitrariamente pequeña. Las defensas de Newton y Leibniz fueron en esencia idénticas. Para sus opositores, ambas debieron de haber sonado como artimañas verbales.

Afortunadamente, los físicos y matemáticos de la época no esperaron a que se

¹⁰ La división entre cero lleva a pruebas falaces. Por ejemplo, podemos «probar» que todos los números son cero. Sea $a = b$. Por lo tanto $a^2 = ab$, así que $a^2 - b^2 = ab - b^2$. Factorizamos para obtener $(a + b)(a - b) = b(a - b)$. Dividimos por $(a - b)$ para deducir que $a + b = b$. Por lo tanto $a = 0$. El error es la división por $(a - b)$, que es 0 ya que consideramos $a = b$.

averiguasen los fundamentos lógicos del cálculo antes de aplicarlo a las fronteras de la ciencia. Tenían una manera alternativa de estar seguros de que estaban haciendo algo sensato: compararlo con las observaciones y los experimentos. El propio Newton inventó el cálculo para precisamente este propósito. Obtuvo leyes para cómo los cuerpos se mueven cuando se les aplica una fuerza y combinó esto con una ley para la fuerza ejercida por la gravedad, para explicar muchos enigmas sobre los planetas y otros cuerpos del Sistema Solar. Su ley de la gravedad es una ecuación tan fundamental en física y astronomía que merece, y tiene, un capítulo solo para ella (el próximo). Su ley del movimiento —estrictamente, un sistema de tres leyes, una de las cuales contiene la mayoría del contenido matemático— nos dirige bastante directamente al cálculo.

Irónicamente, cuando Newton publicó estas leyes y sus aplicaciones científicas en su *Principia*, eliminó todo rastro de cálculo y lo remplazó por argumentos de geometría clásica. Probablemente pensó que la geometría sería más adecuada para la audiencia futura y, si lo hizo, estaba casi con seguridad en lo correcto. Sin embargo, muchas de sus pruebas geométricas estaban o motivadas por el cálculo, o dependían del uso de las técnicas de cálculo para determinar las respuestas correctas sobre las cuales la estrategia de la prueba geométrica recae. Esto es especialmente claro, a los ojos de alguien en la actualidad, en su tratamiento de lo que llamó «cantidades generadas» en el Libro II de *Principia*. Estas cantidades que crecen y decrecen por «movimiento continuo o flujo», las fluxiones de su libro no publicado. Hoy en día las llamaríamos funciones continuas (es más, diferenciables). En lugar de las operaciones explícitas del cálculo, Newton sustituyó un método geométrico de «razones primeras y últimas». Su lema (el nombre dado a un resultado matemático auxiliar que se usa repetidamente pero no tiene un interés intrínseco por sí mismo) de apertura descubre el pastel, porque define la igualdad de estas cantidades que fluyen como:

Las cantidades, y las razones de las cantidades, las cuales en un tiempo finito convergen continuamente a la igualdad, y antes del final de este tiempo se aproximan más cerca la una de la otra que cualquier diferencia dada, se vuelven finalmente iguales.

En *Never at rest*, el biógrafo de Newton, Richard Westfall, explica cómo de radical y novedoso era este lema: «cualquiera que fuera el lenguaje, el concepto... era totalmente moderno, la geometría clásica no había contemplado nada como eso».¹¹ Los contemporáneos de Newton debieron verse en apuros para averiguar lo que Newton estaba insinuando. Berkeley probablemente nunca lo hizo, porque, como veremos en breve, contiene la idea básica necesaria para deshacerse de su objeción.

El cálculo, entonces, estaba jugando un papel influyente entre bastidores en *Principia*, pero no aparecía en escena. Sin embargo, tan pronto como el cálculo asomó la cabeza tras las cortinas, los sucesores intelectuales de Newton rápidamente aplicaron la ingeniería inversa a sus procesos de pensamiento. Reformularon sus ideas principales en el lenguaje del cálculo, porque este proporcionaba un marco más natural y más poderoso, y dispuesto a conquistar el mundo científico.

La pista ya era visible en las leyes de movimiento de Newton. La pregunta que llevó a Newton a esas leyes era filosófica: ¿qué hace que un cuerpo se mueva o cambie su estado de movimiento? La respuesta clásica era de Aristóteles: un cuerpo se mueve porque una fuerza es aplicada sobre él y esto afecta a su velocidad. Aristóteles también afirmó que para mantener un cuerpo en movimiento, la fuerza debe seguir aplicándosele. Puedes probar las afirmaciones de Aristóteles colocando un libro o un objeto similar en una mesa. Si empujas el libro, empieza a moverse, si continuas empujándolo con la misma fuerza, continua deslizándose sobre la mesa a una velocidad aproximadamente constante. Si dejas de empujarlo, el libro deja de moverse. Así que la visión de Aristóteles parece estar acorde con el experimento. Sin embargo, esa concordancia es superficial, porque el empuje no es la única fuerza que actúa sobre el libro. Hay también fricción con la superficie de la mesa. Además, cuanto más rápido se mueve el libro, mayor se hace la fricción —al menos mientras la velocidad del libro permanece razonablemente pequeña—. Cuando el libro se mueve a ritmo constante por la mesa, impulsado por una fuerza constante, la resistencia de la fricción anula la fuerza aplicada, y la fuerza total que actúa en el cuerpo es en realidad cero.

¹¹ Richard Westfall. *Never at Rest*, Cambridge University Press, Cambridge 1980, p. 425.

Newton, siguiendo las ideas previas de Galileo y Descartes, se dio cuenta de eso. La teoría del movimiento que resultó es muy diferente de la de Aristóteles. Las tres leyes de Newton son:

1. **Primera ley:** todo cuerpo continúa en su estado de reposo, o de movimiento uniforme en una línea recta, a menos que sea obligado a cambiar ese estado por una fuerza ejercida sobre él.
2. **Segunda ley:** el cambio de movimiento es proporcional a la fuerza motriz ejercida, y se hace en la dirección de la línea recta en la cual se ejerce la fuerza. (La constante de proporcionalidad es la inversa de la masa del cuerpo, esto es, 1 dividido por esa masa.)
3. **Tercera ley:** para toda acción, hay siempre una reacción opuesta igual.

La primera ley contradice a Aristóteles explícitamente. La tercera ley dice que si empujas algo, eso te empuja a ti. La segunda ley es donde el cálculo aparece. Con «cambio de movimiento» Newton quería decir la tasa en la cual la velocidad del cuerpo cambia: su aceleración. Esto es la derivada de la velocidad con respecto al tiempo, y la derivada segunda del desplazamiento. De modo que la segunda ley de movimiento de Newton especifica la relación entre la posición de un cuerpo y las fuerzas que actúan en él, en la forma de una ecuación diferencial:

$$\text{Derivada segunda de la posición} = \text{fuerza/masa}$$

Para encontrar la posición, tenemos que resolver esta ecuación, deduciendo la posición a partir de su derivada segunda.

Esta línea de pensamiento nos lleva a una explicación simple de las observaciones de Galileo para las bolas que rodaban. El punto crucial es que la aceleración de la bola es constante. Yo afirmé esto previamente usando un cálculo burdo pero efectivo aplicado en intervalos discretos de tiempo, ahora lo podemos hacer del modo adecuado, permitiendo al tiempo que varíe continuamente. La constante está relacionada con la fuerza de gravedad y en el ángulo de la pendiente, pero aquí no necesitamos tanto detalle. Supongamos que la aceleración constante es a . Integrando la función correspondiente, la velocidad bajando la pendiente en un

momento t es $at + b$, donde b es la velocidad cuando el tiempo es cero. Integrando de nuevo, la posición bajando la pendiente es

$$\frac{1}{2}at^2 + bt + c$$

donde c es la posición en el instante cero. En el caso especial $a = 2$, $b = 0$, $c = 0$, las posiciones sucesivas encajan en mi ejemplo simplificado: la posición en ese momento t es t^2 . Un análisis similar recupera el resultado más importante de Galileo: la ruta que sigue un proyectil es una parábola.

Las leyes de movimiento de Newton no solo proporcionaron un modo de calcular cómo los cuerpos se mueven. Nos llevaron a principios físicos profundos y generales. Primordial entre ellos son las «leyes de conservación», que nos dicen que cuando un sistema de cuerpos, no importa cómo de complicado sea, se mueve, ciertas características de ese sistema no cambian. Entre el tumulto del movimiento, unas pocas cosas ni se inmutan. Tres de estas cantidades que se conservan son la energía, el momento y el momento angular.

La energía puede definirse como la capacidad para hacer un trabajo. Cuando un cuerpo se levanta a cierta altura, en contra de la fuerza (constante) de la gravedad, el trabajo hecho para poner ahí es proporcional a la masa del cuerpo, la fuerza de gravedad, y la altura a la cual se levanta. A la inversa, si luego soltamos el cuerpo, puede realizar la misma cantidad de trabajo cuando cae a su altura original. Este tipo de energía se llama energía potencial.

Por sí misma, la energía potencial no sería terriblemente interesante, pero hay una bonita consecuencia matemática de la segunda ley de movimiento de Newton que nos lleva a un segundo tipo de energía: energía cinética. A medida que un cuerpo se mueve, tanto su energía potencial como su energía cinética cambian. Pero el cambio en una compensa exactamente el cambio en la otra. A medida que el cuerpo desciende por la gravedad, se va acelerando. La ley de Newton nos permite calcular cómo cambia su velocidad con la altura. Resulta que el descenso en la energía potencial es exactamente igual a la mitad de la masa multiplicada por el cuadrado de la velocidad. Si le damos a esa cantidad un nombre, energía cinética, entonces la

energía total, potencial más cinética, se conserva. Esta consecuencia matemática de las leyes de Newton prueba que las máquinas con un movimiento perpetuo son imposibles: ningún instrumento mecánico puede no detenerse indefinidamente y hacer el trabajo sin alguna entrada externa de energía.

Físicamente, la energía cinética y la potencial parecen ser dos cosas diferentes; matemáticamente, podemos intercambiar la una con la otra. Es como si el movimiento de algún modo convirtiera la energía potencial en cinética. «Energía», como un término aplicable a ambas, es una abstracción oportuna, cuidadosamente definida de modo que se conserva. Como una analogía, los viajeros pueden convertir euros en dólares. El intercambio de moneda tiene tablas con tasas de cambio, dicho eso, por ejemplo, 1 euro es igual al valor de 1,3061 dólares. También deducen una suma de dinero para ellos mismos. Sujeto a las tecnicidades de las comisiones del banco, etcétera, el valor monetario total envuelto en la transacción se supone que se compensa: el viajero obtiene exactamente la cantidad en dólares que corresponde a su suma original en euros, menos varias deducciones. No obstante, no hay un algo físico en el interior de los billetes que de algún modo intercambie un billete de euros en uno de dólares y algunas monedas. Lo que hace el intercambio es la convención humana de que estos elementos, en particular, tiene un valor monetario.

La energía es un nuevo tipo de cantidad «física». Desde un punto de vista newtoniano, las cantidades tales como posición, tiempo, velocidad, aceleración y masa tiene interpretaciones físicas directas. Puedes medir la posición con una regla, el tiempo con un reloj, la velocidad y la aceleración usando los aparatos correspondientes, y la masa con una balanza. Pero no puedes medir la energía usando un medidor de energía. De acuerdo, puedes medir ciertos tipos específicos de energía. La energía potencial es proporcional a la altura, así que una regla bastará si conoces la fuerza de gravedad. La energía cinética es la mitad de la masa multiplicada por el cuadrado de la velocidad: usa una pesa y un velocímetro. Pero la energía, como concepto, no es tanto un algo físico como una ficción conveniente que ayuda a equilibrar las reglas de la mecánica.

El momento, la segunda cantidad que se conserva, es un concepto simple: la masa multiplicada por la velocidad. Surge cuando hay varios cuerpos. Un ejemplo

importante es un cohete; aquí un cuerpo es el cohete y el otro el combustible. Como el combustible es expulsado por el motor, la conservación del momento implica que el cohete debe moverse en la dirección opuesta. Esto muestra cómo un cohete funciona en el vacío.

El momento angular es similar, pero está relacionado con el giro más que con la velocidad. Es también vital en el estudio de cohetes, de hecho en toda la mecánica, terrestre o celeste. Uno de los mayores enigmas sobre la Luna es su gran momento angular. La teoría actual es que la Luna fue salpicada cuando un planeta del tamaño de Marte golpeó la Tierra alrededor de hace 4.500 millones de años. Esto explica el momento angular, y hasta hace poco era generalmente aceptado, pero ahora parece que la Luna tiene demasiada agua en sus rocas. Un impacto como el planteado debería haber hecho que se evaporase mucha agua.¹² Cualquiera que sea el resultado final, el momento angular es de vital importancia aquí.

El cálculo funciona. Soluciona problemas en física y geometría, da las respuestas correctas. Incluso nos dirige hacia nuevos y fundamentales conceptos físicos como la energía y el momento. Pero eso no responde a la objeción del obispo Berkeley. El cálculo tiene que funcionar como las matemáticas, no tan solo estar de acuerdo con la física. Tanto Newton como Leibniz entendieron que σ o dx no pueden ser ambos, cero y distinto de cero. Newton, cansado de escapar de la trampa lógica, empleó la imagen física de la fluxión. Leibniz habló de infinitesimales. Ambos se refirieron a cantidades que se aproximan a cero sin llegar a él nunca. Pero ¿qué son estos elementos? Irónicamente, la burla de Berkeley sobre «fantasmas de cantidades difuntas» se acercaba a la resolución de este asunto, pero lo que no tuvo en cuenta, y en lo que tanto Newton como Leibniz hicieron énfasis, fue cómo las cantidades fallecían. Hazlas fallecer en el modo correcto y puedes dejar un fantasma perfectamente bien formado. Si Newton y Leibniz hubiesen formulado su intuición en un lenguaje matemático riguroso, puede que Berkeley hubiese entendido qué era a lo que se referían.

La cuestión central es una que Newton no logró responder explícitamente porque parecía demasiado obvia. Recuerda que en el ejemplo donde $y = x^2$, Newton

¹² Erik H. Hauri, Thomas Weinreich, Alberto E. Saal, Malcolm C. Rutherford y James A. Van Orman. «High pre-eruptive water contents preserved in lunar melt inclusions», *Science Online* (26 de mayo de 2011) 1204626. [DOI:10.1126/science.1204626]. Sus resultados resultan polémicos.

obtenía la derivada como $2x + \sigma$, y luego afirmaba que como σ tiende hacia cero, $2x + \sigma$ tiende hacia $2x$. Esto puede parecer obvio, pero no podemos fijar $\sigma = 0$ para probarlo. Es cierto que obtenemos el resultado correcto haciendo eso, pero esto es una pista falsa.¹³ En *Principia* Newton se desliza alrededor de este tema totalmente, remplazando $2x + \sigma$ por su «razón primera» y $2x$ por su «razón última». Pero la clave real para avanzar es abordar el tema de frente. ¿Cómo sabemos que cuando más cerca de cero está σ , más cerca de $2x$ está $2x + \sigma$? Puede parecer un punto bastante pedante, pero si usase ejemplos más complicados, la respuesta correcta podría no parecer tan plausible.

Cuando los matemáticos volvieron a la lógica del cálculo, se dieron cuenta de que esta cuestión aparentemente simple era el quid de la cuestión. Cuando decimos que σ se aproxima a cero, queremos decir que para un número dado positivo y distinto de cero, σ puede escogerse de modo que sea más pequeño que ese número. (Esto es obvio: sea σ la mitad de ese número, por ejemplo.) De manera similar, cuando decimos que $2x + \sigma$ se aproxima a $2x$, queremos decir que la diferencia se aproxima a cero, en el sentido anterior. Como resulta que la diferencia es el propio σ en este caso, esto es todavía más obvio: cualquiera que sea el significado de «se aproxima a cero», claramente σ se aproxima a cero cuando σ se aproxima a cero. Una función más complicada que el cuadrado requeriría un análisis más complicado. La respuesta a esta cuestión clave es establecer el proceso en términos matemáticos formales, evitando por completo ideas como «flujo». Este gran paso adelante llegó con el trabajo del matemático y teólogo de Bohemia, Bernard Bolzano, y el matemático alemán Karl Weierstrass. El trabajo de Bolzano data de 1816, pero no se apreció hasta alrededor de 1870 cuando Weierstrass amplió su formulación a funciones complejas. Su respuesta a Berkeley fue el concepto de un límite. Daré la definición con palabras y dejaré la versión con símbolos para las notas.¹⁴ Digamos que una función $f(h)$ de una variable h tiende a un límite L a medida que h tiende a cero, dado cualquier número positivo distinto de cero, la

¹³ Sin embargo, no es una coincidencia. Funciona para cualquier función diferenciable: una con una derivada continua. Esto incluye todos los polinomios y todas las series de potencias convergentes, tales como las funciones logarítmicas, exponenciales y varias trigonométricas.

¹⁴ La definición moderna es: una función $f(h)$ tiende al límite L a medida que h tiende a cero si para cualquier $\epsilon > 0$ existe $\sigma > 0$ tal que $|h| < \sigma$ implica que $|f(h) - L| < \epsilon$. Usando cualquier $\epsilon > 0$ se evita referirse a algo fluyendo o haciéndose más pequeño; aborda todos los posibles valores de una vez.

diferencia entre $f(h)$ y L puede hacerse más pequeña que ese número escogiendo valores de h distintos de cero suficientemente pequeños. En símbolos:

$$\lim_{h \rightarrow 0} f(h) = L$$

La idea en el núcleo del cálculo es aproximar la tasa de variación de una función en un intervalo pequeño h , y entonces tomar el límite a medida que h tiende a cero. Para una función general $y = f(x)$ este procedimiento lleva a la ecuación que adorna la apertura de este capítulo, pero usando una variable general x en vez del tiempo:

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h}$$

En el numerador vemos el cambio en f , en el denominador está el cambio en x . Esta ecuación define la derivada $f'(x)$ de manera única, siempre y cuando el límite exista. Esto tiene que probarse para cualquier función que consideremos; el límite sí que existe para la mayoría de las funciones estándar (cuadrado, cubo, potencias mayores, funciones logarítmicas, exponenciales, trigonométricas).

En ningún punto en el cálculo dividimos entre cero, porque nunca fijamos $h = 0$. Además, nada aquí fluye realmente. Lo que importa es el rango de valores que h puede asumir, no cómo se mueve a través de ese rango. De modo que la caracterización sarcástica de Berkeley es en realidad certera. El límite L es el fantasma de la cantidad difunta —mi h , la σ de Newton—. Pero la manera de fallecer la cantidad —aproximándose a cero, no alcanzándolo— nos lleva a un fantasma perfectamente sensato y bien definido lógicamente.

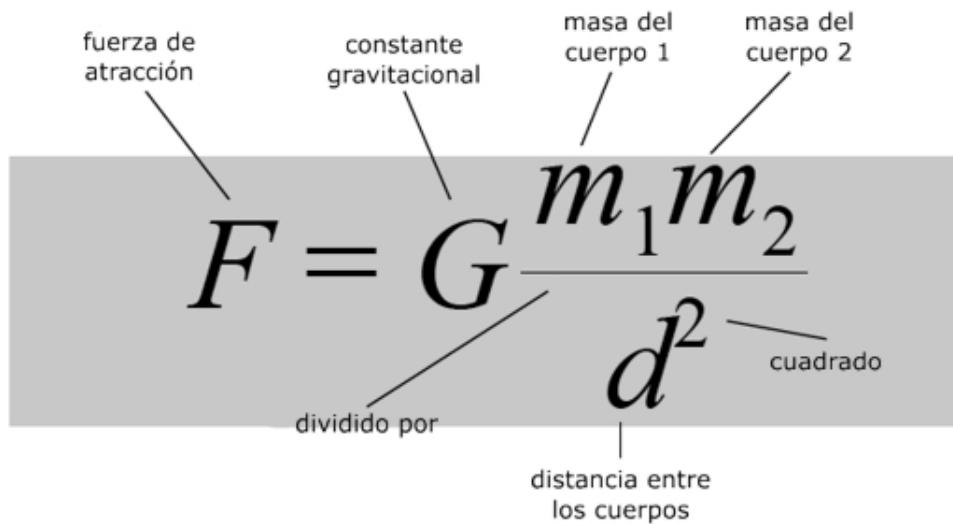
Desde ese momento el cálculo tenía unas bases lógicas sólidas. Se merecía, y adquirió, un nuevo nombre para reflejar su nuevo estatus: análisis.

Hacer un listado de todos los modos en los que se puede aplicar el cálculo es tan factible como hacer una lista de lo que todo en el mundo depende de usar un destornillador. A un nivel computacional simple, las aplicaciones del cálculo incluyen

encontrar la longitud de curvas, áreas de superficies y formas complicadas, volúmenes de sólidos, valores máximos y mínimos y centros de masa. En conjunción con las leyes de la mecánica, el cálculo nos dice cómo averiguar la trayectoria de un cohete espacial, las tensiones en una roca en una zona de subducción que podrían producir un terremoto, el modo en que un edificio vibrará si se produce el terremoto, el modo en que un coche da botes arriba y abajo en su suspensión, el tiempo que tarda una infección bacteriológica en extenderse, el modo en que una herida quirúrgica se cura, y las fuerzas que actúan en un puente en suspensión cuando hace mucho viento.

Muchas de estas aplicaciones provienen de la profunda estructura de las leyes de Newton: son modelos de la naturaleza formulados como ecuaciones diferenciales. Estas son ecuaciones que implican derivadas de una función desconocida, y se necesitan técnicas del cálculo para resolverlas. No diré nada más por ahora, porque cada capítulo del 8 en adelante implica al cálculo explícitamente, principalmente bajo la apariencia de ecuaciones diferenciales. La única excepción es el capítulo 15 sobre la teoría de la información, e incluso otros desarrollos que no menciono también involucran al cálculo. Como el destornillador, el cálculo es una herramienta simple e indispensable en la caja de herramientas de ingenieros y científicos. Más que cualquier otra técnica matemática, ha creado el mundo moderno.

Capítulo 4
El sistema del mundo
Ley de gravitación universal

$$F = G \frac{m_1 m_2}{d^2}$$


The diagram shows the formula for universal gravitation: $F = G \frac{m_1 m_2}{d^2}$. Lines with arrows point from the labels to the corresponding parts of the equation:

- fuerza de atracción (force of attraction) points to F .
- constante gravitacional (gravitational constant) points to G .
- masa del cuerpo 1 (mass of body 1) points to m_1 .
- masa del cuerpo 2 (mass of body 2) points to m_2 .
- cuadrado (square) points to the exponent 2 in d^2 .
- distancia entre los cuerpos (distance between bodies) points to d .
- dividido por (divided by) points to the fraction bar.

¿Qué dice?

Determina la fuerza de atracción gravitacional entre dos cuerpos en términos de sus masas y la distancia entre ellos.

¿Por qué es importante?

Puede aplicarse a cualquier sistema de cuerpos interactuando a través de la fuerza de gravedad, como el Sistema Solar. Nos dice que su movimiento está determinado por una sencilla ley matemática.

¿Qué provocó?

Predicción precisa de eclipses, órbitas planetarias, la reaparición de los cometas, la rotación de las galaxias. Satélites artificiales, mediciones de la Tierra, el telescopio Hubble, observaciones de erupciones solares. Sondas interplanetarias, vehículos motorizados a Marte, comunicaciones vía satélite y la televisión, el Sistema de Posicionamiento Global (GPS).

Las leyes de movimiento captan la relación entre las fuerzas que actúan en un cuerpo y cómo se mueve en respuesta a estas fuerzas. El cálculo proporciona técnicas matemáticas para resolver las ecuaciones resultantes. Se necesita un

ingrediente más para aplicar las leyes: especificar las fuerzas. El aspecto más ambicioso del *Principia* de Newton era hacer precisamente eso para los cuerpos del Sistema Solar: el Sol, los planetas, satélites, asteroides y cometas. La ley de gravitación universal de Newton sintetiza, en una sencilla fórmula matemática, milenios de observaciones astronómicas y teorías. Explica muchas características misteriosas del movimiento planetario e hizo posible predecir los movimientos futuros del Sistema Solar con gran precisión. La teoría de la relatividad general de Einstein finalmente suplanta la teoría newtoniana de la gravedad, en lo que a física fundamental se refiere, pero para casi todos los propósitos prácticos la aproximación newtoniana más simple todavía impera. En la actualidad, las agencias espaciales mundiales, como NASA y ESA, todavía usan las leyes de movimiento y gravitación de Newton para averiguar las trayectorias más efectivas para las naves espaciales.

Fue la ley de gravitación universal, sobre todas las demás, la que justificó su subtítulo: *El sistema del mundo*. Esta ley demostraba el enorme poder de las matemáticas para encontrar patrones escondidos en la naturaleza y revelar simplicidades escondidas tras las complejidades del mundo. Y con el tiempo, a medida que los matemáticos y astrónomos preguntaban cuestiones más difíciles, revelar las complejidades escondidas implícitas en la sencilla ley de Newton. Para apreciar lo que Newton logró, debemos primero remontarnos en el tiempo, para ver cómo culturas anteriores veían las estrellas y los planetas.

Los humanos han estado observando el cielo nocturno desde el amanecer de la historia. Sus impresiones iniciales habrían sido una dispersión aleatoria de puntos de luz brillantes, pero se darían cuenta pronto de que a través de este fondo el brillante orbe de la Luna trazaba un recorrido regular, cambiando de forma a medida que lo hacía. También habrían visto que la mayoría de esos minúsculos destellos brillantes de luz permanecían en los mismos modelos relativos, los cuales ahora llamamos constelaciones. Las estrellas se mueven a través del cielo nocturno, pero se mueven como una única unidad rígida, como si las constelaciones estuviesen pintadas en un bol gigante que rota.¹⁵ Sin embargo, un pequeño número

¹⁵ El libro del Génesis se refiere al «firmamento». La mayoría de los académicos creen que esto proviene de la antigua creencia hebrea de que las estrellas eran luces minúsculas fijadas a una bóveda del Cielo, con la forma de

de estrellas se comportan de un modo bastante diferente, parece que deambulan alrededor del cielo. Sus recorridos son bastante complicados y algunas parecen que regresan sobre sí mismas de tanto en tanto. Estos son los planetas, una palabra que viene del término griego para «vagabundos». En la Antigüedad reconocieron 5 de ellos, ahora llamados Mercurio, Venus, Marte, Júpiter y Saturno. Se mueven en relación con las estrellas fijas a diferentes velocidades, y Saturno es el más lento. Otros fenómenos celestes eran incluso más enigmáticos. De tanto en tanto un cometa aparecía, como si viniese de la nada, siguiendo una estela larga y curva. «Estrellas fugaces» parecerían caer del cielo, como si se hubiesen despegado del bol en el que estaban. No es de extrañar que los primeros humanos atribuyesen estas irregularidades en los cielos a los caprichos de seres sobrenaturales.

Las regularidades podrían resumirse en términos tan obvios que pocos habrían alguna vez soñado con discutirlas. El Sol, las estrellas y los planetas giran alrededor de una Tierra inmóvil. Esto es lo que parece, así es como se siente, así que así es como debe de ser. En la Antigüedad, el cosmos era geocéntrico, la Tierra era el centro. Una solitaria voz cuestionó lo obvio: Aristarco de Samos. Usando principios geométricos y observaciones, Aristarco calculó los tamaños de la Tierra, el Sol y la Luna. Alrededor del año 270 a.C., expuso la primera teoría heliocéntrica: la Tierra y los planetas giran alrededor del Sol. Su teoría rápidamente cayó en desgracia y no resurgió durante casi 2.000 años.

En la época de Ptolomeo, un romano que vivió en Egipto alrededor del 120 d.C., los planteas habían sido domesticados. Sus movimientos no eran caprichosos, sino predecibles. El *Almagesto* (El gran tratado) de Ptolomeo proponía que vivimos en un universo geocéntrico en el cual todo literalmente gira alrededor de la humanidad en combinaciones complejas de círculos llamadas epiciclos, apoyados en esferas de cristal gigantes. Su teoría era errónea, pero los movimientos que predijo eran lo suficientemente precisos para que los errores no se detectasen durante siglos. El sistema de Ptolomeo tenía una atracción filosófica adicional: representaba el cosmos en términos de figuras geométricas perfectas: esferas y círculos. Continuaba la

una semiesfera. Esto es lo que parece el cielo nocturno, el modo en que nuestros sentidos visuales responden a objetos distantes hace que las estrellas aparenten estar a más o menos la misma distancia de nosotros. Muchas culturas, especialmente en Oriente Medio y el lejano Oriente, pensaban en el cielo como un bol que giraba lentamente.

tradición pitagórica. En Europa, la teoría de Ptolomeo permaneció sin cambios durante 1.400 años.

Mientras Europa perdía el tiempo, nuevos avances científicos se llevaban a cabo en otros lugares, especialmente en el mundo árabe, China e India. En el 499, el astrónomo hindú Aryabhata expuso un modelo matemático para el Sistema Solar en el cual la Tierra giraba sobre su eje y los períodos de órbitas planetarias se establecían en relación con el Sol. En el mundo islámico, Alhazen escribió una crítica punzante a la teoría ptolemaica, aunque esta probablemente no se centraba en su naturaleza geocéntrica. Alrededor del año 1000, Abu Rayhan Biruni aportó reflexiones serias sobre la posibilidad de un Sistema Solar heliocéntrico, con la Tierra girando sobre su eje, pero finalmente optó por la ortodoxia de la época, una Tierra fija. Alrededor del 1300, Najm al-Din al-Qazwini al-Katibi propuso una teoría heliocéntrica, pero pronto cambió de opinión.

El gran paso adelante llegó con el trabajo de Nicolás Copérnico, publicado en 1543 como *De Revolutionibus Orbium Coelestium* (Sobre las revoluciones de las esferas celestes). Hay evidencias, en particular el hecho de diagramas casi idénticos etiquetados con las mismas letras, que sugieren que Copérnico estaba, como mínimo, influenciado por al-Kabiti, pero fue mucho más lejos. Propuso un sistema explícitamente heliocéntrico, argumentó que encajaba mejor y más económicamente con las observaciones que la teoría geocéntrica de Ptolomeo y expuso algunas de sus implicaciones filosóficas. Primordial entre ellas era el pensamiento novedoso de que los humanos no eran el centro de las cosas. La Iglesia cristiana vio sus sugerencias como contrarias a la doctrina e hizo lo que pudo para disuadirlo. El heliocentrismo explícito era una herejía.

A pesar de todo prevaleció, porque las evidencias eran muy fuertes. Aparecieron teorías heliocéntricas nuevas y mejores. Luego las esferas se descartaron totalmente en favor de una forma diferente de la geometría clásica: la elipse. Las elipses son formas ovaladas y pruebas indirectas sugieren que se estudiaron por primera vez en la geometría griega por Menecmo alrededor del 350 a.C., junto con las hipérbolas y las parábolas, como secciones de un cono (figura 13). Se dice que Euclides habría escrito cuatro libros sobre las secciones cónicas, aunque nada ha sobrevivido si lo hizo, y Arquímedes investigó algunas de sus propiedades. La

investigación hecha por los griegos sobre el tema alcanzó su clímax alrededor del 240 a.C. con los ocho volúmenes de *Secciones cónicas* de Apolonio de Perga, quien encontró un modo de definir estas curvas simplemente en el plano, evitando la tercera dimensión. Sin embargo, la visión pitagórica de que los círculos y las esferas alcanzaban un grado mayor de perfección que las elipses y otras curvas más complejas persistía.

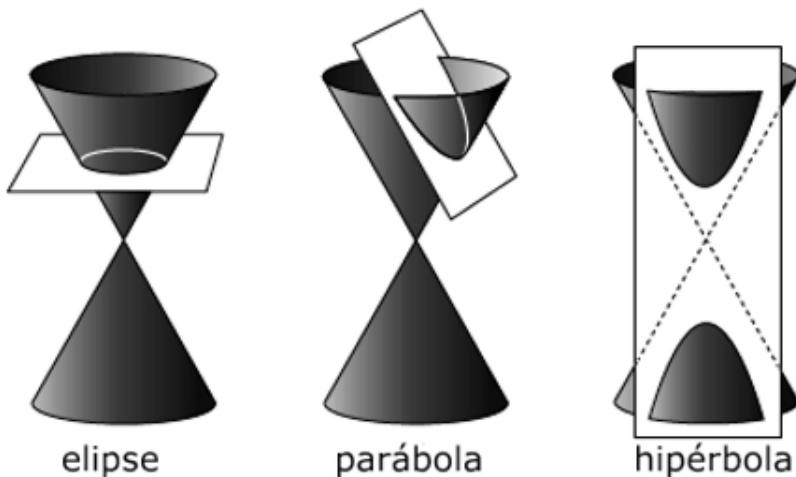


FIGURA 13. Secciones cónicas.

Las elipses consolidaron su papel en la astronomía alrededor del 1600, con el trabajo de Kepler. Sus intereses astronómicos empezaron en la infancia, con seis años fue testigo del gran cometa de 1577,¹⁶ y tres años más tarde vio un eclipse de la Luna. En la Universidad de Tubinga, Kepler demostró un gran talento para las matemáticas y le sacó un uso rentable haciendo horóscopos. En esa época, matemática, astronomía y astrología con frecuencia iban juntas. Combinó un embriagador nivel de misticismo con una atención sensata al detalle matemático. Un ejemplo típico es su *Mysterium Cosmographicum* (El misterio cosmográfico), una defensa vehemente del sistema heliocéntrico publicada en 1569. Combina unos conocimientos claros de la teoría de Copérnico con lo que a ojos actuales es una especulación muy extraña que relaciona las distancias de los planetas conocidos a

¹⁶ El Gran Cometa de 1577 no es el cometa Halley, sino otro de importancia histórica, ahora llamado C/1577 V1. Fue visible a la vista, sin necesidad de telescopio, en 1577 d.C. Brahe observó el cometa y dedujo que los cometas se encontraban fuera de la atmósfera terrestre. El cometa está actualmente a alrededor de 24.000 millones de kilómetros del Sol.

partir del Sol con los sólidos regulares. Durante mucho tiempo Kepler consideró este descubrimiento como uno de los más importantes, revelando los planes del Creador para el universo. Vio sus investigaciones posteriores, las cuales en la actualidad consideramos mucho más importantes, como meras elaboraciones de este plan básico. En la época, una ventaja de la teoría era que explicaba por qué había justamente seis planetas (de Mercurio a Saturno). Entre estas seis órbitas hay cinco huecos, uno por cada sólido regular. Con el descubrimiento de Urano y, más tarde, Neptuno y Plutón (hasta su reciente degradación de su estatus como planeta) esta característica rápidamente se convirtió en un fallo nefasto.

La duradera contribución de Kepler tiene sus raíces en su empleo con Tycho Brahe. Se encontraron por primera vez en 1600. Después de una estancia de dos meses y una discusión acalorada, Kepler negoció un salario aceptable. Después de una racha de problemas en su ciudad natal, Graz, se mudó a Praga, asistiendo a Tycho en el análisis de sus observaciones planetarias, especialmente de Marte. Cuando Tycho inesperadamente falleció en 1601, Kepler asumió su trabajo como matemático imperial para Rodolfo II. Su papel principal era hacer los horóscopos del imperio, pero también tenía tiempo para continuar su análisis de la órbita de Marte. Siguiendo principios epicíclicos tradicionales refinó su modelo hasta el punto en el que sus errores, comparados con la observación, eran normalmente unos escasos dos minutos de arco, el típico error en las propias observaciones. Sin embargo, no se detuvo ahí porque a veces los errores eran mayores, hasta ocho minutos de arco.

Su búsqueda finalmente lo llevó a dos leyes del movimiento de los planetas, publicadas en *Astronomia Nova* (Una nueva astronomía). Durante muchos años, había intentado encajar la órbita de Marte en un ovoide —una curva con forma de huevo, más fina en un extremo que en otro— sin éxito. Quizá esperaba que la órbita fuese más curvada cuando estuviese más cerca del Sol. En 1605, a Kepler se le ocurrió intentarlo con una elipse, redondeada por igual en ambos extremos, y cuál fue su sorpresa al ver que encajaba mucho mejor. Concluyó que todas las órbitas planetarias eran elipses, su primera ley. Su segunda ley describía cómo el planeta se mueve a lo largo de su órbita, afirmando que los planetas barren áreas iguales en tiempos iguales. El libro se publicó en 1609. Kepler dedicó entonces

mucho de su esfuerzo a preparar varias tablas astronómicas, pero volvió a las regularidades de las órbitas planetarias en 1619 en su *Harmonices Mundi* (La armonía del mundo). Este libro tenía algunas ideas que ahora encontramos extrañas, por ejemplo que los planetas emitían sonidos musicales a medida que rotaban alrededor del Sol. Pero también incluye su tercera ley: los cuadrados de los períodos orbitales son proporcionales a los cubos de sus distancias al Sol.

Las tres leyes de Kepler fueron prácticamente sepultadas entre una masa de misticismo, simbolismo religioso y especulaciones filosóficas. Pero representaron un salto hacia delante gigantesco, llevando a Newton a uno de los más grandes descubrimientos científicos de todos los tiempos.

Newton obtuvo su ley de la gravedad a partir de las tres leyes del movimiento de los planetas de Kepler. Afirma que toda partícula en el universo atrae a todas las otras partículas con una fuerza que es proporcional al producto de sus masas e inversamente proporcional al cuadrado de la distancia entre ellas. En símbolos:

$$F = G \frac{m_1 m_2}{d^2}$$

Aquí F es la fuerza de atracción, d es la distancia, las m son las dos masas y G es un número concreto, la constante de gravitación.¹⁷

¿Quién descubrió la ley de gravitación de Newton? Suena como una de esas preguntas que contienen la respuesta, como «¿de qué color es el caballo blanco de Santiago?». Pero una respuesta sensata es el comisario de experimentos en la Royal Society, Robert Hooke. Cuando Newton publicó la ley en 1687, en su *Principia*, Hooke le acusó de plagio. Sin embargo, Newton proporcionaba la primera derivada matemática de órbitas elípticas a partir de la ley, lo cual era vital para establecer su corrección, y Hooke reconoció esto. Además, Newton había citado a Hooke, entre otros, en el libro. Presumiblemente Hooke sentía que merecía más crédito, había sufrido problemas parecidos varias veces antes y era un asunto delicado.

¹⁷ La cifra no se conoció hasta 1798, cuando Henry Cavendish obtuvo un valor razonablemente preciso en un experimento de laboratorio. Es sobre $6,67 \times 10^{-11}$ newton metro cuadrado por kilogramo al cuadrado.

La idea de que los cuerpos se atraían unos a otros había estado merodeando por un tiempo, y también su probable expresión matemática. En 1645 el astrónomo francés Ismaël Boulliau (Bullialdus) escribió su *Astronomia Philolaica* (Astronomía filolaica; Filolao de Crotona era un filósofo griego que pensaba que un fuego central, no la Tierra, era el centro del universo). En él escribió:

Con respecto a la fuerza por la cual el Sol agarra o sostiene a los planetas, y la cual, al ser corpórea, funciona de la misma manera que las manos; se emite en líneas rectas por toda la extensión del mundo y, como especies del Sol, gira con el cuerpo del Sol; ahora, viendo que es corpóreo, se hace más débil y atenúa a distancias o intervalos mayores, y la razón de su descenso según la longitud es la misma que en el caso de la luz, a saber, la proporción duplicada, pero inversamente, de las distancias.

Esta es la famosa dependencia del «cuadrado de la inversa» de la fuerza sobre la distancia. Hay razones simples, aunque inocentes, para esperar dicha fórmula, porque el área de la superficie de una esfera varía según el cuadrado de su radio. Si la misma cantidad de «material» gravitacional se extiende por una esfera cada vez mayor a medida que se separa del Sol, entonces la cantidad recibida en cualquier punto debe variar en proporción inversa al área. Exactamente esto sucede con la luz, y Boulliau asumió, sin muchas pruebas, que la gravedad debía ser análoga. También pensó que los planetas se mueven a lo largo de sus órbitas bajo su propia fuerza, por así decirlo: «Ningún tipo de movimiento presiona sobre los planetas restantes, los cuales son conducidos alrededor por formas individuales con las cuales fueron provistos».

Las contribuciones de Hooke datan de 1666, cuando presentó un artículo a la Royal Society con el título «On gravity» (Sobre la gravedad). Aquí ponía en orden lo que Boulliau había hecho erróneamente, argumentando que una fuerza de atracción proveniente del Sol podía interferir con una tendencia natural de los planetas a moverse en línea recta (como especificaba la tercera ley de movimiento de Newton) y por causa de eso se obtenía una curva. También afirmó que «estas fuerzas de atracción son mucho más potentes funcionando, cuanto más cercano está el cuerpo a sus propios centros», mostrando que pensaba que la fuerza decaía con la

distancia. Pero no le dijo a nadie más la forma matemática para este decrecimiento hasta 1679, cuando escribió a Newton: «la atracción siempre está en una proporción duplicada de la distancia desde el centro recíproca». En la misma carta dijo que esto implica que la velocidad de un planeta varía como el recíproco de su distancia desde el Sol. Lo cual es erróneo.

Cuando Hooke se quejó de que Newton le había robado su ley, Newton no estaba cogiendo nada de ella, e indicó que había discutido la idea con Christopher Wren antes de que Hooke le hubiese enviado su carta. Para demostrar conocimientos previos, citó a Boulliau, y también a Giovanni Borelli, un psicólogo y físico matemático italiano. Borelli había sugerido que tres fuerzas se combinan para crear movimiento planetario: una fuerza interior causada por el deseo de los planetas de aproximarse al Sol, una fuerza lateral causada por la luz del Sol, y una fuerza exterior causada por la rotación del Sol. Acertó una de tres, y eso siendo generoso. El principal punto de Newton, generalmente considerado decisivo, es que fuera lo que fuera que Hooke había hecho, no había deducido la forma exacta de las órbitas a partir de la atracción de la ley del cuadrado de la inversa. Newton lo había hecho. De hecho, había deducido las tres leyes del movimiento planetario de Kepler: órbitas elípticas, barrido de áreas iguales en intervalos de tiempo iguales, con el cuadrado del período siendo proporcional al cubo de la distancia. «Sin mis demostraciones», insistía Newton, la ley del cuadrado de la inversa, «un filósofo sensato no podría creer que fuese ni mucho menos correcta». Pero también aceptaba que «el señor Hooke es todavía un extraño» a esta prueba. Una característica clave de la argumentación de Newton es que no se aplica tan solo a un punto, sino a una esfera. Esta extensión, la cual es crucial para el movimiento de los planetas, había supuesto un considerable esfuerzo para Newton. Su prueba geométrica es una aplicación del cálculo integral disfrazada, y estaba, con razón, orgulloso de ella. Hay también evidencias de que Newton había estado pensando sobre dicha cuestión durante bastante tiempo.

De cualquier forma, es frecuente referirse a ella como ley de gravitación universal de Newton, y esto hace justicia a la importancia de su contribución.

El aspecto más importante de la ley de gravitación universal no es la ley del cuadrado de la inversa como tal. Es la afirmación de que la gravedad actúa

universalmente. Dos cuerpos cualquiera, en cualquier lugar del universo, se atraen el uno al otro. Por supuesto se necesita una ley de fuerzas precisa (el cuadrado de la inversa) para obtener resultados precisos, pero sin la universalidad, no sabes cómo escribir las ecuaciones para cualquier sistema con más de dos cuerpos. Casi todos los sistemas interesantes, como el propio Sistema Solar, o la sutil estructura del movimiento de la Luna bajo la influencia de (al menos) el Sol y la Tierra, implica más de dos cuerpos, así que la ley de Newton habría sido casi inútil si solo se hubiese aplicado al contexto en el cual la dedujo inicialmente.

¿Qué motivó esta visión de universalidad? En sus *Memoirs of Sir Isaac Newton's Life* (Memorias de la vida de Sir Isaac Newton) de 1752, William Stukeley cuenta una historia que Newton le había contado a él en 1726:

La noción de gravedad... era ocasionada por la caída de una manzana, mientras que estaba sentado con un humor contemplativo. ¿Por qué debería esa manzana siempre descender perpendicularmente a la tierra?, pensó. ¿Por qué no va hacia los lados o hacia arriba, sino constantemente al centro de la Tierra? Ciertamente la razón es que la Tierra tira de ella. Debe haber una fuerza tiradora en la materia. Y la suma de la potencia tiradora en la materia de la Tierra debe estar en el centro de la Tierra, no en cualquier lado de la Tierra. Por lo tanto, ¿cae la manzana perpendicularmente o hacia el centro? Si la materia tira de este modo de la materia, debe estar en proporción a su cantidad. Por lo tanto la manzana tira de la Tierra, del mismo modo en que la Tierra tira de la manzana.

Si la historia es la verdad literal o una invención oportuna que Newton se inventó para ayudar a explicar sus ideas más adelante, no está totalmente claro, pero parece razonable tomar el cuento en serio porque la idea no acaba con manzanas. La manzana era importante para Newton porque le hizo darse cuenta de que la misma ley de fuerza puede explicar tanto el movimiento de la manzana como el movimiento de la Luna. La única diferencia es que la Luna también se mueve hacia los lados, esta es la razón por la cual se sostiene. Realmente, está siempre cayendo hacia la Tierra, pero el movimiento lateral causa que la superficie de la Tierra se caiga también. Newton, siendo Newton, no podía detenerse con este argumento

cualitativo. Hizo las cuentas, las comparó con las observaciones y se quedó satisfecho creyendo que su idea debía ser correcta.

Si la gravedad actúa en la manzana, la Luna y la Tierra, como una característica inherente de la materia, entonces probablemente actúa sobre todo.

No es posible verificar la universalidad de las fuerzas de gravedad directamente, tendrías que estudiar todos los pares de cuerpos del universo entero, y encontrar el modo de eliminar la influencia de los otros cuerpos. Pero así no es como funciona la ciencia. En su lugar, emplea una mezcla de inferencia y observaciones. La universalidad es una hipótesis, capaz de ser falsificada cada vez que se aplica. Cada vez que sobrevive a una falsificación, un modo extravagante de decir que da buenos resultados, la justificación para usarla se hace un poco más fuerte. Si (como en este caso) sobrevive a miles de dichas pruebas, la justificación se hace realmente fuerte. Sin embargo, la hipótesis nunca se puede verificar; hasta donde sabemos, el próximo experimento podría producir un resultado incompatible. Quizá en algún punto de una galaxia lejana, muy lejos hay una pizca de materia, un átomo, que no es atraído por todo lo demás. Si es así, nunca lo encontraríamos, por lo tanto no alteraría nuestros cálculos. La propia ley del cuadrado de la inversa es extremadamente difícil de verificar directamente, esto es, midiendo realmente la fuerza de atracción. En su lugar, aplicamos la ley a un sistema que podemos medir, usándolo para predecir órbitas y luego comprobar si las predicciones coinciden con las observaciones.

Incluso concediendo la universalidad, no es suficiente para escribir una ley de atracción precisa. Esto solo produce una ecuación para describir el movimiento. Para encontrar el propio movimiento, hay que resolver la ecuación. Incluso para dos cuerpos, esto no es directo y sencillo, e incluso teniendo en mente que sabía por adelantado qué respuesta esperaba, la deducción de Newton de órbitas elípticas es una hazaña. Explica por qué las tres leyes de Kepler proporcionan una descripción muy precisa de la órbita de cada planeta. También explica por qué esa descripción no es exacta: otros cuerpos del Sistema Solar, otros diferentes del Sol y el propio planeta afectan al movimiento. Para explicar estas alteraciones, tienes que resolver las ecuaciones de movimiento para tres o más cuerpos. En particular, si quieres predecir el movimiento de la Luna con una precisión alta, tienes que incluir el Sol y

la Tierra en tus ecuaciones. Los efectos de los otros planetas, especialmente Júpiter, tampoco son totalmente desdeñables, pero solo aparecen a largo plazo. Así, fuertes por el éxito de Newton con el movimiento de dos cuerpos bajo la gravedad, los matemáticos y físicos pasaron al siguiente caso: tres cuerpos. Su optimismo inicial se disipó rápidamente: el caso de tres cuerpos resultó ser muy diferente del caso de dos cuerpos. De hecho, se resiste a una solución.

Era posible con frecuencia calcular buenas aproximaciones al movimiento (las cuales a menudo solucionaban el problema para propósitos prácticos), pero ya no parecía ser una fórmula exacta. Este problema aquejaba incluso a versiones simplificadas, tales como el problema de tres cuerpos restringido. Supongamos que un planeta describe una órbita alrededor de una estrella en un círculo perfecto; ¿cómo se moverá una mota de polvo de masa insignificante?

Calculando órbitas aproximadas para tres o más cuerpos, a mano, usando lápiz y papel, era más o menos factible, pero muy laborioso. Los matemáticos diseñaron innumerables trucos y atajos, que llevaban a un entendimiento razonable de varios fenómenos astronómicos. Solo a finales del siglo XIX la verdadera complejidad del problema de tres cuerpos se hizo evidente, cuando Henri Poincaré se dio cuenta de que la geometría que implicaba era, por fuerza, extraordinariamente complicada. Y solo a finales del siglo XX la llegada de ordenadores potentes redujo la labor de los cálculos a mano, permitiendo predicciones precisas de movimiento del Sistema Solar a largo plazo.

El gran avance de Poincaré —si puede llamarse así, ya que en la época parece que decía a todos que el problema era imposible y no tenía sentido buscar una solución— sucedió porque compitió para un premio matemático. Óscar II, rey de Suecia y Noruega, anunció una competición para celebrar su sesenta cumpleaños en 1889. Teniendo en cuenta el consejo del matemático Gösta Mittag-Leffler, el rey escogió el problema general de muchos cuerpos moviéndose arbitrariamente bajo la gravedad newtoniana. Ya que se tenía claro que una fórmula explícita semejante a la elipse para los dos cuerpos era un objetivo poco realista, el requisito era laxo: el premio se concedería por un método de aproximación de un tipo muy específico. Concretamente, el movimiento debía determinarse como una serie infinita, dando resultados tan precisos como quisiéramos si se incluían los términos suficientes.

Poincaré no respondió a esta cuestión. En su lugar, su memoria sobre el tema, publicada en 1890, proporcionaba pruebas de que no podría tenerse ese tipo de respuesta, ni siquiera para tres cuerpos: estrella, planeta y una partícula de polvo. Pensando en la geometría de soluciones hipotéticas, Poincaré descubrió que en algunos casos la órbita de la partícula de polvo debía ser extremadamente compleja y enrevesada. Entonces prácticamente se echó las manos a la cabeza con horror e hizo la pesimista afirmación de que «cuando uno trata de describir la figura formada por estas dos curvas y su infinidad de intersecciones, cada una de las cuales corresponde a una solución doblemente asintótica, estas intersecciones forman un tipo de red, maraña o malla infinitamente ceñida... Uno se da de brúces con la complejidad de esta figura que no me atrevo siquiera a dibujar».

Ahora vemos el trabajo de Poincaré como un gran avance, y pasamos por alto su pesimismo, porque la geometría complicada que le desalentó de resolver alguna vez el problema realmente proporciona una comprensión poderosa si se desarrolla e interpreta adecuadamente. La geometría compleja de la dinámica asociada resultó ser uno de los primeros ejemplos del caos: la existencia, en ecuaciones no aleatorias, de soluciones tan complicadas que de algún modo parecen ser aleatorias (véase el capítulo 16).

Hay varias ironías en la historia. La historiadora matemática June Barrow-Green descubrió que la versión publicada de la memoria ganadora de Poincaré del premio no era la que ganó el premio.¹⁸ Esta versión anterior contenía un error importante, pasando por alto las soluciones caóticas. El trabajo estaba en una etapa de prueba cuando un avergonzado Poincaré se dio cuenta de su metedura de pata, y pagó por una nueva impresión de una versión correcta. Casi todas las copias de la original fueron destruidas, pero una permaneció guardada en los archivos del Instituto de Mittag-Leffler en Suecia, donde Barrow-Green la encontró.

También resultó que la presencia del caos de hecho no excluía las series como soluciones, sino que estas eran válidas casi siempre en vez de serlo siempre. Karl Frithiof Sundman, un matemático finés, descubrió esto en 1912 para el problema de tres cuerpos, usando series formadas a partir de potencias de la raíz cúbica del tiempo. (Las potencias del tiempo no lo resolverán.) Las series convergen —tienen

¹⁸ June Barrow-Green. *Poincaré and the Three Body Problem*, American Mathematical Society, Providence 1997.

una suma perceptible— a menos que el estado inicial tenga un momento angular igual a cero, pero dichos estados son tremadamente raros, en el sentido de que una elección aleatoria del momento angular es casi siempre distinta de cero. En 1991, el matemático chino Qiudong Wang amplió estos resultados para cualquier número de cuerpos, pero no clasificó las excepciones raras cuando las series no convergen. Dicha clasificación es probable que sea muy complicada, debe incluir soluciones donde los cuerpos escapan al infinito en un tiempo finito, u oscilan incluso más rápido, ambos casos pueden darse para cinco o más cuerpos.

La ley de gravitación universal se aplica rutinariamente al diseño de órbitas para misiones espaciales. Aquí incluso la dinámica de dos cuerpos es útil por sí misma. En sus inicios, la exploración del Sistema Solar principalmente usaba órbitas de dos cuerpos, segmentos de elipses. Quemando combustible, la nave espacial podía pasar de una elipse a otra diferente. Pero a medida que los objetivos de los programas espaciales se hacían más ambiciosos, se necesitaban métodos más eficientes. Vinieron gracias a la dinámica de muchos cuerpos, normalmente tres cuerpos, pero ocasionalmente pueden llegar a ser cinco. Los nuevos métodos de caos y dinámica topológica se convirtieron en las bases de las soluciones prácticas a los problemas de ingeniería.

Todo empezó con una pregunta simple: ¿cuál es la ruta más eficiente para ir de la Tierra a la Luna o los planetas? La respuesta clásica, conocida como una órbita de transferencia de Hohmann (figura 14) empieza con una órbita circular alrededor de la Tierra, y luego sigue con parte de una elipse larga y fina para unir con una segunda órbita circular alrededor del destino. Este método fue empleado durante las misiones espaciales Apolo de las décadas sesenta y setenta del siglo XX, pero para muchos tipos de misión tenía un inconveniente. La nave espacial debe lanzarse fuera de la órbita terrestre y luego reducir la velocidad al entrar en la órbita lunar; esto malgasta combustible. Hay alternativas que implican muchos bucles alrededor de la Tierra, una transición por el punto entre la Tierra y la Luna donde sus campos de gravedad se anulan y muchos bucles alrededor de la Luna. Pero trayectorias como esta llevan mucho más tiempo que las elipses de Hohmann, de modo que no se usaron en las misiones tripuladas Apolo donde la comida y el oxígeno, y por consiguiente el tiempo, eran oro. Para misiones no tripuladas, sin embargo, el

tiempo es relativamente barato, mientras que cualquier cosa que añade peso al total de la nave espacial, incluyendo el combustible, cuesta dinero.

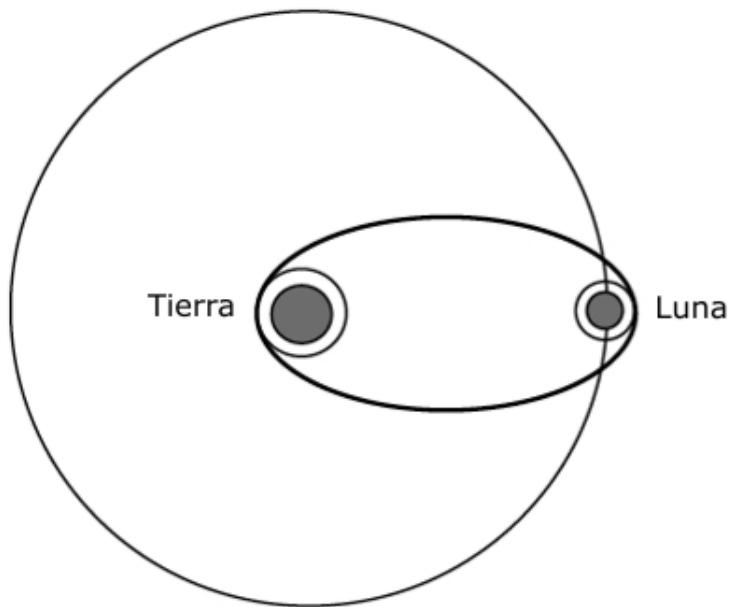


FIGURA 14. Órbita de transferencia de Hohmann de una órbita terrestre baja a la órbita lunar.

Considerando una nueva perspectiva de la ley de gravitación universal y la segunda ley de movimiento de Newton, los matemáticos e ingenieros espaciales han descubierto recientemente una nueva y destacable aproximación a los viajes interplanetarios para que sean eficientes en lo que respecta al combustible.

Viajar en tubo.

Es una idea sacada directamente de la ciencia ficción. En *La estrella de Pandora* de 2004, Peter Hamilton describe un futuro donde la gente viaja a los planetas rodeando estrellas distantes en tren, donde las vías del tren van a través de un agujero de gusano, un atajo a través del espacio-tiempo. En sus series de *El hombre lente* de 1934 a 1948, Edward Elmer «Doc» Smith se inventaba un metro hiperespacial, el cual alienígenas malvados usaban para invadir mundos humanos desde la cuarta dimensión.

Aunque todavía no tenemos agujeros de gusano o alienígenas de la cuarta dimensión, se ha descubierto que los planetas y satélites del Sistema Solar están

conectados unos a otros por una red de túneles, cuya definición matemática requiere muchas más dimensiones que cuatro. Los tubos proporcionan rutas que son energético-eficientes de un mundo a otro. Solo pueden verse a través de ojos matemáticos, porque no están hechos de materia, sus paredes son niveles de energía. Si pudiésemos visualizar el paisaje siempre cambiante de los campos de gravedad que controlan cómo los planetas se mueven, seríamos capaces de ver los tubos, haciendo remolinos, junto con los planetas a medida que se mueven en su órbita alrededor del Sol.

Los tubos explican algunas dinámicas orbitales misteriosas. Considera, por ejemplo, el cometa llamado Oterma. Hace un siglo, la órbita de Oterma estaba fuera de la de Júpiter. Pero después de un encuentro cercano con el planeta gigante, la órbita del cometa se cambió a la de Júpiter. Después de otro encuentro cercano, se volvió a salir. Podemos predecir con seguridad que Oterma continuará cambiando su órbita de este modo cada pocas décadas, no porque rompa la ley de Newton, sino porque la obedece.

Esto no tiene nada que ver con elipses ordenadas. Las órbitas que predecía la gravedad newtoniana son elípticas solo cuando no hay otros cuerpos que ejerzan una atracción gravitacional significativa. Pero el Sistema Solar está lleno de otros cuerpos y pueden suponer una diferencia enorme, y sorprendente. Es aquí cuando los tubos entran en la historia. La órbita de Oterma está dentro de dos tubos, los cuales se cruzan cerca de Júpiter. Un tubo está ubicado dentro de la órbita de Júpiter y el otro fuera. Encierran órbitas especiales en resonancia 3:2 y 2:3 con Júpiter, esto quiere decir que un cuerpo con dicha órbita girará alrededor del Sol tres veces por cada dos revoluciones de Júpiter, o dos veces por cada tres. En la confluencia de tubos cerca de Júpiter, el cometa puede cambiarse de tubo, o no, dependiendo de efectos bastante sutiles de la gravedad joviana y solar. Pero una vez dentro del tubo, Oterma se queda ahí hasta que el tubo vuelve al punto de cruce. Como un tren que tiene que quedarse en la vía, pero puede cambiar de ruta a otra vía si alguien cambia los puntos, Oterma tiene alguna libertad de cambio de su itinerario, pero no mucha (figura 15).

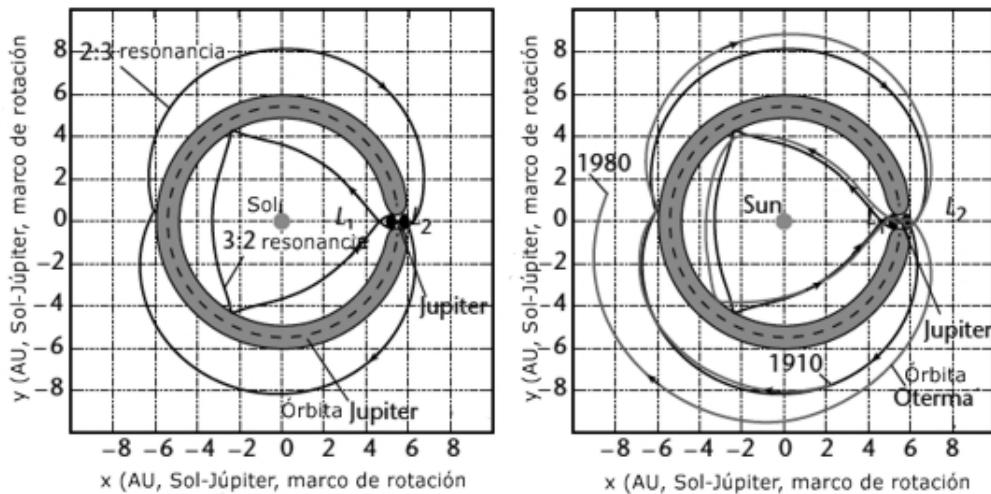


FIGURA 15. A la izquierda: dos órbitas periódicas en resonancia 2:3 y 3:2 con Júpiter, conectadas por puntos de Lagrange. A la derecha: órbita real del cometa Oterma, 1910-1980.

Los tubos y sus confluencias podrían parecer extraños, pero son características naturales e importantes de la geografía gravitacional del Sistema Solar. Los constructores de las vías de tren de la época victoriana entendieron la necesidad de explotar características naturales del terreno, llevando las vías del tren por valles y a lo largo de curvas de nivel, y excavando túneles a través de colinas en vez de llevar el tren sobre la cima. Una razón era que los trenes tienden a resbalar en pendientes pronunciadas, pero la principal era la energía. Subir una colina, en contra de la fuerza de la gravedad, cuesta energía, la cual se manifiesta como un incremento del consumo de combustible, lo cual cuesta dinero.

Es muy parecido a los viajes interplanetarios. Imagina una nave moviéndose por el espacio. Su siguiente destino no depende solamente de donde está ahora, también depende de lo rápido que se mueva y en qué dirección. Se necesitan tres números para especificar la posición de la nave espacial, por ejemplo, su dirección desde la Tierra, la cual requiere dos números (los astrónomos usan ascensión recta y declinación, que son las análogas a la longitud y la latitud para la esfera celestial, la supuesta esfera formada por el cielo nocturno) y su distancia desde la Tierra. Se necesitan más de tres números para especificar su velocidad en estas tres direcciones. De modo que la nave viaja a través de un terreno matemático que

tiene seis dimensiones en lugar de dos.

Un terreno natural no es llano, tiene colinas y valles. Se necesita energía para escalar una colina, pero un tren puede ganar energía deslizándose cuesta abajo hacia un valle. De hecho, entran en juego dos tipos de energía. La altura sobre el nivel del mar determina la energía potencial del tren, la cual representa el trabajo hecho contra la fuerza de gravedad. Cuanto más alto vaya, mayor será la energía potencial que se debe crear. El segundo tipo es la energía cinética, la cual se corresponde con la celeridad. Cuanto más rápido vaya, mayor se hace la energía cinética. Cuando el tren está bajando la colina y acelera, intercambia energía potencial por energía cinética. Cuando sube una colina y baja su velocidad, el intercambio es a la inversa. La energía total es constante, de modo que la trayectoria del tren es análoga a una curva de nivel en la superficie de energía potencial. Sin embargo, los trenes tienen una tercera fuente de energía: carbón, gasoil o electricidad. Gastando combustible, un tren puede subir una pendiente o acelerar, liberándose de su trayectoria natural que se mueve libremente. La energía total todavía no puede cambiar, pero todo lo demás es negociable.

Es muy parecido con la nave espacial. Los campos de gravitación combinados del Sol, los planetas y otros cuerpos del Sistema Solar proporcionan energía potencial. La velocidad de la nave se corresponde con la energía cinética. Y su fuerza motriz, ya sea propergol, iones o presión luminosa, añade una fuente de energía extra, la cual puede apagarse o encenderse según se quiera. La ruta seguida por la nave espacial es un tipo de curva de nivel en la superficie de energía potencial correspondiente y a lo largo de esa ruta la energía total permanece constante. Y algunos tipos de curvas de nivel están rodeadas por tubos, que se corresponden con sus niveles de energía cercanos.

Los ingenieros de trenes victorianos eran también conscientes de que la superficie terrestre tenía características especiales: picos, valles, puertos de montaña; los cuales tiene un gran efecto en rutas eficientes para las vías del tren, porque constituyen un tipo de esqueleto para la geometría global de las curvas de nivel. Por ejemplo, cerca de un pico o en el fondo de un valle, las curvas de nivel están muy cerca unas de otras. En los picos la energía potencial está localmente en un máximo, en un valle, está en un mínimo local. Los puertos combinan características

de ambos, siendo un máximo en una dirección, pero un mínimo en otra. De manera similar, las superficies de energía potencial del Sistema Solar tienen características especiales. Las más obvias son los propios planetas y lunas, que se encuentran en el fondo de los pozos gravitacionales, como los valles. Igualmente importantes, pero menos visibles, son los picos y los puertos de las superficies de energía potencial. Todas estas características organizan la geometría global y, con ello, los tubos. Las superficies de energía potencial tienen otras características atractivas para los turistas, en particular, los puntos de Lagrange. Imagina un sistema formado solo por la Tierra y la Luna. En 1772 Joseph-Louis Lagrange descubrió que en cualquier instante hay precisamente cinco lugares donde los campos gravitacionales de dos cuerpos, junto con la fuerza centrífuga, se anulan totalmente. Tres están alineados con la Tierra y la Luna: L1 está entre ellos, L2 está en la cara más alejada de la Luna y L3 está en la cara alejada de la Tierra. El matemático suizo Leonhard Euler ya había descubierto esto alrededor de 1750. Pero también estaban L4 y L5, conocidos como puntos troyanos, los cuales están en la misma órbita que la Luna pero 60 grados por delante o por detrás de ella. A medida que la Luna rota alrededor de la Tierra, los puntos de Lagrange rotan con ella. Otros pares de cuerpos también tienen puntos de Lagrange: Tierra/Sol, Júpiter/Sol, Titán/Saturno. La antigua transferencia de órbita de Hohmann está construida a partir de piezas de círculos y elipses, que son las trayectorias naturales para sistemas de dos cuerpos. Las nuevas rutas basadas en tubos están construidas a partir de piezas de las trayectorias naturales de los sistemas de tres cuerpos, tales como Sol/Tierra/nave espacial. Los puntos de Lagrange juegan un papel especial, justo como los picos y los puertos lo hacían para las vías del tren, son los cruces donde los tubos se encuentran. L1 es un gran lugar para hacer pequeños cambios de recorrido, porque la dinámica natural de la nave espacial cerca de L1 es caótica (figura 16). El caos tiene una característica útil (véase el capítulo 16): cambios muy pequeños en la posición o velocidad pueden crear grandes cambios en la trayectoria. De modo que es fácil redireccionar a la nave en una manera combustible-eficiente, aunque posiblemente lenta.

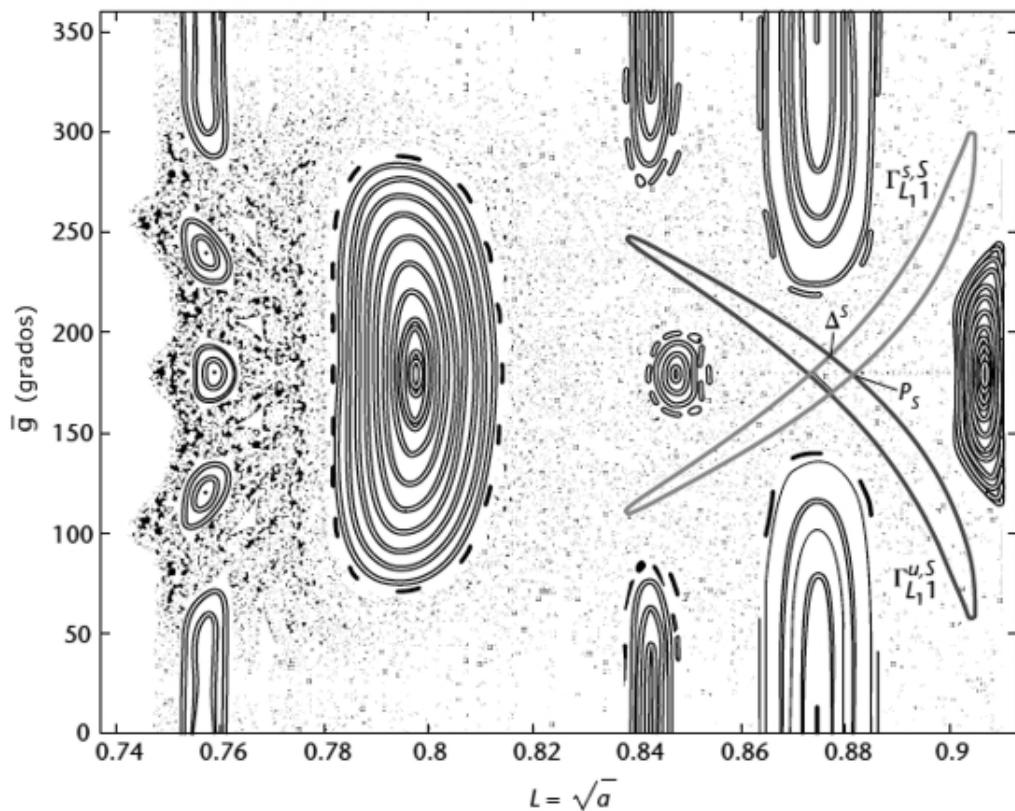


FIGURA 16. Caos cerca de Júpiter. El diagrama muestra una sección de órbitas. Los bucles anidados son órbitas cuasiperiódicas y la restante región punteada es una órbita caótica. Los dos bucles finos que se cruzan en la derecha son secciones de tubos.

La primera persona que se tomó esta idea en serio fue el matemático nacido en Alemania Edward Belbruno, un analista de órbitas en el Jet Propulsion Laboratory (Laboratorio de propulsión a chorro) desde 1985 hasta 1990. Se dio cuenta de que las dinámicas caóticas en sistemas de muchos cuerpos proporcionan una oportunidad para órbitas de transferencia de baja energía novedosas, llamando a la teoría técnica de límites borrosos. En 1991, puso sus ideas en práctica. Hiten, una sonda espacial japonesa, había estado inspeccionando la Luna y había completado la misión planeada, volviendo a la órbita de la Tierra. Belbruno diseñó una nueva órbita que la llevaría de vuelta a la Luna a pesar de estar quedándose sin combustible. Después de aproximarse a la Luna como se pretendía, Hiten visitó sus puntos L4 y L5 para buscar polvo cósmico que quizás se había quedado atrapado ahí.

Un truco similar se usó en 1985 para redireccionar la casi muerta International Sun-Earth Explorer ISEE-3 para encontrarse con el cometa Giacobini-Zinner, y fue usado de nuevo por la misión Génesis de la NASA para traer muestras de viento solar. Los matemáticos e ingenieros querían repetir el truco, y encontrar otros del mismo tipo, lo que significa averiguar qué es lo que realmente lo hace funcionar. Resultó que eran tubos.

La idea subyacente es simple pero inteligente. Esos lugares especiales en las superficies de energía potencial que recuerdan a los puertos de montaña crean cuellos de botella que aspirantes a viajeros no pueden evitar fácilmente. En la Antigüedad descubrieron, a fuerza de palos, que incluso aunque consuma energía subir un puerto, consume más energía seguir cualquier otra ruta, a menos que puedas rodear la montaña en una dirección totalmente diferente. El puerto es la mejor de las opciones.

En las superficies de energía potencial, los análogos a los puertos incluyen puntos de Lagrange. Asociadas a ellos hay rutas entrantes muy específicas, que son el modo más eficiente de subir el puerto. Hay también rutas salientes igualmente específicas, análogas a las rutas naturales de bajada del puerto. Para seguir estas rutas de entrada y salida exactamente, tienes que viajar justo a la velocidad correcta, aunque si tu velocidad es ligeramente diferente puedes todavía permanecer cerca de estas rutas. A finales de la década de los sesenta del siglo XX, los matemáticos americanos Charles Conley y Richard McGehee pusieron en práctica el trabajo pionero de Belbruno, señalando que cada una de dichas rutas está rodeada por un conjunto anidado de tubos, uno dentro de otro. Cada tubo se corresponde con una elección concreta de velocidad; cuanto más lejos está de la velocidad óptima, más ancho es el tubo. En la superficie de un tubo dado cualquiera, la energía total es constante, pero las constantes difieren de un tubo a otro. Algo así como una curva de nivel que está a una altura constante pero la altura es diferente para cada curva de nivel.

Entonces, el modo de planear un perfil eficiente de la misión es calcular qué tubos son relevantes para el destino elegido. Luego haces la ruta para la nave a lo largo del interior del primer tubo entrante y cuando llega al punto de Lagrange asociado provoca un rápido arranque en los motores para redirigirlo a lo largo del tubo

saliente más apropiado (figura 17). Ese tubo fluye de modo natural al correspondiente tubo entrante del siguiente punto de cambio... y allá va.

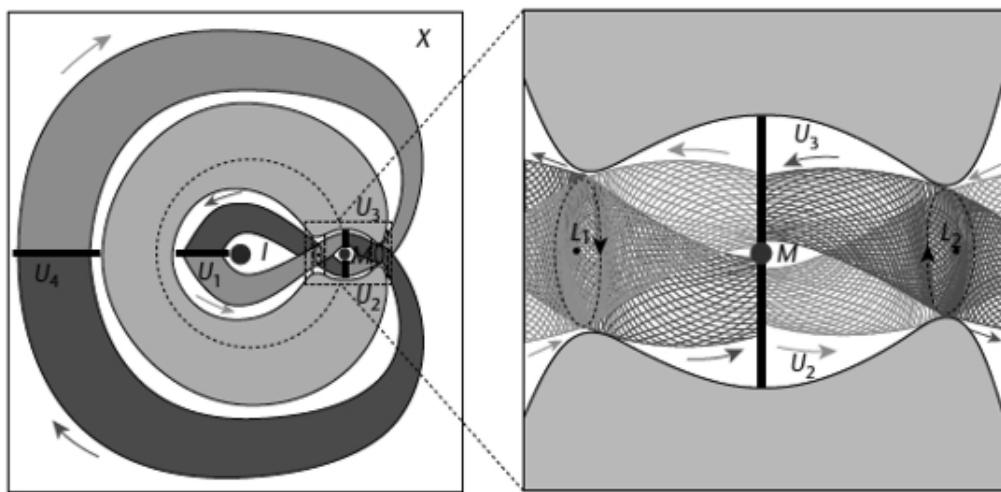


FIGURA 17. A la izquierda: tubos cruzándose cerca de Júpiter. A la derecha: primer plano de la región donde los tubos se cruzan.

Planos para futuras misiones tubulares ya se están preparando. En el año 2000 Wang Sang Koon, Martin Lo, Jerrold Marsden y Shane Ross usaron la técnica de los tubos para encontrar un «Petit Grand Tour» de las lunas de Júpiter, acabando con una órbita de captura alrededor de Europa, la cual había sido muy escurridiza con métodos anteriores. La ruta supone un empuje gravitacional cerca de Ganímedes seguido por un viaje en tubo a Europa. Una ruta más compleja, que requiere incluso menos energía, incluye también a Calisto. Hace uso de otra característica de la superficie de energía potencial: resonancias. Estas se dan cuando, por ejemplo, dos satélites repetidamente vuelven a las mismas posiciones relativas, pero uno da dos vueltas a Júpiter mientras que el otro da tres vueltas. Cualquier número pequeño puede remplazar a 2 y 3. Esta ruta usa dinámica de cinco cuerpos: Júpiter, las tres lunas y la nave espacial.

En 2005, Michael Dellnitz, Oliver Junge, Marcus Post y Bianca Thiere usaron tubos para planear una misión energético-eficiente de la Tierra a Venus. El principal tubo aquí une el punto L1 del sistema Tierra/Sol con el punto L2 del sistema Sol/Venus. En comparación, esta ruta usa solo un tercio del combustible necesario por la misión

Venus Express de la Agencia Espacial Europea, porque puede usar motores de bajo empuje; el precio pagado es una prolongación del tiempo de tránsito de 150 días a alrededor de 650 días.

La influencia de los tubos quizá vaya más lejos. En un trabajo no publicado, Dellnitz ha descubierto pruebas de un sistema natural de tubos que conectan Júpiter a cada uno de los planetas interiores. Esta estructura notable, ahora llamada la Superautopista interplanetaria, insinúa que Júpiter, durante mucho tiempo conocido por ser el planeta dominante del Sistema Solar, también desempeña el papel de una gran estación central celestial. Sus tubos podrían haber organizado la formación de todo el Sistema Solar, determinando los espacios de los planetas interiores.

¿Por qué los tubos no se descubrieron antes? Hasta hace muy poco, se carecía de dos cosas vitales. Una era ordenadores potentes, capaces de llevar a cabo los cálculos necesarios para muchos cuerpos. Son demasiado engorrosos para hacerlos a mano. La otra, incluso más importante, era un entendimiento matemático profundo de la geografía de una superficie de energía potencial. Sin este triunfo imaginativo de los métodos matemáticos modernos, los ordenadores no tendrían nada que calcular. Y sin la ley de gravitación universal, los métodos matemáticos nunca se habrían concebido.

Capítulo 5
Presagio del mundo ideal
Raíz cuadrada de menos uno

$$i^2 = -1$$

¿Qué dice?

Aunque debería ser imposible, el cuadrado del número i es menos uno.

¿Por qué es importante?

Llevó a la creación de los números complejos, los cuales a su vez llevaron al análisis complejo, una de las áreas más potentes de las matemáticas.

¿Qué provocó?

Métodos mejorados para calcular tablas trigonométricas. Generalizaciones de casi todas las matemáticas al reino complejo. Métodos más potentes para comprender ondas, calor, electricidad y magnetismo. Las bases matemáticas de la mecánica cuántica.

La Italia del Renacimiento era un estercolero de políticos y violencia. El norte del país estaba controlado por una docena de ciudades Estado enfrentadas, entre ellas Milán, Florencia, Pisa, Génova y Venecia. En el sur, güelfos y gibelinos estaban en conflicto mientras papas y emperadores del Sacro Imperio Romano Germánico luchaban por la supremacía. Bandas de mercenarios deambulaban por el campo, los pueblos eran arrasados, y las ciudades costeras libraban unas contra otras una guerra naval. En 1454 Milán, Nápoles y Florencia firmaron el Tratado de Lodi, y la

paz reinó durante las siguientes cuatro décadas, pero el papado seguía envuelto en corrupción política. Esta era la época de los Borgia, famosos por envenenar a cualquiera que se pusiese en su camino por la búsqueda de poder político y religioso, pero también era la época de Leonardo da Vinci, Brunelleschi, Piero della Francesca, Tiziano y Tintoretto. En medio de un ambiente de intrigas y asesinatos, suposiciones que se habían mantenido durante mucho tiempo eran puestas en duda. Arte importante y ciencia importante florecieron en simbiosis, cada uno nutriéndose del otro.

Matemáticas importantes también florecieron. En 1545, el académico ludópata Girolamo Cardano estaba escribiendo un texto de álgebra, y encontró un nuevo tipo de número, uno tan desconcertante que lo declaró «tan sutil como inútil» y desechó la idea. Rafael Bombelli tenía una comprensión sólida de los libros de álgebra de Cardano, pero encontró la exposición confusa y decidió que podía mejorarlala. Alrededor de 1572, se había dado cuenta de algo enigmático: aunque estos números nuevos y desconcertantes no tenían sentido, podían usarse en cálculos algebraicos y nos llevaban a resultados que eran correctos y demostrables.

Durante siglos, los matemáticos se vieron envueltos en una relación amor-odio con estos «números imaginarios», como son todavía conocidos en la actualidad. El nombre delata una actitud ambivalente: no son números reales, los números habituales que encontramos en aritmética, pero en casi todos los sentidos se comportan como ellos. La principal diferencia es que cuando haces el cuadrado de un número imaginario, el resultado es negativo. Pero eso no debería ser posible, porque los cuadrados son siempre positivos.

No fue hasta el siglo XVIII cuando los matemáticos comprendieron qué eran los números imaginarios. No fue hasta el siglo XIX cuando empezaron a sentirse cómodos con ellos. Pero en el momento en que el estatus lógico de los números imaginarios se vio que era totalmente comparable al de los números reales más tradicionales, los imaginarios se habían convertido en indispensables para toda la matemática y la ciencia, y la cuestión de su significado difícilmente parecía interesar ya a nadie. A finales del siglo XIX y principios del XX, el interés resurgido en los cimientos de las matemáticas llevó a repensar el concepto de número y se vio que los números «reales» tradicionales no eran más reales que los imaginarios.

Lógicamente, los dos tipos de números eran tan parecidos como dos gotas de agua. Ambos eran construcciones de la mente humana, ambos representaban, pero no eran sinónimos de, aspectos de la naturaleza. Pero representaban la realidad en modos diferentes y contextos diferentes.

En la segunda mitad del siglo XX, los números imaginarios eran simplemente parte esencial de la caja de herramientas mental de todo matemático y científico. Se incorporaron a la mecánica cuántica de una manera tan fundamental que hacer física sin ellos es como escalar la cara norte del Eiger sin cuerdas. Incluso así, los números imaginarios raras veces se enseñan en las escuelas. Las cuentas son bastante fáciles, pero la sofisticación mental necesaria para apreciar por qué merece la pena estudiar los imaginarios es todavía demasiado para la gran mayoría de los estudiantes. Muy pocos adultos, incluso con formación, son conscientes cuán profundamente nuestra sociedad depende de números que no representan cantidades, longitudes, áreas o cantidades de dinero. La tecnología más moderna, desde la luz eléctrica a las cámaras digitales, no podría haberse inventado sin ellos. Permíteme volver atrás a una pregunta crucial. ¿Por qué los cuadrados son siempre positivos?

En el Renacimiento, donde las ecuaciones eran generalmente reformuladas para hacer positivo todo número en ellas, no habrían expresado la pregunta de este modo. Habrían dicho que si sumas un número a un cuadrado, entonces tienes un número mayor, no se puede obtener cero. Pero incluso si permites que haya números negativos, como hacemos ahora, los cuadrados todavía tienen que ser positivos. Y este es el porqué.

Los números reales pueden ser positivos o negativos. Sin embargo, el cuadrado de cualquier número real, cualquiera que sea su signo, es siempre positivo, porque el producto de dos números negativos es positivo. Así tanto 3×3 como -3×-3 tienen el mismo resultado: 9. Por lo tanto 9 tiene dos raíces cuadradas, 3 y -3.

¿Qué pasa con -9? ¿Cuáles son sus raíces cuadradas?

No tiene ninguna.

Todo parece terriblemente injusto, los números positivos acaparan dos raíces cuadradas, mientras que los números negativos se quedan sin ellas. Es tentador cambiar la regla para multiplicar dos números negativos de modo que, por ejemplo,

$-3 \times -3 = -9$. Entonces los números positivos y negativos tendrían cada uno una raíz cuadrada; además, tendrían el mismo signo que su cuadrado, lo cual parece pulcro y ordenado. Pero esta línea de razonamiento tentadora tiene un inconveniente no buscado: echa por tierra las reglas habituales de la aritmética. El problema es que -9 ya es el resultado de 3×-3 , una consecuencia de las reglas habituales de la aritmética y un hecho que casi todo el mundo acepta contento. Si insistimos en que -3×-3 sea también -9 , entonces $-3 \times -3 = 3 \times -3$. Hay varios modos de comprobar que esto causa problemas, el más simple es dividir por -3 y obtenemos $3 = -3$.

Por supuesto puedes cambiar las reglas de la aritmética. Pero ahora todo se vuelve complicado y lioso. Una solución más creativa es conservar las reglas de la aritmética y extender el sistema de números reales permitiendo los imaginarios. Sorprendentemente —y nadie podía haber anticipado esto, solo tienes que seguir el pensamiento lógico— este paso audaz nos lleva a un sistema de números consistente y bello, con multitud de usos. Ahora todos los números excepto 0 tienen dos raíces cuadradas, siendo una la opuesta de la otra. Esto es cierto incluso para el nuevo tipo de números; una ampliación del sistema basta. Se tardó un tiempo en tener esto claro, pero en retrospectiva, tiene un aire de inevitabilidad. Los números imaginarios, imposibles como eran, se negaron a irse. Parecían no tener sentido, pero continuaban surgiendo en los cálculos. A veces el uso de los números imaginarios hacía los cálculos más simples, y el resultado era más completo y más satisfactorio. Siempre que se obtenía una respuesta usando los números imaginarios, pero no los involucraba explícitamente, se podía verificar independientemente, y resultaba ser correcta. Pero cuando la respuesta sí involucraba números imaginarios explícitos, parecía no tener sentido y, con frecuencia, era contradictoria lógicamente. El enigma se cocinó a fuego lento durante doscientos años y cuando finalmente rompió a hervir, los resultados fueron explosivos.

Cardano es conocido como un académico ludópata porque ambas actividades desempeñaron un papel importante en su vida. Era tanto un genio como un granuja. Su vida consiste en una serie desconcertante de altos, muy altos y bajos, muy bajos. Su madre intentó abortar, su hijo fue decapitado por matar a su esposa

(la del hijo) y él (Cardano) perdió jugando la fortuna familiar. Fue acusado de herejía por hacer el horóscopo de Jesús. Entre medias, también se convirtió en rector de la Universidad de Padua, fue elegido para el Colegio de Físicos de Milán, ganó 2.000 coronas de oro por curar el asma del arzobispo de Saint Andrew y recibió una pensión del papa Gregorio XIII. Inventó la cerradura con combinación y suspensiones universales para sostener un giroscopio, y escribió numerosos libros, incluyendo una autobiografía extraordinaria *De Vita Propria* (Mi propia vida). El libro que es relevante para nuestro relato es *Ars Magna*, de 1545. Su título significa «el gran arte» y se refiere al álgebra. En él, Cardano recopila las ideas algebraicas más avanzadas de su época, incluyendo métodos nuevos y espectaculares para resolver ecuaciones, algunos inventados por un estudiante suyo, algunos obtenidos de otros en circunstancias controvertidas.

El álgebra, en su sentido familiar de las escuelas de matemáticas, es un sistema para representar números simbólicamente. Sus raíces se remontan al griego Diofanto de Alejandría alrededor del 250 d.C., cuya *Arithmetica* empleaba símbolos para describir modos de resolver ecuaciones. La mayoría del trabajo era verbal: «encontrar dos números cuya suma es 10 y cuyo producto es 24».

date	author	notation
c.250	Diofanto	$\Delta^{\gamma} \alpha \zeta \beta \overset{\circ}{M} \gamma$
c.825	Al-Khowârizmî	potencia más dos veces más tres [en árabe]
1545	Cardano	cuadrado más dos veces más tres [en italiano]
1572	Bombelli	$3p \cdot 2 \overset{1}{\cup} p \cdot 1 \overset{2}{\cup}$
1585	Stevin	$3 + 2^{\oplus} + 1^{\oplus}$
1591	Viète	$x \text{ quadr.} + x \text{ 2} + 3$
1637	Descartes, Gauss	$xx + 2x + 3$
1670	Bachet de Méziriac	$Q + 2N + 3$
1765	Euler, moderno	$x^2 + 2x + 3$

TABLA 1. El desarrollo de la notación algebraica

Pero Diofanto resumió los métodos que usaba para encontrar las soluciones (en este caso 4 y 6) de manera simbólica. Los símbolos (véase la tabla 1) eran muy diferentes de los que usamos hoy en día, y la mayoría eran abreviaturas, pero fue

un comienzo. Cardano usaba palabras principalmente, con unos pocos símbolos para raíces y, de nuevo, los símbolos apenas se parecen a los que actualmente se usan. Autores posteriores se centraron, bastante caprichosamente, en la notación actual, la mayoría de la cual fue normalizada por Euler en sus numerosos libros de texto. Sin embargo, Gauss usaba xx en lugar de x^2 aún en 1800.

Los temas más importantes en el *Ars Magna* eran métodos nuevos para resolver ecuaciones cúbicas y de cuarto grado. Que son como ecuaciones cuadráticas, las cuales la mayoría de nosotros vimos en la escuela, pero más complicadas. Una ecuación cuadrática plantea una relación que envuelve una cantidad desconocida, normalmente se simboliza con la letra x , y su cuadrado, x^2 . Un ejemplo típico es:

$$x^2 - 5x + 6 = 0$$

De palabra se dice: «el cuadrado de la incógnita, menos 5 veces la incógnita, más 6 es igual a cero». Dada una ecuación con una incógnita, nuestra tarea es resolver la ecuación, es decir, encontrar el valor o valores de la incógnita que hacen la ecuación correcta.

Para un valor de x escogido aleatoriamente, esta ecuación lo normal es que sea falsa. Por ejemplo, si probamos con $x = 1$, entonces

$$x^2 - 5x + 6 = 1 - 5 + 6 = 2$$

que es distinto de cero. Pero para contadas elecciones de x , la ecuación es cierta. Por ejemplo, cuando $x = 2$, tenemos

$$x^2 - 5x + 6 = 4 - 10 + 6 = 0$$

¡Pero esta no es la única solución! Cuando $x = 3$, tenemos

$$x^2 - 5x + 6 = 9 - 15 + 6 = 0$$

también. Hay dos soluciones $x = 2$ y $x = 3$, y se puede demostrar que no hay otras.

Una ecuación cuadrática puede tener dos soluciones, una o ninguna (en números reales). Por ejemplo, $x^2 - 2x + 1 = 0$ tiene una única solución, $x = 1$, y $x^2 + 1 = 0$ no tiene solución en los números reales.

La obra maestra de Cardano proporciona métodos para solucionar ecuaciones cúbicas, que junto con x y x^2 también involucran el cubo de una incógnita, x^3 , y ecuaciones de grado cuatro, donde también aparece x^4 . El álgebra se hace muy complicada, incluso con simbolismos modernos ocupa una página o dos calcular las respuestas. Cardano no se metió con las ecuaciones de grado cinco, aquellas en las que aparece x^5 , porque no sabía cómo resolverlas. Mucho más tarde se probó que no existen soluciones (del tipo de las que Cardano hubiera querido); aunque se pueden calcular soluciones numéricas muy precisas en cualquier caso particular, no hay fórmula general para ellas, a menos que te inventes nuevos símbolos específicamente para la tarea.

Voy a escribir unas pocas fórmulas algebraicas, porque creo que el tema tiene más sentido si no intentamos evitarlas. No necesitas seguir los detalles, pero me gustaría mostrarte qué aspecto tiene todo esto. Usando símbolos modernos, podemos escribir la solución de Cardano para las ecuaciones cúbicas en un caso especial, cuando $x^3 + ax + b = 0$, donde a y b son unos números concretos (si x^2 está presente, un ingenioso truco nos libra de él, de modo que este caso en realidad vale para todos los casos). La respuesta es:

$$x = \sqrt[3]{-\frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}} + \sqrt[3]{-\frac{b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}$$

Esto puede parecer larguísimo, pero es mucho más simple que muchas fórmulas algebraicas. Nos dice cómo calcular la incógnita x averiguando el cuadrado de b y el cubo de a , sumando unas pocas fracciones y calculando un par de raíces cuadradas (el símbolo $\sqrt{}$) y un par de raíces cúbicas (el símbolo $\sqrt[3]{}$). La raíz cúbica de un número es cualquier número que haya que elevar al cubo para obtener ese número. El descubrimiento de la solución para las ecuaciones cúbicas involucra al menos a

otros tres matemáticos, uno de los cuales se quejó amargamente porque Cardano había prometido no revelar su secreto. La historia, aunque fascinante, es también complicada de contar aquí.¹⁹ La ecuación de grado cuatro fue resuelta por el alumno de Cardano, Lodovico Ferrari. Te ahorraré la todavía más complicada fórmula para ecuaciones de grado cuatro.

Los resultados presentados en el *Ars Magna* eran un triunfo matemático, la culminación de una historia que abarcaba milenios. En Babilonia sabían cómo resolver ecuaciones cuadráticas alrededor del 1500 a.C., quizá antes. En la Grecia Clásica y Omar Khayyam sabían métodos geométricos para resolver ecuaciones cúbicas, pero soluciones algebraicas a ecuaciones cúbicas, dejando aparte las de grado cuatro, no tenían precedentes. De un golpe, los matemáticos aventajaron sus orígenes clásicos.

No obstante, había una pequeña pega. Cardano se dio cuenta de ella, y varias personas trataron de explicarla, todos fracasaron. A veces el método funciona de manera brillante; otras veces, la fórmula es tan enigmática como el oráculo de Delfos. Supón que aplicamos la fórmula de Cardano a la ecuación $x^3 - 15x - 4 = 0$. El resultado es:

$$x = \sqrt[3]{2 + \sqrt{-121}} + \sqrt[3]{2 - \sqrt{-121}}$$

¹⁹ En 1535, los matemáticos Antonio Fior y Niccolò Fontana (apodado Tartaglia, «el tartamudo») participaron en un concurso público. Se pusieron ecuaciones cúbicas para resolver el uno al otro, y Tartaglia venció a Fior rotundamente. En esa época, las ecuaciones cúbicas estaban clasificadas en tres tipos distintos, porque los números negativos no se reconocían. Fior sabía cómo resolver solo un tipo, inicialmente Tartaglia sabía cómo resolver un tipo diferente, pero poco antes del concurso averiguó cómo resolver los otros tipos. Entonces le puso a Fior solo los tipos que sabía que Fior no podría resolver. El concurso llegó a los oídos de Cardano, que estaba trabajando en su libro de texto de álgebra, y se dio cuenta de que Fior y Tartaglia sabían cómo resolver ecuaciones cúbicas. Este resultado mejoraría enormemente el libro, así que le pidió a Tartaglia que le revelase sus métodos.

Finalmente Tartaglia le reveló el secreto, más tarde afirmando que Cardano había prometido no hacerlo nunca público. Pero el método apareció en *Ars Magna*, así que Tartaglia acusó a Cardano de plagio. Sin embargo, Cardano tenía una excusa, y también tenía una buena razón para romper su promesa. Su estudiante Lodovico Ferrari había encontrado cómo solucionar ecuaciones de cuarto grado, un descubrimiento, igualmente, novedoso y espectacular, y Cardano lo quería también en su libro. Sin embargo, el método de Ferrari necesitaba la solución de una ecuación cónica asociada, así que Cardano no podía publicar el trabajo de Ferrari sin publicar también el de Tartaglia.

Luego supo que Fior era un estudiante de Scipio del Ferro, de quien se rumoreaba que había solucionado los tres tipos de ecuaciones cúbicas, y le había contado a Fior solo la solución para un tipo. Los artículos no publicados de Del Ferro estaban en posesión de Annibale del Nave. De modo que Cardano y Ferrari fueron a Bolonia en 1543 a consultar a Del Nave y en los papeles encontraron las soluciones para los tres tipos. Así que Cardano podía, de manera honesta, decir que estaba publicando el método de Del Ferro, no el de Tartaglia. Tartaglia todavía se sintió traicionado y publicó una larga y amarga diatriba contra Cardano. Ferrare le retó a un debate público y le ganó sin despeinarse. Tartaglia nunca recuperó realmente su reputación después de eso.

Como -121 es negativo, no tiene raíz cuadrada. Para agravar el misterio, hay una solución perfectamente buena, $x = 4$. La fórmula no la da.

Se arrojó algo de luz en 1572 cuando Bombelli publicó *L'Algebra*. Su objetivo principal era clarificar el libro de Cardano, pero cuando llegó a este particular tema peliagudo, descubrió algo que Cardano había pasado por alto. Si ignoras lo que significa el símbolo, y tan solo realizas los cálculos rutinarios, las reglas estándar del álgebra muestran que:

$$(2 + \sqrt{-1})^3 = 2 + \sqrt{-121}$$

Por lo tanto, estás autorizado a escribir:

$$\sqrt[3]{2 + \sqrt{-121}} = 2 + \sqrt{-1}$$

De manera similar:

$$\sqrt[3]{2 + \sqrt{-121}} = 2 - \sqrt{-1}$$

Ahora la fórmula que desconcertó a Cardano puede reescribirse como:

$$(2 + \sqrt{-1}) + (2 - \sqrt{-1})$$

Que es igual a 4 porque las raíces cuadradas problemáticas se anulan. Así, los cálculos formales y sin sentido de Bombelli conseguían la respuesta correcta. Que era un número real perfectamente normal.

De alguna manera, pretendiendo que las raíces cuadradas de números negativos tienen sentido, incluso aunque obviamente no lo tienen, se podía llegar a respuestas sensatas. ¿Por qué?

Para responder a esta pregunta, los matemáticos tuvieron que desarrollar modos buenos de pensar en las raíces cuadradas de cantidades negativas, y hacer cálculos con ellas. Escritores anteriores, entre ellos Descartes y Newton, interpretaron estos números «imaginarios» como un signo de que un problema no tenía solución. Si querías encontrar un número cuyo cuadrado fuese menos uno, la solución formal «raíz cuadrada de menos uno» era imaginaria, así que no existía solución. Pero los cálculos de Bombelli implicaban que había más para los imaginarios que eso. Podían usarse para encontrar soluciones, podían surgir como parte del cálculo de soluciones que sí existían.

Leibniz no tenía dudas sobre la importancia de los números imaginarios. En 1702 escribió: «El Espíritu Santo encontró una salida sublime en esa maravilla del análisis, ese presagio del mundo ideal, ese anfibio entre el ser y el no ser, la cual llamamos la raíz imaginaria de la unidad negativa». Pero la elocuencia de esta afirmación fracasa en ocultar un problema fundamental: no tenía ni idea de qué eran realmente los números imaginarios.

Una de las primeras personas que planteó una representación sensata de los números complejos fue Wallis. La imagen de los números reales extendiéndose a lo largo de una línea, como puntos marcados en una regla, era ya algo común. En 1673, Wallis sugirió que el número complejo $x + iy$ debería pensarse como un punto en un plano. Dibujó una línea en el plano e identificó puntos en esta línea con los números reales en el modo habitual. Luego pensó en $x + iy$ como un punto que está a un lado de la línea a una distancia y del punto x .

La idea de Wallis fue ignorada totalmente, o peor, criticada. François Daviet de Foncenex, escribiendo sobre los imaginarios en 1758, dijo que pensar en los imaginarios a medida que formaban una línea en ángulos rectos con la línea real no tenía sentido. Pero finalmente la idea resurgió en una forma ligeramente más explícita. De hecho, tres personas propusieron exactamente el mismo método para representar números complejos en intervalos de unos pocos años (figura 18). Una fue un topógrafo noruego, otra un matemático francés y otra un matemático alemán. Respectivamente eran: Caspar Wessel, quien lo publicó en 1797, Jean-Robert Argand en 1806, y Gauss en 1811. Básicamente decían lo mismo que Wallis, pero añadían una segunda línea a la imagen, un eje imaginario en ángulo recto con

el real. A lo largo de este segundo eje viven los números imaginarios: i , $2i$, $3i$, etcétera. Un número complejo general, como $3 + 2i$, se encuentra en el plano, tres unidades a lo largo del eje real y dos a lo largo del imaginario.

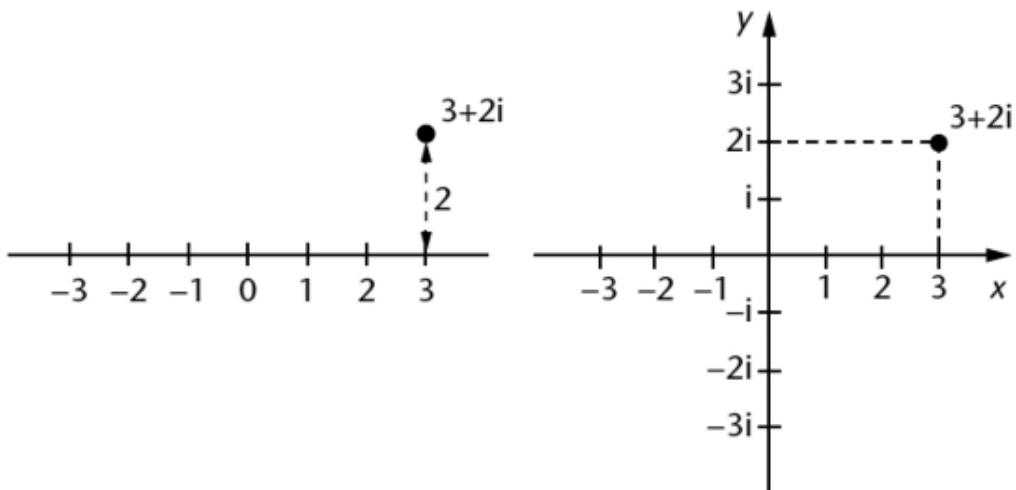


FIGURA 18. El plano complejo. A la izquierda: según Wallis. A la derecha: según Wessel, Argand y Gauss.

Esta representación geométrica funcionaba muy bien, pero no explicaba por qué los números complejos forman un sistema lógicamente consistente. No nos dice en qué sentido son números. Tan solo proporciona un modo de visualizarlos. Esto define tanto qué es un número complejo, como un dibujo de una línea recta define un número real. Proporcionaba algún tipo de apoyo psicológico, una conexión ligeramente artificial entre esos imaginarios locos y el mundo real, pero nada más. Lo que convenció a los matemáticos de que deberían tomarse en serio los números imaginarios no fue una descripción lógica de qué eran. Fue una prueba abrumadora de que fuera lo que fueran, las matemáticas podían hacer un buen uso de ellos. Uno no hace preguntas difíciles sobre las bases filosóficas de una idea cuando la está usando todos los días para resolver problemas y puede ver que da las respuestas correctas. Las preguntas fundamentales todavía tienen algún interés, por supuesto, pero se quedan en un segundo plano respecto a asuntos pragmáticos sobre usar la nueva idea para resolver problemas antiguos y nuevos.

Los números imaginarios, y el sistema de números complejos que engendran,

consolidaron su lugar en las matemáticas cuando unos pocos pioneros volvieron su atención al análisis complejo: cálculo (capítulo 3) pero con números complejos en lugar de con reales. El primer paso era extender todas las funciones habituales: potenciales, logarítmicas, exponenciales, trigonométricas, al reino de los complejos. ¿Qué es $\sin z$ cuando $z = x + iy$ es complejo? ¿Qué es e^z o $\log z$?

Lógicamente, estas cosas pueden ser lo que queramos que sean. Estamos operando en un nuevo dominio donde las viejas ideas no se aplican. No tiene mucho sentido, por ejemplo, pensar en un triángulo rectángulo cuyos lados tiene longitudes complejas, así que la definición geométrica de la función seno es irrelevante. Podemos respirar hondo, insistir en que $\sin z$ tiene su valor habitual cuando z es real, pero que es igual a 42 cuando z no es real; listo. Pero sería una definición bastante tonta, no porque sea imprecisa, sino porque no soporta una relación sensata con la original para números reales. Un requerimiento para extender una definición es que debe estar acorde con la anterior cuando se aplica a los números reales, pero eso no es suficiente. Es cierto para mi tonta extensión del seno. Otro requerimiento es que el nuevo concepto debería conservar tantas características del antiguo como sea posible, debería, de algún modo, ser «natural».

¿Qué propiedades del seno y del coseno queremos preservar? Es de suponer que nos gustaría que todas las bellas fórmulas de la trigonometría sigan siendo válidas, como $\sin 2z = 2 \sin z \cos z$. Esto impone una restricción pero no ayuda. Una propiedad más interesante, obtenida usando el análisis (la formulación rigurosa del cálculo) es la existencia de una serie infinita:

$$\sin z = z - \frac{z^3}{1 \cdot 2 \cdot 3} + \frac{z^5}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} + \frac{z^7}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7} + \dots$$

(La suma de dicha serie está definida para ser el límite de la suma de un número finito de términos a medida que el número de términos aumenta indefinidamente.) Hay una serie parecida a esta para el coseno:

$$\cos z = 1 - \frac{z^2}{1 \cdot 2} + \frac{z^4}{1 \cdot 2 \cdot 3 \cdot 4} - \frac{z^6}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} + \dots$$

y las dos están obviamente relacionadas de algún modo con la serie para la exponencial:

$$e^z = 1 + z + \frac{z^2}{1 \cdot 2} + \frac{z^3}{1 \cdot 2 \cdot 3} + \frac{z^4}{1 \cdot 2 \cdot 3 \cdot 4} + \dots$$

Estas series pueden parecer complicadas, pero tienen una característica atractiva: sabemos cómo hacer que tengan sentido para los números complejos. Todo lo que involucran son potencias de enteros (que obtenemos repitiendo una multiplicación) y un tema técnico de convergencia (dando sentido a la suma infinita). Ambos temas se extienden de manera natural al reino de los complejos y tienen todas las propiedades esperadas. Así que podemos definir senos y cosenos de números complejos usando la misma serie que funciona en el caso real.

Ya que todas las fórmulas habituales de trigonometría son consecuencia de esta serie, estas fórmulas automáticamente siguen funcionando bien. Al igual que realidades básicas del cálculo, tales como «la derivada del seno es el coseno». También $e^{z+w} = e^z e^w$. Esto es tan grato que a los matemáticos les alegró haberse decidido por las definiciones de las series. Y una vez hicieron eso, muchísimo más necesariamente tenía que encajar con ello. Si sigues tu intuición, podrás descubrir adónde lleva.

Por ejemplo, esas tres series se parecen mucho. De hecho, si remplazas z por iz en la serie para la exponencial, puedes dividir la serie que resulta en dos partes, y lo que obtienes son precisamente las series para el seno y el coseno. Así la definición de series implica que:

$$e^{iz} = \cos z + i \sin z$$

También puedes expresar tanto el seno como el coseno usando exponentiales:

$$\cos z = \frac{e^{iz} + e^{-iz}}{2} \quad \operatorname{sen} z = \frac{e^{iz} - e^{-iz}}{2i}$$

Esta relación escondida es extraordinariamente bella. Pero nunca habrías sospechado que algo así pudiese existir, si permanecieses atascado en el reino de los reales. Semejanzas curiosas entre fórmulas trigonométricas y exponenciales (por ejemplo, sus series infinitas) permanecerían justo como eso. Vistas a través de las gafas de los complejos, de repente todo encaja.

Una de las ecuaciones más bellas, aunque enigmática, en el mundo de las matemáticas surge casi por accidente. En las series trigonométricas, el número z (cuando es real) tiene que medirse en radianes, para lo cual los 360° de un círculo completo pasan a ser 2π radianes. En particular, el ángulo 180° es π radianes. Además, $\operatorname{sen} \pi = 0$ y $\cos \pi = -1$. Por lo tanto:

$$e^{i\pi} = \cos \pi + i \operatorname{sen} \pi = -1$$

El número imaginario i une los dos números más notables en matemáticas, e y π , en una única y elegante ecuación. Si nunca antes has visto esto y tienes algo de sensibilidad matemática, los pelillos en tu cuello se levantan y pican por toda tu columna vertebral. Esta ecuación, que se atribuye a Euler, normalmente aparece en la cima de las listas en encuestas para la ecuación más bella en matemáticas. Eso no significa que sea la ecuación más bella, sino muestra cuánto la aprecian los matemáticos.

Armados con las funciones complejas y conociendo sus propiedades, los matemáticos del siglo XIX descubrieron algo notable: podían usar estas cosas para resolver ecuaciones diferenciales en física matemática. Podían aplicar el método a la electricidad estática, magnetismo y dinámica de fluidos. No solo eso: era fácil.

En el capítulo 3 hablamos de funciones (reglas matemáticas que asignan a un número dado, un número que le corresponde, como su cuadrado o seno.) Las funciones complejas están definidas del mismo modo, pero ahora se permite a los números que están involucrados que sean complejos. El método para resolver

ecuaciones diferenciales era maravillosamente sencillo. Todo lo que tenías que hacer era tomar una función compleja, llámala $f(z)$, y dividirla en sus partes real e imaginaria:

$$f(z) = u(z) + iv(z)$$

Ahora tienes dos funciones de valores reales, u y v , definidas para cualquier z en el plano complejo. Además, cualquiera que sea la función con la que empieces, estas dos funciones que la componen satisfacen ecuaciones diferenciales encontradas en física. En una interpretación flujo-fluido, por ejemplo, u y v determinan las líneas de flujo. En una interpretación electrostática, las dos componentes determinan el campo eléctrico y cómo una pequeña partícula cargada se movería; en una interpretación magnética, determinan el campo magnético y las líneas de fuerza.

Daré solo un ejemplo: un imán de barra. La mayoría de nosotros recordamos haber visto un experimento famoso en el cual un imán se colocaba tras una hoja de papel y limaduras de hierro se esparcían por toda la hoja. Automáticamente se alineaban para mostrar las líneas de la fuerza magnética asociada con el imán, los caminos que un imán minúsculo de prueba seguiría si se colocase en el campo magnético. Las curvas tienen el aspecto que se ve en la figura 19 (a la izquierda).

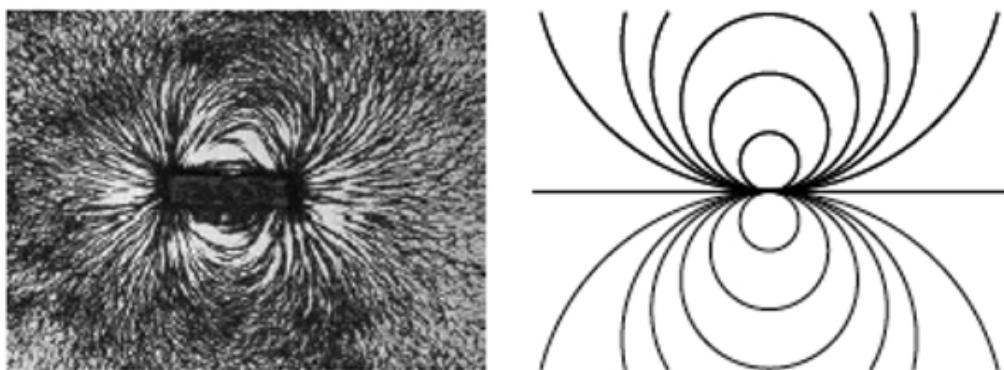


FIGURA 19. A la izquierda: campo magnético de un imán de barra. A la derecha: campo obtenido usando análisis complejo.

Para obtener la imagen usando funciones complejas, tan solo hacemos $f(z) = 1/z$.

Las líneas de fuerza resultan ser círculos tangentes al eje real, como en la figura 19 (a la derecha). Este es el aspecto que tendrían los campos magnéticos de una barra de imán muy pequeña. Una elección más complicada de la función se corresponde con un imán de un tamaño finito. Escojo esta función para que todo sea lo más simple posible.

Esto era maravilloso. Había infinidad de funciones con las que trabajar. Decidías en qué función fijarte, encontrabas sus partes real e imaginaria, averiguabas su geometría... y, iquién lo iba a decir!, habías resuelto un problema en magnetismo, o electricidad, o dinámica de fluidos. La experiencia pronto dijo qué función usar para cada problema. El logaritmo era una fuente, menos el logaritmo un sumidero a través del cual el fluido desaparecía como el desagüe en el fregadero de una cocina, ¡ multiplicado por el logaritmo era un vórtice donde el fluido da vueltas y más vueltas... ¡Era magia! Había un método que podía producir como churros solución tras solución a problemas que de otro modo serían opacos. Además venía con una garantía de éxito, y si te preocupaba todo el tema del análisis complejo, podías comprobar directamente que los resultados que obtenías realmente representaban soluciones.

Esto era solo el principio. Además de soluciones especiales, podía probar principios generales, patrones escondidos en las leyes físicas. Podías analizar ondas y resolver ecuaciones diferenciales. Podías transformar formas en otras formas, usando ecuaciones complejas y las mismas ecuaciones transformaban las líneas de flujo alrededor de ellas.

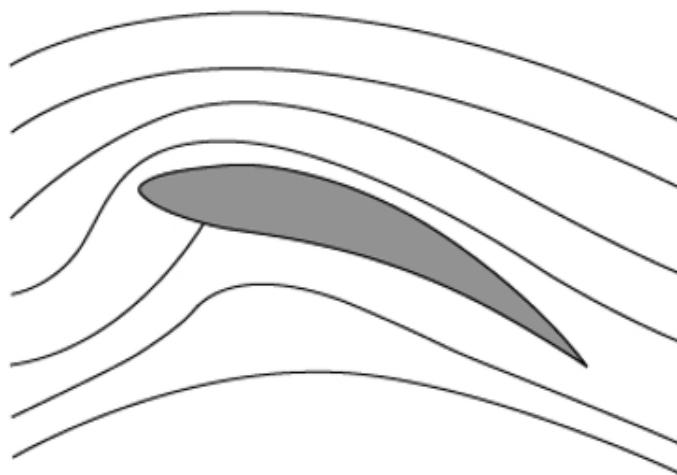


FIGURA 20. Fluido que pasa delante de un ala obtenido a partir de la transformación de Joukowski.

El método estaba limitado a sistemas en el plano, porque era donde un número complejo vivía de manera natural, pero el método era un regalo divino cuando, anteriormente, incluso problemas en el plano estaban fuera del alcance. Hoy en día, se enseña a todo ingeniero cómo usar el análisis complejo para resolver problemas prácticos en los primeros cursos de la universidad. La transformación de Joukowski $z + 1/z$ transforma un círculo en un perfil alar, la sección transversal de un ala de avión rudimentaria (véase la figura 20). Por tanto, convierte el flujo de delante de un círculo, fácil de encontrar si sabes los trucos del oficio, en el flujo de delante de un perfil alar. Este cálculo, y mejoras más realistas, fueron importantes en los principios de la aerodinámica y el diseño de aviones.

Esta riqueza de experiencia práctica hizo los temas de fundamentos irrelevantes. ¿Por qué mirar la boca de un caballo regalado? Tenía que haber un significado sensato para los números complejos, de otra manera no funcionarían. La mayoría de los científicos y matemáticos estaban mucho más interesados en sacar el oro que en establecer exactamente de dónde venía y qué lo distinguía del oro falso. Pero unos pocos persistieron. Finalmente, el matemático irlandés William Rowan Hamilton terminó definitivamente con todo el asunto. Tomó la representación geométrica propuesta por Wessel, Argand y Gauss y la expresó en coordenadas. Un número complejo era un par de números reales (x, y) . Los números reales eran los de la forma $(x, 0)$. El imaginario i era $(0, 1)$. Había fórmulas sencillas para sumar y multiplicar estos pares. Si te preocupaba alguna ley del álgebra, tal como la propiedad conmutativa $ab = ba$, podías de manera rutinaria calcular ambos lados como pares, y asegurarte que eran lo mismo (lo eran). Si identificabas $(x, 0)$ con una simple x , incrustabas los números reales en los complejos. Mejor todavía, $x + iy$ funcionaba como el par (x, y) .

Esto no era solo una representación, sino una definición. Un número complejo, dijo Hamilton, no es nada más ni nada menos que un par de números reales ordinarios. Lo que los hizo tan útiles fue una elección inspirada en las reglas para sumarlos y multiplicarlos. Esto realmente era algo común; era cómo los usabas lo que producía

la magia. Con esta genialidad simple, Hamilton zanjó siglos de una discusión acalorada y debate filosófico. Pero para entonces, los matemáticos se habían acostumbrado tanto a trabajar con números complejos y funciones que a nadie le preocupaba ya. Todo lo que necesitabas recordar era que $i^2 = -1$.

Capítulo 6
Agujeros, nudos y movimientos
Fórmula de Euler para los poliedros

$$F - E + V = 2$$

numero de caras numero de aristas número de vértices

¿Qué dice?

El número de caras, aristas y vértices de un sólido no son independientes, sino que están relacionados de un modo sencillo.

¿Por qué es importante?

Distingue entre sólidos con diferentes topologías usando el ejemplo más temprano de un invariante topológico. Esto allanó el camino para técnicas más generales y potentes, creando una rama nueva en las matemáticas.

¿Qué provocó?

Una de las áreas más importantes y potentes de la matemática pura: la topología, la cual estudia propiedades geométricas que no cambian tras deformaciones continuas. Los ejemplos incluyen superficies, nudos y enlaces. La mayoría de las aplicaciones son indirectas, pero su influencia en la sombra es vital. Nos ayuda a entender cómo las enzimas actúan sobre el ADN en una célula y por qué el movimiento de los cuerpos celestes puede ser caótico.

A medida que el siglo XIX se aproximaba a su fin, los matemáticos empezaban a desarrollar un nuevo tipo de geometría, una en la cual los conceptos familiares como longitud y ángulo no jugaban ningún papel en absoluto y no se hacía distinción entre triángulos, cuadrados y círculos. Inicialmente se llamó *análisis situs*, el análisis de la posición, pero los matemáticos rápidamente fijaron otro nombre:

topología.

La topología tiene sus raíces en un patrón numérico curioso del que Descartes se dio cuenta en 1639 cuando pensaba en los cinco sólidos regulares de Euclides. Descartes era un polímata nacido en Francia que pasó la mayoría de su vida en la entonces República Holandesa, los Países Bajos actuales. Su fama proviene principalmente de su filosofía, la cual resultó tan influyente que durante mucho tiempo la filosofía occidental consistía en gran parte en responder a Descartes. No siempre estaba de acuerdo, como puedes suponer, pero, no obstante, sí motivada por sus argumentos. Su cita jugosa *cogito ergo sum* (Pienso, luego existo) forma parte del conocimiento cultural popular. Pero los intereses de Descartes se extendían más allá de la filosofía hasta la ciencia y las matemáticas.

En 1639, Descartes volcó su atención en los sólidos regulares y fue entonces cuando se dio cuenta de un patrón numérico curioso. Un cubo tiene 6 caras, 12 aristas y 8 vértices, la operación $6 - 12 + 8$ es igual a 2. Un dodecaedro tiene 12 caras, 30 aristas y 20 vértices, la operación $12 - 30 + 20 = 2$. Un icosaedro tiene 20 caras, 30 aristas y 12 vértices, la operación $20 - 30 + 12 = 2$. La misma relación se cumple para el tetraedro y el octaedro. De hecho, se cumple para un sólido de cualquier forma, regular o no. Si el sólido tiene C caras, A aristas y V vértices, entonces $C - A + V = 2$. Descartes vio esta fórmula como una curiosidad menor y no la publicó. Solo mucho más tarde los matemáticos realmente vieron esta ecuación pequeña y simple como uno de los primeros pasos vacilantes hacia la historia del gran éxito en las matemáticas del siglo XX, el ascenso inexorable de la topología. En el siglo XIX, los tres pilares de la matemática pura eran el álgebra, el análisis y la geometría. A finales del siglo XX, eran el álgebra, el análisis y la topología.

La topología es calificada como «la geometría de la lámina elástica» porque es el tipo de geometría que sería apropiada para figuras dibujadas en una página elástica, de modo que las líneas se pueden curvar, contraer o estirar, y los círculos pueden aplastarse de modo que se conviertan en triángulos o cuadrados. Lo único que importa es la continuidad: no está permitido partir la hoja. Podría parecer sorprendente que algo tan raro pudiese tener alguna importancia, pero la continuidad es un aspecto básico del mundo natural y una característica

fundamental de las matemáticas. Hoy en día, generalmente usamos la topología indirectamente, como una de las técnicas matemáticas entre otras muchas. No encuentras nada obviamente topológico en tu cocina. Sin embargo, una compañía japonesa comercializó un lavaplatos caótico, el cual, según su departamento de marketing, limpiaba los platos más eficientemente, y nuestra comprensión del caos depende de la topología. También lo hacen aspectos importantes de la teoría cuántica de campos y la simbólica molécula del ADN. Pero cuando Descartes contó las características más obvias de los sólidos regulares y se dio cuenta de que no eran independientes, todo esto pertenecía a un futuro lejano.

Se le dejó al incansable Euler, el matemático más prolífico en la historia, probar y publicar esta relación, lo cual hizo en 1750 y 1751. Esbozaré una versión moderna. La expresión $C - A + V$ puede parecer bastante arbitraria, pero tiene una estructura muy interesante. Las caras (C) son polígonos de dimensión 2, las aristas (A) son líneas, así que tienen dimensión 1, y los vértices (V) son puntos, de dimensión 0. Los signos de la expresión alternan, $+ - +$, con $+$ asignada a las características de dimensión par y $-$ a aquellas de dimensión impar. Esto implica que puedes simplificar un sólido fusionando sus caras o eliminando sus aristas y vértices y estos cambios no alterarán el número $C - A + V$ siempre que cada vez que elimines una cara, también elimines una arista o cada vez que elimines un vértice, también elimines una arista. Los signos que se alternan quieren decir que los cambios de este tipo se compensan.

Ahora explicaré cómo esta estructura inteligente hace que la prueba funcione. La figura 21 muestra las etapas clave. Considera tu sólido. Defórmalo para obtener una bella esfera redondeada, con sus aristas siendo curvas en esa esfera. Si dos caras se encuentran en un eje común, entonces puedes eliminar la arista y fundir las dos caras en una. Ya que esta unión reduce tanto C como A en una unidad, $C - A + V$ no cambia. Sigue haciendo esto hasta que lo reduzcas a una sola cara, la cual cubre casi toda la esfera. Además de esta cara, te quedas solo con aristas y vértices. Estos deben formar un árbol, una cadena sin curvas cerradas, porque cualquier curva cerrada en una esfera separa al menos dos caras: una dentro y otra fuera. Las ramas de este árbol son las aristas que quedan del sólido, y se unen en los vértices que quedan. En esta etapa solo queda una cara: la esfera completa, menos

el árbol. Algunas ramas de este árbol se conectan con otras ramas en ambos extremos, pero algunos, en los extremos, terminan en un vértice, al cual ninguna otra rama está vinculada. Si eliminas una de esas ramas finales junto con ese vértice, entonces el árbol se hace más pequeño, pero como tanto A como V decrecen una unidad, $C - A + V$ de nuevo no sufre cambios.



FIGURA 21. Etapas clave en la simplificación de un sólido. De izquierda a derecha: (1) Principio. (2) Fusionando caras adyacentes. (3) El árbol que queda cuando todas las caras se han fusionado. (4) Eliminando una arista y un vértice del árbol. (5) Fin.

Este proceso continúa hasta que te quedas con un único vértice sobre una esfera, que por lo demás no tiene ninguna característica especial. Ahora $V = 1$, $A = 0$ y $C = 1$. De modo que, $C - A + V = 1 - 0 + 1 = 2$. Pero como cada paso deja sin cambios a $C - A + V$, su valor al principio debe haber sido también 2, que es lo que queremos probar.

Es una idea ingeniosa y contiene el germen de un principio de gran alcance. La prueba tiene dos ingredientes. Uno es un proceso de simplificación: eliminar tanto una cara como una arista adyacente o un vértice y una arista con la que se corta. El otro es un invariante, una expresión matemática que permanece sin cambios siempre que lleves a cabo un paso en el proceso de simplificación. Cuando estos dos ingredientes coexisten, puedes calcular el valor del invariante para cualquier objeto inicial simplificándolo tanto como puedas y luego calculando el valor del invariante para esta versión simplificada. Como es un invariante, los dos valores deben ser iguales. Como el resultado final es sencillo, el invariante es fácil de calcular.

Ahora tengo que admitir que me he guardado un asunto técnico en la manga. La fórmula de Descartes en realidad no se cumple para cualquier sólido. El sólido más conocido que falla es un marco de fotos. Piensa en un marco de fotos hecho con cuatro trozos de madera, cada uno rectangular en su corte transversal, unidos en

las cuatro esquinas por ingletes de 45° como en la figura 22 (izquierda). Cada trozo de madera aporta 4 caras, así que $C = 16$. Cada trozo también aporta 4 aristas, pero el inglete crea 4 más en cada esquina, así que $A = 32$. Cada esquina comprende 4 vértices, así que $V = 16$. Por lo tanto $C - A + V = 0$.

¿Qué es lo que está mal?

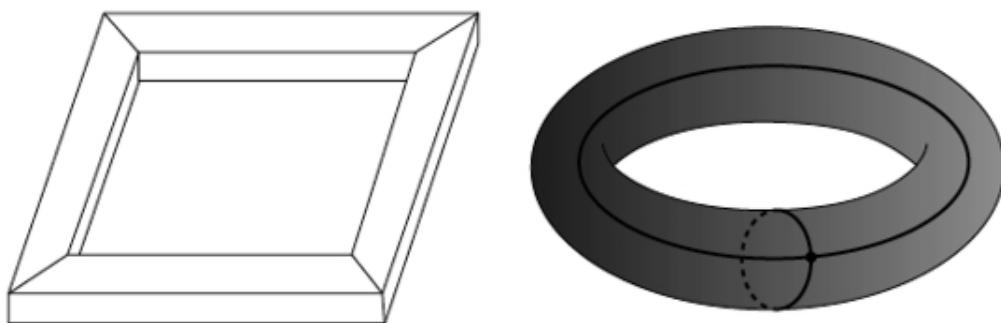


FIGURA 22. A la izquierda: un marco con $C - A + V = 0$. A la derecha: configuración final cuando el marco es redondeado y simplificado.

No hay ningún problema con $C - A + V$ siendo invariante. Tampoco es demasiado problema el proceso de simplificación. Pero si lo aplicas en el marco, siempre anulando una cara con una arista, o un vértice con una arista, entonces la configuración final simplificada no es un único vértice sobre una única cara. Anulando los elementos unos con otros del modo más obvio, lo que obtienes es la figura 22 (derecha) con $C = 1$, $V = 1$, $A = 2$. He redondeado las caras y las aristas por razones que rápidamente se harán evidentes. En esta etapa eliminar una arista solo funde la única cara restante consigo misma, así que los cambios en los números ya no se cancelan. Esta es la razón por la que paramos, pero tenemos el éxito asegurado de todos modos, para esta configuración, $C - A + V = 0$. De manera que el método funciona perfectamente. Es tan solo que da un resultado diferente para el marco. Debe haber alguna diferencia fundamental entre un marco y un cubo y el invariante $C - A + V$ lo recoge.

La diferencia resulta ser topológica. Antes, en mi versión de la prueba de Euler, te dije que consideraras un sólido y «defórmalo para obtener una bella esfera redondeada». Pero eso no es posible para el marco. No tiene la forma de una esfera

incluso después de haberlo simplificado. Es un toro, que parece un flotador hinchable con un agujero en medio. El agujero es también claramente visible en la forma original, es donde iría la foto. Una esfera, por el contrario, no tiene agujeros. El agujero en el marco es la razón de que el proceso de simplificación nos lleve a un resultado diferente. Sin embargo, finalmente podemos cantar victoria porque $C - A + V$ es todavía invariante. De modo que la prueba nos dice que cualquier sólido que se deforma en un toro satisfará la ecuación ligeramente diferente $C - A + V = 0$. En consecuencia, tenemos las bases de una prueba rigurosa de que un toro no puede convertirse en una esfera, esto es, las dos superficies son topológicamente diferentes.

Por supuesto esto es intuitivamente obvio, pero ahora podemos apoyar la intuición en la lógica. Del mismo modo que Euclides empezó a partir de propiedades obvias de puntos y rectas y los formalizó en una teoría de la geometría rigurosa, los matemáticos de los siglos XIX y XX podían ahora desarrollar una teoría de la topología formal y rigurosa.



FIGURA 23. A la izquierda: toro de 2 agujeros. A la derecha: toro de tres agujeros.

No había que ser brillante para saber por dónde empezar. Existen sólidos como un toro pero con dos o más agujeros, como en la figura 23, y el mismo invariante debería decírnos algo útil sobre ellos. Resulta que cualquier sólido que se deforma en un toro de 2 agujeros satisface $C - A + V = -2$, cualquier sólido que se deforma en un toro de 3 agujeros satisface $C - A + V = -4$ y, en general, cualquier sólido deformable en un toro de g -agujeros satisface $C - A + V = 2 - 2g$. El símbolo de g viene por «género», el nombre técnico para el número de agujeros. Continuar con la línea de pensamiento que Descartes y Euler empezaron lleva a una conexión entre una propiedad cuantitativa de los sólidos, el número de caras, vértices y aristas y una propiedad cualitativa, tener agujeros. Llamamos a $C - A + V$ la característica de Euler del sólido; observa que depende solo del sólido que estemos

considerando y no de cómo recortemos sus caras, aristas y vértices. Esto lo hace una característica intrínseca del sólido en sí mismo.

De acuerdo, contamos el número de agujeros, una operación cuantitativa, pero «agujero» en sí mismo es cualitativo en el sentido de que no es obviamente una característica del sólido en absoluto. Intuitivamente, es una región en el espacio donde no hay sólido. Pero no cualquier región. Después de todo, esa descripción aplica a todo el espacio que rodea al sólido, y nadie lo consideraría todo como un agujero. Y también aplica a todo el espacio que rodea a la esfera... la cual no tiene un agujero. De hecho, cuanto más piensas en qué es un agujero, más te das cuenta de que es un poco peliagudo definir uno. Mi ejemplo favorito para mostrar justo cómo de confuso se vuelve todo es la forma en la figura 24, conocida como «agujero a través de un agujero en un agujero». Aparentemente puedes pasar un agujero por otro agujero, lo cual es realmente un agujero en un tercer agujero.

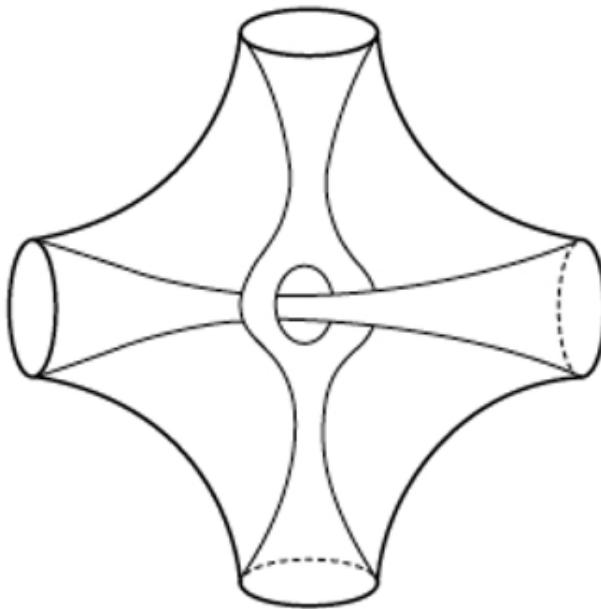


FIGURA 24. Agujero a través de un agujero en un agujero.

En ese camino se encuentra la locura.

No importaría mucho si los sólidos con agujeros en ellos nunca apareciesen en un lugar importante. Pero a finales del siglo XIX, aparecían por todas partes en matemáticas: en el análisis complejo, en la geometría algebraica y en la geometría

diferencial de Riemann. Peor, se hicieron protagonistas equivalencias de sólidos de dimensiones mayores en todas las áreas de las matemáticas pura y aplicada. Como ya apunté, la dinámica del Sistema Solar necesita 6 dimensiones por cuerpo y tienen análogos con agujeros de dimensiones mayores. De algún modo, era necesario traer un atisbo de orden al área. Y la respuesta resultó ser... invariantes. La idea de un invariante topológico se remonta al trabajo de Gauss en magnetismo. Estaba interesado en cómo las líneas del campo magnético y eléctrico podían vincularse unas con otras, y definió un número de enlace, que cuenta cuántas veces una línea de campo se enrosca con otra. Esto es un invariante topológico; permanece igual si las curvas son continuamente deformadas. Encontró una fórmula para este número usando cálculo integral y expresó muchas veces su deseo de una comprensión mejor de las «propiedades básicas de la geometría» de diagramas. No es coincidencia que las primeras incursiones serias en dicho entendimiento viniesen a través del trabajo de uno de los estudiantes de Gauss, Johann Listing, y el asistente de Gauss, August Möbius. El *Vorstudien zur Topologie* (Estudios en Topología) de Listing de 1847 introdujo la palabra «topología» y Möbius hizo explícito el papel de las transformaciones continuas.

Listing tuvo una idea brillante: buscar generalizaciones de la fórmula de Euler. La expresión $C - A + V$ es un invariante combinatorio: una característica de un modo específico de describir un sólido, basado en recortar las caras, aristas y vértices. El número g de agujeros es un invariante topológico: algo que no cambia no importa cómo se deforme el sólido mientras que la deformación sea continua. Un invariante topológico captura una característica cualitativa conceptual de una forma, uno combinatorio proporciona un método para calcularlo. Los dos juntos son muy poderosos, porque podemos usar el invariante conceptual para pensar sobre las formas y la versión combinatoria para precisar de qué estamos hablando.

De hecho, la fórmula nos permite esquivar el difícil asunto de definir «agujero». En su lugar, definimos «el número de agujeros» como un paquete, sin definir agujero o contar cuántos hay. ¿Cómo? Fácil. Tan solo reescribe la versión generalizada de la fórmula de Euler $C - A + V = 2 - 2g$ de la forma:

$$g = 1 - C/2 + A/2 - V/2$$

Ahora calculamos g dibujando las caras y demás de nuestro sólido, contando C , A y V , y sustituyendo estos valores en la fórmula. Como la expresión es un invariante, no importa cómo troceemos el sólido, siempre obtendremos la misma respuesta. Pero nada de lo que hacemos depende de tener una definición de agujero. En su lugar, «número de agujeros» se convierte en una interpretación en términos intuitivos, derivada de observar ejemplos simples donde sentimos que sabemos lo que la frase significaría.

Puede parecer como un engaño, pero hace incursiones significativas en una cuestión fundamental en topología: ¿cuándo una forma puede deformarse de modo continuo para convertirse en otra? Esto es, por lo que a los topólogos respecta, ¿son las dos formas iguales o no? Si son iguales, sus invariantes deben también ser el mismo; por el contrario, si los invariantes son diferentes, también lo son las formas. (No obstante, a veces dos formas podrían tener el mismo invariante, pero ser diferentes, depende del invariante.) Como una esfera tiene característica de Euler 2, y un toro tiene característica de Euler 0, no hay modo de deformar una esfera de manera continua para convertirla en un toro. Esto puede parecer obvio por el agujero... pero hemos visto las aguas turbulentas a las cuales ese modo de pensar nos puede llevar. No tienes que interpretar la característica de Euler con el fin de usarla para distinguir formas, y esto es decisivo.

De modo menos obvio, la característica de Euler muestra que el misterio «agujero a través de un agujero en un agujero» (figura 24) es realmente como un toro de 3 agujeros disfrazado. La mayoría de la complejidad aparente no es producto de la topología intrínseca de una superficie, sino del modo en que hemos escogido incrustarla en el espacio.

El primer teorema realmente significativo en topología se originó a partir de la fórmula para la característica de Euler. Era una clasificación completa de superficies, formas curvas bidimensionales como la superficie de una esfera o un toro. También se impusieron un par de condiciones técnicas: la superficie no debería tener límites y debería ser de extensión finita (la jerga es «compacta»).

Para este fin una superficie se describe intrínsecamente, esto es, no se concibe como existente en algún espacio que la rodea. Un modo de hacer esto es ver la

superficie como un número de regiones poligonales (las cuales topológicamente son equivalentes a círculos) que están pegadas unas a otras a lo largo de sus aristas según unas reglas específicas, como las instrucciones de «pegar la lengüeta de A con la lengüeta de B» que tienes cuando montas un recortable de cartón. Una esfera, por ejemplo, puede describirse usando dos círculos, pegados el uno con el otro a lo largo de su borde. Un círculo se convierte en el hemisferio norte, el otro en el hemisferio sur. Un toro tiene una descripción especialmente elegante como un cuadrado con las aristas opuestas pegadas la una con la otra. Esta construcción se puede visualizar en un espacio circundante (figura 25), lo que explica por qué crea un toro, pero las matemáticas se pueden llevar a cabo usando tan solo el cuadrado y las reglas de pegado, y esto ofrece ventajas precisamente porque es intrínseco.

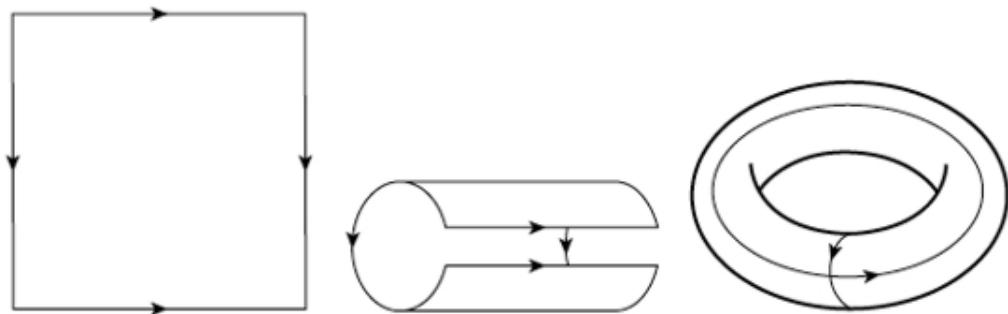


FIGURA 25. Pegando las aristas de un cuadrado hacemos un toro.

La posibilidad de pegar trozos de un borde nos lleva a un fenómeno bastante extraño: superficies con solo una cara. El ejemplo más famoso es la banda de Möbius, presentada por Möbius y Listing en 1858, la cual es una tira rectangular cuyos extremos están pegados con un giro de 180° (normalmente llamado medio giro, basado en la convención de que 360° constituye un giro completo). La banda de Möbius (véase la figura 26, izquierda), tiene una arista, que consiste en las aristas del rectángulo que no se han pegado a nada. Esta es la única arista, porque las dos aristas separadas del rectángulo están conectadas en una curva cerrada por el medio giro, lo que las une de punta a punta.

Es posible hacer un modelo de la banda de Möbius a partir de un papel, porque se incrusta de manera natural en un espacio tridimensional. La banda tiene solo una cara, en el sentido de que si empiezas a pintar una de sus superficies, y sigues por

ella, finalmente cubrirás la superficie entera, por delante y por detrás. Esto sucede porque el medio giro conecta la parte de delante con la de detrás. Esto no es una descripción intrínseca, porque se apoya en la incrustación de la banda en el espacio, pero hay una equivalente, una propiedad más técnica conocida como orientabilidad, que es intrínseca.

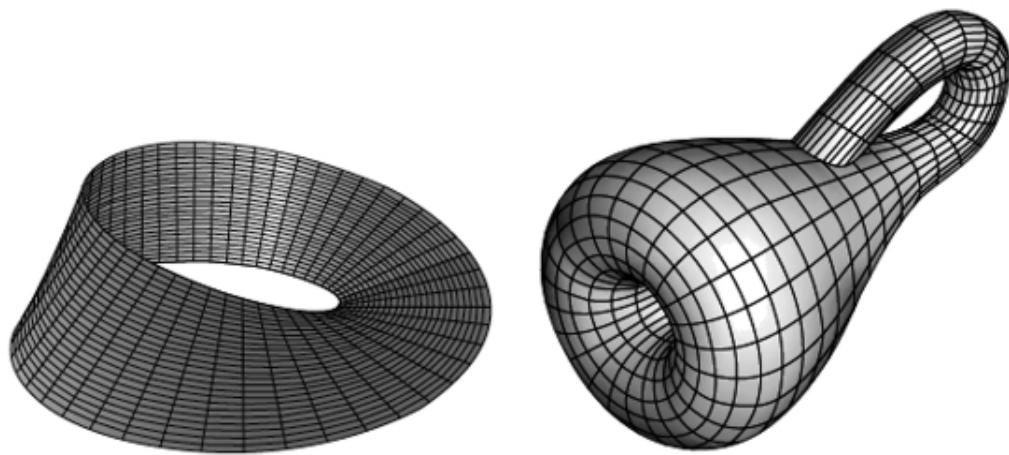


FIGURA 26. A la izquierda: banda de Möbius. A la derecha: botella de Klein. La aparente intersección consigo misma ocurre porque el dibujo se incrusta en un espacio tridimensional.

Existe una superficie con una única cara relacionada, que no tiene ninguna arista (figura 26, derecha). Surge si pegamos dos caras de un rectángulo juntas como una banda de Möbius y pegamos los otros dos lados juntos sin ningún giro. Cualquier modelo en un espacio tridimensional tiene que pasar a través de sí mismo, incluso aunque desde un punto de vista intrínseco las reglas de pegado no introducen ninguna autointersección. Si esta superficie se dibuja con dicho cruce, parece una botella cuyo cuello ha sido metido a través de la pared lateral y unido al fondo. Fue inventada por Felix Klein, y es conocida como la botella de Klein, casi con seguridad una broma basada en un juego de palabras alemán, cambiando *Kleinsche Fläche* (superficie de Klein) por *Kleinsche Flasche* (botella de Klein).

La botella de Klein no tiene bordes y es compacta, de modo que cualquier clasificación de superficies debe incluirla. Es la más conocida de una familia entera de superficies de una cara y sorprendentemente no es la más simple. El honor le

corresponde al plano proyectivo, que surge si pegas los pares de los lados opuestos de un cuadrado uno con otro, con un medio giro cada uno. (Esto es difícil de hacer con papel porque el papel es demasiado rígido; como la botella de Klein, se necesita que la superficie se interseque consigo misma. Se hace mejor «conceptualmente», esto es, dibujando imágenes en el cuadrado y recordando las reglas de pegado cuando las líneas se salen de las aristas y «se envuelven alrededor».) El teorema de clasificación de superficies, probado por Johann Listing alrededor de 1860, nos lleva a dos familias de superficies. Las que tienen dos caras son la esfera, toro, toro de 2 agujeros, toro de 3 agujeros, etcétera. Las que tiene solo una cara forman una familia infinita similar, empezando con el plano proyectivo y la botella de Klein. Se pueden obtener cortando un pequeño círculo de la superficie de dos caras correspondiente y pegando en su lugar una banda de Möbius.

Las superficies aparecen de manera natural en muchas áreas de las matemáticas. Son importantes en el análisis complejo, donde las superficies están asociadas con singularidades, puntos en los cuales las funciones se comportan de un modo extraño, por ejemplo, la derivada no existe. Las singularidades son la clave de muchos problemas en el análisis complejo, en cierto sentido capturan la esencia de la función. Como las singularidades están asociadas con superficies, la topología de las superficies proporciona una técnica importante para el análisis complejo. Históricamente, esto motivó la clasificación.

La mayoría de la topología moderna es sumamente abstracta, y mucho sucede en cuatro o más dimensiones. Podemos hacernos una idea de la materia en un entorno más familiar: los nudos. En el mundo real, un nudo es una maraña atada en un trozo de cuerda. Los topólogos necesitan un modo de evitar que el nudo se escape por los extremos una vez ha sido atado, de modo que unen los extremos de la cuerda formando una curva cerrada. Ahora un nudo es tan solo un círculo incrustado en el espacio. Intrínsecamente, un nudo es topológicamente idéntico a un círculo, pero en esta ocasión lo que cuenta es cómo el círculo se coloca dentro de su espacio circundante. Esto podría parecer contrario al espíritu de la topología, pero la esencia de un nudo recae en la relación entre la lazada de la cuerda y el espacio que la rodea. Considerando no solo la lazada, sino cómo se relaciona con el espacio, la topología puede abordar cuestiones importantes sobre los nudos. Entre

ellas están:

- ¿Cómo sabemos que un nudo está realmente anudado?
- ¿Cómo podemos distinguir topológicamente nudos diferentes?
- ¿Podemos clasificar todos los nudos posibles?

La experiencia nos dice que hay muchos tipos de nudos diferentes. La figura 27 muestra unos pocos de ellos: el nudo simple o de trébol, el nudo de rizo, el nudo de abuelita, nudo con forma de 8, nudo Stevedore, etcétera. Está también el nudo trivial o no-nudo, una curva circular ordinaria; como el propio nombre refleja, esta curva no está anudada. Muchos tipos de nudos diferentes han sido usados por generaciones de marineros, montañeros, y *boy-scouts*. Cualquier teoría topológica debería, por supuesto, reflejar, esta riqueza de experiencia, pero todo tiene que probarse, rigurosamente, dentro del entorno formal de la topología, justo como Euclides tuvo que probar el teorema de Pitágoras en lugar de tan solo dibujar unos pocos triángulos y medirlos. Sorprendentemente, la primera prueba topológica de que los nudos existen, en el sentido de que hay una incrustación en el círculo que no puede deformarse y convertirse en el nudo trivial, apareció por primera vez en 1926 en el *Knoten und Gruppen* (Nudos y grupos) del matemático alemán Kurt Reidemeister. La palabra «grupo» es un término técnico en el álgebra abstracta, que rápidamente se convirtió en la fuente más efectiva de los invariantes topológicos. En 1927, Reidemeister, e independientemente el americano James Waddell Alexander, en colaboración con su estudiante G.B. Briggs, encontró una prueba más simple de la existencia de nudos usando el «diagrama de nudos». Esto es una caricatura del nudo, dibujado con pequeños cortes en la curva para mostrar cómo las hebras separadas se solapan, como en la figura 27. Los cortes no están presentes en el propio nudo, pero representan su estructura tridimensional en un diagrama bidimensional. Ahora podemos usar los cortes para dividir el diagrama de nudos en un número de piezas definidas, sus componentes, y luego podemos manipular el diagrama y ver qué ocurre a las componentes.

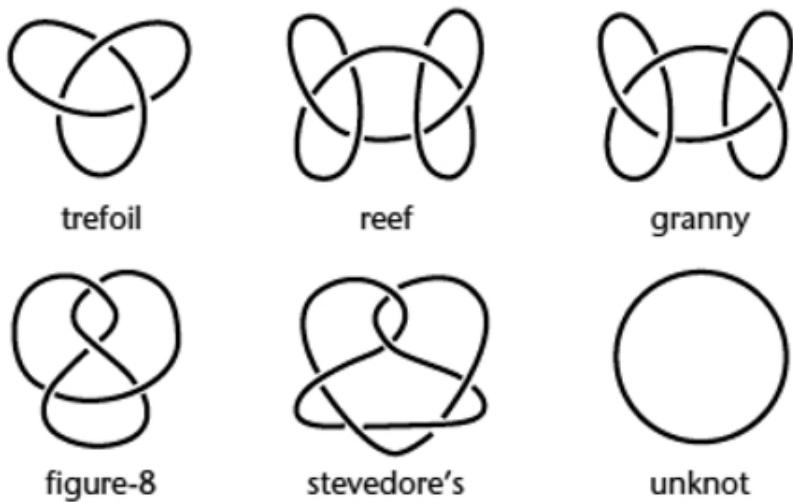


FIGURA 27. Cinco nudos y el no-nudo.

Si vuelves atrás, a cómo usaba la invariancia de la característica de Euler, verás que simplificaba el sólido usando una serie de movimientos especiales: unir dos caras eliminando una arista, unir dos aristas eliminando un punto. El mismo truco se aplica en los diagramas de nudos, pero ahora necesitas tres tipos de movimiento para simplificarlos, llamados movimientos de Reidemeister (figura 28). Cada movimiento puede llevarse a cabo en cualquier dirección: añadir o eliminar nudos, sobreponer dos hebras o separarlas, mover una hebra a través del lugar donde otras dos se cruzan.

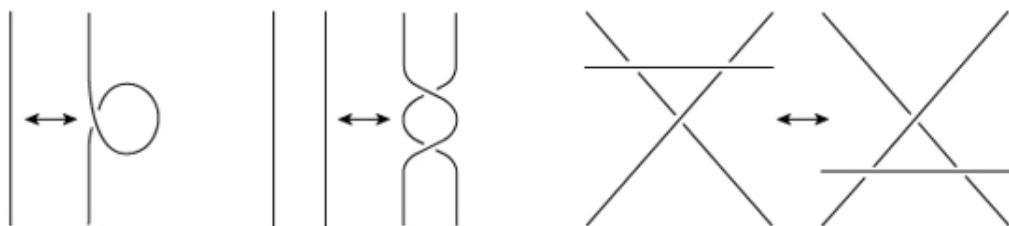


FIGURA 28. Movimientos de Reidemeister.

Con algunos arreglos preliminares para ordenar el diagrama de nudos, tales como modificar los lugares donde las curvas se solapan si eso sucede, se puede probar que cualquier deformación de un nudo se puede representar como una serie finita de movimientos de Reidemeister aplicados a su diagrama. Ahora podemos jugar el

juego de Euler, todo lo que tenemos que hacer es encontrar un invariante. Entre ellos está el grupo fundamental de un nudo, pero hay un invariante mucho más simple que prueba que el trébol realmente es un nudo. Puedo explicarlo en términos de colorear las componentes separadas en un diagrama de nudos. Empiezo con un diagrama ligeramente más complicado, con una curva extra, con el propósito de ilustrar algunas características de la idea (figura 29).

El giro extra crea cuatro componentes separadas. Supón que coloreo las componentes usando tres colores, por ejemplo, rojo, amarillo y azul (en la figura aparecen como negro, gris claro y gris oscuro). Entonces este coloreado obedece dos reglas simples:

- Al menos se usan dos colores distintos. (Realmente se usan tres, pero esta es información extra que no necesito.)
- En cada cruce, cualesquiera que sean las tres hebras cerca del cruce, todas tienen diferentes colores o todas son del mismo color. Cerca del cruce provocado por mi curva extra, las tres componentes son amarillas. Dos de estas componentes (en amarillo) se juntan en otro punto, pero cerca del cruce están separadas.

La observación maravillosa es que si un diagrama de nudos se puede colorear usando tres colores, obedeciendo estas dos reglas, entonces esto mismo es cierto después de cualquier movimiento de Reidemeister. Puedes probar esto muy fácilmente averiguando cómo los movimientos de Reidemeister afectan a los colores. Por ejemplo, si deshago mi curva extra en el dibujo entonces puedo dejar los colores sin cambios y se sigue cumpliendo todo. ¿Por qué esto es maravilloso? Porque prueba que el trébol realmente está anudado. Supón, en pro del argumento, que se puede desanudar; entonces alguna sucesión de movimientos de Reidemeister lo convierte en una curva sin nudos. Como el trébol se puede colorear obedeciendo a las dos reglas, lo mismo

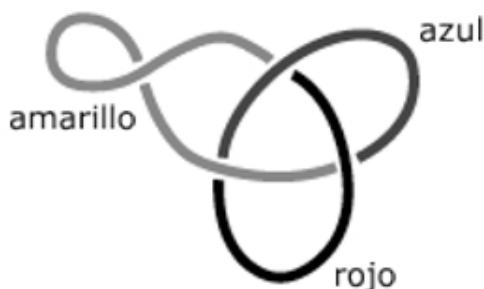


FIGURA 29. Coloreando un nudo de trébol con un giro extra.

debe aplicar a la curva sin nudos. Pero una curva sin nudos consiste en una sola hebra sin solapamientos, así que el único modo de colorearlo es usando el mismo color en todas partes. Pero esto viola la primera regla. Incurre en una contradicción, así que no puede existir dicha sucesión de movimientos de Reidemeister, es decir, el trébol no se puede desanudar.

Esto prueba que el trébol está anudado, pero no lo distingue de otros nudos como el nudo de rizo o el nudo de Stevedore. Uno de los primeros modos efectivos de hacer esto fue inventado por Alexander. Se deriva de métodos de Reidemeister de álgebra abstracta, pero nos lleva a un invariante que es algebraico en el sentido más común del álgebra escolar. Se llama el polinomio de Alexander, y asocia a cualquier nudo una fórmula formada a partir de potencias de una variable x . Estrictamente hablando, el término «polinomio» se aplica solo cuando las potencias son enteros positivos, pero aquí también permitimos potencias negativas. La tabla 2 ofrece unos cuantos de los polinomios de Alexander. Si dos nudos en la lista tienen diferentes polinomios de Alexander, y en este caso todos lo tienen excepto el de rizo y el de la abuelita, entonces los nudos deben ser topológicamente diferentes. Lo opuesto no es cierto: el de rizo y el de la abuelita tienen el mismo polinomio de Alexander, pero en 1952 Ralph Fox probó que eran topológicamente diferentes. La prueba requiere topología sorprendentemente complicada. Fue mucho más difícil de lo que nadie se esperaba.

knot	Alexander polynomial
Unknot	1
Trefoil	$x - 1 + x^{-1}$
Figure-8	$-x + 3 - x^{-1}$
Reef	$x^2 - 2x + 3 - 2x^{-1} + x^{-2}$
Granny	$x^2 - 2x + 3 - 2x^{-1} + x^{-2}$
Stevedore's knot	$-2x + 5 - 2x^{-1}$

TABLA 2. Polinomios de Alexander de nudos

Después de 1960, la teoría de nudos entró en el estancamiento topológico, detenida

en un vasto océano de cuestiones sin resolver, esperando un aliento de perspicacia creativa. Llegó en 1984, cuando el matemático neozelandés Vaughan Jones tuvo una idea tan simple que podría habersele ocurrido a cualquiera a partir de Reidemeister. Jones no era un teórico de nudos, ni siquiera era topólogo. Era un analista, trabajando sobre álgebra de operadores, un área con fuertes vínculos con la física matemática. No fue una sorpresa total que la idea se aplicase a nudos, porque los matemáticos y los físicos ya sabían de las conexiones interesantes entre álgebra de operadores y trenzas, que son un tipo especial de nudo con varias hebras. El nuevo invariante de nudos que inventó, llamado el polinomio de Jones, se define también usando el diagrama de nudos y tres tipos de movimiento. Sin embargo, los movimientos no conservan el tipo de nudo topológico, no conservan el nuevo «polinomio de Jones». Sin embargo, aunque parezca mentira, puede hacerse que la idea funcione, y el polinomio de Jones es un invariante de nudos.

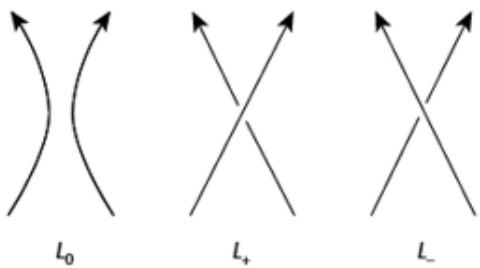


FIGURA 30. Movimientos de Jones.

Para este invariante, tenemos que escoger una dirección concreta a lo largo del nudo, que se muestra con una flecha. El polinomio de Jones $V(x)$ se define como uno para el nudo trivial. Dado cualquier nudo L_0 , acerca dos hebras separadas sin cambiar ningún cruce en su diagrama. Ten

cuidado de alinear las direcciones como se indica, esto es la razón de que la flecha sea necesaria, y el proceso no funcione sin ella. Reemplaza esa región de L_0 con dos hebras que se crucen en los dos modos posibles (figura 30). Sean los diagramas de nudos resultantes $L+$ y $L-$. Ahora definimos:

$$(x^{1/2} - x^{-1/2}) V(L_0) = x^{-1} V(L+) - x V(L-)$$

Empezando con el nudo trivial y aplicando dichos movimientos en el modo correcto, puedes averiguar el polinomio de Jones para cualquier nudo. Misteriosamente, resulta ser un invariante topológico. Y supera al tradicional polinomio de Alexander; por ejemplo, puede distinguir el nudo de rizo del de la abuelita, porque tienen diferentes polinomios de Jones.

El descubrimiento de Jones le hizo ganar la medalla Fields, el premio más prestigioso en matemáticas. También desencadenó el arranque de nuevos invariantes de nudos. En 1985, cuatro grupos de matemáticos diferentes, ocho personas en total, descubrieron simultáneamente la misma generalización del polinomio de Jones y presentaron sus artículos independientemente a la misma revista. Las cuatro pruebas eran diferentes, y el editor convenció a los ocho autores para unir fuerzas y publicar un artículo combinado. Su invariante es con frecuencia llamada polinomio HOMFLY, por sus iniciales. Pero incluso los polinomios de Jones y HOMFLY no respondieron completamente a los tres problemas de la teoría de nudos. No se sabe si un nudo con un polinomio de Jones 1 debe ser trivial, aunque muchos topólogos creen que esto es probablemente cierto. Existen nudos distintos topológicamente con el mismo polinomio de Jones; el ejemplo más simple conocido tiene diez cruces en su diagrama de nudos. Una clasificación sistemática de todos los posibles nudos sigue siendo una quimera matemática.

Es bonita, pero ¿es útil? La topología tiene muchos usos, pero normalmente son indirectos. Los principios topológicos proporcionan entendimiento sobre otras áreas más directamente aplicables. Por ejemplo, nuestra comprensión del caos se fundamenta en propiedades topológicas de sistemas dinámicos, tales como el extraño comportamiento del que Poincaré se dio cuenta cuando reescribió su memoria premiada (capítulo 4). La superautopista interplanetaria es una característica topológica de la dinámica del Sistema Solar.

Aplicaciones más esotéricas de la topología surgen en las fronteras de la física fundamental. Aquí el consumidor principal de la topología son los teóricos cuánticos de campos, porque la teoría de supercuerdas, la ansiada unificación de la mecánica cuántica y la relatividad, está basada en la topología. Aquí analogías del polinomio de Jones en teoría de nudos surgen en el contexto de los diagramas de Feynman, los cuales muestran cómo partículas cuánticas, como los electrones y fotones, se mueven a través del espacio-tiempo, colisionando, fusionándose y rompiéndose. Un diagrama de Feynman es un poco como un diagrama de nudos, y las ideas de Jones se pueden extender a este contexto.

Para mí, una de las aplicaciones más fascinantes de la topología es su creciente uso en biología, ayudándonos a entender el funcionamiento de la molécula de la vida, el

ADN. La topología se presenta porque el ADN es una doble hélice, como dos escaleras de caracol enroscándose la una con la otra. Las dos hebras están intrincadamente entrelazadas, y procesos biológicos importantes, en particular el modo en que una célula copia su ADN cuando se divide, tienen que tener en cuenta esta topología compleja. Cuando Francis Crick y James Watson publicaron su trabajo sobre la estructura molecular del ADN en 1953, acabaron con una breve alusión a un posible mecanismo de copiado, supuestamente involucrado en la división celular, en la cual las dos hebras se separan y cada una es usada como una plantilla para una nueva copia. Eran reacios a afirmar demasiado, porque eran conscientes de que había obstáculos topológicos para separar hebras entrelazadas. Si hubiesen sido demasiado específicos sobre su propuesta podrían haber enturbiado las aguas en una etapa temprana.

Según resultaron las cosas, Crick y Watson tenían razón. Los obstáculos topológicos eran reales, pero la evolución había proporcionado métodos para vencerlos, tales como enzimas especiales que cortan y pegan hebras de ADN. No es coincidencia que una de estas se llame topoisomerasa. En la década de los noventa del siglo pasado, los matemáticos y los biólogos moleculares usaron la topología para analizar los giros y vueltas del ADN, y para estudiar cómo funciona en la célula,

donde el método habitual de difracción de rayos X no puede usarse porque requiere que el ADN esté en forma cristalina.

Algunas enzimas, llamadas recombinasas, cortan las dos hebras de ADN y las vuelven a unir de un modo diferente. Para determinar cómo dicha enzima actúa cuando están en una célula, los biólogos aplican la enzima a un bucle cerrado de ADN. Luego observan la forma del bucle modificada usando un microscopio electrónico. Si la enzima une hebras distintas, la imagen es un nudo (figura 31).

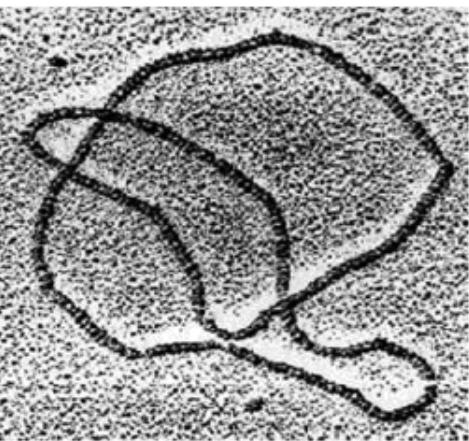


FIGURA 31. Bucle de ADN formando un nudo de trébol.

31). Si la enzima mantiene las hebras separadas, la imagen muestra dos bucles enlazados. Métodos procedentes de la teoría de nudos, como el polinomio de Jones

y otra teoría conocida como «de enredos», hacen posible averiguar qué nudos y lazos se dan y esto proporciona información detallada sobre qué hace la enzima. También hacen nuevas predicciones que han sido verificadas experimentalmente, proporcionando cierta confianza para creer que el mecanismo indicado por los cálculos topológicos es correcto.²⁰

Teniendo todo esto en cuenta, no te tropezarás con la topología en tu vida diaria, aparte del lavaplatos que mencioné al principio del capítulo. Pero entre bastidores, la topología informa a todas las corrientes principales de las matemáticas, posibilitando el desarrollo de otras técnicas con usos prácticos más obvios. Este es el motivo de que los matemáticos consideren que la topología tiene una gran importancia, mientras el resto del mundo difícilmente ha oído hablar de ella.

²⁰ Resumido en el capítulo 12 de *Las matemáticas de la vida* de Ian Stewart, Crítica, Barcelona 2011.

Capítulo 7

Patrones del azar

Distribución normal

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The diagram shows the formula for the cumulative distribution function of a normal distribution, $\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, with various mathematical terms labeled:

- probabilidad...** points to the first term $\Phi(x)$.
- ...de obtener este número** points to the second term x .
- igual** points to the third term μ .
- uno** points to the fourth term 1 .
- dividido por** points to the fifth term σ .
- menos** points to the sixth term $(x-\mu)$.
- media** points to the seventh term μ .
- cuadrado** points to the eighth term $(x-\mu)^2$.
- desviación estandar** points to the ninth term σ^2 .
- dos** points to the tenth term 2 .
- elevación a potencia** points to the eleventh term $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$.
- 2.71828** points to the twelfth term e .
- desviación estandar** points to the thirteenth term σ .
- raíz cuadrada** points to the fourteenth term $\sqrt{2\pi}$.
- dos** points to the fifteenth term $\sqrt{2}$.
- 3.14159** points to the sixteenth term π .

¿Qué dice?

La probabilidad de observar un valor concreto de un dato es mayor cerca del valor de la media y se desvanece rápidamente a medida que la diferencia con la media incrementa. Cómo de rápido se desvanece depende de una cantidad llamada desviación estándar.

¿Por qué es importante?

Define una familia especial de distribuciones de probabilidad con forma de campana, que son, con frecuencia, modelos buenos para observaciones comunes del mundo real.

¿Qué provocó?

El concepto de «hombre medio», testes de la importancia de los resultados experimentales, como pruebas médicas, y una tendencia desafortunada a tomar por defecto la campana de Gauss como si nada más existiese.

Las matemáticas tratan sobre patrones. El funcionamiento aleatorio del azar parece estar tan alejado de los patrones como te puedas imaginar. De hecho, una de las

definiciones actuales de «aleatorio» se reduce a «carencia de cualquier patrón apreciable». Los matemáticos han estado investigando patrones en geometría, álgebra y análisis durante siglos antes de darse cuenta de que incluso la aleatoriedad tiene sus propios patrones. Pero los patrones del azar en absoluto están en conflicto con la idea de que los sucesos aleatorios no tienen patrón, porque las regularidades de los sucesos aleatorios son estadísticas. Son características de toda una serie de sucesos, tales como el comportamiento medio a largo plazo de ensayos. No nos dicen nada sobre qué suceso ocurre en cada instante. Por ejemplo, si tiras un dado²¹ repetidamente, entonces alrededor de un sexto de las veces obtendrás 1, y lo mismo se cumple para 2, 3, 4, 5 y 6 —un patrón estadístico claro—. Pero esto no nos dice nada sobre qué número aparecerá en el próximo lanzamiento.

No fue hasta el siglo XIX cuando los matemáticos y científicos se dieron cuenta de la importancia de los patrones estadísticos en los sucesos del azar. Incluso las acciones humanas, como el suicidio o el divorcio, están sujetas a leyes cuantitativas, en promedio y a largo plazo. Llevó tiempo acostumbrarse a lo que parece en un principio contradecir el libre albedrío. Pero en la actualidad estas regularidades estadísticas conforman las bases de ensayos médicos, políticas sociales, primas de seguros, evaluación de riesgos y el deporte profesional.

Y los juegos de azar, que es donde todo empezó.

Todo fue iniciado, de manera apropiada, por el académico ludópata Girolamo Cardano. Al ser algo gandul, Cardano ganaba el dinero que necesitaba apostando en partidas de ajedrez y juegos de azar. Aplicaba su poderoso intelecto a ambos. El ajedrez no depende del azar, ganar depende de una buena memoria para posiciones estándar y movimientos, y un sexto sentido para flujo total del juego. En un juego de azar, sin embargo, el jugador está sujeto a los caprichos de la diosa Fortuna. Cardano se dio cuenta de que podía aplicar su talento matemático con buenos resultados incluso en esta relación tempestuosa. Podía mejorar su rendimiento en los juegos de azar adquiriendo una mejor comprensión de las probabilidades —las posibilidades de ganar o perder— de la que sus oponentes tenían. Escribió un libro

²¹ Sí, sé que esto es el plural de 'die', pero hoy en día todo el mundo lo usa para el singular, así, y he renunciado a la lucha contra esta tendencia. Podría ser peor: alguien me acaba de enviar un e-mail usando cuidadosamente 'dices' para el singular y 'die' para el plural. (para la versión en inglés)

sobre el tema, *Liber de Ludo Aleae* (Libro sobre los juegos de azar). No se publicó hasta 1633. Su contenido académico es el primer tratamiento sistemático de las matemáticas de la probabilidad. Su contenido menos honroso es un capítulo sobre cómo engañar y salir impune de ello.

Uno de los principios fundamentales de Cardano era que en una apuesta justa, las apuestas deberían ser proporcionales al número de modos en el cual cada jugador puede ganar. Por ejemplo, supón que los jugadores tiran un dado, y el primer jugador gana si sale un 6, mientras el segundo jugador gana si sale cualquier otro resultado. El juego sería sumamente injusto si cada uno apuesta la misma cantidad para jugar al juego, porque el primer jugador tiene solo un modo de ganar, mientras que el segundo tiene cinco. Sin embargo, si el primer jugador apuesta 1 € y el segundo apuesta 5 €, las probabilidades se hacen equitativas. Cardano era consciente de que este método de cálculo de probabilidades justas dependía de que los distintos modos de ganar fuesen igualmente posibles, y en juegos de dados, cartas o lanzamiento de monedas estaba claro cómo garantizar que se aplicaba esta condición. Lanzar una moneda tiene dos resultados, cara o cruz, y estas son igualmente posibles si la moneda es justa. Si la moneda tiende a sacar más caras que cruces, está claramente predisposta de modo no justo. De manera similar los seis resultados de un dado no trucado son igualmente posibles, como lo son los 48 resultados para extraer una carta de una baraja española.

La lógica tras el concepto de imparcialidad aquí es ligeramente circular, porque deducimos parcialidad a partir de un fracaso en la obtención de las condiciones numéricas obvias. Pero estas condiciones están apoyadas en más que un mero conteo. Están basadas en un sentimiento de simetría. Si la moneda es un círculo de metal plano, de densidad uniforme, entonces los dos resultados están relacionados por la simetría de la moneda (dale la vuelta). Para el dado, los seis resultados están relacionados por las simetrías del cubo. Y para las cartas, la simetría relevante es que ninguna carta difiere de manera significativa de otra, excepto por el valor escrito en su cara. Las frecuencias $1/2$, $1/6$ y $1/48$ para cualquier resultado dado dependen de estas simetrías básicas. Una moneda trucada o un dado trucado pueden crearse insertando pesos encubiertos, una carta trucada puede crearse usando marcas sutiles en el reverso que revelen su valor a aquellos que las

conocen.

Hay otros modos de engañar, que involucran juegos de manos, por ejemplo, introducir y sacar un dado trucado del juego antes de que nadie note que siempre da como resultado 6. Pero el modo más seguro de «engañar» —ganar usando subterfugios— es ser totalmente honesto, pero saber las probabilidades mejor que tu oponente. En cierto sentido, estás tomando la instancia moral suprema, pero puedes mejorar tus oportunidades encontrando un oponente lo suficientemente inocente y amañando, no las probabilidades, sino las expectativas de tu oponente sobre las probabilidades. Hay muchos ejemplos donde las probabilidades reales en el juego de azar son significativamente diferentes de las que mucha gente asumiría de manera natural.

Un ejemplo es el juego de la corona y el ancla, al que jugaban mucho los marinos británicos en el siglo XVIII. Usa tres dados, los cuales no tienen los números del 1 al 6, sino seis símbolos: una corona, un ancla, y los cuatro palos de la baraja inglesa: diamantes, picas, tréboles y corazones. Estos símbolos son también marcados en un tapete. Los jugadores apuestan colocando dinero en el tapete y lanzando los tres dados. Si cualquiera de los símbolos a los que han apostado aparece, la banca les paga su apuesta multiplicada por el número de dados en los que aparece el símbolo. Por ejemplo, si apuestan 1 € a la corona y salen dos coronas, entonces gana 2 € en suma a su apuesta. Todo suena muy razonable, pero la teoría de la probabilidad nos dice que a la larga un jugador puede esperar perder un 8 % de su apuesta.

La teoría de la probabilidad empezó a tener éxito cuando atrajo la atención de Blaise Pascal. Pascal era hijo de un recaudador de impuestos de Ruan y un niño prodigo. En 1646 se convirtió al jansenismo, una secta del catolicismo romano que el papa Inocencio X declaró herética en 1655. Un año antes, Pascal había experimentado lo que él llamaba su «segunda conversión», probablemente provocada por un accidente casi fatal cuando sus caballos cayeron por el borde del puente Neuilly y a su carroaje casi le pasa lo mismo. La mayoría de su producción a partir de entonces fue en filosofía religiosa. Pero justo antes del accidente, él y Fermat se estuvieron escribiendo para tratar un problema matemático que tenía que ver con el juego. El Caballero de Meré, un escritor francés que se llamaba a sí mismo caballero aunque no lo era, era un amigo de Pascal, y le preguntó cómo

deberían dividirse las apuestas en una serie de juegos de azar si el concurso tenía que abandonarse en mitad del juego. Esta pregunta no era nueva, se remonta a la Edad Media. Lo que fue nuevo fue la solución. En un intercambio de cartas, Pascal y Fermat encontraron la respuesta correcta. Y por el camino, crearon una nueva rama de las matemáticas: la teoría de la probabilidad.

Un concepto central en su solución era lo que ahora llamamos «esperanza». En un juego de azar, esto es beneficio medio de un jugador a la larga. Por ejemplo, sería 92 céntimos para la corona y el ancla con una apuesta de 1 €. Después de esta segunda conversión, Pascal dejó su pasado en el juego tras él, pero lo usó como ayuda en una famosa argumentación filosófica, la apuesta de Pascal.²² Pascal asumió, jugando a abogado del diablo, que alguien podría considerar la existencia de Dios como muy poco probable. En su *Pensées* (Pensamientos) de 1669, Pascal analiza las consecuencias desde el punto de vista de las probabilidades.

Consideremos el peso de ganar y perder apostando que Dios es (existe). Estimemos estas dos opciones. Si ganas, lo ganas todo, si pierdes, no pierdes nada. Apuesta, entonces, sin duda, a que Él es... Hay por ganar una infinidad de una vida infinitamente feliz, una oportunidad de ganar contra un número finito de oportunidades de perder y lo que apuestas es finito. Y así nuestra proposición es de fuerza infinita, cuando se apuesta algo finito en un juego donde hay riesgos iguales de ganar y perder, y el infinito por ganar.

La teoría de la probabilidad triunfó como un área de las matemáticas completamente desarrollada en 1713 cuando Jacob Bernoulli publicó su *Ars Conjectandi* (El arte de hacer conjeturas). Empezó con la definición de probabilidades de un suceso que funciona habitualmente: la proporción de ocasiones en las que sucederá, a la larga, casi siempre. Digo «definición que funciona» porque esta aproximación a las probabilidades da problemas si tratas de hacerla fundamental. Por ejemplo, supongamos que tengo una moneda no trucada y la lanzo una y otra vez. La mayoría de las veces obtengo una secuencia de aspecto aleatorio de caras y cruces, y si sigo lanzándola durante el tiempo suficiente obtendré cara aproximadamente la mitad de las veces. Sin embargo, rara vez

²² Hay muchas falacias en la argumentación de Pascal. La principal es que se aplicaría a cualquier ser hipotético sobrenatural.

obtengo caras exactamente la mitad de las veces: esto es imposible en un número de lanzamientos impares, por ejemplo. Si trato de modificar la definición tomando inspiración del cálculo, de modo que la probabilidad de obtener caras es el límite de la proporción de caras a medida que el número de lanzamientos tiende a infinito, tengo que probar que este límite existe. Pero solo existe a veces. Por ejemplo, supón que la secuencia de caras y cruces es la siguiente:

$$+ C C + + + C C C C C C + + + + + + + + + + \dots$$

Con una cruz, dos caras, tres cruces, seis caras, doce cruces, etcétera, el número se dobla en cada etapa después de tres cruces. Después de tres lanzamientos la proporción de caras es $2/3$, después de seis lanzamientos es $1/3$, después de doce lanzamientos vuelve a ser $2/3$, después de veinticuatro es $1/3$, ... de modo que la proporción oscila de un lado a otro, entre $2/3$ y $1/3$, y por lo tanto no tiene un límite bien definido. De acuerdo que dicha secuencia de lanzamientos es muy poco probable, pero para definir «poco probable», necesitamos primero definir probabilidades, que es lo que el límite se supone que tiene que lograr. Así que la lógica es circular. Además, incluso si el límite existe, quizás no sea el valor «correcto» de $1/2$. Un caso extremo ocurre cuando la moneda siempre cae con cara. Ahora el límite es 1 . De nuevo, esto es improbabílísimo, pero...

Bernoulli decidió aproximarse a todo el tema desde la dirección opuesta. Empezó simplemente definiendo la probabilidad de caras y cruces como algún número entre 0 y 1 . Digamos que la moneda es justa si $p = 1/2$, y está trucada en caso contrario. Ahora Bernoulli probó un teorema básico, la ley de los grandes números. Introduce una regla razonable para asignar probabilidades a una sucesión de sucesos repetidos. La ley de los grandes números afirma que a la larga, con la excepción de una fracción de ensayos que se hace arbitrariamente pequeña, la proporción de caras tiene límite y ese límite es p . Filosóficamente este teorema muestra que asignando probabilidades —esto es, números— de un modo natural, la interpretación «proporción de casos que se dan a la larga ignorando excepciones raras» es válida. De modo que Bernoulli consideró el punto de vista de que los números asignados como probabilidades proporcionan un modelo matemático

consistente del proceso de lanzar una moneda una y otra vez.

Esta prueba depende de un patrón numérico que era muy familiar a Pascal. Es normalmente llamado el triángulo de Pascal, incluso aunque él no fue la primera persona en fijarse en él. Los historiadores han rastreado su origen hasta el *Chandas Shastra*, un texto sánscrito atribuido a Pingala, escrito en algún momento entre el 500 a.C. y el 200 a.C. El original no ha sobrevivido, pero el trabajo es conocido a través de comentarios hindúes del siglo X. El triángulo de Pascal tiene este aspecto:

$$\begin{array}{c} 1 \\ 1 \ 1 \\ 1 \ 2 \ 1 \\ 1 \ 3 \ 3 \ 1 \\ 1 \ 4 \ 6 \ 4 \ 1 \end{array}$$

Donde todas las filas empiezan y acaban en 1 y cada número es la suma de los dos que están justo encima suyo. Ahora llamamos a estos números coeficientes binomiales, porque aparecen en el álgebra de la expresión binomial (de dos variables) $(p + q)^n$. Concretamente:

$$\begin{aligned} (p + q)^0 &= 1 \\ (p + q)^1 &= p + q \\ (p + q)^2 &= p^2 + 2pq + q^2 \\ (p + q)^3 &= p^3 + 3p^2q + 3pq^2 + q^3 \\ (p + q)^4 &= p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4 \end{aligned}$$

Y el triángulo de Pascal se forma con los coeficientes de términos separados.

La clave del entendimiento de Bernoulli es que si lanzamos una moneda n veces, con una probabilidad p de obtener caras, entonces la probabilidad de un número específico de lanzamientos obteniendo cara es el término correspondiente de $(p + q)^n$ donde $q = 1 - p$. Por ejemplo, supongamos que lanzo la moneda tres veces. Entonces los ocho posibles resultados son:

CCC
 CC+ C+C +CC
 C++ +C+ ++C
 +++

Donde he agrupado las secuencias según el número de caras. De modo que de las ocho secuencias posibles hay:

- 1 secuencia con 3 caras
- 3 secuencias con 2 caras
- 3 secuencias con 1 cara
- 1 secuencia con 0 caras

El vínculo con los coeficientes binomiales no es coincidencia. Si expandes la fórmula algebraica $(C + (+))^3$ pero no juntas los términos unos con otros, tienes

$$CCC + CC(+) + C(+C) + (+)CC + C(+)(+) + (+)C(+) + (+)(+)C + (+)(+)(+)$$

Agrupando los términos según el número de Cs, tenemos entonces:

$$C^3 + 3C^2(+) + 3C(+)^2 + (+)^3$$

Después de eso, se trata de remplazar cada C y (+) por su probabilidad, p o q , respectivamente.

Incluso en este caso, cada extremo CCC y +++ se da solo una vez en ocho pruebas, y números más equitativos se dan en los otros seis. Un cálculo más sofisticado usando propiedades estándar de los coeficientes binomiales prueba la ley de Bernoulli de los grandes números.

Los avances en las matemáticas con frecuencia son provocados por la ignorancia. Cuando los matemáticos no sabían cómo calcular algo importante, encontraban un modo de acercarse sigilosamente a ello indirectamente. En este caso, el problema es calcular estos coeficientes binomiales. Hay una fórmula explícita, pero si, por ejemplo, quieras saber la probabilidad de obtener exactamente 42 caras cuando

lanzas una moneda 100 veces, tienes que hacer 200 multiplicaciones y luego simplificar una fracción muy complicada. (Hay atajos, pero son también liosos.) Mi ordenador me dice en una fracción de segundo que la respuesta es:

$$28.258.808.871.162.574.166.368.460.400p^{42}q^{58}$$

Pero Bernoulli no tenía este lujo. Nadie lo tuvo hasta la década de los sesenta del siglo XX y los sistemas de álgebra computacional no estuvieron realmente disponibles de manera general hasta finales de la década de los ochenta de ese mismo siglo.

Como este tipo de cálculo directo no era viable, los sucesores inmediatos de Bernoulli trataron de encontrar buenas aproximaciones. Alrededor de 1730, Abraham De Moivre obtuvo una fórmula aproximada para las probabilidades involucradas en lanzamientos repetidos de una moneda trucada. Esto llevó a la función error o a la distribución normal, a la que con frecuencia se hace referencia como la «curva de campana» o «campana de Gauss» a causa de su forma. Lo que él probó fue esto. Define la *distribución normal* $\Phi(x)$ con media μ y varianza σ^2 con la fórmula:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Entonces para una n grande, la probabilidad de obtener m caras en n lanzamientos de una moneda trucada está muy cercana a $\Phi(x)$ cuando:

$$x = m/n - p$$

$$\mu = np$$

$$\sigma = npq$$

Aquí «media» se refiere al promedio, y «varianza» es una medida de cómo de dispersos están los datos, el ancho de la campana de Gauss. La raíz cuadrada de la varianza, σ sin más, se llama la desviación estándar. La figura 32 (izquierda)

muestra cómo el valor de $\Phi(x)$ depende de x . La curva se parece un poco a una campana, de ahí el nombre que recibe de manera informal. La campana de Gauss es un ejemplo de una distribución de probabilidad, lo que significa que la probabilidad de obtener datos entre dos valores dados es igual al área bajo la curva y entre las líneas verticales que se corresponden con esos valores. El área total bajo la curva es 1, gracias a ese factor inesperado $\sqrt{2\pi}$.

La idea se entiende mucho más fácilmente usando un ejemplo. La figura 32 (derecha) muestra un gráfico de probabilidades de obtener varios números de caras cuando se lanza una moneda no trucada 15 veces seguidas (barras rectangulares) junto con la curva de campana aproximada.

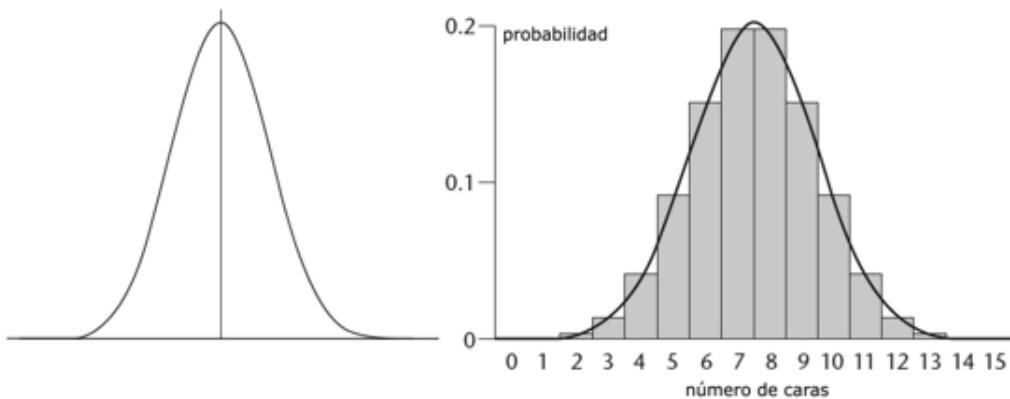


FIGURA 32. A la izquierda: campana de Gauss. A la derecha: cómo aproximar el número de caras en 15 lanzamientos de una moneda no trucada.

La campana de Gauss empezó a adquirir un estatus icónico cuando empezó a aparecer en datos empíricos en las ciencias sociales, no tan solo en las matemáticas teóricas. En 1835 Adolphe Quetelet, un belga quien entre otras cosas fue pionero en métodos cuantitativos en sociología, recogió y analizó grandes cantidades de datos de crímenes, la proporción de divorcios, suicidios, nacimientos, muertes, altura de los humanos, peso, etcétera. Variables que nadie esperaba que se ajustasen a una ley matemática, porque sus causas subyacentes eran demasiado complejas e implicaban elecciones humanas. Considera, por ejemplo, el tormento emocional que lleva a alguien a cometer suicidio. Parece ridículo pensar que esto podría reducirse a una simple fórmula.

Estas objeciones tienen mucho sentido si quieres predecir exactamente quién se matará a sí mismo y en qué momento. Pero cuando Quetelet centró la atención en cuestiones estadísticas, tales como la proporción de suicidios en varios grupos de gente, varias localizaciones y diferentes años, empezó a ver patrones. Esto resultó controvertido: si predices que habrá seis suicidios en París el próximo año, ¿cómo puede tener sentido cuando cada persona involucrada actúa según su propia voluntad? Podrían todos cambiar sus pensamientos. Pero la población formada por aquellos que se matarán no está especificada de antemano, aparece como una consecuencia de las elecciones hechas no solo por aquellos que cometen suicidio, sino por aquellos que piensan sobre ello y no lo hacen.



FIGURA 33. El gráfico de Quetelet de cuánta gente (eje vertical) tiene una altura dada (eje horizontal).

El ejercicio de libre voluntad de la gente en el contexto de muchas otras cosas, las cuales influyen en que decidan libremente; aquí las limitaciones incluyen problemas financieros, problemas de relación, estado mental, formación religiosa... En cualquier caso, la campana de Gauss no hace predicciones exactas, solo expone qué cifra es más probable. Quizá ocurran cinco o siete suicidios, dejando espacio de sobra para que cualquiera ejerza su libre voluntad y cambie de opinión.

Los datos finalmente triunfan: por la razón que sea, la gente en masa se comporta más predeciblemente que los individuos. Quizá el ejemplo más simple sea la altura.

Cuando Quetelet determinó las proporciones de gente con una altura dada, obtuvo una bella campana de Gauss (figura 33). Obtuvo la misma forma de curva para muchas otras variables sociales.

Quetelet estaba tan impresionado con sus resultados que escribió el libro *Sur l'homme et le développement de ses facultés* (Sobre el hombre y el desarrollo de las facultades humanas), publicado en 1835. En él, introduce la noción del «hombre medio», un individuo ficticio que estaba en todos los aspectos en la media. Hace tiempo que se percibió que esto no funcionaba del todo; el «hombre» medio, esto es, una persona, de modo que el cálculo incluye hombres y mujeres, tiene (ligeramente menos que) un pecho, un testículo, 2,3 hijos, etcétera. No obstante, Quetelet vio su hombre medio como el objetivo de la justicia social, no solo una ficción matemática llamativa. No es tan absurdo como suena. Por ejemplo, si la riqueza humana se reparte por igual a todos, entonces todo el mundo tendrá la riqueza media. No es un objetivo práctico, a menos que ocurran cambios sociales enormes, pero alguien con fuertes visiones igualitarias podría defenderlo como un objetivo deseable.

La campana de Gauss rápidamente pasó a ser un ícono en teoría de la probabilidad, especialmente su rama aplicada, la estadística. Había dos razones principales: la campana de Gauss era relativamente simple de calcular, y había una razón teórica para que se diese en la práctica. Una de las principales fuentes para este modo de pensamiento era la astronomía del siglo XVIII. Los datos que se observaban estaban sujetos a errores, causados por ligeras variaciones en aparatos, errores humanos, o simplemente el movimiento del aire de ese momento en la atmósfera. Los astrónomos de la época querían observar los planetas, cometas y asteroides, y calcular sus órbitas, y esto requería encontrar la órbita que encajase mejor con los datos. Cómo encajaba no sería nunca perfecto.

Primero apareció la solución práctica a este problema. Se reducía a lo siguiente: dibuja una línea recta a través de los datos y escoge esta línea de manera que el error total sea lo más pequeño posible. Los errores aquí tienen que considerarse positivos y el modo más fácil de lograr esto mientras mantenemos el álgebra agradable es elevarlos al cuadrado. Así el error total es la suma de los cuadrados de las desviaciones de las observaciones a partir de la línea recta trazada, y la línea

deseada minimiza este error. En 1805 el matemático francés Adrien-Marie Legendre descubrió una fórmula simple para esta línea, haciendo fácil su cálculo. El resultado es el llamado método de los mínimos cuadrados. La figura 34 ilustra el método con datos artificiales relacionados con el estrés (medidos con un cuestionario) y la presión sanguínea. La línea en la imagen, calculada usando la fórmula de Legendre, es la que se ajusta mejor a los datos según la medida del error cuadrático. En diez años, el método de los mínimos cuadrados era estándar entre los astrónomos en Francia, Prusia e Italia. Pasados otros veinte años era estándar en Inglaterra.

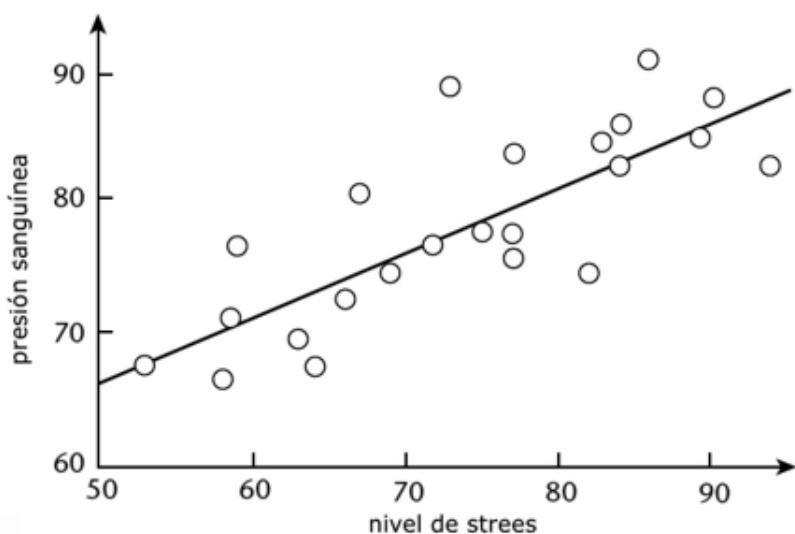


FIGURA 34. Utilización del método de los mínimos cuadrados para relacionar la presión sanguínea y el estrés. Los puntos: los datos. La línea: la línea recta que se ajusta mejor.

Gauss hizo del método de los mínimos cuadrados una piedra angular de su trabajo en mecánica celeste. Llegó al área en 1801, mediante una predicción con éxito de la vuelta del asteroide Ceres después de que se escondiese tras el resplandor del Sol, cuando la mayoría de los astrónomos pensaban que los datos disponibles eran demasiado limitados. Este triunfo selló su reputación matemática entre el público y lo instaló de por vida como profesor de astronomía en la Universidad de Gotinga. Gauss no usó los mínimos cuadrados para esta predicción en particular, sus cálculos se reducen a resolver ecuaciones algebraicas de grado ocho, las cuales obtuvo por

un método numérico inventado expresamente. Pero en su trabajo posterior, culminando en su *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum* (Teoría del movimiento de cuerpos celestes moviéndose en secciones cónicas alrededor del Sol) de 1809, hacía gran énfasis en el método de los mínimos cuadrados. También afirmó que había desarrollado, y usado, la idea diez años antes de Legendre, lo cual causó un poco de revuelo. Sin embargo, era muy probable que fuese cierto y la justificación de Gauss del método era bastante diferente. Legendre lo había visto como un ejercicio en el ajuste de curvas, mientras que Gauss lo vio como un modo de ajustar una distribución de probabilidad. Su justificación de la fórmula asumía que los datos subyacentes, para los cuales se ajustaba la línea recta, seguían una campana de Gauss.

Quedaba justificar la justificación. ¿Por qué deberían estar los errores de observación distribuidos normalmente? En 1810, Laplace aportó una respuesta asombrosa, también motivada por la astronomía. En muchas ramas de la ciencia es normal hacer la misma observación varias veces independientemente y luego tomar la media. De manera que es natural hacer un modelo matemático de este procedimiento. Laplace usó la transformada de Fourier (véase el capítulo 9), para probar que el promedio de muchas observaciones se describe con una campana de Gauss, incluso si las observaciones individuales no lo hacen. Su resultado, el teorema central del límite, fue un punto de inflexión muy importante en probabilidad y estadística, porque proporcionó una justificación teórica para usar la distribución favorita de los matemáticos, la campana de Gauss, en el análisis de los errores experimentales.²³

El teorema central del límite distingue la campana de Gauss como la única distribución de probabilidad apropiada para la media de muchas observaciones repetidas. De ahí que adquiriese el nombre de «distribución normal», y se vio como

²³ El teorema afirma que bajo ciertas (bastante comunes) condiciones, la suma de un número grande de variables aleatorias tendrá una distribución aproximadamente normal. Más precisamente, si (x_1, \dots, x_n) es una secuencia de variables aleatorias independientes distribuidas de manera idéntica, cada una teniendo media μ y varianza σ^2 , entonces el teorema central del límite afirma que

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n x_i - \mu \right)$$

Converge a la distribución normal con media 0 y desviación estándar σ a medida que n se hace arbitrariamente grande.

la elección por defecto para una distribución de probabilidad. No solo la distribución normal tiene unas propiedades matemáticas gratas, sino que hay también razones sólidas para asumirlas como modelo para datos reales. Esta combinación de atributos resultó ser muy atractiva para los científicos que deseaban comprender mejor los fenómenos sociales que habían interesado a Quetelet, ya que ofrecía un modo de analizar los datos a partir de registros oficiales. En 1865, Francis Galton estudió cómo la altura de un niño se relaciona con la altura de sus padres. Esto era parte de un objetivo más amplio: comprender la herencia, cómo las características humanas pasan de padres a hijos. Irónicamente, al principio el teorema central del límite de Laplace llevó a Galton a dudar de la existencia de este tipo de herencia. Y, aunque existiese, probarla sería difícil, porque el teorema central del límite era una espada de doble filo. Quetelet había encontrado una bella campana de Gauss para las alturas, pero parecía decir muy poco sobre los diferentes factores que afectaban a la altura, porque el teorema central del límite predecía una distribución normal en cualquier caso, para cualquier distribución posible de estos factores. Incluso si las características de los padres estaban entre estos factores, podrían ser aplastadas por las otras, tales como la nutrición, salud, estatus social, etcétera.

En 1889, sin embargo, Galton había encontrado una respuesta a este dilema. La prueba del maravilloso teorema de Laplace se apoyaba en calcular el promedio de los efectos de muchos factores distintos, pero estos tienen que satisfacer algunas condiciones rigurosas. En 1875, Galton describió estas condiciones como «sumamente artificiales» y señaló que las influencias al ser una media, deben ser

- (1) *todas independientes en sus efectos,*
- (2) *todas iguales (teniendo la misma distribución de probabilidad),*
- (3) *todas admiten ser tratadas como alternativas simples «sobre el promedio» o «bajo el promedio», y*
- (4) ... *calculadas sobre la suposición de que las influencias de la variable son infinitamente numerosas.*

Ninguna de estas condiciones se aplica a la herencia humana. La condición (4) corresponde a la suposición de Laplace de que el número de factores que se añaden tiende a infinito, de modo que «infinitamente numerosas» es un poco exagerado; no

obstante, lo que establecieron los matemáticos era que para obtener una buena aproximación a la distribución normal, tienes que combinar un número de factores grande. Cada uno de ellos contribuye en una pequeña cantidad al promedio; con, por ejemplo, una centena de factores, cada uno contribuye una centésima de su valor. Galton se refiere a dichos factores como «insignificantes». Cada uno por sí mismo no tiene un efecto significativo.

Había una salida potencial, y Galton la aprovechó. El teorema central del límite proporciona una condición suficiente para que una distribución sea normal, no una necesaria. Aunque estas suposiciones no se cumplan, la distribución que nos ocupa podría todavía ser normal por otras razones. La tarea de Galton era averiguar cuáles podrían ser estas razones. Para tener alguna esperanza de vincularlo con la herencia, tenían que aplicarse a la combinación de unas pocas influencias grandes y dispares, no a un número enorme de influencias insignificantes. Lentamente buscó a tientas su camino hacia una solución y lo encontró a través de dos experimentos, ambos datan de 1877. Una fue un artilugio, la máquina de Galton, en el cual unas bolas caen por una pendiente, rebotando contra un grupo de clavos con las mismas posibilidades de ir a la izquierda y a la derecha. En teoría las bolas deberían apilarse en la parte baja según una distribución binomial, una aproximación discreta a la distribución normal, así que debería, y lo hacen, formar, aproximadamente, un montón con forma de campana, como en la figura 32 (derecha). La clave para comprenderlo fue imaginar que las bolas se detienen temporalmente cuando están bajando. Todavía formarán una campana de Gauss, pero sería más estrecha que la final. Imagina liberar tan solo un compartimento de bolas. Caerían al fondo, distribuyéndose en una campana de Gauss minúscula. Lo mismo ocurre para cualquier otro compartimento. Lo que significa que, al final, la campana de Gauss grande podría verse como una suma de muchas pequeñas. La campana de Gauss se reproduce a sí misma cuando varios factores, cada uno siguiendo su propia campana de Gauss por separado, se combinan.

El factor decisivo llegó cuando Galton crió guisantes. En 1875, distribuyó semillas entre siete amigos. Cada uno recibió 70 semillas, pero uno recibió semillas muy ligeras, otro unas ligeramente más pesadas, etcétera. En 1877, midió los pesos de las semillas de la progenie resultante. Cada grupo está normalmente distribuido,

pero el peso medio difería en cada caso, siendo comparable al peso de cada semilla en el grupo original. Cuando combinó los datos para todos los grupos, los resultados de nuevo estaban normalmente distribuidos, pero la varianza era mayor, la campana de Gauss era más ancha. De nuevo, esto sugería que combinando varias curvas de campana se llegaba a otra campana de Gauss. Galton buscó el origen de la razón matemática para esto. Supón que dos variables aleatorias están normalmente distribuidas, no necesariamente con la misma media o la misma varianza. Entonces su suma está también normalmente distribuida; esto quiere decir que es la suma de las dos medias y su varianza es la suma de dos varianzas. Obviamente lo mismo aplica para la suma de tres, cuatro o más variables aleatorias normalmente distribuidas.

Este teorema funciona cuando un número pequeño de factores se combinan y cada factor puede multiplicarse por una constante, así que realmente funciona para cualquier combinación lineal. La distribución normal es válida incluso cuando el efecto de cada factor es grande. Ahora Galton podía ver cómo este resultado se aplicaba a la herencia. Supongamos que la variable aleatoria dada para la altura de un niño es alguna combinación de las variables aleatorias correspondientes para las alturas de sus padres, y estas siguen una distribución normal. Asumiendo que los factores hereditarios funcionan para la suma, la altura del niño seguirá también una distribución normal.

Galton escribió sus ideas en 1889 bajo el título de *Natural Inheritance* (Herencia natural). En particular, discutió una idea que llamó regresión. Cuando un progenitor alto y uno bajo tienen un niño, la altura media del niño debería ser intermedia, de hecho, debería ser la media de la altura de los padres. Asimismo la varianza debería ser el promedio de las varianzas, pero las varianzas para los padres parecían ser aproximadamente iguales, así que la varianza no cambiaba mucho. A medida que pasaban generaciones sucesivas, la altura media debería «regresar» a un valor fijo a mitad de camino, mientras que la varianza debería permanecer sin demasiados cambios. De modo que la nítida campana de Gauss de Quetelet podía sobrevivir de una generación a otra. Su pico rápidamente se asentaría en un valor fijo, la media total, mientras que su ancho sería igual. Por tanto cada generación debería tener la misma diversidad de alturas, a pesar de la regresión a la media. La diversidad se

mantendría gracias a individuos raros cuya regresión fracasase y era autosuficiente en una población suficientemente grande

Con el papel central de la campana de Gauss firmemente fundamentado en lo que, con el tiempo, se consideraron cimientos sólidos, los estadísticos podían trabajar sobre la percepción de Galton y los trabajadores en otros campos podían aplicar los resultados. Las ciencias sociales fueron uno de los primeros beneficiarios, pero la biología pronto le siguió y las ciencias físicas ya estaban adelantadas en este juego gracias a Legendre, Laplace y Gauss. Pronto una caja de herramientas estadísticas completa estuvo disponible para cualquiera que quisiera extraer patrones a partir de datos. Me centraré tan solo en una técnica, porque se usa de manera rutinaria para determinar la eficacia de medicamentos y procedimientos médicos, además de tener muchas otras aplicaciones. Se llama contraste de hipótesis y su objetivo es evaluar la importancia de patrones aparentes en los datos. Fue descubierta por cuatro personas: los ingleses Ronald Aylmer Fisher, Karl Pearson, su hijo Egon, y el polaco nacido en Rusia y que pasó la mayoría de su vida en América, Jerzy Neyman. Me centraré en Fisher, quien desarrolló las ideas básicas cuando estaba trabajando como estadístico agrícola en la Estación Experimental de Rothamstead, analizando nuevas variedades de plantas.

Supongamos que estás cultivando una variedad nueva de patata. Tus datos sugieren que esta variedad es más resistente a algunas plagas. Pero dichos datos están sujetos a muchas fuentes de error, de modo que no puedes estar completamente seguro de que los números apoyen esa conclusión, ciertamente no tan seguro como un físico que puede hacer medidas muy precisas para eliminar la mayoría de los errores. Fisher se dio cuenta de que el asunto clave era distinguir una diferencia genuina de una que surgiese puramente por casualidad, y que el modo de hacer esto es preguntar cuán probable sería esa diferencia si solo una casualidad estuviese involucrada.

Asume, por ejemplo, que la variedad nueva de patata parece conferir el doble de resistencia, en el sentido de que la proporción de la nueva variedad que sobrevive a las plagas es el doble de la proporción para la variedad antigua. Es concebible que este efecto sea debido al azar y puedas calcular su probabilidad. De hecho, lo que calculas es la probabilidad de un resultado al menos tan extremo como el observado

en los datos. ¿Cuál es la probabilidad de que la proporción de la nueva variedad que sobrevive a la plaga sea al menos dos veces la de la variedad antigua? Incluso se permiten proporciones mayores porque la probabilidad de obtener exactamente dos veces la proporción seguro que es muy pequeña. Cuanto más amplio sea el rango de resultados que incluyas, se hacen más probables los efectos del azar, así que puedes confiar más en tu conclusión si tus cálculos sugieren que no es resultado del azar. Si esta probabilidad obtenida por estos cálculos es baja, digamos 0,05, entonces el resultado es poco probable que sea fruto del azar, se dice que tiene un nivel de significación del 95 %. Si la probabilidad es más baja, por ejemplo 0,01, entonces el resultado es extremadamente poco probable que sea por azar y se dice que su nivel de significación es del 99 %. Los porcentajes indican que si solo interviniese el azar, el resultado no sería tan extremo como el observado en el 95 % de las pruebas, o en el 99 % de ellas.

Fisher describió su método como una comparación entre dos hipótesis distintas: la hipótesis de que los datos son significativos en un nivel establecido, y la llamada hipótesis nula, en la que los resultados se deben al azar. Insistió en que su método no debe ser interpretado como confirmación de la hipótesis de que los datos son significativos, debe ser interpretado como un rechazo de la hipótesis nula. Lo que quiere decir que proporciona evidencias contra los datos que no son significativos. Esto podría parecer una distinción muy fina, ya que la evidencia contra los datos que no son significativos seguramente cuenta como evidencia a favor de que sean significativos. Sin embargo, no es completamente cierto, y la razón es que la hipótesis nula tiene una suposición intrínseca extra. Para calcular la probabilidad de que un resultado tan extremo sea debido al azar, necesitas un modelo teórico. El modo más simple de obtener uno es asumir una distribución de probabilidad específica. Esta suposición se aplica solo en conexión con la hipótesis nula, porque eso es lo que usas para hacer las cuentas. No asumes que los datos están distribuidos normalmente. Pero la distribución por defecto para la hipótesis nula es normal: la campana de Gauss.

Este modelo inherente tiene una consecuencia importante, que «el rechazo a la hipótesis nula» tiende a disimular. La hipótesis nula es «los datos son causa del azar». De modo que es demasiado fácil leer esa afirmación como «rechazo de que

los datos son debidos al azar», lo cual implica que aceptas que no se deben al azar. Aunque, realmente, la hipótesis nula es «los datos son debidos al azar y los efectos del azar están distribuidos normalmente», así que podría haber dos razones para rechazar la hipótesis nula: los datos no se deben al azar, o no siguen una distribución normal. La primera apoya lo significativo que son los datos, pero la segunda no. Dice que puede que estés usando el modelo estadístico equivocado. El trabajo agrícola de Fisher, estaba generalmente lleno de evidencias para distribuciones normales de los datos. De modo que la distinción que estoy haciendo realmente no importa. Aunque en otras aplicaciones del contraste de hipótesis podría importar. Decir que los cálculos rechazan la hipótesis nula sí es cierto, pero debido a que la suposición de una distribución normal no está explícitamente mencionada, es bastante fácil olvidar que necesitas comprobar la normalidad de la distribución de los datos antes de concluir que tus resultados son estadísticamente significativos. A medida que el método es usado por más y más gente entrenada en cómo hacer los cálculos pero no en las suposiciones que hay tras él, existe un peligro creciente de asumir erróneamente que las pruebas muestran que tus datos son significativos. Especialmente cuando la distribución normal se ha convertido en la suposición automática por defecto.

En la conciencia pública, el término «campana de Gauss» está indeleblemente asociado con el polémico libro de 1994 *The bell curve* (La campana de Gauss) escrito por dos norteamericanos, el psicólogo Richard J. Herrnstein y el científico político Charles Murray. El principal tema del libro es un reivindicado vínculo entre la inteligencia, medida por el coeficiente intelectual (CI), y variables sociales como los ingresos, el empleo, los índices de embarazo y el crimen. Los autores argumentan que niveles de CI son mejores prediciendo dichas variables que el estatus social y económico de los padres o su nivel de educación. Las razones para la controversia y los argumentos involucrados son complejos. Un rápido esbozo no puede realmente hacer justicia al debate, pero los temas van directos de vuelta a Quetelet y merecen su mención.

La polémica era inevitable, no importa cuáles podrían haber sido los méritos o deméritos académicos del libro, porque pone el dedo en la llaga: la relación entre raza e inteligencia. Los artículos en los medios tienden a insistir en la propuesta de

que las diferencias en el CI tienen un origen genético predominante, pero el libro era más cuidadoso sobre este vínculo, dejando la interacción entre genes, el entorno y la inteligencia abiertos. Otro tema polémico era un análisis sugiriendo que la estratificación social en los Estados Unidos (y en realidad en cualquier lugar) se incrementó significativamente a lo largo del siglo XX, y que la principal causa fue las diferencias en la inteligencia. Otro más era una serie de recomendaciones políticas para tratar este presunto problema. Una era reducir la inmigración, la cual el libro reivindicaba que estaba bajando el CI medio. Quizá la más polémica era la sugerencia de que las políticas de bienestar social que supuestamente animaban a mujeres pobres a tener hijos deberían detenerse.

Irónicamente, la idea se remonta al propio Galton. Su libro *Hereditary Genius* (Genio hereditario) de 1869 construido sobre escritos anteriores para desarrollar la idea de que «las habilidades naturales de un hombre son derivadas de la herencia, bajo exactamente las mismas limitaciones que la forma y las características físicas de todo el mundo orgánico. Consecuentemente ... sería bastante factible producir una raza altamente dotada de hombre por matrimonios juiciosos durante varias generaciones consecutivas». Afirmaba que la fertilidad era mayor entre los menos inteligentes, pero evitaba cualquier sugerencia de selección deliberada en favor de la inteligencia. En su lugar, expresaba la esperanza de que la sociedad podría cambiar de modo que la gente más inteligente comprendiese la necesidad de tener un montón de niños.

Para muchos, la propuesta de Herrnstein y Murray para manipular el sistema de bienestar estaba incómodamente cerca del movimiento de eugenesia de principios del siglo XX, por el cual 60.000 norteamericanos fueron esterilizados, supuestamente por una enfermedad mental. La eugenesia pasó a estar ampliamente desacreditada cuando se empezó a asociar con la Alemania nazi y el holocausto, y muchas de sus prácticas son ahora consideradas violaciones de la legislación de los derechos humanos, en algunos casos ascendiendo a crímenes contra la humanidad. Las propuestas de engendrar humanos de manera selectiva son generalmente vistas como racismo intrínsecamente. Varios científicos sociales refrendaron las conclusiones científicas del libro pero cuestionaron la carga de racismo; algunos de ellos estaban menos seguros sobre las propuestas políticas.

The Bell Curve inició un debate prolongado sobre los métodos usados para recoger datos, los métodos matemáticos usados para analizarlos, la interpretación de los resultados y las sugerencias políticas basadas en estas interpretaciones. Un grupo de trabajo seleccionado por la American Psychological Association concluyó que algunos resultados del libro eran válidos: las puntuaciones del CI son buenas para predecir logros académicos, esto está correlacionado con el estatus laboral y no hay diferencias significativas en los resultados de hombres y mujeres. Por otro lado, el informe del grupo de trabajo reafirmó que tanto genes como entorno influyen en la puntuación del CI y no encontró evidencias significativas de que las diferencias raciales en las puntuaciones del CI estén genéticamente determinadas.

Otros críticos han argumentado que hay errores en la metodología científica, tales como ignorar datos que no convenían, y que el estudio y algunas respuestas podrían de algún modo haber sido motivados políticamente. Por ejemplo, es cierto que la estratificación social se ha incrementado dramáticamente en Estados Unidos, pero podría argumentarse que la causa principal es la negativa de los ricos a pagar impuestos, más que las diferencias en la inteligencia. También parece que hay inconsistencia entre el presunto problema y la solución propuesta. Si la pobreza hace que la gente tenga más niños y crees que eso es una cosa mala, ¿a santo de qué querría hacerlos todavía más pobres?

Una parte importante del fondo, con frecuencia ignorado, es la definición del CI. Más que ser algo directamente medible, como la altura o el peso, el CI es deducido estadísticamente a partir de test. Los sujetos se exponen a las preguntas y sus puntuaciones son analizadas usando un descendiente del método de los mínimos cuadrados llamado análisis de la varianza. Como el método de los mínimos cuadrados, esta técnica asume que los datos se distribuyen según la distribución normal, y busca aislar aquellos factores que determinan la mayor cantidad de variabilidad en los datos y son por tanto los más importantes para modelar los datos. En 1904, el psicólogo Charles Spearman aplicó esta técnica a varios testes de inteligencia diferentes. Observó que las puntuaciones que los sujetos obtenían en test diferentes estaban altamente correlacionadas, es decir, si alguien lo hacía bien en uno de los test, tendía a hacerlo bien en todos. Intuitivamente, parecían estar midiendo la misma cosa. El análisis de Spearman mostró que un único factor común

—una variable matemática, a la cual llamó g , que significaba «inteligencia general»— explicaba casi todo sobre la correlación. El CI es una versión estandarizada de la g de Spearman.

Una cuestión clave es si g es una cantidad real o una ficción matemática. La respuesta es complicada a causa de los métodos usados para escoger las pruebas para el CI. Estas asumen que la distribución de inteligencia «correcta» en la población es la normal (la campana de Gauss epónima), y calibra los test manipulando las puntuaciones matemáticamente para estandarizar la media y la desviación estándar. Un peligro potencial aquí es que obtienes lo que esperas porque sigues los pasos para filtrar cualquier cosa que lo contradijera. Stephen Jay Gould hizo una crítica extensiva de dichos peligros en 1981 en *The Mismeasure of Man* (La falsa medida del hombre), señalando entre otras cosas que puntuaciones sin filtrar en test del CI con frecuencia no siguen una distribución normal para nada. La principal razón para pensar que g representa una característica genuina de la inteligencia humana es que es el único factor: matemáticamente define una única dimensión. Si muchos test diferentes parecen todos estar midiendo la misma cosa, es tentador concluir que la cosa que nos concierne debe ser real. Si no lo es, ¿por qué todos los resultados serían tan similares? Parte de la respuesta podría ser que los resultados de los test de CI se reducen a una puntuación numérica única. Esto comprime un conjunto de preguntas multidimensional y actitudes potenciales en una respuesta unidimensional. Además, los test han sido seleccionados de modo que la puntuación esté correlacionada fuertemente con la visión de respuestas inteligentes de quien lo diseña, si no, nadie consideraría usarlo.

Por analogía, imagina recoger datos de varios aspectos diferentes del «tamaño» en el reino animal. Uno podría medir la masa, otro la altura, otro la longitud, ancho, diámetro de la pata trasera izquierda, tamaño de los dientes, etcétera. Cada una de dichas medidas sería un único número. En general estarían íntimamente correlacionados: animales altos tienden a pesar más, a tener dientes mayores, patas más gruesas... Si pasas los datos a través de un análisis de la varianza, encontrarías muy probablemente que una única combinación de estos datos explica la vasta mayoría de la variabilidad, justo como la g de Spearman lo hace para diferentes medidas de cosas aunque estén relacionadas con la inteligencia.

¿Implicaría necesariamente esto que todas estas características de los animales tiene la misma causa subyacente? ¿Que una cosa controla todas? ¿Quizá, posiblemente, un nivel de la hormona del crecimiento? Pero probablemente no. La riqueza de la forma animal no se condensa cómodamente en un único número. Muchas otras características no se correlacionan con el tamaño en absoluto: la habilidad para volar, tener líneas o puntos, comer carne o vegetación. La combinación de medidas especial y única que cuenta para la mayoría de la variabilidad podría ser una consecuencia matemática de los métodos usados para encontrarla, especialmente si esas variables fueron escogidas, como ocurre aquí, por tener mucho en común para empezar.

Volviendo a Spearman, vemos que su muy pregonada g podría ser unidimensional porque los test de CI son unidimensionales. El CI es un método estadístico, conveniente matemáticamente, para cuantificar tipos específicos de habilidades para resolver problemas, pero no necesariamente se corresponde con un atributo real del cerebro humano, y no necesariamente representa lo que sea que queremos decir con «inteligencia».

Centrándonos en un único tema, el CI, y usándolo para establecer políticas, *The Bell Curve* ignora el contexto más amplio. Incluso si fuese sensible a manipular genéticamente la población de una nación, ¿por qué restringir este proceso a los pobres? Incluso si de promedio los pobres tiene un CI más bajo que los ricos, un niño pobre brillante superaría a uno rico tonto algún día, a pesar de las obvias ventajas sociales y educacionales de las que los hijos de los ricos disfrutan. ¿Por qué recurrir a los cortes en bienestar cuando podrías dirigirte más exactamente hacia lo que reivindicas que es el problema real: la inteligencia en sí misma? ¿Por qué no mejorar la educación? De hecho, ¿por qué dirigir tus políticas hacia un incremento de la inteligencia? Hay muchos otros rasgos humanos deseables. ¿Por qué no reducir la credulidad, la agresividad o la avaricia?

Es un error pensar en un modelo matemático como si fuera la realidad. En las ciencias físicas, donde los modelos con frecuencia se ajustan a la realidad muy bien, esto podría ser un modo conveniente de pensar porque causa poco daño. Pero en las ciencias sociales, los modelos con frecuencia son poco mejores que caricaturas. La elección del título para *The Bell Curve* alude a esta tendencia a refundir el

modelo con la realidad. La idea de que el CI es algún tipo de medida precisa de la habilidad humana, simplemente porque tiene un pedigrí matemático, comete el mismo error. No es sensato basar políticas sociales radicales y muy polémicas en modelos matemáticos erróneos y simplistas. El tema central real sobre *The Bell Curve*, uno que trata extensamente pero sin darse cuenta, es que habilidad, inteligencia y sabiduría no son lo mismo.

La teoría de la probabilidad se usa de manera generalizada en ensayos médicos de medicamentos y tratamientos nuevos para probar la significación estadística de los datos. Las pruebas están, con frecuencia, pero no siempre, basadas en la suposición de que la distribución subyacente es normal. Un ejemplo típico es la detección de conglomerados de cáncer. Un conglomerado, para algunas enfermedades, es un grupo en el que la enfermedad se da con más frecuencia de lo esperado en el total de la población. El conglomerado puede ser geográfico, o puede referirse más metafóricamente a gente con un estilo de vida particular o un período de tiempo específico. Por ejemplo, luchadores profesionales retirados o niños nacidos entre 1960 y 1970.

Conglomerados aparentes podrían ser debidos totalmente al azar. Los números aleatorios están raras veces distribuidos en un modo aproximadamente uniforme, en vez de eso, con frecuencia se agrupan unos con otros. En simulaciones aleatorias de la Lotería Nacional de Reino Unido, donde seis números entre el 1 y el 49 se extraen aleatoriamente, más de la mitad parecen mostrar algún tipo de patrón regular como ser dos números consecutivos o tres números separados por la misma cantidad, por ejemplo, 5, 9, 13. Contrario a la intuición común, lo aleatorio se agrupa. Cuando se encuentra un conglomerado claro, las autoridades médicas tratan de evaluar si se debe al azar o si podría haber alguna posible conexión causal. Hace tiempo, la mayoría de los hijos de pilotos de combate israelíes eran niños. Sería fácil pensar en posibles explicaciones —los pilotos son muy viriles y hombres viriles engendran más chicos (por cierto, no es verdad), los pilotos están expuestos a más radiación de la normal, experimentan fuerzas G mayores—, pero este fenómeno es efímero, igual que un conglomerado aleatorio. En datos posteriores desapareció. En cualquier población de gente, siempre es probable que haya más niños de un sexo que de otro, exactamente la misma cantidad es muy

improbable. Para evaluar el significado del conglomerado, se debe seguir observando y ver si persiste.

No obstante, este aplazamiento no puede continuarse indefinidamente, especialmente si el conglomerado tiene que ver con enfermedades serias. El sida fue primero detectado como un conglomerado de casos de neumonía en hombres homosexuales de Norteamérica en la década de los ochenta del siglo XX, por ejemplo. Las fibras de amianto como una causa de una forma de cáncer de pulmón, el mesotelioma, apareció primero como un conglomerado entre antiguos trabajadores de amianto. De manera que los métodos estadísticos se usan para evaluar cuán probables serían dichos conglomerados si surgiesen por razones aleatorias. Los métodos de Fisher de contraste de hipótesis, y métodos relacionados, se usan ampliamente con ese propósito.

La teoría de la probabilidad es también fundamental para nuestra comprensión del riesgo. Esta palabra tiene un significado técnico concreto. Se refiere al potencial para que alguna acción nos lleve a un resultado no deseado. Por ejemplo, volar en un avión podría llevar a estar involucrado en un accidente, fumar cigarrillos podría llevar al cáncer de pulmón, construir una central nuclear podría llevar a liberar radiación en un accidente o ataque terrorista, construir un dique para una central hidroeléctrica podría causar muertes si el dique se derrumba. «Acción» aquí puede referirse a no hacer nada: no vacunar a un niño podría llevar a que muera de una enfermedad, por ejemplo. En este caso hay también un riesgo asociado con vacunar al niño, como puede ser una reacción alérgica. En el conjunto de toda la población este riesgo es pequeño, pero para grupos específicos puede ser mayor.

Se emplean muchos conceptos diferentes de riesgo en contextos diferentes. La definición matemática habitual es que el riesgo asociado con alguna acción, o ausencia de ella, es la probabilidad de un resultado adverso, multiplicado por la pérdida en la que se incurría. Según esta definición una entre diez probabilidades de matar a diez personas tiene el mismo nivel de riesgo que la probabilidad de una entre un millón de matar a un millón de personas. La definición matemática es racional en el sentido de que hay un fundamento específico tras ella, pero eso no significa que sea necesariamente sensata. Ya hemos visto que la «probabilidad» se refiere a largo plazo, pero para sucesos raros el largo plazo es en realidad muy

largo. Los humanos, y sus sociedades, pueden adaptarse a pequeños números de muertes repetidos, pero un país que de repente pierde un millón de personas de una vez podría estar en problemas serios, porque todos los servicios públicos y la industria estarían bajo una severa presión. Sería de poco consuelo decir que en los próximos 10 millones de años, las muertes totales en los dos casos serían comparables. De modo que se están desarrollando métodos nuevos para cuantificar riesgos en dichos casos.

Los métodos estadísticos, derivados de cuestiones sobre el juego, tienen una variedad enorme de usos. Proporcionan herramientas para el análisis social, médico y de datos científicos. Como todas las herramientas, lo que sucede depende de cómo se usen. Cualquiera que utilice métodos estadísticos necesita ser consciente de las suposiciones que hay tras estos métodos, y sus implicaciones.

Introducir números ciegamente en un ordenador y tomar los resultados como palabra de Dios, sin comprender las limitaciones de los métodos que se usan, es una receta para el desastre. El uso legítimo de la estadística, sin embargo, ha mejorado nuestro mundo de manera irreconocible. Y todo empezó con la campana de Gauss.

Capítulo 8

Buenas vibraciones

Ecuación de onda

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

desplazamiento

The diagram shows the wave equation $\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$ inside a grey rectangular box. Four lines point from labels to specific parts of the equation: one from 'derivada parcial de segundo orden con respecto al tiempo' to the left term $\frac{\partial^2 u}{\partial t^2}$; one from 'velocidad cuadrado' to the right term c^2 ; one from 'derivada parcial de segundo orden con respecto al espacio' to the right denominator $\frac{\partial^2 u}{\partial x^2}$; and one from 'desplazamiento' to the variable u .

¿Qué dice?

La **aceleración** de un pequeño segmento de la cuerda de un violín es proporcional al desplazamiento medio de los segmentos vecinos.

¿Por qué es importante?

Predice que la cuerda se moverá en ondas, y se generaliza de manera natural a otros sistemas físicos en los cuales aparecen ondas.

¿Qué provocó?

Grandes avances en nuestra comprensión de las ondas de agua, sonido, luz, vibraciones elásticas... Los sismólogos usan versiones modificadas de ella para deducir la estructura del interior de la Tierra a partir de cómo vibra. Compañías petrolíferas usan métodos similares para encontrar petróleo. En el capítulo 11 veremos cómo predijo la existencia de ondas electromagnéticas, que llevaron a la radio, la televisión, el radar y las comunicaciones modernas.

Vivimos en un mundo de ondas. Nuestras orejas detectan ondas de compresión en el aire, llamamos a esto «oído». Nuestros ojos detectan ondas de radiación electromagnética, llamamos a esto «vista». Cuando un terremoto azota un pueblo o una ciudad, la destrucción la causan ondas en el cuerpo sólido de la Tierra. Cuando un barco se balancea en el océano, está reaccionando a las ondas en el agua. Los surfistas usan las ondas del mar como diversión; la radio, la televisión y gran parte

de las redes de teléfonos móviles usan las ondas de la radiación electromagnética, similares a las que vemos, pero de longitudes de onda diferentes. Los microondas... bueno, el nombre lo dice todo, ¿no?

Con tantos ejemplos prácticos de ondas afectando a nuestra vida diaria, incluso desde hace siglos, los matemáticos que decidieron poner en práctica el descubrimiento épico de Newton de que la naturaleza tiene leyes difícilmente podrían evitar empezar a pensar en ondas. Aunque lo que les hizo empezar vino del arte, concretamente de la música. ¿Cómo la cuerda de un violín crea un sonido? ¿Qué lo provoca?

Había una razón para empezar con los violines, el tipo de razón que atrae a los matemáticos, aunque no a los gobiernos u hombres de negocios que dudan en invertir en matemáticas y esperan una retribución rápida. La cuerda de un violín puede ser modelada razonablemente como una línea infinitamente fina, y puede asumirse que su movimiento, el cual es claramente la causa del sonido que el instrumento hace, tiene lugar en un plano. Esto hace el problema «de dimensión baja», lo que quiere decir que hay posibilidades de resolverlo. Una vez has comprendido este ejemplo sencillo de ondas, es muy probable que la comprensión pueda transferirse, con frecuencia en pequeñas etapas, a ejemplos de ondas más realistas y más prácticos.

La alternativa, invertir precipitadamente en problemas sumamente complejos, puede parecer atractiva a políticos y capitanes de la industria, pero normalmente acaba estancándose en complejidades. A las matemáticas, la simplicidad les da alas y, si es necesario, los matemáticos la crearán artificialmente para proporcionar una ruta de entrada a problemas más complejos. Con desprecio se refieren a tales modelos como «juegos», pero estos son juguetes con un propósito serio. Los modelos de juguete de ondas nos llevaron al mundo actual de la electrónica y comunicaciones globales a gran velocidad, aviones de pasajeros de fuselaje ancho y satélites artificiales, la radio, la televisión, sistemas de aviso de tsunamis... pero nunca habríamos logrado ninguna de estas cosas si no fuese porque unos pocos matemáticos empezaron resolviendo cómo funciona un violín, usando un modelo que no era realista, ni siquiera para un violín.

Los pitagóricos creían que el mundo se basaba en números, con ello quieren decir

números naturales o las proporciones entre números naturales. Algunas de sus creencias tendían hacia lo místico, confiriendo a números específicos atributos humanos: 2 era el hombre, 3 la mujer, 5 simbolizaba el matrimonio, etcétera. El número 10 era muy importante para los pitagóricos porque era $1 + 2 + 3 + 4$ y creían que había cuatro elementos: tierra, aire, fuego y agua. Este tipo de especulación choca con la mentalidad moderna y resulta ligeramente disparatado, (bueno, al menos, con mi mentalidad), pero era razonable en una época en que los humanos estaban tan solo empezando a investigar el mundo que les rodeaba, buscando patrones cruciales. Fue necesario algo de tiempo para averiguar qué patrones eran importantes y cuáles eran escoria.

Uno de los grandes triunfos de la visión del mundo pitagórico vino de la música. Circulan varias historias; según una de ellas, Pitágoras estaba pasando por una herrería y se dio cuenta de que los martillos de diferentes tamaños hacían sonidos de tonos diferentes, y que los martillos relacionados por números sencillos —uno era el doble en tamaño que otro, por ejemplo— hacían sonidos que estaban en armonía. Por muy bonita que sea la historia, cualquiera que realmente lo intente con martillos reales descubrirá que los trabajos que se llevan a cabo en una herrería no son especialmente musicales, y los martillos tienen una forma bastante complicada para vibrar en armonía. Pero hay una pizca de verdad; en el conjunto, objetos pequeños emiten sonidos de tonos más altos que los grandes.

Las historias tienen una base más fuerte cuando se refieren a una serie de experimentos que los pitagóricos realizaron usando una cuerda estirada, un instrumento musical rudimentario conocido como monocordio. Tenemos conocimiento de estos experimentos porque Ptolomeo los documentó en su *Harmónicos* alrededor del año 150 d.C. Moviendo un soporte por varias posiciones a lo largo de la cuerda, los pitagóricos descubrieron que cuando dos cuerdas con la misma tensión tienen longitudes en una razón simple, como 2:1 o 3:2, producen notas armoniosas inusuales. Proporciones más complejas eran discordantes y desagradables al oído. Científicos posteriores llevaron estas ideas más lejos, probablemente un poco demasiado lejos; lo que nos parece agradable depende de la física del oído, que es más complicada que la de una sola cuerda, y también tiene una dimensión cultural porque los oídos de niños que están creciendo se entrena-

ser expuestos a los sonidos que son comunes en su sociedad. Predigo que los niños de hoy en día serán inusualmente sensibles a las diferencias en los tonos de llamada de los teléfonos móviles. Sin embargo, hay una historia científica sólida tras estas complejidades, y mucho de ello confirma y explica los descubrimientos tempranos de los pitagóricos con su instrumento experimental monocorde.

Los músicos describen pares de notas en términos de intervalo entre ellas, una medida de cuántos pasos las separan en la escala musical. El intervalo más fundamental es la octava, ocho teclas blancas en un piano. Las notas separadas una octava suenan muy similar, excepto que una nota es más alta que otra, y son extremadamente armoniosas. De hecho, tanto es así, que las armonías basadas en la octava pueden parecer un poco sosas. En un violín, el modo de tocar la nota una octava mayor en una cuerda suelta es presionar la mitad de esa cuerda contra el diapasón. Una cuerda la mitad de larga toca una nota una octava más alta. De modo que la octava está asociada con una razón numérica sencilla, 2:1.

Otros intervalos armoniosos están también asociados con proporciones numéricas simples. Las más importantes para la música occidental son la cuarta, una razón de 4:3, y la quinta, una razón de 3:2. Los nombres tienen sentido si consideras una escala musical de todas las notas C D E F G A B C.* Con C como base, la nota correspondiente a la cuarta es F, la quinta es G y la octava C. Si numeramos las notas consecutivamente con la base como 1, estos son respectivamente la 4^a, 5^a y 8^a notas a lo largo de la escala. La geometría es especialmente clara en un instrumento como una guitarra, que tiene segmentos de metal, «trastes», insertados en las posiciones relevantes. El traste para la cuarta está en un cuarto de la longitud de la cuerda, que para una quinta está a un tercio de la longitud y la octava está en la mitad. Puedes comprobar esto con un metro.

Estas proporciones ofrecen una base teórica para la escala musical y nos llevan a la(s) escala(s) más usadas en la actualidad en la mayoría de la música occidental. La historia es compleja, así que daré una versión simplificada. Por comodidad más adelante, a partir de ahora, reescribiré una razón 3:2 como una fracción 3/2. Empieza en una nota base y asciende en quintos, para obtener cuerdas de longitudes:

$$1 \frac{3}{2} \left(\frac{3}{2}\right)^2 \left(\frac{3}{2}\right)^3 \left(\frac{3}{2}\right)^4 \left(\frac{3}{2}\right)^5$$

Operando, estas fracciones son:

$$1 \frac{3}{2} \frac{9}{4} \frac{27}{8} \frac{81}{16} \frac{243}{32}$$

Todas estas notas, excepto las dos primeras, son demasiado agudas para permanecer en una octava, pero podemos bajarlas en una o más octavas, dividiendo repetidamente las fracciones por 2 hasta que el resultado quede entre 1 y 2. Esto nos lleva a las fracciones:

$$1 \frac{3}{2} \frac{9}{8} \frac{27}{16} \frac{81}{64} \frac{243}{128}$$

Finalmente, ordenándolas de manera ascendente, tenemos:

$$1 \frac{9}{8} \frac{81}{64} \frac{3}{2} \frac{27}{16} \frac{243}{128}$$

Esto se corresponde de manera bastante próxima a las notas C D E G A B en un piano. Observa que F no está. De hecho, al oírlo, el hueco entre 81/64 y 3/2 suena más amplio que los otros. Para llenar ese hueco, insertamos 4/3, la razón para la cuarta, que es muy próxima a F en el piano. También es útil completar la escala con una segunda C, una octava por encima, en razón de 2. Ahora obtenemos una escala musical basada totalmente en cuartos, quintos y octavos, con todos en las proporciones.

$$\begin{array}{ccccccccc}
 & 9 & 81 & 4 & 3 & 27 & 243 & \\
 1 & \frac{9}{8} & \frac{81}{64} & \frac{4}{3} & \frac{3}{2} & \frac{27}{16} & \frac{243}{128} & 2 \\
 & C & D & E & F & G & A & B & C
 \end{array}$$

La longitud es inversamente proporcional al tono, así tendríamos que invertir las fracciones para obtener las longitudes correspondientes.

Hasta ahora hemos explicado todas las teclas blancas del piano, pero también hay teclas negras. Estas aparecen porque los números sucesivos en la escala tienen dos proporciones diferentes: $9/8$ (llamado tono) y $256/243$ (semitono). Por ejemplo, teniendo en cuenta la proporcionalidad inversa, $9/8$ de $81/64$ son $9/8$, pero $81/64$ de $4/3$ son $256/243$. Los nombres «tono» y «semitono» indican una comparación de intervalos aproximada. Numéricamente son 1,125 y 1,05. El primero es más grande, de modo que un tono se corresponde con un cambio mayor en la altura que un semitono. Dos semitonos dan una razón de 1,05², que es casi 1,11, no lejos de 1,25. Así que dos semitonos están cerca de un tono. No muy cerca, lo admito.

Continuando con esta pauta, podemos dividir cada tono en dos intervalos, cada uno próximo a un semitono, para obtener una escala de 12 notas. Esto puede hacerse de varios modos, obteniendo resultados ligeramente diferentes. Comoquiera que se haga, puede haber problemas sutiles pero audibles cuando se cambia la clave de una pieza de música; los intervalos cambian ligeramente si, por ejemplo, subimos todas las notas un semitono. Este efecto podría haberse evitado si hubiésemos escogido una razón específica para un semitono y lo arreglásemos para que su duodécima potencia fuese 2. Entonces dos semitonos harían un tono exacto, 12 semitonos harían una octava y podrías cambiar la escala subiendo o bajando todas las notas una cantidad fija.

Existe dicho número, concretamente la raíz doce de 2, que es alrededor de 1,059, y nos lleva a la denominada «escala temperada». Es un convenio, por ejemplo, en la escala temperada la razón $4/3$ para un cuarto es $1,0595 = 1,335$ en lugar de $4/3 = 1,333$. Un músico muy entrenado puede detectar la diferencia, pero es fácil hacerse a ella y la mayoría de nosotros nunca nos daríamos cuenta.

La teoría pitagórica de la armonía en la naturaleza, entonces, está realmente incluida en las bases de la música occidental. Para explicar por qué proporcionales

simples van mano a mano con la armonía musical tenemos que echar un vistazo a la física de una cuerda vibrando. La psicología de la percepción humana también entra en juego, pero no todavía.

La clave es la segunda ley del movimiento de Newton, que relaciona aceleración con fuerza. También necesitas saber cómo la fuerza ejercida por una cuerda bajo tensión cambia a medida que la cuerda se mueve, estirándose o contrayéndose ligeramente. Para esto, usamos algo que el reticente contrincante de Newton, Hooke, descubrió en 1660, llamado la ley de Hooke: el cambio en la longitud de un muelle es proporcional a la fuerza ejercida en él (una cuerda de violín es de hecho un tipo de muelle, de modo que aplica la misma ley). Todavía queda un obstáculo. Podemos aplicar la ley de Newton a un sistema compuesto de un número finito de masas; obtenemos una ecuación por masa, y entonces lo hacemos lo mejor que podemos para resolver el sistema resultante. Pero una cuerda de violín es un continuo, una línea compuesta de infinidad de puntos. De modo que los matemáticos de la época pensaron en la cuerda como un gran número de masas de puntos muy juntas, ligadas unas a otras por los muelles de la ley de Hooke. Escribieron las ecuaciones ligeramente simplificadas para que se pudieran resolver y las resolvieron; finalmente permitieron que el número de masas se hiciese arbitrariamente grande, y calcularon qué pasaba con la solución.

Johann Bernoulli llevó a cabo estos pasos en 1727, y el resultado fue extraordinariamente bonito, considerando qué dificultades se estaban escondiendo. Para evitar confusión en la descripción que sigue, imagina que el violín está apoyado sobre su parte trasera con la cuerda horizontal. Si tiras de la cuerda, vibra arriba y abajo en ángulos rectos con el violín. Esta es la imagen a tener en mente. El uso del arco provoca que la cuerda vibre hacia los lados, y la presencia del arco es confusa. En el modelo matemático, todo lo que tenemos es una cuerda, fija en los extremos, y ningún violín; la cuerda vibra arriba y abajo en el plano. En este sistema, Bernoulli encontró que la forma de la cuerda vibrante, en cualquier instante de tiempo, era una curva sinusoidal. La amplitud de la vibración —la altura máxima de la curva— también seguía una curva sinusoidal en el tiempo en vez del espacio. Usando la simbología, su solución era como $\sin ct \sin x$, donde c es una constante (figura 35). La parte espacial, $\sin x$, nos dice la forma, pero está

multiplicada por un factor, $\sin ct$, en el tiempo t . La fórmula dice que la cuerda vibra arriba y abajo, repitiendo el mismo movimiento una y otra vez. El período de oscilación, el tiempo entre las repeticiones sucesivas, es $2\pi/c$.

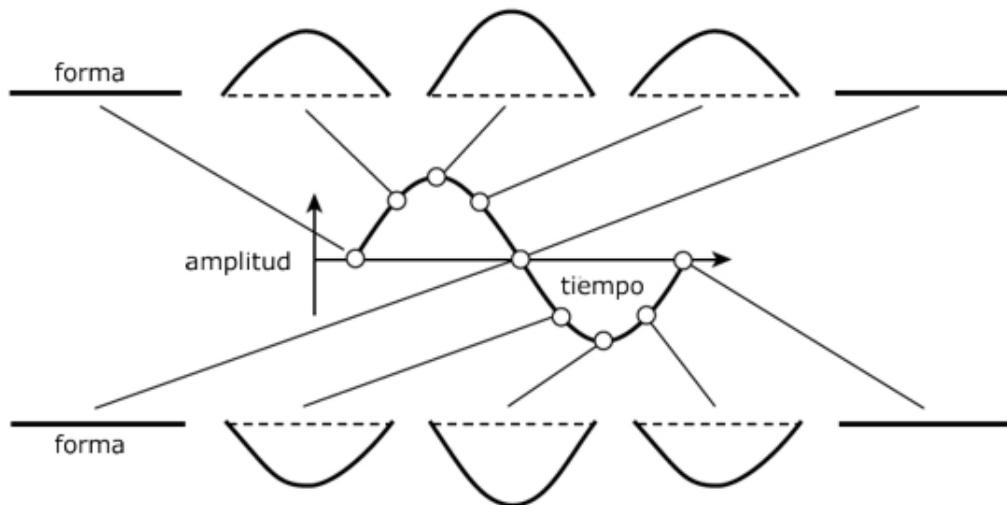


FIGURA 35. Instantáneas sucesivas de una cuerda vibrando. La forma es una curva sinusoidal en cada instante. La amplitud también varía sinusoidalmente con el tiempo.

Esta era la solución más simple que Bernoulli obtuvo, pero había otras; todas ellas curvas sinusoidales, diferentes «modos» de vibrar, con 1, 2, 3 o más ondas a lo largo de la longitud de la cuerda (figura 36). De nuevo, la curva sinusoidal era una instantánea de la forma en cualquier instante, y su amplitud se multiplicaba por un factor que dependía del tiempo, que también variaba sinusoidalmente. Las fórmulas eran $\sin 2ct \sin 2x$, $\sin 3ct \sin 3x$, etcétera. Los períodos de vibración eran $2\pi/2c$, $2\pi/3c$, etcétera; de modo que cuantas más ondas había, más rápido se movía la cuerda.

La cuerda está siempre inmóvil en sus extremos, por la construcción del instrumento y las suposiciones del modelo matemático. En todos los modos excepto el primero, hay puntos adicionales donde la cuerda no está vibrando, y estos se dan donde la curva se cruza con el eje horizontal. Estos «nodos» son la razón matemática para que ocurran proporciones numéricas sencillas en los experimentos pitagóricos. Por ejemplo, como el segundo y el tercer modos de vibración se dan en la misma cuerda, el hueco entre nodos sucesivos en la curva del segundo modo es

el hueco en la curva del tercer modo multiplicado por $3/2$. Esto explica por qué las proporciones como 3:2 surgen de manera natural a partir de las dinámicas del muelle que vibra, pero no por qué estas proporciones son armoniosas mientras que otras no lo son. Antes de abordar esta cuestión, introducimos el tema principal de este capítulo: la ecuación de onda.



FIGURA 36. Instantáneas de modos 1, 2, 3 de una cuerda vibrando. En cada caso, la cuerda vibra arriba y abajo y su amplitud varía sinusoidalmente con el tiempo.

Cuántas más ondas hay, más rápida es la vibración.

La ecuación de onda surge a partir de la segunda ley de movimiento de Newton si aplicamos la aproximación de Bernoulli al nivel de ecuaciones más que al de las soluciones. En 1746, Jean Le Rond D'Alembert siguió un procedimiento estándar, tratando una cuerda de violín vibrando como una colección de masas puntuales, pero en vez de resolver las ecuaciones y buscar un patrón cuando el número de masas tendía a infinito, calculó qué sucedía con las propias ecuaciones. Obtuvo una ecuación que describe cómo cambia la forma de la cuerda en el tiempo. Pero antes de que te muestre qué aspecto tiene, necesitamos una idea nueva, llamada una «derivada parcial».

Imagínate a ti mismo en medio del océano, observando las olas que pasan con diferentes formas y tamaños. A medida que lo hacen, te balanceas arriba y abajo. Físicamente puedes describir cómo está cambiando lo que te rodea de diferentes maneras. En particular, te puedes centrar en el cambio en el tiempo o en los cambios en el espacio. A medida que el tiempo pasa en tu posición, la velocidad a la que tu altura cambia con respecto al tiempo es la derivada (en el sentido del cálculo, capítulo 3) de tu altura, también con respecto al tiempo. Pero esto no describe la forma del océano que está a tu alrededor, solo cómo de altas son las olas cuando pasan por donde estás tú. Para describir la forma, puedes congelar el tiempo (conceptualmente) y calcular cómo de altas son las olas, no solo en tu

posición, sino en todas las cercanas. Entonces puedes usar el cálculo para determinar cómo de abruptamente las olas se inclinan en tu localización. ¿Están en un pico o en un hoyo? Si es así, la pendiente es cero. ¿Estás a medio camino del lado que desciende de la ola? Si es así, la pendiente es bastante grande. En términos del cálculo, puedes poner un número a esa pendiente calculando la derivada de la altura de las olas respecto al espacio.

Si una función u depende solo de una variable, llamémosla x , escribimos la derivada como du/dx : «un pequeño cambio en u dividido por un pequeño cambio en x ». Pero en el contexto de las olas del mar la función u , la altura de la ola, no solo depende del espacio x , sino que también depende del tiempo t . En cualquier instante fijo del tiempo, podemos todavía calcular du/dx , que nos dice la pendiente local de la ola. Pero en lugar de fijar el tiempo y permitir que el espacio varíe, podemos fijar el espacio y permitir que el tiempo varíe, esto nos dice la velocidad a la que estamos balanceándonos. Podemos usar la notación du/dt para esta «derivada del tiempo» e interpretarla como «un pequeño cambio en u dividido por un pequeño cambio en t ». Pero esta notación esconde una ambigüedad, el pequeño cambio en la altura, du , podría ser, y normalmente es, diferente en los dos casos. Si olvidas eso, es probable que hagas tus cálculos mal. Cuando estamos derivando con respecto al espacio, permitimos a la variable del espacio cambiar un poco y ver cómo la altura cambia; cuando estamos derivando con respecto al tiempo, permitimos a la variable del tiempo cambiar un poco y ver cómo la altura cambia. No hay ninguna razón por la que los cambios en el tiempo deban ser iguales a los cambios en el espacio.

De modo que los matemáticos decidieron recordarse a sí mismos esta ambigüedad cambiando el símbolo d por algo que no (directamente) les hiciese pensar en «pequeño cambio». Se decidieron por una d curvada muy linda, escrita ∂ . Luego escribieron las dos derivadas como $\partial u / \partial x$ y $\partial u / \partial t$. Puedes argumentar que esto no es un gran avance, porque es igual de fácil confundir los dos significados diferentes de ∂u . Hay dos respuestas a esta crítica. Una es que en este contexto se supone que no piensas en ∂u como un pequeño cambio específico en u . La otra es que usando un símbolo nuevo y chic te acuerdas de que no te tienes que confundir. La segunda respuesta definitivamente funciona; tan pronto como ves ∂ , te dice que estarás viendo las tasas de variación con respecto a varias variables diferentes.

Estas tasas de variación son llamadas derivadas parciales, porque conceptualmente solo cambias parte del conjunto de variables, manteniendo el resto fijas.

Cuando D'Alembert calculó su ecuación para la cuerda vibrando, se enfrentó justo a esta situación. La forma de la cuerda dependía del espacio —cuánta longitud de cuerda observes— y del tiempo. La segunda ley de movimiento de Newton le dijo que la aceleración de un pequeño segmento de cuerda es proporcional a la fuerza que actúa sobre él. La aceleración es una (segunda) derivada del tiempo. Pero la fuerza está causada por los segmentos vecinos de la cuerda tirando del segmento en el que estamos interesados, y «vecinos» quiere decir pequeños cambios en el espacio. Cuando calculó estas fuerzas, obtuvo la ecuación:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

Donde $u(x, t)$ es la posición vertical en la localización x en la cuerda en el momento t , y c es una constante relacionada con la tensión en la cuerda y cómo de elástica es. Los cálculos eran realmente más fáciles que los de Bernoulli, porque evitaban introducir características especiales de soluciones particulares.²⁴

La elegante fórmula de D'Alembert es la ecuación de onda. Como la segunda ley de Newton, es una ecuación diferencial y está involucrada la derivada (segunda) de u .

²⁴ Observa estas tres masas consecutivas, numeradas $n - 1$, n , $n + 1$. Supongamos que en el tiempo t , se desplazan las distancias $u_{n-1}(t)$, $u_n(t)$ y $u_{n+1}(t)$ desde sus posiciones iniciales en el eje horizontal. Por la segunda ley de Newton, la aceleración de cada masa es proporcional a las fuerzas que actúan sobre ella. Haz la suposición simplificada de que cada masa se mueve a través de una distancia muy pequeña solo en dirección vertical. Para una aproximación muy buena, la fuerza que la masa $n - 1$ ejerce sobre la masa n es entonces proporcional a la diferencia $u_{n-1}(t) - u_n(t)$, y de modo similar la fuerza que la masa $n + 1$ ejerce sobre la masa n es proporcional a la diferencia $u_{n+1}(t) - u_n(t)$. Sumándolas, la fuerza total ejercida en la masa n es proporcional a $u_{n-1}(t) - 2u_n(t) + u_{n+1}(t)$. Esta es la diferencia entre $u_{n-1}(t) - u_n(t)$ y $u_n(t) - u_{n+1}(t)$, y cada una de estas expresiones es también la diferencia entre las posiciones de masas consecutivas. De modo que la fuerza ejercida en la masa n es una diferencia entre diferencias.

Ahora supongamos que las masas están muy cerca la una de la otra. En cálculo, una diferencia, dividida por una constante pequeña adecuada, es una aproximación a una derivada. Una diferencia entre diferencias es una aproximación a una derivada de una derivada, es decir, una segunda derivada. En el límite de infinitud de masas puntuales, infinitesimalmente juntas, la fuerza ejercida en un punto dado del muelle es por tanto proporcional a $\partial^2 u / \partial x^2$, donde x es la coordenada del espacio medida a lo largo de la longitud de la cuerda. Por la segunda ley de Newton esto es proporcional a la aceleración en ángulo recto de esa línea, que es la segunda derivada del tiempo $\partial^2 u / \partial t^2$. Escribiendo la constante de proporcionalidad como c^2 obtenemos:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

Donde $u(x, t)$ es la posición vertical de la localización x en la cuerda en el momento t .

Ya que hay derivadas parciales, es una ecuación en derivadas parciales. La segunda derivada del espacio representa la fuerza neta actuando en la cuerda, y la segunda derivada del tiempo es la aceleración. La ecuación de onda sienta un precedente: la mayoría de las ecuaciones clave de la física matemática clásica, y muchas de la moderna, son ecuaciones en derivadas parciales.

Una vez D'Alembert había escrito su ecuación de onda, estaba en situación de resolverla. Esta tarea era mucho más fácil porque resultó ser una ecuación lineal. Las ecuaciones en derivadas parciales tienen muchas soluciones, habitualmente infinidad, porque cada estado inicial lleva a una solución distinta. Por ejemplo, la cuerda del violín puede en principio curvarse en cualquier forma que quieras, antes de soltarse y que la ecuación de ondas tome el mando. «Lineal» quiere decir que si $u(x, t)$ y $v(x, t)$ son soluciones, entonces lo es cualquier combinación lineal

$$au(x, t) + bv(x, t)$$

donde a y b son constantes. Otro término es «superposición». La linealidad de la ecuación de onda es producto de la aproximación que Bernoulli y D'Alembert tuvieron que hacer para obtener algo que pudiesen resolver; todas las alteraciones se presuponían pequeñas. Ahora una combinación lineal de desplazamientos de masas individuales puede ser una buena aproximación de la fuerza ejercida por la cuerda. Una aproximación mejor llevaría a una ecuación en derivadas parciales no lineal, y la vida sería muchísimo más complicada. A la larga, estas complicaciones tienen que afrontarse de frente, pero los precursores ya tenían suficiente con lo que lidiar, así que trabajaron con una ecuación aproximada, pero muy elegante, y restringieron su atención a ondas de amplitud pequeña. Funcionaba muy bien. De hecho, también funcionaba bastante bien para ondas de amplitudes mayores, un plus de suerte.

D'Alembert sabía que estaba en la senda correcta porque encontró soluciones en las cuales una forma fija viajaba a través de la cuerda, justo como una onda.²⁵ La velocidad de la onda resultó ser la constante c en la ecuación. La onda puede

²⁵ Para una animación véase: http://en.wikipedia.org/wiki/Wave_equation (o la versión española: http://es.wikipedia.org/wiki/Ecuaci%C3%B3n_de_onda).

moverse tanto a izquierda como a derecha, y aquí el principio de superposición entra en juego. D'Alembert probó que toda solución es una superposición de dos ondas, una desplazándose hacia la izquierda y la otra hacia la derecha. Además, cada onda separada podía tener cualquier forma fuera la que fuera.²⁶ Las ondas de posición encontradas en la cuerda del violín, con extremos fijos, resultaron ser una combinación de dos ondas con la misma forma, siendo la una la inversa de la otra, con una desplazándose a la izquierda y la otra (al revés) desplazándose a la derecha. En los extremos, las dos ondas se anulaban la una a la otra; picos de una coincidían con hoyos de la otra. De modo que cumplían con las condiciones de contorno físicas.

Los matemáticos ahora tenían un empacho de soluciones. Había dos modos de resolver la ecuación: la de Bernoulli, que llevaba a los senos y cosenos, y la de D'Alembert, que llevaba a ondas con cualquier forma que se desease. Al principio, parecía como si la solución de D'Alembert fuera a ser más general; senos y cosenos son funciones, pero la mayoría de las funciones no son senos y cosenos.

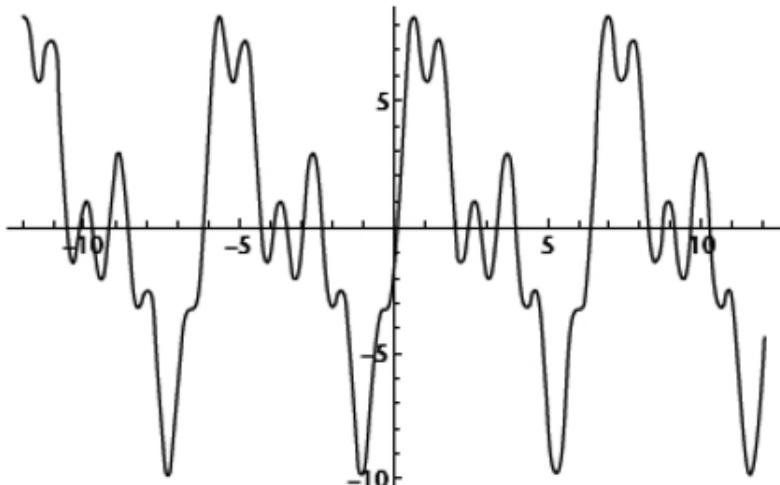


FIGURA 37. Combinación típica de senos y cosenos con varias amplitudes y frecuencias.

²⁶ En símbolos, las soluciones son precisamente las expresiones:

$$u(x, t) = f(x - ct) + g(x + ct)$$

para cualquier función f y g .

Sin embargo, la ecuación de onda es lineal, así que podrías combinar las soluciones de Bernoulli añadiendo múltiplos constantes. Para mantenerlo simple considera solo un instante de un tiempo fijo, librándote de la dependencia del tiempo. La figura 37 muestra $5 \operatorname{sen} x + 4 \operatorname{sen} 2x - 2 \cos 6x$, por ejemplo. Tiene una forma bastante irregular, y se curva mucho, pero es todavía suave y ondulada.

Lo que molestaba a los matemáticos más reflexivos era que algunas funciones eran muy abruptas y con picos, y no puedes obtenerlas como una combinación lineal de senos y cosenos. Bueno, no si usas una cantidad finita de términos, y eso sugería una salida. Una serie infinita convergente de senos y cosenos (una cuya suma en el infinito tenga sentido) también satisface la ecuación de onda. ¿Permitiría esto funciones dentadas a la vez que suaves? Los matemáticos importantes discutieron sobre esta cuestión, que finalmente llegó a un punto crítico cuando el mismo tema apareció en la teoría del calor. Los problemas sobre el flujo del calor involucraban de manera natural funciones discontinuas, con saltos repentinos, que eran incluso peores que las dentadas. Contaré esa historia en el capítulo 9, pero el resultado es que la mayoría de ondas con formas «razonables» pueden representarse por una serie infinita de senos y cosenos, de modo que pueden aproximarse tanto como se quiera por combinaciones finitas de senos y cosenos.

Los senos y cosenos explican las proporciones armoniosas que tanto impresionaron a los pitagóricos. Estas formas especiales de ondas son importantes en la teoría del sonido porque representan tonos «puros», notas sueltas en un instrumento ideal, por así decirlo. Cualquier instrumento real produce mezcla de notas puras. Si tiras de la cuerda de un violín, la nota principal que oyes es la onda de $\operatorname{sen} x$, pero superpuesta hay un poco de $\operatorname{sen} 2x$, quizás algo de $\operatorname{sen} 3x$, etcétera. La nota principal se llama la fundamental y las otras son sus armónicos. El número delante de x se llama el número de onda. Los cálculos de Bernoulli nos dicen que el número de onda es proporcional a la frecuencia, el número de veces que la cuerda vibra, para esa particular onda sinusoidal, durante una oscilación individual de la fundamental.

En concreto, $\operatorname{sen} 2x$ tiene dos veces la frecuencia de $\operatorname{sen} x$. ¿Cómo hace que suene? Es la nota una octava más alta. Es la nota que suena más armoniosa cuando se toca al lado de la fundamental. Si observas la forma de una cuerda durante el segundo

modo ($\sin 2x$) en la figura 36, te darás cuenta de que cruza el eje en sus puntos medios así como en los dos extremos. Permanece fija en ese punto, conocido como nodo. Si colocas tu dedo en ese punto, las dos mitades de la cuerda todavía serán capaces de vibrar siguiendo el patrón de $\sin 2x$, pero no el de $\sin x$. Esto explica el descubrimiento pitagórico de que una cuerda la mitad de larga produce una nota una octava mayor. Una explicación similar se ocupa de otras de proporciones simples que descubrieron, todas están asociadas con las curvas sinusoidales cuyas frecuencias tienen esa razón y dichas curvas encajan unas con otras pulcramente en una cuerda de una longitud fija cuyos extremos no está permitido que se muevan.

¿Por qué suenan armoniosas estas proporciones? Parte de la explicación es que las ondas senos con frecuencias que no están en proporciones simples producen un efecto llamado «batimiento» cuando se superponen. Por ejemplo, una proporción como 11: 23 se corresponde con $\sin 11x + \sin 23x$, que tiene el aspecto de la figura 38, con muchos cambios repentinos en la forma. Otra parte es que el oído responde a sonidos entrantes aproximadamente del mismo modo que la cuerda del violín. El oído también vibra. Cuando dos notas batan, el sonido correspondiente es como un zumbido que se va haciendo repetidamente más fuerte y más suave. De manera que no suena armonioso.

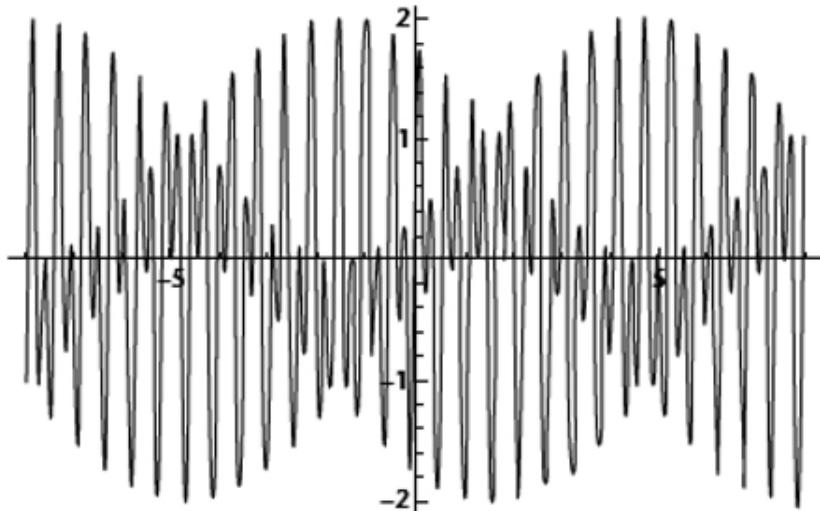


FIGURA 38. Batimientos.

Sin embargo, hay una tercera parte de la explicación: los oídos de los bebés se van compenetrando con los sonidos que oyen con más frecuencia. Hay más conexiones nerviosas del cerebro al oído de las que hay en la otra dirección. Así el cerebro ajusta la respuesta del oído a los sonidos entrantes. En otras palabras, lo que consideramos que es armonioso tiene una dimensión cultural. Pero las proporciones más simples son armoniosas de manera natural, de modo que la mayoría de las culturas las usan.

Los matemáticos primero obtuvieron la ecuación de onda en la versión más simple que se les ocurrió: una recta vibrando, un sistema unidimensional. Las aplicaciones realistas requieren una teoría más general, hacer modelos de ondas en dos y tres dimensiones. Incluso aunque nos quedemos en el terreno de la música, un tambor necesita dos dimensiones para hacer un modelo de los patrones según los cuales vibra la piel del tambor. Lo mismo se aplica para las olas marinas en la superficie del océano. Cuando hay un terremoto, toda la Tierra repica como una campana, y nuestro planeta es tridimensional. Muchas otras áreas de la física contienen modelos con dos y tres dimensiones. Extender la ecuación de onda a dimensiones mayores resultó ser directo y sencillo, todo lo que tenías que hacer era repetir el mismo tipo de cálculos que habían funcionado para la cuerda del violín. Al haber aprendido a jugar con la versión más simple del juego, no era difícil jugar con él de verdad.

En tres dimensiones, por ejemplo, usamos tres coordenadas espaciales (x, y, z) y el tiempo t . La onda está descrita por una función u que depende de estas cuatro coordenadas. Por ejemplo, esto podría describir la presión en un cuerpo de aire a medida que la onda de sonido pasa a través de él. Haciendo las mismas suposiciones que D'Alembert, en concreto que la amplitud de la alteración es pequeña, la misma aproximación nos lleva a una ecuación igualmente bella.

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right)$$

La fórmula dentro de los paréntesis se llama laplaciano, y se corresponde con la diferencia media entre el valor de u en el punto en cuestión y su valor cerca. Esta

expresión aparece con tanta frecuencia en la física matemática que tiene su propio símbolo especial: $\nabla^2 u$. Para obtener el laplaciano en dos dimensiones, tan solo omitimos el término con z y nos lleva a la ecuación de onda en esa versión.

La principal novedad en dimensiones mayores es que la forma con la que las ondas aparecen, llamada el dominio de la ecuación, puede ser complicada. En una dimensión la única forma relacionada es un intervalo, un segmento de la recta. Sin embargo, en dos dimensiones, puede ser cualquier forma que puedas dibujar en el plano y en tres dimensiones, cualquier forma en el espacio. Puedes hacer un modelo de un tambor cuadrado, un tambor rectangular, un tambor circular,²⁷ o un tambor con la forma de la silueta de un gato. Para los terremotos, podrías emplear un dominio esférico, o para una precisión mayor, un elipsoide ligeramente aplastado en los polos. Si estás diseñando un coche y quieres eliminar vibraciones no deseadas, tu dominio debería tener la forma de un coche, o cualquier parte del coche en la que los ingenieros se quieran centrar.

Para cualquier forma de dominio escogida, hay funciones análogas a los senos y cosenos de Bernoulli, los patrones de vibración más simples. Estos patrones se llaman modos, o modos normales si quieres dejar totalmente claro de lo que estás hablando. Todas las otras ondas se pueden obtener al superponer los modos normales, de nuevo usando una serie infinita si es necesario. Las frecuencias de los modos normales representan las frecuencias de vibración naturales del dominio. Si el dominio es rectangular, estas son funciones trigonométricas de la forma $\sin mx \cos ny$, para enteros m y n , produciendo ondas con formas como las de la figura 39 (izquierda). Si es un círculo, están determinadas por funciones nuevas, llamadas funciones de Bessel, con formas más interesantes, figura 39 (derecha). Las matemáticas resultantes se aplican no solo a tambores, sino a olas del mar, ondas de sonido, ondas electromagnéticas como las de la luz (capítulo 11), incluso ondas cuánticas (capítulo 14). Es fundamental en todas estas áreas. El laplaciano también aparece en ecuaciones para otros fenómenos físicos, en concreto, campos de gravedad, eléctricos y magnéticos. El truco favorito de los matemáticos de empezar con un modelo de juguete, uno tan simple que no es posible que sea

²⁷ Animaciones para los primeros pocos modos normales de un tambor circular se pueden encontrar en http://en.wikipedia.org/wiki/Vibrations_of_a_circular_drum. Hay animaciones de tambores circulares y rectangulares en: <http://www.mobiusilearn.com/viewcasestudies.aspx?id=2432>

realista, amortiza mucho tiempo en el caso de las ondas.

Esta es una razón por la que no es sabio juzgar la idea matemática por el contexto en el cual surge por primera vez. Hacer un modelo de la cuerda de un violín podía parecer que no tenía sentido cuando lo que querías era comprender terremotos. Pero si te metes de lleno en lo más complicado, y tratas de enfrentarte con todas las complejidades de los terremotos reales, te ahogarás. Debes empezar mojándote los pies en una zona poco profunda y ganar confianza para hacer unos largos en la piscina. Entonces estarás listo para un trampolín alto.

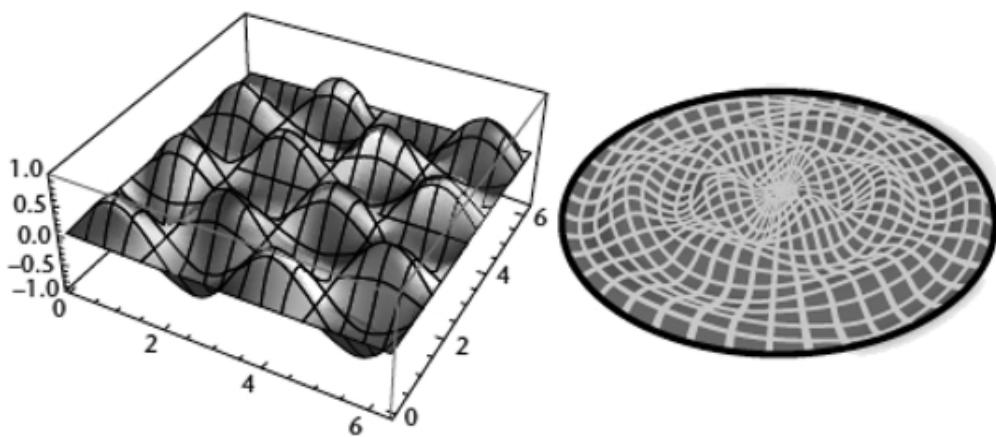


FIGURA 39. A la izquierda: instantánea del primer modo de vibración de un tambor rectangular, con número de onda 2 y 3. A la derecha: instantánea del primer modo de vibración de un tambor circular.

La ecuación de ondas fue un éxito espectacular, y en algunas áreas de la física describe una muy buena aproximación a la realidad. Sin embargo, obtenerla requiere varias suposiciones de simplificación. Cuando estas suposiciones no son realistas, las mismas ideas físicas se pueden modificar para que se adapten al contexto, llevando a diferentes versiones de la ecuación de onda.

Los terremotos son un ejemplo típico. Aquí el principal problema no es la suposición de D'Alembert de que la amplitud de la onda es pequeña, sino los cambios en las propiedades físicas del dominio. Estas propiedades pueden tener un efecto fuerte en las ondas sísmicas, vibraciones que se desplazan a través de la Tierra. Entendiendo estos efectos, podemos mirar a lo más profundo de nuestro planeta y averiguar de

qué está hecho.

Hay dos tipos principales de ondas sísmicas: ondas primarias o de presión y ondas secundarias o de superficie, normalmente abreviadas como ondas P y ondas S. (Hay otras muchas; esto es una simplificación, cubriendo algunos de los puntos esenciales.) Ambas pueden darse en un medio sólido, pero las ondas S no se dan en fluidos. Las ondas P son ondas de presión, análogas a las ondas de sonido en el aire, y los cambios de presión indican la dirección en la que la onda se propaga. Dichas ondas se dice que son longitudinales. Las ondas S son ondas transversales, cambian en ángulos rectos respecto a la dirección del desplazamiento, como las ondas de una cuerda de violín. Provocan que los sólidos se partan, es decir, se deformen como una baraja de cartas que se empuja por los laterales de modo que las cartas se desplazan a lo largo una de la otra. Los fluidos no se comportan como barajas de cartas.

Cuando ocurre un terremoto, envía ambos tipos de onda. Las ondas P viajan más rápido, de modo que un sismólogo en algún punto de la superficie de la Tierra observa estas primero. Luego, llegan las ondas S, más lentas. En 1906, el geólogo inglés Richard Oldham explotó esta diferencia para hacer un descubrimiento importante sobre el interior de nuestro planeta. En términos generales, la Tierra tiene un núcleo de hierro, rodeado por un manto rocoso, y los continentes flotan sobre ese manto. Oldham sugirió que las capas externas del núcleo deben ser líquidas. Si es así, las ondas S no pueden pasar a través de estas regiones, pero las ondas P sí que pueden. De modo que existe una especie de sombra de ondas S, y puedes averiguar dónde está observando señales de los terremotos. El matemático inglés Harold Jeffreys resolvió los detalles en 1926 y confirmó que Oldham tenía razón.

Si un terremoto es lo suficientemente grande, puede causar que el planeta entero vibre en uno de sus modos normales, los análogos para la Tierra de los senos y cosenos para un violín. El planeta entero repica como una campana, en un sentido que sería literal si solo pudiésemos oír las frecuencias involucradas muy bajas. Instrumentos lo suficiente sensibles para registrar estos modos surgieron en los años sesenta del siglo XX, y se usaron para observar los dos terremotos más potentes registrados científicamente por aquel entonces. Fueron el terremoto en

Chile de 1960 (magnitud 9,5) y el terremoto en Alaska de 1964 (magnitud 9,2). El primero mató a alrededor de 5.000 personas; el segundo mató a alrededor de 130 gracias a su localización remota. Ambos causaron tsunamis y provocaron enormes daños. Ambos ofrecieron una visión sin precedentes del interior más profundo de la Tierra, provocando los modos de vibración básicos de la Tierra.

Versiones sofisticadas de la ecuación de ondas han dado a los sismólogos la habilidad para ver qué está sucediendo a cientos de kilómetros bajo nuestros pies. Pueden hacer un mapa de las placas tectónicas de la Tierra a medida que una se desliza bajo otra, lo que se conoce como subducción. La subducción provoca terremotos, especialmente los conocidos como megaterremotos, como los dos que se acaban de mencionar. También provoca la elevación de cadenas montañosas a lo largo de los límites de los continentes como los Andes, y volcanes, donde la placa llega tan al fondo que empieza a fundirse y el magma sube a la superficie. Un descubrimiento reciente es que las placas no necesitan subducirse como un todo, sino que pueden romperse en bloques gigantescos, hundiéndose bajo el manto a profundidades diferentes.

El mayor premio en esta área sería un modo fiable de predecir terremotos y erupciones volcánicas. Lo que está resultando escurridizo, porque las condiciones que desencadenan dichos sucesos son combinaciones complejas de muchos factores en muchas localizaciones. Sin embargo, se han hecho algunos progresos, y la versión de los sismólogos de la ecuación de onda respalda muchos de los métodos que se están investigando.

Las mismas ecuaciones tienen aplicaciones más comerciales. Las compañías petrolíferas buscan oro negro, a pocos kilómetros bajo tierra, provocando explosiones en la superficie para diseñar la geología subyacente, usando el eco que vuelve a partir de las ondas sísmicas generadas. El principal problema matemático aquí es reconstruir la geología a partir de las señales recibidas, que es un poco como usar la ecuación de onda hacia atrás. En lugar de resolver la ecuación en un dominio conocido y calcular las ondas que causa, los matemáticos usan los patrones de ondas observados para reconstruir las características geológicas del dominio. Como es frecuente en estos casos, trabajar hacia atrás —resolver el problema inverso, en la jerga— es más difícil que ir en el otro sentido. Pero existen métodos

prácticos. Una de las compañías petrolíferas más importantes realiza estos cálculos un cuarto de un millón de veces cada día.

Perforar en busca de petróleo tiene sus propios problemas, como dejó claro el reventón en la plataforma petrolífera Deepwater Horizon en 2010. Pero por el momento, la sociedad humana depende muchísimo del petróleo, y llevaría décadas reducir esto de manera significativa, incluso aunque todo el mundo quisiera. La próxima vez que llenes tu depósito, acuérdate de los pioneros matemáticos que quisieron saber cómo un violín producía su sonido. No era un problema práctico entonces, y sigue sin serlo hoy en día. Pero sin sus descubrimientos, tu coche no te llevaría a ningún lado.

Capítulo 9

Ondas e instantes

Transformada de Fourier

The diagram shows the Fourier Transform equation:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-2\pi i x \xi} dx$$

Annotations explain various components:

- transformada de función**: transform of function
- frecuencia**: frequency
- integral**: integral
- menos infinito**: minus infinity
- infinito**: infinity
- función**: function
- espacio**: space
- 2.718...**: value of e
- 3.141...**: value of π
- raíz cuadrada de -1**: square root of -1 (i)
- frecuencia**: frequency (repeated)

¿Qué dice?

Cualquier patrón en el espacio y el tiempo se puede pensar como una superposición de patrones sinusoidales con diferentes frecuencias.

¿Por qué es importante?

Las frecuencias constituyentes se pueden usar para analizar los patrones, hacerlas a medida, extraer características importantes y eliminar ruido aleatorio.

¿Qué provocó?

La técnica de Fourier se usa muchísimo, por ejemplo, en tratamiento de imágenes y mecánica cuántica. Se usa para encontrar la estructura de moléculas biológicas grandes como el ADN, para comprimir datos de imágenes en fotografía digital, para limpiar grabaciones de audio viejas o dañadas y para analizar terremotos. Variantes modernas se usan para almacenar datos de huellas digitales de manera eficiente y mejorar escáneres médicos.

El *Principia* de Newton abrió la puerta al estudio matemático de la naturaleza, pero sus compatriotas estaban demasiado obsesionados con la discusión de la prioridad del cálculo como para encontrar qué se hallaba más allá. Mientras los mejores de Inglaterra estaban furiosos con lo que percibían como acusaciones vergonzosas sobre el más importante matemático vivo del país —gran parte de ello

probablemente fuese culpa suya por escuchar a amigos bienintencionados pero tontos—, sus colegas del continente estaban extendiendo las ideas de Newton sobre las leyes de la naturaleza a la mayoría de las ciencias físicas. A la ecuación de onda rápidamente le siguieron ecuaciones extraordinariamente similares para la gravitación, la electrostática, la elasticidad y el flujo del calor. Muchas llevan el nombre de sus inventores: la ecuación de Laplace, la ecuación de Poisson. La ecuación para el calor no, carga con el nombre carente de imaginación y no del todo preciso de «ecuación del calor». Fue introducida por Joseph Fourier y sus ideas llevaron a la creación de un área nueva de las matemáticas cuyas ramificaciones se extendieron mucho más allá de su fuente original. La ecuación de onda podría haber sido el desencadenante de estas ideas, donde métodos similares estaban deambulando en la conciencia matemática colectiva, pero la historia optó por el calor.

El método nuevo tenía un comienzo prometedor; en 1807 Fourier envió un artículo sobre el flujo del calor a la Academia de Ciencias francesa, basado en una nueva ecuación en derivadas parciales. Aunque ese prestigioso organismo rechazó publicar el trabajo, animó a Fourier a desarrollar más sus ideas e intentarlo de nuevo. En esa época, la academia ofrecía un premio anual para investigaciones en cualquier tema que encontrasen lo suficientemente interesante y escogió el calor como tema del premio de 1812. Fourier presentó debidamente su artículo revisado y ampliado, y ganó. Su ecuación del calor tiene el siguiente aspecto:

$$\frac{\partial u}{\partial t} = \alpha \frac{\partial^2 u}{\partial x^2}$$

Aquí $u(x, t)$ es la temperatura de una varilla de metal en la posición x y el tiempo t , considerando la varilla como infinitamente fina, y α es una constante, la difusividad térmica. De modo que realmente podría llamarse la ecuación de la temperatura. También desarrolló una versión para dimensiones mayores:

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u$$

Válida en cualquier región específica del plano o el espacio.

La ecuación del calor guarda un parecido asombroso con la ecuación de onda, con una diferencia crucial. La ecuación de onda usa la derivada segunda del tiempo $\partial^2 u / \partial t^2$, pero en la ecuación del calor esto es remplazado por la derivada primera $\partial u / \partial t$. Este cambio puede parecer pequeño, pero su significado físico es enorme. El calor no persiste indefinidamente, en el sentido de que una cuerda de violín vibrando continúa haciéndolo para siempre (según la ecuación de onda, la cual supone que no hay fricción u otro amortiguamiento). En cambio, el calor se disipa, se extingue a medida que el tiempo pasa, a menos que haya alguna fuente de calor que pueda recargarlo. Así que un problema típico podría ser: calienta un extremo de una varilla para mantener su temperatura estable, enfriá el otro extremo para hacer lo mismo, y averigua cómo la temperatura varía a lo largo de la varilla cuando se asienta en un estado estable. La respuesta es que cae exponencialmente. Otro problema típico es especificar el perfil de temperatura inicial a lo largo de la varilla, y luego preguntarse cómo cambia con el paso del tiempo. Quizá la mitad izquierda empiece a mayor temperatura y la mitad derecha esté más fría, la ecuación entonces nos dice cómo el calor se esparce de la parte caliente a la parte más fría. El aspecto más fascinante de la memoria premiada de Fourier no era la ecuación, sino cómo la resolvía. Cuando el perfil inicial es una función trigonométrica, como $\sin x$, es fácil (para quienes tienen experiencia en la materia) resolver la ecuación, y la respuesta es $e^{-at} \sin x$. Esto recuerda al modo fundamental de la ecuación de onda, pero allí la fórmula era $\sin ct \sin x$. La oscilación eterna de una cuerda de violín, correspondiente al factor $\sin ct$, ha sido remplazada por una exponencial, y el signo menos en el exponente, $-at$, nos dice que el perfil entero de la temperatura se extingue a la misma velocidad a lo largo de la varilla. (La diferencia física aquí es que la onda conserva la energía pero el flujo del calor no.) De manera similar, para un perfil $\sin 5x$, por ejemplo, la solución es $e^{-25at} \sin 5x$, que también se extingue, pero a un ritmo mucho más rápido. El 25 es 5^2 y este es un ejemplo de

un patrón general, aplicable a perfiles iniciales de la forma $\sin nx$ o $\cos nx$.²⁸ Para resolver la ecuación del calor, tan solo multiplicamos por e^{-n^2at} .

Ahora la historia sigue el mismo guión general que la ecuación de onda.

La ecuación del calor es lineal, de modo que podemos superponer las soluciones. Si el perfil inicial es:

$$u(x, 0) = \sin x + \sin 5x$$

Entonces la solución es:

$$u(x, t) = e^{-at} \sin x + e^{-25at} \sin 5x$$

y cada modo se desvanece a una velocidad diferente. Pero perfiles iniciales como este son un poco artificiales. Para resolver el problema que mencioné antes, queremos un perfil inicial donde $u(x, 0) = 1$ para la mitad de la varilla, y -1 para la otra mitad. Este perfil es discontinuo, una onda cuadrada en terminología de ingeniería. Pero las curvas del seno y el coseno son continuas. De modo que ninguna superposición de las curvas del seno y el coseno puede representar una onda cuadrada.

Ninguna superposición finita, desde luego. Pero, de nuevo, ¿qué pasa si permitimos infinitos términos? Entonces podemos intentar expresar el perfil inicial como una serie infinita de la forma:

$$\begin{aligned} u(x, 0) &= a_0 + a_1 \cos x + a_2 \cos 2x + a_3 \cos 3x + \dots \\ &\quad \dots + b_1 \sin x + b_2 \sin 2x + b_3 \sin 3x + \dots \end{aligned}$$

para las constantes adecuadas $a_0, a_1, a_2, a_3, \dots, b_1, b_2, b_3, \dots$ (No hay b_0 porque $\sin 0x = 0$.) Ahora parece posible obtener una onda cuadrada (véase la figura 40).

²⁸ Supón que $u(x, t) = e^{-n^2at} \sin nx$. Entonces:

$$\frac{\partial u}{\partial t} - n^2 a e^{-n^2at} \sin nx = \frac{\partial^2 u}{\partial x^2}$$

Por lo tanto, $u(x, t)$ satisface la ecuación del calor.

De hecho, la mayoría de los coeficientes pueden igualarse a 0. Solo se necesitan los b_n para n impar, y en este caso de la onda cuadrada $b_n = 8/n\pi$.

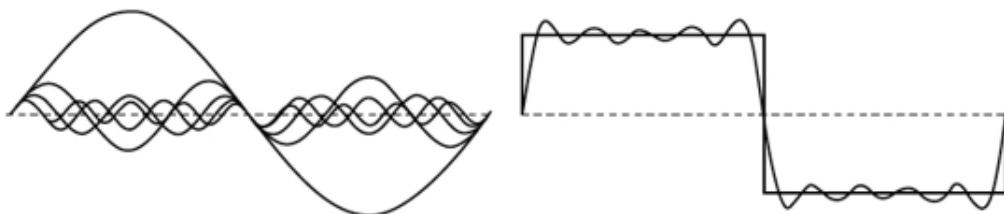


FIGURA 40. Cómo obtener una onda cuadrada a partir de senos y cosenos. A la izquierda: las ondas sinusoidales que lo componen. A la derecha: su suma y una onda cuadrada. Aquí mostramos unos pocos de los primeros términos de la serie de Fourier. Términos adicionales hacen la aproximación a la onda cuadrada incluso mejor.

Fourier incluso tenía fórmulas generales para los coeficientes a_n y b_n para un perfil general $f(x)$, en términos de integrales:

$$a_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \cos(nx) dx \quad b_n = \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(nx) dx$$

Después de una larga caminata a través de ampliaciones de series de potencias de funciones trigonométricas, se dio cuenta de que había un modo más simple de obtener estas fórmulas. Si tomas dos funciones trigonométricas diferentes, digamos $\cos 2x$ y $\sin 5x$, multiplicas la una por la otra y las integras entre 0 y 2π , el resultado es cero. Este incluso es el caso cuando tienen el aspecto de $\cos 5x$ y $\sin 5x$. Pero si son la misma, por ejemplo, iguales a $\sin 5x$, la integral de sus productos no es cero. De hecho, es π . Si empiezas suponiendo que $f(x)$ es la suma de una serie trigonométrica, multiplica todo por $\sin 5x$, e intégralo, todos los términos desaparecen excepto el que corresponde a $\sin 5x$, concretamente $b_5 \sin 5x$. Aquí la integral es π . Divide entre este resultado y tienes la fórmula de Fourier para b_5 . Y lo mismo se aplica para todos los otros coeficientes.

Aunque ganó el premio de la academia, la memoria de Fourier fue duramente

criticada por no ser lo suficientemente rigurosa, y la academia rechazó publicarla. Esto era muy inusual y molestó muchísimo a Fourier, pero la academia se mantuvo firme. Fourier estaba indignado. La intuición física le decía que estaba en lo correcto, y si introducías sus series en esta ecuación, era claramente una solución. Funcionaba. El problema real era que inconscientemente había abierto una vieja herida. Como vimos en el capítulo 8, Euler y Bernoulli habían estado discutiendo durante años por un tema parecido para la ecuación de onda, remplazando la disipación exponencial de Fourier en el tiempo por una oscilación sinusoidal sin fin en la amplitud de onda. Los temas matemáticos subyacentes eran idénticos. De hecho, Euler ya había publicado las fórmulas integrales para los coeficientes en el contexto de la ecuación de onda.

Sin embargo, Euler nunca afirmó que la fórmula funcionase para funciones discontinuas $f(x)$, la característica más polémica del trabajo de Fourier. El modelo de la cuerda del violín no envolvía condiciones iniciales discontinuas de ningún modo, eso habría sido el modelo para una cuerda rota, la cual no vibraría en absoluto. Pero para el calor, era natural considerar tener una región de una varilla a una temperatura y una región adyacente a otra temperatura diferente. En la práctica, la transición sería suave y muy pronunciada, pero un modelo discontinuo era razonable y más conveniente para los cálculos. De hecho, la solución a la ecuación del calor explicaba por qué la transición rápidamente se convertiría en suave y muy pronunciada a medida que el calor se difundiera por los lados. De modo que un tema por el que Euler no había necesitado preocuparse se estaba convirtiendo en inevitable, y Fourier sufrió las consecuencias.

Los matemáticos empezaban a darse cuenta de que las series infinitas eran bestias peligrosas. No siempre se comportaban como agradables sumas finitas. Finalmente, estas liosas complejidades se arreglaron, pero hacerlo requirió una visión nueva de las matemáticas y cientos de años de trabajo duro. En la época de Fourier, todo el mundo pensaba que ya sabía qué eran las integrales, funciones y series infinitas, pero en realidad todo era bastante vago —«la conozco cuando la veo»—. Así que cuando Fourier entregó el artículo que hizo época, había buenas razones para que los directivos de la academia fuesen cautelosos. No hubo quien los moviese, así que en 1822 Fourier sorteó sus objeciones publicando su trabajo como un libro, *Théorie*

analytique de la chaleur (Teoría analítica del calor). En 1824 lo nombraron secretario de la academia, se burló de todas las críticas y publicó su memoria original de 1811, sin cambios, en la prestigiosa publicación de la academia.

Sabemos ahora que aunque Fourier tenía razón en esencia, sus críticos tenían buenas razones para preocuparse por el rigor. Los problemas eran sutiles y las respuestas no eran terriblemente intuitivas. El análisis de Fourier, como lo llamamos ahora, funciona muy bien, pero tiene cualidades ocultas de las cuales Fourier no era consciente.

La cuestión parece ser: ¿cuándo las series de Fourier convergen a la función que supuestamente representan? Es decir, si consideras más y más términos, ¿la aproximación a la función es mejor? Incluso Fourier sabía que la respuesta no era «siempre». Parecía ser «habitualmente, pero con posibles problemas de discontinuidades». Por ejemplo, en su punto medio, donde la temperatura da un salto, la serie de Fourier de la onda cuadrada converge, pero al número equivocado. La suma es 0, pero la onda cuadrada toma valor 1.

Para los propósitos más físicos, no importa mucho si cambias el valor de una función en un punto aislado. La onda cuadrada, así modificada, todavía parece cuadrada. Tan solo hace algo ligeramente diferente en la discontinuidad. Para Fourier, este tipo de asunto no importaba realmente. Él estaba haciendo un modelo del flujo del calor, y no importaba si el modelo era un poco artificial, o necesitaba cambios técnicos que no tenían efectos importantes en el resultado final. Pero el asunto de la convergencia no podía desestimarse tan a la ligera, porque las funciones pueden tener discontinuidades mucho más complicadas que una onda cuadrada.

Sin embargo, Fourier estaba reivindicando que su método funcionaba para cualquier función, de modo que debería poder aplicarse incluso a funciones como: $f(x) = 0$ cuando x es racional, $f(x) = 1$ cuando x es irracional. Esta función es discontinua en todas partes. Para dichas funciones, en esa época, no estaba ni siquiera claro lo que significaba la integral. Y eso resultaba ser la causa real de la polémica. Nadie había definido qué era una integral, no para funciones extrañas como esta. Peor, nadie había definido qué era una función. E incluso aunque pudieses arreglar estos descuidos, no era solo un tema de si la serie de Fourier convergía. La dificultad real

era resolver en qué sentido convergía.

Resolver estos temas era complicado. Requería una nueva teoría de integración, aportada por Henri Lebesgue, una reformulación de los fundamentos de las matemáticas en términos de la teoría de conjuntos, empezada por Georg Cantor y que sacó a la luz varios problemas complicados totalmente nuevos, nuevas percepciones importantes a partir de personajes destacados como Riemann, y una dosis de abstracción del siglo XX para arreglar los problemas de convergencia. El veredicto final fue que, con las interpretaciones correctas, la idea de Fourier podría hacerse rigurosa. Funcionaba para una clase de funciones muy amplia, aunque no universal. La pregunta correcta no era si las series convergían a $f(x)$ para cada valor de x , todo estaba bien siempre que los valores excepcionales de x donde no convergía fuesen suficientemente raros, en un sentido preciso pero técnico. Si la función era continua, la serie convergía para cualquier x . En una discontinuidad de salto, como el cambio de 1 a -1 en la onda cuadrada, la serie convergía muy democráticamente a la media de los valores que están inmediatamente a ambos lados del salto. Pero las series siempre convergían a la función con la interpretación correcta de «converger». Convergía como un todo, más que punto por punto. Establecer esto rigurosamente dependía de encontrar el modo correcto de medir la distancia entre dos funciones. Con todo esto en juego, las series de Fourier sí que resolvían la ecuación del calor. Pero su importancia real era mucho más amplia, y el principal beneficiario fuera de las matemáticas puras no era la física del calor sino la ingeniería. Especialmente la ingeniería electrónica.

En su forma más general, el método de Fourier representa una señal, determinada por una función f , como una combinación de ondas de todas las frecuencias posibles. Esto se llama la transformada de Fourier de la onda. Reemplaza la señal original por su espectro: una lista de las amplitudes y frecuencias para los senos y cosenos que la componen, codificando la misma información de un modo diferente; los ingenieros hablan de transformación del dominio del tiempo en dominio de la frecuencia. Cuando los datos se representan de modos diferentes, las operaciones que eran difíciles o imposibles en una representación pueden convertirse en fáciles en la otra. Por ejemplo, puedes empezar con una conversación telefónica, formar su transformada de Fourier y eliminar todas las partes de las señales cuyas

componentes de Fourier tienen frecuencias demasiado altas o demasiado bajas como para que el oído humano las oiga. Esto hace posible enviar más conversación a través de los mismos canales de comunicación, y es una razón por la que las facturas de teléfono actuales son, hablando en términos relativos, tan bajas. No puedes jugar a este juego con la señal original sin transformar, porque no tiene la «frecuencia» como una característica obvia. No sabes qué eliminar.

Una aplicación de esta técnica es diseñar edificios que resistan terremotos. La transformada de Fourier de las vibraciones producidas por un terremoto típico revela, entre otras cosas, las frecuencias a las cuales es mayor la energía transmitida por el suelo cuando se mueve. Un edificio tiene sus modos propios de vibración naturales, donde resonará con el terremoto, esto es, responderá con una fuerza inusual. De modo que los primeros pasos sensatos hacia un edificio a prueba de terremotos es estar seguro de que las frecuencias preferidas del edificio son diferentes a las de los terremotos. Las frecuencias de los terremotos se pueden obtener a partir de la observación, las del edificio se pueden calcular usando un modelo informático.

Esto es solo uno de los muchos modos en los que, escondida entre bastidores, la transformada de Fourier afecta a nuestras vidas. La gente que vive o trabaja en edificios en zonas de terremotos no necesita saber cómo calcular la transformada de Fourier, pero su posibilidad de sobrevivir a un terremoto mejora considerablemente porque alguna gente lo hace. La transformada de Fourier se ha convertido en una herramienta de la rutina en ciencias e ingeniería; sus aplicaciones incluyen eliminar ruido de grabaciones de sonido antiguas, como chasquidos en discos de vinilo rayados, encontrar la estructura de grandes moléculas bioquímicas como el ADN usando la difracción de rayos X, mejorar la recepción de la radio, recoger fotografías tomadas desde el aire, sistemas de sonar como los usados por los submarinos, y prevenir vibraciones no deseadas en coches en la etapa de diseño. Me centraré tan solo en uno de los miles de usos diarios de la magnífica percepción de Fourier, uno que la mayoría de nosotros aprovechamos sin darnos cuenta cada vez que vamos de vacaciones: la fotografía digital.

En un viaje reciente a Camboya hice alrededor de 1.400 fotografías, usando una cámara digital, y todas estaban en una tarjeta de memoria de 2 GB con espacio

para alrededor de 400 fotos más. No hago fotografías con una resolución especialmente alta, así que cada archivo de foto es de más o menos 1,1 MB. Pero las imágenes están llenas de color, no presentan ninguna pixelación perceptible en una pantalla de ordenador de 27 pulgadas, de modo que la pérdida en la calidad no es obvia. De algún modo, mi cámara se las apaña para meter en una única tarjeta de 2 GB dos veces más datos de los que la tarjeta posiblemente pudiese albergar. Es como echar un litro de leche en una huevera. Aunque todo encaja. La pregunta es: ¿cómo?

La respuesta es la compresión de datos. La información que especifica la imagen es procesada para reducir su cantidad. Algo de este proceso es «sin pérdida», que quiere decir que la información original sin procesar puede, si es necesario, recuperarse a partir de la versión comprimida. Esto es posible porque la mayoría de las imágenes del mundo real contienen información redundante. Grandes bloques de cielo, por ejemplo, son con frecuencia del mismo tono de azul (bueno, lo son donde tendemos a ir). En vez de repetir la información del color y el brillo para un píxel azul una y otra vez, podrías almacenar las coordenadas de dos esquinas opuestas de un rectángulo y un pequeño código que signifique «el color en toda esta región es azul». Así no es exactamente como se hace, por supuesto, pero muestra por qué la compresión sin pérdida es a veces posible. Cuando no lo es, la compresión «con pérdida» es con frecuencia aceptable. El ojo humano no es especialmente sensible a ciertas características de las imágenes, y estas características pueden grabarse en una escala más gruesa sin que la mayoría de nosotros lo notemos, especialmente si no tenemos la imagen original para compararla. Comprimir la información de este modo es como hacer huevos revueltos: es fácil de hacer en una dirección y hace el trabajo necesario, pero no es posible dar marcha atrás. La información no redundante se pierde. Era solo información que para empezar no hacía mucho, debido a cómo funciona la visión humana.

Mi cámara, como la mayoría de las automáticas, guarda las imágenes en ficheros con etiquetas como P1020339.JPG. El sufijo se refiere a JPEG, *Joint Photographic Experts Group* (Grupo de expertos fotográficos unido), e indica que un sistema particular de compresión de datos ha sido usado. Software para manipular e

imprimir fotos, como Photoshop o iPhoto, se escriben de modo que puedan decodificar el formato JPEG y convertir los datos en una imagen. Millones de nosotros usamos archivos JPEG regularmente, menos son conscientes de que están comprimidos, y menos todavía se preguntan cómo se hace. No es una crítica, lo que quiero decir es que no tienes que saber cómo funciona para usarlo. La cámara y el software se encargan de todo por ti. Pero es con frecuencia sensato tener una idea aproximada de qué hace el software, y cómo lo hace, aunque sea solo para descubrir hasta qué punto son ingeniosos algunos. Puedes saltarte los detalles aquí si quieras; me gustaría que apreciaras tan solo cuántas matemáticas van en cada imagen en la tarjeta de memoria de tu cámara, saber exactamente cuál es menos importante.

El formato JPEG²⁹ combina cinco pasos de compresión diferentes. El primero convierte la información del color y el brillo, que empieza siendo tres intensidades para el rojo, verde y azul, en tres equivalentes matemáticamente que son más acordes al modo en que el cerebro humano percibe las imágenes. Uno (luminancia) representa el brillo total, que se vería con una versión de la misma imagen en blanco y negro o a escala de grises. Las otras dos (crominancia) son las diferencias entre esta y las cantidades de luz roja y azul, respectivamente.

A continuación, los datos de crominancia se hacen toscos; reducidos a rangos más pequeños de valores numéricos. Este paso solo reduce a la mitad la cantidad de datos. No hace un daño perceptible porque el sistema visual humano es mucho menos sensible a las diferencias de color de lo que lo es la cámara.

El tercer paso usa una variante para la transformada de Fourier. Esto no funciona con una señal que cambia en el tiempo, sino con un patrón en dos dimensiones del espacio. Las matemáticas son prácticamente idénticas. El espacio afectado es un sub-bloque de píxeles de 8×8 de la imagen. Por simplicidad piensa tan solo en la componente de la luminancia; la misma idea se aplica también a la información del color. Empezamos con un bloque de 64 píxeles y para cada uno de ellos necesitamos almacenar un número, el valor de la luminancia para ese pixel. La transformada de coseno discreta, un caso especial de la transformada de Fourier,

²⁹ Esto es codificación JFIF, usada para la web. El código EXIF, para cámaras, también incluye «metadata» describiendo los ajustes de la cámara, como la fecha, la hora y la exposición.

descompone la imagen en una superposición de imágenes «a rayas» estándar. En la mitad de ellas las rayas son horizontales, en la otra mitad son verticales. Están espaciadas en intervalos diferentes, como los diferentes armónicos en la transformada de Fourier habitual, y sus valores de la escala de grises son una aproximación cercana a la curva del coseno. En coordenadas sobre el bloque, son versiones discretas de $\cos mx \cos ny$ para distintos enteros m y n (véase la figura 41).

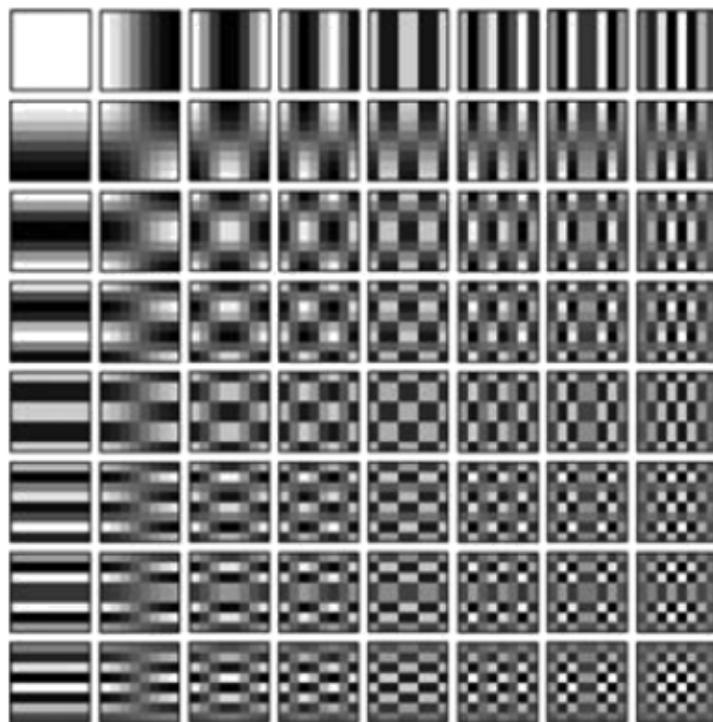


FIGURA 41. Los 64 patrones básicos a partir de los cuales cualquier bloque de 8×8 píxeles se puede obtener.

Este paso allana el camino para el paso cuatro, una segunda explotación de las deficiencias de la visión humana. Somos más sensibles a las variaciones en el brillo (o color) en regiones grandes de lo que lo somos a variaciones muy poco separadas. De modo que los patrones en la figura se pueden registrar con menor precisión a medida que la separación de las rayas se hace más fina. Esto comprime los datos más. El quinto y último paso usa un «código de Huffman» para expresar la lista de intensidades de los 64 patrones básicos de manera más eficiente.

Cada vez que haces una foto digital usando JPEG, la electrónica en tu cámara hace todas estas cosas, excepto quizá el paso uno. (Los profesionales se están decantando ahora por los archivos RAW, que graban los datos reales sin compresión, junto con los «metadatos» habituales como la fecha, hora, exposición, etcétera. Los archivos en este formato ocupan más memoria, pero la memoria se va haciendo más grande y más barata mes a mes, así que eso ya no importa.) Un ojo entrenado puede encontrar la pérdida de calidad de imagen creada por la compresión JPEG cuando la cantidad de datos se reduce a alrededor de un 10 % respecto a la original, y un ojo no entrenado puede verlo claramente el momento en que el tamaño del archivo se ve reducido a un 2-3 %. De modo que tu cámara puede grabar alrededor de diez veces más las imágenes que caben en una tarjeta de memoria, comparado con los datos de la imagen sin tratar, antes de que nadie que no sea experto lo note.

Gracias a aplicaciones como esta, el análisis de Fourier se ha convertido en algo intuitivo entre los ingenieros y los científicos, pero para algunos propósitos la técnica tiene un fallo importante: los senos y los cosenos no tienen fin. El método de Fourier se encuentra con problemas cuando trata de representar una señal compacta. Requiere una cantidad enorme de senos y cosenos imitar un instante localizado. El problema no es obtener la forma básica correcta del instante, sino hacer todo lo exterior al instante igual a cero. Tienes que erradicar la infinidad de largas colas de onda de todos esos senos y cosenos, lo cual haces añadiendo todavía más senos y cosenos de alta frecuencia en un desesperado esfuerzo por anular la basura no deseada. De modo que la transformada de Fourier no vale para señales del tipo instante; la versión transformada es más complicada, y necesita más datos para describirla, que la original.

Lo que lo salva es la generalidad del método de Fourier. Los senos y los cosenos funcionan porque satisfacen una condición simple: son matemáticamente independientes. Formalmente esto quiere decir que son ortogonales; en un sentido abstracto pero significativo, forman ángulos rectos los unos con los otros. Aquí es donde el truco de Euler, finalmente redescubierto por Fourier, entra en juego. Multiplicar dos de las ondas sinusoidales básicas la una por la otra e integrarlas en un período es un modo de medir cuán relacionadas están. Si este número es

grande, son muy similares; si es cero (la condición para la ortogonalidad), son independientes. El análisis de Fourier funciona porque sus ondas básicas son ambas ortogonales y completas; son independientes y hay suficientes para representar cualquier señal si se superponen del modo adecuado. En efecto, proporcionan un sistema de coordenadas en el espacio de todas las señales, como los tres ejes habituales del espacio ordinario. La principal característica nueva es que ahora tenemos infinidad de ejes; uno para cada onda básica. Pero esto no causa muchas dificultades matemáticamente, una vez te acostumbras. Solo quiere decir que tienes que trabajar con series infinitas en lugar de sumas finitas, y preocuparte un poco sobre cuándo converge la serie.

Incluso en espacios de dimensión finita, hay muchos sistemas de coordenadas diferentes, por ejemplo, los ejes pueden rotarse para apuntar en nuevas direcciones. No es sorprendente descubrir que, en un espacio de señales de dimensión infinita, hay sistemas de coordenadas alternativos que difieren completamente del de Fourier. Uno de los descubrimientos más importantes en toda el área, en los años recientes, es un nuevo sistema de coordenadas en el cual las ondas básicas son confinadas a regiones limitadas del espacio. Se llaman ondículas, y pueden representar instantes de manera muy eficiente porque son instantes.

No fue hasta recientemente cuando alguien se dio cuenta de que un análisis como el de Fourier para los instantes era posible. Empezar es directo y sencillo: escoge una forma concreta de un instante, la ondícula madre (figura 42). Entonces genera ondículas hijas (y nietas, bisnietas, etcétera) deslizando la ondícula madre hacia los lados en varias posiciones, y expandiéndola o comprimiéndola mediante un cambio en la escala. Del mismo modo, el seno y el coseno básico de Fourier son «sinusoidículas madre» y los senos y cosenos de frecuencias mayores son los hijos. Al ser periódicas, estas curvas no pueden ser



FIGURA 42. Ondícula de Daubechies.

como instantes.

Las ondículas están diseñadas para describir datos del tipo de instantes de manera eficiente. Además, como las ondículas hijas y nietas son tan solo versiones a otra escala de la madre, es posible centrarse en niveles concretos del detalle. Si no quieras ver estructuras a pequeña escala, tan solo elimina todas las ondículas bisnietas de la transformada de la ondícula. Para representar un leopardo por ondículas, necesitas unas pocas grandes para obtener el cuerpo correctamente, y más pequeñas para los ojos, la nariz y, por supuesto, los lunares, y algunas muy pequeñas para cada pelo individualmente. Para comprimir los datos que representan al leopardo, podrías decidir que los pelos individuales no importan, así que bastaría con eliminar esas ondículas en concreto. Lo bueno es que la imagen todavía parece un leopardo y todavía tiene lunares. Si tratas de hacer esto con la transformada de Fourier de un leopardo, entonces la lista de componentes es enorme, no está claro qué elementos deberías eliminar, y probablemente no reconocieses el resultado como un leopardo.

Todo está muy bien, pero ¿qué forma debería tener la ondícula madre? Durante mucho tiempo nadie pudo calcularlo, ni siquiera mostrar que existía una forma buena. Pero a principios de los ochenta en el siglo XX, el geofísico Jean Morlet y el físico matemático Alexander Grossmann encontraron la primera ondícula madre adecuada. En 1985 Yves Meyer encontró una ondícula madre mejor, y en 1987 Ingrid Daubechies, una matemática de los Laboratorios Bell, destapó el asunto por completo. Aunque las ondículas madre anteriores parecían adecuadas para los instantes, todas tenían una pequeñísima cola matemática que serpenteaba hasta el infinito. Daubechies encontró una ondícula madre sin ningún tipo de cola; fuera de algún intervalo, la madre era siempre exactamente cero, un instante genuino, confinado por completo a una región finita del espacio.

Las características como las de los instantes de las ondículas las hacen especialmente buenas para la compresión de imágenes. Uno de los primeros usos prácticos a gran escala fue almacenar huellas dactilares, y el cliente fue el Federal Bureau of Investigation. La base de datos de huellas dactilares del FBI contiene 300 millones de registros, cada uno de 10 huellas dactilares, que se almacenan originalmente como impresiones de tinta en tarjetas de papel. Este no es un medio

de almacenaje cómodo, así que los registros se han modernizado con la digitalización de las imágenes y el almacenamiento de los resultados en un ordenador. Ventajas obvias que incluyen la capacidad de organizar una búsqueda automatizada rápida de las huellas que coinciden con las encontradas en el escenario de un crimen.

El archivo del ordenador para cada tarjeta de huella dactilar es de 10 megabytes: 80 millones de dígitos binarios. De modo que el archivo entero ocupa 3.000 terabytes de memoria: 24.000 billones de dígitos binarios. Para hacer las cosas peor, el número de nuevos conjuntos de huellas dactilares crece en 30.000 cada día, así que la necesidad de almacenaje crecería en 2,4 billones de dígitos binarios cada día. El FBI, de manera sensata, decidió que necesitaban algún método para la compresión de datos. JPEG no era adecuado por varias razones, así que en 2002 el FBI decidió desarrollar un nuevo sistema de compresión usando ondículas, el algoritmo de cuantificación escalar de ondículas, representado por WSQ (siglas que provienen del nombre inglés *wavelet scalar quantization*). El WSQ reduce los datos un 5 % de su tamaño al eliminar pequeños detalles en toda la imagen. Esto es irrelevante para la capacidad de los ojos, como también para la de los ordenadores, para reconocer una huella dactilar.

Hay también muchas aplicaciones recientes de ondículas en imágenes médicas. Los hospitales emplean ahora varios tipos de escáner diferentes, que ensamblan secciones transversales bidimensionales del cuerpo humano u órganos importantes como el cerebro. Las técnicas incluyen CT (tomografía computarizada), PET (tomografía por emisión de positrones) e IRM (imagen por resonancia magnética). En la tomografía, la máquina observa la densidad total de los tejidos, o una cantidad similar, en una única dirección a través del cuerpo, un poco como lo que verías desde una posición fija si todos los tejidos fuesen ligeramente transparentes. Una imagen bidimensional puede reconstruirse aplicando algo de matemáticas inteligentes a toda una serie de dichas «proyecciones», tomadas desde muchos ángulos diferentes. En CT, cada proyección requiere una exposición de rayos X, de modo que hay buenas razones para limitar la cantidad de datos adquiridos. En todos esos métodos de escaneo, cuantos menos datos menos tiempo se necesita para recopilarlos, así que más pacientes pueden usar la misma cantidad de

equipamiento. Por otro lado, imágenes buenas necesitan más datos de modo que el método de reconstrucción pueda funcionar de modo más efectivo. Las ondículas proporcionan una vía media, en la cual se reduce la cantidad de datos pero obtenemos imágenes igualmente aceptables. Tomando una transformada de ondícula, eliminando las componentes no deseadas y «detransformando» a una imagen de nuevo, una imagen pobre puede ser suavizada y limpiada. Las ondículas también mejoran los métodos por los que los escáneres adquieren los datos al principio.

De hecho, están apareciendo ondículas casi por todas partes. Los investigadores en áreas tan alejadas como la geofísica y la ingeniería eléctrica están subiéndolas a bordo y poniéndolas a trabajar en sus propios campos. Ronald Coifman y Victor Wickerhauser las han usado para eliminar ruido no deseado de grabaciones; un triunfo reciente fue una actuación de Brahms tocando una de sus propias danzas húngaras. Originalmente se grabó en un cilindro de cera en 1889, que se había derretido parcialmente, fue regrabada en un disco a 78 rpm. Coifman empezó a partir de una retransmisión radiofónica del disco, en la cual la música era prácticamente inaudible entre el ruido de alrededor. Después de la limpieza con ondículas, podías oír lo que Brahms estaba tocando, no perfectamente, pero al menos era audible. Es una trayectoria impresionante para una idea que surgió inicialmente en la física del flujo de calor hace 200 años y su publicación fue rechazada.

Capítulo 10

La ascensión de la humanidad

Ecuación de Navier-Stokes

$$\rho \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) = -\nabla p + \nabla \cdot \mathbf{T} + \mathbf{f}$$

densidad velocidad presión estrés fuerzas del cuerpo
 derivada en el tiempo producto escalar gradiente divergencia

¿Qué dice?

Es la segunda ley de movimiento de Newton disfrazada. La parte izquierda es la aceleración de una región pequeña de un fluido. La parte derecha son las fuerzas que actúan en ella: presión, tensión y las fuerzas internas de los cuerpos.

¿Por qué es importante?

Proporciona un modo realmente preciso de calcular cómo los fluidos se mueven. Esto es una característica clave en innumerables problemas científicos y tecnológicos.

¿Qué provocó?

Aviones de pasajeros modernos, submarinos rápidos y silenciosos, coches de Fórmula 1 que se mantienen en la pista a velocidades altas y avances médicos en el flujo sanguíneo en venas y arterias. Métodos computacionales para resolver ecuaciones, conocidos como mecánica de fluidos computacional o CFD (por su nombre en inglés *computational fluid dynamics*), son muy usados por ingenieros para mejorar la tecnología en sus áreas.

Vista desde el espacio, la Tierra es una bonita esfera resplandeciente azul y blanca con parches verdes y marrones, bastante diferente a cualquier otro planeta en el Sistema Solar, es más, a cualquiera de los 500 planetas más conocidos hasta ahora

que están girando en torno a otras estrellas. La propia palabra «Tierra» rápidamente trae esta imagen a la mente. Aunque hace poco más de cincuenta años, la imagen casi universal para la misma palabra habría sido un puñado de tierra, tierra en el sentido de la jardinería. Antes del siglo XX, la gente miraba al cielo y se preguntaba por las estrellas y los planetas, pero lo hacían a ras de suelo. Que el hombre pudiese volar no era más que un sueño, tema de mitos y leyendas. Difícilmente nadie pensaba en viajar a otro mundo.

Unos pocos pioneros intrépidos empezaron a escalar lentamente al cielo. Los chinos fueron los primeros. Alrededor del 500 a.C., Lu Ban inventó un pájaro de madera que podría haber sido un planeador primitivo. En el 559 d.C., el presuntuoso Gao Yang ató con una correa a una cometa a Yuan Huangtou, el hijo del emperador, contra su voluntad, para espiar al enemigo desde arriba. Yuan sobrevivió a la experiencia pero fue ejecutado más tarde. Con el descubrimiento en el siglo XVII del hidrógeno, las ganas de volar se extendieron por Europa, inspirando a unos pocos individuos valientes a ascender a los tramos más bajos de la atmósfera terrestre en globos. El hidrógeno es un explosivo y en 1783 los hermanos franceses, Joseph-Michel y Jacques Étienne Montgolfier dieron una demostración pública de su nueva idea mucho más segura, el globo aerostático, primero con un vuelo de prueba sin tripulación, y luego con Étienne como piloto.

El ritmo del progreso, y las alturas a las cuales los humanos podrían ascender, empezaron a incrementar rápidamente. En 1903, Orville y Wilbur Wright hicieron el primer vuelo con motor en un aeroplano. La primera aerolínea, DELAG (Deutsche Luftschiffahrts-Aktiengesellschaft), empezó a operar en 1910, haciendo vuelos de pasajeros de Frankfurt a Baden-Baden y Düsseldorf usando dirigibles hechos por Zeppelin Corporation. En 1914 la St. Petersburg-Tampa Airboat Line hacía vuelos comerciales para pasajeros entre las dos ciudades de Florida, un viaje que duraba 23 minutos en el hidroavión de Tony Jannus. Los vuelos comerciales rápidamente se hicieron frecuentes, y llegó el avión a reacción; el De Havilland Comet empezó sus vuelos regulares en 1952, pero la fatiga del metal causó varios accidentes, y el Boeing 707 se convirtió en el líder desde su lanzamiento en 1958.

Personas corrientes podían ahora de manera rutinaria encontrarse a una altitud de 8 kilómetros, su límite hasta este día, al menos hasta que Virgin Galactic empiece sus

vuelos suborbitales. Vuelos militares y aviones experimentales subían a alturas mayores. Los vuelos espaciales, hasta la fecha el sueño de unos pocos visionarios, empezaron a ser una propuesta plausible. En 1961, el astronauta soviético Yuri Gagarin recorrió la órbita terrestre en el primer vuelo tripulado a bordo del *Vostok 1*. En 1969, la misión Apolo 11 de la NASA llevó a dos astronautas americanos, Neil Armstrong y Buzz Aldrin, a la Luna. El transbordador espacial empezó a funcionar en 1982, y mientras recortes en el presupuesto le impedían lograr sus objetivos iniciales —un vehículo reutilizable de respuesta rápida— se convirtió en uno de los caballos de batalla de los vuelos suborbitales, junto con la nave espacial rusa *Soyuz*. *Atlantis* ha hecho ahora el último vuelo del programa del transbordador espacial, pero se están planificando vehículos nuevos, principalmente por compañías privadas. Europa, India, China y Japón tienen sus programas y agencias espaciales propias.

El ascenso literal de la humanidad ha cambiado nuestra visión de quiénes somos y dónde vivimos, la principal razón de por qué «Tierra» ahora significa globo blanquiazul. Estos colores nos dan una pista de nuestra recién descubierta habilidad para volar. El azul es el agua, y el blanco es el vapor de agua en forma de nubes. La Tierra es un mundo de agua, con océanos, mares, ríos y lagos. Lo que mejor hace el agua es fluir, con frecuencia a lugares donde no es querida. Este fluir podría ser lluvia goteando desde un tejado o el poderoso torrente de una cascada. Puede ser tranquilo y claro, o agitado y turbulento; el fluir estable del Nilo a lo largo de lo que de otro modo sería desierto, o la espumosa agua blanca de sus seis cataratas.

Fueron los patrones formados por el agua, o, de manera más general, cualquier fluido en movimiento, lo que atrajo la atención de los matemáticos en el siglo XIX, cuando obtuvieron las primeras ecuaciones para el flujo de un fluido. El fluido vital para los vuelos es menos visible que el agua, pero tan omnipresente como ella: el aire. El flujo del aire es más complejo matemáticamente, porque el aire puede comprimirse. Modificando sus ecuaciones de modo que se apliquen a un fluido comprimible, los matemáticos iniciaron la ciencia que finalmente haría despegar la Era de la Aviación: la aerodinámica. Los primeros pioneros quizás volasen a ojo, pero las aerolíneas comerciales y el transbordador espacial vuelan porque los ingenieros han hecho los cálculos para hacerlos seguros y fiables (a menos que ocurra algo

imprevisible ocasionalmente). El diseño de aviones necesita una comprensión profunda de las matemáticas del flujo de fluidos. Y el pionero en la mecánica de fluidos fue el célebre matemático Leonhard Euler, que murió en el año que los Montgolfier hicieron su primer vuelo en globo.

Hay pocas áreas de las matemáticas hacia las que el prolífico Euler no dirigiese su atención. Se ha sugerido que la política era una razón para su producción prodigiosa y versátil, o más exactamente, el evitarla. Trabajó en Rusia durante muchos años, en la corte de Catalina la Grande, y un modo efectivo de evitar ser pillado en conspiraciones políticas, con consecuencias potencialmente desastrosas, era estar tan ocupado con las matemáticas que nadie creería que tenía tiempo libre para la política. Si esto es lo que estaba haciendo, tenemos que agradecer a la corte de Catalina muchos descubrimientos maravillosos. Pero yo me inclino a pensar que Euler era prolífico porque tenía ese tipo de mente. Creó cantidades enormes de matemáticas porque no podía ser de otro modo.

Había predecesores. Arquímedes estudió la estabilidad de cuerpos flotantes hace más de 2.200 años. En 1738 el matemático holandés Daniel Bernoulli publicó *Hydrodynamica* (Hidrodinámica), que contenía el principio de que los fluidos fluyen más rápido en regiones donde la presión es más baja. El principio de Bernoulli es con frecuencia invocado hoy para explicar por qué un avión puede volar: el ala se hace con una forma tal que el aire fluye más rápido a lo largo de la superficie de arriba, bajando la presión y creando la elevación. Esta explicación es un poco simplista, y muchos otros factores están involucrados en el vuelo, pero ilustra la cercana relación entre los principios matemáticos básicos y el diseño práctico de aviones. Bernoulli plasmó su principio en una ecuación algebraica relacionando velocidad y presión en un fluido incompresible.

En 1757, Euler volcó su mente fértil en el flujo de fluidos, publicando un artículo «Principes généraux du mouvement des fluides» (Principios generales del movimiento de fluidos) en las *Memorias* de la Academia de Berlín. Era el primer intento serio de hacer un modelo del flujo de fluidos usando una ecuación en derivadas parciales. Para mantener el problema dentro de unos límites razonables, Euler hizo algunas suposiciones que lo simplificaban; en particular, asumió que el fluido no era comprimible, era como el agua más que como el aire, y tenía

viscosidad cero, no era pegajoso. Estas suposiciones le permitieron encontrar algunas soluciones, pero también hizo su ecuación bastante poco realista. La ecuación de Euler se usa todavía hoy en día para algunos tipos de problemas, pero en su conjunto es demasiado simple para tener mucho uso práctico.

Dos científicos presentaron una ecuación más realista. Claude-Louis Navier era un físico e ingeniero francés; George Gabriel Stokes era un físico y matemático irlandés. Navier obtuvo un sistema de ecuaciones en derivadas parciales para el flujo de un fluido viscoso en 1822; Stokes empezó a publicar sobre el tema veinte años más tarde. El modelo resultante del flujo de fluidos es ahora conocido como ecuación de Navier-Stokes (con frecuencia se usa el plural porque la ecuación se plantea en términos de un vector, de modo que tiene varias componentes). Esta ecuación es tan precisa que en la actualidad los ingenieros a menudo usan soluciones informáticas en lugar de realizar pruebas físicas en túneles de viento. Esta técnica, conocida como mecánica de fluidos computacional, es ahora estándar en cualquier problema en el que haya flujo de fluidos: la aerodinámica del transbordador espacial, el diseño de coches de Fórmula 1 y coches comunes y la circulación sanguínea a través del cuerpo humano o un corazón artificial.

Hay dos modos de mirar la geometría de un fluido. Uno es seguir los movimientos de minúsculas partículas individuales de un fluido y ver adónde van. El otro es centrarse en las velocidades de dichas partículas; cómo de rápido, y en qué dirección, se están moviendo en cada instante. Los dos están muy relacionados, pero la relación es difícil de esclarecer excepto en aproximaciones numéricas. Una de las agudezas de Euler, Navier y Stokes fue darse cuenta de que todo parece mucho más simple en términos de las velocidades. El flujo de un fluido se comprende mejor en términos de campo de velocidades: una descripción matemática de cómo la velocidad varía de un punto a otro en el espacio y de un instante a otro en el tiempo. De modo que Euler, Navier y Stokes escribieron ecuaciones describiendo el campo de velocidad. Entonces, los patrones reales de flujo de un fluido pueden calcularse, al menos con una buena aproximación.

La ecuación de Navier-Stokes tiene este aspecto:

$$\rho \left(\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{v} \right) = -\nabla p + \nabla \cdot \mathbf{T} + \mathbf{f}$$

donde ρ es la densidad del fluido, \mathbf{v} es su campo de velocidad, p es la presión, \mathbf{T} determina la tensión, y \mathbf{f} representa las fuerzas del cuerpo, fuerzas que actúan por toda la región, no solo en su superficie. El punto es una operación entre vectores, y ∇ es una expresión en derivadas parciales, en concreto:

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z} \right)$$

La ecuación se obtiene a partir de física básica. Como con la ecuación de onda, un primer paso crucial es aplicar la segunda ley del movimiento de Newton para relacionar el movimiento de una partícula del fluido con las fuerzas que actúan sobre ella. La fuerza principal es la tensión elástica y esta tiene dos componentes principales: las fuerzas de fricción causadas por la viscosidad del fluido, y los efectos de la presión, tanto positivos (compresión) como negativos (rarefacción). Hay también fuerzas del cuerpo, que son producto de la aceleración de las propias partículas del fluido. Combinando toda esta información llegamos a la ecuación de Navier-Stokes, que puede verse como una exposición de la ley de conservación del momento en este contexto particular. La física subyacente es impecable, y el modelo es lo suficiente realista al incluir la mayoría de los factores significativos; razón por la que concuerda con la realidad tan bien. Como todas las ecuaciones tradicionales de la física matemática clásica es un modelo continuo: asume que el fluido es divisible infinitamente.

Esto es quizá el punto principal donde la ecuación de Navier-Stokes pierde contacto con la realidad, pero la discrepancia solo aparece cuando el movimiento envuelve cambios rápidos a la escala de moléculas individuales. Dichos movimientos a pequeña escala son importantes en un contexto fundamental: las turbulencias. Si abres un grifo y dejas el agua fluir lentamente, sale un hilillo de agua liso. Sin embargo, abre el grifo del todo y normalmente lo que tienes es un chorro de agua espumoso y a borbotones. Flujos espumosos similares ocurren en los rápidos de un

río. Este efecto es conocido como turbulencia, y aquellos de nosotros que vuelan regularmente son conscientes de sus efectos cuando ocurren en el aire. Se siente como si el avión estuviese yendo por una carretera llena de baches.

Resolver la ecuación de Navier-Stokes es difícil. Hasta que se inventaron ordenadores realmente rápidos, era tan difícil que los matemáticos no tenían otra alternativa que recurrir a atajos y aproximaciones. Pero es que debería ser difícil si piensas en lo que un fluido real puede hacer. Tan solo tienes que echar un vistazo al agua fluyendo en un riachuelo o las olas rompiendo en la playa, para ver que el fluido puede fluir de maneras extremadamente complejas. Hay olas y torbellinos, patrones de onda y remolinos, y estructuras fascinantes como el macareo del Severn, en el suroeste de Inglaterra, cuando sube la marea. Estos patrones de flujo de fluidos han sido la fuente de innumerables investigaciones matemáticas, aunque una de las mayores y más básicas cuestiones en el área sigue sin resolverse: ¿hay alguna garantía matemática de que las soluciones de la ecuación de Navier-Stokes realmente existan, válidas para todo tiempo futuro? Hay un premio de un millón de dólares para quien sea capaz de resolverlo, uno de los siete problemas del milenio de los premios del Instituto Clay, escogidos por representar los problemas matemáticos más importantes sin resolver de nuestra era. La respuesta es «sí» en un flujo bidimensional, pero nadie sabe la respuesta para un flujo tridimensional.

A pesar de esto, la ecuación de Navier-Stokes proporciona un modelo útil del flujo de turbulencias porque las moléculas son extremadamente pequeñas. Vórtices turbulentos de unos pocos milímetros ya capturan muchas de las principales características de las turbulencias, mientras que una molécula es mucho más pequeña, así que un modelo continuo sigue siendo apropiado. El problema principal provocado por la turbulencia es práctico: hace prácticamente imposible resolver la ecuación de Navier-Stokes numéricamente, porque un ordenador no puede manejar cálculos infinitamente complejos. Las soluciones numéricas de ecuaciones en derivadas parciales usan una rejilla, dividen el espacio en regiones discretas y el tiempo en intervalos discretos. Para abarcar el amplio rango de escalas en las que las turbulencias operan —sus vórtices grandes, los medianos, bajando directamente a los de escala milimétrica— necesitas una rejilla informática increíblemente fina. Por esta razón, los ingenieros con frecuencia usan en su lugar modelos estadísticos

de turbulencias.

La ecuación de Navier-Stokes ha revolucionado el transporte moderno. Quizá su mayor influencia es en el diseño de aviones de pasajeros, porque no solo hace que estos vuelen de manera eficiente, sino que tienen que volar de manera estable y fiable. El diseño de barcos también se beneficia de la ecuación, porque el agua es un fluido. Incluso coches familiares ordinarios están ahora diseñados sobre principios aerodinámicos, no solo porque les hace parecer elegantes y modernos, sino porque el consumo eficiente de combustible tiene que ver con minimizar la resistencia causada por el flujo de aire que pasa por el vehículo. Un modo de reducir tu huella de carbono es conducir un coche eficiente en el sentido aerodinámico. Por supuesto hay otros modos, que van desde coches más pequeños y más lentos a motores eléctricos, o conducir menos. Algunas de las grandes mejoras en las cifras de consumo de combustible han resultado de mejorar la tecnología del motor, otras de una aerodinámica mejor.

En los inicios del diseño de aviones, los pioneros montaron sus aeroplanos usando cálculos aproximados, intuición física y ensayo-error. Cuando tu objetivo era volar más de un centenar de metros a no más de tres metros de altura, eso era suficiente. La primera vez que *Wright Flyer I* despegó realmente, en lugar de calarse y estrellarse después de tres segundos en el aire, recorrió 36,576 metros a una velocidad por debajo de los 11,263 km/h. Orville, el piloto, en esa ocasión se las arregló para mantenerlo en el aire durante unos asombrosos 12 segundos. Pero el tamaño del avión de pasajeros creció rápidamente, por razones económicas; cuanta más gente puedas llevar en un vuelo, más beneficio puedes obtener. Pronto el diseño de aviones tuvo que basarse en métodos más racionales y fiables. La ciencia de la aerodinámica había nacido y sus herramientas matemáticas básicas eran las ecuaciones para el flujo de fluidos. Como el aire es tanto viscoso como comprimible, la ecuación de Navier-Stokes, o alguna simplificación que tenga sentido en un problema dado, cobró protagonismo a medida que la teoría avanzaba. Sin embargo, resolver estas ecuaciones, sin contar con ordenadores modernos, era prácticamente imposible. De modo que los ingenieros recurrieron a un ordenador analógico: poner modelos de aviones en un túnel de viento. Usando unas pocas propiedades generales de las ecuaciones para calcular cómo cambian las variables a

medida que la escala del modelo cambia, este método proporciona información básica de modo rápido y fiable. La mayoría de los equipos de Fórmula 1 en la actualidad usan túneles de viento para probar sus diseños y evaluar mejoras potenciales, pero la potencia de los ordenadores es ahora tan grande que la mayoría también usan CFD. Por ejemplo, la figura 43 muestra un cálculo de CFD del flujo del aire pasando por un coche de BMW Sauber. Cuando escribo esto, un equipo, Virgin Racing, usa solo CFD, pero también usarán un túnel de viento el próximo año.

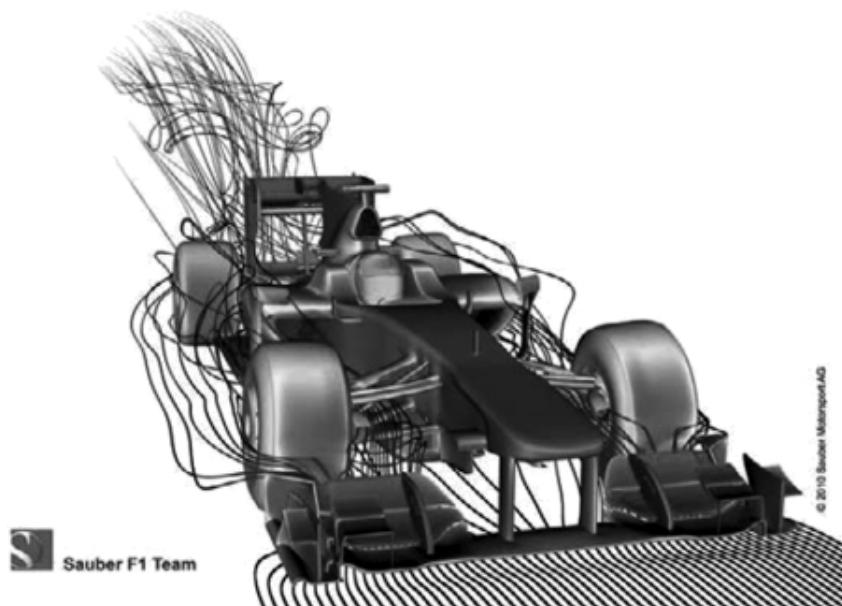


FIGURA 43. Cálculo del flujo del aire que pasa por un coche de Fórmula 1.

Los túneles de viento no son tremadamente convenientes, son caros de construir y mantener, y necesitan muchos modelos a escala. Quizá la mayor dificultad es hacer medidas precisas del flujo del aire sin que les afecte. Si pones un instrumento en el túnel de viento para medir, digamos, la presión del aire, entonces el propio instrumento altera el flujo. Quizá la mayor ventaja práctica del CFD es que puedes calcular el flujo sin alterarlo. Cualquier cosa que quieras medir es fácil de conseguir. Además, puedes modificar el diseño del coche o un componente en el software, que es mucho más rápido y barato que hacerlo en muchos de los diferentes modelos. De todos modos, los procesos de fabricación modernos con frecuencia involucran modelos informáticos en la etapa de diseño.

El vuelo supersónico, en el que el avión va más rápido que el sonido, es

especialmente difícil de estudiar usando modelos en un túnel de viento, porque las velocidades del viento son muy grandes. A esas velocidades, el aire no se puede apartar del avión tan rápido como el avión se empuja a sí mismo a través del aire, y esto provoca ondas expansivas —discontinuidades repentinas en la presión del aire, oídas en la tierra como un estampido—. Este problema con el entorno fue una razón por la que el avión anglo-francés Concorde, el único avión comercial supersónico que ha operado alguna vez, tuvo un éxito limitado; no estaba permitido volar a velocidades supersónicas excepto sobre el océano. La CFD es muy usada para predecir el flujo del aire al pasar un avión supersónico.

Hay alrededor de 600 millones de coches en el planeta y decenas de miles de aviones civiles, así que aunque estas aplicaciones de la CFD puedan parecer alta tecnología, son importantes en la vida diaria. Otros modos de usar la CFD tienen una dimensión más humana. Los investigadores médicos la usan mucho para entender el flujo sanguíneo en el cuerpo humano, por ejemplo. Las disfunciones cardíacas son una de las causas principales de muerte en el mundo desarrollado, y pueden desencadenarse por problemas con el propio corazón o por arterias obstruidas, que interrumpen el flujo sanguíneo y pueden causar coágulos. Las matemáticas del flujo sanguíneo en el cuerpo humano son especialmente intratables analíticamente porque las paredes de las arterias son elásticas. Si ya es bastante difícil calcular el movimiento de un fluido a través de un tubo rígido, es mucho más difícil si el tubo puede cambiar su forma dependiendo de la presión que el fluido ejerce, ya que en ese caso el dominio para el cálculo no es el mismo a medida que el tiempo pasa. La forma del dominio afecta el patrón del flujo del fluido y simultáneamente el patrón del flujo del fluido afecta a la forma del dominio. Las matemáticas hechas con lápiz y papel no pueden manejar este tipo de bucle de retroalimentación.

La CFD es ideal para este tipo de problema porque los ordenadores pueden realizar billones de cálculos cada segundo. La ecuación tiene que modificarse para incluir los efectos de las paredes elásticas, pero esto es principalmente un asunto de extracción de los principios necesarios de la teoría de la elasticidad, otra parte de la mecánica de medios continuos clásica muy desarrollada. Por ejemplo, un cálculo de la CFD de cómo la sangre fluye por la aorta, la arteria principal que llega al corazón,

ha sido llevado a cabo en la École Polytechnique Fédérale de Lausanne en Suiza. Los resultados proporcionan información que puede ayudar a los doctores a obtener una comprensión mejor de los problemas cardiovasculares.

También ayuda a los ingenieros a desarrollar aparatos médicos mejorados como *stents*, pequeños tubos de malla metálica que mantienen las arterias abiertas. Suncica Canic ha usado la CFD y modelos de propiedades elásticas para diseñar mejores *stents*, obteniendo un teorema matemático que provocó que se abandonase un diseño y sugirió diseños mejores. Los modelos de este tipo han llegado a ser tan precisos que la Agencia de alimentos y medicamentos de EE.UU. está considerando exigir a cualquier grupo que diseñe *stents* que realice modelos matemáticos antes de realizar ensayos clínicos. Los matemáticos y los doctores están uniéndose fuerzas para usar la ecuación de Navier-Stokes con el fin de obtener predicciones y tratamientos mejores para las causas principales de los ataques de corazón.

Otra aplicación relacionada son las operaciones de bypass coronario, en las cuales se elimina una vena de algún lugar en el cuerpo y se injerta en la arteria coronaria. La geometría del injerto tiene un gran efecto en el flujo sanguíneo. Esto a su vez afecta a la coagulación, que es más probable si el flujo tiene vórtices porque la sangre puede quedarse atrapada en un vórtice y no circular adecuadamente. De modo que aquí vemos un vínculo directo entre la geometría del flujo y los problemas médicos potenciales.

La ecuación de Navier-Stokes tiene otra aplicación: el cambio climático, también conocido como calentamiento global. El clima y el tiempo están relacionados, pero son diferentes. El tiempo es lo que pasa en un lugar dado en un momento dado. Puede estar lloviendo en Londres, nevando en Nueva York o hacer un calor achicharrante en el Sahara. El tiempo es claramente impredecible, y hay buenas razones matemáticas para esto: véase el capítulo 16 sobre el caos. Sin embargo, mucha de su impredecibilidad concierne a cambios a pequeña escala, tanto en el espacio como en el tiempo; detalles sutiles. Si el hombre del tiempo de la televisión predice chubascos en tu ciudad mañana por la tarde y suceden seis horas antes y a 20 kilómetros, él cree que hizo un buen trabajo y tú apenas estás impresionado. El clima es la «textura» del tiempo a largo plazo, cómo la lluvia y la temperatura se

comportan cuando se hace el promedio de períodos largos, décadas incluso. Debido a que el clima calcula el promedio de estas discrepancias, es paradójicamente más fácil de predecir. Las dificultades son todavía considerables, y mucho de la literatura científica investiga posibles fuentes de error intentando mejorar los modelos.

El cambio climático es un tema polémico políticamente, a pesar de un gran consenso científico en que la actividad humana durante el siglo pasado más o menos ha provocado que la temperatura media de la Tierra suba. El incremento hasta la fecha suena pequeño, alrededor de 0,75º Celsius durante el siglo XX, pero el clima es muy sensible a los cambios de temperatura en una escala global. Tienden a hacer el tiempo más extremo, haciéndose más comunes las sequías e inundaciones.

El «calentamiento global» no implica que la temperatura cambie la misma cantidad minúscula en todas partes. Al contrario, hay grandes fluctuaciones de un lugar a otro y de un momento a otro. En 2010, Gran Bretaña experimentó su invierno más frío en 31 años, dando lugar a que el *Daily Express* publicase el titular «y todavía afirman que es calentamiento global». Da la casualidad de que 2010 junto con 2005 son los años más calurosos registrados en todo el planeta.³⁰ Así que tenían razón. De hecho, la ola fría fue provocada por la corriente en chorro que cambió de posición, empujando aire frío al sur desde el Ártico, y esto sucedió porque el Ártico estaba inusualmente cálido. Dos semanas de helada en el centro de Londres no desacredita el calentamiento global. Curiosamente, el mismo periódico informó que el domingo de Pascua de 2011 fue el más cálido registrado, pero no hizo ninguna conexión con el calentamiento global. En esa ocasión distinguieron correctamente tiempo de clima. Estoy fascinado por el enfoque selectivo.

De modo similar, «cambio climático» no simplemente significa que el clima esté cambiando. Ha hecho eso sin la ayuda humana de manera repetitiva, principalmente en períodos de tiempo largos, gracias a cenizas volcánicas y gases, a las variaciones a largo plazo en la órbita terrestre alrededor del Sol, e incluso a la India colisionando con Asia para crear la cordillera del Himalaya. En el contexto en que está actualmente debatiéndose, «cambio climático» es la expresión corta para «cambio climático antropogénico», cambios en el clima global causados por la

³⁰ <http://www.nasa.gov/topics/earth/features/2010-warmest-year.html>

actividad humana. Las principales causas son la producción de dos gases: dióxido de carbono y metano. Son gases de efecto invernadero: atrapan las radiaciones entrantes (calor) del Sol. La física básica implica que cuantos más de estos gases contenga la atmósfera, más calor atrapa; aunque el planeta irradie algo de calor lejos, en general estará más caliente. El calentamiento global fue predicho, sobre estas bases, en la década de los cincuenta del siglo XX, y el incremento de la temperatura pronosticada es acorde con lo que se ha observado.

La evidencia de que los niveles de dióxido de carbono han incrementado drásticamente viene de muchas fuentes. La más directa son los núcleos de hielo. Cuando la nieve cae en las regiones polares, se aglutina para formar hielo, con la nieve más reciente en la cima y la más vieja en el fondo. El aire está atrapado en el hielo, y las condiciones que prevalecen ahí lo dejan prácticamente sin cambios durante períodos de tiempo muy largos, manteniendo el aire original en el interior y el más reciente en el exterior. Con cuidado, es posible medir la composición del aire atrapado y determinar, con mucha exactitud, la fecha en la que se quedó atrapado. Las mediciones hechas en la Antártida muestran que la concentración de dióxido de carbono en la atmósfera era prácticamente constante durante los pasados 100.000 años, excepto por los últimos 200, cuando se disparó en un 30 %. La fuente del exceso de dióxido de carbono puede deducirse a partir de las proporciones del carbono-13, uno de los isótopos (formas atómicas diferentes) del carbono. La actividad humana es con mucho la explicación más probable.

La principal razón por la que los escépticos tienen todavía ligeros motivos para serlo es la complejidad de la predicción climática. La cual hay que hacerla usando modelos matemáticos, porque es sobre el futuro. Ningún modelo puede incluir cada una de todas las características del mundo real, y si lo hiciese, nunca podrías calcular qué predice, porque ningún ordenador podría simularlo. Toda discrepancia entre el modelo y la realidad, no obstante insignificantes, es música para los oídos escépticos. Con certeza hay espacio para las diferencias de opinión sobre los posibles efectos del cambio climático, o qué deberíamos hacer para mitigarlo. Pero enterrar nuestras cabezas en la tierra como el aveSTRUZ no es una opción sensata.

Los dos aspectos vitales del clima son la atmósfera y los océanos. Ambos son fluidos, y ambos pueden estudiarse usando la ecuación de Navier-Stokes. En 2010,

el principal organismo de financiación de ciencia de Reino Unido, el Consejo de Investigación de Ciencias Físicas e Ingeniería, publicó un documento sobre el cambio climático, señalando las matemáticas como una fuerza unificadora: «todos los investigadores en meteorología, física, geografía y una gran cantidad de otros campos aportan su pericia, pero las matemáticas son el lenguaje que unifica y permite a estos diferentes grupos de gente implementar sus ideas en los modelos climáticos». El documento también explica que: «los secretos del sistema climático están guardados bajo llave en la ecuación de Navier-Stokes, pero es demasiado compleja para resolverse directamente». En su lugar, los modelos de clima usan métodos numéricos para calcular el flujo del fluido en los puntos de una rejilla tridimensional que cubre el planeta desde la profundidad de los océanos hasta las cotas superiores de la atmósfera. La separación horizontal de la rejilla es 100 kilómetros, algo más pequeño haría los cálculos poco prácticos. Ordenadores más rápidos no ayudarán mucho, así que el mejor modo de proceder es pensar más. Los matemáticos están trabajando en modos más eficientes de resolver la ecuación de Navier-Stokes numéricamente.

La ecuación de Navier-Stokes es solo parte del rompecabezas del clima. Otros factores incluyen el flujo de calor en, y entre, los océanos y la atmósfera, el efecto de las nubes, contribuciones no humanas tales como los volcanes, incluso emisiones de los aviones en la atmósfera. A los escépticos les gusta enfatizar dichos factores para sugerir que los modelos son erróneos, pero la mayoría de ellos se sabe que son irrelevantes. Por ejemplo, cada año los volcanes aportan un escaso 0,6 % al dióxido de carbono producido por la actividad humana. Todos los modelos principales sugieren que hay un problema serio, y los humanos lo han provocado. La principal cuestión es cuánto se calentará el planeta, y cómo de desastroso resultará. Como predicciones perfectas son imposibles, interesa a todo el mundo asegurarse de que nuestros modelos climáticos son los mejores que podemos concebir, de modo que podemos tomar las acciones apropiadas. A medida que los glaciares se derriten, el paso del Noroeste se abre mientras el hielo del Ártico disminuye y las capas de hielo de la Antártida se están rompiendo y deslizando al océano, no podemos durante más tiempo correr el riesgo de creer que no necesitamos hacer nada y todo se arreglará solo.

Capítulo 11
Ondas en el éter
Ecuaciones de Maxwell

The diagram shows the four Maxwell equations with their physical interpretations:

- $\nabla \cdot \mathbf{E} = 0$: campo eléctrico (electric field) and divergencia (divergence).
- $\nabla \times \mathbf{E} = -\frac{1}{c} \frac{\partial \mathbf{H}}{\partial t}$: campo magnético (magnetic field), bucle (loop), velocidad de la luz (speed of light), and tasa de cambio con respecto al tiempo (rate of change with respect to time).
- $\nabla \cdot \mathbf{H} = 0$: campo magnético (magnetic field).
- $\nabla \times \mathbf{H} = \frac{1}{c} \frac{\partial \mathbf{E}}{\partial t}$: campo eléctrico (electric field).

¿Qué dicen?

La electricidad y el magnetismo no pueden desvanecerse sin más. Una región de un campo eléctrico girando crea un campo magnético perpendicular al giro. Una región de un campo magnético girando crea un campo eléctrico perpendicular al giro, pero en el sentido opuesto.

¿Por qué es importante?

Fue la primera unificación importante de fuerzas físicas, mostrando que la electricidad y el magnetismo están íntimamente interrelacionados.

¿Qué provocó?

La predicción de que las ondas electromagnéticas existen, desplazándose a la velocidad de la luz, de modo que la propia luz es una de dichas ondas. Esto motivó la invención de la radio, el radar, la televisión, las conexiones inalámbricas para los ordenadores y la mayoría de las comunicaciones modernas.

Al comienzo del siglo XIX la mayoría de la gente iluminaba sus casas usando velas y faroles. El alumbrado de gas, que data de 1790, se usaba ocasionalmente en casas y locales de negocios, principalmente por inventores y empresarios. El alumbrado de gas en las calles empezó a usarse en París en 1820. En esa época, el modo

estándar de enviar mensajes era escribir una carta y enviarla en un carro tirado por caballos; para mensajes urgentes, con el caballo sin más, omitiendo el carro. La principal alternativa, generalmente restringida a comunicaciones militares y oficiales, era el telégrafo óptico. Este usaba un semáforo: un aparato mecánico colocado en torres, que podían representar letras o palabras en código colocando brazos rígidos en varios ángulos. Estas configuraciones podían ser vistas a través de un telescopio y transmitidas a la siguiente torre en la secuencia. El primer sistema extenso de este tipo data de 1792, cuando el ingeniero francés Claude Chappe construyó 556 torres para crear una red de 4.800 kilómetros a través de casi toda Francia. Estuvo en uso durante sesenta años.

Pasado un centenar de años, las casas y las calles tenían alumbrado eléctrico, el telégrafo eléctrico había venido y se había ido, y la gente podía hablar la una con la otra por teléfono. Los físicos habían demostrado las comunicaciones por radio en sus laboratorios, y un empresario ya había montado una fábrica que vendía radios al público. Dos científicos hicieron el principal descubrimiento que desencadenó esta revolución social y tecnológica. Uno fue el inglés Michael Faraday, que estableció la física básica del electromagnetismo, una combinación que establecía un vínculo fuerte entre los fenómenos, previamente separados, de la electricidad y el magnetismo. El otro fue el escocés James Clerk Maxwell, quien convirtió las teorías mecánicas de Faraday en ecuaciones matemáticas y las usó para predecir la existencia de radiofrecuencias desplazándose a la velocidad de la luz.

La Royal Institution en Londres es un edificio imponente, liderado por columnas clásicas, escondido en una calle lateral cerca de Piccadilly Circus. Hoy su actividad principal es albergar eventos de divulgación de ciencia para el público, pero cuando se fundó en 1799, su competencia también incluía «difundir el conocimiento y facilitar la introducción general de invenciones mecánicas útiles». Cuando John «Mad Jack» Fuller creó una cátedra de Química en la Royal Institution, su primer titular no fue un académico. Fue el hijo de un aspirante a herrero, que se había formado como aprendiz de librero. El puesto le permitió leer vorazmente, a pesar de la escasez de dinero de su familia, y *Conversations on Chemistry* (Conversaciones sobre Química) de Jane Marcet y *The Improvement of the Mind* (La mejora de la mente) de Isaac Watts inspiraron un profundo interés en la ciencia en general y en

la electricidad en particular.

El joven era Michael Faraday. Había asistido a clases en la Royal Institution impartidas por el eminente químico Humphry Davy, y envió al profesor 300 páginas de apuntes. Poco después, Davy tuvo un accidente que dañó su vista y pidió a Faraday que fuese su secretario. Luego un asistente en la Royal Institution fue despedido, y Davy sugirió a Faraday como su sustituto, poniéndolo a trabajar en la química del cloro.

La Royal Institution permitió a Faraday dedicarse a sus propios intereses científicos también, y este llevó a cabo innumerables experimentos sobre un tema que se había descubierto recientemente, la electricidad. En 1821 aprendió del trabajo del científico danés Hans Christian Ørsted, vinculando la electricidad a un fenómeno mucho más antiguo, el magnetismo. Faraday explotó este vínculo para inventar un motor eléctrico, pero Davy se enfadó cuando no le concedió ningún crédito, y mandó a Faraday a trabajar en otras cosas. Davy murió en 1831, y dos años más tarde Faraday empezó una serie de experimentos en electricidad y magnetismo que sellaron su reputación como uno de los más grandes científicos de todos los tiempos. Sus investigaciones exhaustivas estaban parcialmente motivadas por la necesidad de proponer grandes números de experimentos novedosos para instruir al ciudadano de a pie y entretenér a la *crème de la crème*, como parte de la competencia de la Royal Institution de alentar al público a comprender la ciencia.

Entre los inventos de Faraday había métodos para convertir la electricidad en magnetismo y ambos en movimiento (un motor) y para convertir movimiento en electricidad (un generador). De esto sacó provecho su gran descubrimiento, la inducción electromagnética. Si el material que puede conducir electricidad se mueve a través de un campo magnético, una corriente eléctrica fluirá a través de él. Faraday descubrió esto en 1831. Francesco Zantedeschi ya se había percatado del efecto en 1829, y Joseph Henry también lo descubrió un poco más tarde. Pero Henry se retrasó en publicar su descubrimiento y Faraday llevó la idea mucho más lejos de lo que Zantedeschi había hecho. El trabajo de Faraday fue mucho más allá de la competencia de la Royal Institution de facilitar invenciones mecánicas útiles, creando máquinas innovadoras que explotaran fronteras en física. Esto llevó, bastante directamente, a la energía eléctrica, la luz y otros miles de artilugios.

Cuando otros tomaron el relevo, toda la colección de equipamiento electrónico y eléctrico moderno irrumpió en escena, empezando con la radio, siguiendo con la televisión, radar y comunicaciones a larga distancia. Fue Faraday, más que cualquier otro individuo solo, quien creó el mundo de la tecnología moderna, con la ayuda de nuevas ideas vitales de centenares de ingenieros, científicos y hombres de negocios de talento.

Al pertenecer a la clase trabajadora y carecer de la educación normal de un caballero, Faraday aprendió por sí mismo ciencia, pero no matemáticas. Desarrolló sus propias teorías para explicar y guiar sus experimentos, pero dependía de analogías mecánicas y máquinas conceptuales, no de fórmulas y ecuaciones. Su trabajo ocupó el lugar que se merecía en la física básica gracias a la intervención de uno de los mayores intelectos científicos de Escocia, James Clerk Maxwell.

Maxwell nació el mismo año en que Faraday anunció el descubrimiento de la inducción electromagnética. Una aplicación, el telégrafo electromagnético, le siguió rápidamente, gracias a Gauss y su asistente Wilhelm Weber. Gauss quería usar cables para llevar señales eléctricas entre el observatorio de Gotinga, donde vivía, y el Instituto de Física, a un kilómetro de distancia, donde Weber trabajaba. Proféticamente, Gauss simplificó la técnica anterior que distinguía las letras del alfabeto —un cable por letra— introduciendo un código binario usando corrientes positivas y negativas (véase el capítulo 15). En 1839, la compañía Great Western Railway estaba enviando mensajes por telégrafo desde Paddington a West Drayton, una distancia de 21 kilómetros. En el mismo año, Samuel Morse, independientemente inventó su propio telégrafo eléctrico en EE.UU., empleando el código Morse (inventado por su asistente Alfred Vail) y enviando su primer mensaje en 1838.

En 1876, tres años antes de que Maxwell muriese, Alexander Graham Bell sacó la primera patente de un nuevo aparato, el telégrafo acústico. Era un artilugio que convertía el sonido, especialmente el habla, en impulsos eléctricos y los transmitía por un cable a un receptor, que los volvía a convertir en sonido. Ahora lo conocemos como el teléfono. No fue la primera persona en concebir tal cosa, ni siquiera en construirla, pero sí quien tuvo la primera patente. Thomas Edison mejoró el diseño con su micrófono de carbón en 1878. Un año más tarde, Edison

desarrolló la bombilla con filamentos de carbono y se consolidó como el inventor de la luz eléctrica en la sabiduría popular. En honor a la verdad, fue precedido por al menos 23 inventores, el más conocido es Joseph Swan, que había patentado su versión en 1878. En 1880, un año después de la muerte de Maxwell, la ciudad de Wabash, Illinois, se convirtió en la primera en usar alumbrado eléctrico en sus calles.

Estas revoluciones en la comunicación y la luz deben mucho a Faraday, la generación de energía eléctrica también debe mucho a Maxwell. Pero el legado de mayor alcance de Maxwell fue hacer que el teléfono pareciese un juguete infantil. Y esto fue producto, directa e inevitablemente, de sus ecuaciones para el electromagnetismo.

Maxwell nació en una familia con talento, pero excéntrica, de Edimburgo, que incluía abogados, jueces, músicos, políticos, poetas, especuladores de la minería y hombres de negocios. Cuando era adolescente empezó a sucumbir a los encantos de las matemáticas, ganando una competición escolar con un trabajo sobre cómo construir óvalos usando clavos e hilo. A los dieciséis, fue a la Universidad de Edimburgo, donde estudió matemáticas y experimentó con la química, el magnetismo y la óptica. Publicó artículos en matemática pura y aplicada en la revista de la Royal Society of Edinburgh. En 1850 su carrera matemática dio un giro más serio y se trasladó a la Universidad de Cambridge, donde fue preparado personalmente por William Hopkins para los *tripos* de matemáticas. Los *tripos* de la época consistían en resolver complicados problemas, que con frecuencia implicaban trucos inteligentes y cálculos extensos, contra reloj. Más tarde Godfrey Harold Hardy, uno de los mejores matemáticos de Inglaterra y catedrático de Cambridge, tendría una importante visión de cómo hacer matemáticas creativas, e hincar los codos por un examen peliagudo no lo era. En 1926, comentó que su objetivo no era «reformar los *tripos*, sino destruirlos». Pero Maxwell hincó los codos y prosperó en la competitiva atmósfera, probablemente porque tenía ese tipo de mente.

También continuó sus experimentos extraños, entre otras cosas tratar de averiguar cómo un gato siempre cae de pie, incluso cuando está sujeto patas arriba solo unos pocos centímetros sobre la cama. La dificultad es que esto parece violar la mecánica newtoniana; el gato tiene que rotar 180 grados, pero no tiene nada contra lo que

empujarse. El mecanismo exacto se le escapaba y no se averiguó hasta que el doctor francés Jules Marey hizo una serie de fotografías de un gato cayendo en 1894. El secreto es que el gato no es rígido; retuerce su parte delantera y trasera en sentidos opuestos y la trasera de nuevo, mientras extiende y contrae sus patas para evitar que estos movimientos se contrarresten.³¹

Maxwell obtuvo su licenciatura en Matemáticas y continuó como posgraduado en el Trinity College. Ahí leyó *Experimental Researches* (Investigaciones experimentales) de Faraday y trabajó sobre la electricidad y el magnetismo. Aceptó una cátedra de Filosofía Natural en Aberdeen, investigando los anillos de Saturno y la dinámica de las moléculas en los gases. En 1860 se trasladó al King's College de Londres y aquí podría haberse visto con Faraday algunas veces. Ahora Maxwell emprendía su búsqueda más influyente: formular unas bases matemáticas para las teorías y experimentos de Faraday.

En la época, la mayoría de los físicos que trabajaban sobre electricidad y magnetismo estaban buscando analogías con la gravedad. Parecía sensato: cargas eléctricas opuestas que se atraen la una a la otra con una fuerza que, como la gravedad, es proporcional al cuadrado de la inversa de la distancia que los separa. Como las cargas se repelen la una a la otra con una fuerza variante similar, y lo mismo aplica para el magnetismo, donde las cargas son remplazadas por polos magnéticos. El modo estándar de pensar era que la gravedad era una fuerza a través de la cual un cuerpo actuaba misteriosamente sobre otro cuerpo lejano, sin que nada pasase entre ellos; se asumía que la electricidad y el magnetismo actuaban de la misma manera. Faraday tuvo una idea diferente: ambos son «campos», fenómenos que llenan el espacio y pueden detectarse por las fuerzas que producen.

¿Qué es un campo? Maxwell pudo hacer progresos pequeños hasta que pudo describir el concepto matemáticamente. Pero Faraday, que carecería de formación matemática, había planteado sus teorías en términos de estructuras geométricas, tales como «líneas de fuerza» a lo largo de las cuales los campos tiran y empujan. El primer gran avance de Maxwell fue reformular estas ideas por analogía con las

³¹ Donald McDonald. «How does a cat fall on its feet?», *New Scientist* 7, n.º 189 (1960) 1647-1649. Véase también: http://en.wikipedia.org/wiki/Cat_righting_reflex

matemáticas del flujo de fluidos, donde el campo a todos los efectos es el fluido. Las líneas de fuerza eran entonces análogas a las rutas seguidas por las moléculas del fluido; la fuerza del campo eléctrico o magnético era análoga a la velocidad del fluido. De modo informal, un campo era un fluido invisible, matemáticamente se comportaba exactamente como eso, fuera lo que fuera realmente. Maxwell tomó prestadas ideas de las matemáticas de fluidos y las modificó para describir el magnetismo. Su modelo explicaba las propiedades principales observadas en la electricidad.

No contento con su intento inicial, continuó para incluir no solo el magnetismo, sino su relación con la electricidad. Cuando el fluido eléctrico fluía, esto afectaba al magnético y viceversa. Para campos magnéticos Maxwell usó la imagen mental de vórtices minúsculos girando en el espacio. Los campos eléctricos estaban, de manera similar, compuestos de minúsculas esferas cargadas. Siguiendo esta analogía y las matemáticas que resultaban, Maxwell empezó a entender cómo un cambio en la fuerza eléctrica podía crear un campo magnético. A medida que las esferas de la electricidad se mueven, provocan que los vórtices magnéticos giren, como un aficionado al fútbol pasando por un torniquete. El aficionado se mueve sin girar; el torniquete gira sin moverse.

Maxwell no estaba satisfecho del todo con esta analogía y dijo: «Yo no lo presento ... como un modo de conexión existente en la naturaleza ... Es, sin embargo ... concebible mecánicamente y fácilmente investigable, y sirve para enfatizar las conexiones mecánicas reales entre el fenómeno electromagnético conocido». Para mostrar qué quería decir, usó el modelo para explicar por qué cables paralelos con corrientes eléctricas opuestas se repelen, y también explicó el descubrimiento crucial de Faraday de la inducción electromagnética.

El siguiente paso era conservar las matemáticas mientras se deshacía de los artifios mecánicos que impulsaron la analogía. Esto equivalía a escribir las ecuaciones para las interacciones básicas entre los campos eléctrico y magnético, obtenidas a partir del modelo mecánico, pero separadas de su origen. Maxwell logró su objetivo en 1864 en su famoso artículo «A dynamical theory of the electromagnetic field» (Una teoría dinámica del campo electromagnético).

Ahora interpretamos sus ecuaciones usando vectores, que son cantidades que no

solo tienen un tamaño, sino que tienen una dirección. El más familiar es la velocidad: el tamaño es la celeridad, cuán rápido se mueve el objeto; la dirección es la dirección a lo largo de la que se mueve. La dirección sí que importa realmente; un cuerpo moviéndose verticalmente hacia arriba a 10 km/s se comporta de un modo muy diferente a uno que se mueve verticalmente hacia abajo a 10 km/s. Matemáticamente, un vector se representa por sus tres componentes: su efecto a lo largo de tres ejes que son perpendiculares entre ellos, como norte/sur, este/oeste y arriba/abajo. De modo que lo mínimo es un vector que sea un conjunto (x, y, z) compuesto de tres números (figura 44). Por ejemplo, la velocidad de un fluido en un punto dado es un vector. Por el contrario, la presión en un punto dado es un único número; el término técnico usado para distinguirlo de un vector es «escalar».

En estos términos, ¿qué es el campo eléctrico? Desde la perspectiva de Faraday, está determinado por líneas de fuerza eléctrica. En la analogía de Maxwell, estas son líneas de flujo de un fluido eléctrico. Una línea de flujo nos dice en qué dirección está fluyendo el fluido, y ya que una molécula se mueve a lo largo de una línea de flujo, podemos observar

también su celeridad. Por lo tanto, para cada punto en el espacio, la línea de flujo pasando a

través de ese punto determina un vector, que describe la velocidad y dirección del fluido eléctrico, esto es, la fuerza y dirección del campo eléctrico en ese punto. A la inversa, si conocemos estas velocidades y direcciones, para cada punto en el espacio, podemos deducir qué aspecto tiene la línea de flujo, de modo que en principio conocemos el campo eléctrico.

En resumen: el campo eléctrico es un sistema de vectores, uno por cada punto en el espacio. Cada vector prescribe la intensidad y dirección de la fuerza eléctrica (ejecutada sobre una minúscula partícula cargada de prueba) en ese punto. Los matemáticos llaman a dicha cantidad un campo vectorial, es una función que asigna a cada punto en el espacio el vector correspondiente. De modo similar, el campo magnético está determinado por las líneas de fuerza magnéticas; es el campo

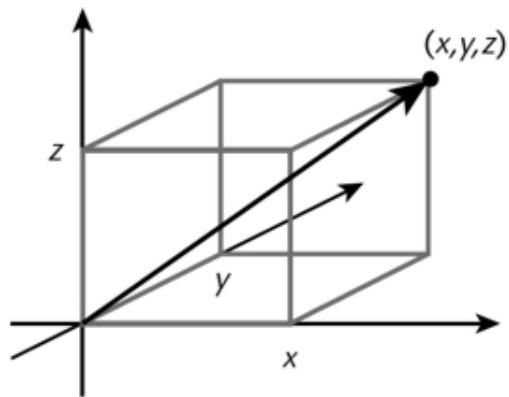


FIGURA 44. Un vector tridimensional.

vectorial correspondiente a las fuerzas que se ejercerían en una minúscula partícula magnética de prueba.

Una vez resuelto qué eran los campos magnéticos y eléctricos, Maxwell podía escribir ecuaciones describiendo qué hacían. Ahora expresamos estas ecuaciones usando dos operadores vectoriales, conocidos como divergencia y rotacional. Maxwell usó fórmulas específicas que envolvían las tres componentes de los campos eléctrico y magnético. En el caso especial en el que no hay alambres conductores ni placas metálicas, ni imanes, y todo sucede en el vacío, la ecuación adopta una forma ligeramente más simple, y restringiré la discusión a este caso.

Dos de las ecuaciones nos dicen que los fluidos eléctricos y magnéticos son incomprimibles, esto es, la electricidad y el magnetismo no pueden evaporarse sin más, tienen que ir a algún sitio. Esto se traslada como «la divergencia es cero», lo que nos lleva a las ecuaciones.

$$\nabla \cdot E = 0 \quad \nabla \cdot H = 0$$

Donde el triángulo del revés y el punto son la notación para la divergencia. Dos ecuaciones más nos dicen que cuando una región de un campo eléctrico gira en un círculo pequeño, crea un campo magnético perpendicular al plano de ese círculo, y de manera similar una región de un campo magnético girando crea un campo eléctrico perpendicular al plano de ese círculo. Hay un giro curioso: los campos eléctrico y magnético apuntan en direcciones opuestas para una dirección dada del giro. Las ecuaciones son:

$$\nabla \times E = -\frac{1}{c} \frac{\partial H}{\partial t} \quad \nabla \times H = \frac{1}{c} \frac{\partial E}{\partial t}$$

Donde ahora el triángulo del revés y el aspa son la notación para el rotacional. El símbolo t es para el tiempo y $\partial/\partial t$ es la tasa de variación con respecto al tiempo. Observa que la primera ecuación tiene un signo menos, pero la segunda no, esto representa las orientaciones opuestas que mencioné.

¿Qué es c ? Es una constante, la proporción de unidades electromagnéticas frente a electrostáticas. Experimentalmente la proporción está justo por debajo de 300.000 en unidades de kilómetros divididas por segundos. Maxwell inmediatamente reconoció este número: es la velocidad de la luz en el vacío. ¿Por qué aparece esa cantidad? Decidió averiguarlo. Una pista, remontándonos a Newton, y desarrollada por otros, fue el descubrimiento de que la luz era algún tipo de onda. Pero nadie sabía en qué consistía la onda.

Un cálculo sencillo proporciona la respuesta. Una vez sabes las ecuaciones para el electromagnetismo, puedes resolverlas para predecir cómo los campos eléctrico y magnético se comportan en diferentes circunstancias. También puedes obtener consecuencias matemáticas generales. Por ejemplo, el segundo par de ecuaciones relaciona E y H ; cualquier matemático inmediatamente tratará de obtener ecuaciones que contengan solo E y solo H , porque eso nos permite concentrarnos en cada campo por separado. Considerando sus consecuencias épicas, esta tarea resulta ser ridículamente sencilla (si estás familiarizado con el cálculo vectorial). He puesto el trabajo detallado en las Notas,³² pero aquí va un resumen rápido. Siguiendo nuestro instinto, empezamos con la tercera ecuación, que relaciona el rotacional de E con la derivada respecto al tiempo de H . No tenemos ninguna otra ecuación que envuelva la derivada respecto al tiempo de H , pero tenemos una que

³² El rotacional de ambos lados de la tercera ecuación da:

$$\nabla \times \nabla \times E = \frac{1}{c} \frac{\partial (\nabla \times H)}{\partial t}$$

El cálculo vectorial nos dice que la parte izquierda de esta ecuación se simplifica a:

$$\nabla \times \nabla \times E = \nabla (\nabla \cdot E) - \nabla^2 E = -\nabla^2 E$$

Donde también usamos la primera ecuación. Aquí ∇^2 es el operador laplaciano. Usando la cuarta ecuación, la parte derecha se convierte en:

$$-\frac{1}{c} \frac{\partial (\nabla \times H)}{\partial t} = -\frac{1}{c} \frac{\partial}{\partial t} \left(\frac{1}{c} \frac{\partial E}{\partial t} \right) = -\frac{1}{c^2} \frac{\partial^2 E}{\partial t^2}$$

Cancelando un signo menos con el otro y multiplicando por c^2 damos con la ecuación de onda para E :

$$\frac{\partial^2 E}{\partial t^2} = c^2 \nabla^2 E$$

Un cálculo similar revela la ecuación de onda para H .

envuelve el rotacional de \mathbf{H} , concretamente, la cuarta ecuación. Esto sugiere que deberíamos tomar la tercera ecuación y la forma rotacional de ambos lados. Entonces aplicamos la cuarta ecuación, simplificamos y aparece:

$$\frac{\partial^2 \mathbf{E}}{\partial t^2} = c^2 \nabla^2 \mathbf{E}$$

¡La ecuación de onda!

El mismo truco aplicado al rotacional de \mathbf{H} produce la misma ecuación con \mathbf{H} en lugar de \mathbf{E} . (El signo menos se aplica dos veces, de modo que desaparece.) Así tanto los campos eléctricos como los magnéticos, en el vacío, obedecen a la ecuación de onda. Como la misma constante c se da en cada ecuación de onda, ambos se desplazan a la misma velocidad, concretamente c . De modo que este pequeño cálculo predice que tanto el campo eléctrico como el magnético pueden simultáneamente sostener una onda, haciéndola una onda electromagnética, en la que los dos campos varían sincronizados. Y la velocidad de esa onda es... la velocidad de la luz.

Es otra de esas preguntas con truco. ¿Qué viaja a la velocidad de la luz? Esta vez la respuesta es lo que esperas: la luz. Pero hay una implicación trascendental: la luz es una onda electromagnética.

Esto son unas noticias estupendas. No hay razón, anterior a la obtención por parte de Maxwell de sus ecuaciones, para imaginar un vínculo tan importante entre la luz, la electricidad y el magnetismo. Pero hay más. La luz llega en muchos colores diferentes y, una vez sabes que la luz es una onda, puedes averiguar que esto se corresponde con ondas con diferentes longitudes de onda (la distancia entre picos sucesivos). La ecuación de onda no impone condiciones sobre la longitud de onda, de modo que puede ser cualquiera. Las longitudes de onda de la luz visible están restringidas a un rango pequeño, a causa de la química de los pigmentos detectores de luz de los ojos. Los físicos ya conocían la «luz invisible», ultravioleta e infrarrojos. Estas, por supuesto, tenían longitudes de onda justo fuera del rango visible. Ahora las ecuaciones de Maxwell llevan a una predicción drástica: deberían existir también ondas electromagnéticas con otras longitudes de onda. Posiblemente

pueda darse cualquier longitud de onda, larga o corta (figura 45).

Nadie había esperado esto, pero tan pronto como la teoría dijo que debía suceder, podían hacerse experimentos y buscarlo. Una de las personas que experimentó sobre ello fue un alemán, Heinrich Hertz. En 1886, construyó un aparato que podía generar radiofrecuencias y otro que podía recibirlas. El transmisor era poco más que una máquina que podía producir una chispa de alto voltaje; la teoría indicaba que dicha chispa emitiría radiofrecuencias. El receptor era un circuito circular de alambre de cobre, cuyo tamaño se escogió para resonar con las ondas entrantes. Un pequeño hueco en el circuito, de unos pocos cientos de milímetros, revelaría esas ondas produciendo chispas minúsculas. En 1887, Hertz hizo el experimento y fue un éxito. Investigó muchas características diferentes de las radiofrecuencias. También midió su velocidad, obteniendo una respuesta cercana a la velocidad de la luz, que confirmó la predicción de Maxwell y confirmó que su aparato realmente estaba detectando ondas electromagnéticas.

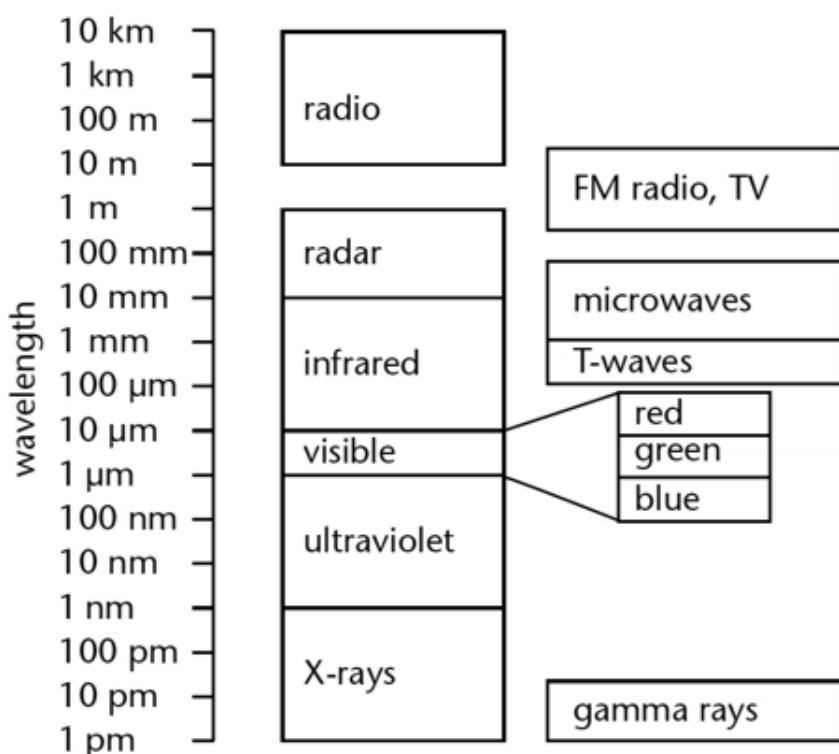


FIGURA 45. El espectro electromagnético.

Hertz sabía que su trabajo era importante para la física y lo publicó en *Electric Waves: being researches on the propagation of electric action with finite velocity through space* (Ondas eléctricas: siendo investigadas sobre la propagación de la acción eléctrica con la velocidad finita a través del espacio). Pero nunca se le ocurrió que la idea podía tener usos prácticos. Cuando se le preguntó, respondió: «Para nada tiene algún tipo de uso ... solo un experimento que prueba que el Maestro Maxwell tenía razón, tan solo tenemos estas misteriosas ondas electromagnéticas que no podemos ver a simple vista. Pero están ahí». Cuando se insistió sobre su visión de las implicaciones, dijo: «Nada, supongo».

¿Fue una falta de imaginación o solo una carencia de interés? Es difícil de decir. Pero el experimento «inútil» de Hertz, confirmando la predicción de Maxwell de la radiación electromagnética, llevaría rápidamente a una invención que hizo que el teléfono pareciese un juguete de niños.

La radio.

La radio hace uso de un rango especialmente fascinante del espectro: ondas con longitud de onda mucho más larga que la luz. Sería posible que dichas ondas conservasen su estructura durante largas distancias. La idea clave, la que Hertz pasó por alto, es simple: si pudiésemos de algún modo imprimir una señal en una onda de ese tipo, podríamos hablarle al mundo.

Otros físicos, ingenieros y empresarios fueron más imaginativos y rápidamente descubrieron el potencial de la radio. Para darse cuenta de ese potencial, sin embargo, tuvieron que resolver unos cuantos problemas técnicos. Necesitaron un transmisor que pudiese producir una señal lo suficientemente potente, y algo para recibirla. El aparato de Hertz estaba restringido a una distancia de unos pocos metros, puedes entender por qué no propuso la comunicación como una posible aplicación. Otro problema fue cómo marcar una señal. Un tercero era hasta dónde se podía enviar la señal, lo que bien podría estar limitado por la curvatura de la Tierra. Si una línea recta entre el transmisor y el receptor golpea el suelo, es de suponer que esto podría bloquear la señal. Más tarde resultó que la naturaleza había sido amable con nosotros, y la ionosfera refleja la radiofrecuencia en un rango amplio de longitudes de onda, pero, de todos modos, antes de que esto se descubriese, había maneras obvias de rodear el problema potencial. Puedes

construir torres altas y poner los transmisores y receptores en ellas. Retransmitiendo señales de una torre a otra, puedes enviar mensajes alrededor del globo muy rápido.

Hay dos maneras relativamente obvias de marcar una señal en una radiofrecuencia. Puedes hacer que la amplitud varíe o puedes hacer que la frecuencia varíe. Estos métodos se llaman amplitud modulada y frecuencia modulada: AM y FM. Ambas se usaron y ambas todavía existen. Eso solucionaba un problema. Antes de 1893, el ingeniero serbio Nikola Tesla había inventado y construido todos los artilugios principales necesarios para la transmisión por radio, y había demostrado sus métodos en público. En 1894, Oliver Lodge y Alexander Muirhead enviaron una señal de radio del laboratorio de Clarendon en Oxford a un auditorio cercano. Un año más tarde el inventor italiano Guglielmo Marconi transmitió señales a través de una distancia de 1,5 kilómetros usando aparatos nuevos que había inventado. El gobierno italiano rechazó financiar más trabajos, de modo que Marconi se trasladó a Inglaterra. Con el apoyo de British Post Office (el Correos británico), pronto mejoró el rango a 16 kilómetros. Experimentos adicionales llevaron a la ley de Marconi: la distancia a través de la cual se pueden enviar señales es aproximadamente proporcional al cuadrado de la altura de la antena que la transmite. Si se hace una torre el doble de alta, la señal va cuatro veces más lejos. Esto era una buena noticia, sugería que transmisiones de largo alcance deberían ser viables. Marconi estableció una estación para transmitir en la Isla de Wight en Reino Unido en 1897, y abrió una fábrica al año siguiente, fabricando lo que llamaba *wirelesses* (inalámbricos). Todavía se le llamaba así en 1952, cuando escuchaba el *Goon Show* y a *Dan Dare* en la *wireless* en mi habitación, pero ya entonces también nos referíamos al aparato como «la radio». Por supuesto, la palabra *wireless* ha vuelto a ponerse de moda, pero ahora es el vínculo entre tu ordenador y el teclado, ratón, módem y router de Internet lo que es *wireless*, inalámbrico, más que el vínculo de tu receptor con un transmisor lejano. Eso todavía es la radio.

Inicialmente Marconi era propietario de las principales patentes de radio, pero las perdió frente a Tesla en 1943 en una batalla judicial. Los avances tecnológicos rápidamente hicieron estas patentes obsoletas. Desde 1906 hasta la década de los cincuenta, la componente electrónica vital de una radio era la válvula de vacío, una

especie de pequeña bombilla, de manera que las radios tenían que ser grandes y voluminosas. El transistor, un aparato mucho más pequeño y más resistente, lo inventó en 1947 en los Laboratorios Bell un equipo de ingenieros en el que estaban William Shockley, Walter Brattain y John Bardeen (véase el capítulo 14). En 1954, los transistores estaban en el mercado, pero la radio ya estaba perdiendo su primacía como un medio de entretenimiento.

En 1953, yo ya había visto el futuro. Fue la coronación de la reina Isabel II, y mi tía en Tonbridge tenía...iun equipo de televisión! Así que nos amontonamos en el coche destalado de mi padre y condujimos 65 kilómetros para ver el evento. Si soy honesto, fue más impresionante gracias a *Bill and Ben the Flowerpot Men* que por la coronación, pero desde ese momento la radio dejó de ser el arquetipo de entretenimiento casero moderno. Pronto también nosotros tuvimos una televisión. Cualquiera que haya crecido con una pantalla de TV plana en color de 48 pulgadas y alta definición y miles de canales estará horrorizado al oír que en esa época la imagen era en blanco y negro y de alrededor de 12 pulgadas, y (en Reino Unido) había exactamente un canal, la BBC. Cuando veíamos «la televisión» realmente quería decir *la televisión*.

El entretenimiento fue solo una aplicación de la radiofrecuencia. Fue también fundamental para el ejército, para las comunicaciones y para otros propósitos. La invención del radar (del inglés *radio detection and ranging*, que significa «detección y medición de distancias por radio») bien podría haber ganado la Segunda Guerra Mundial para los aliados. Este aparato de alto secreto hizo posible detectar aviones, especialmente aviones enemigos, haciendo rebotar señales de radio en ellos y observando las ondas que se reflejaban. El mito urbano de que las zanahorias son buenas para tu vista se originó en una desinformación durante la guerra, intentando hacer que los nazis dejases de preguntarse por qué los británicos se estaban volviendo tan buenos descubriendo bombarderos cuando atacaban. El radar también tiene uso en períodos de paz. Permite a los controladores aéreos determinar dónde están los aviones, para así prevenir colisiones; cuando hay niebla guía a los aviones de pasajeros hasta la pista; avisa a los pilotos de turbulencias inminentes. Los arqueólogos usan georradares para localizar probables ubicaciones de restos de tumbas y estructuras antiguas.

Los rayos X, primero estudiados sistemáticamente por Wilhelm Röntgen en 1875, tienen longitudes de onda mucho más cortas que la luz. Esto los hace más energéticos, de modo que pueden pasar a través de objetos opacos, en particular el cuerpo humano. Los doctores podrían usar rayos X para detectar huesos rotos y otros problemas fisiológicos, y todavía lo hacen, aunque métodos modernos son más sofisticados y someten al paciente a radiaciones mucho menos dañinas. Los escáneres de rayos X pueden ahora recrear imágenes tridimensionales de un cuerpo humano, o parte de él, en un ordenador. Otros tipos de escáner pueden hacer lo mismo usando otras ramas de la física.

Las microondas son maneras eficientes de enviar señales telefónicas y también aparecen en la cocina, en los hornos microondas, un modo rápido de calentar comida. Una de las últimas aplicaciones en aparecer se emplea en la seguridad de los aeropuertos. La radiación terahertz, también conocida como rayos T, puede atravesar la ropa e incluso cavidades del cuerpo. Los agentes de aduanas pueden usarlos para descubrir traficantes de drogas y terroristas. Su uso es un poco controvertido, ya que equivale a un registro exhaustivo electrónico, pero la mayoría de nosotros parece que pensamos que es un precio pequeño que hay que pagar si eso evita que se haga explotar un avión o que la cocaína llegue a las calles. Los rayos T también son útiles para los historiadores de arte, porque pueden desvelar murales cubiertos por capas de yeso. Los fabricantes y transportistas comerciales pueden usar los rayos T para inspeccionar productos sin sacarlos de sus cajas.

El espectro electromagnético es tan versátil, y tan efectivo, que su influencia está relacionada ahora con prácticamente todas las esferas de la actividad humana. Hace posibles cosas que a cualquier generación anterior le parecerían un milagro. Requirió de un gran número de gente de todas las profesiones convertir las posibilidades inherentes en las ecuaciones matemáticas en artilugios reales y sistemas comerciales. Pero nada de esto fue posible hasta que alguien se dio cuenta de que la electricidad y el magnetismo podían unir fuerzas para crear una onda. Toda la colección de comunicaciones modernas, desde la radio y la televisión hasta el radar y vinculaciones de microondas para los teléfonos móviles, fue luego inevitable. Y todo es producto de cuatro ecuaciones y un par de líneas de cálculo vectorial básico.

Las ecuaciones de Maxwell no solo cambiaron el mundo, sino que establecieron uno nuevo.

Capítulo 12
La ley y el desorden
Segunda ley de la termodinámica

$$dS \geq 0$$

¿Qué dice?

La cantidad de desorden en un sistema termodinámico siempre aumenta.

¿Por qué es importante?

Pone límites a cuánto trabajo útil puede extraerse a partir del calor.

¿Qué provocó?

Mejores máquinas de vapor, estimaciones de la eficiencia de energía renovable, el escenario de «la gran congelación», la prueba de que la materia está hecha de átomos, y conexiones paradójicas con la flecha del tiempo.

En mayo de 1959, el físico y novelista C.P. Snow dio una conferencia con el título *The Two Cultures* (Las dos culturas), que provocó una extensa controversia. La respuesta del destacado crítico literario F.R. Leavis fue la típica del otro bando de la discusión, dijo rotundamente que había solo una cultura: la suya. Snow sugería que las ciencias y las humanidades habían perdido contacto la una con la otra, y argumentaba que esto estaba haciendo muy difícil solucionar los problemas del mundo. Vemos lo mismo hoy en día con la negación del cambio climático y los ataques a la evolución. La motivación puede ser diferente, pero las barreras culturales ayudan a que prosperen estos sinsentidos, aunque es la política quien lo maneja.

Snow estaba en particular descontento con lo que veía como los estándares de la

educación en declive, y afirmó:

Un buen número de veces he estado presente en reuniones de gente que, por los estándares de la cultura tradicional, son consideradas eruditos y que han expresado con un entusiasmo considerable su incredulidad sobre el analfabetismo de los científicos. Una o dos veces se me ha provocado y he preguntado a quienes me acompañaban cuántos de ellos podrían explicar la segunda ley de la termodinámica, la ley de la entropía. La respuesta era fría, y también negativa. Aunque estaba preguntando algo que es más o menos el equivalente científico de: «¿has leído algo de Shakespeare?».

Quizá sentía que estaba pidiendo demasiado —muchos científicos cualificados no pueden enunciar la segunda ley de la termodinámica—. Así que más tarde añadió: Ahora creo que incluso aunque hubiese hecho una pregunta más simple, como qué quieras decir con masa, o aceleración, que son el equivalente científico a «¿sabes leer?», no más de una décima parte de los eruditos habrían sentido que yo estaba hablando el mismo idioma. De modo que la gran estructura de la física moderna se construye, y la mayoría de la gente más lista del mundo occidental tiene más o menos la misma capacidad para comprenderlo que la que habrían tenido nuestros antepasados de la época neolítica.

Tomando a Snow al pie de la letra, mi objetivo en este capítulo es sacarnos del Neolítico. La palabra «termodinámica» da una pista: parece querer decir la dinámica del calor. ¿Puede el calor ser dinámico? Sí, el calor puede fluir. Puede moverse de un lugar a otro, de un objeto a otro. Sal fuera en un día de invierno y pronto sentirás frío. Fourier había escrito el primer modelo serio del flujo del calor (capítulo 9), e hizo algunas matemáticas bellas. Pero la principal razón por la que los científicos se comenzaron a interesar por el flujo del calor fue un objeto tecnológico modernísimo y muy rentable: la máquina de vapor.

Hay una historia de James Watt de niño repetida con frecuencia; sentado en la cocina de su madre viendo cómo el vapor hacía subir la tapa de una tetera, y su repentino golpe de inspiración: el vapor puede realizar trabajo. De modo que cuando creció, inventó la máquina de vapor. Es material inspirador, pero como muchas de estas historias, es solo palabrería. Watt no inventó la máquina de vapor y no aprendió acerca del poder del vapor hasta que fue un adulto. La conclusión de

la historia sobre el poder del vapor es cierta, pero incluso en la época de Watt, estaba ya muy visto.

Alrededor del 15 a.C., el arquitecto e ingeniero romano Vitruvio describió una máquina llamada eolípila en su *De Architectura*, y el matemático e ingeniero griego Herón de Alejandría construyó una un siglo más tarde. Era una esfera hueca con algo de agua dentro, y dos tubos sobresaliendo, curvados en un ángulo como en la figura 46. Calienta la esfera y el agua se convierte en vapor, se escapa a través de los extremos de los tubos y la reacción hace a la esfera girar. Fue la primera máquina a vapor y demostró que el vapor podía hacer un trabajo, pero Herón no hizo nada con ello, más allá de entretenerte a la gente. Hizo una máquina parecida usando aire caliente en una cámara cerrada para tirar de una cuerda que abría las puertas de un templo. Esta máquina tuvo una aplicación práctica, produciendo un milagro religioso, pero no era una máquina de vapor.

Watt aprendió que el vapor podía ser una fuente de fuerza en 1762 cuando tenía veintiséis años. No lo descubrió observando una tetera; su amigo John Robison, un profesor de filosofía natural en la Universidad de Edimburgo, le habló sobre ello. Pero el poder del vapor práctico era mucho más viejo. Su descubrimiento se atribuye con frecuencia al ingeniero y arquitecto italiano Giovanni Branca, cuya *Le Machine* (La máquina) de 1629 contenía 63 grabados de madera de aparatos mecánicos. Uno muestra una rueda con pedales que giraría sobre su propio eje cuando el vapor de una tubería chocase con sus paletas. Branca hizo conjeturas sobre lo útil que podría ser esta máquina para moler harina, subir agua y cortar madera en pedazos, aunque probablemente nunca se construyese. Era más un experimento mental, un sueño mecánico imposible como la máquina voladora de



FIGURA 46. La eolípila de Herón.

Leonardo da Vinci.

En cualquier caso, a Branca se le había anticipado Taqi al-Din Muhammad ibn Ma'ruf al-Shami al-Asadi, quien vivió alrededor de 1550 en el Imperio otomano y es reconocido ampliamente como el mayor científico de su época. Sus logros son impresionantes. Trabajó en todo, desde la astrología a la zoología, incluyendo relojería, medicina, filosofía y teología, y escribió más de 90 libros. En su *Al-turuq al-samiyya fi al-alat al-ruhaniyya* (Los métodos sublimes de las máquinas espirituales) de 1551, al-Din describió una turbina de vapor primitiva, diciendo que podría usarse para girar carne asada en un asador.

La primera máquina de vapor verdaderamente práctica fue una bomba de agua inventada por Thomas Savery en 1698. La primera en obtener beneficios comerciales, construida por Thomas Newcomen en 1712, disparó la Revolución Industrial. Pero la máquina de Newcomen era muy poco eficiente. La contribución de Watt fue introducir un condensador separado para el vapor, reduciendo la pérdida de calor. Desarrollada con dinero proporcionado por el emprendedor Matthew Boulton, este nuevo tipo de máquina solo usaba una cuarta parte de carbón, lo que suponía un ahorro enorme. La máquina de Boulton y Watt comenzó a fabricarse en 1775, más de 220 años después del libro de al-Din. En 1776, tres estaban listas y funcionando: una en una mina de carbón en Tipton, una en una siderurgia en Shropshire y otra en Londres.

Las máquinas de vapor realizaban varias tareas industriales, pero con mucha diferencia la más común era bombear agua de las minas. Costaba mucho dinero crear una mina, pero a medida que las capas altas se quedaban sin trabajo y, los operadores estaban forzados a cavar más profundo en la tierra y llegaban al nivel freático. Merecía la pena gastar mucho dinero en bombear el agua fuera, ya que la alternativa era cerrar la mina y empezar de nuevo en otro lugar, y eso podría no ser ni siquiera factible. Pero nadie quería pagar más de lo que tenía que pagar, así que los fabricantes que pudiesen diseñar y construir una máquina de vapor más eficiente monopolizarían el mercado. De modo que la pregunta básica de cómo de eficiente podría ser una máquina de vapor reclamaba a gritos atención. Su respuesta hizo más que describir los límites de las máquinas de vapor; creó una rama nueva de la física, cuyas aplicaciones casi no tenían límites. La nueva física

arrojó luz, sobre todo, desde los gases hasta la estructura de todo el universo; se aplicó no solo a la materia inanimada de la física y la química, sino también a los procesos complejos de la propia vida. Se llamó termodinámica: el movimiento del calor. Y, al igual que la ley de la conservación de la energía en mecánica descartó las máquinas mecánicas en perpetuo movimiento, las leyes de la termodinámica descartaron máquinas similares usando calor.

Una de esas leyes, la primera ley de la termodinámica, revela una nueva forma de energía asociada al calor, y extiende la ley de conservación de la energía (capítulo 3) en el nuevo reino de los motores térmicos. Otra, sin ningún precedente previo, muestra que algunas maneras potenciales de intercambio de calor, las cuales no entran en conflicto con la conservación de la energía, eran no obstante imposibles porque tendrían que crear orden a partir del desorden. Esto era la segunda ley de la termodinámica.

Termodinámica es la física matemática de los gases. Explica cómo características a gran escala, como la temperatura y la presión, surgen a raíz del modo en que las moléculas de un gas interactúan. El tema empieza con una serie de leyes de la naturaleza relacionadas con la temperatura, la presión y el volumen. Esta versión se llama termodinámica clásica y no hay moléculas involucradas, en esa época pocos científicos creían en ellas. Más tarde, las leyes de los gases se consolidaron añadiendo otra capa a la explicación, basada en un modelo matemático simple que involucraba moléculas de modo explícito. Las moléculas de los gases eran imaginadas como esferas minúsculas que rebocaban unas contra otras como bolas de billar totalmente elásticas, sin perder energía en la colisión. Aunque las moléculas no son esféricas, este modelo resultaba ser notablemente efectivo. Se llama la teoría cinética de los gases, y condujo a la prueba experimental de que las moléculas existían.

Estas primeras leyes de los gases surgieron a rachas durante un período de cerca de cincuenta años, y se atribuyen principalmente al físico y químico irlandés Robert Boyle, al matemático y pionero en globos francés Jacques Alexandre César Charles, y al físico y químico francés Joseph Louis Gay-Lussac. Sin embargo, muchos de los descubrimientos fueron hechos por otros. En 1834, el ingeniero y físico francés Émile Clapeyron combinó todas estas leyes en una, la ley de los gases ideales, que

ahora escribimos como:

$$pV = RT$$

Aquí p es la presión, V es el volumen, T es la temperatura y R es una constante. La ecuación afirma que la presión por el volumen es proporcional a la temperatura. Supuso mucho trabajo con muchos gases diferentes para confirmar, experimentalmente, cada ley por separado, y la síntesis global de Clapeyron. La palabra «ideal» aparece porque los gases reales no obedecen la ley en todas las circunstancias, especialmente a presiones altas donde las fuerzas interatómicas entran en juego. Pero la versión ideal era lo suficientemente buena para diseñar máquinas de vapor.

La termodinámica está condensada en un número de leyes más generales, que no dependen de la forma exacta de la ley del gas. Sin embargo, sí se necesita que haya algo de dicha ley, porque la temperatura, la presión y el volumen no son independientes. Tiene que haber alguna relación entre ellos, pero no importa mucho cuál.

La primera ley de la termodinámica surge de la ley mecánica de la conservación de la energía. En el capítulo 3, vimos que hay dos tipos distintos de energía en mecánica clásica: la energía cinética, determinada por la masa y la velocidad, y la energía potencial, determinada por el efecto de fuerzas como la gravedad. Ninguno de estos tipos de energía se conserva por sí solo. Si dejas caer un balón, la velocidad sube, de ese modo gana energía cinética. También cae, perdiendo energía potencial. La segunda ley del movimiento de Newton, implica que estos dos cambios se contrarrestan el uno con el otro de manera exacta, de modo que la energía total no cambia durante el movimiento.

No obstante, esta no es la historia completa. Si pones un libro en una mesa y le das un empujón, su energía potencial no cambia siempre que la mesa esté en horizontal. Pero su velocidad sí que cambia, después de un incremento inicial producido por la fuerza con la que lo empujas, el libro rápidamente se ralentiza y acaba en reposo. De modo que la energía cinética empieza en un valor inicial distinto de cero justo después del empujón, y luego desciende a cero. La energía

total, por lo tanto, también decrece, de manera que no se conserva la energía. ¿Dónde se ha ido? ¿Por qué el libro se para? Según la primera ley de Newton, el libro debería continuar moviéndose, a menos que alguna fuerza se oponga. La fuerza es la fricción entre el libro y la mesa. Pero ¿qué es la fricción?

La fricción ocurre cuando superficies rugosas se frotan la una contra la otra. La superficie rugosa del libro tiene pedacitos que sobresalen ligeramente. Estos entran en contacto con partes de la mesa que también sobresalen ligeramente. El libro se roza con la mesa y la mesa, obedeciendo la tercera ley de Newton, se resiste. Esto crea una fuerza que se opone al movimiento del libro, de modo que disminuye su velocidad y pierde energía. Así que, ¿adónde se va la energía? Quizá la conservación simplemente no se aplica. Alternativamente, la energía está todavía merodeando por algún lugar, pasando inadvertida. Y esto es lo que la primera ley de la termodinámica nos dice: la energía desaparecida se presenta como calor. Tanto el libro como la mesa se calientan ligeramente. Los humanos hemos sabido que la fricción crea calor ya desde que algún listillo descubrió cómo frotar dos palos el uno contra el otro para empezar un fuego. Si deslizas tus manos por una cuerda demasiado rápido, te las quemarás a causa de la fricción con la cuerda. Había un montón de indicios. La primera ley de la termodinámica afirma que el calor es una forma de energía, y la energía, con esta ampliación, se conserva en los procesos termodinámicos.

La primera ley de la termodinámica pone límites a lo que puedes hacer con un motor térmico. La cantidad de energía cinética que puedes sacar, en la forma de movimiento, no puede ser más que la cantidad de energía que introduces como calor. Pero resultó que había una restricción adicional en cómo un motor térmico puede convertir energía térmica en energía cinética de modo eficiente; no solo el apunte práctico de que algo de energía siempre se pierde, sino un límite teórico que impide que toda la energía térmica sea convertida en movimiento. Solo alguna de ella, la energía «libre», puede ser convertida. La segunda ley de la termodinámica convierte esta idea en un principio general, pero nos hará falta un rato para llegar a ello. La limitación fue descubierta por Nicolas Léonard Sadi Carnot en 1824, en un modelo simple de cómo funciona una máquina de vapor: el ciclo de Carnot.

Para entender el ciclo de Carnot es importante distinguir entre calor y temperatura.

En la vida cotidiana, decimos que algo está caliente si su temperatura es alta, y así confundimos los dos conceptos. En la termodinámica clásica, ningún concepto es tan sencillo. La temperatura es una propiedad de un fluido, pero el calor solo tiene sentido como una medida de la transferencia de energía entre fluidos, y no es una propiedad intrínseca del estado del fluido (esto es, la temperatura, presión y volumen). En la teoría cinética, la temperatura de un fluido es la energía cinética media de sus moléculas, y la cantidad de calor transferido entre fluidos es el cambio en la energía cinética total de sus moléculas. En cierto sentido, el calor es un poco como la energía potencial, que se define en relación con una altura de referencia arbitraria; esto introduce una constante arbitraria, de modo que «la» energía potencial de un cuerpo no está definida de manera única. Pero cuando el cuerpo cambia de altura, la diferencia en las energías potenciales es la misma sea cual sea la altura de referencia usada, porque la constante la contrarresta. En resumen, la medición del calor cambia, pero la medición de la temperatura se estipula. Las dos están vinculadas; la transferencia de calor es posible solo cuando los fluidos afectados tienen temperaturas diferentes, y entonces se transfiere del más caliente al más frío. Esto es llamado con frecuencia el principio cero de la termodinámica porque lógicamente precede a la primera ley, pero históricamente fue reconocido más tarde.

La temperatura puede medirse usando un termómetro, que se aprovecha de la expansión de un fluido, como el mercurio, causada por el incremento de la temperatura. El calor puede medirse usando su relación con la temperatura. En un fluido de prueba estándar, como el agua, cada grado que aumenta la temperatura de un gramo de fluido se corresponde con un incremento fijo en el contenido de calor. Esta cantidad es llamada el calor específico del fluido, que en el agua es una caloría por gramo por grado Celsius. Observa que el incremento de calor es un cambio, no un estado, como exige la definición de calor.

Podemos visualizar el ciclo de Carnot pensando en una cámara que contiene gas, con un émbolo móvil en un extremo. El ciclo consta de cuatro pasos:

1. Calienta el gas tan rápidamente que su temperatura no cambie. Se expande, realizando trabajo sobre el émbolo.
2. Permite al gas expandirse más, reduciendo la presión. El gas se enfriá.

3. Comprime el gas tan rápidamente que su temperatura no cambie. El émbolo ahora realiza trabajo sobre el gas.
4. Permite al gas expandirse más, incrementando la presión. El gas vuelve a su temperatura original.

En un ciclo de Carnot, el calor introducido en el primer paso transfiere energía cinética al émbolo, permitiendo a este hacer el trabajo. La cantidad de energía transferida puede calcularse en términos de la cantidad de calor introducido y la diferencia de temperatura entre el gas y lo que lo rodea. El teorema de Carnot prueba que, en principio, un ciclo de Carnot es el modo más eficiente de convertir calor en trabajo. Esto pone un límite riguroso sobre la eficiencia de cualquier motor térmico y, en particular, sobre una máquina de vapor.

En un diagrama que muestra la presión y el volumen del gas, un ciclo de Carnot tiene el aspecto de la figura 47 (izquierda). El físico y matemático alemán Rudolf Clausius descubrió una manera más simple de visualizar el ciclo, figura 47 (derecha). Ahora los ejes son la temperatura y una cantidad nueva y fundamental llamada entropía. Con estas coordenadas, el ciclo se hace un rectángulo y la cantidad de trabajo realizado es justo el área del rectángulo.

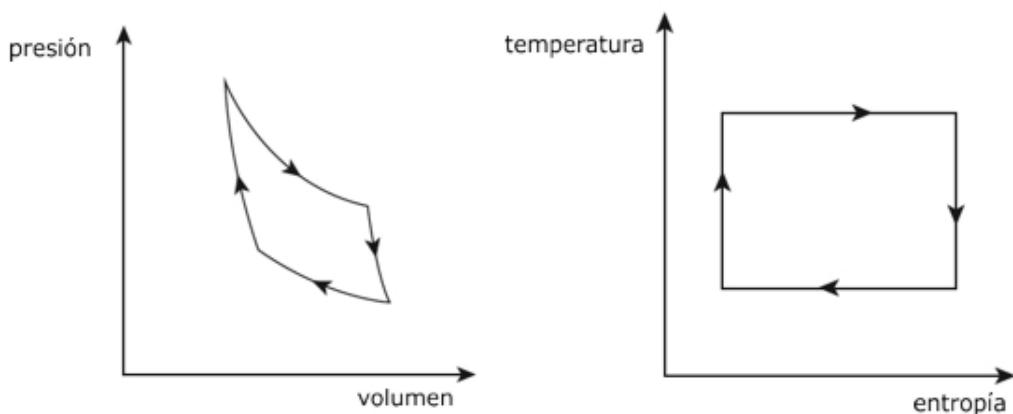


FIGURA 47. El ciclo de Carnot. A la izquierda en términos de presión y volumen. A la derecha en términos de temperatura y entropía.

La entropía es como el calor: está definida en términos de un cambio de estado, no un estado como tal. Supón que un fluido en algún estado inicial cambia a un nuevo

estado. Entonces la diferencia de entropía entre los dos estados es el cambio total en la cantidad «calor dividido entre temperatura». En símbolos, para un pequeño paso a lo largo de un camino entre los dos estados, la entropía S está relacionada con el calor Q y la temperatura T por la ecuación diferencial $dS = dQ/T$. El cambio en la entropía es el cambio en el calor por unidad de temperatura. Un cambio grande de estado puede representarse como una serie de pequeños, de modo que sumamos todos estos cambios pequeños en la entropía para obtener el cambio total de entropía. El cálculo nos dice que el modo de hacer esto es usar una integral.³³

Una vez definida la entropía, la segunda ley de la termodinámica es muy simple. Afirma que en cualquier proceso termodinámico físicamente factible, la entropía de un sistema aislado debe siempre aumentar.³⁴ En símbolos, $dS \geq 0$. Por ejemplo, supón que dividimos una habitación con una mampara móvil, ponemos oxígeno en un lado de la mampara y nitrógeno en el otro. Cada gas tiene una entropía concreta, relacionada con algún estado de referencia inicial. Ahora elimina la mampara, permitiendo a los dos mezclarse. El sistema combinado también tiene una entropía concreta, relacionada con los mismos estados de referencia iniciales. Y la entropía del sistema combinado es siempre mayor que la suma de las entropías de los dos gases por separado.

La termodinámica clásica es fenomenológica: describe lo que puedes medir, pero no está basada en ninguna teoría coherente del proceso implicado. Ese fue el siguiente paso en la teoría cinética de los gases, promovido por Daniel Bernoulli en 1738. Esta teoría proporciona una explicación física de la presión, la temperatura, las leyes de los gases y esa cantidad misteriosa de la entropía. La idea básica, muy polémica en su época, es que un gas consiste en un gran número de moléculas idénticas, que van dando tumbos por el espacio y ocasionalmente chocan unas con otras. Ser un gas significa que las moléculas no están muy apretujadas, de modo

³³ En concreto,

$$S_B - S_A = \int_A^B \frac{dQ}{T}$$

Donde S_A y S_B son las entropías en los estados A y B .

³⁴ La segunda ley de la termodinámica es técnicamente una desigualdad, no una ecuación. He incluido la segunda ley en este libro porque su posición central en la ciencia demandaba su inclusión. Es, sin lugar a dudas, una fórmula matemática, una interpretación amplia de «ecuación» que se extiende más allá de la literatura científica técnica. La fórmula a la que se hace alusión en la nota 1 de este capítulo, usando una integral, es una ecuación genuina. Define el cambio de entropía, pero la segunda ley nos dice cuál es su característica más importante.

que cualquier molécula dada pasa mucho de su tiempo viajando a través del vacío a una velocidad constante en línea recta. (Digo «vacío» incluso aunque estemos hablando de un gas, porque esto es en lo que consiste el espacio entre moléculas.) Como las moléculas, aunque sean muy pequeñas, tienen un tamaño distinto de cero, de vez en cuando dos de ellas colisionan. La teoría cinética hace la suposición simplificadora de que rebotan como dos bolas de billar que chocan, y que estas bolas son totalmente elásticas, de manera que no se pierde ninguna energía en la colisión. Entre otras cosas, esto implica que las moléculas se mantienen rebotando para siempre.

Cuando Bernoulli propuso el modelo por primera vez, la ley de la conservación de la energía no estaba establecida y la elasticidad parecía poco probable. La teoría gradualmente ganó apoyo de un pequeño número de científicos, que desarrollaron sus propias versiones y añadieron varias ideas nuevas, pero su trabajo fue ignorado casi universalmente. El químico y físico alemán August Krönig escribió un libro sobre el tema en 1856, simplificando la física al no permitir a las moléculas rotar. Clausius eliminó esta simplificación un año más tarde. Afirmó que había llegado a sus resultados independientemente, y ahora está considerado como uno de los primeros fundadores significativos de la teoría cinética. Propuso uno de los conceptos clave de la teoría, el camino libre medio de una molécula: con qué rapidez se desplaza, de media, entre colisiones sucesivas.

Tanto Krönig como Clausius dedujeron la ley de los gases ideales de la teoría cinética. Las tres variables clave son volumen, presión y temperatura. El volumen está determinado por el envase que contiene al gas, establece las «condiciones de frontera» que afecta a cómo el gas se comporta, pero no es una característica del gas como tal. La presión es la fuerza media (por unidad cuadrada de área) ejercida por las moléculas del gas cuando colisionan con las paredes del envase. Esto depende de cuántas moléculas están dentro del envase y cómo de rápido se mueven. (No se mueven todas a la misma velocidad.) Más interesante es la temperatura. Esta también depende de la rapidez con que se estén moviendo las moléculas del gas, y es proporcional a la energía cinética media de las moléculas. Deducir la ley de Boyle, el caso especial de la ley de los gases ideales para una temperatura constante, es especialmente sencillo. Con una temperatura fija, la

distribución de las velocidades no cambia, de modo que la presión está determinada por cuántas moléculas golpean la pared. Si reduces el volumen, el número de moléculas por unidad cúbica de espacio sube, y la posibilidad de que cualquier molécula golpee la pared aumenta también. Un volumen más pequeño quiere decir un gas más denso, que quiere decir más moléculas golpeando la pared, y este argumento puede hacerse cuantitativo. Argumentos similares pero más complicados dan lugar a la ley de los gases ideales en toda su gloria, siempre y cuando las moléculas no se aplasten unas contra otras con demasiada fuerza. De modo que ahora había unas bases teóricas más profundas para la ley de Boyle, basada en la teoría de moléculas.

A Maxwell le inspiró el trabajo de Clausius, y en 1859 puso la teoría cinética sobre fundamentos matemáticos escribiendo una fórmula para la probabilidad de que una molécula se desplazase a una velocidad dada. Se basa en la distribución normal o campana de Gauss (capítulo 7). La fórmula de Maxwell parece haber sido el primer ejemplo de una ley física basada en la probabilidad. Le siguió el físico austriaco Ludwig Boltzmann, quien desarrolló la misma fórmula, ahora llamada la distribución de Maxwell-Boltzmann. Boltzmann reinterpretó la termodinámica en términos de la teoría cinética de gases, fundando lo que ahora se llama mecánica estadística. En particular, dio con una interpretación nueva de entropía, relacionando el concepto de termodinámica con una característica estadística de las moléculas en el gas.

Todas las cantidades termodinámicas tradicionales, como la temperatura, presión, calor y entropía, se refieren a propiedades medias a gran escala del gas. Sin embargo, la estructura menuda consiste en muchas moléculas que pasan zumbando por todas partes y chocan unas contra otras. El mismo estado a gran escala puede surgir de innumerables estados diferentes a pequeña escala, debido a que las diferencias menores en la escala pequeña se compensan con la media.

Boltzmann, por lo tanto, distinguía macroestados de microestados del sistema: promedios a gran escala y el estado real de las moléculas. Usando esto, mostró que la entropía, un macroestado, puede interpretarse como una característica estadística de microestados. Lo expresó en la ecuación:

$$S = k \log W$$

Donde S es la entropía del sistema, W es el número de microestados distintos que pueden dar lugar al macroestado total, y k es una constante. Ahora se llama constante de Boltzmann, y su valor es $1,38 \times 10^{-23}$ julios por grado Kelvin.

Es esta fórmula la que motiva la interpretación de la entropía como desorden. La idea es que menos microestados se corresponden con un macroestado más ordenado que con uno desordenado, y podemos comprender por qué si pensamos en barajas de cartas. Para simplificar, supón que tenemos solo seis cartas marcadas con 2, 3, 4, J, Q, K. Ponlas en dos montones separados, con las cartas de valor bajo en un montón y las figuras en el otro. Esto es una disposición ordenada. De hecho, mantiene restos de un orden si barajás cada montón, pero mantienes los montones por separado, porque aunque barajes, las cartas de valor bajo están en un montón y las figuras están en el otro. Sin embargo, si barajás los dos montones juntos, los dos tipos de cartas pueden mezclarse, con disposiciones como 4QK2J3. Intuitivamente, estas disposiciones revueltas son más desordenadas.

Veamos cómo se relaciona esto con la fórmula de Boltzmann. Hay 36 modos de colocar las cartas en dos montones, seis para cada montón. Pero hay 720 modos ($6! = 1 \times 2 \times 3 \times 4 \times 5 \times 6$) de colocar las 6 cartas. El tipo de orden de las cartas que permitimos (dos montones o uno) es análogo al macroestado de un sistema termodinámico. El orden exacto es el microestado. El macroestado más ordenado tiene 36 microestados, el menos ordenado tiene 720. Así que cuantos más microestados haya, menos ordenado pasa a estar el macroestado correspondiente. Ya que cuanto mayor es un número, mayor es su logaritmo, cuanto mayor es el logaritmo del número de microestados, más desordenado está el macroestado. En este caso:

$$\log 36 = 3,58$$

$$\log 720 = 6,58$$

Estas son realmente las entropías de los dos macroestados. La constante de Boltzmann solo escala los valores para adecuarlos al formalismo de la

termodinámica cuando estamos tratando con gases.

Los dos montones de cartas son como dos estados termodinámicos que no interactúan, como una caja con un tabique separando dos gases. Las entropías individuales de cada uno son $\log 6$, así que la entropía total es $2 \cdot \log 6$, que es igual a $\log 36$. De modo que el logaritmo hace a la entropía aditiva para sistemas que no interactúan; para obtener la entropía de un sistema combinado (pero que no está interactuando), suma las entropías sueltas. Si ahora permitimos a los sistemas interactuar (eliminamos el tabique) la entropía incrementa a $\log 720$.

Cuantas más cartas hay, más pronunciado se hace este efecto. Divide una baraja francesa estándar de 52 cartas en dos montones, con todas las cartas rojas en un montón y todas las negras en otro. Esta disposición puede darse de $(26!)^2$ modos, lo que es alrededor de $1,63 \times 10^{53}$. Barajando los dos montones, obtenemos 52! microestados, aproximadamente $8,07 \times 10^{67}$. Los logaritmos son 122,53 y 156,36 respectivamente y, de nuevo, el segundo es mayor.

Las ideas de Boltzmann no fueron recibidas con grandes vítores. A un nivel técnico, la termodinámica estaba plagada de asuntos conceptuales difíciles. Uno era el significado exacto de «microestado». La posición y la velocidad de una molécula son variables continuas, capaces de tomar infinidad de valores, pero Boltzmann necesitaba un número finito de microestados para poder contar cuántos había y luego calcular el logaritmo. Así que estas variables tenían que ser «toscas» en cierto modo, dividiendo el continuo de los posibles valores en un número finito de intervalos muy pequeños. Otro asunto, de naturaleza más filosófica, era la flecha del tiempo, un conflicto aparente entre la dinámica reversible del tiempo de los microestados y el tiempo unidireccional de los macroestados, determinado por el incremento de la entropía. Los dos asuntos están relacionados, como veremos en breve.

Sin embargo, el mayor obstáculo para la aceptación de la teoría era la idea de que la materia está hecha de partículas extremadamente pequeñas, los átomos. Este concepto, y la palabra átomo, que significa «indivisible», se remonta a la Grecia Clásica, aunque todavía alrededor de 1900 la mayoría de los físicos no creían que la materia estuviese hecha de átomos. De modo que tampoco creían en las moléculas y una teoría de gases basada en ellas era obviamente un sinsentido. Maxwell,

Boltzmann y otros pioneros de la teoría cinética estaban convencidos de que las moléculas y los átomos eran reales, pero para los escépticos, la teoría atómica era solo un modo conveniente de imaginarse la materia. No se habían observado átomos nunca, de modo que no había evidencias científicas de que existiesen. Las moléculas, combinaciones específicas de átomos, eran igualmente polémicas. Sí, la teoría atómica encajaba con todo tipo de datos experimentales en química, pero no había prueba de que los átomos existiesen.

Una de las cosas que finalmente convenció a la mayoría de objetores fue el uso de la teoría cinética para hacer predicciones sobre el movimiento browniano. Este efecto fue descubierto por un botánico escocés, Robert Brown.³⁵ Fue pionero en el uso del microscopio, descubriendo, entre otras cosas, la existencia de los núcleos de una célula, ahora conocidos por ser el almacén de su información genética. En 1827, Brown estaba viendo a través de su microscopio los granos de polen en un fluido y descubrió partículas incluso más pequeñas que habían sido expulsadas por el polen. Estas partículas diminutas se mueven de un lado a otro de una manera aleatoria, y al principio Brown se preguntó si eran alguna forma diminuta de vida. Sin embargo, sus experimentos mostraron el mismo efecto en las partículas obtenidas de materia no viva, de modo que fuese lo que fuese lo que causaba el movimiento, no tenía que estar vivo. En la época, nadie sabía qué causaba este efecto. Ahora sabemos que las partículas expulsadas por el polen son orgánulos, subsistemas minúsculos de células con funciones específicas, en este caso, para fabricar almidón y grasas. E interpretamos su movimiento aleatorio como la prueba para la teoría de que la materia está hecha de átomos.

El vínculo con los átomos viene de modelos matemáticos del movimiento browniano, que primero aparecieron en un trabajo estadístico del astrónomo y actuaria danés Thorvald Thiele en 1880. El gran avance fue hecho por Einstein en 1905 y el científico polaco Marian Smoluchowski en 1906. De manera independiente propusieron una explicación física para el movimiento browniano: los átomos del fluido en el que las partículas están flotando están aleatoriamente chocando con las partículas y dándoles patadas diminutas. Partiendo de esto, Einstein usó un modelo

³⁵ A Brown se le adelantó el fisiólogo holandés Jan Ingenhousz, quien vio el mismo fenómeno en el polvo de carbón flotando en la superficie del alcohol, pero no propuso ninguna teoría para explicar lo que había visto.

matemático para hacer predicciones cuantitativas sobre la estadística del movimiento, que fueron confirmadas por Jean Baptiste Perrin en 1908-1909.

Boltzmann se suicidó en 1906, justo cuando el mundo científico estaba empezando a apreciar que las bases de su teoría eran reales.

En la formulación de Boltzmann de termodinámica, las moléculas en un gas son análogas a las cartas en una baraja, y la dinámica natural de las moléculas es análoga a barajar. Supongamos que en algún momento todas las moléculas de oxígeno en una habitación se concentran en un extremo, y todas las de nitrógeno en el otro. Esto es un estado termodinámico ordenado, como los dos montones de cartas separados. Sin embargo, después de un período muy corto, colisiones aleatorias mezclarán todas las moléculas, más o menos uniformemente, por toda la habitación, como barajar las cartas. Acabamos de ver que este proceso habitualmente provoca que la entropía se incremente. Esta es la imagen ortodoxa del incremento implacable de la entropía, y es la interpretación estándar de la segunda ley: «la cantidad de desorden en el universo incrementa a ritmo constante». Estoy bastante seguro de que esta caracterización de la segunda ley habría satisfecho a Snow si alguien la hubiese ofrecido. En esta forma, una consecuencia dramática de la segunda ley es el escenario de la «Gran congelación», en el cual todo el universo se acabará convirtiendo en un gas tibio con una estructura para nada interesante.

La entropía, y el formalismo matemático que la acompaña, proporciona un modelo excelente para muchas cosas. Explica por qué los motores térmicos pueden alcanzar solo un nivel de eficiencia concreto, que evita que los ingenieros gasten un tiempo y dinero valiosos buscando resultados que no van a ningún lado. Esto no solo es cierto para las máquinas a vapor de la época victoriana, también se aplica para los motores de los coches modernos. El diseño de motores es una de las áreas prácticas que se ha visto beneficiada por el conocimiento de las leyes de la termodinámica. Los frigoríficos son otra. Usan reacciones químicas para transferir calor fuera de la comida en la nevera. Tiene que ir a algún lado, con frecuencia puedes sentir el calor saliendo del exterior del compartimento del motor del frigorífico. Lo mismo ocurre con el aire acondicionado. La generación de energía es otra aplicación. En una central de energía de carbón, de gas o nuclear, lo que es generado inicialmente es

calor. El calor crea vapor, que activa una turbina. La turbina, siguiendo principios que se remontan a Faraday, convierte el movimiento en electricidad.

La segunda ley de la termodinámica también determina la cantidad de energía que podemos esperar extraer de recursos renovables como el viento o las olas. El cambio climático ha añadido una nueva urgencia a esta cuestión, porque las fuentes de energía renovable producen menos dióxido de carbono que las convencionales. Incluso las centrales nucleares tienen un gran impacto de carbono, porque el combustible tiene que hacerse, transportarse y almacenarse cuando ya no es útil pero todavía es radiactivo. Cuando escribo esto hay un vivo debate sobre la cantidad máxima de energía que podemos extraer del océano y la atmósfera sin causar el tipo de cambio que estamos intentando evitar. Está basado en las estimaciones termodinámicas de la cantidad de energía libre en esos sistemas naturales. Esto es un asunto importante; si las renovables en principio no pueden aportar la energía que necesitamos, tenemos que buscar en otro lado. Los paneles solares, que extraen energía directamente de la luz del sol, no se ven afectados directamente por los límites de la termodinámica, pero incluso estos implican procesos de fabricación y todo eso. En este momento, ver dichos límites como un obstáculo serio recae en algunas simplificaciones radicales, e incluso aunque sean correctas, los cálculos no descartan las renovables como una fuente para la mayoría de la energía del mundo. Pero merece la pena recordar que de manera similar cálculos amplios sobre la producción de dióxido de carbono, realizados en la década de 1950, han probado ser sorprendentemente precisos en la predicción del calentamiento global.

La segunda ley funciona de manera brillante en su contexto original, el comportamiento de los gases, pero parece entrar en conflicto con las ricas complejidades de nuestro planeta, en concreto, la vida. Parece excluir la complejidad y organización exhibida por los sistemas vivos. De modo que la segunda ley es a veces invocada para atacar la evolución darwiniana. Sin embargo, la física de las máquinas de vapor no es particularmente apropiada para el estudio de la vida. En la teoría cinética de gases, las fuerzas que actúan entre las moléculas son de corto alcance (activas solo cuando las moléculas colisionan) y repulsivas (rebotan). Pero la mayoría de las fuerzas de la naturaleza no son así. Por ejemplo,

la gravedad actúa en distancias enormes y es atractiva. La expansión del universo a partir del Big Bang no ha emborronado la materia convirtiéndola en un gas uniforme. En su lugar, la materia se ha agrupado: planetas, estrellas, galaxias, supercúmulos... Las fuerzas que mantienen las moléculas unidas son también atractivas, excepto en distancias muy cortas, donde se hacen repulsivas, lo que evita que las moléculas colapsen, pero su alcance efectivo es bastante corto. Para sistemas como estos, el modelo termodinámico de subsistemas independientes cuyas interacciones se encienden pero no se apagan es simplemente irrelevante. Las características de la termodinámica tampoco se aplican, o se hace tan a largo plazo que no son el modelo de nada interesante.

Entonces, las leyes de la termodinámica sustentan muchas cosas que damos por hecho. Y la interpretación de la entropía como «desorden» nos ayuda a entender esas leyes y ganar un sentimiento intuitivo para sus bases físicas. Sin embargo, hay ocasiones en las que interpretar la entropía como desorden parece llevar a paradojas. Esto es una esfera más filosófica del discurso, y es fascinante.

Uno de los misterios más profundos de la física es la flecha del tiempo. El tiempo parece fluir en una dirección concreta. Sin embargo, parece posible lógica y matemáticamente para el tiempo fluir hacia atrás, una posibilidad explotada por libros como *La flecha del tiempo* de Martin Amis, la novela mucho más temprana *El mundo contra reloj* de Philip K. Dick, y la serie de televisión de la BBC *Enano rojo*, cuyos protagonistas memorablemente bebieron cerveza y participan en una pelea de bar en el tiempo marcha atrás. De modo que, ¿por qué no puede fluir el tiempo en el otro sentido? A primera vista, la termodinámica ofrece una explicación simple para la flecha del tiempo: es la dirección del incremento de la entropía. Los procesos termodinámicos son irreversibles: el oxígeno y el nitrógeno se mezclarán espontáneamente, pero no se *desmezclarán* espontáneamente.

Sin embargo, aquí hay un enigma, porque cualquier sistema mecánico clásico, como las moléculas en una habitación, es reversible en el tiempo. Si continúas barajando un montón de cartas de modo aleatorio, entonces finalmente volverán a su orden original. En las ecuaciones matemáticas, si en algún instante las velocidades de todas las partículas se invierten simultáneamente, entonces el sistema remontará sus pasos al revés en el tiempo. El universo entero puede rebotar, obedeciendo las

mismas ecuaciones en ambas direcciones. De modo que, ¿por qué nunca vemos un huevo *desrevueltándose*?

La respuesta termodinámica habitual es: un huevo revuelto está más desordenado que uno no revuelto, la entropía aumenta, y ese es el modo de fluir del tiempo. Pero hay una razón sutil por la que los huevos no se *desrevueltan*: es muy, muy, muy poco probable que el universo rebote de la manera necesaria. La probabilidad de que suceda es ridículamente pequeña. De modo que la discrepancia entre el incremento de la entropía y la reversibilidad del tiempo viene de las condiciones iniciales, no de las ecuaciones. Las ecuaciones para moléculas en movimiento son reversibles en el tiempo, pero las condiciones iniciales no. Cuando invertimos el tiempo debemos usar condiciones «iniciales» dadas por el estado final del movimiento del tiempo hacia delante.

La distinción más importante aquí es entre la simetría de las ecuaciones y la simetría de sus soluciones. Las ecuaciones para moléculas que rebotan tienen simetría reversible en el tiempo, pero las soluciones individuales pueden tener una flecha del tiempo definitiva. Cuanto más puedes deducir sobre una solución, a partir de la reversibilidad del tiempo de la ecuación, es que debe existir también otra solución que es reversible en el tiempo de la primera. Si Alicia lanza una pelota a Roberto, la solución reversible en el tiempo es Roberto lanzándole una pelota a Alicia. De modo similar, como las ecuaciones de la mecánica permiten a un vaso caer al suelo y romperse en mil pedazos, deben también permitir una solución en la cual miles de fragmentos de cristal misteriosamente se muevan juntos y se unan entre sí formando un vaso intacto y que salte en el aire.

Claramente hay algo extraño en eso, y requiere investigarlo. No tenemos problema con Roberto y Alicia lanzándose un balón en cualquier sentido. Vemos cosas así cada día. Pero no vemos un vaso romperse y luego recomponerse por sí solo. No vemos un huevo *desrevueltándose*.

Supón que rompemos un vaso y grabamos el resultado. Empezamos con un estado ordenado y simple: un vaso intacto. Cae al suelo, donde el impacto hace que se rompa en pedazos y lanza esos pedazos por todo el suelo. Estos se van frenando hasta que se detienen. Todo parece completamente normal. Ahora rebobina la película. Trozos de cristal, que resultan tener justo la forma correcta para encajar

unos con otros, están esparcidos por el suelo. Espontáneamente empiezan a moverse. Se mueven justo a la velocidad correcta, y justo en la dirección correcta, para encontrarse. Se ensamblan formando un vaso, que se dirige hacia el cielo. No parece que esté bien.

De hecho, como está descrito, no es correcto. Varias leyes de la mecánica parecen violarse, entre ellas la conservación del momento y la conservación de la energía. Masas que no están en movimiento no pueden de repente moverse. Un vaso no puede ganar energía de ninguna parte para saltar en el aire.

Ah, sí... pero eso es porque no estamos observando con la suficiente atención. El vaso no salta en el aire *motu proprio*. El suelo empieza a vibrar, y las vibraciones se juntan para dar al vaso una repentina patada al aire. Los trozos de cristal de manera similar fueron impulsados a moverse por ondas entrantes de vibración del suelo. Si trazamos esas vibraciones hacia atrás, se extienden, y parecen extinguirse. Finalmente la fricción disipa todo movimiento... Oh, sí, fricción. ¿Qué sucede con la energía cinética cuando hay fricción? Se convierte en calor. Así que hemos omitido algunos detalles del escenario del tiempo reversible. El momento y la energía se mantienen en equilibrio, pero las cantidades omitidas provienen del suelo perdiendo calor.

En principio, podemos establecer un sistema hacia delante en el tiempo para imitar el vaso en el tiempo invertido. Tan solo tenemos que inducir a las moléculas en el suelo a colisionar justo del modo correcto para liberar algo de su calor como movimiento del suelo, dar una patada a los trozos de cristal justo en el modo correcto, luego arrojar el vaso al aire. El asunto no es que sea imposible en principio, si lo fuera, la reversibilidad del tiempo fracasaría. Sino que es imposible en la práctica, ya que no hay modo de controlar tantas moléculas de un modo tan exacto.

Esto también es un tema sobre las condiciones frontera, en este caso las condiciones iniciales. Las condiciones iniciales para el experimento del vaso rompiéndose son fáciles de implementar, y el equipo es fácil de adquirir. Todo es muy robusto también; usa otro vaso, láñzalo desde una altura diferente... sucederá prácticamente lo mismo. El experimento del vaso ensamblándose, por el contrario, necesita un control extraordinariamente preciso de infinidad de moléculas

individuales y trozos de cristal hechos con sumo cuidado. Y sin que todo ese equipo de control moleste a una sola molécula. Es por esto por lo que no podemos hacerlo en realidad.

No obstante, observa cómo estamos pensando aquí; nos estamos centrando en condiciones iniciales. Eso establece una flecha del tiempo, el resto de la acción viene después del comienzo. Si viésemos las condiciones finales del vaso rompiéndose, bajando al nivel molecular, serían tan complejas que nadie en su sano juicio consideraría intentar replicarlas.

Las matemáticas de la entropía esquivan estas consideraciones a escala muy pequeña. Permite a las vibraciones extinguirse pero no aumentar. Permite a la fricción convertirse en calor, pero no al calor convertirse en fricción. La discrepancia entre la segunda ley de la termodinámica y la reversibilidad microscópica se plantea a partir de algo tosco, las suposiciones para hacer el modelo hechas cuando se pasó de la detallada descripción molecular a la estadística. Estas suposiciones implícitamente especifican una flecha del tiempo, se permite que las perturbaciones a gran escala se extingan bajo un plano perceptible *a medida que el tiempo avanza*, pero no se permite a las perturbaciones a pequeña escala seguir el escenario del tiempo reversible. Una vez la dinámica pasa a través de esta trampilla temporal, no se permite que vuelva.

Si la entropía siempre aumenta, ¿cómo la gallina jamás creó el huevo ordenado con el que empezar? Una explicación común, avanzada por el físico austriaco Erwin Schrödinger en 1944 en un libro breve y precioso llamado *¿Qué es la vida?*, es que los sistemas vivos de algún modo toman prestado orden de su entorno y lo devuelven haciendo el entorno incluso más desordenado de lo que de otro modo habría estado. Este orden extra se corresponde a la «entropía negativa», que la gallina puede usar para hacer un huevo sin violar la segunda ley. En el capítulo 15 veremos que la entropía negativa puede, en las circunstancias apropiadas, ser pensada como información, y se reivindica con frecuencia que la gallina accede a la información, proporcionada por su ADN, por ejemplo, para obtener la entropía negativa necesaria. Sin embargo, la identificación de la información con entropía negativa solo tiene sentido en unos contextos muy específicos, y las actividades de las criaturas vivas no son uno de ellos. Los organismos crean orden a través de

procesos que llevan a cabo, pero estos procesos no son termodinámicos. Las gallinas no acceden a algún almacén de orden para hacer que las reglas de la termodinámica se equilibren, usan procesos para los cuales el modelo termodinámico es inapropiado, y tiran las reglas porque no se aplican.

El escenario en el cual un huevo es creado al tomar prestada entropía sería apropiado si el proceso que la gallina usó era el inverso en el tiempo de un huevo rompiéndose en sus moléculas constituyentes. A primera vista esto es plausible de un modo vago, porque las moléculas que finalmente forman el huevo están esparcidas por todo el entorno; se unen en la gallina, donde los procesos bioquímicos las ponen juntas de una manera ordenada para formar un huevo. Sin embargo, hay una diferencia en las condiciones iniciales. Si diste una vuelta de antemano etiquetando moléculas en el entorno de la gallina, para decir «esta acabará en el huevo en tal o cual localización», estarías a todos los efectos creando condiciones iniciales tan complejas e improbables como las de hacer un huevo *desrevuelto*. Pero la gallina no funciona así. Algunas moléculas podrían haber hecho el mismo trabajo: una molécula de carbonato de calcio es tan buena para hacer la cáscara como cualquier otra. De modo que la gallina no está creando orden del desorden. El orden se asigna al resultado final del proceso de hacer el huevo, como barajar las cartas en un orden aleatorio y luego numerarlas 1, 2, 3, etcétera, con un rotulador. Sorprendente, iestán en orden numérico!

Para estar seguro, el huevo parece más ordenado que sus ingredientes, incluso si tenemos en cuenta esta diferencia en las condiciones iniciales. Pero eso es porque el proceso que hace un huevo no es termodinámico. De hecho, muchos procesos físicos hacen huevos *desrevueltos*. Un ejemplo es el modo en que los minerales disueltos en agua puede crear stalactitas y stalagmitas en las cuevas. Si especificamos la forma exacta de la stalactita que queremos, por adelantado, estaríamos en la misma posición que alguien tratando de recomponer un vaso roto. Pero si estamos dispuestos a conformarnos con cualquier stalactita vieja, obtenemos una: orden del desorden. Estos dos términos son con frecuencia usados de una manera descuidada. Lo que importa son qué tipo de orden y qué tipo de desorden. Dicho esto, sigo sin esperar ver un huevo *desrevueltarse*. No hay un modo factible de establecer las condiciones iniciales necesarias. Lo mejor que

podemos hacer es convertir el huevo revuelto en comida para las gallinas y esperar que el ave ponga uno nuevo.

De hecho, hay una razón por la que no veríamos un huevo *desrevueltarse*, incluso si el mundo fuese marcha atrás. Como nosotros y nuestras memorias somos parte del sistema que está siendo invertido, no estaríamos seguros de qué sentido del tiempo está «realmente» ocurriendo. Nuestro sentido del fluir del tiempo está producido por las memorias, patrones psicoquímicos en el cerebro. En el lenguaje convencional, el cerebro almacena registros del pasado, pero no del futuro. Imagina que haces una serie de instantáneas del cerebro observando un huevo siendo revuelto, junto con su memoria del proceso. En una etapa el cerebro recuerda un huevo frío y que no estaba revuelto, y algo de su historia cuando lo cogimos del frigorífico y lo pusimos en la sartén. En otra etapa recuerda haber batido el huevo con un tenedor y haberlo movido de la nevera a la sartén.

Si ahora todo el universo gira al revés, invertimos el orden en que las memorias ocurren en tiempo «real». Pero no invertimos el orden de una memoria dada en el cerebro. Al principio (en el tiempo invertido) del proceso que *desrevuelve* el huevo, el cerebro no recuerda el «pasado» del huevo, cómo aparece de la boca en la cuchara, cómo fue *desbatido* y cómo gradualmente construyó un huevo completo... En su lugar, el registro en el cerebro en ese momento es uno en el que recuerda haber golpeado un huevo para abrirlo, junto con el proceso de moverlo del frigorífico a la sartén y hacerlo revuelto. Pero este recuerdo es exactamente el mismo que el de los registros en el escenario en el que el tiempo va hacia delante. Lo mismo ocurre para todas las otras instantáneas de la memoria. Nuestra percepción del mundo depende de lo que observemos ahora, y qué memorias guarde nuestro cerebro ahora. En un universo con un tiempo invertido, en realidad recordaríamos el futuro, no el pasado.

La paradoja de la reversibilidad del tiempo y la entropía no son problemas sobre el mundo real. Son problemas sobre las suposiciones que hacemos cuando intentamos hacer un modelo de ellas.

Capítulo 13
Una cosa es absoluta
Relatividad

The diagram shows the famous equation $E = mc^2$ in a large serif font. Four arrows point from labels to specific parts of the equation:

- An arrow points to the term m with the label "masa" (mass).
- An arrow points to the term c^2 with the label "cuadrado" (square).
- An arrow points to the entire term mc^2 with the label "energía remanente de la masa" (rest energy).
- An arrow points to the term c with the label "velocidad de la luz" (speed of light).

¿Qué dice?

La materia contiene energía igual a su masa multiplicada por el cuadrado de la velocidad de la luz.

¿Por qué es importante?

La velocidad de la luz es enorme y su cuadrado es absolutamente monumental. Un kilogramo de materia liberaría alrededor del 40 % de la energía en el arma nuclear más grande que jamás ha explotado. Es parte de un paquete de ecuaciones que cambiaron nuestra visión del espacio, tiempo, materia y gravedad.

¿Qué provocó?

Indudablemente, física radicalmente nueva. Armas nucleares... bueno, solo quizá, aunque no tan directamente o de manera concluyente como los mitos urbanos reclaman. Agujeros negros, el Big Bang, GPS y navegación vía satélite.

Al igual que Albert Einstein, con su mata de pelo alborotada, es el científico arquetípico en la cultura popular, su ecuación $E = mc^2$ es la ecuación arquetípica. Es una creencia generalizada que la ecuación llevó a la invención de las armas nucleares, que viene de la teoría de la relatividad de Einstein y que esa teoría (obviamente) tiene algo que ver con varias cosas que son relativas. De hecho, muchos relativistas sociales felizmente corean «todo es relativo», y piensan que

tiene algo que ver con Einstein.

No tiene nada que ver. Einstein hizo su teoría «de la relatividad» porque era una modificación de las reglas para el movimiento relativo que se habían usado tradicionalmente en la mecánica newtoniana, donde el movimiento es relativo, dependiendo de un modo simple e intuitivo del sistema de referencia en el que se observa. Einstein tuvo que retocar ligeramente la relatividad newtoniana para que tuviese sentido un descubrimiento experimental desconcertante: que un fenómeno físico particular no es relativo para nada, sino absoluto. A partir de aquí obtuvo un nuevo tipo de física en la cual los objetos se encogen cuando se mueven muy rápido, el tiempo avanza a paso de tortuga y la masa aumenta sin límite. Una extensión que incorpora la gravedad nos ha dado una comprensión mejor de la que ya teníamos de los orígenes del universo y la estructura del cosmos. Está basada en la idea de que el espacio y el tiempo pueden curvarse.

La relatividad es real. El Sistema de Posicionamiento Global (GPS, usado entre otras cosas para la navegación vía satélite de los coches) funciona porque se hicieron las correcciones para los resultados relativistas. Lo mismo es aplicable a los aceleradores de partículas como el Gran Colisionador de Hadrones, que ha anunciado el descubrimiento del bosón de Higgs en 2012, que se cree que es el origen de la masa. Las comunicaciones modernas se han hecho tan rápidas que los mercados están empezando a correr contra una limitación relativista: la velocidad de la luz. Esto es lo más rápido que cualquier mensaje, como por ejemplo una orden en Internet de comprar o vender acciones, puede viajar. Algunos ven esto como una oportunidad de cerrar un trato nanosegundos antes que la competencia, pero, hasta cierto punto, los resultados relativistas no han tenido un efecto serio en las finanzas internacionales. Sin embargo, la gente ya ha calculado las mejores localizaciones para mercados de valores o franquicias. Es solo una cuestión de tiempo.

De cualquier forma, no solo la relatividad no es relativa, incluso la icónica ecuación no es lo que parece. Al principio, cuando Einstein obtuvo la idea física que representa, no la escribió en el modo habitual. No es una consecuencia matemática de la relatividad, aunque se convierte en una si se aceptan varias suposiciones y definiciones físicas. Es quizá típico de la cultura humana que nuestra ecuación más

icónica no es, y no fue, lo que parece ser, y tampoco lo es la teoría que la vio nacer. Incluso la conexión con armamento nuclear no es clara, y su influencia histórica en la primera bomba atómica fue pequeña comparada con el peso político de Einstein como el científico icónico.

La «relatividad» cubre dos teorías distintas pero relacionadas: relatividad especial y relatividad general. Usaré la famosa ecuación de Einstein como una excusa para hablar de ambas. La relatividad especial es sobre el espacio, el tiempo y la materia en ausencia de gravedad; la relatividad general también tiene la gravedad en cuenta. Las dos teorías son parte de una gran imagen, pero Einstein tardó diez años de esfuerzo intenso en descubrir cómo modificar la relatividad especial para incorporar la gravedad. Ambas teorías estaban inspiradas por las dificultades en reconciliar la física newtoniana con las observaciones, pero la fórmula icónica surgió en la relatividad especial.

La física parecía bastante sencilla e intuitiva en la época de Newton. El espacio era espacio, el tiempo era tiempo y los dos nunca se deberían encontrar. La geometría del espacio era la de Euclides. El tiempo era independiente del espacio, el mismo para todos los que observaban, siempre que tuvieran sus relojes sincronizados. La masa y el tamaño de un cuerpo no cambiaban cuando se movían y el tiempo siempre pasaba al mismo ritmo por todas partes. Pero cuando Einstein había acabado de reformular la física, todas estas afirmaciones, tan intuitivas que es muy difícil imaginar cómo cualquiera de ellas podría fracasar representando la realidad, resultaron ser erróneas.

No eran totalmente erróneas, por supuesto. Si no hubiesen tenido sentido, entonces el trabajo de Newton nunca habría despegado. La imagen newtoniana del universo físico es una aproximación, no una descripción exacta. La aproximación es extremadamente precisa siempre que todo lo involucrado se esté moviendo lo suficientemente lento, y en la mayoría de las circunstancias del día a día ese es el caso. Incluso un avión caza, viajando al doble de la velocidad del sonido, se está moviendo lentamente para este fin. Pero una de las cosas que sí juega un papel en la vida diaria se mueve realmente muy rápido, y establece el criterio para todas las otras velocidades: la luz. Newton y sus sucesores habían demostrado que la luz era una onda y las ecuaciones de Maxwell lo confirmaron. Pero la naturaleza de onda de

la luz destapó un nuevo tema. Las olas del mar son ondas en el agua, las ondas del sonido son ondas en el aire, los terremotos son ondas en la Tierra. Así que las ondas de luz eran ondas en... ¿qué?

Matemáticamente, son ondas en el campo electromagnético, que se asume llena todo el espacio. Cuando el campo electromagnético está excitado, persuadido para sustentar electricidad y magnetismo, observamos una onda. Pero ¿qué sucede cuando no está excitado? Sin ondas, un océano, todavía sería un océano, el aire todavía sería aire y la Tierra todavía sería la Tierra. Análogamente, el campo electromagnético todavía sería... el campo electromagnético. Pero no puedes observar el campo electromagnético si no hay electricidad o magnetismo ocurriendo. Si no puedes observarlo, ¿qué es? ¿Existe en realidad?

Todas las ondas conocidas en física, excepto las del campo electromagnético, son ondas sobre algo tangible. Los tres tipos de onda —las olas, las del aire y los terremotos— son ondas de movimiento. El medio se mueve de arriba abajo o de un lado a otro, pero normalmente no se desplaza con la onda. (Ata una cuerda larga a una pared y agita un extremo, una onda se desplaza por la cuerda, pero la cuerda no se desplaza por la cuerda.) Hay excepciones: cuando el aire viaja conjuntamente con la onda, lo llamamos «viento», y las olas del mar mueven el agua en la playa cuando se topan con una. Pero incluso aunque describamos un tsunami como una pared de agua en movimiento, no rueda por encima del océano como un balón rueda por un campo. Generalmente, el agua en una localización determinada se mueve arriba y abajo. Es la localización del «arriba» lo que se mueve. Hasta que el agua se acerca a la orilla, entonces lo que tienes es algo mucho más parecido a una pared en movimiento.

La luz, y las ondas electromagnéticas en general, no parecían ser ondas en algo tangible. En la época de Maxwell, y durante cincuenta años o más después, eso era inquietante. La ley de la gravedad de Newton había sido muy criticada porque implicaba que la gravedad, de algún modo, «actúa en la distancia», tan milagroso en un principio filosófico como dar una patada a un balón y marcar gol cuando estás sentado en las gradas. Decir que es transmitido por «el campo gravitacional» no explica realmente qué está sucediendo. Lo mismo sucede con el electromagnetismo. Así que los físicos se convencieron con la idea de que había algún medio, nadie

sabía qué, lo llamaron el «éter luminoso» o simplemente «éter», que sustentaba a las ondas electromagnéticas. Las vibraciones se desplazan más rápido cuanto más rígido es el medio y la luz era muy rápida, de modo que el éter tenía que ser sumamente rígido. Aunque los planetas se podían mover a través de él sin resistencia. Al no haberse detectado con facilidad, el éter no debía tener masa, ni viscosidad, ser incompresible y ser totalmente transparente a todas las formas de radiación.

Era una combinación de atributos sobrecogedora, pero casi todos los físicos asumieron que el éter existía, porque la luz claramente hacía lo que hacía. Algo tenía que llevar la onda. Además, la existencia del éter podía en principio detectarse, porque otra característica de la luz sugería un modo de observarlo. En un vacío, la luz se mueve con una velocidad fija c . La mecánica newtoniana había enseñado a todo físico a preguntar: ¿velocidad relativa a qué? Si mides la velocidad en dos sistemas de referencia diferentes, uno moviéndose con respecto a otro, obtienes respuestas diferentes. La constancia de la velocidad de la luz sugería una respuesta obvia: relativa al éter. Pero esto era un poco simplista, porque dos sistemas de referencia que se están moviendo uno con respecto al otro no pueden ser ambos relativos al éter en reposo.

A medida que la Tierra se abre paso a través del éter, que milagrosamente no se resiste, va dando vueltas alrededor del Sol. En puntos opuestos de su órbita, se está moviendo en direcciones opuestas. De modo que por la mecánica newtoniana, la velocidad de la luz debería variar entre los dos extremos: c más una contribución del movimiento de la Tierra relativo al éter, y c menos la misma contribución. Mide la velocidad, mídela seis meses más tarde, encuentra la diferencia, si hay una, y has probado que el éter existe. A finales del siglo XIX, se llevaron a cabo muchos experimentos siguiendo esta línea, pero los resultados no eran concluyentes. Por lo tanto no había diferencia, o había una pero el método experimental no era suficientemente preciso. Peor, la Tierra quizá estuviese arrastrando el éter con ella. Esto explicaría simultáneamente por qué la Tierra podía moverse a través de un medio tan rígido sin resistencia e implicaría que no deberíamos ver ninguna diferencia en la velocidad de la luz. El movimiento de la Tierra relativo al éter sería siempre cero.

En 1887, Albert Michelson y Edward Morley llevaron a cabo uno de los experimentos físicos más famosos de todos los tiempos. Sus aparatos fueron diseñados para detectar variaciones extremadamente pequeñas en la velocidad de la luz en dos direcciones, perpendicular la una a la otra. Como la Tierra se estaba moviendo en relación con el éter, no podría moverse con las mismas velocidades relativas en dos direcciones diferentes... a menos que se diese la coincidencia de que se estuviese moviendo a lo largo de la línea que dividía en dos estas direcciones, en cuyo caso bastaría que rotases el aparato un poco y lo intentases de nuevo.

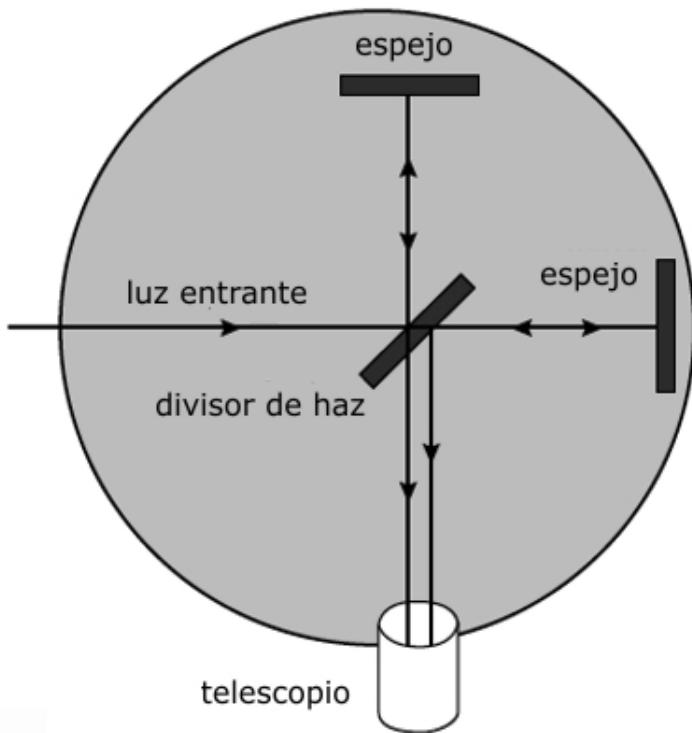


FIGURA 48. Experimento de Michelson-Morley.

El aparato (figura 48), era lo suficientemente pequeño para que pudiese estar en una mesa de laboratorio. Usaba un espejo semiplateado para dividir un rayo de luz en dos partes, una pasando a través del espejo y la otra reflejándose en un ángulo recto. Cada rayo por separado se reflejaba de vuelta a lo largo de su trayectoria y los dos rayos combinados de nuevo chocaban con un detector. El aparato se ajustaba para hacer las trayectorias de la misma longitud. El rayo original se establecía para ser coherente, lo que quiere decir que sus ondas estaban en

sincronía unas con otras, todas tenían la misma fase, picos coincidiendo con picos. Cualquier diferencia entre la velocidad de la luz en las direcciones seguidas por los dos rayos causaría que sus fases cambiasen la una con respecto a la otra, de modo que sus picos estarían en lugares diferentes. Esto causaría interferencia entre las dos ondas, resultando un patrón a rayas de «franjas de difracción». El movimiento de la Tierra relativo al éter causaría que las franjas se moviesen. El efecto sería diminuto; una vez determinado lo que se conocía del movimiento de la Tierra respecto al Sol, las franjas de difracción se moverían alrededor de un 4 % de ancho de una franja. Usando múltiples reflexiones, esto podría aumentar al 40 %, que es lo suficientemente grande para detectarse. Para evitar la posible coincidencia de la Tierra moviéndose exactamente a lo largo del bisector de los dos rayos, Michelson y Morley hacían flotar el aparato en un baño de mercurio, de manera que podía girarse fácil y rápidamente. Debería entonces ser posible observar las franjas moviéndose con igual rapidez.

Fue un experimento cuidadoso y preciso. Su resultado fue totalmente negativo. Las franjas no se movieron el 40 % de su ancho. Hasta donde alguien podría decir con certeza, no se movieron para nada. Experimentos posteriores, capaces de detectar una alteración de un 0,07 % del ancho de la franja, también dieron un resultado negativo. El éter no existía.

Este resultado no solo descartaba al éter, también amenazaba con descartar la teoría de Maxwell del electromagnetismo. Implicaba que la luz no se comportaba de una manera newtoniana, relativa a sistemas de referencia móviles. Este problema puede remontarse directamente a las propiedades matemáticas de las ecuaciones de Maxwell y cómo se transforman en relación con un sistema móvil. El físico y químico irlandés George FitzGerald y el físico holandés Hendrik Lorenz sugirieron de manera independiente (en 1892 y 1895, respectivamente) un modo audaz de sortear el problema. Si un cuerpo moviéndose se contrae ligeramente en su dirección de movimiento, justo la cantidad correcta, entonces el cambio en la fase que el experimento de Michelson-Morley estaba esperando detectar se contrarrestaría de manera exacta con el cambio en la longitud de la trayectoria que la luz estaba siguiendo. Lorenz mostró que esta «contracción de Lorenz-FitzGerald» solucionaba, también, las dificultades matemáticas de las ecuaciones de Maxwell. El

descubrimiento conjunto mostró que los resultados de los experimentos en electromagnetismo, incluyendo la luz, no dependían del movimiento relativo del sistema de referencia. Poincaré, que también había estado trabajando en una línea similar, añadió su convincente peso intelectual a la idea.

El escenario estaba ahora preparado para Einstein. En 1905, desarrolló y amplió las especulaciones previas sobre una nueva teoría de movimiento relativo en un artículo «*On the electrodynamics of moving bodies*» (Sobre la electrodinámica de los cuerpos en movimiento). Su trabajo fue más allá que el de sus predecesores en dos sentidos. Mostró que el cambio necesario para la formulación matemática del movimiento relativo era más que un truco para arreglar el electromagnetismo. Se necesitaba para todas las leyes físicas. Entendió que las nuevas matemáticas debían ser una descripción genuina de la realidad, con el mismo estatus filosófico que había sido acordado para la descripción newtoniana reinante, pero proporcionando una concordancia mejor con los experimentos. Era física real.

La visión del movimiento relativo empleada por Newton se remontaba incluso más allá, a Galileo. En su *Dialogo sopra i due massimi sistemi del mondo* (Diálogo sobre los dos máximos sistemas del mundo) Galileo hablaba de un barco viajando a velocidad constante en un mar totalmente en calma, y argumentaba que ningún experimento en mecánica llevado a cabo bajo cubierta podría revelar que el barco se estaba moviendo. Esto es el principio de la relatividad de Galileo: en mecánica, no hay diferencia entre las observaciones hechas en dos sistemas que se están moviendo con una velocidad uniforme uno con respecto al otro. En concreto, no hay un sistema especial de referencia que esté «en reposo». El punto de arranque de Einstein fue el mismo principio, pero con una vuelta de tuerca extra: debe aplicarse no solo a la mecánica, sino a todas las leyes de la física. Entre ellas, por supuesto, están las ecuaciones de Maxwell y la constancia de la velocidad de la luz.

Para Einstein, el experimento de Michelson-Morley era una pequeña pieza de evidencia extra, pero no era la prueba. La prueba de que su nueva teoría era correcta recaía en su principio extendido de la relatividad y lo que implicaba para la estructura matemática de las leyes de la física. Si puedes aceptar el principio, todo lo demás lo sigue. Este es el motivo de que la teoría fuese conocida como «relatividad». No porque «todo es relativo», sino porque tienes en cuenta la manera

en la que todo es relativo. Y no es lo que esperabas.

Esta versión de la teoría de Einstein es conocida como relatividad especial porque se aplica solo en sistemas de referencia que se están moviendo uniformemente el uno con respecto al otro. Entre sus consecuencias están las contracciones de Lorenz-FitzGerald, ahora interpretadas como una característica necesaria del espacio-tiempo. De hecho, había tres efectos relacionados. Si un sistema de referencia se está moviendo de manera uniforme respecto a otro, entonces las longitudes medidas en el sistema se contraen a lo largo de la dirección del movimiento, la masa aumenta y el tiempo avanza más lentamente. Estos tres efectos están atados unos a otros por las leyes básicas de conservación de la energía y el momento, una vez aceptas uno de ellos, los otros son consecuencias lógicas.

La formulación técnica de estos efectos es una fórmula que describe cómo se relacionan mediciones en un sistema con las correspondientes en el otro. El resumen es: si un cuerpo pudiese moverse próximo a la velocidad de la luz, entonces su longitud se haría muy pequeña, el tiempo avanzaría muy lentamente y su masa se haría muy grande. Tan solo daré una idea de las matemáticas: la descripción física no debería tomarse demasiado literalmente y llevaría demasiado establecerla en el lenguaje correcto. Todo viene de... el teorema de Pitágoras. Una de las ecuaciones más antiguas en ciencias lleva a una de las más nuevas.

Supón que una nave espacial está pasando por encima de nuestras cabezas con velocidad v y la tripulación realiza un experimento. Envía una secuencia de luz desde el suelo de la cabina al techo y la medición del tiempo es T . Mientras un observador en el suelo observa el experimento a través de un telescopio (asumiendo que la nave espacial es transparente) y mide el tiempo como t .

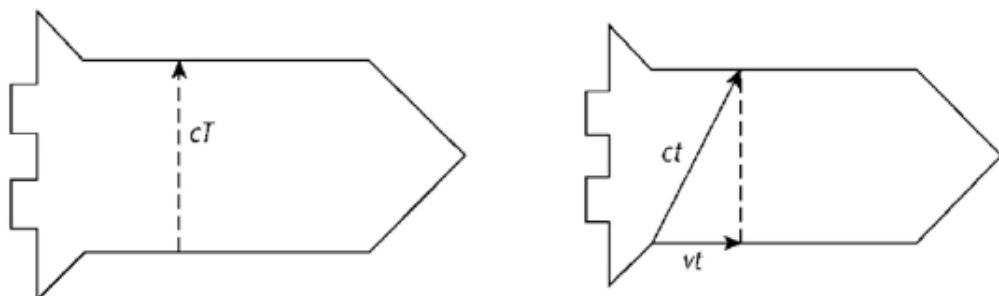


FIGURA 49. A la izquierda: el experimento en el sistema de referencia de la

tripulación. A la derecha: el mismo experimento en el sistema de referencia del observador desde el suelo. El gris muestra la posición de la nave vista desde el suelo cuando el rayo de luz empieza su viaje; el negro muestra la posición de la nave cuando la luz completa su viaje.

La figura 49 (izquierda) muestra la geometría del experimento desde el punto de vista de la tripulación. Para ellos, la luz ha ido derecha hacia arriba. Como la luz viaja a una velocidad c , la distancia que se desplaza es cT , mostrada con la flecha punteada. La figura 49 (derecha) muestra la geometría del experimento desde el punto de vista del observador desde el suelo. La nave espacial se ha movido una distancia vt , de modo que la luz se ha desplazado diagonalmente. Como la luz también se desplaza a una velocidad c para el observador desde el suelo, la diagonal tiene una longitud ct . Pero la línea de puntos tiene la misma longitud que la línea de puntos en la primera imagen, en concreto, cT . Por el teorema de Pitágoras:

$$(ct)^2 = (cT)^2 + (vt)^2$$

Despejamos T , y obtenemos:

$$T = t \sqrt{1 - \frac{v^2}{c^2}}$$

Que es más pequeño que t .

Para obtener la contracción de Lorenz-FitzGerald, ahora nos imaginamos que la nave espacial viaja a un planeta a una distancia de la Tierra x a velocidad v . Entonces el tiempo transcurrido es $t = x/v$. Pero la fórmula previa muestra que para la tripulación, el tiempo que tarda es T , no t . Para ellos, la distancia X debe satisfacer $T = X/v$. Por lo tanto:

$$x = x \sqrt{1 - \frac{v^2}{c^2}}$$

Que es más pequeña que x .

La obtención del cambio de masa es ligeramente más complicada y depende de una interpretación concreta de la masa, «masa en reposo», de modo que no daré detalles. La fórmula es:

$$m = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}}$$

Que es mayor que m_0 .

Estas ecuaciones nos dicen que hay algo muy especial en la velocidad de la luz (y por tanto en la luz). Una consecuencia importante de este formalismo es que la velocidad de la luz es una barrera impenetrable. Si un cuerpo arranca más lento que la luz, no puede acelerarse a una velocidad mayor que la de la luz. En septiembre de 2011, físicos que trabajaban en Italia anunciaron que las partículas subatómicas llamadas neutrinos parecían estar viajando más rápido que la luz.³⁶ Su observación es polémica, pero si se confirma, llevará a una nueva física importante.

³⁶ En el Laboratori Nazionali del Gran Sasso, en Italia, hay un detector de partículas de 1.300 toneladas llamado OPERA (acrónimo inglés para *Oscillation Project with Emulsion-tRacking Apparatus*). Durante más de dos años rastreó 16.000 neutrinos producidos en el CERN, el laboratorio europeo de física de partículas en Ginebra. Los neutrinos son partículas subatómicas eléctricamente neutras con una masa muy pequeña, y pueden pasar a través de materia ordinaria con facilidad. Los resultados fueron desconcertantes: de media los neutrinos completaban el viaje de 730 kilómetros en 60 nanosegundos (mil millonésimas de segundo) más rápido de lo que lo habrían hecho si hubiesen viajado a la velocidad de la luz. Las mediciones eran precisas con un margen de error de 10 nanosegundos, pero ahí se encuentra la posibilidad de algún error sistemático en el modo en que se calcularon e interpretaron los tiempos, que es sumamente complejo.

Los resultados han sido publicados online: «Measurement of the neutrino velocity with the OPERA detector in the CNGS beam» por OPERA Collaboration, <http://arxiv.org/abs/1109.4897>

Este artículo no reivindica haber refutado la relatividad, simplemente presenta sus observaciones como algo que el equipo no puede explicar con la física convencional. Un informe no técnico puede encontrarse en: <http://www.nature.com/news/2011/110922/full/news.2011.554.html>

Una posible fuente de error sistemático, relacionado con las diferencias en las fuerzas de gravedad en los dos laboratorios, se propone en: <http://www.nature.com/news/2011/111005/full/news.2011.575.html>, pero el equipo de OPERA cuestiona esta sugerencia.

La mayoría de los físicos creen que, a pesar del gran cuidado tenido por los investigadores, hay un error sistemático implicado. En concreto, observaciones de neutrinos anteriores de una supernova parecen entrar en conflicto con las nuevas. La resolución de la polémica necesitará experimentos independientes y estos requerirán varios años. Los físicos teóricos están ya analizando explicaciones en potencia que van de extensiones menores muy conocidas del modelo estándar de la física de partículas a una exótica nueva física en la cual el universo tiene más dimensiones que las cuatro habituales. En el momento en que leas esto, la historia ya habrá avanzado.

Pitágoras aparece en la relatividad de otras maneras. Una es la formulación de la relatividad especial en términos de la geometría del espacio-tiempo, originalmente introducida por Hermann Minkowski. El espacio ordinario newtoniano puede capturarse matemáticamente haciendo corresponder sus puntos con tres coordenadas (x, y, z) , y definiendo la distancia d entre dicho punto y otro (X, Y, Z) con el teorema de Pitágoras:

$$d^2 = (x - X)^2 + (y - Y)^2 + (z - Z)^2$$

Ahora hacemos la raíz cuadrada para obtener d . El espacio-tiempo de Minkowski es similar, pero ahora hay cuatro coordenadas (x, y, z, t) , tres del espacio más una del tiempo, y un punto se llama suceso, una localización en el espacio, observado en un tiempo específico. La fórmula de la distancia es muy similar:

$$d^2 = (x - X)^2 + (y - Y)^2 + (z - Z)^2 - c^2(t - T)^2$$

El factor c^2 es solo una consecuencia de las unidades usadas para medir el tiempo, pero el signo menos delante es crucial. La «distancia» d es llamada el intervalo y la raíz cuadrada es real solo cuando la parte derecha de la ecuación es positiva. Lo que se reduce a la distancia espacial entre los dos sucesos siendo mayor que la diferencia temporal (en unidades correctas: años luz y años, por ejemplo). Eso, a su vez, significa que en principio un cuerpo podría desplazarse desde el primer punto en el espacio en el primer tiempo y llegar al segundo punto en el espacio en el segundo tiempo, sin ir más rápido que la luz.

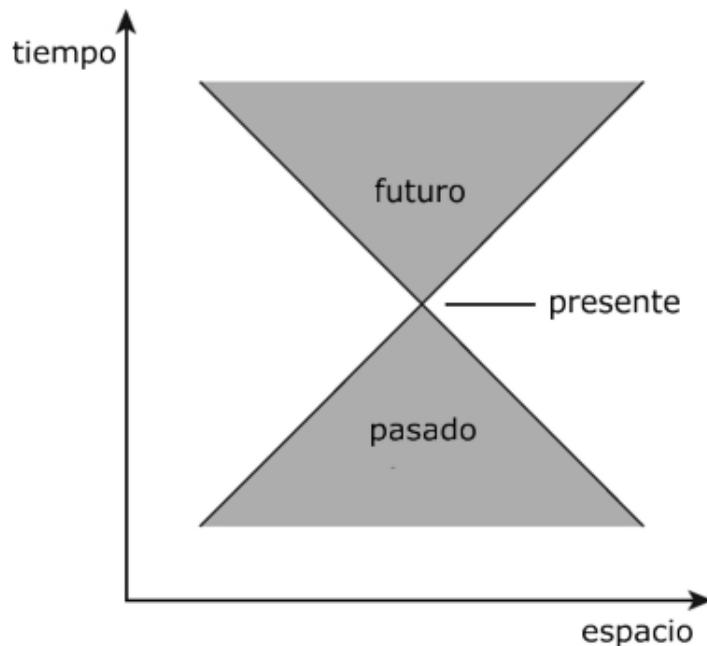


FIGURA 50. Espacio-tiempo de Minkowski, con el espacio mostrado como unidimensional.

En otras palabras, el intervalo es real si, y solo si, es físicamente posible, en principio, desplazarse entre dos sucesos. El intervalo es cero si, y solo si, la luz puede desplazarse entre ellos. Esta región físicamente accesible se llama el cono de luz de un suceso y viene en dos partes: el pasado y el futuro. La figura 50 muestra la geometría cuando el espacio está reducido a una dimensión.

Ahora te he mostrado tres ecuaciones relativistas y bosquejado cómo surgen, pero ninguna de ellas es la icónica ecuación de Einstein. Sin embargo, ahora estamos listos para entender cómo la obtuvo, una vez apreciemos una innovación más de la física de principios del siglo XX. Como hemos visto, los físicos habían realizado con anterioridad experimentos para demostrar de manera concluyente que la luz era una onda, y Maxwell había demostrado que era una onda electromagnética. Sin embargo, en 1905, estuvo claro que a pesar del peso de la evidencia para la naturaleza en forma de onda de la luz, hay circunstancias en las cuales se comporta como una partícula. Ese año Einstein usó esta idea para explicar algunas características del efecto fotoeléctrico, en el que la luz que choca con un metal apropiado genera electricidad. Argumentó que los experimentos tenían sentido solo

si la luz llegaba en paquetes discretos, a todos los efectos, partículas. Ahora se llaman fotones.

Este desconcertante descubrimiento era uno de los pasos clave hacia la mecánica cuántica y diré más sobre ello en el capítulo 14. Curiosamente, esta idea de la mecánica cuántica intrínsecamente era vital para la formulación de la relatividad de Einstein. Para obtener su ecuación relacionando masa con energía, Einstein pensó en qué le sucede a un cuerpo que emite un par de fotones. Para simplificar los cálculos, restringió la atención a una dimensión del espacio, de modo que el cuerpo se movía a lo largo de una línea recta. Esta simplificación no afecta a la respuesta. La idea básica es considerar el sistema en dos sistemas de referencia diferentes.³⁷ Uno se mueve con el cuerpo, de modo que el cuerpo parece estar parado en ese sistema. El otro sistema se mueve con una velocidad relativa al cuerpo pequeña pero distinta de cero. Permíteme llamarlos el sistema estacionario y el sistema en movimiento. Son como la nave espacial (en su propio sistema de referencia está detenida) y mi observador terrestre (para quien la nave parece estar en movimiento).

Einstein asumió que los dos fotones son igualmente energéticos, pero emitidos en direcciones opuestas. Sus velocidades son iguales y opuestas, de modo que la velocidad del cuerpo (en cualquier marco) no cambia cuando los fotones se emiten. Entonces calculaba la energía del sistema antes de que el cuerpo emitiese el par de fotones, y después. Al asumir que la energía debe conservarse, obtuvo una expresión que relaciona el cambio en la energía del cuerpo, provocada por la emisión de fotones, con el cambio en su masa (relativista). El resultado fue:

$$(cambio en la energía) = (cambio en la masa) \times c^2$$

³⁷ Una explicación rigurosa la da Terence Tao en su website: <http://terrytao.wordpress.com/2007/12/28/einsteins-derivation-of-emc2/>

La obtención de la ecuación supone cinco pasos:

- a. Describir cómo las coordenadas del espacio y el tiempo se transforman cuando el marco de referencia se cambia.
- b. Usar esta descripción para calcular cómo la frecuencia de un fotón se transforma cuando el marco de referencia se cambia.
- c. Usar la ley de Planck para calcular cómo se transforman la energía y el momento de un fotón.
- d. Aplicar la conservación de la energía y el momento para calcular cómo la energía y el momento de un cuerpo en movimiento se transforman.
- e. Fijar el valor de una constante de lo contrario arbitraria en el cálculo comparando los resultados con la física newtoniana cuando la velocidad del cuerpo es pequeña.

Haciendo la suposición razonable de que un cuerpo de masa cero tiene energía cero, entonces tenemos que:

$$\text{energía} = \text{masa} \times c^2$$

Esto, por supuesto, es la fórmula famosa, en la que E simboliza la energía y m la masa.

Además de hacer los cálculos, Einstein tuvo que interpretar su significado. En concreto, expuso que en un sistema para el cual el cuerpo está en reposo, la energía dada por la fórmula debería considerarse como su energía «interna», que posee porque está hecho de partículas subatómicas, y cada una tiene su propia energía. En un sistema en movimiento, hay también una contribución de la energía cinética. Hay otras sutilezas matemáticas también, tales como el uso de una velocidad pequeña y aproximaciones a las fórmulas exactas.

A Einstein con frecuencia se le atribuye el mérito, si esa es la palabra, de la comprensión de que una bomba atómica liberaría cantidades de energía tremendas. Ciertamente esa impresión dio la revista *Time* en julio de 1946 cuando puso su cara en la cubierta con una nube de hongo atómica tras él con su ecuación icónica. La conexión entre la ecuación y una explosión enorme parece clara; la ecuación nos dice que la energía inherente en cualquier objeto es su masa multiplicada por el cuadrado de la velocidad de la luz. Como la velocidad de la luz es enorme, su cuadrado es todavía mayor, lo cual identifica mucha energía en una pequeña cantidad de materia. La energía en un gramo de materia resulta ser 90 terajulios, equivalente más o menos a la producción de un día de electricidad de una central de energía nuclear.

Sin embargo, no ocurre así. La energía liberada en una bomba atómica es solo una pequeña fracción de la masa en reposo relativista y los físicos ya son conscientes, por motivos experimentales, que ciertas reacciones nucleares podrían liberar mucha energía. El principal problema técnico era mantener junto un pedazo de material radiactivo adecuado durante el tiempo suficiente para que se provocase una reacción en cadena, que creciese exponencialmente, en la cual la descomposición

de un átomo radiactivo sea la causante de emitir radiaciones que desencadenen el mismo efecto en otros átomos. Sin embargo, la ecuación de Einstein rápidamente comenzó a establecerse en la mente del público como la progenitora de la bomba atómica. El Informe Smyth, un documento del gobierno de Estados Unidos dado a conocer al público para explicar la bomba atómica, coloca la ecuación en su segunda página. Sospecho que lo que pasó es lo que Jack Cohen y yo hemos llamado «mentiras a los niños», historias simplificadas dichas con propósitos legítimos, que pavimentan el camino a una explicación más precisa.³⁸ Esto muestra cómo funciona la educación: la historia completa es siempre demasiado complicada para cualquiera excepto para los expertos y ellos saben tanto que no creen la mayoría de ello.

Sin embargo, la ecuación de Einstein no puede ser descartada sin más. Sí desempeñó un papel en el desarrollo de armas nucleares. La noción de fisión nuclear, que impulsa la bomba atómica, surge de discusiones entre los físicos Lise Meitner y Otto Frisch en la Alemania nazi en 1938. Estaban intentando comprender las fuerzas que mantienen al átomo junto, las cuales son un poco como la tensión de la superficie de una gota de un líquido. Estaban fuera paseando, discutiendo sobre física y aplicaron la ecuación de Einstein para averiguar si la fisión era posible por motivos energéticos. Frisch posteriormente escribió:³⁹

Ambos nos sentamos en el tronco de un árbol y empezamos a calcular en pedacitos de papel ... Cuando las dos gotas se separan, se alejarán por repulsión eléctrica, sobre 200 MeV en total. Afortunadamente Lise Meitner recordó cómo calcular las masas de los núcleos ... y calculó que los dos núcleos formados ... serían más ligeros alrededor de un quinto de la masa de un protón ... según la fórmula de Einstein $E = mc^2$... la masa era justo equivalente a 200 MeV. ¡Todo encajaba!

Aunque $E = mc^2$ no era directamente responsable de la bomba atómica, fue uno de los grandes descubrimientos en física que llevó a una comprensión teórica efectiva de las reacciones nucleares. El papel más importante de Einstein en lo que a la bomba atómica se refiere fue político. Alentado por Leo Szilard, Einstein escribió al presidente Roosevelt advirtiendo que los nazis podrían estar desarrollando armas atómicas y explicándole su asombrosa potencia. Su reputación e influencia eran

³⁸ Ian Stewart y Jack Cohen. *Figments of Reality*, Cambridge University Press, Cambridge 1997, p. 37.

³⁹ http://en.wikipedia.org/wiki/Mass_Energy_equivalence

enormes y el presidente hizo caso de la advertencia. El Proyecto Manhattan, Hiroshima y Nagasaki, y la consiguiente Guerra Fría fueron solo algunas de las consecuencias.

Einstein no estaba satisfecho con la relatividad especial. Proporcionaba una teoría unificada del espacio, tiempo, materia y electromagnetismo, pero excluía un ingrediente fundamental.

La gravedad.

Einstein creía que «todas las leyes de la física» deben satisfacer su versión extendida del principio de relatividad de Galileo. La ley de la gravedad seguramente debería estar entre ellas. Pero eso no ocurría en la versión actual de la relatividad. La ley de la inversa del cuadrado de Newton no se transformaba correctamente entre sistemas de referencia. Así que Einstein decidió que tenía que cambiar la ley de Newton. Ya había cambiado prácticamente todo lo demás en el universo newtoniano, así que ¿por qué no?

Le llevó diez años. Su punto de partida fue averiguar las implicaciones del principio de relatividad para un observador moviéndose libremente bajo la influencia de la gravedad, por ejemplo, en un ascensor que cae libremente. Finalmente se dirigió a una formulación apropiada. En esto recibió la ayuda de un amigo cercano, el matemático Marcel Grossmann, que lo dirigió hacia un campo de las matemáticas que estaba creciendo rápidamente: la geometría diferencial. Este había sido desarrollado a partir del concepto de variedad de Riemann y su caracterización de curvatura, discutida en el capítulo 1. Ahí mencioné que la métrica de Riemann puede escribirse como una matriz 3×3 , y que técnicamente este es un tensor simétrico. Una escuela de matemáticos italianos, en particular Tullio Levi-Civita y Gregorio Ricci-Curbastro, retomaron las ideas de Riemann y las transformaron en el cálculo tensorial.

Desde 1912, Einstein estaba convencido de que la clave para una teoría relativista de la gravedad le requería reformular sus ideas usando el cálculo tensorial, pero en un espacio-tiempo tetradiimensional más que en un espacio tridimensional. Los matemáticos estaban felizmente siguiendo a Riemann y permitiendo cualquier número de dimensiones, de modo que ya habían establecido las cosas en una

generalidad más que suficiente. Para acortar la historia, finalmente obtuvo lo que ahora llamamos las ecuaciones del campo de Einstein, que escribió como:

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} = \kappa T_{\mu\nu}$$

Aquí R , g y T son tensores —cantidades que definen las propiedades físicas y se transforman según las reglas de la geometría diferencial— y κ es una constante. Los subíndices μ y ν repasan las cuatro coordenadas del espacio-tiempo, de manera que cada tensor es una tabla 4×4 de 16 números. Ambos son simétricos, lo que quiere decir que no cambian cuando μ y ν se intercambian, lo que reduce la lista a 10 números distintos. Así que la fórmula realmente esconde 10 ecuaciones, motivo por el que con frecuencia nos referimos a ellas usando el plural, comparable a las ecuaciones de Maxwell. R es la métrica de Riemann, define la forma del espacio-tiempo. g es el tensor de curvatura de Ricci, que es una modificación de la noción de curvatura de Riemann. Y T es el tensor de energía-momento, que describe cómo estas dos cantidades fundamentales dependen del suceso espacio-tiempo que nos ocupa. Einstein presentó su ecuación a la Academia de Ciencias prusiana en 1915. Llamó a su nuevo trabajo la teoría de la relatividad general.

Podemos interpretar las ecuaciones de Einstein geométricamente, y cuando lo hacemos, proporcionan una nueva aproximación a la gravedad. La innovación básica es que la gravedad no está representada como una fuerza, sino como la curvatura del espacio-tiempo. En ausencia de gravedad, el espacio-tiempo se reduce al espacio de Minkowski. La fórmula para el intervalo determina el tensor de curvatura correspondiente. Su interpretación es «no curvada», de la misma manera que el teorema de Pitágoras se aplica a un plano llano, pero no a un espacio no euclíadiano curvado positiva o negativamente. El espacio-tiempo de Minkowski es plano. Pero cuando aparece la gravedad, el espacio-tiempo se curva.

El modo habitual de imaginarlo es olvidarse del tiempo, bajar las dimensiones del espacio a dos y obtener algo como la figura 51 (izquierda). El plano llano del espacio(-tiempo) de Minkowski está distorsionado, mostrado aquí por una curva concreta que crea una depresión. Lejos de la estrella, la materia o la luz se

desplazan en una línea recta (punteada). Pero la curvatura provoca que la trayectoria se curve. De hecho, parece superficialmente como si alguna fuerza proveniente de la estrella atrajese la materia hacia ella. Pero no hay fuerza, solo espacio-tiempo combado. Sin embargo, esta imagen de la curvatura deforma el espacio a lo largo de una dimensión extra, que no se necesita matemáticamente. Una imagen alternativa es dibujar una rejilla de geodésicas, las trayectorias más cortas que distan lo mismo unas de otras según la métrica curva. Estas se amontonan donde la curvatura es mayor (figura 51, derecha).

Si la curvatura del espacio-tiempo es pequeña, esto es, si lo que (en la imagen antigua) pensamos como fuerzas gravitacionales no son demasiado grandes, entonces esta formulación nos lleva a la ley de la gravedad de Newton. Comparando las dos teorías, la constante de Einstein κ resulta ser $8\pi G/c^4$, donde G es la constante gravitacional de Newton.

Esto vincula la nueva teoría con la anterior y prueba que en la mayoría de los casos la nueva estará de acuerdo con la antigua. La nueva física interesante se da cuando esta ya no es cierta, cuando la gravedad es grande. Cuando Einstein propuso esta teoría, cualquier prueba de relatividad tenía que hacerse fuera del laboratorio, a gran escala. Lo que quiere decir astronomía.

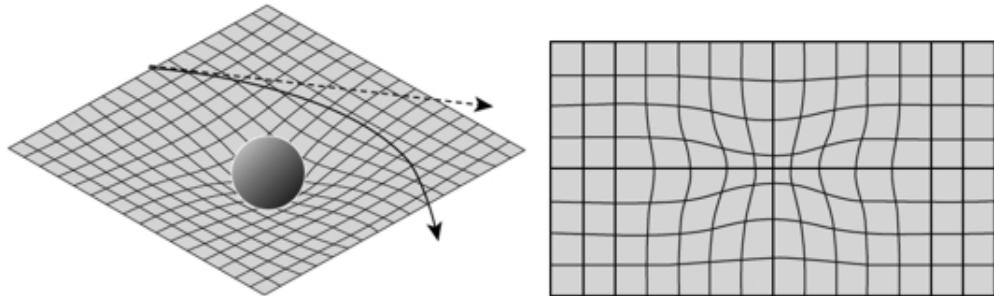


FIGURA 51. A la izquierda: espacio combado cerca de una estrella, y cómo curva la trayectoria de materia o luz que están pasando. A la derecha: imagen alternativa usando una rejilla de geodésicas, las cuales se amontonan en regiones de curvaturas más altas.

Einstein, por lo tanto, fue buscando peculiaridades inexplicables en el movimiento

de los planetas, efectos que no concordasen con Newton. Encontró una que podría ser adecuada: una característica confusa de la órbita de Mercurio, el planeta más cercano al Sol, sujeto a las mayores fuerzas gravitacionales, y en consecuencia, si Einstein tenía razón, dentro de una región de curvatura mayor.

Como todos los planetas, Mercurio sigue una trayectoria que es muy próxima a una elipse, de modo que algunos puntos en su órbita están más próximos al Sol que otros. Los más próximos de todos se llaman su perihelio («cerca del Sol» en griego). La localización exacta de este perihelio había sido observada durante muchos años, y había algo extraño en ella. El perihelio lentamente rotaba alrededor del Sol, un efecto llamado precesión; en efecto, el eje largo de la elipse orbital estaba lentamente cambiando de dirección. Eso estaba bien, las leyes de Newton lo predecían, porque Mercurio no es el único planeta del Sistema Solar y los otros planetas estaban lentamente cambiando su órbita. El problema era que los cálculos newtonianos daban la velocidad de precesión equivocada. Los ejes estaban rotando demasiado rápido.

Esto se sabía desde 1840, cuando François Arago, director del Observatorio de París, pidió a Urbain Le Verrier que calculase la órbita de Mercurio usando las leyes de movimiento y gravitación de Newton. Pero cuando los resultados se probaron observando el cronometraje exacto de un tránsito de Mercurio —un paso delante del Sol, visto desde la Tierra— vieron que estaban equivocados. Le Verrier decidió intentarlo de nuevo, eliminando las potenciales fuentes de error, y en 1859 publicó sus nuevos resultados. En el modelo newtoniano, la velocidad de precesión tenía un margen de error de un 0,7 %. La diferencia comparada con las observaciones era minúscula: 38 segundos de arco cada siglo (más tarde se modificó a 43 arcosegundos). No es mucho, menos que una diezmilésima de un grado por año, pero era suficiente para interesar a Le Verrier. En 1846 había construido su reputación analizando irregularidades en la órbita de Urano y prediciendo la existencia, y localización, de un planeta, por aquel entonces por descubrir: Neptuno. Ahora estaba esperando repetir la hazaña. Interpretó el movimiento inesperado del perihelio como la prueba de que algún mundo desconocido estaba alterando la órbita de Mercurio. Hizo los cálculos y predijo la existencia de un pequeño planeta con una órbita más próxima al Sol que la de Mercurio. Incluso tenía un nombre para

él: Vulcano, el dios romano del fuego.

Observar a Vulcano, si existía, sería difícil. El resplandor del Sol era un obstáculo, de modo que la mejor apuesta era coger a Vulcano en tránsito, donde hubiese un diminuto punto oscuro en el brillante disco solar. Poco después de la predicción de Le Verrier, un astrónomo aficionado llamado Edmond Lescarbault notificó al famoso astrónomo que acababa de ver justo eso. Inicialmente asumió que el punto debía ser una mancha solar, pero se movía a la velocidad equivocada. En 1860, Le Verrier anunció el descubrimiento de Vulcano a la Academia de Ciencias de París, y el gobierno premio a Lescarbault con la prestigiosa Legión de Honor.

En medio del clamor, algunos astrónomos no acababan de estar convencidos. Uno fue Emmanuel Liais, quien había estado estudiando el Sol con un equipo mucho mejor que el de Lescarbault. Su reputación estaba en peligro: había estado observando el Sol para el gobierno brasileño, habría sido vergonzoso no haber visto algo de tanta importancia. Él negó rotundamente que hubiese algún tránsito. Durante un tiempo, todo fue muy confuso. Los aficionados reclamaban repetidamente que habían visto a Vulcano, a veces, años antes de que Le Verrier anunciase su predicción. En 1878, James Watson, un profesional, y Lewis Swift, un aficionado, dijeron que habían visto un planeta como Vulcano durante un eclipse solar. Le Verrier había muerto un año antes, todavía convencido de que había descubierto un nuevo planeta cerca del Sol, pero sin sus nuevos cálculos de órbitas y predicciones de tránsitos entusiastas, ninguno de los cuales había sucedido, el interés en Vulcano rápidamente se desvaneció. Los astrónomos se hicieron escépticos.

En 1915, Einstein dio el tiro de gracia. Reanalizó el movimiento usando la relatividad general, sin asumir ningún planeta nuevo, y un cálculo sencillo y transparente le llevó a un valor de 43 segundos de arco para la precesión, la cifra exacta obtenida actualizando los cálculos originales de Le Verrier. Un moderno cálculo newtoniano predice una precesión de 5.560 arcosegundos por siglo, pero las observaciones dan 5.600. La diferencia es 40 segundos de arco, de modo que alrededor de 3 arcosegundos por siglo sigue sin aparecer. El anuncio de Einstein hizo dos cosas: fue visto como una confirmación de la relatividad, y en lo que a la

mayoría de los astrónomos se refería, relegaba a Vulcano al montón de desechos.⁴⁰ Otra verificación astronómica famosa de la relatividad general es la predicción de Einstein de que el Sol curva la luz. La gravitación newtoniana también predice esto, pero la relatividad general predice una cantidad de curvamiento que es dos veces mayor. El eclipse solar total de 1919 proporcionó una oportunidad para distinguir los dos, y Sir Arthur Eddington organizó una expedición, finalmente anunciando que Einstein se imponía. Esto fue aceptado con entusiasmo en la época, pero más tarde se hizo claro que los datos eran pobres y el resultado fue cuestionado. Observaciones independientes adicionales de 1922 parecían estar de acuerdo con la predicción relativista, como lo estuvo un reanálisis posterior de los datos de Eddington. En la década de los sesenta del siglo XX, se hizo posible hacer las observaciones para radiaciones de radiofrecuencia y, solo entonces, fue seguro que los datos sí que mostraban una desviación dos veces mayor que la predicha por Newton e igual a la que predijo Einstein.

Las predicciones más dramáticas de la relatividad general surgen en una escala mucho más grande: los agujeros negros, que nacen cuando una estrella de gran masa se colapsa bajo su propia gravitación, y el universo en expansión, actualmente explicado por el Big Bang.

Las soluciones para las ecuaciones de Einstein son geometrías del espacio-tiempo. Estas representarían el universo como un todo, o alguna parte de él, asumiendo que esté aislado gravitacionalmente de modo que el resto del universo no tiene un efecto importante. Esto es análogo a suposiciones newtonianas anteriores de que solo dos cuerpos están interaccionando, por ejemplo. Como las ecuaciones de campo de Einstein involucran a diez variables, las soluciones explícitas en términos de fórmulas matemáticas son raras. Hoy podemos solucionar las ecuaciones numéricamente, pero eso era una quimera antes de la década de los sesenta del siglo pasado, porque los ordenadores no existían o eran demasiado limitados para resultar útiles. El modo estándar de simplificar ecuaciones es invocar a la simetría.

⁴⁰ Unos pocos no lo vieron de ese modo. Henry Courten, reanalizando las fotografías del eclipse solar de 1970, informó de la existencia de al menos siete cuerpos muy diminutos en órbitas rodeando al Sol muy cercanas a él, quizás la evidencia de un cinturón de asteroides interior poco poblado. No se ha encontrado ninguna prueba concluyente de su existencia, y habría sido de menos de 60 kilómetros de ancho. Los objetos vistos en las fotografías podrían tan solo haber sido pequeños cometas o asteroides pasando en órbitas excéntricas. Fuesen lo que fuesen, no eran Vulcano.

Supón que las condiciones iniciales para el espacio-tiempo son simétricas esféricamente, esto es, todas las cantidades físicas dependen solo de la distancia al centro. Entonces el número de variables en cualquier modelo se reduce considerablemente. En 1916 el astrofísico alemán Karl Schwarzschild hizo esta suposición para las ecuaciones de Einstein y se las arregló para resolver las ecuaciones resultantes con una fórmula exacta, conocida como la métrica Schwarzschild. Su fórmula tiene una característica curiosa: una singularidad. La solución se hace infinita a una distancia concreta del centro, llamada radio de Schwarzschild. Al principio se asumía que esta singularidad era algún tipo de artefacto matemático, y su significado físico era tema de una discusión importante. Ahora la interpretamos como el horizonte de sucesos de un agujero negro.

Imagina una estrella con una masa tan grande que su radiación no puede contrarrestar su campo gravitacional. La estrella empezará a contraerse, absorbida por su propia masa. Cuanto más densa se hace, más fuerte es este efecto, de manera que la contracción se vuelve cada vez más rápida. La velocidad de escape de la estrella, la velocidad con la que un objeto se debe mover para escapar del campo gravitacional, también aumenta. La métrica de Schwarzschild nos dice que, en alguna etapa, la velocidad de escape se hace igual a la de la luz. Ahora nada puede escapar, porque nada puede viajar más rápido que la luz. La estrella se ha convertido en un agujero negro, y el radio de Schwarzschild nos indica la región de la cual nada puede escapar, limitado por el horizonte de sucesos del agujero negro. La física de los agujeros negros es compleja, y no hay espacio para hacerle justicia aquí. Basta con decir que la mayoría de los cosmólogos están ahora satisfechos con que la predicción es válida, que el universo contiene innumerables agujeros negros y, de hecho, que al menos uno merodea en el corazón de nuestra galaxia. En realidad, de la mayoría de las galaxias.

En 1917, Einstein aplicó sus ecuaciones a todo el universo, asumiendo otro tipo de simetría: homogeneidad. El universo debería tener el mismo aspecto (en una escala lo suficientemente grande) en todos los puntos en el espacio y el tiempo. Para entonces, había modificado las ecuaciones para incluir una «constante cosmológica» Λ , y averiguado el significado de la constante κ . Entonces escribió las ecuaciones así:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4} T_{\mu\nu}$$

Las soluciones tuvieron una implicación sorprendente: el universo debería contraerse a medida que el tiempo pasase. Esto forzó a Einstein a añadir el término que involucra a la constante cosmológica; él estaba buscando un universo sin cambios y estable y al ajustar la constante al valor correcto, podía evitar el modelo de universo que se contrae hasta un punto. En 1922, Alexander Friedmann encontró otra ecuación, que predecía que el universo debería expandirse y no necesitaba una constante cosmológica. También predijo la velocidad de expansión. Einstein todavía no estaba contento, quería que el universo fuese estable e inmutable.

Por una vez, la imaginación de Einstein le falló. En 1929, los astrónomos americanos Edwin Hubble y Milton Humason encontraron pruebas de que el universo se está expandiendo. Galaxias distantes se están alejando de nosotros, como muestran los cambios en la frecuencia de la luz que emiten, el famoso efecto Doppler, en el cual el sonido de una ambulancia que va muy rápido disminuye cuando se aleja, porque las ondas de sonido se ven afectadas por la velocidad relativa del emisor y el receptor. Ahora las ondas son electromagnéticas y la física es relativista, pero hay todavía un efecto Doppler. No solo galaxias distantes se alejan de nosotros, sino que cuanto más distantes están, más rápido se alejan.

Recorriendo la expansión hacia atrás en el tiempo, resulta que en algún punto en el pasado, todo el universo era esencialmente tan solo un punto. Antes de eso, no existía en absoluto. En ese punto primigenio, tanto el espacio como el tiempo se originaron en el famoso Big Bang, una teoría propuesta por el matemático francés Georges Lemaître en 1927, y casi universalmente ignorado. Cuando los radiotelescopios observaron la radiación del fondo de las microondas cosmológicas en 1964, a una temperatura en la que se ajustaban al modelo del Big Bang, los cosmólogos decidieron que Lemaître había estado en lo correcto después de todo. De nuevo, el tema se merece un libro propio, y mucho se ha escrito. Basta decir que nuestra teoría actual aceptada mayoritariamente es una elaboración del escenario del Big Bang.

El conocimiento científico, sin embargo, es siempre provisional. Nuevos descubrimientos pueden cambiarlo. El Big Bang ha sido el paradigma cosmológico aceptado durante los últimos 30 años, pero está empezando a mostrar algunas fisuras. Varios descubrimientos o bien arrojan serias dudas sobre la teoría, o bien necesitan nuevas partículas y fuerzas físicas que han sido deducidas pero no observadas. Hay tres fuentes de dificultad principales. Las resumiré primero y luego discutiré cada una con más detalle. La primera son las curvas de rotación galáctica, que sugieren que la mayoría de la materia en el universo está perdida. La propuesta actual es que esto es un signo de un nuevo tipo de materia, materia oscura, que constituye alrededor del 90 % de la materia en el universo, y es diferente a cualquier materia ya observada directamente sobre la Tierra. La segunda es una aceleración en la expansión del universo, que necesita una nueva fuerza, energía oscura, de origen desconocido pero modelada usando la constante cosmológica de Einstein. La tercera es una colección de asuntos teóricos relacionados con la teoría popular de la inflación cósmica, que explica por qué el universo observable es tan uniforme. La teoría encaja con las observaciones, pero su lógica interna parece débil.

La materia oscura primero. En 1938, el efecto Doppler se usaba para medir la velocidad de las galaxias en grupo, y el resultado no era consistente con la gravitación newtoniana. Como las galaxias están separadas grandes distancias, el espacio-tiempo es casi plano y la gravedad newtoniana es un buen modelo. Fritz Zwicky sugirió que debe haber alguna materia no observada para explicar la discrepancia, y se llamó materia oscura porque no podía ser vista en fotografías. En 1959, usando el efecto Doppler para medir la velocidad de rotación de las estrellas en la galaxia M33, Louise Volders descubrió que la curva de rotación observada, un trazo de la velocidad respecto a la distancia al centro, era también inconsistente con la gravitación newtoniana, que de nuevo es un buen modelo. La velocidad en lugar de decaer en grandes distancias, permaneció casi constante (figura 52). El mismo problema aparece para muchas otras galaxias.

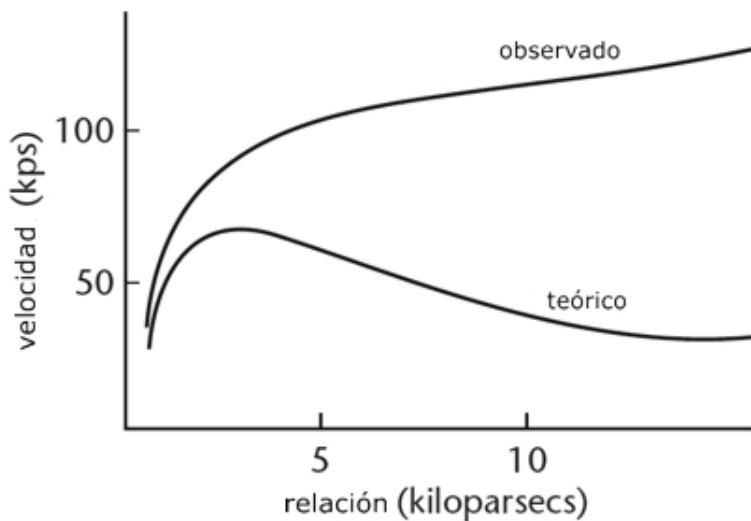


FIGURA 52. Curvas de rotación galácticas para M33: teoría y observaciones.

Si existe, la materia oscura debe ser diferente de la materia «bariónica» común, las partículas observadas en experimentos en la Tierra. Su existencia es aceptada por la mayoría de los cosmólogos, que argumentan que la materia oscura explica varias anomalías diferentes en las observaciones, no solo las curvas de rotación. Se han sugerido partículas candidatas, como WIMPs (de sus siglas en inglés *weakly interacting massive particles*, partículas masivas de interacción débil), pero hasta ahora estas partículas no se han detectado en experimentos. La distribución de la materia oscura alrededor de las galaxias se ha trazado asumiendo que la materia oscura existe y averiguando dónde tiene que estar para hacer las curvas de rotación planas. Generalmente parece formar dos esferas de proporciones galácticas, una sobre el plano de la galaxia y otra bajo él, como una mancuerna gigante. Esto es un poco como predecir la existencia de Neptuno a partir de discrepancias en la órbita de Urano, pero dichas predicciones requerían confirmación: hubo que encontrar a Neptuno.

La energía oscura se propone de modo similar para explicar los resultados de 1998 del High-Z Supernova Search Team, que esperaban encontrar pruebas de que la expansión del universo se está haciendo más lenta a medida que se va perdiendo el impulso inicial del Big Bang. En su lugar, las observaciones indicaron que la expansión del universo está acelerándose, un hallazgo confirmado por el Supernova

Cosmology Projects en 1999. Es como si alguna fuerza antigravedad dominase el espacio, empujando las galaxias lejos unas de otras a una velocidad cada vez mayor. Esta fuerza no es ninguna de las cuatro fuerzas básicas de la física: gravedad, electromagnetismo, fuerza nuclear fuerte, fuerza nuclear débil. Se llamó energía oscura. De nuevo, su existencia parece resolver otros problemas cosmológicos.

La inflación cósmica fue propuesta por el físico norteamericano Alan Guth en 1980 para explicar por qué el universo es extremadamente uniforme en sus propiedades físicas a escalas muy grandes. La teoría mostró que el Big Bang debería haber producido un universo que fuese mucho más curvo. Guth sugirió que un «campo inflatón» (no hay una errata, no es de inflación, está pensado para ser un campo cuántico escalar que se corresponda con una partícula hipotética, el inflatón) provocó que el primer universo se expandiese con una rapidez extrema. Entre 10^{-36} y 10^{-32} segundos después del Big Bang, el volumen del universo se multiplicó por el alucinante factor de 10^{78} . El campo inflatón no se ha observado (esto requeriría energías inviablemente altas), pero la inflación cósmica explica muchas características del universo, y encaja con las observaciones tan bien, que la mayoría de los cosmólogos están convencidos de que existe.

No es sorprendente que la materia oscura, la energía oscura y la inflación cósmica sean populares entre los cosmólogos, porque les permiten seguir usando sus modelos físicos favoritos, y los resultados están acorde con las observaciones. Pero las cosas están empezando a desmoronarse.

Las distribuciones de la materia oscura no proporcionan una explicación satisfactoria de las curvas de rotación. Se necesitan cantidades enormes de materia oscura para mantener la curva de rotación plana fuera de las grandes distancias observadas. La materia oscura tiene que tener un momento angular tan grande que es poco realista, lo que es inconsistente con las teorías comunes de formación de la galaxia. La misma distribución inicial bastante especial de materia oscura se necesita en cada galaxia, lo cual parece improbable. La forma de mancuerna es inestable porque coloca la masa adicional en la parte externa de la galaxia.

A la energía oscura le va mejor, y está pensada para ser algún tipo de energía del vacío mecánico-cuántica, que surge a partir de fluctuaciones en el vacío. Sin

embargo, los cálculos actuales del tamaño de la energía del vacío son del orden de 10122 veces más grande, lo que son malas noticias incluso por los estándares de la cosmología.⁴¹

Los principales problemas que afectan a la inflación cósmica no son observaciones, ya que encaja con estas sorprendentemente bien, sino sus fundamentos lógicos. La mayoría de los escenarios de la inflación llevarían a un universo que difiere considerablemente del nuestro, lo que cuentan son las condiciones iniciales en el momento del Big Bang. Con el propósito de ajustarse a las observaciones, la inflación necesita que el estado temprano del universo sea muy especial. Sin embargo, hay también condiciones iniciales muy especiales que producen un universo justo como el nuestro sin invocar a la inflación cósmica. Aunque ambos conjuntos de condiciones son extremadamente raros, los cálculos realizados por Roger Penrose⁴² muestran que las condiciones iniciales que no necesitan la inflación cósmica superan en número a las que produce la inflación por un factor de un gúgolplex, diez elevado a 10 elevado a 100. De modo que explicar el actual estado del universo sin inflación cósmica sería mucho más convincente que explicarlo con ella.

Los cálculos de Penrose se apoyan en la termodinámica, que podría no ser un modelo apropiado, pero una aproximación alternativa, llevada a cabo por Gary Gibbons y Neil Turok, lleva a la misma conclusión. Esto es «desenrollar» el universo de vuelta a su estado inicial. Resulta que la mayoría de los estados iniciales potenciales no involucran un período de inflación, y aquellos que lo necesitan son una proporción extremadamente pequeña. Pero el mayor problema de todos es que cuando la inflación cósmica se casa con la mecánica cuántica, predice que las fluctuaciones cuánticas desencadenarán ocasionalmente la inflación cósmica en una pequeña región de un universo aparentemente estable. Aunque dichas fluctuaciones son raras, la inflación es tan rápida y tan enorme que el resultado neto son islas diminutas de espacio-tiempo normal rodeadas por regiones en constante crecimiento con inflación fuera de control. En esas regiones, las constantes

⁴¹ La energía del vacío en un centímetro cúbico de espacio libre se estima que es 10^{-15} julios. Según la electrodinámica cuántica debería ser en teoría 10^{107} julios, un error del orden de 10122.

http://en.wikipedia.org/wiki/Vacuum_energy

⁴² El trabajo de Penrose se presenta en Paul Davies. *The Mind of God*, Simon & Schuster, Nueva York 1992.

fundamentales de la física pueden ser diferentes de sus valores en nuestro universo. De hecho, cualquier cosa es posible. ¿Puede una teoría que predice «cualquier cosa» ser probada científicamente?

Hay alternativas, y se está empezando a analizarlas como si necesitasen ser tomadas en serio. La materia oscura podría no ser otro Neptuno, sino otro Vulcano, un intento de explicar una anomalía gravitacional invocando una materia nueva, cuando lo que realmente necesita cambiarse es la ley de la gravitación.

La principal propuesta bien desarrollada es MOND, de las siglas en inglés de dinámica newtoniana modificada, *modified Newtonian dynamics*, propuesta por el físico israelí Mordehai Milgrom en 1983. Esto, de hecho, no modifica la ley de la gravedad, sino la segunda ley de movimiento de Newton. Asume que la aceleración no es proporcional a la fuerza cuando la aceleración es muy pequeña. Hay una tendencia entre los cosmólogos a asumir que las únicas teorías alternativas viables son la materia oscura o la MOND, de modo que si MOND no concuerda con las observaciones, eso deja sola a la materia oscura. Sin embargo, hay muchos modos potenciales de modificar la ley de la gravedad, y es improbable que encontremos el correcto de manera inmediata. El fallecimiento de MOND se ha proclamado varias veces, pero en investigaciones adicionales no se ha encontrado un defecto decisivo todavía. El principal problema con MOND, a mi entender, es que pone en sus ecuaciones lo que espera obtener, es como Einstein modificando la ley de Newton para cambiar la fórmula cerca de una masa grande. En su lugar, encontró un modo radicalmente nuevo de pensar en la gravedad, la curvatura del espacio-tiempo.

Incluso si mantenemos la relatividad general y su aproximación newtoniana, podría no haber necesidad de la energía oscura. En 2009, usando las matemáticas de ondas de choque, los matemáticos norteamericanos Joel Smoller y Blake Temple mostraron que hay soluciones para las ecuaciones del campo de Einstein en las que la métrica se expande a un ritmo que se va acelerando.⁴³ Estas soluciones muestran que cambios pequeños en el modelo estándar podrían explicar la aceleración de galaxias observada sin invocar a la energía oscura.

⁴³ Joel Smoller y Blake Temple. «A one parameter family of expanding wave solutions of the Einstein equations that induces an anomalous acceleration into the standard model of cosmology», <http://arxiv.org/abs/0901.1639>

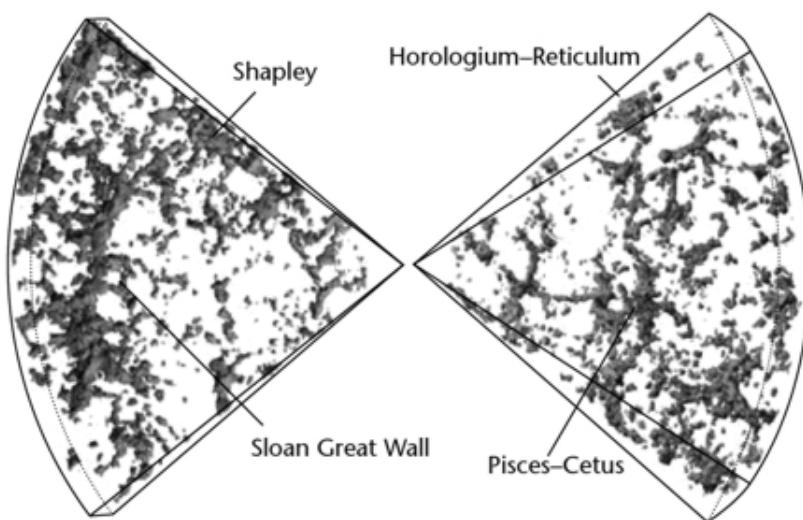


FIGURA 53. La rugosidad del universo.

Los modelos del universo de relatividad general asumen que se forma una variedad, esto es, en escalas muy grandes la estructura se alisa. Sin embargo, la distribución de la materia observada en el universo a escalas muy grandes es a pedazos, como la Gran Muralla Sloan, un filamento compuesto de galaxias de 1.370 millones de años luz de largo (figura 53). Los cosmólogos creen que en escalas todavía mayores el alisamiento se hará aparente, pero hasta la fecha, cada vez que el rango de observaciones se ha extendido, la forma rugosa ha persistido.

Robert MacKay y Colin Rourke, dos matemáticos británicos, han argumentado que un universo rugoso en el cual hay muchas fuentes de gran curvatura locales podría explicar todos los misterios cosmológicos.⁴⁴ Dicha estructura está más cercana a lo que se observa que cualquier alisamiento a gran escala y es consistente con el principio general de que el universo debería ser parecido en todas partes. En dicho universo no habría necesidad de ningún Big Bang, de hecho, todo podría estar en un estado estable y ser mucho, mucho, mucho más viejo que la cifra actual de 13.800 millones de años. Las galaxias individuales irían a través de un ciclo de vida, sobreviviendo relativamente invariables durante alrededor de 10^{16} años. Tendrían un enorme agujero negro central. Las curvas de rotación galácticas serían planas debido al arrastre de la inercia, una consecuencia de la relatividad general en la que

⁴⁴ R.S. MacKay y C.P. Rourke, «A new paradigm for the universe, borrador», Universidad de Warwick, 2011. Para más detalles, véanse los artículos listados en <http://msp.warwick.ac.uk/~cpr/paradigm/>

un cuerpo enorme que está rotando arrastra con él el espacio-tiempo alrededor de su entorno. El corrimiento al rojo observado en cuásares estaría provocado por un gran campo gravitacional, no por el efecto Doppler, y no sería indicativo de un universo en expansión. Esta teoría ha avanzado mucho gracias al astrónomo norteamericano Halton Arp, y nunca se ha refutado satisfactoriamente. El modelo alternativo incluso indica una temperatura de 5 K para el fondo de microondas cosmológicas, la principal evidencia (aparte del corrimiento al rojo interpretado como expansión) para el Big Bang.

MacKay y Rourke dicen que su propuesta «anula prácticamente todo principio de la cosmología actual. Sin embargo, no contradice cualquiera de las pruebas experimentales». Bien podría estar equivocada, pero el punto fascinante es que puedes mantener las ecuaciones del campo de Einstein sin cambios, prescindiendo de la materia oscura, la energía oscura y la inflación cósmica, y todavía obtiene un comportamiento razonable como todas esas observaciones misteriosas. De modo que cualquiera que sea el destino de la teoría, sugiere que los cosmólogos deberían considerar modelos matemáticos más imaginativos antes de recurrir a físicas nuevas aunque sin apoyo. La materia oscura, la energía oscura, la inflación, cada una de ellas necesita de físicas radicalmente nuevas que nadie ha observado... En ciencia, un *deus ex machina* genera escepticismo. Tres serían considerados intolerables en cualquier otra rama que no fuese la cosmología. Para ser justos, es difícil hacer experimentos sobre todo el universo, de modo que teorías que encajen con las observaciones haciendo especulaciones son más o menos todo lo que puede hacerse. Pero imagina qué sucedería si un biólogo explicase la vida con algún «campo de vida» que no se puede observar, sin mencionar sugerir que un nuevo tipo de «material vital» y un nuevo tipo de «energía vital» fuesen también necesarias, mientras no proporciona pruebas de que alguna de ellas existe.

Dejando a un lado el reino desconcertante de la cosmología, ahora hay maneras más caseras de comprobar la relatividad, tanto la especial como la general, a una escala humana. La relatividad especial puede comprobarse en el laboratorio, y técnicas de medición modernas proporcionan una precisión exquisita. Aceleradores de partículas como el Gran Colisionador de Hadrones simplemente no funcionarían a menos que los diseñadores tuvieran la relatividad especial en cuenta, porque las

partículas que dan vueltas en estas máquinas lo hacen a velocidades muy próximas a la de la luz. La mayoría de las pruebas de la relatividad general son todavía astronómicas, y van desde lentes gravitacionales a dinámica de púlsares, con un nivel de precisión alto. Un experimento reciente de la NASA en una órbita baja terrestre, usando giroscopios de alta precisión, confirmó la existencia del efecto de arrastre por inercia, pero fracasó en alcanzar la precisión deseada debido a unos efectos electrostáticos inesperados. Cuando se corrigieron los datos para este problema, otros experimentos ya habían logrado los mismos resultados.

Sin embargo, un ejemplo de dinámica relativista, tanto especial como general, está más cerca de nosotros: la navegación vía satélite de los coches. Los sistemas de navegación por satélite usados por conductores calculan la posición de los coches usando señales de una red de 24 satélites en órbita, el Sistema de Posicionamiento Global o GPS. El GPS es sorprendentemente preciso, y funciona porque la electrónica moderna puede manejar con seguridad y medir instantes de tiempo muy diminutos. Está basado en señales de tiempo muy precisas, pulsaciones emitidas por los satélites y recogidas en la tierra. Comparando las señales desde varios satélites se triangula la localización del receptor con un margen de error de unos pocos metros. Este nivel de precisión requiere saber el tiempo con un margen de error de 25 nanosegundos. La dinámica newtoniana no da localizaciones correctas, porque dos efectos que no se tienen en cuenta en las ecuaciones de Newton alteran el flujo del tiempo: el movimiento de los satélites y el campo gravitatorio de la Tierra.

La relatividad especial aborda el movimiento y predice que los relojes atómicos de los satélites deberían perder 7 microsegundos por día comparados con los relojes de la tierra, debido a la dilatación del tiempo relativista. La relatividad general predice una ganancia de 45 microsegundos por día provocados por la gravedad terrestre. El resultado neto es que los relojes en los satélites ganan 38 microsegundos por día debido a razones relativistas. Por pequeño que esto pueda parecer, su efecto en las señales de GPS no es de ninguna manera insignificante. Un error de 38 microsegundos es 38.000 nanosegundos, alrededor de 1.500 veces el error que el GPS puede tolerar. Si el software calculó la localización de tu coche usando la dinámica newtoniana, tu navegación por satélite será inservible rápidamente,

porque el error crecería a una velocidad de 10 kilómetros por día. Si contamos 10 minutos a partir de ahora, el GPS newtoniano te situaría en la calle equivocada, y mañana en la ciudad equivocada. En una semana estarías en la comunidad equivocada y en un mes en el país equivocado. Dentro de un año, estarías en el planeta equivocado. Si no crees en la relatividad, pero usas navegación por satélite para planear tus viajes, tienes que dar algunas explicaciones.

Capítulo 14
Rareza cuántica
Ecuación de Schrödinger

$$i\hbar \frac{\partial}{\partial t} \Psi = \hat{H} \Psi$$

raíz cuadrada de menos 1 tasa de cambio función onda cuántica
 constante de Planck, dividida por 2π respecto al tiempo operador Hamiltoniano

¿Qué dice?

La ecuación modela la materia no como una partícula, sino como una onda, y describe cómo estas ondas se propagan.

¿Por qué es importante?

La ecuación de Schrödinger es fundamental para la mecánica cuántica, que junto con la relatividad general constituyen en la actualidad las teorías más efectivas del universo físico.

¿Qué provocó?

Una revisión radical de la física del mundo a escalas muy pequeñas, en las cuales cada objeto tiene una «función de onda» que describe una nube de probabilidad de posibles estados. A este nivel el mundo es incierto intrínsecamente. Intentos de relacionar el mundo microscópico cuántico con nuestro mundo macroscópico clásico llevaron a temas filosóficos que todavía tienen eco. Pero experimentalmente, la teoría cuántica funciona maravillosamente bien y los láseres y chips de los ordenadores actuales no funcionarían sin ella.

En 1900, el gran físico Lord Kelvin declaró que la entonces actual teoría del calor y la luz, que se consideraba que era una descripción casi completa de la naturaleza,

estaba «oscurecida por dos nubes. La primera tiene que ver con la pregunta: ¿cómo podría la Tierra moverse a través de un sólido elástico, como el que en esencia es el éter lumínico? La segunda es la doctrina de Maxwell-Boltzmann con respecto a la partición de la energía». El instinto de Kelvin para un problema importante era certero. En el capítulo 13, vimos cómo la primera pregunta llevaba, y era resuelta, a la relatividad. Ahora veremos cómo la segunda llevó al otro gran pilar de la física de nuestros días, la teoría cuántica.

El mundo cuántico es notablemente raro. Muchos físicos sienten que si no aprecias cómo de raro es, no lo aprecias en absoluto. Se ha dicho mucho con respecto a esa opinión, porque el mundo cuántico es tan diferente de nuestro mundo a una cómoda escala humana, que incluso los conceptos más simples cambian tanto que resultan irreconocibles. Es, por ejemplo, un mundo en el que la luz es tanto una partícula como una onda. Es un mundo donde un gato en una caja puede estar vivo y muerto al mismo tiempo... hasta que abres la caja, esto es, cuando de repente la función de onda del desafortunado animal «choca» con un estado u otro. En el multiverso cuántico, existe una copia de nuestro universo en el cual Hitler pierde la Segunda Guerra Mundial, y otra en el cual la gana. Lo que ocurre es que nosotros vivimos, esto es, existimos como funciones de onda cuánticas, en el primero. Otra versión de nosotros, igual de real, pero inaccesible a nuestros sentidos, vive en el otro.

La mecánica cuántica es definitivamente rara. Aunque si es *tan* rara es tema aparte. Todo empezó con bombillas. Esto era adecuado, porque era una de las aplicaciones más espectaculares que surgía de las materias florecientes de la electricidad y el magnetismo, que Maxwell tan brillantemente unificó. En 1894 un físico alemán llamado Max Planck fue contratado por una compañía eléctrica para diseñar la bombilla más eficiente que fuese posible, una que diese la máxima luz consumiendo la menor energía eléctrica. Vio que la clave de esta cuestión era un asunto fundamental en la física, planteado en 1859 por otro físico alemán, Gustav Kirchhoff. Incumbe a una construcción teórica conocida como un cuerpo negro, que absorbe toda la radiación electromagnética con la que se encuentra. La gran pregunta era: ¿cómo dicho cuerpo emite radiación? No puede almacenarla toda, alguna tiene que volver a salir fuera. En concreto, ¿cómo la intensidad de la radiación emitida depende de su frecuencia y la temperatura del cuerpo?

Ya había una respuesta de la termodinámica, en la cual un cuerpo negro puede modelarse como una caja cuyas paredes son espejos perfectos. La radiación electromagnética rebota de un lado a otro, reflejada por los espejos. ¿Cómo está distribuida la energía en la caja entre las diferentes frecuencias cuando el sistema ha establecido un estado de equilibrio? En 1876, Boltzmann probó el «teorema de equipartición»: la energía es distribuida igualmente en cada componente independiente del movimiento. Estas componentes son justo como las ondas básicas en una cuerda de un violín: modos normales.

Había solo un problema con esta respuesta: no era posible que fuese correcta. Implicaba que el total de energía radiada por todas las frecuencias debe ser infinita. Esta conclusión paradójica se hizo conocida como la catástrofe ultravioleta; ultravioleta porque eso era el principio del rango de alta frecuencia, y catástrofe porque lo era. Ningún cuerpo real puede emitir una cantidad infinita de energía. Aunque Planck era consciente de este problema, no le molestaba, porque, de todos modos, no creía en el teorema de la equipartición. Irónicamente, su trabajo resolvió la paradoja y acabó con la catástrofe ultravioleta, pero solo se dio cuenta de esto más tarde. Usó observaciones experimentales de cómo la energía depende de la frecuencia, y adecuó una fórmula matemática a los datos. Su fórmula, obtenida a principios de 1900, inicialmente no tenía ninguna bases físicas. Era tan solo una fórmula que funcionaba. Pero más tarde el mismo año, intentó conciliar su fórmula con la de la termodinámica clásica y decidió que los niveles de energía de los modos de vibración del cuerpo negro no podían formar un continuo, como la termodinámica asumía. En su lugar, estos niveles tenían que ser discretos, separados por huecos minúsculos. De hecho, para cualquier frecuencia dada, la energía tenía que ser un múltiplo entero de esa frecuencia, multiplicado por una constante muy pequeña. Ahora llamamos a este número la constante de Planck y la representamos como h . Su valor, en unidades de julios por segundo, es $6,62606957(29) \times 10^{-34}$, donde las cifras entre paréntesis podrían ser incorrectas. Este valor se deduce a partir de la relación teórica entre la constante de Planck y otras cantidades que son más fáciles de medir. La primera de dichas mediciones fue hecha por Robert Millikan usando el efecto fotoeléctrico, descrito más abajo. Los diminutos paquetes de energía que ahora llamamos cuantos, del latín *quantus*,

«cuánto».

La constante de Planck puede ser pequeñísima, pero si el conjunto de niveles de energía para una frecuencia dada es discreto, la energía total resulta ser finita. De modo que la catástrofe ultravioleta era un signo de que un modelo continuo fracasaba en el intento de reflejar la naturaleza. Y eso implicaba que la naturaleza, a escalas muy pequeñas, debe ser discreta. Inicialmente esto no se le ocurrió a Planck; él pensaba en sus niveles de energía discretos como un truco matemático para obtener una fórmula práctica. De hecho, Boltzmann había considerado una idea parecida en 1877, pero no llegó a ningún lado con ella. Todo cambió cuando Einstein puso en marcha su fértil imaginación, y la física entró en un nuevo reino. En 1905, el mismo año en que hizo su trabajo en la relatividad especial, investigó el efecto fotoeléctrico, en el cual la luz que golpea un metal adecuado lo provoca para emitir electrones. Tres años antes Philipp Lenard había notado que cuando la luz tiene una frecuencia mayor, los electrones tienen energías más altas. Pero la teoría de ondas de la luz, ampliamente confirmada por Maxwell, implica que la energía de los electrones debería depender de la intensidad de la luz, no de su frecuencia. Einstein se dio cuenta de que los cuantos de Planck explicarían la discrepancia. Sugirió que la luz, más que ser una onda, estaba compuesta de partículas diminutas, ahora llamadas fotones. La energía en un único fotón, de una frecuencia dada, debería ser la frecuencia multiplicada por la constante de Planck, justo como uno de los cuantos de Planck. Un fotón era un cuanto de luz.

Hay un problema obvio con la teoría de Einstein del efecto fotoeléctrico: asume que la luz es una partícula. Pero había pruebas abundantes de que la luz era una onda. Por otro lado, el efecto fotoeléctrico era incompatible con la luz siendo una onda. Así que, ¿era la luz una onda o una partícula?

Sí.

Era, o tenía aspectos que se manifestaban como ambas. En algunos experimentos, la luz parecía comportarse como una onda. En otros, se comportaba como una partícula. A medida que los físicos iban comprendiendo las escalas muy pequeñas del universo, decidieron que la luz no era la única cosa que tenía esta naturaleza dual extraña, a veces partícula, a veces onda. Toda la materia la tenía. Le llamaron la dualidad onda corpúsculo. La primera persona en captar esta naturaleza dual de

la materia fue Louis-Victor de Broglie, en 1924. Reformuló las leyes de Planck en términos no de energía, sino de momento, y propuso que el momento del punto de vista como partícula y la frecuencia del punto de vista como onda deberían estar relacionados; multiplica uno por otro y obtienes la constante de Planck. Tres años más tarde, se probó que tenía razón, al menos para los electrones. Por un lado, los electrones son partículas, y pueden observarse comportándose de ese modo. Por otro lado, se difractan como ondas. En 1988, los átomos de sodio también se pudieron ver comportándose como una onda.

La materia no era ni una partícula ni una onda, sino un poco de ambas, un «ondúsculo».

Se concibieron varias imágenes más o menos intuitivas de esta naturaleza dual de la materia. En una, una partícula es un montón de ondas localizadas, conocido como un paquete de ondas (figura 54). El paquete es un todo que puede comportarse como una partícula, pero algunos experimentos pueden probar su estructura interna como de onda. La atención cambió de proporcionar imágenes para los «ondúsculos» a averiguar cómo se comportaban. La búsqueda rápidamente alcanzó este objetivo, y apareció la ecuación central de la teoría cuántica.

La ecuación lleva el nombre de Erwin Schrödinger. En 1927, ampliando el trabajo de varios físicos, entre los que destaca Werner Heisenberg, Schrödinger escribió una ecuación diferencial para cualquier función de onda cuántica. Tenía el siguiente aspecto:

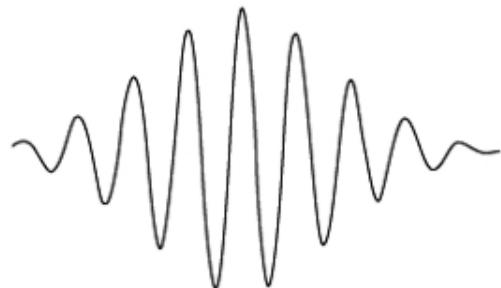


FIGURA 54. Paquete de ondas.

$$\textcolor{blue}{i\hbar} \frac{\partial \Psi}{\partial t} = \hat{H}\Psi$$

Aquí Ψ (la letra mayúscula griega psi) es la forma de la onda, t es el tiempo (de modo que $\partial/\partial t$ aplicado a Ψ da su tasa de variación con respecto al tiempo), \hat{H} es una expresión llamada el operador Hamiltoniano, y \hbar es $h/2\pi$, donde h es la

constante de Planck. ¿E i? Esa era la característica más rara de todas. Es la raíz cuadrada de menos uno (capítulo 5). La ecuación de Schrödinger se aplica a ondas definidas en los números complejos, no solo los números reales como ocurre en la ecuación de onda común.

¿Ondas en qué? La ecuación de onda clásica (capítulo 8) define ondas en el espacio, y su solución es una función numérica del espacio y el tiempo. Lo mismo aplica para la ecuación de Schrödinger, pero ahora la función de onda Ψ toma valores complejos, no solo reales. Es un poco como una ola del mar cuya altura es $2 + 3i$. La aparición de i es de muchas maneras la característica más misteriosa y profunda de la mecánica cuántica. Previamente i había aparecido en las soluciones de ecuaciones, y en métodos para encontrar esas soluciones, pero aquí era parte de la ecuación, una característica explícita de la ley física.

Un modo de interpretar esto es que las ondas cuánticas están vinculadas a pares de ondas reales, como si mis olas complejas fuesen realmente dos olas, una de altura 2 y la otra de altura 3, con las dos direcciones formando un ángulo recto. Pero no es tan sencillo, porque las dos olas no tienen una forma fija. A medida que el tiempo pasa, se dan cíclicamente a través de toda una serie de formas, y cada una está misteriosamente vinculada a la otra. Es un poco como los componentes eléctricos y magnéticos de una onda de luz, pero ahora la electricidad puede y de hecho «rota» en magnetismo, y a la inversa. Las dos ondas son dos caras de una forma única, que da vueltas uniformemente alrededor de la circunferencia goniométrica en el plano complejo. Tanto la parte real como la imaginaria de esta forma rotatoria cambian de un modo muy concreto: se combinan en cantidades que varían sinusoidalmente. Matemáticamente esto lleva a la idea de que una función de onda cuántica tiene un tipo especial de fase. La interpretación física de la fase es similar, pero se diferencia, al papel de la fase en la ecuación de onda clásica.

¿Recuerdas cómo los trucos de Fourier solucionaron tanto la ecuación del calor como la ecuación de onda? Algunas soluciones especiales, los senos y cosenos de Fourier, tienen propiedades matemáticas especialmente agradables. Todas las otras soluciones, aunque complicadas, son superposiciones de estos modos normales. Podemos resolver la ecuación de Schrödinger usando una idea parecida, pero ahora los patrones básicos son más complicados que senos y cosenos. Se llaman

autofunciones y pueden distinguirse de todas las otras soluciones. En lugar de ser alguna función general del espacio y el tiempo, una autofunción es una función definida solo en el espacio, multiplicada por una que solo depende del tiempo. Las variables del espacio y el tiempo, en la jerga, son separables. Las autofunciones dependen del operador hamiltoniano, que es una descripción matemática del sistema físico que nos ocupa. Sistemas diferentes —un electrón en un pozo de potencial, un par de fotones que chocan, lo que sea— tienen diferentes operadores hamiltonianos, por tanto, diferentes autofunciones.

Para simplificar, considera una onda estacionaria para la ecuación de onda clásica, una cuerda de violín vibrando, cuyos extremos están inmovilizados. En todos los instantes de tiempo, la forma de la cuerda es casi la misma, pero la amplitud se modula: multiplicada por un factor que varía sinusoidalmente con el tiempo, como en la figura 35 (página 177). La fase compleja de la función de onda cuántica es parecida, pero más difícil de visualizar. Para cualquier autofunción individual, el efecto de la fase cuántica es solo un desplazamiento de la coordenada del tiempo. Para una superposición de varias autofunciones, divide la función de onda en estas componentes, ten en cuenta cada una como una parte puramente espacial multiplicada por una parte puramente temporal, gira la parte temporal alrededor de la circunferencia goniométrica en el plano complejo a la velocidad adecuada y pon juntas las piezas. Cada autofunción separada tiene una amplitud compleja, y esto se modula en su propia frecuencia particular.

Puede sonar complicado, pero sería completamente incomprendible si no divides la función de onda en autofunciones. Al menos de este modo hay alguna opción.

A pesar de las complejidades, la mecánica cuántica sería solo una versión extravagante de la ecuación de onda clásica, que resultaría ser dos ondas en vez de una, si no fuera por un giro desconcertante. Puedes observar ondas clásicas y ver de qué forma son, incluso si son superposiciones de varios modos de Fourier. Pero en mecánica cuántica, nunca puedes observar la función de onda entera. Todo lo que puedes observar para una ocasión dada es una única componente de sus autofunciones. En términos generales, si intentas medir dos de estas componentes a la vez, el proceso de medición en una de ellas molesta al de la otra.

Esto inmediatamente presenta un asunto filosófico difícil. Si no puedes observar la

función de onda por completo, ¿existe realmente? ¿Es un objeto físico genuino, o solo una ficción matemática conveniente? ¿Es una cantidad no observable científicamente coherente? Es aquí cuando el famoso gato de Schrödinger aparece en la historia. Surge debido a un modo estándar de interpretar lo que es una medición cuántica, llamada interpretación de Copenhague.⁴⁵

Imagina un sistema cuántico en algún estado superpuesto, por ejemplo, un electrón cuyo estado es una mezcla de espín arriba y espín abajo, que son estados puros definidos por autofunciones. (No importa lo que espín arriba y espín abajo significan.) Sin embargo, cuando observas el estado, lo que tienes es o el espín arriba o el espín abajo. No puedes observar la superposición. Además, una vez has observados uno de ellos, por ejemplo el espín arriba, eso se convierte en el estado real del electrón. De algún modo tu medición parece haber forzado a la superposición a cambiar a una concreta de sus autofunciones. Esta interpretación de Copenhague toma esta afirmación literalmente: tu proceso de medición ha colapsado la función de onda original en una única autofunción pura.

Si observas muchos electrones, a veces tienes un espín arriba, otras veces un espín abajo. Puedes deducir la probabilidad de que el electrón esté en uno de esos estados. De modo que la propia función de onda puede interpretarse como un tipo de nube de probabilidad. No muestra el estado real del electrón, muestra cuán probable es que cuando lo midas, obtengas un resultado concreto. Pero eso lo hace un patrón estadístico, no una cosa real. No prueba más que la función de onda es real de lo que las mediciones de Quetelet de la altura humana prueban que el desarrollo de un embrión posee algún tipo de campana de Gauss.

La interpretación de Copenhague es sencilla, refleja qué sucede en los experimentos, y no hace suposiciones detalladas sobre qué sucede cuando observas un sistema cuántico. Por estas razones, la mayoría de los físicos activos están muy contentos de usarlo. Pero algunos no lo estaban al principio cuando la teoría estaba siendo discutida exhaustivamente y algunos todavía no lo están. Y uno de los

⁴⁵ La interpretación de Copenhague se dice habitualmente que surgió de las discusiones entre Niels Bohr, Werner Heisenberg, Max Born, y otros, en la mitad de la segunda década del siglo XX. Adquirió el nombre porque Bohr era danés, pero ninguno de los físicos involucrados usó el término en esa época. Don Howard ha sugerido que el nombre, y el punto de vista que engloba, aparece por primera vez a mediados de ese siglo, probablemente a través de Heisenberg. Véase D. Howard. «Who Invented the "Copenhagen Interpretation"? A Study in Mythology», *Philosophy of Science* 71(2004) 669-682.

disidentes era el propio Schrödinger.

En 1935, Schrödinger estaba preocupado por la interpretación de Copenhague. Podía ver que funcionaba, a un nivel paradigmático, para sistemas cuánticos como los electrones y los fotones. Pero el mundo alrededor de él, incluso aunque en lo más profundo de su interior era justo una masa furiosa de partículas cuánticas, parecía diferente. Buscando un modo de hacer la diferencia tan manifiesta como pudiese, Schrödinger se inventó una disquisición teórica en la que una partícula cuántica tenía un efecto dramático y obvio en un gato.

Imagina una caja, la cual cuando está cerrada es impermeable a todas las interacciones cuánticas. Dentro de ella, pon un átomo de materia radiactiva, un detector de radiaciones, un frasco de veneno y un gato vivo. Ahora cierra la caja y espera. En algún punto el átomo radiactivo se descompondrá y emitirá una partícula de radiación. El detector la localizará y está amañado para que cuando eso suceda, provoque que el frasco se rompa y libere el veneno dentro. Esto mata al gato.

En mecánica cuántica, la descomposición de un átomo radiactivo es un suceso aleatorio. Desde fuera, ningún observador puede decir si el átomo se ha desintegrado o no. Si lo ha hecho, el gato está muerto, si no, está vivo. Según la interpretación de Copenhague, hasta que alguien observe el átomo, es una superposición de dos estados cuánticos: desintegrado y no desintegrado. Lo mismo aplica para los estados del detector, el tarro y el gato. De modo que el gato está en una superposición de dos estados: muerto y vivo.

Como la caja es impermeable a todas las interacciones cuánticas, el único modo de saber si el átomo se ha desintegrado y matado al gato es abrir la caja. La interpretación de Copenhague nos dice que en el momento en que hagamos esto, las funciones de onda se colapsan y el gato de repente pasa a un estado puro: o muerto o vivo. Sin embargo, dentro de la caja no hay diferencia con el mundo exterior, donde nunca observamos un gato que está en un estado superpuesto vivo/muerto. De modo que antes de que abramos la caja y observemos su contenido, debe haber dentro bien un gato muerto o bien uno vivo.

Schrödinger pensó en esta disquisición teórica como una crítica de la interpretación de Copenhague. Los sistemas cuánticos microscópicos obedecen el principio de superposición y pueden existir en estados mixtos, los macroscópicos no pueden. Al

vincular un sistema microscópico, el átomo, a uno macroscópico, el gato, Schrödinger estaba señalando lo que creía que era un defecto en la interpretación de Copenhague: era absurdo cuando se aplicaba a un gato. Debió de sorprenderse cuando prácticamente la mayoría de los físicos le respondieron: «Sí, Erwin, tienes toda la razón, hasta que alguien abre la caja, el gato realmente está a la vez muerto y vivo». Especialmente cuando cayó en la cuenta de que no podía averiguar quién tenía razón, incluso si abría la caja. Observaría o bien un gato vivo o bien uno muerto. Podría inferir que el gato había estado en ese estado antes de que abriese la caja, pero no podía estar seguro. El resultado observable era consistente con la interpretación de Copenhague.

Muy bien, añade una cámara al contenido de la caja y graba lo que realmente sucede. Eso resolverá el asunto. «Ah, no», respondieron los físicos. «Solo se puede ver lo que la cámara ha grabado después de abrir la caja. Antes de eso, la grabación está en un estado superpuesto: conteniendo una película de un gato vivo y conteniendo una película de uno muerto.»

La interpretación de Copenhague dejó libres a los físicos para hacer sus cálculos y averiguar qué predecía la mecánica cuántica sin enfrentarse a la dificultad, si no imposibilidad, de cómo surgía el mundo clásico a partir de un sustrato cuántico; cómo un instrumento macroscópico, inimaginablemente complejo en una escala cuántica, de algún modo hacía una medición de un estado cuántico. Como la interpretación de Copenhague hacía el trabajo, no estaban realmente interesados en cuestiones filosóficas. De modo que a generaciones de físicos se les enseñó que Schrödinger había inventado su gato para mostrar que la superposición cuántica se extendía también al mundo macroscópico; exactamente lo contrario de lo que Schrödinger había estado intentando decirles.

No es realmente una gran sorpresa que la materia se comporte de manera extraña al nivel de los electrones y los átomos. Podríamos inicialmente rebelarnos ante ese pensamiento fuera de lo común, pero si un electrón es realmente un diminuto montón de ondas más que un diminuto montón de cosas, podemos aprender a vivir con ello. Si eso significa que el estado del electrón es en sí mismo un poco raro, dando vueltas no solo sobre un eje hacia arriba o un eje hacia abajo, sino un poco de ambos, podemos vivir también con eso. Y si las limitaciones de nuestros

aparatos de medida implican que nunca podemos pillar al electrón haciendo ese tipo de cosa, que cualquier medición que hagamos necesariamente lo deja en alguno de los estados puros, arriba o abajo, entonces así es como es. Si lo mismo aplica para un átomo radiactivo, y los estados son «desintegrado» o «no desintegrado», porque las partículas que lo componen tienen estados tan escurridizos como los del electrón, podemos incluso aceptar que el propio átomo, en su totalidad, podría ser una superposición de esos estados hasta que hagamos una medición. Pero un gato es un gato, y parece ser una extensión muy grande de la imaginación concebir que el animal puede estar tanto vivo como muerto a la vez, y solo milagrosamente se convierten en uno u otro cuando abrimos la caja que lo contiene. Si la realidad cuántica necesita un gato con la superposición vivo/muerto, ¿por qué es tan tímida que no nos permite observar dicho estado?

Hay razones sólidas en el formalismo de la teoría cuántica que (hasta hace muy poco) necesitan alguna medición, algo «observable», para ser una autofunción. Hay razones incluso más sólidas de por qué el estado de un sistema cuántico debería ser una onda, obedeciendo a la ecuación de Schrödinger. ¿Cómo puedes obtener una a partir de la otra? La interpretación de Copenhague declara que de algún modo (no preguntes cómo) el proceso de medición colapsa la función de onda superpuesta compleja en una sola de sus autofunciones. Habiendo sido provisto con esta forma de vocablos, tu tarea como físico es ponerte a hacer mediciones, calcular autofunciones, etcétera, y dejar de hacer preguntas raras. Funciona sorprendentemente bien, si mides el éxito por la obtención de respuestas que concuerdan con el experimento. Y todo habría estado bien si la ecuación de Schrödinger permitiese a la función de onda comportarse de este modo, pero no lo hace. En *La realidad oculta*, Brian Greene lo plantea de este modo: «pero basta un pequeño empujón para revelar rápidamente que hay un aspecto incómodo ... el colapso instantáneo de una onda ... no puede salir de las matemáticas de Schrödinger». En su lugar, la interpretación de Copenhague era un añadido práctico a la teoría, un modo de manejar mediciones sin comprender o enfrentarse a qué era realmente.

Todo esto está muy bien, pero no es lo que Schrödinger estaba intentando señalar. Introdujo un gato, en lugar de un electrón o un átomo, porque puso lo que

consideraba era el principal asunto a destacar. Un gato pertenece al mundo macroscópico en el que vivimos, en el cual la materia no se comporta del modo en que la mecánica cuántica requiere. No vemos gatos superpuestos.⁴⁶ Schrödinger estaba preguntando por qué nuestro universo común «clásico» fracasa en asemejarse a la realidad cuántica subyacente. Si todo aquello de lo que está construido el mundo puede existir en estados superpuestos, ¿por qué el universo parece clásico? Muchos físicos han realizado experimentos maravillosos mostrando que los electrones y los átomos realmente no se comportan del modo que la física cuántica y Copenhague dicen que deberían comportarse. Pero esto no capta la idea: tienes que hacerlo con un gato. Los teóricos se preguntan si el gato podría observarse en su propio estado o si alguien más podría secretamente abrir la caja y escribir qué había dentro. Concluyeron, siguiendo la misma lógica que Schrödinger, que si el gato observase su estado entonces la caja contendría una superposición de un gato muerto que ha cometido suicidio al observarse a sí mismo y un gato vivo que se ha observado a sí mismo para vivir, hasta que el observador legítimo (un físico) abriese la caja. Entonces todo el tinglado pasa a ser uno u otro. De manera similar, el amigo se hace una superposición de dos amigos: uno que ha visto un gato muerto, mientras el otro ha visto uno vivo, hasta que un físico abre la caja, provocando que el estado del amigo colapse. Podrías proceder de este modo hasta que el estado de todo el universo fuese una superposición de un universo con un gato muerto y un universo con un gato vivo, y entonces el estado del universo se colapsaría cuando un físico abriese la caja.

Era todo un poco embarazoso. Los físicos podían seguir con su trabajo sin averiguarlo, podían incluso negar que hubiese algo que tuviesen que averiguar, pero faltaba algo. Por ejemplo, ¿qué nos ocurre si un físico alienígena abre una caja en el planeta Apellobetnees III? ¿De repente descubrimos que realmente nos hicimos saltar por los aires a nosotros mismos en una guerra nuclear cuando la crisis de los misiles en Cuba del 1962 se intensificó, y hemos estado viviendo en un tiempo prestado desde entonces?

El proceso de medición no es una operación matemática pulcra y ordenada que la

⁴⁶ Mi gato, *Harlequin*, puede con frecuencia observarse en una superposición de estados «dormido» y «roncando», pero esto probablemente no cuenta.

interpretación de Copenhague asuma. Cuando se pide describir cómo el aparato llega a su decisión, la interpretación de Copenhague responde «lo hace». La imagen de la función de onda colapsando en una única autofunción describe la entrada y la salida del proceso de medición, pero no cómo obtener uno a partir de otro. Pero cuando haces una medida real, no agitas sin más una varita mágica y haces que la función de onda desobedezca a la ecuación de Schrödinger y colapse. En su lugar, haces algo tan tremadamente complicado, desde un punto de vista cuántico, que es obviamente imposible hacer un modelo de manera realista. Para medir un espín de un electrón, por ejemplo, lo haces interactuar con una pieza apropiada del aparato, la cual tiene una aguja que se mueve o bien a la posición «arriba», o bien a «abajo». O una pantalla numérica, o una señal enviada a un ordenador... Este aparato produce un estado, y solo uno. No ves la aguja en una superposición de arriba y abajo.

Estamos habituados a esto, porque así es como el mundo clásico funciona. Pero por debajo se supone que hay un mundo cuántico. Reemplaza el gato por el aparato de espines y sí que debería existir en un estado superpuesto. El aparato, visto como un sistema cuántico, es extraordinariamente complicado. Contiene cantidades ingentes de partículas, entre 10²⁵ y 10³⁰, en una estimación *grosso modo*. La medición surge de algún modo a partir de la interacción de ese único electrón con ese montón de partículas. La admiración por la pericia de la compañía que manufactura el instrumento debe ser infinita, extraer algo coherente de algo tan lioso es casi increíble. Es como intentar averiguar la talla de zapato de alguien haciéndolo pasear por una ciudad. Pero si eres listo (arréglatelas para que se encuentre una zapatería) puedes obtener un resultado sensato, y un diseñador de instrumentos listo puede producir mediciones significativas de espines de electrones. Pero no es una posibilidad realista hacer un modelo en detalle de cómo funciona dicho instrumento como un sistema cuántico genuino. Hay mucho que detallar, el mayor ordenador del mundo se bloquearía. Eso hace difícil analizar un proceso de medición real usando la ecuación de Schrödinger.

Incluso así, sí tenemos cierta comprensión de cómo nuestro mundo clásico surge a partir de uno cuántico subyacente. Empecemos con una versión simple, un rayo de luz chocando con un espejo. La respuesta clásica, la ley de Snell, afirma que el rayo

reflejado rebota con el mismo ángulo con el que lo golpea. En su libro *QED* sobre electrodinámica cuántica, el físico Richard Feynman explicó que esto no es lo que sucede en el mundo cuántico. El rayo es realmente una corriente de fotones y cada fotón puede rebotar hacia cualquier punto. Sin embargo, si superpones todas las cosas posibles que el fotón podría hacer, entonces obtienes la ley de Snell. La sobrecogedora proporción de fotones rebota en ángulos muy próximos al ángulo con el que golpean. Feynman incluso se las arregló para mostrar el porqué sin usar ninguna matemática complicada, pero tras sus cálculos hay una idea matemática general: el principio de la fase estacionaria. Si superpones todos los estados cuánticos para un sistema óptico, obtienes el resultado clásico en el cual el rayo de luz sigue la trayectoria más corta, medida para un tiempo determinado. Puedes incluso añadir toda la parafernalia para decorar la trayectoria del rayo con las clásicas franjas de difracción de la onda óptica.

Este ejemplo muestra, muy explícitamente, que la superposición de todos los mundos posibles, en este marco óptico, produce el mundo clásico. La característica más importante no es tanto la geometría detallada del rayo de luz, sino el hecho de que produzca solo un único mundo en el nivel clásico. Bajando al detalle cuántico de cada uno de los fotones, puedes observar toda la parafernalia de la superposición, autofunciones, etcétera. Pero en la escala humana, todo se anula, bueno, se agrupa, para producir un nítido mundo clásico.

La otra parte de la explicación se llama decoherencia. Hemos visto que las ondas cuánticas tienen una fase y también una amplitud. Es una fase muy rara, un número complejo, pero es una fase no obstante. La fase es absolutamente crucial para cualquier superposición. Si tomas dos estados superpuestos, cambias la fase de uno y los juntas, lo que obtienes no tiene que ver con el original. Si haces lo mismo con muchas componentes, la onda recreada podría ser casi cualquier cosa. La pérdida de información de la fase deriva en cualquier suposición como la del gato de Schrödinger. No solo pierdes la pista de si está vivo o muerto, sino que no puedes afirmar que sea un gato. Cuando las ondas cuánticas dejan de tener buenas relaciones entre las fases, hay decoherencia, empiezan a comportarse más como la física clásica y las superposiciones pierden cualquier significado. Lo que hace que haya decoherencia son las interacciones con las partículas que las rodean. Así es

probablemente como el aparato puede medir el espín del electrón y obtener un resultado concreto único.

Ambas aproximaciones llevan a la misma conclusión: la física clásica es lo que observas si consideras una visión a escala humana de un sistema cuántico muy complicado con cantidades ingentes de partículas. Métodos experimentales especiales, artilugios especiales, podrían preservar algunos de los efectos cuánticos, haciéndolos encajar en nuestra cómoda existencia clásica, pero los sistemas genéricos cuánticos rápidamente dejan de parecer cuánticos a medida que nos movemos a escalas más grandes de comportamiento.

Ese es un modo de resolver el destino del pobre gato. Solo si la caja es totalmente impermeable a la decoherencia cuántica, el experimento puede producir el gato superpuesto y dichas cajas no existen. ¿De qué las harías?

Pero hay otro modo, uno que va al extremo opuesto. Dije antes que «podrías proceder de este modo hasta que el estado de todo el universo fuese una superposición». En 1957, Hugh Everett Jr. señaló que en cierto sentido, tienes que hacerlo. El único modo de proporcionar un modelo cuántico exacto de un sistema es considerar su función de onda. Todo el mundo era feliz haciéndolo así cuando el sistema era un electrón, o un átomo, o (de manera más polémica) un gato. Everett tomó como sistema el universo entero.

Argumentó que no tenías elección si eso era de lo que querías hacer un modelo. Nada menor que el universo puede ser realmente aislado. Todo interactúa con todo lo demás. Y descubrió que si das ese paso, entonces el problema del gato y la relación paradójica entre la realidad cuántica y la clásica se resuelve fácilmente. La función de onda cuántica del universo no es un modo normal puro, sino una superposición de todos los modos normales posibles. Aunque no podemos calcular dichas cosas (no podemos para un gato, y un universo es una pizca más complicado), podemos razonar acerca de ellas. De hecho, estamos representando el universo, mecánica y cuánticamente, como una combinación de todas las cosas posibles que un universo puede hacer.

El resultado fue que la función de onda del gato no tiene que colapsar para dar una única observación clásica. Puede permanecer totalmente sin cambios, sin violar la ecuación de Schrödinger. En su lugar, hay dos universos que coexisten. En uno, el

gato muere, en el otro, no lo hace. Cuando abres la caja, hay en consecuencia dos tú y dos cajas. Una de ella es parte de la función de onda de un universo con un gato muerto, la otra es parte de una función de onda diferente con un gato vivo. En lugar de un único mundo clásico que de algún modo surge de la superposición de posibilidades cuánticas, tenemos un amplio rango de mundos clásicos, cada uno correspondiente a una posibilidad cuántica.

La versión original de Everett, que llamó la formulación del estado relativo, captó la atención popular en la década de los setenta del siglo XX gracias Bryce DeWitt, quien le dio un nombre más pegadizo: los universos paralelos de la mecánica cuántica. Es con frecuencia escenificado en términos históricos, por ejemplo, que hay un universo en el que Adolf Hitler ganó la Segunda Guerra Mundial y otro en el que no. El universo en el que estoy escribiendo este libro es el segundo, pero en algún lugar en el reino cuántico otro Ian Stewart está escribiendo un libro muy similar a este, pero en alemán, recordando a sus lectores que están en el universo en el que Hitler ganó. Matemáticamente, la interpretación de Everett puede verse como un equivalente lógico de la mecánica cuántica convencional, y lleva, en interpretaciones más limitadas, a modos eficientes de resolver problemas físicos. Su formalismo, por lo tanto, sobrevivirá a cualquier prueba experimental a la que sobreviva la mecánica cuántica convencional. De modo que eso implica que estos universos paralelos, o «mundos alternativos» como también se les llama, ¿realmente existen? ¿Hay otro yo tecleando felizmente en un teclado de ordenador en un mundo donde Hitler ganó? ¿O es el montaje de una ficción matemática conveniente?

Hay un problema obvio: ¿cómo podemos estar seguros de que en un mundo dominado por el sueño de Hitler, el Reich de los mil años, ordenadores como el que estoy usando existirían? Claramente debe haber muchos más universos que dos y los sucesos en ellos deben seguir patrones clásicos coherentes. Así que quizá Stewart-2 no existe, pero sí Hitler-2. Una descripción común de la formación y evolución de universos paralelos los supone «separándose» siempre que hay una elección de un estado cuántico. Greene señala que esta imagen es errónea: nada se separa. La función de onda del universo ha sido, y siempre será, separada. Sus autofunciones están ahí; imaginamos una división cuando seleccionamos una de

ellas, pero el objetivo de la explicación de Everett es que nada en la función de onda realmente cambia.

Con eso como salvedad, un número sorprendente de físicos cuánticos aceptaron la interpretación de los universos paralelos. El gato de Schrödinger realmente está vivo y muerto. Hitler realmente gana y pierde. Una versión de nosotros vive en uno de esos universos, en otros no. Esto es lo que las matemáticas dicen. No es una interpretación, un modo conveniente de arreglar los cálculos. Es tan real como tú y yo. Es tú y yo.

No estoy convencido. Aunque no es la superposición lo que me molesta. No encuentro la existencia de un mundo nazi paralelo impensable o imposible.⁴⁷ Pero sí rechazo, enérgicamente, la idea de que puedes separar una función de onda cuántica según una narración histórica de escala humana. La separación matemática se da al nivel de estados cuánticos de las partículas que lo constituyen. La mayoría de las combinaciones de los estados de partículas no tienen sentido sea cual sea la narración humana. Una alternativa simple a una muerte de un gato no es un gato vivo. Es un gato muerto con un electrón en un estado diferente. Las alternativas complejas son mucho más numerosas que un gato vivo. Incluyen un gato que de repente explota sin razón aparente, uno que se convierte en un florero, uno que es elegido presidente de los Estados Unidos y uno que sobrevive aunque el átomo radiactivo libere el veneno. Esos gatos alternativos son útiles retóricamente, pero poco representativos. La mayoría de las alternativas no son ni siquiera gatos, de hecho, son indescriptibles en términos clásicos. Si es así, la mayoría de los Stewart alternativos no son reconocibles como personas, de hecho como nada, y casi todos los que existen lo hacen en un mundo que no tiene sentido para nada en términos humanos. De modo que la posibilidad de otra versión de un pequeño viejo yo, que resulta vivir en otro mundo que tiene sentido narrativo para un ser humano, es insignificante.

El universo podría bien ser una superposición increíblemente compleja de un estado alternativo. Si piensas que la mecánica cuántica es fundamentalmente correcta, tiene que serlo. En 1983, el físico Stephen Hawking dijo que la interpretación de los

⁴⁷ Dos novelas de ciencia ficción sobre esto son *El hombre en el castillo* de Philip K. Dick y *El sueño de hierro* de Norman Spinrad. *SS-GB* del escritor de misterio Len Deighton está también ambientado en una Inglaterra contrafactual regida por los nazis.

universos paralelos era «evidentemente correcta» en este sentido. Pero eso no quiere decir que exista una superposición de universos en los cuales un gato está vivo o muerto y Hitler gana o no gana. No hay razón para suponer que los componentes matemáticos pueden separarse en conjuntos que encajan unos con otros para crear narraciones humanas. Hawking descartó interpretaciones narrativas del formalismo de los universos paralelos, diciendo que «todo lo que hace, realmente, es calcular probabilidades condicionales, en otras palabras, la probabilidad de que A suceda, dada B. Creo que esto es todo lo que la interpretación de los universos paralelos es. Alguna gente lo recubre con mucho misticismo sobre la función de onda dividiéndose en partes diferentes. Pero todo lo que estás calculando es una probabilidad condicional».

Merece la pena comparar el relato de Hitler con la historia de Feynman del rayo de luz. A la manera de los Hitlers alternativos, Feynman nos estaría diciendo que hay un mundo clásico donde los rayos de luz rebotan en el espejo con el mismo ángulo con el que lo golpean, otro mundo clásico en el cual rebotan con un ángulo que se desvía un grado, otro donde se desvía dos grados, etcétera. Pero no lo hizo. Nos dijo que hay un único mundo clásico, que surge a partir de la superposición de alternativas cuánticas. Podría haber innumerables mundos paralelos a nivel cuántico, pero estos no se corresponden de manera significativa con los mundos paralelos que son descriptibles a un nivel clásico. La ley de Snell es válida en cualquier mundo clásico. Si no lo fuese, el mundo no podría ser clásico. Como explicó Feynman para los rayos de luz, el mundo clásico surge cuando superpones todas las alternativas cuánticas. Hay solo una de dichas superposiciones, de modo que solo hay un universo clásico. El nuestro.

La mecánica cuántica no está confinada al laboratorio. Toda la electrónica moderna depende de ella. La tecnología de semiconductores, las bases de todos los circuitos integrados —chips de silicio— es mecánica cuántica. Sin la física de lo cuántico, nadie habría soñado que dichos aparatos podrían funcionar. Ordenadores, teléfonos móviles, reproductores de CD, consolas de videojuegos, coches, neveras, hornos, prácticamente todos los electrodomésticos modernos contienen chips de memoria para almacenar las instrucciones que hacen que estos aparatos hagan lo que queremos. Muchos contienen sistemas de circuitos más complejos, como

microprocesadores, un ordenador entero en un chip. La mayoría de los chips de memoria son variaciones sobre el primer aparato semiconductor verdadero: el transistor.

En la tercera década del siglo XX, los físicos norteamericanos Eugene Wigner y Frederick Seitz analizaron cómo los electrones se movían a través de un cristal, un problema que necesita mecánica cuántica. Descubrieron algunas de las características básicas de los semiconductores. Algunos materiales son conductores de la electricidad: los electrones pueden fluir a través de ellos con facilidad. Los metales son buenos conductores, y en el día a día el uso del hilo de cobre es común para este propósito. Los aislantes no permiten a los electrones fluir, de modo que paran el flujo de electricidad. Los plásticos que revisten los cables eléctricos, para impedir que nos electrocutemos con los cables de la TV, son aislantes. Los semiconductores son un poco de ambos, dependiendo de las circunstancias. El silicio es el más conocido y actualmente el que se usa más ampliamente, pero otros elementos como el antimonio, el arsénico, el boro, el carbono, el germanio y el selenio son también semiconductores. Porque los semiconductores pueden cambiarse de un estado a otro, pueden usarse para manipular corrientes eléctricas, y esto es la base de todos los circuitos electrónicos.

Wigner y Seitz descubrieron que las propiedades de los semiconductores dependen de los niveles de energía de los electrones en ellos, y estos niveles pueden controlarse «dopando» el material semiconductor básico añadiéndole pequeñas cantidades de impurezas específicas. Dos tipos importantes son los semiconductores de tipo P, que llevan la corriente como el flujo de los electrones, y de tipo N, en los cuales la corriente fluye en el sentido opuesto a los electrones, llevados por «huecos», lugares donde hay menos electrones de lo normal. En 1947, John Bardeen y Walter Brattain en los Laboratorios Bell descubrieron que un cristal de germanio podía actuar como un amplificador. Si una corriente eléctrica lo alimentaba, la corriente resultante era mayor. William Shockley, líder del Solid State Physics Group (Grupo de la física de estados sólidos), se dio cuenta de cuán importante podría ser esto, e inició un proyecto para investigar los semiconductores. Por esto apareció el transistor, abreviatura de «transfer resistor» (resistencia de transferencia). Hubo algunas patentes anteriores pero no aparatos

que funcionasen o artículos publicados. Técnicamente el artilugio del Laboratorio Bell era un JEFT (del inglés *junction gate field-effect transistor*, transistor de efecto de campo de juntura, figura 55). Desde este gran paso adelante, se han inventado muchos tipos de transistores diferentes. Texas Instruments fabricó el primer transistor de silicio en 1954. El mismo año vio la luz un ordenador basado en transistores, TRIDAC, construido por el ejército de EE.UU. El tamaño era de tres pies cúbicos y su necesidad de energía era la misma que la de una bombilla. Fue uno de los primeros pasos en un enorme programa militar estadounidense para desarrollar alternativas a la electrónica del tubo de vacío, la cual era demasiado engorrosa, frágil y poco fidedigna para el uso militar.

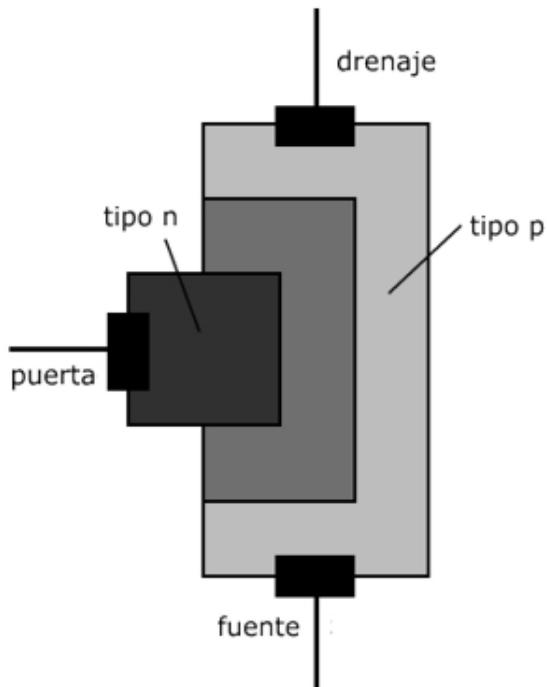


FIGURA 55. Estructura de un JEFT. La fuente y el sumidero son los extremos, en una capa de tipo P, mientras la puerta está en una capa de tipo N que controla el flujo. Si piensas en el flujo de electrones de la fuente al sumidero como una manguera, la puerta presiona la manguera, incrementando la presión (voltaje) en el sumidero.

Como la tecnología de semiconductores está basada en dopar silicio o sustancias

similares con impurezas, se presta a la minituarización. Los circuitos pueden construirse en capas sobre un sustrato de silicio, pero bombardeando la superficie con la impureza deseada, y grabando aparte regiones no queridas con ácido. Las áreas afectadas están determinadas por marcas producidas fotográficamente, y estas pueden reducirse a tamaños muy pequeños usando lentes ópticas. De todo esto surgió la electrónica actual, incluyendo los chips de memoria que pueden almacenar miles de millones de bytes de información y microprocesadores muy rápidos que orquestan la actividad de los ordenadores.

Otra aplicación de la mecánica cuántica que está por todas partes es el láser. Se trata un artefacto que emite un fuerte haz de luz coherente: uno en la cual las ondas de luz guardan una relación entre sus fases. Consiste en una cavidad óptica con espejos al final, llenos con algo que reacciona a la luz de una longitud de onda específica produciendo más luz de la misma longitud de onda, un amplificador de luz. Bombea energía para empezar a rodar el proceso, permite a la luz rebotar de un lado a otro a lo largo de la cavidad, ampliándose todo el tiempo, y cuando alcanza una intensidad suficientemente alta, la deja salir. El medio activo puede ser un fluido, un gas, un cristal o un semiconductor. Materiales diferentes funcionan para longitudes de onda diferentes. El proceso de amplificación depende de la mecánica cuántica de los átomos. Los electrones en los átomos pueden existir en diferentes estados de energía, y pueden intercambiarse entre ellos absorbiendo o emitiendo fotones.

LASER significa ampliación de luz por una emisión de radiación estimulada (en inglés *Light Amplification by Stimulated Emission of Radiation*). Cuando se inventó el primer laser, fue ampliamente ridiculizado como una respuesta buscando un problema. Esto fue poco imaginativo; una vez hubo una solución, toda una serie de problemas apropiados aparecieron rápidamente. Producir un haz de luz coherente es tecnología básica y siempre ha estado vinculada a tener usos, del mismo modo que un martillo mejorado automáticamente encontraría muchos usos. Cuando inventas tecnología genérica, no tienes que tener una aplicación específica en mente. Hoy en día usamos láser con tantos propósitos que es imposible hacer una lista de todos ellos. Hay usos prosaicos como los punteros láser para las clases y láser para cortar papel en manualidades. Los reproductores de CD, los

reproductores de DVD y Blue-ray, todos usan láser para medir distancias y ángulos. Los astrónomos usan láser para medir la distancia de la Tierra a la Luna. Los cirujanos usan láser para cortar con cuidado tejidos delicados. El tratamiento láser para los ojos es una rutina, para reparar retinas separadas y remodelar la superficie de la córnea para una visión correcta en lugar de usar gafas o lentillas. El sistema antimisiles de «La guerra de las galaxias»* estaba pensado para usar láseres potentes para disparar a los misiles enemigos, y aunque nunca se construyó, algunos de los láseres sí. Los usos militares de los láseres, similares a la pistola de rayos de la ciencia ficción barata, están siendo investigados ahora. Y podría incluso ser posible lanzar vehículos espaciales desde la Tierra haciéndolos montar en un rayo láser potente.

Nuevos usos de la mecánica cuántica aparecen casi todas las semanas. Uno de los últimos son los puntos cuánticos, piezas minúsculas de semiconductores cuyas propiedades electrónicas, incluyendo la luz que emiten, varía según su tamaño y forma. Pueden, por lo tanto, adaptarse a tener muchas características deseables. Ya tiene una variedad de aplicaciones, incluyendo la representación de imágenes en biología, donde pueden remplazar a los tintes tradicionales (y con frecuencia tóxicos). Además su imagen es mucho mejor, emitiendo una luz más brillante.

Traspasando la línea, algunos ingenieros y físicos están trabajando en los componentes básicos de un ordenador cuántico. En dicho dispositivo, los estados binarios de 0 y 1 pueden ser superpuestos en cualquier combinación, prácticamente permitiendo a los cómputos adquirir ambos valores al mismo tiempo. Esto permitiría realizar muchos cálculos diferentes en paralelo, acelerándolos enormemente. Se han concebido algoritmos teóricos, llevando a cabo tareas como la división de un número en sus factores primos. Los ordenadores convencionales se topán con problemas cuando los números tienen más de un centenar de dígitos o así, pero un ordenador cuántico debería ser capaz de factorizar números mucho mayores con facilidad. El principal obstáculo para la computación cuántica es la decoherencia, que destruye los estados superpuestos. El gato de Schrödinger se está vengando por su tratamiento inhumano.

Capítulo 15

Códigos, comunicaciones y ordenadores

Teoría de la información

$$H = - \sum_{X} p(X) \log p(X)$$

información / suma simbolo de probabilidad
 X símbolos logaritmo base 2

¿Qué dice?

Define cuánta información contiene un mensaje, en términos de las probabilidades con las que los símbolos que lo componen tienen la posibilidad de darse.

¿Por qué es importante?

Es la ecuación que marca el comienzo de la era de la información. Estableció los límites en la eficiencia de las comunicaciones, permitiendo a los ingenieros dejar de buscar códigos que fuesen demasiado efectivos para existir. Es básica en las comunicaciones digitales de hoy en día: teléfonos, CDs, DVDs, Internet.

¿Qué provocó?

Códigos eficientes de detección y corrección de errores, usados en todo, desde CDs a sondas espaciales. Las aplicaciones incluyen estadística, inteligencia artificial, criptografía, y obtener significado de la secuencia de ADN.

En 1977, la NASA lanzó dos sondas espaciales, *Voyager 1* y *2*. Los planetas del Sistema Solar se habían colocado a sí mismos en posiciones favorables poco comunes, haciendo posible encontrar órbitas razonablemente eficientes que permitiesen a las sondas visitar varios planetas. El objetivo inicial era examinar Júpiter y Saturno, pero si las sondas resistían, sus trayectorias los harían pasar por Urano y Neptuno. *Voyager 1* podría haber ido hasta Plutón (en esa época

considerado un planeta, e igualmente interesante, de hecho sin ningún tipo de cambio, ahora no lo es), pero una alternativa, la misteriosa luna de Saturno, Titán, tuvo prioridad. Ambas sondas fueron espectacularmente exitosas y *Voyager 1* es ahora el objeto hecho por el hombre más distante de la Tierra, a más de 16 mil millones de kilómetros y todavía enviando datos.

La fuerza de la señal decae con el cuadrado de la distancia, de modo que la señal recibida en la Tierra es del orden de 10–20 veces la fuerza con la que se recibiría si la distancia fuese un kilómetro, esto es, cien trillónes más débil. *Voyager 1* debe tener un transmisor realmente potente... No, es una sonda espacial diminuta. Está propulsada por un isótopo radiactivo, plutonio-238, pero aun así la potencia total disponible es ahora alrededor de un octavo de la de un hervidor de agua eléctrico típico. Hay dos razones de por qué podemos todavía obtener información útil de la sonda: los potentes receptores en la Tierra, y los códigos especiales usados para proteger los datos de errores provocados por factores extraños como interferencias.

Voyager 1 puede enviar datos usando dos sistemas diferentes. Uno, el canal de baja velocidad, puede enviar 40 dígitos binarios, 0s o 1s, cada segundo, pero no permite codificar para tratar con errores potenciales. El otro, el canal de alta velocidad, puede transmitir hasta 120.000 dígitos binarios cada segundo, y estos están codificados de manera que los errores pueden descubrirse y corregirse siempre que no sean demasiado frecuentes. El precio pagado por esta habilidad es que los mensajes son el doble de largos de lo que serían de otro modo, de manera que solo llevan la mitad de datos de los que podrían llevar. Como los errores pueden arruinar los datos, este es un precio que merece la pena pagar.

Los códigos de este tipo son ampliamente usados en todas las comunicaciones modernas: misiones espaciales, teléfonos fijos, Internet, CD, DVD y Blue-ray, etcétera. Sin ellos, todas las comunicaciones serían propensas a errores, lo que no sería aceptable si, por ejemplo, estás usando Internet para pagar una factura. Si tu orden de pagar 20 € se recibe como 200 €, no sería agradable. Un reproductor de CD usa unas lentes diminutas, que enfocan un rayo láser sobre unas pistas muy finitas impresas en el material del disco. La lente se mantiene a una distancia pequeñísima sobre el disco que gira. Y aun así puedes escuchar un CD mientras

conduces por una carretera llena de baches, porque la señal está codificada de un modo que permite al sistema electrónico encontrar los errores y corregirlos mientras el disco está sonando. Hay otros trucos, también, pero este es el fundamental.

Nuestra era de la información se sustenta en señales digitales, cadenas largas de 0s y 1s, pulsaciones y no pulsaciones de electricidad o radio. El equipamiento que envía, recibe y almacena las señales depende de circuitos electrónicos muy pequeños y muy precisos sobre láminas de silicio, «chips». A pesar de todo lo ingenioso del diseño y fabricación del circuito, ninguno funcionaría sin códigos de detección y corrección de errores. Y fue en este contexto donde el término «información» dejó de ser una palabra informal para «conocimientos» y se convirtió en una cantidad numérica medible. Y eso proporcionó limitaciones fundamentales en la eficiencia con la que los códigos pueden modificar mensajes para protegerlos contra errores. Conocer estas limitaciones ahorró a los ingenieros mucha pérdida de tiempo, tratando de inventar códigos que serían tan eficientes que serían imposibles. Proporcionó las bases para la cultura de la información actual.

Soy lo suficientemente mayor para recordar cuando el único modo de telefonear a alguien en otro país (ihorror de horrores!) era hacer una reserva por adelantado con la compañía telefónica —en Reino Unido solo había una, Post Office Telephones—, a una hora y de una duración concreta. Por ejemplo, diez minutos a las 3:45 pm el 11 de enero. Y valía una fortuna. Hace unas pocas semanas un amigo y yo hicimos una entrevista que duró una hora para una convención de ciencia ficción en Australia desde Reino Unido, usando Skype™. Fue gratis, y enviaba vídeo además de sonido. Han cambiado muchas cosas en cincuenta años. En la actualidad, intercambiamos información *online* con amigos, tanto los reales como los falsos que gran número de personas coleccionan como mariposas usando las redes sociales. Ya no compramos CD de música o DVD de películas, compramos la información que contienen, transferida a través de Internet. Los libros apuntan en la misma dirección. Las compañías de investigación de mercados amasan enormes cantidades de información sobre nuestros hábitos de consumo e intentan usarla para influenciar en lo que compramos. Incluso en medicina hay un énfasis creciente en la información que está contenida en nuestro ADN. Con frecuencia la actitud

parece ser que si tienes la información necesaria para hacer algo, entonces eso solo es suficiente; no necesitas realmente hacerlo, o incluso saber cómo hacerlo.

No hay duda de que la revolución de la información ha transformado nuestras vidas, y pueden darse buenos argumentos de que en términos generales, los beneficios pesan más que las desventajas, incluso aunque las últimas incluyan la pérdida de privacidad, el potencial acceso fraudulento a nuestras cuentas bancarias desde cualquier lugar del mundo a un clic de un ratón y virus informáticos que pueden inutilizar un banco o una central nuclear.

¿Qué es información? ¿Por qué tiene tanto poder? Y ¿es realmente lo que dice ser?

El concepto de información como una cantidad medible surge a partir de los laboratorios de investigación de Bell Telephone Company, el principal proveedor de servicios telefónicos en Estados Unidos desde 1877 hasta su división en 1984 en base a las leyes antimonopolio. Entre sus ingenieros estaba Claude Shannon, un primo lejano del famoso inventor Edison. La asignatura que mejor se le daba a Shannon en la escuela eran las matemáticas, y tenía una gran aptitud para construir dispositivos mecánicos. En la época que estaba trabajando para Bell Labs, era matemático y criptógrafo, además de ingeniero electrónico. Fue uno de los primeros en aplicar la lógica matemática, denominada álgebra booleana, a circuitos informáticos. Usó esta técnica para simplificar el diseño de circuitos de conmutación usados por los sistemas telefónicos, y luego lo amplió a otros problemas en el diseño de circuitos.

Durante la Segunda Guerra Mundial trabajó en códigos y comunicaciones secretas y desarrolló algunas ideas fundamentales que se presentaron en un memorándum clasificado para Bell en 1945 bajo el título de «A mathematical theory of cryptography» (Una teoría matemática de la criptografía). En 1948, publicó parte de su trabajo en una publicación abierta y el artículo de 1945, desclasificado, se publicó poco después. Con material adicional de Warren Weaver, apareció en 1949 como *The Mathematical Theory of Communication* (La teoría matemática de la comunicación).

Shannon quería saber cómo transmitir mensajes de modo efectivo cuando el canal de transmisión estaba sujeto a errores aleatorios, «ruido» en la jerga de ingenieros. Todas las comunicaciones prácticas sufren de ruido, ya sea de un equipo

defectuoso, de rayos cósmicos o variabilidad inevitable en los componentes de los circuitos. Una solución es reducir el ruido fabricando equipamiento mejor, si es posible. Una alternativa es codificar las señales usando procedimientos matemáticos que pueden detectar errores, e incluso corregirlos.

El código de detección de errores más simple es enviar el mismo mensaje dos veces. Si recibes:

- el mismo masaje dos veces
- el mismo mensaje dos veces

entonces hay claramente un error en la tercera palabra; pero sin entender español, no está claro qué versión es la correcta. Una tercera repetición decidiría el asunto por mayoría y se convertiría en un código de corrección de errores. Cómo de efectivos o precisos son dichos códigos depende de la probabilidad, y naturaleza, de los errores. Si el canal de comunicación es muy ruidoso, por ejemplo, entonces las tres versiones del mensaje podrían ser tan enrevesadas que sería imposible reconstruirlo.

En la práctica la mera repetición es demasiado simple: hay modos más eficientes de codificar mensajes para descubrir y corregir errores. El punto de arranque de Shannon era establecer con exactitud el significado de eficiencia. Todos esos códigos remplazan el mensaje original por uno más largo. Los dos códigos anteriores doblan o triplican la longitud. Mensajes más largos tardan más en enviarse, cuestan más, ocupan más memoria, y obstruyen el canal de comunicación. De manera que la eficiencia, para una tasa dada de detección o corrección de error, puede cuantificarse como la proporción de la longitud del mensaje codificado con respecto al original.

El asunto principal, para Shannon, era determinar las limitaciones inherentes de dichos códigos. Supón que un ingeniero ha creado un código nuevo. ¿Había algún modo de decidir si era lo mejor que podían obtener o sería posible alguna mejora? Shannon empezó cuantificando cuánta información contiene un mensaje. Haciendo eso, hizo que «información» pasase de ser una metáfora vaga a un concepto científico.

Hay dos modos distintos de representar un número. Puede definirse por una

secuencia de símbolos, por ejemplo, sus dígitos decimales, o puede corresponderse con alguna cantidad física, como la longitud de una vara o el voltaje en un cable. Las representaciones del primer tipo son digitales, las del segundo son analógicas. En la tercera década del siglo XX, los cálculos de científicos e ingenieros con frecuencia se realizaban usando ordenadores analógicos, porque en esa época estos eran más fáciles de diseñar y construir. Circuitos electrónicos simples pueden, por ejemplo, sumar o multiplicar dos voltajes. Sin embargo, las máquinas de este tipo carecen de precisión y los ordenadores digitales empezaron a aparecer. Muy rápidamente estuvo claro que la representación más conveniente de números no era la decimal, la de base 10, sino la binaria, la de base 2. En la notación decimal, hay diez símbolos para los dígitos, 0-9, y cada dígito multiplica por diez su valor por cada paso que se mueve a la izquierda. De modo que 157 representa:

$$1 \times 10^2 + 5 \times 10^1 + 7 \times 10^0$$

La notación binaria emplea el mismo principio básico, pero ahora hay solo dos dígitos, 0 y 1. Un número binario como 10011101 codifica, de forma simbólica, el número:

$$1 \times 2^7 + 0 \times 2^6 + 0 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$$

De modo que cada dígito dobla su valor por cada paso que se mueva a la izquierda. En decimales, este número es igual a 157, así que hemos escrito el mismo número de dos formas diferentes, usando dos tipos de notación diferentes.

La notación binaria es ideal para los sistemas electrónicos porque es mucho más fácil de distinguir entre dos posibles valores de una corriente, o un voltaje, o un campo magnético, de lo que es distinguir entre más de dos. En términos rudimentarios, 0 puede significar «no corriente eléctrica» y 1 puede significar «algo de corriente eléctrica», 0 puede significar «no campo magnético» y 1 puede significar «algo de campo magnético», etcétera. En la práctica los ingenieros fijan un valor umbral, y entonces 0 significa «bajo el umbral» y 1 significa «sobre el umbral». Manteniendo los valores reales usados para 0 y 1 suficientemente

alejados, y fijando el umbral entre ellos, hay muy poco peligro de confundir 0 con 1. De modo que los dispositivos basados en notación binaria son robustos. Esto es lo que les hace digitales.

Con los primeros ordenadores, los ingenieros tuvieron que luchar por mantener las variables del circuito dentro de límites razonables, y el binario hizo sus vidas mucho más fáciles. Los circuitos modernos en chips de silicio son lo suficientemente precisos para permitir otras opciones, como base 3, pero el diseño de ordenadores digitales ha estado basado en notación binaria durante tanto tiempo que generalmente tiene sentido seguir con el binario, incluso si hay alternativas que podrían funcionar. Los circuitos modernos son también muy pequeños y muy rápidos. Sin dicho avance tecnológico en la fabricación de circuitos, el mundo tendría unos pocos miles de ordenadores, en vez de millardos. Thomas J. Watson, fundador de IBM, una vez dijo que no creía que hubiera mercado para más de cinco ordenadores en todo el mundo. En ese momento, parecía que estaba hablando con sensatez, porque en esa época los ordenadores más potentes eran más o menos del tamaño de una casa, consumían tanta electricidad como un pueblo pequeño, y costaban decenas de millones de dólares. Solo grandes organizaciones gubernamentales, como la armada de Estados Unidos, podían permitírselos, o hacer uso suficiente de ellos. Hoy en día un teléfono móvil básico anticuado contiene más potencia informática que cualquiera de los que estaban disponibles cuando Watson hizo su comentario.

La elección de la representación binaria para ordenadores digitales, por lo tanto también para los mensajes digitales transmitidos entre ordenadores, y más tarde entre casi cualquier par de aparatos electrónicos en el planeta, llevó a la unidad básica de información: el bit. El nombre es la abreviatura de «dígito binario», y un bit de información es un 0 o un 1. Es razonable definir la información «contenida en» una secuencia de dígitos binarios como el número total de dígitos en la secuencia. De modo que la secuencia de 8 dígitos 10011101 contiene 8 bits de información.

Shannon se dio cuenta de que el conteo sencillo de bits tiene sentido como una medida de información solo si ceros y unos son como caras y cruces con una moneda no trucada, esto es, hay la misma probabilidad de que ocurran. Supón que

sabemos que en algunas circunstancias específicas el 0 se da nueve veces sobre diez, y el 1 solo una vez. Cuando leemos la cadena de dígitos, esperamos que la mayoría de dígitos sea 0. Si esa expectativa se confirma, no hemos recibido mucha información, porque esto es lo que esperamos de todos modos. Sin embargo, si vemos 1, eso expresa mucha más información, porque no esperábamos eso para nada.

Podemos sacar ventaja de esto codificando la misma información más eficientemente. Si 0 se da con probabilidad 9/10 y 1 con probabilidad 1/10, podemos definir un nuevo código como este:

- 000 → 00 (lo usamos siempre que sea posible)
- 00 → 01 (si no quedan 000)
- 0 → 10 (si no quedan 00)
- 1 → 11 (siempre)

Lo que quiero decir aquí es que en un mensaje como:

00000000100010000010000001000000000

primero se parte de izquierda a derecha en bloques que se leen como 000, 00, 0 o 1. Con cadenas consecutivas de ceros, usamos 000 siempre que podamos. Si no, lo que está a la izquierda es o 00 o 0, seguido por un 1. Así que aquí el mensaje se divide como:

000-000-00-1-000-1-000-00-1-000-000-1-000-000-000

Y la versión codificada es:

00-00-01-11-00-11-00-01-11-00-00-11-11-00-00-00

El mensaje original tiene 35 dígitos, pero la versión codificada solo 32. La cantidad de información parece haber decrecido.

A veces la versión codificada podría ser más larga, por ejemplo, 111 se convierte en

111111. Pero eso es raro porque 1 se da solo una vez cada diez de media. Habrá bastantes 000, que se reducen a 00. Cualquier resto 00 cambia a 01, la misma longitud, un resto 0 incrementa la longitud en uno al cambiarse a 00. El resultado es que a la larga, para mensajes escogidos aleatoriamente con las probabilidades dadas para 0 y 1, la versión codificada es más corta.

Mi código aquí es muy sencillo, y una elección más inteligente puede acortar el mensaje todavía más. Una de las muchas cuestiones que Shannon quería responder era: ¿cómo de eficiente pueden ser los códigos de este tipo? Si conoces la lista de símbolos que están siendo usados para crear un mensaje, y también sabes cómo de probable es cada símbolo, ¿cuánto puedes acortar el mensaje usando un código apropiado? Su solución era una ecuación, definiendo la cantidad de información en términos de estas probabilidades.

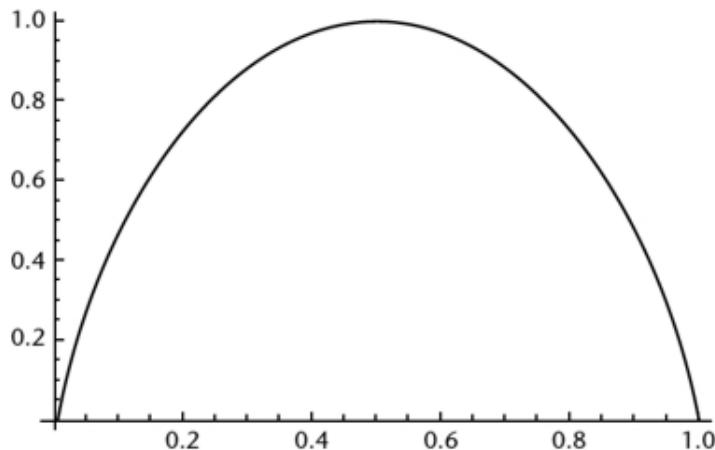


FIGURA 56. Cómo la información H de Shannon depende de p . H avanza verticalmente y p horizontalmente.

Supón, para simplificar, que el mensaje usa solo dos símbolos 0 y 1, pero ahora estos son como lanzamientos de una moneda trucada, de modo que 0 tiene probabilidad p de ocurrir, y 1 tiene probabilidad $q = 1 - p$. El análisis de Shannon le llevó a una fórmula para el contenido de la información: debería definirse como:

$$H = -p \log_2 p - q \log_2 q$$

donde \log_2 es el logaritmo de base 2.

A primera vista esto no parece demasiado intuitivo. Explicaré cómo Shannon llegó a ello en su momento, pero el tema principal que hay que apreciar en esta etapa es cómo H se comporta a medida que p varía de 0 a 1, tal como se muestra en la figura 56. El valor de H aumenta suavemente de 0 a 1 a medida que p crece de 0 a $\frac{1}{2}$, y luego cae simétricamente de nuevo a 0 a medida que p va de $\frac{1}{2}$ a 1.

Shannon señaló varias «propiedades interesantes» de H , así definidas:

- Si $p = 0$, en cuyo caso solo se dará el símbolo 1, la información H es cero. Esto es, si estamos seguros de qué símbolo se nos va a transmitir, recibirla no expresa ninguna información en absoluto.
- Lo mismo aplica cuando $p = 1$. Solo se dará el símbolo 0 y, de nuevo, no recibiremos ninguna información.
- La cantidad de información es la mayor cuando $p = q = \frac{1}{2}$, y corresponde al lanzamiento de una moneda no trucada. En este caso:

$$H = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = -\log_2 \frac{1}{2} = 1$$

Esto es, un lanzamiento de una moneda no trucada transmite un bit de información, como estábamos originalmente asumiendo antes de empezar a preocuparnos por comprimir los mensajes codificados y las monedas trucadas.

- En todos los otros casos, recibir un símbolo transmite menos información que un bit.
- Cuanto más trucada está la moneda, menos información transmite el resultado de un lanzamiento.
- La fórmula trata los dos símbolos exactamente del mismo modo. Si intercambiamos p y q , entonces H se queda igual.

Todas estas propiedades corresponden a nuestro sentido intuitivo de cuánta información recibimos cuando se dice el resultado de un lanzamiento de moneda. Eso hace de la fórmula una definición razonable que funciona. Shannon proporcionó luego una fundación sólida para su definición haciendo una lista de varios principios

básicos que cualquier medida de contenido de información debería obedecer, obteniendo una fórmula única que los satisfacía. Su sistema era muy general: el mensaje podía escoger entre un número diferente de símbolos, que se daban con probabilidades p_1, p_2, \dots, p_n , donde n es el número de símbolos. La información H transmitida por la elección de uno de estos símbolos debería satisfacer:

- H es una función continua de p_1, p_2, \dots, p_n . Esto es, cambios pequeños en las probabilidades deberían llevar a pequeños cambios en la cantidad de información.
- Si todas las probabilidades son iguales, lo que implica que son todas $1/n$, entonces H debería aumentar si n se hace mayor. Esto es, si estás escogiendo entre 3 símbolos, todos igual de probables, entonces la información que recibes debería ser más que si la elección fuese entre dos símbolos igual de probables; una elección entre 4 símbolos debería transmitir más información que una elección entre 3 símbolos, y así sucesivamente.
- Si hay un modo natural de desglosar una elección en dos elecciones sucesivas, entonces el H original debería ser una combinación simple de los nuevos H .

Esta condición final se entiende mucho más fácilmente usando un ejemplo, y he puesto uno en las Notas⁴⁸. Shannon probó que la única función H que obedece sus

⁴⁸ Supón que tiro un dado y asingo los símbolos a, b, c de este modo:

- a. En el dado sale 1, 2, o 3
- b. En el dado sale 4 o 5
- c. En el dado sale 6

El símbolo a se da con una probabilidad de $1/2$, el símbolo b tiene probabilidad $1/3$, y el símbolo c tiene probabilidad $1/6$. Entonces mi fórmula, cualquiera que esta sea, asignará un contenido de información $H(1/2, 1/3, 1/6)$.

Sin embargo, podría pensar en este experimento de un modo diferente. Primero decido si en el dado sale algo menor o igual que 3, o algo mayor. Llama estas posibilidades q y r , de modo que:

- q En el dado sale 1, 2 o 3
- r En el dado sale 4, 5 o 6

Ahora q tiene probabilidad $1/2$ y r tiene probabilidad $1/2$. Cada uno transmite información $H(1/2, 1/2)$. El caso q es mi a original, y el caso r son mis b y c originales. Puedo dividir el caso r en b y c , y sus probabilidades son $2/3$ y $1/3$ asumiendo que ha ocurrido r . Si ahora consideramos solo este caso, la información transmitida por cualquiera que resulte ser b y c es $H(2/3, 1/3)$. Shannon ahora insiste en que la información original debería estar relacionada con la información en estos subcasos como sigue:

$$H(1/2, 1/3, 1/6) = H(1/2, 1/2) + 1/2 H(2/3, 1/3)$$

Véase la figura 61.

tres principios es:

$$H(p_1, p_2, \dots, p_n) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \dots - p_n \log_2 p_n$$

O una constante múltiplo de esta expresión, que básicamente solo cambia la unidad de información, como cambiar de pies a metros.

Hay una buena razón para tomar la constante como 1, y la ilustraré con un caso sencillo. Piensa en las cuatro cadenas binarias 00, 01, 10, 11 como símbolos por sí mismos. Si 0 y 1 son igualmente probables, cada cadena tiene la misma probabilidad, concretamente $\frac{1}{4}$. La cantidad de información transmitida por una elección de una cadena es por lo tanto:

$$\begin{aligned} H(1/4, 1/4, 1/4, 1/4) &= \\ &= -\frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{4} \log_2 \frac{1}{4} = \\ &= -\log_2 \frac{1}{4} = 2 \end{aligned}$$

Esto es, 2 bits. Que es un número sensato para la información en una cadena binaria de longitud 2 cuando las elecciones 0 y 1 son igual de probables. Del mismo modo, si los símbolos son todos cadenas binarias de longitud n , y fijamos la constante en 1, entonces la información contenida es n bits. Observa que cuando $n = 2$, obtenemos la fórmula representada en la figura 56. La prueba del teorema de Shannon es demasiado complicada para ponerla aquí, pero muestra que si aceptas las tres condiciones de Shannon, entonces hay un único modo natural de cuantificar

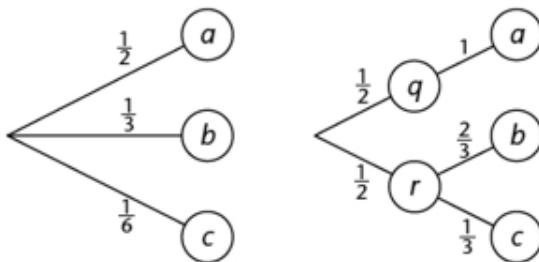


FIGURA 61. Elecciones combinadas de modos diferentes. La información debería ser la misma en cada caso.

El factor $\frac{1}{2}$ delante de la H final está presente porque esta segunda elección se da solo la mitad de las veces, concretamente cuando se escoge r en la primera etapa. No existe este factor delante de H justo después del signo igual, porque esto se refiere a la elección que siempre se hace, entre q y r .

la información.⁴⁹ La ecuación en sí misma es meramente una definición: lo que cuenta es cómo responde en la práctica.

Shannon usó su ecuación para probar que hay un límite fundamental en cuánta información puede transmitir un canal de información. Supongamos que estás transmitiendo una señal digital a lo largo de una línea de teléfono, cuya capacidad para llevar un mensaje es como mucho C bits por segundo. Esta capacidad está determinada por el número de dígitos binarios que la línea de teléfonos puede transmitir, y no está relacionada con las probabilidades de varias señales. Supón que el mensaje está siendo generado a partir de símbolos con el contenido de información H , también medido en bits por segundo. El teorema de Shannon responde a la pregunta: si el canal es ruidoso, ¿puede la señal codificarse de modo que la proporción de errores sea tan pequeña como queramos? La respuesta es que esto es siempre posible, no importa cuál sea el nivel de ruido, si H es menor o igual que C . Esto no es posible si H es mayor que C . De hecho, la proporción de errores no puede reducirse por debajo de la diferencia $H - C$, no importa el código que se emplee, pero existen códigos que se acercan tanto como quieras a esa tasa de error.

La prueba de Shannon de su teorema demuestra que los códigos del tipo necesario existen, en cada uno de sus dos casos, pero la prueba no nos dice cuáles son esos códigos. Una rama entera de la ciencia de la información, una mezcla de matemáticas, computación e ingeniería electrónica, está dedicada a encontrar códigos eficientes para propósitos específicos. Se llama la teoría de códigos. Los métodos para dar con estos códigos son muy diversos, aprovechándose de muchas áreas de las matemáticas. Son estos métodos los que se incorporan en nuestros aparatos electrónicos, ya sea un *smartphone* o el transmisor de la *Voyager 1*. La gente lleva de un lado a otro cantidades significativas de sofisticada álgebra abstracta en sus bolsillos, en la forma de software que implementa códigos de detección de errores para teléfonos móviles.

Intentaré transmitir el sabor de la teoría de códigos sin enredarme mucho en las complejidades. Uno de los conceptos más influenciables en la teoría relaciona

⁴⁹ Véase el capítulo 2 de C.E. Shannon y W. Weaver. *The Mathematical Theory of Communication*, University of Illinois Press, Urbana, 1964.

códigos con geometría multidimensional. Fue publicado por Richard Hamming en 1950 en un famoso artículo: «Error detecting and error correcting codes» (Códigos de detección y corrección de errores). En su forma más simple, proporciona una comparación entre cadenas de dígitos binarios. Considera dos de dichas cadenas, por ejemplo, 10011101 y 10110101. Compara los bits correspondientes y cuenta cuántas veces son diferentes, tal que así:

10 0 11101
101 1 0101

donde he marcado con negrita las diferencias. Aquí hay dos localizaciones en las que la cadena de bits difiere. Llamamos a este número la distancia de Hamming entre dos cadenas. Puede pensarse como el número más pequeño de errores de un bit que puede convertir una cadena en otra. De modo que está muy relacionado con el probable efecto de los errores, si estos se dan en una tasa media conocida. Lo que sugiere que podría proporcionar algo de comprensión sobre cómo detectar dichos errores y, quizá, cómo corregirlos.

La geometría multidimensional entra en juego porque las cadenas de una longitud fija pueden asociarse con los vértices de un «hipercubo» multidimensional. Riemann nos enseñó cómo pensar en dichos espacios pensando en una lista de números. Por ejemplo, un espacio de cuatro dimensiones consiste en todas las listas de cuatro números posibles: (x_1, x_2, x_3, x_4) . Cada lista se considera que representa un punto en el espacio, y todas las listas posibles pueden, en principio, darse. Cada una de las x son las coordenadas del punto. Si el espacio tiene 157 dimensiones, tiene que usar listas de 157 números: $(x_1, x_2, \dots, x_{157})$. Es con frecuencia útil especificar cuán separadas están estas listas. En la geometría «plana» de Euclides, esto se hacía usando una generalización simple del teorema de Pitágoras. Supón que tenemos un segundo punto $(y_1, y_2, \dots, y_{157})$ en nuestro espacio 157-dimensional. Entonces la distancia entre los dos puntos es la raíz cuadrada de la suma de los cuadrados de las diferencias entre las coordenadas correspondientes. Esto es:

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_{107} - y_{107})^2}$$

Si el espacio es curvado, se puede usar en su lugar la idea de Riemann de una métrica.

La idea de Hamming es hacer algo muy similar, pero los valores de las coordenadas se restringen solo a 0 y 1. Entonces $(x_1 - y_1)^2$ es 0 si x_1 e y_1 son lo mismo, y 1 si no lo son, y lo mismo aplica para $(x_2 - y_2)^2$, etcétera. También omitió la raíz cuadrada, la cual cambia la respuesta, pero en compensación el resultado es siempre un número natural, igual a la distancia de Hamming. Esta noción tiene todas las propiedades que hacen la «distancia» útil, como ser cero solo cuando las dos cadenas son idénticas, y asegurar que la longitud de cualquier lado de un «triángulo» (un conjunto de tres cadenas) es menor o igual que la suma de las longitudes de los otros dos lados.

Podemos dibujar imágenes de todas las cadenas de bits de longitudes 2, 3 y 4 (y con más esfuerzo y menos claridad, 5, 6 y posiblemente incluso 10, aunque nadie la encontraría útil). Los diagramas resultantes se muestran en la figura 57.

Los dos primeros son reconocibles como un cuadrado y un cubo (proyectado en un plano porque tiene que imprimirse en una hoja de papel). El tercero es un hipercubo, el análogo en 4 dimensiones y, de nuevo, tiene que proyectarse en un plano. Las líneas rectas uniendo los puntos tienen longitud de Hamming 1, las dos cadenas en cada extremo difieren en precisamente una localización, una coordenada. La distancia de Hamming entre dos cadenas cualesquiera es el número de dichas líneas en la ruta más corta que las conecta.

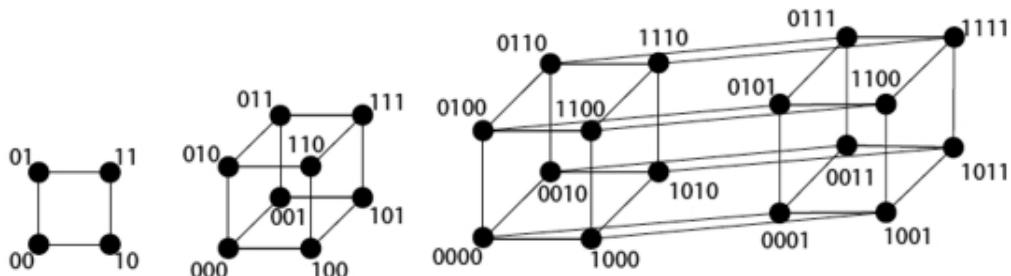


FIGURA 57. Los espacios de todas las cadenas de bits de longitudes 2, 3 y 4.

Supón que estamos pensando en cadenas de 3 bits, presentes en las esquinas de un cubo. Selecciona una de las cadenas, por ejemplo 101. Supón que la tasa de errores es como mucho un bit cada tres. Entonces esta cadena podría o bien transmitirse sin cambios, o bien podría acabar como cualquiera de estas: 001, 111 o 100. Cada una de estas difiere de la cadena original en tan solo una localización, de modo que su distancia de Hamming de la cadena original es 1. La esfera consiste en solo tres puntos, y si estuviésemos trabajando en un espacio de 157 dimensiones con un radio de 5, por ejemplo, ni siquiera parecería demasiado esférica. Pero juega un rol similar a una esfera ordinaria: tiene una forma bastante compacta, y contiene exactamente los puntos cuya distancia desde el centro es menor o igual que el radio.

Supón que usamos las esferas para construir un código, de modo que cada esfera se corresponde con un símbolo nuevo, y ese símbolo está codificado con las coordenadas del centro de la esfera. Supón además que estas esferas no se superponen. Por ejemplo, podría introducir un símbolo *a* para la esfera con centro en 1010. Esta esfera contiene cuatro cadenas: 101, 001, 111 y 100. Si recibo cualquiera de estas cuatro cadenas, sé que el símbolo era originalmente *a*. Al menos, eso es cierto siempre que mis otros símbolos se correspondan de un modo similar a las esferas que no tienen ningún punto en común con esta.

Ahora la geometría comienza a hacerse útil. En el cubo hay ocho puntos (cadenas) y cada esfera contiene cuatro de ellos. Si intento encajar las esferas en el cubo, sin que se superpongan, el mejor resultado que puedo lograr es con dos de ellas, porque $8/4 = 2$. Realmente puedo encontrar otra, concretamente la esfera centrada en 010. Esta contiene 010, 110, 000, 011, ninguno de los cuales está en la primera esfera. De manera que se puede introducir un segundo símbolo *b* asociado con esta esfera. Mi código de corrección de errores escrito con los símbolos *a* y *b* ahora remplaza todo *a* con 101, y todo *b* con 010. Si recibo, digamos:

101-010-100-101-000

entonces puedo decodificar el mensaje original como:

$$a-b-a-a-b$$

a pesar de los errores en la tercera y la quinta cadena. Acabo de ver cuáles de mis dos esferas pertenecen las cadenas erróneas.

Todo está muy bien, pero esto multiplica la longitud del mensaje por 3, y ya sabemos un modo más fácil de lograr el mismo resultado: repetir el mensaje tres veces. Pero la misma idea adquiere un nuevo significado si trabajamos en espacios de dimensiones mayores. Con cadenas de longitud 4, el hipercubo, hay 16 cadenas, y cada esfera contiene 5 puntos. De modo que podría ser posible encajar tres esferas sin que se solapen. Si lo intentas, resulta que no es posible realmente, dos encajan pero el hueco que queda tiene la forma equivocada. Pero cada vez más números funcionan a nuestro favor. El espacio de cadenas de longitud 5 contiene 32 cadenas, y cada esfera usa solo 6 de ellas, posiblemente haya hueco para 5, y si no, una posibilidad mejor de encajar 4. Longitud 6 nos da 64 puntos y esferas que usan 7, de modo que pueden encajar hasta 9 esferas.

A partir de este punto son necesarios muchos detalles complicados para averiguar solo qué es lo que es posible, y ayuda a desarrollar métodos más sofisticados. Pero lo que estamos observando es análogo, en el espacio de cadenas, a las maneras más eficientes de agrupar esferas unas con otras. Y esto es una vieja área de las matemáticas, sobre la que se conoce bastante. Algunas de esas técnicas pueden transferirse de la geometría euclídea a las distancias de Hamming, y cuando eso no funciona, podemos inventar nuevos métodos más apropiados para la geometría de cadenas. Como ejemplo, Hamming inventó un código nuevo, más eficiente que cualquiera conocido en la época, el cual codifica cadenas de 4 bits, convirtiéndolas en cadenas de 7 bits. Puede detectarse y corregirse cualquier error de un único bit. Modificado a un código de 8 bits, puede detectarse, pero no corregirse, cualquier error de 2 bits.

Este código se llama el código de Hamming. No lo describiré, pero hagamos las operaciones para ver si podría ser posible. Hay 16 cadenas de longitud 4, y 128 de longitud 7. Las esferas de radio 1 en el hipercubo de 7 dimensiones contienen 8 puntos. Y $128/8 = 16$. De modo que con suficiente ingenio, podría ser posible meter las 16 esferas necesarias en el hipercubo de 7 dimensiones. Tendrían que encajar

exactamente, porque no queda hueco libre. Al parecer, existe dicha colocación, y Hamming la encontró. Sin la geometría multidimensional como ayuda, sería difícil adivinar que existe, por no hablar de encontrarla. Posible, pero duro. Incluso con la geometría no es obvio.

El concepto de Shannon de información proporciona un límite en cómo de eficientes pueden ser los códigos. La teoría de códigos hace la otra mitad del trabajo: encontrar códigos que sean lo más eficientes posible. Las herramientas más importante aquí vienen del álgebra abstracta. Este es el estudio de las estructuras matemáticas que comparten las características aritméticas básicas de los números enteros o reales, pero difieren de ellos de maneras significativas. En aritmética, podemos sumar números, restarlos y multiplicarlos, para obtener números del mismo tipo. Para los números reales, podemos también dividir por cualquier cosa diferente de cero para obtener un número real. Esto no es posible para los enteros, porque por ejemplo $\frac{1}{2}$ no es un entero. Sin embargo, las fracciones son posibles si pasamos al sistema más grande de los números racionales. En los sistemas numéricos habituales, se soportan varias leyes algebraicas, por ejemplo, la propiedad conmutativa de la adición, que afirma que $2 + 3 = 3 + 2$ y lo mismo aplica para cualquier par de números.

Los sistemas comunes comparten estas propiedades algebraicas con los menos comunes. El ejemplo más simple usa solo dos números, 0 y 1. Las sumas y los productos están definidos solo como para los enteros, con una excepción: insistimos en que $1 + 1 = 0$, no 2. A pesar de esta modificación, todas las leyes habituales del álgebra sobreviven. El sistema tiene solo dos «elementos», dos objetos como los números. Hay exactamente uno de dichos sistemas siempre que el número de elementos sea una potencia de cualquier número primo: 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, etcétera. Dichos sistemas se llaman cuerpos finitos o campos de Galois, por el matemático francés Évariste Galois, que los clasificó alrededor de 1830. Porque tienen un número finito de elementos, son adecuados para las comunicaciones digitales, y las potencias de 2 son especialmente convenientes debido a la notación binaria.

Los cuerpos finitos llevan a sistemas codificados llamados códigos de Reed-Solomon, por Irving Reed y Gustave Solomon, quienes los inventaron en 1960. Se

usan en los aparatos electrónicos de consumo, especialmente en CDs y DVDs. Son códigos de corrección de errores basados en propiedades algebraicas de polinomios, cuyos coeficientes se toman de los cuerpos finitos. La señal una vez codificada, audio o vídeo, se usa para construir un polinomio. Si el polinomio tienen grado n , esto es, si la mayor potencia que aparece es x^n , entonces el polinomio puede reconstruirse a partir de sus valores en n puntos cualquiera. Si especificamos los valores en más de n puntos, podemos perder o modificar algunos de los valores sin perder la pista de qué polinomio es. Si el número de errores no es demasiado grande, es todavía posible averiguar qué polinomio es, y decodificarlo para obtener los datos originales.

En la práctica la señal se representa como una serie de bloques de dígitos binarios. Una elección popular usa 255 bytes (una cadena de 8 bits) por bloque. De estos, 223 bytes codifican la señal, mientras los restantes 32 bytes son los «símbolos de paridad», que nos dicen si varias combinaciones de dígitos en los datos incorruptos son pares o impares. Este código de Reed-Solomon en concreto puede corregir hasta 16 errores por bloque, una tasa de error menor del 1 %.

Siempre que conduces a lo largo de una carretera llena de baches con un CD en el reproductor del coche, estás usando álgebra abstracta, en la forma del código de Reed-Solomon, para asegurarte de que la música se escucha limpia y clara, en lugar de entrecortada y chirriante, quizá saltándose algunas partes.

La teoría de la información se usa ampliamente en criptografía y criptoanálisis, códigos secretos y métodos para descifrarlos. El propio Shannon la usó para estimar la cantidad de mensajes codificados que deben ser interceptados para tener alguna posibilidad de descifrar el código. Mantener la información secreta resulta ser más difícil de lo que se esperaba, y la teoría de la información arroja luz sobre este problema, tanto desde el punto de vista de la gente que quiere mantenerla en secreto como del de aquellos que quieren averiguar qué es. El asunto es importante no solo en el ejército, sino para cualquiera que use Internet para comprar o contrate la banca telefónica.

La teoría de la información ahora juega un rol importante en biología, particularmente en el análisis de la secuencia del ADN. La molécula del ADN es una doble hélice, formada por dos hebras que se enroscan una alrededor de la otra.

Cada hebra es una secuencia de bases, moléculas especiales que se dan en cuatro tipos: adenina, guanina, tiamina y citosina. De modo que el ADN es como un mensaje codificado escrito usando los cuatro símbolos posibles: A, G, T y C. El genoma humano, por ejemplo, tiene una longitud de 3.000 millones de bases. Los biólogos pueden ahora encontrar las secuencias del ADN de innumerables organismos a una velocidad cada vez más rápida, llevando a una nueva área de la informática: la bioinformática. Esta se centra en métodos para manejar los datos biológicos de manera eficiente y efectiva, y una de sus herramientas básicas es la teoría de la información.

Un asunto más complicado es la calidad de la información, más que la cantidad. Los mensajes «dos más dos son cuatro» y «dos más dos son cinco» contienen exactamente la misma cantidad de información, pero uno es cierto y el otro es falso. Los cantos de alabanza por la era de la información ignoran la verdad incómoda de que mucha de la información merodeando en Internet es desinformación. Hay *websites* que las llevan criminales que quieren robar tu dinero, o negacionistas que quieren remplazar la ciencia sólida por lo que sea que tienen metido entre ceja y ceja.

El concepto vital aquí no es la información como tal, sino el significado. Tres mil millones de bases de ADN de la información del ADN humano carecen, literalmente, de significado a menos que puedas averiguar cómo afectan a nuestro cuerpo y comportamiento. En el décimo aniversario de la finalización del Proyecto Genoma Humano, varias publicaciones científicas destacadas examinaron los progresos médicos resultantes, hasta el momento, de la enumeración de las bases del ADN humano. El tono general fue silenciado: se habían encontrado unas pocas nuevas curas para enfermedades, pero no en la cantidad originalmente predicha. Extraer significado de la información del ADN ha resultado ser más duro de lo que los biólogos habían esperado. El Proyecto Genoma Humano era un primer paso necesario, pero, más que resolverlos, solo ha revelado lo difíciles que son dichos problemas.

La noción de la información ha escapado de la ingeniería electrónica y ha invadido muchas otras áreas de la ciencia, ambas tanto como una metáfora y como un concepto técnico. La fórmula para la información se parece mucho a la de la

entropía en la aproximación de Boltzmann a la termodinámica, las principales diferencias son los logaritmos de base 2 en lugar de los logaritmos neperianos, y un cambio en el signo. Se puede formalizar esta similitud, y la entropía se puede interpretar como «información perdida». De modo que la entropía de un gas aumenta porque perdemos la noción de dónde están las moléculas exactamente, y con qué rapidez se mueven. La relación entre la entropía y la información tiene que fijarse con mucho cuidado; aunque las fórmulas son muy parecidas, el contexto en el que se aplican es diferente. La entropía de la termodinámica es una propiedad a gran escala del estado de un gas, pero la información es una propiedad de una fuente que produce señales, no de una señal como tal. En 1957, el físico americano Edwin Jaynes, un experto en mecánica estadística, resumió la relación: la entropía de la termodinámica puede verse como una aplicación de la información de Shannon, pero la entropía en sí misma no debería identificarse con información perdida sin especificar el contexto correcto. Si esta distinción se tiene en mente, hay varios contextos válidos en los que la entropía puede verse como una pérdida de información. Justo como el incremento de entropía pone límites en la eficiencia de los motores a vapor, la interpretación entrópica de la información pone límites a la eficiencia de los cómputos. Por ejemplo, se necesitan al menos $5,8 \times 10^{-23}$ julios de energía para convertir un bit de 0 a 1 o viceversa a la temperatura del helio líquido, cualquiera que sea el método que se utilice.

Los problemas surgen cuando las palabras «información» y «entropía» se usan en un sentido más metafórico. Los biólogos con frecuencia dicen que el ADN determina «la información» necesaria para formar un organismo. Hay un sentido en el que esto es casi correcto: elimina el «la». Sin embargo, la interpretación metafórica de información sugiere que una vez que conoces el ADN, entonces lo conoces todo sobre ese organismo. Después de todo, tienes *la* información, ¿no? Y durante un tiempo muchos biólogos pensaron que esta afirmación era cercana a la verdad. Sin embargo, ahora sabemos que es demasiado optimista. Incluso aunque la información en el ADN realmente especificase al organismo de manera única, todavía necesitaría averiguar cómo crece y qué hace el ADN. Pero se necesita mucho más que una lista de código de ADN para crear un organismo, los conocidos como factores epigenéticos deben, también, tenerse en cuenta. Estos incluyen

«cambios» químicos que hacen activo o inactivo un segmento de código de ADN, pero también factores totalmente diferentes que se transmiten de padres a hijos. Para los seres humanos, estos factores incluyen la cultura en la que crecen. De modo que no resulta ser tan superficial el uso de términos técnicos como «información».

Capítulo 16
El desequilibrio de la naturaleza
Teoría del caos

The diagram shows the logistic map equation $X_{t+1} = kX_t(1-X_t)$ on a light gray background. Four arrows point from text labels to specific parts of the equation:

- An arrow from "tamaño de la población" points to the variable X_t .
- An arrow from "de la generación siguiente" points to the term X_{t+1} .
- An arrow from "tasa de crecimiento sin restricción" points to the coefficient k .
- An arrow from "ahora" points to the term $(1-X_t)$.
- A vertical line labeled "tamaño de la población" is positioned under the X_t term.
- A vertical line labeled "ahora" is positioned under the $(1-X_t)$ term.

¿Qué dice?

Hace un modelo de cómo una población de criaturas vivas cambia de una generación a la siguiente, cuando hay límites en los recursos disponibles.

¿Por qué es importante?

Es una de las ecuaciones más simples que puede generar el caos determinista, comportamiento aparentemente aleatorio con causas no aleatorias.

¿Qué provocó?

La comprensión de que ecuaciones no lineales sencillas pueden crear dinámicas muy complejas, y que esa aleatoriedad aparente podría ocultar un orden escondido. Popularmente conocida como teoría del caos, este descubrimiento tiene innumerables aplicaciones en toda la ciencia, incluyendo el movimiento de los planetas del Sistema Solar, la predicción del tiempo, la dinámica de poblaciones en ecología, las estrellas variables, el modelado de terremotos y trayectorias eficientes para las sondas espaciales.

La metáfora del equilibrio de la naturaleza aparece inmediatamente de manera natural como una descripción de qué haría el mundo si los malvados humanos dejasesen de interferir. La naturaleza, dejando que se las arregle por su cuenta, establecería un estado de perfecta armonía. Los arrecifes de coral esconderían

siempre las mismas especies de peces de colores en cantidades similares, los conejos y los zorros aprenderían a compartir los campos y los bosques de modo que los zorros se alimentasen bien, la mayoría de los conejos sobreviviesen y ninguna población se disparase o extinguiese. El mundo se establecería en un estado fijo y se quedaría ahí. Hasta que el siguiente gran meteorito o un supervolcán alterasen el equilibrio.

Es una metáfora común, peligrosamente cercana a ser un cliché. También muy engañosa. El equilibrio de la naturaleza claramente se tambalea.

Hemos estado aquí antes. Cuando Poincaré estaba trabajando en el premio del rey Óscar, la sabiduría popular sostenía que un Sistema Solar estable es uno en el que los planetas siguen prácticamente las mismas órbitas para siempre, dando o recibiendo una pequeña e inocua sacudida. Técnicamente esto no es un estado estable, sino uno en el que cada planeta repite movimientos parecidos una y otra vez, sujeto a perturbaciones menores provocadas por los demás, pero no desviándose enormemente de lo que habría hecho sin ellos. La dinámica es «cuasiperiódica», combinando varios movimientos periódicos separados cuyos períodos no son todos múltiplos del mismo intervalo de tiempo. En el reino de los planetas, esto es lo más cerca a «estable» que se puede esperar.

Pero las dinámicas no son así, como Poincaré tardíamente, y a su costa, averiguó. Podía, en las circunstancias correctas, ser caótico. Las ecuaciones no tienen términos aleatorios explícitos, de modo que en principio el estado presente determina completamente el estado futuro, aunque paradójicamente el movimiento actual podría aparentar ser aleatorio. De hecho, si haces preguntas muy generales como «¿sobre qué lado del Sol estará?», la respuesta podrá ser realmente una serie aleatoria de observaciones. Solo si pudieses observar de muy, muy, muy cerca, serías capaz de ver que el movimiento realmente estaba completamente determinado.

Este fue el primer indicio de lo que ahora llamamos «caos», que es el modo corto para «caos determinista», y bastante diferente de «aleatorio», incluso aunque eso es lo que puede parecer. La dinámica caótica ha escondido patrones, pero son sutiles, difieren de lo que de manera natural podríamos pensar en medir. Solo con la comprensión de las causas del caos podemos extraer esos patrones de un

revoltijo irregular de datos.

Como siempre en ciencia, hubo unos pocos precursores aislados, generalmente vistos como curiosidades menores, que no merecían una atención seria. Solo en la década de los sesenta del siglo XX, los matemáticos, físicos e ingenieros empezaron a darse cuenta de cómo de natural es el caos en la dinámica, y cómo difiere radicalmente de cualquier cosa imaginada en la ciencia clásica. Todavía estamos aprendiendo a apreciar qué nos dice, y qué hacer con ello. Pero la dinámica caótica, la «teoría del caos» en lenguaje popular, ya invade la mayoría de las áreas de la ciencia. Podría incluso tener cosas que decírnos sobre la economía y las ciencias sociales. No es la respuesta a todo, solo los críticos reclamaron alguna vez que lo fuese, y eso fue para hacer más fácil derribarla. El caos ha sobrevivido a todos esos ataques y por una buena razón: es absolutamente fundamental para todo comportamiento gobernado por las ecuaciones diferenciales, y estas son lo básico de las leyes físicas.

Hay caos también en biología. Uno de los primeros en apreciar que esto podría ser posible fue el ecologista australiano Robert May, ahora Lord May de Oxford y expresidente de la Royal Society. Intentaba comprender cómo las poblaciones de varias especies cambiaban a lo largo del tiempo en los sistemas naturales como los arrecifes de coral y los bosques. En 1975, May escribió un pequeño artículo para la revista *Nature*, señalando que las ecuaciones usadas habitualmente en los cambios de modelo de las poblaciones de animales y plantas podían producir caos. May no afirmó que los modelos que estaba discutiendo eran representaciones precisas de lo que las poblaciones reales hacían. Su argumento era más general: el caos era natural en modelos de ese tipo y tenía que tenerse en cuenta.

La consecuencia más importante del caos es que ese comportamiento irregular no necesitaba causas irregulares. Previamente, si los ecólogos notaban que alguna población de animales estaba fluctuando sin orden, buscarían alguna causa externa, también supuesta de estar fluctuando sin orden y, generalmente, etiquetada como «aleatoriedad». El tiempo, quizás, o una repentina afluencia de depredadores de algún lugar. Los ejemplos de May mostraron que el funcionamiento interno de las poblaciones de animales podía generar irregularidades sin la ayuda externa.

Su principal ejemplo fue la ecuación que decora la apertura de este capítulo. Es

llamada la ecuación logística, y es un modelo simple de una población de animales en la que el tamaño de cada generación está determinado por el de la anterior. «Discreta» significa que el flujo del tiempo se cuenta en generaciones, y es de este modo un entero. Así que el modelo es similar a la ecuación diferencial, en la que el tiempo es una variable continua, pero conceptualmente y computacionalmente más simple. La población se mide como una fracción de algún valor mayor total, y puede de ese modo representarse por un número real que se encuentra entre 0 (extinción) y 1 (el máximo teórico que el sistema puede mantener). Permitiendo al tiempo t avanzar en pasos enteros, correspondientes a generaciones, este número es x_t en la generación t . La ecuación logística afirma que:

$$x_{t+1} = kx_t(1 - x_t)$$

donde k es una constante. Podemos interpretar k como la tasa de crecimiento de la población cuando los recursos cada vez menores no la frenan.⁵⁰

Empezamos el modelo en el tiempo 0 con una población inicial x_0 . Luego usamos la ecuación con $t = 0$ para calcular x_1 , luego hacemos $t = 1$ y hallamos x_2 , y así sucesivamente. Sin ni siquiera hacer los cálculos, podemos ya ver que, para cualquier tasa de crecimiento fija k , el tamaño de la población de la generación cero determina totalmente los tamaños de todas las generaciones sucesivas. De modo que el modelo es determinista: el conocimiento del presente determina el futuro de manera única y exacta.

Así que, ¿qué es el futuro? La metáfora del «equilibrio de la naturaleza» sugiere que la población debería fijar un estado estable. Podemos incluso calcular qué estado estable debería ser; basta establecer la población en el momento $t + 1$ para que sea la misma que en el momento t . Esto conduce a dos estados estables: poblaciones 0 y poblaciones $1 - \frac{1}{k}$. Una población de tamaño 0 está extinguida, de modo que otro valor debería aplicarse a las poblaciones existentes. Desafortunadamente, aunque esto es un estado estable, puede ser inestable. Si es así, entonces en la

⁵⁰ Si la población x_t es relativamente pequeña, de modo que es cercana a cero, entonces $1 - x_t$ es cercano a 1. La siguiente generación tendrá, por lo tanto, un tamaño cercano a kx_t , que es k veces tan grande como la actual. A medida que el tamaño de la población aumenta, el factor extra $1 - x_t$ hace la tasa de crecimiento real más pequeña, y cae a cero a medida que la población se aproxima a su máximo teórico.

práctica nunca lo verás; es como intentar equilibrar un lápiz verticalmente sobre la punta afilada. La alteración más ligera provocará su caída. Los cálculos muestran que el estado estable es inestable cuando k es mayor que 3.

Entonces, ¿qué vemos en la práctica? La figura 58 muestra una «serie de tiempo» típica para la población cuando $k = 4$. No es estable, está disperso. Sin embargo, si observas de cerca, hay pistas de que la dinámica no es totalmente aleatoria. Siempre que la población se hace realmente grande, inmediatamente cae estrepitosamente a un valor muy bajo, y luego crece de manera regular (más o menos exponencialmente) durante las siguientes dos o tres generaciones (véanse las flechas cortas en la figura 58). Y sucede algo interesante siempre que la población se mantenga cerca del entorno de 0,75; oscila alternativamente por encima y por debajo del valor, y las oscilaciones crecen dando una forma de zigzag característica, que se hace más ancha a la derecha (véanse las flechas largas de la figura).

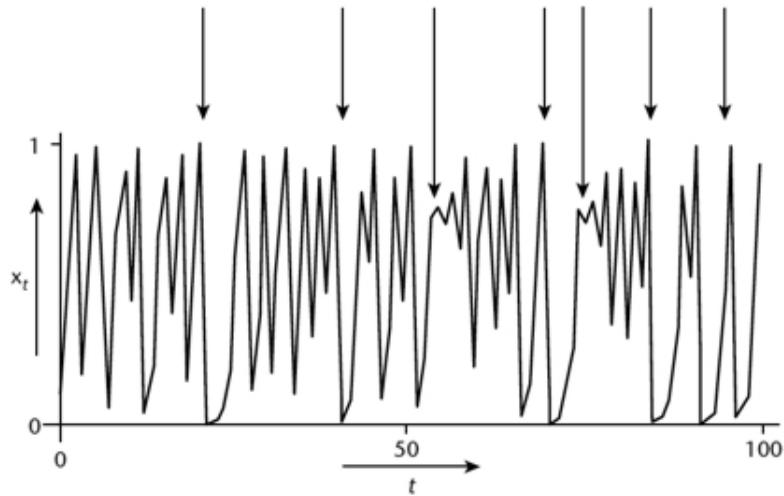


FIGURA 58. Oscilaciones caóticas en un modelo de población animal. Las flechas cortas muestran caídas estrepitosas seguidas por períodos cortos de crecimiento exponencial. Las flechas más largas muestran oscilaciones inestables.

A pesar de estos patrones, en cierto modo, cuando no se tienen en cuenta los detalles, el comportamiento es realmente aleatorio. Supón que asignamos el símbolo C (caras) siempre que la población sea mayor que 0,5, y + (cruces) cuando

es menor que 0,5. Este conjunto de datos concreto empieza con la secuencia $+C+C+CC+CC++CC$ y continúa de modo impredecible, justo como una secuencia aleatoria de lanzamientos de moneda. Esta manera de embrutecer los datos, observando rangos específicos de valores y anotando solo a qué rango pertenece la población, se llama dinámica simbólica. En este caso, es posible probar que, para la mayoría de los valores de población iniciales x_0 , la secuencia de caras y cruces es, en todos los aspectos, como una secuencia típica de lanzamientos aleatorios de una moneda no trucada. Solo cuando observamos los valores exactos, empezamos a ver algunos patrones.

Es un descubrimiento sorprendente. Un sistema dinámico puede ser totalmente determinista, con patrones visibles en los datos detallados, aunque una visión *grosso modo* de los mismos datos puede ser aleatoria, en un sentido demostrable y riguroso. Determinismo y aleatoriedad no son opuestos. En algunas circunstancias, pueden ser totalmente compatibles.

May no inventó la ecuación logística, y no descubrió sus propiedades sorprendentes. No reivindicó haber hecho ninguna de esas cosas. Su propósito era alertar a los investigadores de las ciencias de la vida, especialmente a los ecologistas, de los descubrimientos excepcionales que surgían en las ciencias físicas y las matemáticas; descubrimientos que fundamentalmente cambian el modo en que los científicos deberían pensar en los datos observables. Nosotros, los humanos, podemos tener problemas resolviendo ecuaciones basadas en reglas simples, pero la naturaleza no tiene que resolver ecuaciones del modo en que nosotros lo hacemos. Tan solo obedece las reglas. Así puede hacer cosas que nos parecen complicadas, por razones simples.

El caos surgió de una aproximación topológica a la dinámica, dirigida en concreto por el matemático americano Stephen Smale y el matemático ruso Vladimir Arnold en la década de los sesenta del siglo XX. Ambos estaban tratando de averiguar qué tipos de comportamiento eran típicos en ecuaciones diferenciales. Smale estaba motivado por los extraños resultados de Poincaré en el problema de los tres cuerpos (capítulo 4), y Arnold estaba inspirado por los descubrimientos relacionados de su antiguo supervisor de la investigación, Andrei Kolmogorov. Ambos rápidamente se dieron cuenta de por qué el caos es común: es una consecuencia natural de la

geometría de las ecuaciones diferenciales, como veremos en un momento.

A medida que el interés en el caos se extendía, se descubrían ejemplos al acecho que habían pasado inadvertidos en artículos científicos anteriores. Previamente considerados como efectos raros aislados, estos ejemplos ahora encajaban en una teoría más amplia. En la década de los cuarenta del siglo XX, los matemáticos ingleses John Littlewood y Mary Cartwright habían visto trazas de caos en osciladores electrónicos. En 1958, Tsuneji Rikitake, de la Asociación para el desarrollo de la predicción de terremotos de Tokio, había encontrado comportamiento caótico en un modelo de la dinamo del campo magnético terrestre. Y en 1963, el meteorólogo americano Edward Lorenz había determinado la naturaleza de la dinámica caótica con un considerable detalle, en un modelo sencillo de convección atmosférica motivado por la predicción del tiempo. Estos y otros pioneros han indicado el camino, ahora todos sus disparatados descubrimientos estaban empezando a encajar unos con otros.

En concreto, las circunstancias que llevaron al caos, en lugar de a algo más sencillo, resultaron ser geométricas más que algebraicas. En el modelo logístico con $k = 4$, ambos extremos de la población, 0 y 1, se mueven a 0 en la siguiente generación, mientras que el punto medio, $\frac{1}{2}$, se mueve a 1. De modo que en cada paso de tiempo, el intervalo de 0 a 1 se estira el doble de su longitud, se dobla por la mitad, y se planta en su localización original. Esto es lo que hace un cocinero al amasar cuando hace pan, y pensando en la masa siendo amasada, ganamos comprensión del caos. Imagina una mota diminuta en la masa logística, una uva pasa, por ejemplo. Supón que aparece en ciclos periódicos, de modo que después de un cierto número de estiramientos y pliegues, vuelve a donde empezó. Ahora podemos ver por qué este punto es inestable. Imagina otra uva pasa, inicialmente muy cercana a la primera. Cada estiramiento la aleja. Aunque durante un tiempo, no se mueve lo suficientemente lejos como para dejar de seguir la pista a la primera. Cuando la masa se dobla, ambas pasas acaban en la misma capa. La siguiente vez, la segunda pasa se ha movido todavía más lejos de la primera. Esto es por qué el estado periódico es inestable; los estiramientos mueven todos los puntos cercanos lejos de ella, no hacia ella. Finalmente la expansión se hace tan grande que las dos pasas acaban en diferentes capas cuando la masa se dobla. Después de eso, sus destinos

son bastante independientes el uno del otro. ¿Por qué un cocinero amasa la masa? Para mezclar los ingredientes (incluyendo el aire atrapado). Si mezclas las cosas, las partículas individuales tienen que moverse de un modo muy irregular. Partículas que empiezan cerca la una de la otra acaban apartadas; puntos que estaban muy separados puede que al doblarse acaben cerca el uno del otro. En resumen, el caos es el resultado natural de mezclar.

Puedes pensar que no tienes nada caótico en tu cocina, excepto un lavaplatos quizá. Es falso. Probablemente tienes varios aparatos caóticos: un robot de cocina, un batidor de huevos. La cuchilla del robot de cocina sigue una regla sencilla: girar y girar, rápido. La comida interactúa con la cuchilla, debería hacer algo sencillo también. Pero no gira y gira, sino que se mezcla. A medida que la cuchilla corta la comida, algunos trozos van hacia un lado y otros hacia otro; localmente la comida se separa. Pero no escapa del bol, de modo que se vuelve a plegar sobre sí misma. Smale y Arnold se dieron cuenta de que toda la dinámica caótica es así. No expresaron sus resultados en ese lenguaje, a saber, «separarse» era «exponente positivo de Lyapunov» y «volverse a plegar» era «el sistema tiene un dominio compacto». Pero con lenguaje imaginativo, lo que estaban diciendo es que el caos es como una masa trabajada.

Esto también explica algo más, de lo que se dio cuenta Lorenz en 1963. La dinámica caótica es sensible a las condiciones iniciales. Por muy cerca que estén al principio las dos uvas pasas, finalmente se separarán tanto que sus movimientos posteriores serán independientes. Este fenómeno se llama con frecuencia el efecto mariposa: una mariposa agita sus alas y un mes más tarde el tiempo es totalmente diferente de lo que habría sido de otro modo. La frase normalmente se le atribuye a Lorenz. No es suya, pero aparece algo similar en el título de una de sus conferencias. Sin embargo, algún otro inventó el título para él, y la conferencia no fue sobre el famoso artículo de 1963, sino sobre uno menos conocido del mismo año.

Como sea que se denomine al fenómeno, tiene una consecuencia práctica importante. Aunque la dinámica caótica es en principio determinista, en la práctica se hace impredecible muy rápido, porque cualquier duda en el estado inicial exacto crece exponencialmente rápido. Hay un horizonte de predicciones más allá del cual el futuro no puede presagiarse. Para el tiempo, un sistema común cuyos modelos

informáticos estándar son conocidos por ser caóticos, este horizonte es unos pocos días por delante. Para el Sistema Solar, es decenas de millones de años por delante. Para juguetes de laboratorio sencillos, como un péndulo doble (un péndulo colgando del extremo de otro) es unos pocos segundos más allá. La suposición que ha existido durante mucho tiempo de que «determinista» y «predecible» eran lo mismo está equivocada. Sería válido si el estado presente de un sistema pudiese medirse con una precisión perfecta, pero eso no es posible.

La previsibilidad a corto plazo del caos puede usarse para distinguirla de la aleatoriedad pura. Se han diseñado muchas técnicas diferentes para hacer esta distinción y averiguar la dinámica subyacente si el sistema se está comportando de manera determinística pero caótica.

El caos tiene ahora aplicaciones en todas las ramas de la ciencia, desde la astronomía a la zoología. En el capítulo 4, vimos cómo está llevando a trayectorias nuevas y más eficientes para las misiones espaciales. En términos más generales, los astrónomos Jack Wisdom y Jacques Laskar han demostrado que la dinámica del Sistema Solar es caótica. Si quieres saber cuál será el paradero de Plutón en su órbita en el año 10.000.000, olvídalos. También han demostrado que las mareas de la Luna estabilizan la Tierra contra las influencias que de otro modo llevarían a un movimiento caótico, provocando rápidos cambios de clima de períodos cálidos a épocas de hielo y de vuelta al período cálido. De modo que la teoría del caos demuestra que, sin la Luna, la Tierra sería un lugar bastante desagradable para vivir. Esta característica de nuestro vecindario planetario es con frecuencia usada para argumentar que la evolución de la vida en un planeta necesita una Luna estabilizadora, pero esto es una exageración. La vida en los océanos apenas notaría si los ejes del planeta cambian en un período de millones de años. La vida en tierra firme tendría mucho tiempo para emigrar a otro lugar, a menos que se viese atrapada en algún lugar que careciese de una ruta terrestre a un lugar donde las condiciones fuesen más adecuadas. El cambio climático está sucediendo mucho más rápido ahora que cualquier cambio que pudiese ser provocado por un cambio en la inclinación de los ejes.

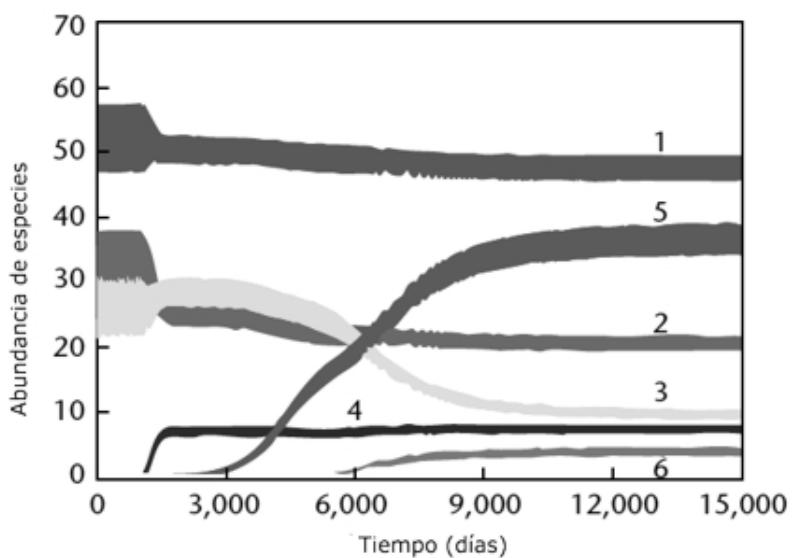


FIGURA 59. Seis especies compartiendo tres recursos. Las bandas son oscilaciones caóticas espaciadas estrechamente. Cortesía de Jef Huisman y Franz Weissing.

La sugerencia de May de que la dinámica de población irregular en un ecosistema podría a veces provocar el caos interno, más que la aleatoriedad superflua, ha sido verificada en versiones de laboratorio de varios ecosistemas del mundo real. En 1995, un equipo liderado por el ecologista norteamericano James Cushing encontró dinámicas caóticas en poblaciones de falsos gorgojos de la harina, *Tribolium castaneum*, las cuales pueden infestar provisiones de harina.⁵¹ En 1999, los biólogos holandeses Jef Huisman y Franz Weissing aplicaron el caos a la «paradoja del plancton», la diversidad inesperada de las especies de plancton.⁵² Un principio estándar en ecología, el principio de exclusión competitiva, afirma que un ecosistema no puede contener más especies que el número de nichos medioambientales, modos de ganarse la vida. El plancton parece violar este principio, el número de nichos es pequeño, pero el número de especies es de miles. Rastrearon esto hasta una laguna en la deducción del principio de exclusión competitiva: la suposición de que las poblaciones son estables. Si las poblaciones pueden cambiar con el tiempo, entonces la deducción matemática a partir de los

⁵¹ R.F. Costantino, R.A. Desharnais, J.M. Cushing y B. Dennis. «Chaotic dynamics in an insect population», *Science* 275 (1997) 389-391.

⁵² J. Huisman y F.J. Weissing. «Biodiversity of plankton by species oscillations and chaos», *Nature* 402 (1999) 407-410.

modelos usuales falla e, intuitivamente, especies diferentes pueden ocupar el mismo nicho haciendo turnos, no por cooperación consciente, sino por una especie tomando el mando temporalmente sobre la otra y experimentando un *boom* en la población, mientras la especie desplazada se reduce a una población pequeña (figura 59).

En 2008, el equipo de Huisman publicó los resultados de un experimento de laboratorio con una ecología de miniatura basada en una encontrada en el mar Báltico, que involucraba bacterias y varios tipos de plancton. Un estudio de seis años reveló dinámicas caóticas en las cuales las poblaciones fluctuaban frenéticamente, con frecuencia haciéndose 100 veces tan grandes durante un tiempo y luego colapsando. Los métodos habituales para detectar caos confirmaron su presencia. Había incluso un efecto mariposa: el horizonte de predicciones del sistema era de unas pocas semanas.⁵³

Hay aplicaciones del caos que afectan a la vida diaria, pero la mayoría ocurren en procesos de fabricación o servicios públicos, más que estar incorporadas en aparatos. El descubrimiento del efecto mariposa ha cambiado el modo en que las predicciones del tiempo se llevan a cabo. En lugar de poner todo el esfuerzo informático en perfeccionar una única predicción, los meteorólogos ahora realizan muchos partes meteorológicos, haciendo diferentes cambios aleatorios minúsculos en las observaciones proporcionadas por los globos meteorológicos y los satélites antes de empezar a realizarlos. Si todos estos partes están de acuerdo, entonces la predicción es probable que sea precisa, si difieren de manera significativa, el tiempo está en un estado menos predecible. Los propios partes meteorológicos se han mejorado con otros avances, en particular calculando la influencia de los océanos en el estado de la atmósfera; pero el papel principal del caos ha sido advertir a los meteorólogos que no esperen demasiado y cuantificar cómo de probable es que un parte meteorológico esté en lo correcto.

Las aplicaciones industriales incluyen una comprensión mejor de los procesos de mezclado, que son ampliamente usados para hacer medicamentos en pastillas o mezclar los ingredientes de una comida. La medicina activa en una pastilla

⁵³ E. Benincà, J. Huisman, R. Heerkloss, K.D. Jöhnk, P. Branco, E.H. Van Nes, M. Scheffer y S.P. Ellner. «Chaos in a long-term experiment with a plankton community», *Nature* 451 (2008) 822-825.

normalmente se da en cantidades muy pequeñas y tiene que mezclarse con alguna sustancia inerte. Es importante tener suficiente del ingrediente activo en cada píldora, pero no demasiado. Una máquina de mezclado es como un robot de cocina gigante y, como el robot de cocina, su dinámica es determinista pero caótica. Las matemáticas del caos han proporcionado una comprensión nueva de los procesos de mezclado y llevan a algunos diseños mejorados. Los métodos usados para detectar el caos en datos han inspirado nuevos equipos de pruebas para el metal usado para hacer muelles, mejorando la eficiencia en la fabricación de muelles y alambre. El humilde muelle tiene muchos usos vitales: puede encontrarse en colchones, coches, reproductores de DVD, incluso bolígrafos. El control del caos, una técnica que usa el efecto mariposa para mantener el comportamiento dinámico estable, está resultando prometedor en el diseño de marcapasos más eficientes y menos intrusivos.

Aunque, sobre todo, el principal impacto del caos ha sido en el pensamiento científico. En los más o menos cuarenta años desde que su existencia empezase a ser ampliamente apreciada, el caos ha cambiado de ser una curiosidad matemática menor a una característica básica de la ciencia. Ahora podemos estudiar muchas de las irregularidades de la naturaleza sin reorganizar las estadísticas, sacando con cuidado los patrones escondidos que caracterizan el caos determinista. Esto es solo uno de los modos en los que la teoría de sistemas dinámicos moderna, con su énfasis en el comportamiento no lineal, está provocando una revolución silenciosa en el modo en el que los científicos piensan en el mundo.

Capítulo 17

La fórmula de Midas

Ecuación de Black-Scholes

$$\frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} - rV = 0$$

¿Qué dice?

Describe cómo el precio de un derivado financiero cambia en el tiempo, basándose en el principio de que cuando el precio es correcto, el derivado no conlleva riesgo y nadie puede sacar beneficio vendiéndolo a un precio diferente.

¿Por qué es importante?

Hace posible comerciar un derivado antes de que venza asignándole un valor «racional» acordado, de modo que puede convertirse en una mercancía virtual por derecho propio.

¿Qué provocó?

Crecimiento masivo del sector financiero, instrumentos financieros cada vez más complejos, aumento repentino, salpicado con quiebras, en la prosperidad económica, los turbulentos mercados de valores de los noventa del siglo pasado, la crisis financiera del 2008-2009, y la depresión económica actual.

Desde el cambio de siglo, la mayor fuente de crecimiento en el sector financiero ha estado en los instrumentos financieros conocidos como derivados. Los derivados no son dinero, ni son inversiones en acciones. Son inversiones en inversiones,

promesas sobre promesas. Los operadores de derivados usan dinero virtual, números en un ordenador. Lo toman prestado de inversores que probablemente lo han tomado prestado en alguna otra parte. Con frecuencia no lo han tomado prestado del todo, ni siquiera virtualmente: han presionado el botón del ratón para estar de acuerdo con que tomarán prestado el dinero si alguna vez es necesario. Pero no tienen intención de permitir que sea necesario, venderán el derivado antes de que suceda. El prestamista, el hipotético prestamista, como el préstamo nunca se llevará a cabo, por la misma razón, probablemente tampoco tiene realmente el dinero. Esto son las finanzas en un reino de fantasía, aunque se ha convertido en una práctica estándar del sistema bancario del mundo.

Desafortunadamente, las consecuencias de las operaciones con derivados, al final, se convierten en dinero real y gente real sufriendo. La trampa funciona la mayoría del tiempo, porque la desconexión con la realidad no tiene un efecto notable, que no sea otro que hacer a unos pocos banqueros y operadores extremadamente ricos a medida que desvían fondos de dinero real desde la reserva virtual. Hasta que las cosas van mal. Entonces llegan las consecuencias, deudas virtuales que tienen que pagarse con dinero real. Por todos los demás, naturalmente.

Esto es lo que desencadenó la crisis bancaria del 2008-2009, por la cual la economía mundial está todavía tambaleándose. Tipos de interés bajos y enormes primas personales animaron a los banqueros y sus bancos a apostar sumas de dinero virtual cada vez mayores en derivados cada vez más complejos, seguros a la larga, o así lo creían, en el mercado inmobiliario, casas y negocios. Como la oferta de propiedad y gente adecuada para comprar empezó a agotarse, los líderes del mundo financiero necesitaban encontrar nuevos modos de convencer a los accionistas de que estaban creando beneficios, para justificar y financiar sus primas. De modo que empezaron a comerciar paquetes de deuda, también supuestamente seguros, en algún momento futuro, sobre propiedad real. Mantener ese esquema demandó la compra continua de bienes, para aumentar la reserva de garantías. De modo que los bancos empezaron a vender hipotecas a gente cuya habilidad para pagarlas era cada vez más dudosa. Esto era el mercado de hipotecas subprime, donde «subprime» es un eufemismo para «probabilidad de impago». Lo que pronto se convirtió en «certeza de impago».

Los bancos se comportaron como uno de esos personajes de dibujos que se aleja del borde del precipicio, se sostiene en el aire hasta que mira abajo, y solo entonces cae en picado a la tierra. Todo parecía marchar bien hasta que los banqueros se preguntaron a sí mismos si múltiples cálculos con dinero inexistente y recursos sobrevalorados eran sostenibles, se preguntaron cuál era el valor real de sus propiedades en derivados, y se dieron cuenta de que no tenían ni idea. Excepto que definitivamente era mucho menos de lo que habían dicho a los accionistas y los organismos reguladores del gobierno.

A medida que la espantosa verdad iba saliendo, la confianza caía en picado. Esto abatió el mercado inmobiliario, de modo que los recursos contra los cuales las deudas estaban aseguradas empezaron a perder su valor. En este punto, todo el sistema se vio atrapado en un ciclo de retroalimentación positivo, en el cual cada revisión a la baja del valor provocaba que fuese revisada todavía más a la baja. El resultado final fue la pérdida de alrededor de 17 billones de dólares. Enfrentados a la posibilidad del colapso total del sistema financiero mundial, destrozando los ahorros de los inversionistas y haciendo que la Gran Depresión de 1929 pareciese un patio de recreo, los gobiernos se vieron forzados a financiar a los bancos, los cuales estaban al borde de la bancarrota. Se dejó que uno, Lehman Brothers, se hundiese, pero la pérdida de confianza fue tan grande que parecía poco inteligente repetir la lección. De modo que los contribuyentes apoquinaron el dinero, y mucho de él era dinero real. Los bancos cogieron el dinero con ambas manos, y luego intentaron fingir que la catástrofe no había sido por su culpa. Culparon a los organismos reguladores de los gobiernos, a pesar de haber hecho campaña contra las regulaciones: un caso interesante de «es culpa vuestra, vosotros nos dejasteis hacerlo».

¿Cómo sucedió el mayor desastre financiero de la historia de la humanidad?

Podría decirse que uno de los colaboradores fue una ecuación matemática.

Los derivados más sencillos existen desde hace mucho tiempo. Son conocidos como futuros y opciones, y se remontan al siglo XVIII en el mercado de arroz de Dojima en Osaka, Japón. El mercado fue fundado en 1697, una época de gran prosperidad económica en Japón, cuando se pagaba a las clases altas, los samuráis, con arroz, no con dinero. Naturalmente apareció una clase de agentes de arroz que

comerciaban con arroz como si fuese dinero. A medida que los comerciantes de Osaka aumentaban su control sobre el arroz, el alimento básico del país, sus actividades tenían un efecto dominó en el precio del artículo. Al mismo tiempo, el sistema financiero estaba empezando a cambiar a dinero en metálico, y la combinación resultaba mortal. En 1730, el precio del arroz estaba por los suelos. Irónicamente, el desencadenante fueron las cosechas pobres. Los samuráis, todavía empeñados en el pago en arroz, pero atentos al crecimiento del dinero, empezaron a tener pánico. Su «moneda» favorita estaba perdiendo su valor rápidamente. Los comerciantes agravaron el problema manteniendo artificialmente el arroz fuera del mercado, acumulando grandes cantidades en almacenes. Aunque puede parecer que esto incrementaría el valor monetario del arroz, tuvo el efecto opuesto, porque los samuráis estaban tratando el arroz como una moneda. No podrían comer nada remotamente aproximado a la cantidad de arroz que poseían. De modo que mientras la gente ordinaria pasaba hambre, los comerciantes almacenaban arroz. El arroz se hizo tan escaso que el dinero en papel lo sustituyó, y rápidamente se hizo más deseable que el arroz porque era posible realmente ponerle la mano encima. Pronto los comerciantes de Dojima estaban dirigiendo lo que equivalía a un sistema bancario gigantesco, teniendo cuentas con la gente adinerada y determinando el valor de intercambio entre arroz y dinero en papel.

Finalmente el gobierno se dio cuenta de que este arreglo daba demasiado poder a los comerciantes de arroz y reorganizó el mercado de arroz junto con la mayoría de las otras partes de la economía del país. En 1939 el Mercado de arroz fue remplazado por la Agencia gubernamental de arroz. Pero mientras existió el Mercado de arroz, los comerciantes inventaron un nuevo tipo de contrato para nivelar las grandes oscilaciones en el precio del arroz. Los firmantes garantizaban comprar (o vender) una cantidad específica de arroz en una fecha futura concreta por un precio concreto. Hoy esos instrumentos se conocen como futuros y opciones. Supón que un comerciante está de acuerdo en comprar arroz dentro de seis meses a un precio acordado. Si el precio de mercado ha subido sobre el acordado en el momento en que la opción vence, consigue el arroz barato e inmediatamente lo vende con ganancias. Por otro lado, si el precio es más bajo, se ha comprometido a comprar arroz a un precio mayor de su valor en el mercado y pierde dinero.

Los granjeros encuentran dichos instrumentos útiles porque realmente quieren vender una mercancía real: arroz. La gente que usa el arroz para comer, o fabricar comestibles que lo usan, quiere comprar la mercancía. En este tipo de transacción, el contrato reduce el riesgo para ambas partes, aunque a un precio. Equivale a una forma de seguro, un mercado garantizado a un precio garantizado, independientemente de los cambios en los valores del mercado. Merece la pena pagar un pequeño recargo para evitar la incertidumbre. Pero la mayoría de los inversores adquieren los contratos en futuros de arroz con el único propósito de hacer dinero, y la última cosa que el inversor quiere es toneladas y toneladas de arroz. Siempre lo vendían antes de que tuviesen que recibirla. De modo que el principal papel de los futuros era alimentar la especulación financiera, y esto se hacía peor por el uso del arroz como moneda. Al igual que el patrón actual del oro crea artificialmente precios altos para una sustancia (el oro) que tiene poco valor intrínseco, y de ese modo alimenta la demanda por ello, así el precio del arroz empezó a estar gobernado por el comercio de futuros más que por el comercio del propio arroz. Los contratos eran una forma de juego de apuestas; pronto los propios contratos adquirieron un valor y podían comerciarse como si fuesen mercancías reales. Además, aunque la cantidad de arroz estaba limitada por lo que los granjeros podían plantar, no había límite para el número de contratos de arroz que podían expedirse.

El mayor mercado de valores del mundo fue rápido encontrando una oportunidad de vender humo y ganar dinero en metálico, y se han comerciado futuros desde entonces. Al principio, esta práctica no causaba por sí misma problemas económicos enormes, aunque a veces llevó a la inestabilidad más que a la estabilidad que con frecuencia es reivindicada para justificar el sistema. Pero alrededor del año 2000, el sector financiero mundial empezó a inventar variantes cada vez más elaboradas sobre el tema de los futuros, «derivados» complejos cuyo valor estaba basado en hipotéticos movimientos futuros de algún recurso. A diferencia de los futuros, para los cuales al menos el recurso era real, los derivados podían estar basados en un recurso que fuese en sí mismo un derivado. Los bancos ya no estaban comprando y vendiendo apuestas sobre el precio futuro de una mercancía como el arroz, estaban comprando y vendiendo apuestas sobre el precio futuro de una apuesta.

Rápidamente se convirtió en un gran negocio. En 1998, el sistema financiero internacional comerció aproximadamente 100 billones de dólares en derivados. En 2007 esto había crecido a miles de billones de dólares. Billones, miles de billones... sabemos que estos son números grandes, pero ¿cómo de grandes? Para poner esta cifra en contexto, el valor total de todos los productos hechos por las industrias manufactureras durante los últimos mil años es alrededor de 100 billones de dólares americanos, ajustado por la inflación. Esto es una décima parte de los derivados comercializados en un año. Ciento es que el grueso de la producción industrial se ha dado en los pasados cincuenta años, pero incluso así, esto es una cantidad asombrosa. Lo que quiere decir, en concreto, que las ventas de derivados consisten casi totalmente en dinero que realmente no existe, dinero virtual, números en un ordenador, sin vínculo con nada en el mundo real. De hecho, estas ventas tienen que ser virtuales: la cantidad total de dinero en circulación, en el mundo, es totalmente insuficiente para pagar las cantidades que están siendo intercambiadas con el clic de un ratón. Por gente que no tiene interés en la mercancía que les ocupa y no sabrían qué hacer con ella si la recibiesen, usando dinero que realmente no poseen.

No necesitas ser ingeniero aeroespacial para sospechar que esta es una receta para el desastre. Aunque durante una década la economía mundial creció sin cesar a lomos del comercio de derivados. No solo podías obtener una hipoteca para comprar una casa; podías obtener más de lo que la casa valía. El banco ni se molestaba en comprobar cuáles eran tus verdaderos ingresos, o qué otras deudas tenías. Podías conseguir una hipoteca autocertificada —que quiere decir que le decías al banco que podías permitírtelo y que no hiciese preguntas raras—, del 125 %, y gastar el dinero extra en unas vacaciones, un coche, una operación de cirugía o cajas de cerveza. Los bancos hicieron un esfuerzo especial para persuadir a los clientes para que contratasen préstamos, aunque no los necesitasen.

Lo que creían que los salvaría si uno de los solicitantes del préstamo no pagaba sus cuotas era sencillo. Esos préstamos estaban asegurados sobre tu casa. Los precios de las casas fueron aumentando, de modo que ese 25 % perdido de patrimonio pronto se haría real; si no pagabas, el banco podía embargar tu casa, venderla y obtener su préstamo de nuevo. Parecía infalible. Por supuesto, no lo era. Los

banqueros no se preguntaron qué sucedería con el precio de los bienes inmuebles si cientos de bancos estaban todos intentando vender millones de casas a la vez. Ni siquiera se preguntaron si los precios podrían continuar subiendo significativamente más rápido que la inflación. Realmente parecían pensar que los precios de las casas podían subir un 10-15 % en términos reales cada año indefinidamente. Todavía estaban instando a los organismos reguladores a relajar las reglas y permitirles prestar incluso más dinero cuando el mercado inmobiliario ya había tocado fondo. Muchos de los modelos matemáticos actuales más sofisticados de los sistemas financieros pueden tener origen en el movimiento browniano, mencionado en el capítulo 12. Cuando vieron a través de un microscopio pequeñas partículas suspendidas en un fluido que se movían de un lado a otro de modo irregular, Einstein y Smoluchowski desarrollaron modelos matemáticos de este proceso y los usaron para establecer la existencia de átomos. El modelo habitual asume que las partículas reciben patadas aleatorias a través de distancias cuya distribución de probabilidad es normal, una campana de Gauss. La dirección de cada patada está distribuida uniformemente, cualquier dirección tiene la misma probabilidad de suceder. Este proceso se llama un camino aleatorio. El modelo del movimiento browniano es una versión continua de dichos caminos aleatorios, en los que los tamaños de las patadas y el tiempo entre patadas sucesivas se hace arbitrariamente pequeño. Intuitivamente, consideramos infinitas patadas infinitesimales.

Las propiedades estadísticas del movimiento browniano, a lo largo de un gran número de pruebas, están determinadas por una distribución de probabilidad, que da las posibilidades de que las partículas acaben en una localización concreta después de un tiempo dado. La distribución tiene simetría radial; la probabilidad depende solo de lo lejos que esté el punto del origen. Inicialmente la partícula es muy posible que esté cerca del origen, pero a medida que el tiempo pasa, el rango de posiciones posibles se amplía, ya que la partícula tiene más opciones de explorar regiones distantes en el espacio. Sorprendentemente, la evolución del tiempo de esta distribución de probabilidad obedece a la ecuación del calor, que en este contexto con frecuencia se llama ecuación de difusión. De modo que la probabilidad se extiende como el calor.

Después de que Einstein y Smoluchowski publicasen su trabajo, resultó que mucho del contenido matemático había sido obtenido con anterioridad, en 1900, por el matemático francés Louis Bachelier en su tesis de doctorado. Pero Bachelier tenía una aplicación diferente en mente, los mercados de valores y opciones. El título de su tesis era *Théorie de la speculation* (Teoría de la especulación). El trabajo no fue recibido con gran entusiasmo, probablemente porque el tema estaba fuera del rango normal de las matemáticas en esa época. El director de tesis de Bachelier era el célebre y formidable matemático Henri Poincaré, quien declaró que el trabajo era «muy original». Él también descubrió el pastel de algún modo, añadiendo, con referencia a la parte de la tesis que obtenía la distribución normal para errores: «Es lamentable que el señor Bachelier no desarrollase más esta parte de su tesis». Lo que cualquier matemático interpretaría como «este era el punto donde las matemáticas empezaban a ponerse realmente interesantes y si tan solo hubiese trabajado más eso, en lugar de las enmarañadas ideas sobre el mercado de valores, habría sido fácil darle una nota mucho mejor». La tesis recibió una calificación de «honorable», un aprobado, y fue incluso publicada. Pero no obtuvo la nota máxima de «très honorable».

A todos los efectos, Bachelier determinaba el principio de que las fluctuaciones del mercado de valores siguen un camino aleatorio. Los tamaños de fluctuaciones sucesivas se ajustan a una campana de Gauss, y la media y la desviación estándar pueden estimarse a partir de los datos de mercado. Una implicación es que fluctuaciones grandes son muy improbables. La razón es que las colas de la distribución normal disminuyen muy rápido, más rápido que exponencialmente. La campana de Gauss decrece hacia cero a una velocidad que es exponencial en el cuadrado de x . Los estadísticos (y físicos y analistas de mercados) hablan de fluctuaciones dos sigma, tres sigma, etcétera. Aquí sigma (σ) es la desviación estándar, una medida de cómo de ancha es la campana de Gauss. Una fluctuación 3 sigma, por ejemplo, es una que se desvía de la media al menos tres veces la desviación estándar. Las matemáticas de la campana de Gauss asignan probabilidades a estos «sucesos extremos» (véase la tabla 3).

TABLA 3. Probabilidades de los sucesos *n-sigma*

Tamaño mínimo de la fluctuación	Probabilidad
σ	0.3174
2σ	0.0456
3σ	0.0027
4σ	0.000063
5σ	0.0000006

El resultado del modelo de movimiento browniano de Bachelier es que fluctuaciones grandes del mercado de valores son tan raras que en la práctica nunca deberían darse. La tabla 3 muestra que un suceso 5-sigma, por ejemplo, se espera que ocurra 6 veces en 10 millones de intentos. Sin embargo, los datos del mercado de valores muestran que son mucho más comunes que eso. Las acciones en Cisco Systems, un líder mundial en comunicaciones, han experimentado diez sucesos 5-sigma en los últimos veinte años, mientras que el movimiento browniano predice 0,003 de ellos. Seleccioné esta compañía aleatoriamente y no es de ningún modo inusual. En el lunes negro (19 de octubre de 1987) el mercado de valores mundial perdió más de un 20 % de su valor en unas pocas horas, un suceso tan extremo debería haber sido prácticamente imposible.

Los datos sugieren de manera inequívoca que los sucesos extremos no son de ninguna manera próximos a ser tan raros como el movimiento browniano predice. La distribución de probabilidad no se desvanece exponencialmente (o más rápido), se desvanece como una curva del modo x^{-a} para alguna constante positiva a . En la jerga financiera, dicha distribución se dice que tiene una cola pesada. Las colas pesadas indican niveles de riesgo mayor. Si tu inversión tiene una rentabilidad esperada de 5-sigma, entonces asumiendo el movimiento browniano, las posibilidades de que fracase son menos de una entre un millón. Pero si las colas son pesadas, podría ser mucho mayor, quizás una entre cien. Eso la hace una apuesta mucho más pobre.

Un término relacionado, hecho popular por Nassim Nicholas Taleb, un experto en matemáticas financieras, es «eventos del cisne negro». Su libro de 2007 *El cisne*

negro se convirtió en un gran *best seller*. En la Antigüedad, todos los cisnes conocidos eran blancos. El poeta Décimo Junio Juvenal se refiere a algo como «un raro pájaro en las tierras, y muy parecido a un cisne negro», para indicar que era imposible. La frase se usaba mucho en el siglo XVI, tanto como nosotros podríamos referirnos a un cerdo volando. Pero en 1697, cuando el explorador holandés Willem de Vlamingh fue al bien llamado río Swan en la Australia occidental, encontró montones de cisnes negros. La frase cambió su significado, y ahora se refiere a una suposición que parece estar basada en hechos, pero podría en cualquier momento resultar ser totalmente errónea.

Estos análisis iniciales de los mercados en términos matemáticos fomentaron la seductora idea de que se podía hacer un modelo matemático del mercado, creando un modo racional y seguro de hacer cantidades de dinero ilimitadas. En 1973, parecía que el sueño podría hacerse real, cuando Fischer Black y Myron Scholes introdujeron un método para poner precio a las opciones: la ecuación de Black-Scholes. Robert Merton proporcionó un análisis matemático de su modelo el mismo año y lo amplió. La ecuación es:

$$\frac{1}{2}(\sigma S)^2 \frac{\partial^2 V}{\partial S^2} + rS \frac{\partial V}{\partial S} + \frac{\partial V}{\partial t} - rV = 0$$

Hay cinco cantidades distintas involucradas: el tiempo t , el precio S de la mercancía, el precio V del derivado, que depende de S y de t , la tasa de interés libre de riesgo r (el interés teórico que puede ganarse con una inversión con riesgo cero, como los bonos del Estado) y la volatilidad σ^2 de las acciones. Es también matemáticamente sofisticada, es una ecuación en derivadas parciales de segundo orden como las ecuaciones de onda y del calor. Expresa la tasa de variación del precio del derivado con respecto al tiempo como una combinación lineal de tres términos: el precio del propio derivado, lo rápido que cambia en relación al precio de la acción y cómo ese cambio acelera. Las otras variables aparecen en los coeficientes de estos términos. Si los términos que representan el precio de los derivados y su tasa de variación se omitiesen, la ecuación sería exactamente la ecuación del calor, describiendo cómo el precio de la opción se extiende por el

espacio de precios de las acciones. Esto determina el origen de la suposición del movimiento browniano de Bachelier. Los otros términos tienen en cuenta factores adicionales.

La ecuación de Black-Scholes se obtuvo como una consecuencia de un número de suposiciones financieras simplificadas, por ejemplo, que no hay coste de transacción y no hay límites en las ventas al descubierto, y que es posible prestar y tomar prestado dinero a una tasa de interés conocida, fija y libre de riesgo. La aproximación se llama teoría del arbitraje, y su núcleo matemático vuelve a Bachelier. Asume que los precios de mercado se comportan estadísticamente como el movimiento browniano, en el cual tanto el ritmo de la deriva como la volatilidad del mercado son constantes. La deriva es el movimiento de la media y la volatilidad es jerga financiera para desviación estándar, una medida de la divergencia media a partir de la media. Esta suposición es tan común en la literatura financiera que se ha convertido en un estándar del sector.

Hay dos tipos principales de opciones. En una opción de venta, el comprador de la opción compra lo razonable para vender una mercancía o instrumento financiero en un momento específico por un precio acordado, si así lo desea. Una opción de compra es similar, pero otorga el derecho de comprar en vez de vender. La ecuación de Black-Scholes tiene soluciones explícitas: una fórmula para las opciones de venta, otra para las opciones de compra.⁵⁴ Si dichas fórmulas no hubiesen existido, la ecuación podría todavía haberse resuelto numéricamente e implementarse como *software*. Sin embargo, las fórmulas hacen sencillo calcular el precio recomendado, además de proporcionar una comprensión teórica importante.

La ecuación de Black-Scholes fue concebida para llevar a los mercados futuros

⁵⁴ El valor de una opción de compra es:

$$C(s, t) = N(d_1)S - N(d_2)Ke^{-r(T-t)}$$

Donde

$$d_1 = \frac{\log(S/K) + (r + \sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}$$

$$d_2 = d_1 - \sigma\sqrt{T-t}$$

El precio de la opción de venta correspondiente es:

$$P(s, t) = [N(d_1) - 1]S + [1 - N(d_2)]Ke^{-r(T-t)}$$

Donde $N(dj)$ es la función de distribución acumulada de la distribución normal estándar para $j = 1, 2$, y $T - t$ es el tiempo para el vencimiento.

cierta racionalidad, lo cual hace de modo efectivo bajo condiciones de mercado normales. Proporciona un modo sistemático de calcular el valor de una opción antes de que venza. Entonces se puede vender. Supón, por ejemplo, que un comerciante contrata una compra de 1.000 toneladas de arroz dentro de 12 meses por un precio de 500 por tonelada, una opción de compra. Después de cinco meses, decide vender la opción a alguien que esté dispuesto a comprarla. Todo el mundo sabe cómo el precio del mercado para el arroz ha estado cambiando, de modo que ¿cuánto cuesta ese contrato justo ahora? Si empiezas a comerciar dichas opciones sin saber la respuesta, vas a tener problemas. Si la operación pierde dinero, estás abierto a la acusación de que obtuviste un precio equivocado y tu trabajo podría estar en riesgo. De modo que, ¿cuál debería ser el precio? Comerciar improvisando sobre la marcha deja de ser una opción cuando las cantidades involucradas son de billones. Tiene que haber un modo acordado de poner precio a una opción en un momento determinado antes de que venza. La ecuación hace justo eso. Proporciona una fórmula que cualquiera puede usar, y si tu jefe usa la misma fórmula, obtendrá el mismo resultado que tú, siempre que no cometas errores de cálculo. En la práctica, los dos usaríais un programa informático estándar.

La ecuación era tan efectiva que hizo que Merton y Scholes ganasen el Nobel en Economía en 1997.⁵⁵ Black había muerto para entonces y las reglas del premio prohíben premios póstumos, pero su contribución fue explícitamente citada por la Academia Sueca. La efectividad de la ecuación depende del propio comportamiento del mercado. Si las suposiciones tras el modelo dejan de cumplirse, no era inteligente seguir usándola. Pero a medida que el tiempo pasaba y la confianza crecía, muchos banqueros y comerciantes se olvidaron de eso, usaron la ecuación como un tipo de talismán, un poco de magia matemática que los protegía contra las críticas. Black-Scholes no solo te proporcionaba un precio que es razonable bajo condiciones normales, también te cubría las espaldas si la operación era un desastre. No me culpe, jefe, usé la fórmula estándar del sector.

El sector de las finanzas fue rápido en ver las ventajas de la ecuación de Black-Scholes y sus soluciones, e igualmente rápido en desarrollar una gran cantidad de ecuaciones relacionadas con diferentes suposiciones dirigidas a diferentes

⁵⁵ En sentido estricto, el premio Sveriges Riksbank en Ciencias Económicas en memoria de Alfred Nobel.

instrumentos financieros. El entonces sedado mundo de la banca convencional podía usar las ecuaciones para justificar préstamos y transacciones, siempre manteniendo los ojos abiertos ante problemas potenciales. Pero los negocios menos convencionales los seguirían pronto, y estos tenían la fe de un verdadero converso. Para ellos, la posibilidad de que el modelo fuese mal era inconcebible. Pasó a conocerse como la fórmula de Midas, una receta para convertir todo en oro. Pero el sector financiero se olvidó de cómo acaba la historia del rey Midas.

La niña mimada de las finanzas, durante varios años, era una compañía llamada Long Term Capital Management (LTCM). Era un fondo de inversión, un fondo privado que extendió sus inversiones de un modo que estaba pensado para proteger a los inversores cuando el mercado cayese y obtener grandes beneficios cuando subiese. Se especializó en estrategias comerciales basadas en modelos matemáticos, incluyendo la ecuación de Black-Scholes y sus extensiones, junto con técnicas como el arbitraje, que explota las discrepancias entre los precios de bonos y el valor al que pueden realmente llevarse a cabo. Inicialmente LTCM tuvo un éxito espectacular, generando ganancias en la región del 40 % por año hasta 1998. En ese momento perdió 4.600 millones de dólares en menos de cuatro meses, y el Banco de la Reserva Federal convocó a sus mayores acreedores para financiarlo con alrededor de 3.600 millones de dólares. Finalmente los bancos involucrados recuperaron su dinero, pero LTCM fue cerrada en el 2000.

¿Qué fue mal? Hay tantas teorías como comentaristas económicos, pero el consenso es que la causa aproximada del fracaso de LTCM fue la crisis económica rusa de 1998. Los mercados occidentales habían invertido mucho en Rusia, cuya economía era tremadamente dependiente de las exportaciones de petróleo. La crisis económica asiática de 1997 provocó que el precio del petróleo cayese repentinamente, y víctima principal fue la economía rusa. El Banco Mundial dio un préstamo de 22.600 millones de dólares para apoyar a Rusia.

La causa final de la desaparición de LTCM ya era visible desde el día en que empezó a comerciar. Tan pronto como la realidad dejó de obedecer las suposiciones del modelo, LTCM estaba en un serio apuro. La crisis económica rusa fastidió todo y derribó casi todas las suposiciones. Algunos factores tuvieron un efecto más grande que otros. La volatilidad en aumento fue uno de ellos. Otro fue la suposición de que

las fluctuaciones extremas ocurren difícilmente: ninguna cola pesada. Pero la crisis sembró los mercados de confusión, y, en el pánico, los precios cayeron enormemente, muchos sigmas, en segundos. Debido a que todos los factores afectados estaban interrelacionados, estos sucesos desencadenaron otros cambios rápidos, tan rápidos que era imposible que los inversores pudiesen saber el estado del mercado en un instante. Incluso aunque quisieran comportarse razonalmente, lo cual la gente no hace en un estado de pánico general, no tenían bases sobre las que hacerlo.

Si el modelo browniano es correcto, los sucesos tan extremos como la crisis financiera rusa no deberían suceder con más frecuencia que una vez en un siglo. Puedo recordar siete a partir de mi experiencia personal en los últimos 40 años: exceso de inversión en el mercado inmobiliario, la ex Unión Soviética, Brasil, mercado inmobiliario (de nuevo), mercado inmobiliario (de nuevo otra vez), empresas punto com y... ah, sí, mercado inmobiliario.

En retrospectiva, el colapso de LTCM fue una advertencia. Se tomó debida nota de los peligros de comerciar usando una fórmula en un mundo que no obedecía las suposiciones convenientes tras la fórmula y rápidamente se ignoraron. La retrospección está muy bien, pero cualquiera puede ver el peligro después de que haya ocurrido una crisis. ¿Qué pasa con la previsión? La afirmación ortodoxa sobre la reciente crisis financiera global es que, como el primer cisne con plumas negras, nadie la vio venir.

Esos no es totalmente cierto.

El Congreso Internacional de Matemáticos es la mayor convención matemática mundial, y tiene lugar cada cuatro años. En agosto de 2002, tuvo lugar en Beijing, y Mary Poovey, catedrática de humanidades y directora del Institute for the Production of Knowledge en la Universidad de Nueva York, dio una conferencia titulada «Can number ensure honesty?»⁵⁶ (¿Pueden los números garantizar la honestidad?). El subtítulo era «Unrealistic expectations and the US accounting scandal» (Expectativas no realistas y el escándalo en las cuentas de EE.UU.), y describía la reciente aparición de un «nuevo eje de poder» en las relaciones

⁵⁶ M. Poovey. «Can numbers ensure honesty? Unrealistic expectations and the U.S. accounting scandal», *Notices of the American Mathematical Society* 50 (2003) 27-35.

mundiales.

Este eje pasa por las grandes corporaciones multinacionales, muchas de las cuales evitan los impuestos nacionales constituyéndose en paraísos fiscales como Hong Kong. Pasa por bancos de inversión, a través de organizaciones no gubernamentales como el Fondo Monetario Internacional, a través de fondos del Estado y fondos de pensiones corporativos, y a través del bolsillo de los inversores ordinarios. Este eje de poder financiero contribuye a catástrofes económicas como la debacle de 1998 en Japón y los impagos en Argentina en 2001, y deja su rastro en las rotaciones de los índices bursátiles como el Dow Jones y el FTSE 100 de la Bolsa de Londres.

Continuó diciendo que este nuevo eje de poderes no es intrínsecamente ni bueno ni malo; lo que importa es cómo ejerce su poder. Ayudó al crecimiento del nivel de vida en China, que muchos de nosotros consideraríamos como beneficioso. También animó al abandono mundial de las sociedades de bienestar, remplazándolas por una cultura de accionistas, que muchos de nosotros consideraríamos como dañino. Un ejemplo menos polémico de un mal resultado es el escándalo de Enron, que quebró en 2001. Enron era una compañía energética con sede en Texas, y su colapso llevó a lo que entonces fue la mayor bancarrota en la historia de los Estados Unidos, y una pérdida para los accionistas de 11.000 millones de dólares. Enron fue otra advertencia, esta vez sobre las leyes liberalizadoras del mercado. De nuevo, pocos hicieron caso de la advertencia.

Poovey lo hizo. Señaló el contraste entre el sistema financiero tradicional, basado en la producción de bienes reales, y el emergente, basado en la inversión, el mercado de divisas y «apuestas complejas sobre si los precios futuros subirán o bajarán». En 1995 esta economía del dinero virtual había superado la economía real de la fabricación. El nuevo eje de poder estaba deliberadamente confundiendo dinero real y virtual; cifras arbitrarias en las cuentas de compañías y dinero y artículos reales. Esta tendencia, defendía, estaba llevando a una cultura en la que los valores tanto de los bienes como de los instrumentos financieros se estaban haciendo terriblemente inestables, propensos a explotar o colapsar con el clic de un ratón.

El artículo ilustraba estos puntos usando cinco técnicas e instrumentos financieros

comunes, como «marcar el mercado», en la que una compañía establece una sociedad con un subsidiario. El subsidiario compra una participación en los beneficios futuros de la empresa matriz, el dinero involucrado se registra entonces como ingresos inmediatos por la empresa matriz mientras que el riesgo se relega al balance general del subsidiario. Enron usó esta técnica cuando cambió su estrategia de marketing de vender energía a vender futuros de energía. El gran problema con adelantar potenciales beneficios futuros de esta manera es que no pueden ponerse como beneficios el año siguiente. La respuesta es repetir la maniobra. Es como tratar de conducir un coche sin frenos presionando cada vez más el acelerador. El resultado inevitable es chocar.

El quinto ejemplo de Poovey fueron los derivados, y este fue el más importante de todos porque las sumas de dinero involucradas eran gigantescas. Su análisis refuerza en buena parte lo que ya he dicho. Su conclusión principal fue: «Las operaciones con futuros y derivados están supeditadas a la creencia de que el mercado de valores se comporta de una manera estadísticamente predecible, en otras palabras, que las ecuaciones matemáticas describen exactamente el mercado». Pero indicó que las pruebas señalan en una dirección totalmente diferente: entre un 75 % y un 90 % de todos los inversores de futuros pierden dinero en un año cualquiera.

En concreto, dos tipos de derivados estuvieron implicados en la creación de los mercados financieros tóxicos de comienzos del siglo XXI: las permutas de incumplimiento crediticio y las obligaciones de deuda colateralizadas. Una permuta de incumplimiento crediticio es una forma de seguro: pagas tu prima y cobras de una compañía aseguradora si alguien no paga una deuda. Pero cualquiera podía hacer dicho seguro sobre cualquier cosa. No tenía que ser la compañía que lo debiese, o a la que se le debiese la deuda. De modo que un fondo de inversión podía, a todos los efectos, apostar que los clientes de un banco no iban a pagar sus hipotecas, y si sucedía, el fondo de inversión ganaría un dineral incluso aunque no fuese una parte en el contrato de la hipoteca. Esto proporcionaba un incentivo para que los especuladores influyesen en las condiciones de mercado e hiciesen los impagos más probables. Una obligación de deuda colateralizada está basada en una colección (cartera) de activos. Estos pueden ser tangibles, como hipotecas

protegidas contra la propiedad real, y pueden ser derivados, o pueden ser una mezcla de ambos. El propietario de los activos vende a los inversores el derecho a una participación de los beneficios de esos activos. El inversor puede jugar sobre seguro y llevarse la primera llamada de los beneficios, pero esto les cuesta más. O pueden asumir el riesgo, pagar menos, y estar más abajo en el orden de picotear para un pago.

Los bancos, los fondos de inversión y otros especuladores comerciaban con ambos tipos de derivados. Se les puso precio usando descendientes de la ecuación de Black-Scholes, de modo que estaban considerados como activos por derecho propio. Los bancos tomaron prestado dinero de otros bancos, de modo que podían prestárselo a la gente que quería hipotecas, aseguraron estos préstamos con propiedades reales y derivados elaborados. Pronto todo el mundo estaba prestando enormes sumas de dinero a todo el mundo, mucho de ello asegurado sobre derivados financieros. Los fondos de inversión y otros especuladores estaban intentando hacer dinero encontrando desastres potenciales y apostando qué les sucedería. El valor de los derivados afectados, y de sus activos reales como la propiedad, se calculaba con frecuencia sobre las bases de marcar el mercado, que está abierto a abusos porque usa procedimientos de contabilidad artificiales y compañías subsidiarias arriesgadas para presentar beneficios futuros estimados como beneficios actuales reales. Prácticamente todo el mundo en el negocio evaluó cómo de arriesgados eran los derivados usando el mismo método, conocido como «valor en riesgo». Este calcula la probabilidad de que la inversión pueda suponer una pérdida que exceda un umbral específico. Por ejemplo, los inversores podrían estar dispuestos a aceptar una pérdida de un millón de dólares si su probabilidad fuese menos de un 5 %, pero no si fuese más probable. Como Black-Scholes, el valor en riesgo asume que no hay colas pesadas. Quizá la peor característica era que todo el sector financiero estaba estimando sus riesgos usando exactamente el mismo método. Si el método era el culpable, esto crearía una desilusión compartida de que el riesgo era bajo cuando en realidad era mucho más alto.

Era un choque de trenes esperando a suceder, un dibujo animado que había caminado un kilómetro más allá del límite del precipicio y permanecía suspendido en medio del aire solo porque se negaba rotundamente a mirar qué había bajo sus

pies. Como Poovey, y otros como ella, había advertido repetidamente, los modelos usados para valorar los productos financieros y estimar sus riesgos incorporaban suposiciones simplificadas que no representaban exactamente los mercados reales y los peligros inherentes a ellos. Los jugadores del mercado financiero ignoraron estos avisos. Seis años más tarde, todos averiguamos por qué esto era un error.

Quizá haya un camino mejor.

La ecuación de Black-Scholes cambió el mundo creando una industria en auge de miles de billones de dólares; su generalización, usada de modo poco inteligente por un pequeño círculo de banqueros, cambió el mundo de nuevo contribuyendo a una quiebra financiera de miles de billones de dólares cuyos efectos cada vez más malignos, que ahora se extienden a economías nacionales enteras, están todavía sintiéndose por todo el mundo. La ecuación pertenece al reino de las matemáticas continuas clásicas, que tiene sus raíces en las ecuaciones en derivadas parciales de la física matemática. Este es un reino en el que las cantidades son infinitamente divisibles, el tiempo fluye de modo continuo y las variables cambian suavemente. La técnica funciona para la física matemática, pero parece menos apropiada para el mundo de las finanzas, donde el dinero viene en paquetes discretos, las operaciones se dan de una en una (aunque muy rápido), y muchas variables pueden cambiar erráticamente.

La ecuación de Black-Scholes está también basada en las suposiciones tradicionales de la economía matemática clásica: información perfecta, racionalidad perfecta, equilibrio de mercado, la ley de la oferta y la demanda. La asignatura se ha enseñado durante décadas como si estas cosas fuesen axiomáticas y muchos economistas cualificados nunca las han cuestionado. Aunque carecen de apoyo empírico convincente. En las pocas ocasiones en que alguien hace experimentos para observar cómo la gente toma sus decisiones económicas, los escenarios clásicos normalmente fallan. Es como si los astrónomos hubiesen pasado los últimos cien años calculando cómo los planetas se mueven basados en lo que creían que era razonable, sin realmente molestarte en comprobar si realmente lo era.

No es que la economía clásica esté completamente equivocada. Pero está equivocada con más frecuencia de lo que sus defensores afirman, y cuando se equivoca, se equivoca muchísimo. De modo que los físicos, matemáticos y

economistas están buscando modelos mejores. Al frente de estos esfuerzos están los modelos basados en la ciencia de la complejidad, una nueva rama de las matemáticas que remplaza el pensamiento continuo clásico por una colección explícita de agentes individuales interactuando según unas reglas específicas.

Un modelo clásico de movimiento de precios de algunas mercancías, por ejemplo, asume que en cualquier instante hay un único precio «justo», en principio conocido por todo el mundo, y que los posibles compradores comparan este precio con una función de utilidad (cómo de útil les es la mercancía); sólo compran la mercancía si su utilidad es mayor que su coste. Un modelo de sistemas complejos es muy diferente. Podría implicar, por ejemplo, diez mil agentes, cada uno con su propia visión de lo que vale la mercancía y cómo de deseable es. Algunos agentes sabrían más que otros, algunos tendrían información más precisa que otros, muchos pertenecerían a pequeñas redes que comercian con información (precisa o no) lo mismo que dinero o bienes.

Ha surgido una variedad de características interesante a partir de estos modelos. Uno es el papel del instinto gregario. Los agentes de bolsa tienden a copiar a otros agentes de bolsa. Si no lo hacen, y resulta que los otros estaban tras algo bueno, sus jefes no estarán contentos. Por otro lado, si siguen a la manada y todo el mundo se equivoca, tienen una buena excusa: es lo que todo el mundo estaba haciendo. Black-Scholes era perfecta para el instinto gregario. De hecho, prácticamente toda crisis económica en el último siglo ha sido llevada al extremo por el instinto gregario. En lugar de que algunos bancos invirtieran en el mercado inmobiliario y otros en la industria, por ejemplo, todos se precipitaron sobre el mercado inmobiliario. Esto sobrecarga el mercado, con demasiado dinero buscando demasiadas pocas propiedades, y todo se hace trizas. De modo que ahora todos se precipitan en préstamos a Brasil o a Rusia, o vuelven a un nuevamente revivido mercado inmobiliario, o pierden la cabeza con compañías punto com, tres niños en una habitación con un ordenador y un módem se valoran diez veces más que un gran fabricante con un producto real, clientes reales y fábricas y oficinas reales. Cuando todo queda patas arriba, todos se precipitan en el mercado de las hipotecas subprime...

Esto no es hipotético. Incluso cuando las repercusiones de la crisis económica global

hacen retumbar las vidas de la gente normal, y las economías nacionales luchan por mantenerse a flote, hay señales de que no se ha aprendido ninguna lección. Una repetición de la moda pasajera de las compañías de Internet está en progreso, ahora encabezada por las webs de redes sociales: Facebook ha sido valorado en 100.000 millones de dólares, y Twitter (la web donde los famosos envían tuits de 140 caracteres a sus devotos seguidores) ha sido valorada en 8.000 millones de dólares a pesar de no haber obtenido beneficios nunca. El Fondo Monetario Internacional ha advertido fuertemente sobre los fondos negociables de mercado (ETFs por sus siglas en inglés, *Exchange Traded Funds*), un modo muy exitoso de invertir en artículos como petróleo, oro o trigo sin realmente comprar nada. Todos ellos han elevado sus precios de modo muy rápido, proporcionando grandes beneficios para fondos de pensión y otros grandes inversores, pero el FMI ha advertido que estos vehículos de inversión tienen «todos los sellos de una burbuja esperando a reventar ... reminiscencias de lo que ha ocurrido en el mercado de las titularizaciones antes de la crisis». Los ETFs son muy parecidos a los derivados que desencadenaron la crisis de crédito, pero asegurados en mercancías en lugar de propiedades. La estampida hacia los ETFs ha llevado a los precios de las mercancías hasta el tope, inflándolos más allá de toda proporción con la demanda real. Ahora mucha gente en el tercer mundo no es capaz de ahorrar para alimentación básica porque los especuladores de los países desarrollados están haciendo grandes apuestas sobre el trigo. La destitución de Hosni Mubarak en Egipto fue, hasta cierto punto, impulsada por los enormes incrementos en el precio del pan.

El principal peligro es que los ETFs están empezando a reinventarse como más derivados, como las obligaciones de deuda colateralizadas y las permutas de incumplimiento crediticio que reventaron la burbuja de las hipotecas subprime. Si la burbuja de las mercancías revienta, podríamos ver una repetición del colapso, basta cambiar la palabra «propiedad» por «mercancías». Los precios de las mercancías son muy volátiles, de modo que los ETFs son inversiones de alto riesgo, no una gran elección para un fondo de pensiones. Así que una vez más, de nuevo, se está animando a los inversores a hacer apuestas cada vez más complejas y cada vez más arriesgadas, a usar dinero que no tienen para comprar participaciones en cosas que no quieren y no pueden usar, con el fin de beneficios especulativos, mientras

que la gente que quiere cosas ya no puede permitírselas.

¿Te acuerdas del mercado de arroz de Dojima?

La economía no es la única área que descubre que sus preciadas teorías tradicionales ya no funcionan en un mundo cada vez más complejo, donde las viejas reglas ya no se aplican. Otra es la ecología, el estudio de los sistemas naturales como los bosques o los arrecifes de coral. De hecho, la economía y la ecología son increíblemente parecidas en muchos aspectos. Algunos de sus parecidos son ilusorios; históricamente con frecuencia cada una ha usado a la otra para justificar sus modelos, en lugar de comparar los modelos con el mundo real. Pero otros son reales; las interacciones entre grandes cantidades de organismos son muy parecidas a las que hay entre grandes cantidades de agentes de bolsa.

Este parecido puede usarse como una analogía, en cuyo caso es peligroso porque las analogías con frecuencia fracasan. O puede usarse como una fuente de inspiración, tomando prestado técnicas para hacer modelos de la ecología y aplicarlas de forma adecuadamente modificada en la economía. En enero de 2011, en la revista *Nature*, Andrew Haldane y Robert May señalaron algunas posibilidades.⁵⁷ Sus argumentos refuerzan varios de los mensajes que han aparecido en este capítulo, y sugieren modos de mejorar la estabilidad de los sistemas económicos.

Haldane y May observaron un aspecto de la crisis económica que yo no he mencionado todavía: cómo los derivados afectan a la estabilidad de los sistemas financieros. Comparan la visión preponderante de los economistas ortodoxos, que mantienen que el mercado automáticamente busca un equilibrio estable, con una visión similar en la ecología de la década de los sesenta del siglo XX, la de que el «equilibrio de la naturaleza» tiende a mantener los ecosistemas estables. De hecho, en esa época muchos ecologistas pensaban que cualquier ecosistema suficientemente complejo sería estable de este modo, y que el comportamiento inestable, como las oscilaciones continuas, implicaba que el sistema no era lo suficiente complejo. Vimos en el capítulo 16 que esto era un error. De hecho, la comprensión actual indica exactamente lo contrario. Supón que un gran número de especies interactúa en un ecosistema. Como la red de interacciones ecológicas se

⁵⁷ A.G. Haldane y R.M. May. «Systemic risk in banking ecosystems», *Nature* 469 (2011) 351-355.

hace más compleja a través de la adición de nuevos vínculos entre especies, o las interacciones se hacen más fuertes, hay un umbral muy marcado más allá del cual el ecosistema deja de ser estable. (Aquí el caos cuenta como estabilidad, las fluctuaciones pueden suceder siempre que permanezcan dentro de unos límites específicos.) Este descubrimiento llevó a los ecologistas a buscar tipos especiales de redes de interacción, inusualmente propicias para la estabilidad.

¿Sería posible transferir estos descubrimientos ecológicos a la economía global? Hay analogías cercanas, con comida o energía en ecología que se corresponden con el dinero en un sistema financiero. Haldane y May eran conscientes de que esta analogía no debería usarse directamente; comentaron: «en los ecosistemas financieros, las fuerzas evolutivas con frecuencia han sido supervivientes de lo más rápido más que de lo más fuerte». Decidieron construir modelos financieros no imitando modelos ecológicos, sino explotando los principios de modelado generales que habían llevado a una mejor comprensión de los ecosistemas.

Desarrollaron varios modelos económicos, mostrando en cada caso que, bajo las circunstancias adecuadas, los sistemas económicos se harían inestables. Los ecologistas lidian con un ecosistema inestable manejándolo de un modo que crea estabilidad. Los epidemiólogos hacen lo mismo con la epidemia de una enfermedad, esto es, por qué, por ejemplo, el gobierno británico desarrolló una política para controlar la epidemia de la fiebre aftosa en 2001 matando rápidamente reses en las granjas cercanas a cualquiera que hubiese resultado positiva para la enfermedad, y deteniendo todo movimiento de reses en el país. De ese modo, la respuesta de los reguladores del gobierno a un sistema financiero inestable debería ser tomar acciones para estabilizarlo. Hasta cierto punto, ahora están haciendo esto, después del pánico inicial tras el cual dieron enormes cantidades de dinero de los contribuyentes a los bancos, pero omitieron imponer condiciones más allá de promesas vagas que no se han mantenido.

Sin embargo, las nuevas regulaciones en buena parte fracasan en dirigirse al problema real, que es el diseño pobre del propio sistema financiero. La facilidad para transferir billones con el clic de un ratón quizás permita beneficios cada vez más rápidos, pero también permite que los impactos se propaguen más rápido, y anima a una complejidad cada vez mayor. Ambas cosas son desestabilizadoras. El no

aplicar impuestos sobre las transacciones financieras permite a los agentes explotar esta velocidad, haciendo apuestas mayores en el mercado, a una velocidad mayor. Esto también tiende a crear inestabilidad. Los ingenieros saben que el modo de obtener una respuesta rápida es usar un sistema inestable; la estabilidad por definición indica una resistencia innata al cambio, mientras que una respuesta rápida requiere lo opuesto. De modo que la búsqueda para beneficios cada vez mayores ha provocado que se desarrolle un sistema financiero cada vez más inestable.

Aunque construyendo de nuevo sobre analogías con ecosistemas, Haldane y May ofrecen algunos ejemplos de cómo la estabilidad podría aumentar. Algunas corresponden a los instintos propios de los reguladores, como requerir a los bancos tener más capital, que los amortigüe contra los impactos. Otros no; un ejemplo es la sugerencia de que los reguladores deberían centrarse no en los riesgos asociados con los bancos individuales, sino en los asociados con el sistema financiero completo. La complejidad del mercado de derivados podría reducirse requiriendo que todas las transacciones pasen a través de una cámara de compensación central. Esto tendría que ser extremadamente robusto, apoyado por todas las naciones importantes, pero si existiese, entonces la propagación de impactos se apaciguaría ya que tiene que pasar a través de ella.

Otra sugerencia es una diversidad cada vez mayor en los métodos de comerciar y en la valoración de riesgos. Una monocultura ecológica es inestable porque cualquier impacto que ocurre es probable que afecte a todo a la vez del mismo modo. Cuando todos los bancos están usando los mismos métodos para evaluar los riesgos, surge el mismo problema: cuando se equivocan, todos se equivocan a la vez. La crisis económica surgió en parte porque todos los bancos principales estaban financiando sus lastres potenciales del mismo modo, evaluando el valor de sus recursos del mismo modo, y evaluando sus riesgos probables del mismo modo. La sugerencia final es la modularidad. Se cree que los ecosistemas se estabilizan a sí mismos organizándose (a través de la evolución) en módulos más o menos autónomos, conectados unos a otros de una manera bastante simple. La modularidad ayuda a prevenir la propagación de impactos. Por eso los reguladores de todo el mundo están pensando seriamente romper los grandes bancos y

reemplazarlos por una cantidad de bancos más pequeños. Como Alan Greenspan, un distinguido economista norteamericano y expresidente de la Reserva Federal de EE.UU., dijo de los bancos: «Si son demasiado grandes para fracasar, son demasiado grandes».

Entonces, ¿fue una ecuación la culpable de la crisis financiera?

Una ecuación es una herramienta, y, como toda herramienta, tiene que ser empuñada por alguien que sabe cómo usarla, y con los fines correctos. La ecuación de Black-Scholes quizá haya contribuido a la crisis, pero solo porque se abusó de ella. No es más responsable del desastre de lo que habría sido el ordenador de un agente de bolsa si su uso llevase a una pérdida catastrófica. La culpa del fracaso de las herramientas debería recaer en aquellos que son responsables de su uso. Existe el peligro de que el sector financiero pueda dar la espalda al análisis matemático, cuando lo que realmente necesita es un rango mejor de modelos y, significativamente, una comprensión sólida de sus limitaciones. El sistema financiero es demasiado complejo para que sea dirigido por coronadas humanas y razonamientos vagos. Necesita desesperadamente más matemáticas, no menos. Pero también necesita aprender cómo usar las matemáticas de manera inteligente, más que algún tipo de talismán mágico.

¿Qué es lo próximo?

Cuando alguien pone por escrito una ecuación, no hay un repentino trueno tras el cual todo es diferente. La mayoría de las ecuaciones tiene poco o ningún efecto (yo las pongo por escrito todo el rato, y créeme, lo sé). Pero incluso las mejores y más influyentes ecuaciones necesitan ayuda para cambiar el mundo: modos eficientes de resolverlas, gente con la imaginación y el instinto para explotar lo que nos quieren decir, mecanismos, recursos, materiales, dinero. Teniendo esto en mente, las ecuaciones han establecido repetidamente nuevas direcciones para la humanidad, y actuado como nuestras guías a medida que las exploramos.

Se necesitaron más que diecisiete ecuaciones para llevarnos a donde estamos en la actualidad. Mi lista es una selección de algunas de las más influyentes, y cada una de ellas requiere de un montón de otras antes de pasar a ser útil en serio. Pero cada una de las diecisiete enteramente merece la inclusión, porque desempeñó un papel fundamental en la historia. Pitágoras llevó a métodos prácticos para medir nuestras tierras y hacernos camino hacia otras nuevas. Newton nos dijo cómo se mueven los planetas y cómo enviar sondas espaciales para explorarlos. Maxwell proporcionó una pista vital que llevó a la radio, televisión y las comunicaciones modernas. Shannon obtuvo límites inevitables de cómo de eficientes pueden ser esas comunicaciones.

Con frecuencia, a lo que nos condujo una ecuación ha sido bastante diferente con respecto a lo que interesaba a su inventor/descubridor. ¿Quién habría predicho en el siglo XV que un número desconcertante, y aparentemente imposible, con el que nos tropezamos mientras resolvíamos problemas de álgebra estaría indeleblemente vinculado al mundo, todavía más desconcertante y aparentemente imposible, de la física cuántica —dejando de lado que esto pavimentaría el camino a instrumentos milagrosos que pueden resolver un millón de problemas de álgebra cada segundo, y permitiéndonos instantáneamente ser vistos y oídos por amigos al otro lado del planeta—? ¿Cómo habría reaccionado Fourier si se le hubiese dicho que su nuevo método para estudiar el flujo del calor se convertiría en máquina del tamaño de una baraja, capaz de pintar imágenes de un modo extraordinariamente preciso y detallado de cualquier cosa a la que esté apuntando, en color, incluso en

movimiento, con miles de ellas contenidas en algo del tamaño de una moneda? Las ecuaciones desencadenaron sucesos, y los sucesos, parafraseando al ex primer ministro británico Harold Macmillan, son los que nos dejan dormir por la noche. Cuando se libera una ecuación revolucionaria, desarrolla una vida propia. Las consecuencias pueden ser buenas o malas, incluso cuando la intención original era benevolente, como lo era para todo el mundo de mis diecisiete ecuaciones. La nueva física de Einstein nos dio un nuevo entendimiento del mundo, pero una de las cosas para las que la usamos fue para las armas nucleares. No tan directamente como reclama el mito popular, pero, no obstante, jugó su parte. La ecuación de Black-Scholes creó un sector financiero vibrante y luego amenazó con destruirlo. Las ecuaciones son lo que hacemos de ellas, y el mundo puede cambiarse para lo peor, del mismo modo que para lo mejor.

Hay muchos tipos de ecuaciones. Algunas son verdades matemáticas, tautologías; piensa en los logaritmos neperianos. Pero las tautologías todavía pueden ser ayudas potentes para el pensamiento y acción humana. Algunas son afirmaciones sobre el mundo físico, que según lo que sabemos podrían haber sido diferentes. Las ecuaciones de este tipo nos hablan de las leyes de la naturaleza, y al resolverlas nos dicen las consecuencias de dichas leyes. Algunas tienen ambos elementos: la ecuación del teorema de Pitágoras es un teorema en la geometría euclíadiana, pero también gobierna las mediciones hechas por los topógrafos y los navegantes. Algunas son poco más que las definiciones, pero i y la información nos dicen mucho una vez que las hemos definido.

Algunas ecuaciones son universalmente válidas. Algunas describen el mundo muy exactamente, pero no perfectamente. Algunas son menos precisas, confinadas a reinos más limitados, aunque ofrecen un entendimiento vital. Algunas son básicamente erróneas sin más, aunque pueden actuar como peldaños hacia algo mejor. Todavía podrían tener un efecto enorme.

Algunas incluso desvelan cuestiones difíciles, de naturaleza filosófica, sobre el mundo en que vivimos y nuestro lugar en él. El problema de las mediciones cuánticas, escenificadas por el desafortunado gato de Schrödinger, es una de ellas. La segunda ley de la termodinámica presenta temas profundos sobre el desorden y la flecha del tiempo. En ambos casos, algunas de las paradojas aparentes pueden

ser resueltas, en parte, pensando menos en el contenido de la ecuación y más en el contexto en el que se aplica. No en los símbolos, sino en las condiciones de contorno. La flecha del tiempo no es un problema sobre la entropía; es un problema sobre el contexto en el cual pensamos en la entropía.

Las ecuaciones existentes pueden adquirir una nueva importancia. La búsqueda de la energía de fusión, como una alternativa limpia a la energía nuclear y los combustibles fósiles, requiere una comprensión de cómo un gas extremadamente caliente, formando un plasma, se mueve en un campo magnético. Los átomos del gas pierden electrones y pasan a estar eléctricamente cargados. De modo que es un problema en la magnetohidrodinámica, y se necesita una combinación de las ecuaciones existentes para fluidos y para electromagnetismo. La combinación llega a un nuevo fenómeno, sugiriendo cómo mantener el plasma estable a la temperatura necesaria para que se produzca la fusión. Las ecuaciones son viejas favoritas.

Hay (o podría haber) una ecuación, sobre todas, por la que los físicos y los cosmólogos darían su brazo derecho por ponerle las manos encima: una Teoría del Todo, la cual en la época de Einstein era llamada una teoría del campo unificado. Esta es la ecuación largamente buscada que unifica la mecánica cuántica y la relatividad, y Einstein pasó sus últimos años en una búsqueda sin frutos para encontrarla. Estas dos teorías son exitosas, pero sus éxitos suceden en dominios diferentes: el muy pequeño y el muy grande. Cuando se solapan, son incompatibles. Por ejemplo, la mecánica cuántica es lineal, la relatividad no lo es. Se busca una ecuación que explique por qué ambas son tan exitosas, pero que haga el trabajo de ambas sin inconsistencias lógicas. Hay muchas candidatas a la teoría del todo, la más conocida es la teoría de supercuerdas. Esta, entre otras cosas, introduce dimensiones extra del espacio; seis (siete en algunas versiones). Las supercuerdas son matemáticamente elegantes, pero no hay pruebas convincentes para ellas como una descripción de la naturaleza. En cualquier caso, es desesperadamente difícil llevar a cabo los cálculos necesarios para extraer predicciones cuantitativas a partir de la teoría de supercuerdas.

Hasta donde sabemos, podría no haber una teoría del todo. Todas nuestras ecuaciones para el mundo físico podrían ser solo modelos demasiado simplificados,

que describen reinos de la naturaleza limitados en un modo que podemos comprender, pero no capturar la estructura profunda de la realidad. Incluso si la naturaleza realmente obedece leyes rígidas, podrían no ser ecuaciones expresables. Incluso si las ecuaciones son relevantes, no necesariamente son simples. Podrían ser tan complicadas que no podamos ni siquiera escribirlas. Los 3.000 millones de bases del ADN del genoma humano son, en cierto sentido, parte de la ecuación para el ser humano. Son parámetros que podrían insertarse en una ecuación más general para el desarrollo biológico. Es (apenas) posible imprimir el genoma en papel, necesitaríamos alrededor de dos mil libros del tamaño de este. Pero cabe en la memoria de un ordenador bastante fácilmente. Sin embargo, es solo una parte diminuta de una hipotética ecuación humana.

Cuando las ecuaciones se hacen complejas, necesitamos ayuda. Los ordenadores ya están extrayendo ecuaciones a partir de grandes conjuntos de datos, en circunstancias donde los métodos humanos habituales fracasan o son demasiado opacos para ser útiles. Una nueva aproximación llamada computación evolutiva extrae patrones significativos: específicamente fórmulas para cantidades que se conservan, cosas que no cambian. Uno de dichos sistemas llamado Eureqa, formulado por Michael Schmidt y Hod Lipson, ha alcanzado cierto éxito. Software como este podría ayudar. O podría no llevar a ningún sitio que realmente importe. Algunos científicos, especialmente aquellos con formación en computación, creen que es el momento de que abandonemos las ecuaciones tradicionales, especialmente las continuas como las ecuaciones ordinarias o en derivadas parciales. El futuro es discreto, se presenta en números enteros, y las ecuaciones deberían dar paso a los algoritmos, recetas para calcular cosas. En lugar de resolver ecuaciones, deberíamos simular el mundo digitalmente ejecutando los algoritmos. De hecho, el propio mundo podría ser digital. Stephen Wolfram defendió esta visión en su polémico libro *A New Kind of Science* (Un nuevo tipo de ciencia), que aboga por un tipo de sistema complejo llamado un autómata celular. Esto es una matriz de células, habitualmente pequeños cuadrados, cada uno existiendo en una variedad de estados distintos. Las células interactúan con sus células vecinas según unas reglas fijadas. Se parece un poco a un juego de ordenador de los ochenta con bloques de colores persiguiéndose unos a otros por la pantalla.

Wolfram expone varias razones por las que los autómatas celulares deberían ser superiores a las ecuaciones matemáticas tradicionales. En concreto, algunos de ellos pueden llevar a cabo cualquier cálculo que pueda realizarse con un ordenador, siendo el más simple el famoso autómata Regla 110. Este puede encontrar dígitos sucesivos de π , resolver ecuaciones del problema de tres cuerpos numéricamente, implementar la fórmula de Black-Scholes para una opción de compra, lo que sea. Los métodos tradicionales para resolver ecuaciones son más limitados. Yo no encuentro este argumento demasiado convincente, porque es también cierto que cualquier autómata celular puede simularse por un sistema dinámico tradicional. Lo que cuenta no es si un sistema matemático puede simular a otro, sino cuál es más efectivo para resolver problemas o proporcionar algún entendimiento. Es más rápido sumar una serie tradicional para π a mano de lo que es calcular el mismo número de dígitos usando el autómata Regla 110.

Sin embargo, es todavía totalmente creíble que podríamos pronto encontrar nuevas leyes de la naturaleza basadas en estructuras y sistemas digitales discretos. El futuro quizás consista en algoritmos, no ecuaciones. Pero hasta que esa época comience, si lo hace, nuestro mayor entendimiento de las leyes de la naturaleza tiene la forma de ecuaciones, y deberíamos aprender a comprenderlas y apreciarlas. Las ecuaciones tienen una trayectoria. Realmente han cambiado el mundo, y lo cambiarán de nuevo.

F I N