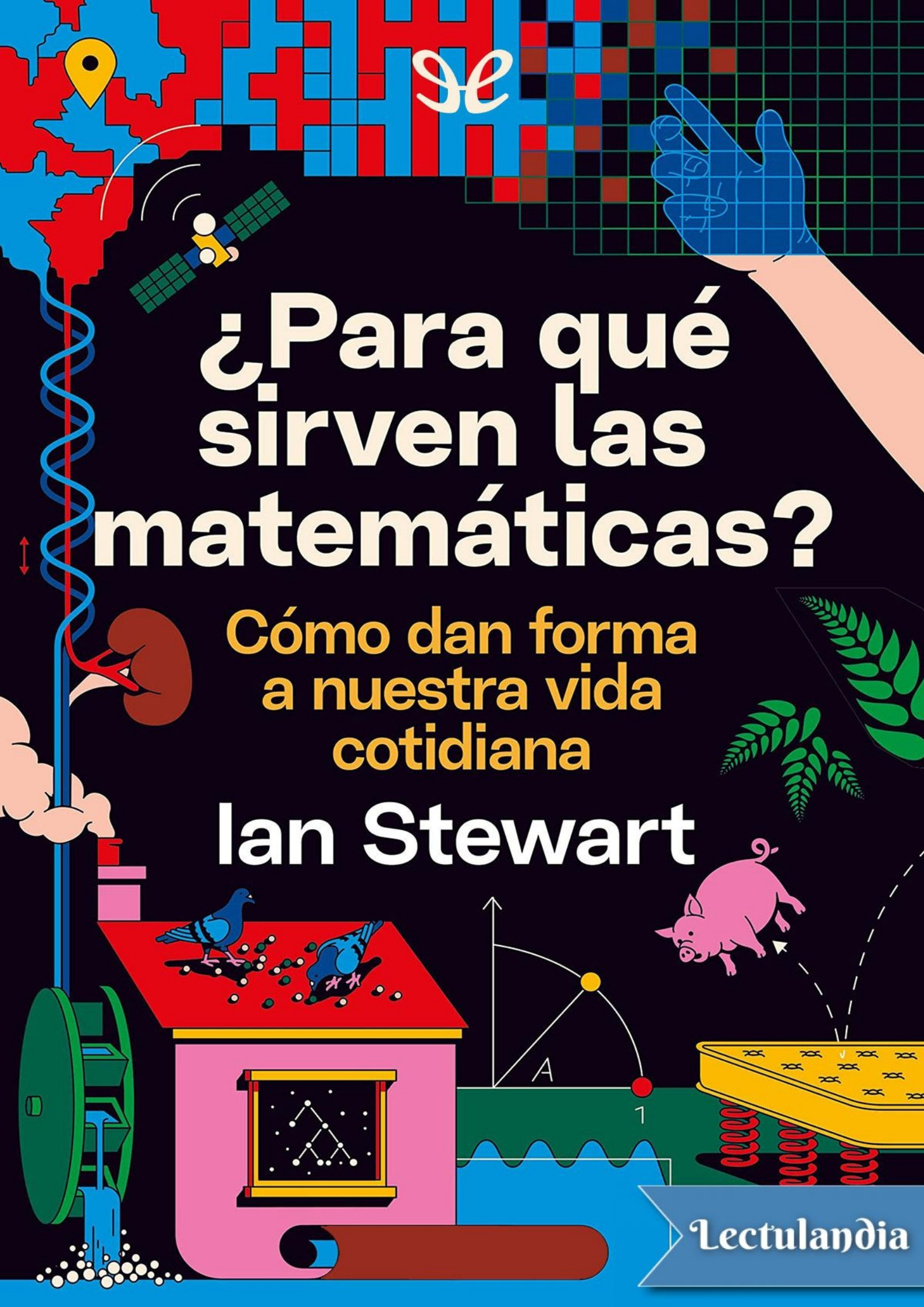


¿Para qué sirven las matemáticas?

Cómo dan forma a nuestra vida cotidiana

Ian Stewart

Lectulandia

Ante la percepción popular de que las matemáticas son inútiles para nuestro día a día, el profesor Ian Stewart nos demuestra que esta disciplina va mucho más allá de los cálculos aburridos que todos recordamos de la escuela y nos propone un curioso recorrido por los usos de las matemáticas que a menudo permanecen ocultos a simple vista, pero contribuyen a nuestras vidas. Desde la trigonometría que mantiene un satélite en órbita hasta los números primos utilizados por los sistemas de seguridad más avanzados del mundo, pasando por los números imaginarios que permiten la realidad aumentada, los métodos más eficientes para los trasplantes de riñón, las aplicaciones en política, la predicción del cambio climático o, incluso, cómo una curva extraña e infinitamente ondulada optimiza las entregas a domicilio; las matemáticas no solo son relevantes para nuestras vidas, sino que sin ellas el mundo moderno como lo conocemos se desmoronaría.

Ian Stewart

¿Para qué sirven las matemáticas?: cómo dan forma a nuestra vida cotidiana

ePub r1.0
Un_Tal_Lucas 04-12-2022

Título original: *What's the Use? The Unreasonable Effectiveness of Mathematics*
Ian Stewart, 2021
Traducción: Miguel A. Pérez

Editor digital: Un_Tal_Lucas
ePub base r2.1


1
Irrazonable eficacia

> El milagro de la idoneidad del lenguaje de las matemáticas para la formulación de las leyes de la física es un regalo maravilloso que ni comprendemos ni merecemos. Deberíamos estar agradecidos por ello y esperar que siga siendo válido en la investigación futura y que se extienda, para bien o para mal, para nuestro placer o incluso para nuestra confusión, a ramas más amplias del saber.
>
> EUGENE WIGNER,
> «La irrazonable eficacia de la matemática en las ciencias naturales»

¿Para qué sirven las matemáticas?
¿Qué hacen por *nosotros*, en nuestra vida cotidiana?

Poco tiempo ha, estas preguntas tenían respuestas sencillas. El ciudadano de a pie empleaba la aritmética básica en todo momento, aunque solo fuese para comprobar el recibo de la compra. Los carpinteros tenían que saber geometría elemental. Agrimensores y navegantes necesitaban también la trigonometría. La ingeniería exigía dominar el cálculo.

Hoy en día, las cosas han cambiado. La caja del supermercado calcula el importe total del recibo, resta las ofertas del día y añade el IVA. Oímos los pitidos cuando el láser escanea los códigos de barras y, mientras el sonido coincida con el paso de los productos, asumimos que los chismes electrónicos saben lo que hacen. Muchas profesiones dependen todavía de un conocimiento matemático amplio, pero incluso en estos casos, la mayor parte de los cálculos se confían a aparatos electrónicos con algoritmos incorporados.

Mi disciplina brilla por su ausencia. Ni siquiera hay un toro al que agarrar por los cuernos.

Sería fácil llegar a la conclusión de que las matemáticas se han quedado anticuadas y obsoletas, pero sería una idea equivocada. Sin ellas, el mundo actual se vendría abajo. Para demostrarlo, voy a traer a colación sus aplicaciones en política, en derecho, en los trasplantes de riñón, en las rutas de reparto de los supermercados, en la seguridad en internet, en los efectos

especiales de las películas y en la fabricación de muelles. Veremos el papel esencial que desempeñan en los escáneres médicos, en la fotografía digital, en la banda ancha por fibra óptica y en la navegación por satélite. También, cómo nos ayudan a predecir los efectos del cambio climático o cómo pueden protegernos frente a terroristas y *hackers* informáticos.

Cabe destacar que muchas de estas aplicaciones dependen de unas matemáticas que surgieron por motivos completamente ajenos a ellas, a menudo solo por el simple atractivo que tiene dejarse llevar por el propio instinto. Mientras me documentaba para este libro, me sorprendí, en repetidas ocasiones, al encontrarme con usos de mi disciplina cuya existencia ni siquiera había imaginado. A menudo, sacaban partido a aspectos que yo no esperaba que tuviesen aplicaciones prácticas, como las curvas que recubren un espacio, los cuaterniones y la topología.

Las matemáticas son un sistema ilimitado de ideas y métodos de una creatividad inmensa. Yacen inmediatamente bajo la superficie de las tecnologías revolucionarias que han hecho que el siglo XXI sea diferente por completo a cualquier época anterior: videojuegos, viajes aéreos internacionales, comunicaciones por satélite, ordenadores, internet o teléfonos móviles[1]. Si se rasca un IPhone, se verá el brillante reflejo de las matemáticas.

Por favor, que nadie lo tome al pie de la letra.

*

Hay una tendencia a asumir que los ordenadores, con sus habilidades casi milagrosas, han dejado obsoletos a los matemáticos, incluso a la disciplina en sí misma. Sin embargo, no pueden sustituir a la persona, del mismo modo que los microscopios no reemplazaron a los biólogos. Los ordenadores han cambiado la manera en que *hacemos* matemáticas, sobre todo al liberarnos de las partes aburridas. Nos permiten disponer de más tiempo para pensar, nos ayudan a encontrar patrones y nos aportan una herramienta potente y novedosa para hacer avanzar la disciplina de manera más rápida y eficaz.

De hecho, un motivo importante por el que las matemáticas se han hecho todavía más esenciales es la omnipresencia de ordenadores baratos y potentes. Su proliferación ha creado nuevas oportunidades para aplicar la disciplina a problemas del mundo real. Métodos que hasta la fecha eran impracticables, debido a los muchos cálculos que exigían, son ahora rutinarios. Los más grandes matemáticos de la época del papel y el lápiz se habrían llevado las

manos a la cabeza, desesperados ante cualquier método que necesitase mil millones de operaciones. Sin embargo, tales métodos se emplean hoy en día de manera rutinaria porque disponemos de una tecnología que puede echar las cuentas en una fracción de segundo.

Hace mucho que los matemáticos se encuentran a la vanguardia de la revolución informática. Junto con muchas otras profesiones, añado presuroso. Piénsese en George Boole, el pionero de la lógica simbólica que constituye la base de la arquitectura de ordenadores actual. Piénsese en Alan Turing y en su universal máquina de Turing, un sistema matemático que puede calcular todo lo calculable. Piénsese en Muhamad al Juarismi, cuyo texto de álgebra del año 820 d. C. enfatizaba la importancia de los procedimientos de cálculo sistemáticos que ahora llevan su nombre: *algoritmos*.

La mayor parte de los algoritmos que otorgan a los ordenadores sus impresionantes habilidades tienen una firme base matemática. Muchas de las técnicas involucradas se han tomado ya listas para su empleo del arsenal existente de ideas matemáticas, tales como PageRank, el algoritmo de Google que cuantifica la importancia de las páginas web y que ha dado lugar a una industria que genera miles de millones de dólares. Incluso los algoritmos de aprendizaje profundo más vistosos de la inteligencia artificial emplean conceptos matemáticos consagrados, tales como matrices y grafos ponderados. Una tarea tan prosaica como buscar una serie concreta de letras en un documento implica, por lo menos en un método muy común, un dispositivo matemático denominado autómata finito.

La participación de las matemáticas en estos fascinantes desarrollos tiende a pasarse por alto. Así que la próxima vez que los medios de comunicación saquen a escena alguna nueva y milagrosa capacidad de los ordenadores, habrá que tener en cuenta que esta lleva un montón de matemáticas escondidas bajo el brazo *y* también de ingeniería, física, química y psicología. Sin el apoyo de este elenco oculto de secundarios, la superestrella digital no sería capaz de lucir sus trucos en público.

*

Es fácil subestimar la importancia de las matemáticas en el mundo moderno porque casi toda su intervención se desarrolla entre bambalinas. Si alguien se pasea por cualquier calle de una ciudad, se verá abrumado por letreros que proclaman la relevancia cotidiana de bancos, fruterías, supermercados, tiendas de ropa, talleres de coches, despachos de abogados, restaurantes de comida

rápida, anticuarios, organizaciones benéficas y otro millar más de actividades y profesiones. Sin embargo, no verá ninguna placa en un portal que anuncie la presencia de un consultor matemático. Ni se venden latas de matemáticas en el supermercado.

No obstante, basta escarbar un poco bajo la superficie para que la relevancia de la disciplina salga enseguida a la luz. Las ecuaciones matemáticas de la aerodinámica son fundamentales para el diseño de aeronaves. La navegación depende de la trigonometría. Es cierto que la manera en que se emplea hoy en día es diferente a como lo hacía Cristóbal Colón porque cobra la forma de dispositivos electrónicos, en vez de papel, lápiz y tablas de navegación, pero los principios subyacentes son los mismos. El desarrollo de nuevos fármacos se basa en la estadística para garantizar que son seguros y eficaces. Las comunicaciones por satélite requieren una profunda comprensión de la dinámica orbital. El pronóstico del tiempo exige resolver las ecuaciones del movimiento de la atmósfera, de la humedad que contiene, de lo caliente o fría que está y de cómo interactúan todos estos factores. Hay miles de ejemplos más. La implicación de las matemáticas pasa desapercibida porque no es necesario conocerla para beneficiarse de sus resultados.

¿Qué es lo que hace que las matemáticas sean tan útiles en una variedad tan amplia de actividades humanas?

No es una pregunta novedosa. En 1959, el físico Eugene Wigner pronunció una célebre conferencia en la Universidad de Nueva York[2] bajo el título de «La irrazonable eficacia de las matemáticas en las ciencias naturales». Se centró en la ciencia, pero podría haber defendido el mismo argumento en la agricultura, la medicina, la política, el deporte… lo que sea. El propio Wigner albergaba la esperanza de que esta eficacia se hiciese extensible a «ramas más amplias del saber». Así ha sido, sin duda.

La palabra fundamental en su título destaca por lo sorprendente: *irrazonable*. La mayoría de los usos de las matemáticas son completamente razonables, una vez que se descubren los métodos implicados en la resolución de un problema importante o en la invención de un aparato útil. Por ejemplo, es del todo razonable que los ingenieros utilicen las ecuaciones de la aerodinámica para diseñar una aeronave. Al fin y al cabo, esta disciplina se creó para eso. Una buena parte de las matemáticas que se emplean para pronosticar el tiempo surgieron con ese propósito en mente. La estadística nació a partir del descubrimiento de patrones de gran escala en datos acerca del comportamiento humano. La cantidad de matemáticas que hacen falta

para diseñar gafas con lentes bifocales es enorme, si bien la mayoría se desarrollaron con la óptica como objetivo.

La capacidad de las matemáticas para resolver problemas importantes se vuelve irrazonable, en el sentido que le da Wigner a este término, cuando no existe una relación como esta entre los motivos originales para el desarrollo del área y su aplicación concreta. Wigner arrancaba su conferencia con una anécdota, que voy a parafrasear y a adornar un poco.

Dos antiguos compañeros de clase se encontraron un día. El primero, un estadístico que analizaba tendencias en la población, le mostró al otro uno de sus artículos científicos, que empezaba con una fórmula común en estadística: la distribución normal o «curva de campana»[3]. Le explicó varios símbolos (este se refiere al tamaño de la población, ese otro es un promedio de la muestra) y la forma en que se podía emplear la fórmula para deducir el tamaño de la población sin tener que contar a todo el mundo. Su compañero de clase pensó que bromeaba, pero no estaba seguro del todo, así que le preguntó por otros símbolos. Al final, llegó a uno que tenía este aspecto: π.

—¿Ese cuál es? Me recuerda a algo.

—Sí, ese es *pi*, la proporción que guarda la longitud de la circunferencia con su diámetro.

—Ya sabía yo que me estabas tomando el pelo —dijo el amigo—. ¿Qué demonios tendrá que ver un círculo con el tamaño de la población?

Lo primero que hay que decir acerca de esta anécdota es que el escepticismo del amigo es del todo razonable. El sentido común dice que dos conceptos tan dispares no pueden tener relación alguna. ¡Por el amor de Dios, uno está sacado de la geometría, el otro se refiere a seres humanos! Lo segundo es que, a pesar del sentido común, esa relación existe. La curva de campana viene dada por una fórmula que resulta que incluye el número π. Y no se trata solo de una aproximación práctica. El número exacto es de verdad el archiconocido π. No obstante, el motivo por el que aparece en el contexto de la curva de campana no tiene nada de intuitivo, ni siquiera para los matemáticos. Se necesita cálculo avanzado para saber de dónde sale, por no hablar del *por qué*.

Permítanme contar otra anécdota sobre π. Hace algunos años, reformamos el baño de la planta de abajo. Spencer, un albañil de versatilidad pasmosa que vino a poner los azulejos, se enteró de que escribo libros divulgativos sobre matemáticas.

—Pues tengo un problema de mates para usted —dijo—. Tengo que cubrir con baldosas un suelo circular y necesito conocer su área para saber

cuántas me van a hacer falta. Había una fórmula que nos enseñaron en el cole…

—π *r* al cuadrado —contesté.

—¡Esa! —Le acababa de recordar cómo usarla.

Se fue tan contento, con la respuesta a su problema de baldosas, una copia firmada de uno de mis libros y habiendo descubierto que las matemáticas que había aprendido en el colegio, al contrario de lo que había creído durante muchos años, resultaban útiles en su profesión actual.

La diferencia entre las dos anécdotas es evidente. En la segunda, π aparece porque se introdujo desde el primer momento para resolver exactamente ese tipo de problema. Es un ejemplo sencillo y directo de la eficacia de las matemáticas. En la primera anécdota también aparece π y resuelve el problema, pero su presencia es sorprendente. Es un ejemplo de eficacia *irrazonable*, de una aplicación de una noción matemática a un área ajena por completo a sus orígenes.

*

En *¿Para qué sirven las matemáticas?* no diré mucho más acerca de los usos razonables de mi disciplina. Son valiosos, interesantes y forman parte del panorama matemático tanto como cualquier otro aspecto. Son igual de importantes, aunque no nos hagan saltar de la silla ni proferir exclamaciones de asombro. También pueden llevar a los mandamases a pensar, de manera errónea, que el único modo de hacer que avance la disciplina es determinar los problemas y luego pedir a los matemáticos que ingenien formas de resolverlos. Desde luego, no hay nada malo en este tipo de investigación dirigida a conseguir logros concretos, pero es como pelear con una mano atada a la espalda. La historia ha demostrado en repetidas ocasiones el valor del segundo puño: el impresionante alcance de la imaginación humana. Lo que otorga a las matemáticas su potencia es la *combinación* de estos dos modos de pensar. Cada uno complementa al otro.

Por ejemplo, en 1736, el gran matemático Leonhard Euler dirigió su atención a un pequeño y curioso rompecabezas acerca de unos paseantes que cruzan puentes. Sabía que era interesante porque parecía exigir un nuevo tipo de geometría que abandonase las nociones habituales de longitud y ángulo. Pero en modo alguno podía prever que la disciplina a la que dio origen con su solución ayudaría, en el siglo XXI, a más pacientes a recibir los trasplantes de riñón necesarios para salvar su vida. Para empezar, este tipo de intervenciones

habría parecido pura fantasía en esa época. Pero incluso de no haber sido así, cualquier relación con el rompecabezas se habría considerado ridícula.

¿Y quién podría haberse imaginado siquiera que el descubrimiento de curvas que recubren un espacio (aquellas que atraviesan *todos* los puntos de una superficie cuadrada) podría ayudar al servicio de comidas a domicilio para personas dependientes a planificar sus rutas de reparto? Desde luego, no los matemáticos que estudiaron estas cuestiones en la década de 1890. Estos estaban interesados en definir conceptos poco intuitivos, como «continuidad» y «dimensión», y se encontraron con que tenían que explicar en primer lugar por qué ciertas creencias matemáticas muy arraigadas podían ser erróneas. Muchos de sus colegas acusaron al proyecto en su conjunto de estar mal planteado y de ser negativo. Con el tiempo, todo el mundo se dio cuenta de que no servía de nada enterrar la cabeza en la arena y asumir que los problemas acabarían por arreglarse, cuando de hecho no era así.

No son solo las matemáticas del pasado las que se emplean de este modo. Los métodos para los trasplantes de riñón dependen de muchos desarrollos modernos de la idea original de Euler, entre ellas potentes algoritmos de optimización combinatoria, que escogen la mejor opción de entre un enorme abanico de posibilidades. La miríada de técnicas matemáticas que se emplean en las películas de animación incluye muchas que datan de la década pasada o menos. Un ejemplo es el «espacio de la forma», un espacio de infinitas dimensiones de curvas que se consideran que son la misma si la única diferencia entre ellas es un cambio de coordenadas. Se utiliza para que las secuencias de animación parezcan más continuas y naturales. La homología persistente, otro desarrollo muy reciente, surgió porque los teóricos matemáticos querían calcular con el ordenador complicados invariantes topológicos que cuentan agujeros multidimensionales en formas geométricas. Resultó que su método también era una manera eficaz de garantizar que las redes de sensores proporcionan una cobertura completa cuando protegen edificios o bases militares frente a terroristas u otros malhechores. Algunos conceptos abstractos de la geometría algebraica («grafos de isogenias supersingulares») pueden hacer que las comunicaciones por internet sean seguras frente a los ordenadores cuánticos. Estos son tan novedosos que en este momento solo existen en versiones rudimentarias, pero podrían dar al traste con los sistemas de encriptación actuales si consiguen desarrollar su potencial.

Las matemáticas no solo dan sorpresas así en ocasiones excepcionales. Lo han convertido en costumbre. De hecho, por lo que respecta a muchos

matemáticos, estos descubrimientos inesperados son sus aplicaciones más interesantes y la principal justificación para considerar que las matemáticas *son* una disciplina y no solo un cajón de sastre de trucos variados, uno para cada tipo de problema.

Wigner dijo a continuación que «la enorme utilidad de la matemática en las ciencias naturales es algo rayano en el misterio, y… no existe ninguna explicación racional para ello». Por supuesto, es verdad que en un principio las matemáticas surgieron a raíz de problemas en la ciencia. Pero a Wigner no le sorprendía la eficacia de la disciplina en las áreas para las que se había diseñado. Lo que le maravillaba era su efectividad en otros campos que no tenían relación aparente. El cálculo surgió a partir de la investigación de Isaac Newton del movimiento de los planetas, así que no es de extrañar que nos ayude a entender cómo se mueven estos. Sin embargo, *es* sorprendente que nos permita hacer estimaciones estadísticas de la población humana, como en la anécdota de Wigner, o que explique los cambios en las capturas de peces en el Adriático durante la primera guerra mundial[4], que rija los precios de las acciones en el sector financiero, que ayude a los ingenieros a diseñar aviones de pasajeros o que sea imprescindible para las telecomunicaciones. Porque el cálculo no se inventó para ninguno de estos propósitos.

Wigner estaba en lo cierto. La forma que tienen las matemáticas de aparecer de manera repetida y sin previo aviso en las ciencias físicas y en la mayoría de las otras áreas de la actividad humana es un misterio. Una solución que se ha propuesto es que el universo está «hecho» de matemáticas y que los seres humanos nos limitamos a dejar al descubierto este componente fundamental. No voy a entrar a debatir esta explicación aquí, pero, de ser correcta, sustituye un misterio por otro todavía más profundo. ¿*Por qué* está hecho el universo de matemáticas?

*

En un nivel más pragmático, puede decirse que las matemáticas tienen varias características que contribuyen a su irrazonable eficacia, en el sentido que le daba Wigner a esta expresión. Estoy de acuerdo en que una de ellas es su estrecha vinculación con las ciencias naturales, que se traslada al mundo de los seres humanos en forma de tecnologías transformadoras. Muchas de las grandes innovaciones matemáticas han surgido, de hecho, a partir de investigaciones científicas. Otras características hunden sus raíces en preocupaciones muy humanas. Los números se desarrollaron a partir de la

contabilidad básica (¿cuántas ovejas tengo?). Geometría *significa* «medición de la tierra» y estaba unida de manera inseparable a los impuestos sobre los campos y, en el antiguo Egipto, a la construcción de pirámides. La trigonometría surgió de la astronomía, de la navegación y de la cartografía.

Sin embargo, por sí solo esto no constituye una explicación satisfactoria. Muchas otras grandes innovaciones matemáticas *no* se han desarrollado a partir de investigaciones científicas ni de problemas humanos concretos. La motivación principal para descubrimientos/inventos tales como los números primos, los complejos, el álgebra abstracta o la topología, fueron la curiosidad humana y la intuición de un patrón. Esta es la segunda razón por la que las matemáticas son eficaces: quienes las practican las emplean para buscar patrones y para desentrañar la estructura subyacente. Persiguen la *belleza*, no en la forma, sino respecto a la lógica. Cuando Newton quiso comprender el movimiento de los planetas, encontró la solución al pensar como un matemático y buscar patrones ocultos enterrados en los datos astronómicos sin procesar. Entonces descubrió su ley de la gravitación universal**[5]**. Muchas de las ideas matemáticas más brillantes han carecido por completo de un motivo en el mundo real. Pierre de Fermat fue un abogado del siglo XVII que se dedicaba a las matemáticas como entretenimiento y que hizo descubrimientos fundamentales en teoría de números: patrones ocultos en el comportamiento de los números naturales ordinarios. Tuvieron que pasar tres siglos hasta que su trabajo en este campo encontró aplicaciones prácticas. Sin embargo, ahora mismo las transacciones comerciales que impulsa internet no serían posibles sin sus aportaciones.

Otra característica de las matemáticas que se ha hecho cada vez más evidente desde finales del siglo XIX es la *generalidad*. Diferentes estructuras matemáticas tienen muchos rasgos en común. Las reglas del álgebra básica son las mismas que las de la aritmética. Distintos tipos de geometría (euclídea, proyectiva, no euclídea… incluso la topología) mantienen una estrecha relación entre sí. Esta unidad oculta puede hacerse explícita al trabajar, desde un primer momento, con estructuras generales que obedecen reglas establecidas. Comprendidas las generalidades, todos los ejemplos especiales pasan a ser evidentes. Esto ahorra mucho trabajo que, de otro modo, habría que malgastar repitiendo en numerosas ocasiones lo que viene a ser lo mismo en un lenguaje ligeramente diferente. Sin embargo, tiene una desventaja: tiende a hacer que la disciplina sea más abstracta. En lugar de tratar de objetos familiares, como los números, las generalidades deben referirse a todo aquello que obedezca las mismas *reglas* que los números, con

nombres tales como «anillo noetheriano», «categoría tensorial» o «espacio vectorial topológico». Cuando se lleva este tipo de abstracción hasta el extremo, puede ser difícil entender lo que *son* las generalidades, por no hablar de cómo emplearlas. Y, sin embargo, son tan útiles que el mundo de los seres humanos ya no funcionaría sin ellas. ¿Quieren Netflix? Pues alguien tiene que echar las cuentas. No es magia, aunque lo parezca.

Una cuarta característica de las matemáticas, muy relevante en este sentido, es su *portabilidad*. Es una consecuencia de la generalidad y el motivo por el que se necesita la abstracción. Independientemente del asunto que lo haya motivado, un concepto o un método matemático posee un nivel de generalidad que a menudo hace que sea aplicable a problemas muy diferentes. Si una cuestión cualquiera puede reformularse en el marco adecuado, entonces es abordable. La manera más sencilla y eficaz de crear matemáticas portátiles es incorporar la portabilidad al diseño desde el primer momento, al hacer explícitas las generalidades.

A lo largo de los últimos dos mil años, las matemáticas han buscado su inspiración en tres fuentes principales: las obras de la naturaleza, las obras de la humanidad y la tendencia característica de nuestra mente a buscar patrones. Estos tres pilares sustentan toda la disciplina. El milagro es que, a pesar de estos motivos multifacéticos, las matemáticas *son todas una sola cosa*. Sean cuales sean los orígenes y los objetivos de cada rama de la disciplina, esta ha acabado ligada de manera inseparable a todas las demás y sus vínculos se hacen cada vez más fuertes y complejos.

Este hecho señala un quinto motivo por el cual las matemáticas son tan eficaces y de modos tan insospechados: su *unidad*. Y de la mano de este va un sexto, del que aportaré abundantes pruebas a lo largo del texto: su *diversidad*.

Realidad, belleza, generalidad, portabilidad, unidad, diversidad. Juntas, conllevan utilidad.

No hay que darle más vueltas.

2
Cómo eligen los políticos a sus votantes

> Ankh-Morpork había coqueteado con diversas formas de gobierno hasta que, al final, se decidió por esa forma de democracia conocida como «un hombre, un voto». El patricio era ese hombre y el voto era el suyo.
>
> TERRY PRATCHETT, *Mort*[6]

La Grecia clásica ha legado muchas cosas al mundo: la poesía, el teatro, la escultura, la filosofía o la lógica. También hemos heredado de ella la geometría y la democracia, que han resultado tener una relación más estrecha de lo que cualquiera se hubiera imaginado, mucho menos los propios griegos. No se puede negar que el sistema político de la antigua Atenas era una forma muy limitada de democracia. Solo podían votar los hombres libres, ni las mujeres ni los esclavos. Aun así, en una época dominada por gobernantes hereditarios, dictadores y tiranos, la democracia ateniense era un avance innegable. Como lo era también la geometría griega, que de la mano de Euclides de Alejandría puso el énfasis en la importancia de exponer con claridad y concisión los supuestos básicos y de derivar todo lo demás a partir de estos de una manera lógica y sistemática.

¿Cómo diantres pueden aplicarse las matemáticas a la política? Esta tiene que ver con relaciones entre seres humanos, acuerdos y obligaciones, mientras que aquellas se refieren a una lógica fría y abstracta. En los círculos políticos, la retórica es más importante que la lógica y parecería que los cálculos inhumanos de las matemáticas están muy alejados de las disputas partidistas. Sin embargo, en democracia estas se desarrollan de acuerdo con unas reglas, que tienen consecuencias que no siempre pueden preverse cuando se establece la norma por primera vez. La obra pionera de Euclides en geometría, recogida en sus famosos *Elementos*, fijó un estándar en lo que se refiere a deducir consecuencias a partir de unas reglas. De hecho, esa sería una buena definición de las matemáticas en su conjunto. Sea como sea, el

caso es que apenas 2500 años después, esta disciplina ha empezado a infiltrarse en el mundo de la política.

Una de las curiosas características de la democracia es que los políticos que profesan devoción por la idea de que es «el pueblo» el que debe tomar las decisiones, se desviven en todo momento para asegurarse de que no sea así. Este comportamiento se remonta a la primera democracia existente, en la Grecia clásica, en la que solo tenían derecho al voto los atenienses varones mayores de edad, cerca de una tercera parte de la población adulta. Desde el momento en que se concibió la idea de elegir los líderes y decidir las políticas mediante el sufragio popular, también surgió el planteamiento, más atractivo incluso, de subvertir todo el proceso, al limitar quién podía participar y la efectividad de su intervención. Es algo fácil de hacer, incluso cuando se otorga un voto a cada elector, porque la eficacia del sufragio depende del contexto en el que se emite y este puede amañarse. Como ha expresado de manera muy diplomática el profesor de periodismo Wayne Dawkins, esto equivale a que sean los gobernantes quienes eligen a sus votantes, en lugar de ser los votantes quienes eligen a sus gobernantes[7].

Es ahí donde entran en juego las matemáticas. No en el tira y afloja del debate político, sino en la estructura de las reglas de este y en el contexto en el que se aplican. El análisis matemático es un arma de doble filo. Puede alumbrar maneras novedosas y astutas de amañar votaciones. Pero también es capaz de sacar a la luz estas prácticas y proporcionar pruebas incontestables de este tipo de manipulaciones, lo que en ocasiones puede evitar que se lleven a cabo.

Las matemáticas dicen también que cualquier sistema democrático debe incorporar un cierto nivel de compromiso. No es posible lograr todo lo que se quiere, por muy apetecible que sea, porque la lista de atributos deseables presenta contradicciones internas.

*

El 26 de marzo de 1812, la *Boston Gazzette* regaló a la lengua inglesa una nueva palabra: *gerrymander*, aunque al principio se escribía *Gerry-mander*. Es un término compuesto, algo a lo que más adelante Lewis Carroll se refirió como «palabras sobretodo», creadas al combinar dos expresiones habituales. En este caso, «*mander*» viene de *salamander* (salamandra en inglés) y «Gerry» se refiere a Elbridge Gerry, gobernador del estado de Massachusetts. No se sabe con certeza quién combinó ambos elementos por primera vez,

aunque basándose en pruebas circunstanciales los historiadores tienden a atribuir la autoría a alguno de los directores de la gaceta: Nathan Hale, Benjamin Russell o John Russell. Ya de paso, hay que decir que, aunque Gerry se pronuncia en inglés con «g», como «gato», *gerrymander* tiene un sonido inicial de «y», como «yerro».

¿Qué hizo Elbridge Gerry para merecer el dudoso honor de verse mezclado con una criatura con apariencia de reptil que, según los mitos medievales, habita en el fuego?

Amañó unas elecciones.

Más en concreto, fue responsable de un proyecto de ley que reformulaba los límites de las circunscripciones de Massachusetts de cara a los comicios para el Senado estatal. Demarcar distritos electorales produce divisiones entre unas zonas y otras de forma natural. Es algo habitual en la mayoría de las democracias, desde hace mucho tiempo. El motivo evidente es la operatividad: es difícil llegar a tomar una decisión si el país entero vota cada propuesta. (Suiza se aproxima a este modelo, porque su Consejo Federal elige unas propuestas que somete al voto de los ciudadanos hasta cuatro veces al año, en lo que vienen a ser una serie de referéndums. Por contra, las mujeres no obtuvieron el derecho al voto en este país hasta 1971, e incluso un cantón se resistió hasta 1991). Una solución que viene de antaño es que los votantes elijan una cantidad mucho menor de representantes y que sean estos quienes tomen las decisiones. Uno de los métodos más justos de hacerlo es la representación proporcional, según la cual el número de escaños de un partido político cualquiera es proporcional a los votos que ha recibido. Lo más habitual es dividir a la población en distritos y hacer que cada circunscripción elija una cierta cantidad de representantes, más o menos proporcional al número de votantes que reside en esa zona.

Por ejemplo, en las elecciones presidenciales estadounidenses, cada estado elige un número concreto de «electores», integrantes del Colegio Electoral. Cada uno de ellos tiene un voto y se decide quién es el presidente por mayoría simple entre estos. Es un sistema que tiene su origen en una época en la que el único modo de hacer llegar un mensaje del interior del país hasta los centros de decisión era mandar una carta con un mensajero a caballo o por diligencia. El ferrocarril de larga distancia y el telégrafo no hicieron su aparición hasta mucho después. En esa época, contabilizar los votos de grandes cantidades de personas era demasiado lento[8]. Pero también es cierto que este sistema cedía el control a la élite que integraba el Colegio Electoral. A su vez, en las elecciones parlamentarias británicas, se divide el país en

circunscripciones (geográficas, en buena medida) y cada una elige un miembro del Parlamento. Después, el partido o coalición de partidos con una mayoría de representantes forma gobierno y nombra a uno de estos, mediante diferentes métodos, para ocupar el cargo de primer ministro. Este cuenta con un poder considerable y en muchos aspectos es equiparable a un presidente.

Hay también un motivo no declarado para filtrar las decisiones democráticas a través de un número reducido de guardianes: es más fácil amañar los resultados. Todos los sistemas como los que se han mencionado tienen defectos inherentes que a menudo redundan en resultados inesperados y que pueden aprovecharse en ocasiones para saltarse a la torera la voluntad del pueblo. En varias elecciones presidenciales recientes en Estados Unidos, el número total de votos populares emitidos a favor del candidato perdedor era mayor que el que había recibido el ganador. Es cierto que el método actual por el que se elige el presidente no depende de estos resultados, pero con las comunicaciones modernas, el único motivo para no adoptar un sistema más justo es que hay un montón de personas poderosas que prefieren que siga así.

Gerry-mander, según una ilustración de 1812 atribuida a Elkanah Tisdale.

El problema subyacente en estos casos es el del «voto inútil». En cada estado, un candidato necesita la mitad de los sufragios más uno (o medio, si el total es impar) para ganar. Ni un solo voto recibido por encima de este umbral supone diferencia alguna en cuanto al resultado en la etapa del Colegio Electoral. Así, en las elecciones de 2016, Donald Trump obtuvo 304 apoyos en esta fase, frente a los 227 de Hillary Clinton, a pesar de que ella obtuvo 2,87 millones de votos populares más que Trump. Así, este se convirtió en el quinto presidente de Estados Unidos elegido pese a haber perdido el voto popular.

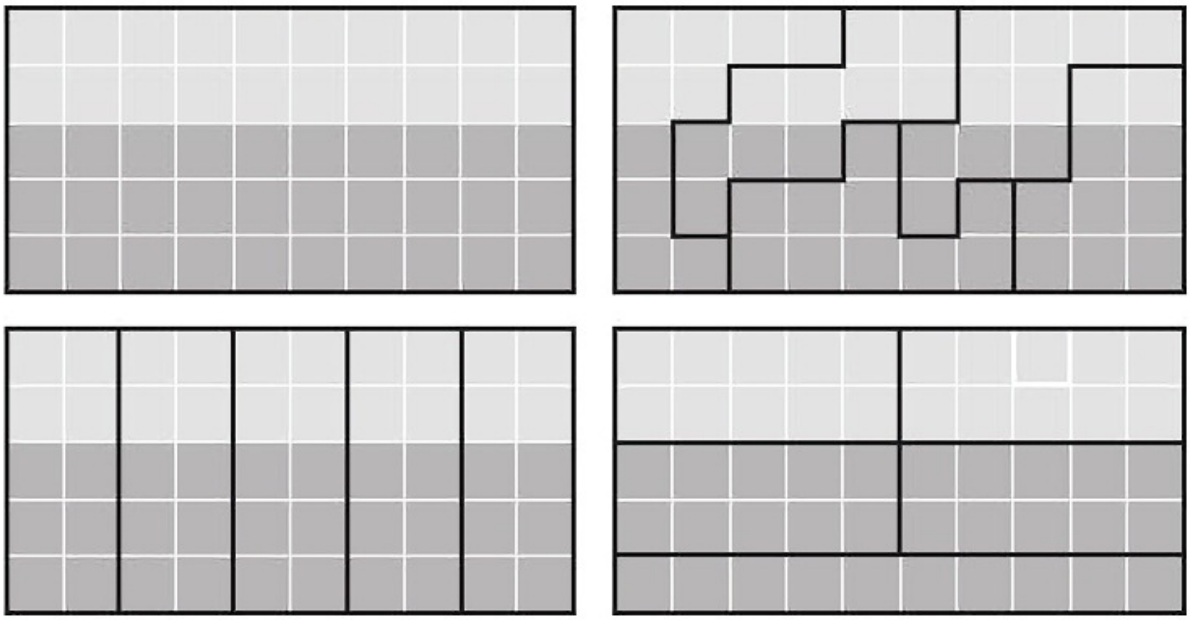
Las fronteras de los estados en Estados Unidos son inamovibles a todos los efectos, así que este no es un problema de demarcación de

circunscripciones. Sin embargo, en otras elecciones, los límites de los distritos sí pueden modificarse. Lo más habitual es que lo haga el partido en el poder, lo que deja al descubierto un fallo más insidioso. En concreto, que dicho partido puede modificar los límites para garantizar que se desperdician cantidades extraordinariamente grandes de sufragios favorables a la oposición. Entran a escena Elbridge Gerry y las elecciones al Senado. Cuando los votantes de Massachusetts vieron el mapa de los distritos electorales, la mayoría parecían del todo normales. Pero uno no. Reunía doce condados del oeste y del norte del estado en una única circunscripción, de gran extensión y forma sinuosa. Para el caricaturista que hizo el dibujo que poco después se publicó en la *Boston Gazzette* (con toda probabilidad el pintor, ilustrador y grabador Elkanah Tisdale), este distrito se parecía mucho a una salamandra.

Gerry pertenecía al Partido Demócrata-Republicano, que se enfrentaba a los federalistas. En las elecciones de 1812, estos ganaron el Congreso y el puesto de gobernador del estado, por lo que Gerry perdió su cargo. Sin embargo, su modificación de los distritos electorales funcionó a las mil maravillas y su partido no tuvo problemas para conservar el Senado estatal.

*

Las matemáticas de la manipulación de circunscripciones parten de fijarse en la manera en que esta se lleva a cabo. Hay dos tácticas principales: concentración y dispersión. La *concentración* consiste en repartir los votos propios de la manera más uniforme posible, de modo que se mantenga una mayoría pequeña pero decisiva en todos los distritos en que se pueda, y se ceden los demás al contrario. ¡Se siente, oposición! La *dispersión* fragmenta los votos del contrincante de modo que pierde todas las circunscripciones posibles. La *representación proporcional*, en la que la cantidad de escaños de cada partido es proporcional a su número total de votos (o tan próxima a esta distribución como sea factible dadas las cifras), evita que se puedan hacer estos trucos y es más justa. No obstante, no es ninguna sorpresa que la Constitución de Estados Unidos la declare ilegal porque, conforme a la normativa actual, los distritos deben tener solo un representante. En 2011, se celebró un referéndum en Reino Unido sobre otra alternativa, el voto individual transferible, pero el resultado fue contrario a esta modificación. Nunca se ha sometido la representación proporcional a referéndum en Reino Unido.



El reparto de Yerimandia. *Esquina superior izquierda*: se reparten cincuenta provincias en cinco circunscripciones de diez cada una. Se sabe que los votantes prefieren el Partido Claro o el Oscuro según el sombreado. *Esquina superior derecha*: la concentración otorga tres circunscripciones a los claros y solo dos a los oscuros. *Esquina inferior izquierda*: la dispersión otorga a los oscuros los cinco distritos. *Esquina inferior derecha*: esta disposición corresponde a la representación proporcional.

A continuación, se verá cómo funcionan la concentración y la dispersión en un ejemplo ficticio, con una geografía y una distribución del voto muy sencillas.

En el estado de Yerimandia hay dos partidos políticos que se presentan a las elecciones: el Claro y el Oscuro. Hay cincuenta provincias que se van a distribuir en cinco circunscripciones. En la última votación, el Partido Claro obtuvo la mayoría en veinte de ellas, todas al norte, mientras que el Oscuro lo hizo en las treinta que hay al sur (esquina superior izquierda). El Gobierno claro, que fue elegido por los pelos en las elecciones anteriores, ha modificado los límites de las circunscripciones en el estado. Ha concentrado más votantes suyos en tres de los distritos (esquina superior derecha), de modo que gana en ellos, y a los oscuros solo les quedan dos. A consecuencia de ello, los oscuros recurren la medida en los tribunales sobre la base de que los contornos de las circunscripciones se han manipulado de manera evidente. Consiguen que se les otorgue el control de la demarcación de límites para las próximas elecciones y usan la dispersión (esquina inferior izquierda) para garantizar que van a ganar en los cinco distritos.

Si las circunscripciones deben incluir diez de las pequeñas regiones cuadradas, lo mejor que pueden conseguir los claros mediante concentración son tres distritos de cinco. Necesitan ganar en seis de las diez regiones en una circunscripción para alzarse con la victoria y controlan veinte de ellas. Eso les

otorga tres veces seis cuadrados más otros dos, que no pueden aprovecharse. A su vez, lo mejor que pueden conseguir los oscuros mediante dispersión son los cinco distritos. Por último, la representación proporcional otorga dos circunscripciones a los claros y otras tres a los oscuros, como se ve en la imagen de la esquina inferior derecha (en la práctica, es imposible lograr una representación proporcional mediante la demarcación de distritos).

*

En los países gobernados por dictadores, o figuras equivalentes, suelen celebrarse elecciones para demostrar al mundo sus credenciales democráticas. Lo habitual es que estén amañadas, e incluso si se admiten impugnaciones ante los tribunales, nunca prosperan porque la justicia también está manipulada. En otros países no solo es posible denunciar en los juzgados cualquier caso concreto de modificación de los distritos electorales, sino que además se cuenta con una posibilidad de ganar, pues las decisiones de los tribunales son en su mayor parte independientes del partido gobernante. Excepto cuando se nombra a los jueces en función de sus simpatías partidistas, desde luego.

En estos casos, el dilema al que se enfrentan los magistrados no es político. El problema estriba en encontrar maneras objetivas de evaluar si se ha producido una manipulación partidista de los distritos electorales. Por cada «experto» que asegure tal cosa con solo echar un vistazo al mapa, siempre se podrá encontrar otro que llegue a la conclusión contraria. Se requieren métodos que sean más objetivos que las opiniones y los argumentos de palabra.

Esto presenta una oportunidad evidente para las matemáticas. Hay fórmulas o algoritmos que pueden cuantificar si los límites de un distrito son justos y razonables, o arbitrarios y partidistas, en un sentido definido con claridad. Desde luego, la confección de estas fórmulas o algoritmos no es en sí misma un proceso objetivo, pero una vez que se han consensuado (lo que es en parte un proceso político), todos los actores implicados saben cuáles son y sus resultados pueden comprobarse de manera independiente. Esto proporciona al tribunal una base lógica para sustentar su decisión.

Una vez que se han comprendido los métodos deshonestos que pueden emplear los políticos para hacer efectiva la modificación partidista de las circunscripciones es factible ingeniar entidades o reglas matemáticas para detectarlos. Sin embargo, no hay reglas infalibles. De hecho, existe una

demostración de que lo contrario es imposible, algo a lo que volveré una vez que se disponga del contexto necesario para apreciar lo que eso significa. En la actualidad, se emplean cinco tipos de enfoques:

- Detectar distritos con contornos extraños.
- Detectar desequilibrios en la proporción de escaños o votos.
- Cuantificar el número de sufragios inútiles que se producen con una división dada y compararlo con lo que se ha decidido que es aceptable dentro de la legalidad.
- Evaluar todos los mapas electorales posibles, estimar el resultado probable en términos de escaños asignados a partir de los datos disponibles del electorado y ver si el mapa propuesto constituye una anomalía estadística.
- Preparar protocolos que garanticen que la decisión final sea justa, que se perciba como tal y que las partes implicadas estén de acuerdo en que lo es.

El quinto enfoque es el más sorprendente, porque es una sorpresa que se pueda hacer en absoluto. Veámoslos uno a uno, dejando el más inesperado para el final.

*

Primero, contornos extraños.

Ya en 1787, James Madison escribió en *El federalista* que «el límite natural de una democracia reside en esa distancia del punto central que justamente permita a los ciudadanos más alejados reunirse tan frecuentemente como lo exijan sus funciones públicas». Tomado de manera literal, proponía que las circunscripciones fuesen más o menos circulares y no tan grandes como para que el tiempo de viaje desde la periferia al centro se volviese inadmisible.

Por ejemplo, supongamos que el apoyo principal de un partido político se concentra en las regiones costeras. Incluir a todos estos votantes en un único distrito electoral daría como resultado un contorno largo, delgado y sinuoso que discurriría a lo largo de toda la costa. Algo artificial por completo en comparación con el resto de las circunscripciones, regulares, compactas y razonables. No sería difícil llegar a la conclusión de que pasa algo raro y de que el contorno se ha trazado para garantizar que se desperdician muchos de los votos de ese partido. Las formas caprichosas de los distritos manipulados

revelan a menudo su carácter partidista, como ocurrió en el caso de la circunscripción original de la que se ha derivado el término.

En círculos legales puede debatirse sobre lo que constituye un contorno inusual hasta que las ranas críen pelo. Por ello, en 1991, los abogados Daniel Polsby y Robert Popper propusieron una manera de cuantificar hasta qué punto es extraño un límite, lo que en la actualidad se conoce como la puntuación de Polsby-Popper[9]. Esta se calcula como

4π por el área del distrito/el cuadrado del perímetro del distrito.

Cualquiera que tenga algo de intuición matemática se verá atraído de forma inmediata por el factor 4π. Del mismo modo que el amigo en la anécdota de Wigner se preguntaba qué tenían que ver las poblaciones con los círculos, podemos dudar de que haya relación entre estos y la demarcación de distritos electorales. Sin embargo, la respuesta es sencilla y directa, lo que es un alivio: el círculo es la más compacta de todas las regiones posibles.

Este hecho en sí tiene una larga historia. Según las fuentes de la antigüedad griega y romana, sobre todo la *Eneida*, el poema épico de Virgilio, y las *Historias filípicas* de Cneo Pompeyo Trogo, la ciudad Estado de Cartago fue fundada por la reina Dido. Juniano Justino resumió en el siglo III d. C. el relato de la historia de Trogo, y que narraba una leyenda sorprendente. Dido y su hermano Pigmalión eran los herederos conjuntos de un rey de la ciudad de Tiro, cuyo nombre no se conserva. Cuando falleció este, los habitantes quisieron que Pigmalión gobernase en solitario, a pesar de su juventud. Dido fue entregada en matrimonio a su tío Acerbas, de quien se rumoreaba que poseía un tesoro escondido. Pigmalión deseaba hacerse con este, por lo que asesinó a Acerbas. Dido hizo como si tirara la supuesta montaña de oro al mar, aunque en realidad arrojó sacos de arena. Temerosa, con buen criterio, de la ira de Pigmalión, emprendió la huida, primero a Chipre y luego a la costa norte de África. Solicitó al rey bereber Jarbas que le otorgase una pequeña parcela de tierra en la que poder descansar durante un tiempo y él aceptó entregarle tanto terreno como pudiese rodear con una piel de buey. Dido cortó la piel en tiras muy finas y la dispuso en círculo alrededor de una colina cercana, que hasta el día de hoy lleva el nombre de Birsa, que significa «piel». El asentamiento se convirtió en la ciudad de Cartago y cuando esta prosperó, Jarbas le dijo a Dido que se tenía que casar con él o afrontar la destrucción de la urbe. Entonces ella encendió una enorme hoguera en la que sacrificó a muchas víctimas bajo la pretensión de hacerlo para honrar a su

primer marido y prepararse así para el matrimonio con Jarbas. Sin embargo, se arrojó a la hoguera y tras proclamar que prefería unirse a su primer esposo antes que someterse a los deseos de Jarbas, se suicidó con una espada.

No se sabe si Dido existió en realidad, aunque es seguro que Pigmalión sí lo hizo y algunas fuentes los mencionan a ambos. Por lo tanto, carece de sentido debatir la precisión histórica de la leyenda. Sea como fuere, oculta un mito matemático en su interior: Dido utilizó la piel para encerrar la colina dentro de un *círculo*. ¿Por qué un círculo? Porque, según afirman los matemáticos, sabía que esta es la forma geométrica que encierra el área más grande para una circunferencia dada[10]. Este hecho, que ostenta el rimbombante nombre de «desigualdad isoperimétrica», se conocía de manera empírica en la antigua Grecia, pero no se demostró de forma rigurosa hasta 1879, cuando Karl Weierstraß, que destacó en el análisis complejo, resolvió una insuficiencia que había en las cinco demostraciones diferentes publicadas por el geómetra Jakob Steiner. Steiner había demostrado que si existe una forma óptima, debe ser el círculo, aunque no consiguió probar la existencia de esta[11].

La desigualdad isoperimétrica afirma que

el cuadrado del perímetro es mayor que o igual a 4π multiplicado por el área.

Esto es válido para cualquier forma en el plano que se comporte lo bastante bien como para tener un perímetro y un área. Es más, la constante 4π es la mejor posible, porque no puede tomarse otra mayor, y «mayor que o igual a» se convierte en una igualdad solo cuando la forma en cuestión es un círculo[12]. La desigualdad isoperimétrica llevó a Polsby y a Popper a sugerir que la cantidad a la que me he referido como puntuación de Polsby-Popper (PP) es una manera eficaz de medir la redondez de una forma. Por ejemplo, la puntuación de las siguientes formas es:

Círculo, puntuación PP = 1
Cuadrado, puntuación PP = 0,78
Triángulo equilátero, puntuación PP = 0,6

La puntuación PP del distrito original manipulado por Gerry es de aproximadamente 0,25.

No obstante, la puntuación PP tiene deficiencias serias. En ocasiones, los contornos inusuales pueden ser inevitables debido a la geografía local, con elementos tales como ríos, lagos, bosques y la línea de la costa. Es más, una

circunscripción puede ser regular y compacta y haber sido manipulada de manera evidente. Un plano de 2011 de los límites electorales para el Senado en Pennsylvania era muy retorcido y poco natural, así que los legisladores republicanos elaboraron una propuesta en 2018 para sustituirlo. Se determinó que los distritos proyectados eran muy compactos conforme a cinco criterios que había especificado la Corte Suprema del estado. Sin embargo, un análisis matemático de las distribuciones de votantes dentro de esas circunscripciones demostró que los contornos eran muy partidistas y que habrían alterado los resultados de la votación.

Incluso la escala a la que se dibuja el mapa puede dar problemas. En este sentido, la pega principal es la geometría de fractales. Un fractal es una figura geométrica con una estructura definida a cualquier escala. Hay muchas formas naturales que parecen ser así o, al menos, que guardan más similitud con ellos que con los triángulos y círculos de Euclides. La línea de la costa y las nubes pueden modelarse de manera muy eficaz como fractales, lo que es un reflejo de su intrincada estructura. El término fue acuñado en 1975 por Benoît Mandelbrot, pionero y promotor de todo este campo de la geometría. La línea de la costa y el curso de los ríos son curvas fractales extremadamente ondulantes y las longitudes que se miden sobre ellas dependen de lo detallada que sea la escala que se emplee para hacer la medición. De hecho, la longitud de una curva de este tipo es infinita desde el punto de vista técnico, lo que traducido a la realidad cotidiana quiere decir que «esta medida crece sin límite cuanto más de cerca se mira». De modo que los abogados pueden discutir de manera interminable sobre la medida del perímetro, por no mencionar la manipulación partidista del distrito.

*

Dado que lo inusual de una forma es tan engañoso, puede intentarse algo más directo. ¿Corresponden los resultados de los sufragios a los patrones estadísticos de voto del electorado?

Si se disputan diez escaños y los votantes están divididos al 60-40, sería de esperar que un partido ganase seis y el otro, cuatro. Si uno de ellos los obtiene todos, es plausible sospechar que ha habido manipulación. Pero no es tan sencillo. Este tipo de resultados son habituales en sistemas electorales uninominales con mayoría simple. En las elecciones generales de 2019 en Reino Unido, el Partido Conservador obtuvo el 44 % de los votos, pero logró 365 escaños de un total de 650, lo que supone el 56 %. Los laboristas

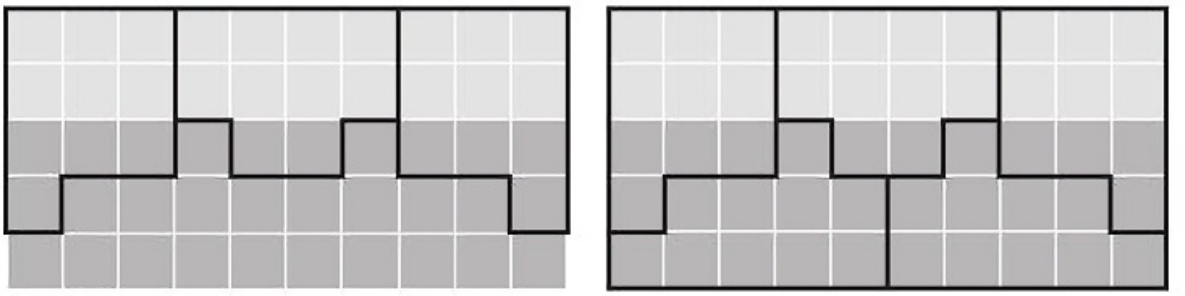
consiguieron el 32 % de las papeletas y el 31 % de los puestos en el parlamento. Los nacionalistas escoceses, con el 4 % de los votos, alcanzaron el 7 % de escaños (aunque este es un caso especial porque sus votantes residen todos en Escocia). El Partido Liberal Demócrata recibió un 12 % de apoyos y un 2 % de asientos parlamentarios. La mayoría de estas discrepancias fueron el resultado de patrones regionales de votación, no de contornos trazados de manera inusual. Después de todo, si unas elecciones para elegir a una persona, digamos a un presidente, se deciden por mayoría simple entre dos partidos, el 50 % de los votos más uno garantiza el 100 % de los resultados.

Pongamos un ejemplo tomado de Estados Unidos. En las elecciones presidenciales y federales de Massachusetts desde 2000, los republicanos han conseguido más de una tercera parte de los votos totales. Sin embargo, la última vez que un republicano consiguió un escaño en la Cámara de Representantes por ese estado fue en 1994. ¿Manipulación? Lo más probable es que no. Si se reparten de manera uniforme por todo Massachusetts ese tercio de votantes republicanos, entonces, sin importar cómo se tracen los contornos de los distritos (a no ser que adopten formas ridículas que serpentean por entre las casas de los ciudadanos individuales y las rodean), la proporción del voto conservador en cualquier circunscripción se situará en torno a una tercera parte. Los demócratas ganarán siempre. Y eso es justo lo que ocurre.

Los matemáticos han demostrado que este tipo de fenómeno puede ser inevitable en unas elecciones en el mundo real sin importar los contornos que se tracen (por lo menos, sin dividir poblaciones individuales). En 2006, Kenneth Chase intentó arrebatar el escaño de Edward Kennedy en el Senado de Estados Unidos, en un momento en que Massachusetts estaba dividido en nueve distritos electorales. Chase obtuvo el 30 % del voto total, pero perdió en las nueve circunscripciones. Los análisis por ordenador de las probabilidades demostraron que no podría haber ganado en ningún conjunto de poblaciones del tamaño de un distrito, ni siquiera aunque estuviesen esparcidas de manera irregular por todo el estado. Sus apoyos estaban distribuidos de un modo bastante uniforme en la mayoría de los sitios. No se podrían haber manipulado los límites para conseguirle una victoria *fuesen los que fuesen* los contornos trazados.

De vuelta en Yerimandia, cuando los oscuros ganaron los cinco distritos, los claros se opusieron a esta redistribución concreta de las circunscripciones alegando que sus contornos rectangulares eran demasiado largos y estrechos,

lo que demostraba sin lugar a duda que los ganadores habían practicado la dispersión. El tribunal sentenció que los distritos debían ser más compactos. Los claros propusieron tres circunscripciones con esta característica y ofrecieron con generosidad dejar que sus contrincantes decidiesen cómo dividir el territorio restante en otras dos más. Pero estos se opusieron, porque hacerlo así otorgaba tres distritos a los claros y solo dos a los oscuros, a pesar de que estos tenían una parte mayor del voto.



Izquierda: propuesta de los claros, con dos distritos a definir por los oscuros. *Derecha*: la opción más compacta posible.

Este reparto saca a la luz dos defectos más en el empleo de la compactibilidad para detectar la manipulación de sus contornos. Aunque no se puede negar que *es* compacto (hasta el momento), otorga a los claros las tres quintas partes de las circunscripciones con dos quintos del voto. Es más, no hay manera de dividir el territorio restante en dos distritos que tengan esta característica. La geografía de Yerimandia hace que sea difícil lograr un reparto compacto y justo a la vez. Tal vez lo haga imposible, depende de las definiciones.

*

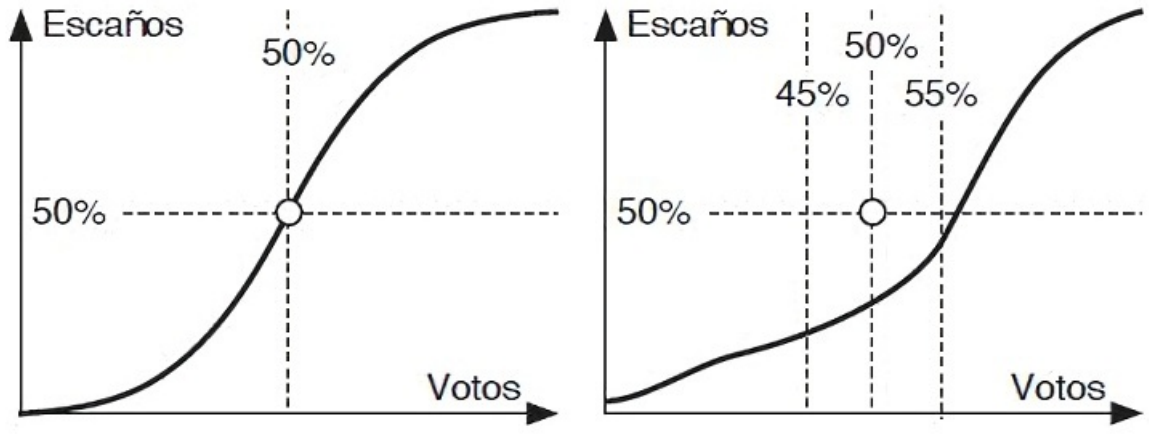
Dado que este enfoque es deficiente, ¿qué otra cosa puede hacerse para detectar la redistribución partidista de circunscripciones? Los datos electorales no solo informan del resultado de una votación, sino de cuál podría haber sido este si los votos obtenidos por cada partido variasen en cantidades concretas. Por ejemplo, si en un solo distrito los oscuros obtienen 6000 apoyos y los claros 4000, ganan los primeros. Si 500 votantes cambiasen para apoyar a la minoría, el resultado sería el mismo, pero si fuesen 1001, habrían dado la vuelta a las elecciones. Si en vez de esto los votos hubiesen sido 5500 para los oscuros y 4500 para los claros, solo haría falta que cambiasen 501 papeletas para alterar el resultado. En resumen, las cifras de las votaciones en un distrito no solo dicen quién gana, sino que informan del margen en el resultado.

Si se realiza este cálculo para cada circunscripción y se combinan los resultados, es posible ver cómo varía el número de escaños obtenidos en función del cambio en los sufragios y obtener una curva de escaños frente a votos (en realidad es un polígono con muchas aristas rectas, pero es más conveniente redondearlo). La imagen de la izquierda muestra de manera aproximada cómo se vería esta gráfica para unas elecciones en las que *no* se han manipulado los límites de los distritos. En concreto, la curva debería cruzar el umbral del 50 % de los escaños para un reparto de votos al 50 % y ser simétrica a ambos lados de este punto al girarla 180°.

La imagen de la derecha muestra la curva de escaños frente a votos para un mapa empleado en las elecciones al Congreso en Pennsylvania, con los apoyos a los demócratas a lo largo del eje horizontal. Estos habrían tenido que conseguir el 57 % de los sufragios para asegurarse el 50 % de los escaños. Consecuentemente, este mapa fue anulado por el Gobierno del estado.

En varias ocasiones, la Corte Suprema de Estados Unidos ha rechazado acusaciones de manipulación de límites a partir de este tipo de cálculo, del mismo modo que ha desestimado denuncias porque los distritos no eran compactos. En el caso LULAC contra Perry de 2006, ordenó que se trazasen de nuevo los contornos de un pequeño número de circunscripciones en Texas, en vista de que una de ellas se había hecho de forma que contravenía la Ley de Derecho al Voto. De hecho, aunque la Corte Suprema ha dictaminado que la manipulación partidista de los límites es inconstitucional, todavía no ha anulado ningún mapa de distribución de distritos completo.

*



Representación de escaños frente a votos. El eje horizontal muestra el porcentaje de votos recibido por un partido y va del 30 % al 70 %. El eje vertical muestra el porcentaje de escaños que se habría obtenido

para esos resultados.

Un motivo muy importante que ha aducido la Corte para fundamentar su decisión negativa ha sido que métodos como el de la curva de escaños frente a votos se basan en hipótesis, en lo que los electores *habrían hecho* en otras circunstancias. Este argumento puede tener sentido entre abogados, pero no desde el punto de vista matemático, porque la curva se deduce a partir de los datos electorales reales mediante un procedimiento definido con precisión. Transferir votos para calcular la curva no depende de lo que podrían haber hecho unos votantes en la realidad. Es como ver un resultado de 101 a 97 en baloncesto y determinar que el partido ha estado ajustado, mientras que un marcador de 120 a 45 implica que no lo fue. No se hace predicción alguna sobre lo que podrían haber hecho los integrantes individuales del equipo si hubiesen jugado mejor o peor. De modo que este caso puede añadirse a la larga y poco decorosa lista de las incapacidades del ámbito legal para entender, o incluso apreciar, las matemáticas básicas. Desde luego, la supuesta naturaleza hipotética de este algoritmo, que se basa por completo en los hechos, proporcionó la excusa perfecta para no anular todo el mapa de Texas.

*

La mejor manera de abordar decisiones legales polémicas no es intentar educar a los jueces, así que quienes buscaban métodos matemáticos para detectar la manipulación de los límites de distrito buscaron otras medidas que no pudiesen rechazarse sobre bases espurias. La manipulación obliga a los simpatizantes de un partido a malgastar muchos de sus votos. Una vez que un candidato obtiene una mayoría, los apoyos adicionales no tienen efecto sobre el resultado. Así que una manera de cuantificar la justicia, o la falta de ella, de una determinación de contornos es exigir que ambos partidos desperdicien de manera aproximada el mismo número de papeletas. En 2015, Nicholas Stephanopoulos y Eric McGhee definieron un método para medir los votos inútiles: la diferencia de eficacia[13]. En el caso de Gill contra Whitford de 2016, un juzgado de Wisconsin declaró ilegal el mapa de la asamblea estatal y la diferencia de eficacia fue determinante en la decisión. Para ver cómo se calcula este parámetro, consideremos una elección simplificada con solo dos candidatos.

Hay dos maneras principales de desperdiciar el voto. Un apoyo otorgado a un candidato perdedor se malgasta porque para el caso no merece la pena ni

molestarse en emitirlo. Uno de más para el ganador, una vez que este ha alcanzado el 50 %, es inútil por el mismo motivo. Estas afirmaciones dependen del resultado final y tienen validez a toro pasado: no puede saberse si el voto se ha desperdiciado hasta que se conoce el resultado. En las elecciones generales de Reino Unido de 2020, el candidato laborista en mi circunscripción obtuvo 19 544 papeletas, mientras que el conservador consiguió 19 143. Ganó el Partido Laborista por 401 votos de un total de 38 687 sufragios emitidos a favor de los dos partidos. Si algún votante individual hubiese decidido no tomarse la molestia de acudir a las urnas, la ventaja habría seguido en 400 votos. Pero si tan solo el 1 % de quienes apoyaron al laborismo hubiesen decidido no acudir, habría ganado el candidato de derechas.

Según la definición de voto desperdiciado, los votantes conservadores habrían malgastado un total de 19 143 sufragios y los laboristas, de 200. La diferencia de eficacia mide el punto hasta el que un partido se ve obligado a desperdiciar más apoyos que el otro. En este caso es:

> El número de votos conservadores desperdiciados menos el número de votos laboristas desperdiciados dividido por el número total de votos.

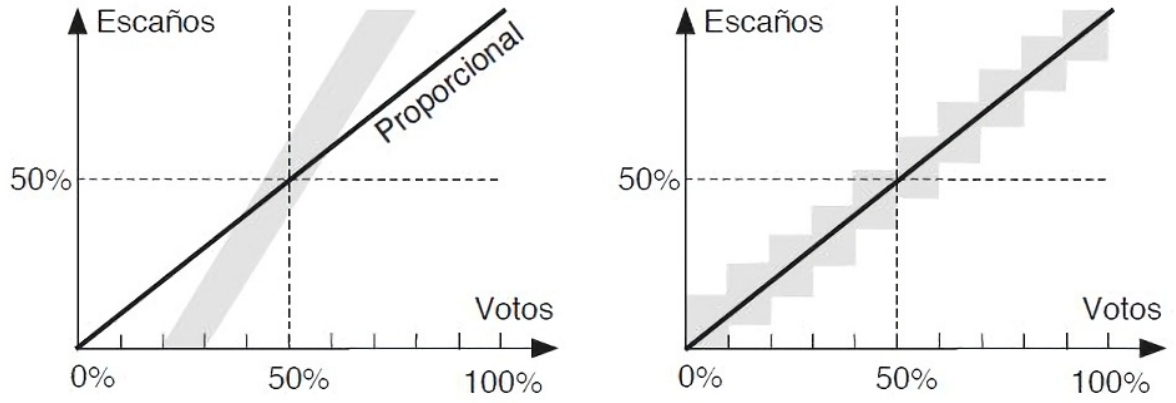
Es decir: (19 143 − 200)/38 687, que es +49 %.

Esto es solo una circunscripción. La idea es calcular la diferencia de eficacia de todas ellas combinadas y hacer que los legisladores fijen un límite legal. Su valor se encuentra siempre entre −50 % y +50 % y una diferencia justa es 0, porque en ese caso ambos partidos desperdician el mismo número de votos. Por eso Stephanopoulos y McGhee sugirieron que un parámetro fuera del intervalo ± 8 % es indicativo de manipulación.

Sin embargo, hay algunos fallos en este baremo. Cuando el resultado es muy ajustado, es inevitable que la diferencia de eficacia sea grande y unos pocos votos la pueden hacer cambiar de cerca de +50 % a cerca de −50 %. Mi circunscripción no había sido manipulada, a pesar de una diferencia de eficacia del +49 %. Con tan solo 201 votantes laboristas que hubiesen apoyado a los conservadores, el valor habría sido −49 %. Si un partido tiene suerte y gana en todas las circunscripciones, parecerá que ha logrado la victoria mediante manipulación. Factores demográficos pueden distorsionar también los números. En el caso de Gill contra Whitford, la defensa señaló, con razón, estos fallos, pero los demandantes argumentaron con éxito que

carecían de relevancia en ese caso concreto. No obstante, como comentario general, son razonables por completo.

En 2015, Mira Bernstein y Moon Duchin[14] se percataron de otros defectos en la diferencia de eficacia y en 2018, Jeffrey Barton sugirió una mejora para eliminarlos[15]. Por ejemplo, supongamos que hay ocho distritos y que los claros obtienen 90 votos en cada uno de ellos, mientras que los oscuros consiguen los 10 restantes. Los primeros desperdician $40 \times 8 = 320$ votos, mientras que los segundos malgastan $10 \times 8 = 80$, de modo que la diferencia de eficacia es $(320 - 80)/800 = 0{,}3 = 30\,\%$. Si se acepta la sugerencia de un umbral del 8 %, este valor del parámetro indica una manipulación partidista *en contra* de los claros, ¡a pesar de que han obtenido los ocho escaños!



Izquierda: gráfica de escaños frente a votos. Se muestra la representación proporcional (línea gruesa) y el intervalo en el que se considera que la diferencia de eficacia es justa (zona sombreada). *Derecha*: gráfica correspondiente a la diferencia de eficacia modificada. La zona sombreada rodea la línea diagonal.

Un segundo escenario pone de relieve otro problema. Supongamos ahora que los claros ganan tres distritos por 51 votos frente a 49, mientras que los oscuros se alzan con la victoria en otros dos, también por 51 a 49. En este caso, los primeros desperdician $1 + 1 + 1 + 49 + 49 = 101$ apoyos y los segundos, $49 + 49 + 49 + 1 + 1 = 149$ votos. La diferencia de eficacia es $(101 - 149)/500 = -0{,}096 = -9{,}6\,\%$, lo que indica una manipulación en contra de los oscuros. No obstante, estos son el partido minoritario y no deberían esperar ganar en más de dos circunscripciones, como así ocurre. Otorgarles otro escaño sería conceder a la opción minoritaria la mayoría de los representantes.

Barton atribuye ambos problemas al empleo de los votos desperdiciados en bruto. En *cualquier* elección, los apoyos de más a favor del ganador son inútiles, sin importar cómo se tracen los límites. Por eso sustituye «votos

desperdiciados» por «votos desperdiciados de manera innecesaria», calcula la proporción de sufragios a favor de cada partido que están destinados a malgastarse y los resta de la definición previa de votos desperdiciados. Con el modelo original, la gráfica de escaños frente a votos produce una banda estrecha en torno a la recta que va del 25 % de los apoyos obtenidos, en la parte inferior, al 75 % en la superior, como se ve en la imagen de la izquierda. En comparación, la línea diagonal muestra la gráfica ideal de la representación proporcional. Ambas coinciden solo en un punto muy cercano a una distribución de papeletas al 50/50. En cambio, con los votos desperdiciados de manera innecesaria, se obtiene la gráfica correspondiente que se muestra a la derecha. Rodea muy de cerca la diagonal, lo que es mucho más razonable.

*

Un método diferente para detectar la manipulación consiste en plantear mapas alternativos y comparar los resultados previsibles, a partir de datos sobre patrones de voto probables para toda la región que se va a dividir en distritos. Si el reparto propuesto por los oscuros les otorga el 70 % de los escaños, pero la mayoría de los modelos alternativos solo les conceden el 45 %, es que traman algo.

El principal problema de esta idea es que, incluso para números realistas de distritos y de divisiones, no es factible dibujar todos los mapas posibles. Se produce una explosión combinatoria y los números crecen a una velocidad vertiginosa. Es más, todos los mapas planteados deben ser acordes con la legislación, lo que introduce restricciones que pueden ser intratables de manera matemática. El caso es que hace ya mucho tiempo que los matemáticos descubrieron una forma de evitar esta explosión combinatoria: las cadenas de Márkov Montecarlo (MCMC, por sus siglas en inglés). En lugar de examinar todos los mapas posibles, las MCMC crean una muestra aleatoria de estos, lo bastante grande como para proporcionar estimaciones precisas. Es algo parecido al modo en que las encuestas de opinión estiman la intención de voto tras preguntar a una muestra aleatoria relativamente pequeña del público.

Los métodos Montecarlo se remontan al Proyecto Manhattan, durante la segunda guerra mundial, para la construcción de una bomba atómica. Un matemático llamado Stanislaw Ulam había estado enfermo y se encontraba convaleciente. Jugaba solitarios con una baraja para pasar el rato. Se preguntó

qué posibilidades tenía de ganar e intentó estimar cuántas formas de ordenar los naipes llevaban a un triunfo sin que sobrasen cartas. Enseguida se dio cuenta de que este enfoque no tenía solución. En su lugar, jugó muchas partidas y contó el número de veces que había ganado. Entonces se percató de que podía emplear un truco parecido con las ecuaciones físicas que necesitaba resolver el Proyecto Manhattan.

Las cadenas de Márkov, que reciben su nombre del matemático ruso Andrei Márkov, son generalizaciones de los recorridos aleatorios (o de borracho). Alguien que se ha tomado unas copas de más se tambalea por la acera y da trompicones hacia delante o hacia atrás de manera aleatoria. En promedio, tras un número dado de pasos, ¿qué distancia ha recorrido? (Respuesta: en promedio, en torno a la raíz cuadrada del número de pasos). Márkov se imaginó un proceso similar en el que se sustituye la acera por una red y en el que se asignan probabilidades a los desplazamientos a lo largo de las aristas de esta. Un aspecto crucial es saber cuál es la probabilidad de encontrarse en una ubicación dada tras desplazarse de un lado a otro durante mucho tiempo. Las cadenas de Márkov son un modelo de muchos problemas del mundo real en los que se producen secuencias de sucesos con probabilidades que dependen de las circunstancias del momento.

Las MCMC son el resultado de combinar ambas nociones: se emplean métodos Montecarlo para tomar una muestra de la lista de probabilidades necesaria. En 2009, el estadístico Persi Diaconis estimó que en torno al 15 % de los análisis estadísticos en ciencia, ingeniería y en el mundo de los negocios funcionan con MCMC, de modo que es lógico probar un método tan potente, útil y con tanta solera en el caso de la manipulación de límites electorales. Es posible emplear recorridos aleatorios *à la* Márkov para generar mapas de distribución de circunscripciones, tomar muestras con Montecarlo y ¡bingo!, ya hay un método estadístico para evaluar si un reparto propuesto es representativo. Unas matemáticas más sofisticadas apuntalan estos métodos, conocidas como teoría ergódica, que garantizan que recorridos aleatorios suficientemente largos son una muestra precisa de la estadística.

Algunos matemáticos han declarado hace poco ante los tribunales sobre las MCMC. En Carolina del Norte, Jonathan Mattingly se basó en una comparación de estimaciones de intervalos razonables de cantidades, tales como los escaños logrados, obtenidas mediante MCMC para argumentar que el mapa que se había elegido era una anomalía estadística muy poco probable, lo que demostraba que era partidista. En Pennsylvania, Wesley Pedgen se sirvió de métodos estadísticos para calcular lo improbable que era que un

mapa neutro desde el punto de vista político produjese resultados peores que otro creado mediante un recorrido aleatorio y para estimar la probabilidad de que algo así ocurriese por puro azar. En ambos casos, los jueces consideraron que la evidencia matemática era creíble.

*

La comprensión matemática de la manipulación de distritos es un arma de doble filo. Puede ayudar a los votantes y a los tribunales a detectar cuándo se produce, pero también puede sugerir maneras más eficaces de llevarla a cabo. Contribuye a hacer que se cumplan las leyes tanto como favorece que no se respeten, o peor aún, que se apliquen de manera partidista. Siempre que se redacta una normativa técnica para impedir algún tipo de abuso hay quien se aprovecha del sistema y escruta el texto en busca de lagunas. La gran ventaja de un enfoque matemático es que hace que las reglas en sí mismas sean claras. También da pie a una posibilidad completamente nueva. En lugar de realizar intentos vanos de persuadir a los intereses políticos opuestos de que se pongan de acuerdo sobre lo que es justo, lo que les da la oportunidad de aprovecharse del sistema, y de vigilar este mediante los tribunales, podría ser más razonable dejar que se *pelearan*. No en un enfrentamiento sin reglas, en el que el poder y el dinero constituyen ventajas enormes, sino en un marco estructurado para garantizar que el resultado es justo, que se percibe como tal y que, además, las partes implicadas no pueden evitar reconocerlo así.

Podría pensarse que se pide demasiado, pero en los últimos tiempos ha surgido toda una rama de las matemáticas dedicada a esta idea: la teoría de la división justa. Esta afirma que puede lograrse lo que en un principio parece imposible mediante marcos para la negociación estructurados con cuidado.

El ejemplo clásico, a partir del que surge todo lo demás, son dos niños que se pelean por una tarta. El problema radica en dividirla entre ellos mediante el empleo de un protocolo (un conjunto de reglas establecidas por adelantado) que sea justo de manera demostrable. La solución clásica es que uno parta los trozos y el otro elija el suyo. Se le dice a Alice que corte la tarta de modo que, a su juicio, ambos pedazos sean del mismo valor. Después, se pide a Bob que elija uno de ellos. No debería poner pegas, porque es él quien hace la elección. Puede elegir un trozo u otro. Alice tampoco puede tener queja alguna: si le parece que Bob ha elegido el pedazo más grande, debería haber empezado por cortar la tarta de otro modo. Si les preocupa quién va primero pueden decidirlo a cara o cruz, aunque eso no es necesario en realidad.

Como la naturaleza humana es como es, no es posible tener la certeza de que los niños van a considerar que el reparto es justo tras realizarlo. En una ocasión mencioné este método en un artículo y un lector me escribió para decirme que lo había probado con sus hijos y que Alice (nombre supuesto) se había quejado enseguida de que el pedazo de Bob (tampoco su verdadero nombre) era el más grande. Cuando su padre le dijo que la culpa era suya por haber cortado de forma desigual, la sugerencia no se recibió con mucho agrado que digamos. A su juicio, esto equivalía a culpar a la víctima, así que el padre les cambió los trozos. Pero enseguida se la oyó lloriquear: «¡El pedazo de Bob *sigue* siendo más grande que el mío!». No obstante, un protocolo de este tipo debería ser suficiente para contentar a los políticos o, al menos, para hacer que cierren la boca y sin duda debería ser aceptable en un tribunal. El juez solo tiene que comprobar que se ha seguido el protocolo de manera correcta.

La característica principal de este tipo de procedimiento es que, en lugar de intentar eliminar el antagonismo mutuo de Alice y de Bob, se *emplea* este para lograr un resultado justo. No se les pide que respeten las reglas, ni se les dice que cooperen ni se propone definición legal artificial alguna de lo que significa «justo». Tan solo se les deja *enfrentarse* entre sí y competir según las reglas. Por supuesto, Alice y Bob tienen que acordar de antemano que van a respetar esas normas, pero siempre van a tener que ponerse de acuerdo en algo y las reglas son justas y transparentes. Es probable que se tenga poca paciencia con cualquiera que se las salte.

Una característica importante de este método de reparto en el que uno corta y el otro elige es que no implica evaluación externa alguna del valor de los trozos de tarta. Emplea las propias valoraciones subjetivas de los participantes al respecto. Solo necesitan quedar conformes con la justicia de su pedazo *según su propio criterio*. En concreto, no es necesario que se pongan de acuerdo sobre el valor de ninguno de los trozos. De hecho, es más fácil hacer una división justa si no lo están. Uno quiere la guinda, el otro el glaseado y a ninguno de los dos le importa el resto: ¡conseguido!

Cuando los matemáticos y los sociólogos empezaron a tomarse en serio este tipo de problemas, surgieron complejidades notables que estaban ocultas. El primer avance se produjo cuando se plantearon cómo deberían hacer tres personas para repartir una tarta. No solo es notoriamente difícil dar con la solución más sencilla, sino que además surge un giro inesperado. Alice, Bob y Charlie pueden estar de acuerdo en que el resultado es justo, en el sentido de que han obtenido al menos una tercera parte de la tarta según sus propias

estimaciones, pero eso no impide que Alice todavía tenga envidia de Bob porque piensa que el trozo de él es más grande. A su juicio, el pedazo de Charlie debe compensar esta diferencia y ser más pequeño que el de ella. Sin embargo, no hay nada contradictorio en eso, porque Bob y Charlie pueden tener ideas diferentes del valor que tienen sus trozos *para ellos mismos*. De modo que tiene sentido buscar un protocolo que no solo sea justo, sino que esté a prueba de envidia. De hecho, es algo que puede lograrse[16].

La década de 1990 fue testigo de avances importantes en nuestra comprensión de los repartos justos y a prueba de envidia, que empezaron con un protocolo que cumplía estas condiciones para divisiones entre cuatro personas, descubierto por Steven Brams y Alan Taylor[17]. Por supuesto, la tarta no es más que una metáfora de cualquier cosa valiosa que sea susceptible de repartirse. La teoría toma en consideración elementos que pueden dividirse en porciones tan pequeñas como se desee (tarta) u otros que se presentan en unidades indivisibles (libros o joyas). Esto hace que sea aplicable a problemas de repartos justos de la vida real y Brams y Taylor explicaban cómo emplear estos métodos para solucionar conflictos en acuerdos de divorcio. Su protocolo del ganador ajustado presenta tres ventajas principales: es equitativo, está a prueba de envidia y es eficaz (u óptimo conforme al principio de Pareto). Es decir, todos los participantes tienen la impresión de que su parte es al menos igual de grande que el promedio, no albergan el deseo de cambiarla por la de nadie más y no existe otro reparto que sea al menos igual de bueno para todos ni mejor para alguien.

Por ejemplo, en un proceso de divorcio se aplicaría como sigue. Después de toda una vida de colaboración en el intercambio de mensajes criptográficos, Alice y Bob se hartan y deciden separarse. A cada uno se le otorgan 100 puntos, que reparten al asignar una cantidad de ellos a cada objeto: la casa, el televisor o el gato. En un primer momento, estos se entregan a quien les haya asignado una mayor puntuación. Esto es eficaz, pero lo habitual es que no sea justo ni esté a prueba de envidia, así que el protocolo pasa a la siguiente etapa. Si cada uno se lleva objetos correspondientes a una cantidad similar de puntos, ambos quedarán satisfechos y se da por terminado el reparto. Si no es así, supongamos que la parte de Alice, es mayor que la de Bob, de acuerdo con sus respectivas valoraciones. Entonces se transfieren objetos de Alice (la *ganadora*) a Bob (el *perdedor*) en un orden tal que garantice que se igualan las puntuaciones de ambos. Tanto las valoraciones como los objetos son discretos, por lo que puede ser necesario subdividir uno de estos, pero el protocolo implica que

este será el caso solo para uno, como mucho (con toda probabilidad será la casa, que tendrá que venderse para repartir el dinero, aunque no si Bob compró acciones de Apple antes de que subiesen en bolsa).

El método del ganador ajustado cumple tres condiciones importantes para una división justa. Tiene garantía de justicia: puede demostrarse que es equitativo y eficaz y que está a prueba de envidia. Funciona a través de valoraciones multilaterales: se tienen en cuenta las preferencias individuales y el valor de las porciones se calcula a partir de sus propias evaluaciones. Por último, es un procedimiento justo: ambos participantes pueden entender y verificar la garantía de justicia de cualquier solución a la que se llegue en última instancia y, si es necesario, un tribunal puede determinar que es justo.

*

En 2009, Zeph Landau, Oneil Reid e Ilona Yershov sugirieron que un enfoque similar podría eliminar el problema del reparto partidista de distritos electorales[18]. Un protocolo que impida que cualquiera de los participantes trace los límites de las circunscripciones para beneficio propio puede parar en seco los rastreros pasos del anfibio de la manipulación. Este método evita hacer consideración alguna sobre el trazado de los mapas y no otorga el poder de imponer estos a ningún actor externo supuestamente neutro. En vez de eso, está configurado de modo que los intereses en liza se equilibran entre sí.

Mejor aún, estos métodos pueden reforzarse para tener en cuenta factores adicionales, tales como la cohesión geográfica y que los distritos sean compactos. Si un organismo externo, tal como una comisión electoral, tiene que pronunciar la última palabra, es posible presentar los resultados del proceso de división como parte de la evidencia en la que basar su decisión. Nadie pretende afirmar que en la vida real estos métodos acaban con cualquier rastro de manipulación, pero funcionan mucho mejor que los modelos existentes y en buena medida evitan la tentación de incurrir en prácticas que son injustas a todas luces.

El protocolo es demasiado complicado como para describirlo en detalle e involucra a un actor independiente que propone una manera de repartir el estado. Se ofrece después a los partidos la posibilidad de alterar este mapa y de subdividir una de las porciones, teniendo en cuenta que se permite a la otra parte hacer lo propio con el resto. De manera alternativa, pueden elegir una opción similar con los papeles invertidos. Es una versión del método en el que uno parte los trozos y el otro elige el suyo, pero con secuencias de corte más

complicadas. Landau, Reid y Yershov han demostrado que su protocolo es justo desde el punto de vista de cada una de las partes. En resumen, ambos participantes se enfrentan entre sí en un juego, solo que este está diseñado para acabar en empate, con cada uno de los jugadores convencido de que lo ha hecho lo mejor posible. Si no es así, debería haber jugado mejor sus cartas.

En 2017, Ariel Procaccia y Pegden mejoraron este protocolo al eliminar al actor independiente, de modo que todo lo deciden las dos partes opuestas. En resumen, un partido político divide un mapa del estado en el número de distritos requerido por ley, con cantidades similares (tanto como sea posible) de votantes en cada uno. Después, la parte opuesta «bloquea» una de las circunscripciones, de modo que ya no puede cambiarse, y traza los contornos de las demás de la manera que desee. Entonces, el primer partido elige un segundo distrito de este nuevo mapa, lo bloquea y vuelve a dibujar el resto. Ambos actores se turnan a la hora de bloquear y volver a trazar los contornos, hasta que todo es inamovible. Así se decide el mapa definitivo que se va a emplear en el reparto de circunscripciones. Si hay, pongamos por caso, 20 distritos, este proceso se repite en 19 ciclos. Pedgen, Procaccia y un estudiante de informática teórica de intercambio, Dingli Yu, demostraron de forma matemática que este protocolo no otorga ventaja alguna al primer participante y que ninguna de las partes puede concentrar poblaciones concretas de votantes en el mismo distrito si la otra no lo desea.

*

Las matemáticas de las elecciones son en la actualidad una disciplina muy amplia y la manipulación de distritos es solo uno de sus aspectos. Se ha hecho mucho trabajo con los diferentes sistemas de votación: uninominal por mayoría simple, voto único transferible, representación proporcional, etcétera. Uno de los temas recurrentes que surgen a partir de estas investigaciones es que si se escribe una lista breve de las propiedades que son deseables en cualquier sistema democrático razonable, resulta que en algunas circunstancias estos requisitos son contradictorios entre sí.

El tatarabuelo de todos estos resultados es el teorema de la imposibilidad de Arrow, que publicó el economista Kenneth Arrow en 1950 y explicó en su libro *Elección social y valores individuales* al año siguiente. Arrow tomaba en consideración un sistema de votación numerado, en el que cada votante asigna una puntuación a una serie de opciones: 1 a su primera elección, 2 a la

siguiente y así en adelante. Estipuló tres criterios de justicia para un sistema de votación así:

- Si *todos* los votantes prefieren una alternativa a las demás, también lo hace el grupo.
- Si no cambian las preferencias de ningún votante entre dos opciones concretas, tampoco lo hacen las del grupo, incluso aunque cambien las preferencias entre otras alternativas.
- No hay un dictador que pueda determinar siempre la preferencia del grupo.

Todos son criterios deseables, pero contradictorios por lógica, como demostraba Arrow a continuación. Eso no quiere decir que un sistema así sea injusto por necesidad, sino que en algunas circunstancias el resultado es contrario al sentido común.

La manipulación de circunscripciones tiene sus propios descendientes del teorema de Arrow. Uno de ellos, publicado por Boris Alexeev y Dustin Mixon[19] en 2018, estipula tres principios para un reparto de distritos justo:

- *Una persona, un voto*: cada circunscripción tiene aproximadamente el mismo número de votantes.
- *Distritos compactos según Polsby-Popper*: todos ellos tienen una puntuación de Polsby-Popper mayor que alguna cantidad definida por ley.
- *Diferencia de eficacia restringida*: más técnico. De manera aproximada, si las poblaciones de dos distritos cualesquiera son como máximo una proporción fija de la población combinada de esos distritos, entonces la diferencia de eficacia es menor que el 50 %.

A continuación, demostraban que ningún sistema de reparto de circunscripciones puede satisfacer en todos los casos estos tres criterios.

Las democracias no pueden ser perfectas. De hecho, es sorprendente que funcionen en absoluto, dado que el objetivo es persuadir a millones de personas, cada una con sus propias opiniones, de que se pongan de acuerdo en temas importantes que les afectan a todas. Las dictaduras son mucho más sencillas. Un dictador, un voto.

3
¡Deja que la paloma conduzca el autobús!

> Por un lado, el conductor de autobús podría estar preocupado por que la paloma no fuese capaz de conducir el vehículo de forma segura. Por el otro, tal vez le preocupase más que la paloma no pudiese hacer una ruta para recoger de manera eficaz a todos los pasajeros en las diversas paradas a través de la ciudad.
>
> BRETT GIBSON, MATTHEW WILKINSON y DEBBIE KELLY,
> *Animal Cognition*

Mo Willems hace dibujos desde que tenía tres años. Le preocupaba que los adultos no fuesen sinceros en sus alabanzas, así que empezó a escribir historias graciosas. Pensaba que las risas forzadas serían más fáciles de detectar. En 1993, se unió al equipo de guionistas y animadores del emblemático programa *Barrio Sésamo*, con el que ganó seis premios Emmy en diez años. Su serie infantil de dibujos para televisión, *Oveja en la ciudad*, estaba protagonizada por Sheep, cuya idílica vida rural se ve arruinada cuando la organización militar secreta del General Concreto lo quiere para su pistola de rayos alimentada por energía ovina. Su primera incursión en la literatura infantil mantuvo el tema animal con *¡No dejes que la paloma conduzca el autobús!*, que le valió una medalla Carnegie por su adaptación animada y una mención de honor Caldecott, que reciben los seleccionados para la Medalla Caldecott. La protagonista es una paloma (obvio) que recurre a todos los trucos en el manual para convencer al lector de que se le debería dejar conducir un autobús cuando el conductor humano de siempre tiene que irse de repente.

El libro de Willems tuvo una secuela científica inesperada en 2012, cuando la muy respetable revista *Animal Cognition* publicó un muy respetable artículo de los muy respetables investigadores Brett Gibson, Matthew Wilkinson y Debbie Kelly. En él demostraban de manera experimental que las palomas pueden encontrar soluciones casi óptimas a casos sencillos de una conocida curiosidad matemática: el problema del viajante. El título de su

artículo fue «¡Deja que la paloma conduzca el autobús! Las palomas pueden planificar rutas futuras en una habitación»[20].

Que no se diga que los científicos no tienen sentido del humor. Ni que los títulos llamativos no ayudan a hacerse notar.

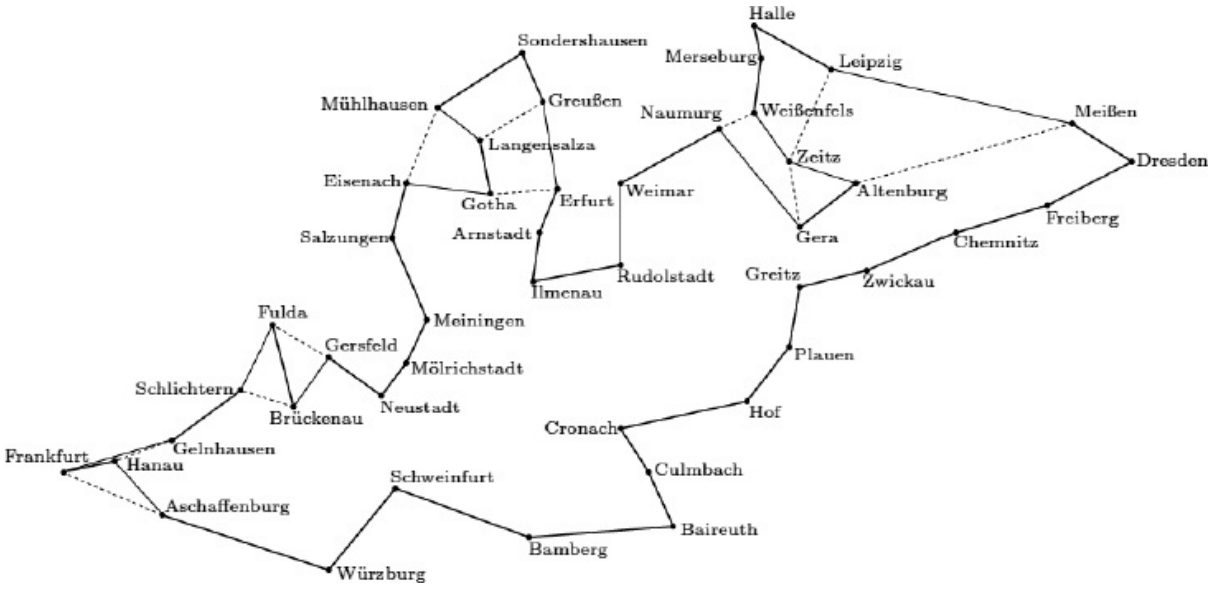
El problema del viajante no es solo una curiosidad. Es un ejemplo muy importante de un tipo de cuestiones con una relevancia práctica enorme denominado optimización combinatoria. Los matemáticos tenemos la costumbre de plantear interrogantes profundos y relevantes en términos en apariencia triviales. Unos congresistas de Estados Unidos han denunciado el derroche de fondos públicos en la teoría de nudos, ignorantes de que esta es un área fundamental de la topología de dimensiones bajas, con aplicaciones en el estudio del ADN y en teoría cuántica. Las técnicas básicas en topología incluyen el teorema de la bola peluda y el teorema del sándwich de jamón, así que supongo que nos lo hemos buscado. Pero no se trata solo de nosotros. No me preocupa el desconocimiento, es algo que le puede ocurrir a cualquiera, pero ¿por qué no *preguntan*[21]?

Sea como sea, el caso de trivialidad significativa al que se refiere este capítulo se origina en un libro útil para (quién lo iba a decir) viajantes de comercio. Vendedores puerta a puerta. Aún los recuerdo, aunque muchos no lo hagan. A menudo vendían aspiradoras. El viajante de comercio alemán de 1832, como cualquier hombre de negocios razonable (y en esa época siempre *era* un hombre), daba mucha importancia a emplear su tiempo con eficacia y a reducir costes. Por suerte, la solución estaba al alcance de la mano en la forma del manual *Der Handlungsreisende – wie er sein soll und was er zu thun hat, um Aufträge zu erhalten und eines glücklichen Erfolgs in seinen Geschäften gewiss zu sein – von einem alten Commis-Voyageur* (El agente comercial itinerante: cómo debe ser y qué tiene que hacer para conseguir pedidos y garantizar el feliz éxito de su negocio. Por un viajante de comercio veterano). Este anciano vendedor peripatético señalaba que:

> El negocio lleva al agente comercial itinerante ahora aquí, luego allí, y es imposible señalar de manera conveniente rutas de viaje que sean adecuadas para todos los casos que se presentan. Sin embargo, a veces, mediante una elección y una preparación cuidadosas del viaje puede ahorrarse tanto tiempo que no creo que podamos evitar dar algunas reglas sobre esto también… El aspecto más importante consiste siempre en visitar cuantos sitios sea posible sin tener que pasar dos veces por el mismo lugar.

El manual no proponía cálculo alguno para resolver este problema, pero daba ejemplos de cinco recorridos que se suponían óptimos alrededor de Alemania (uno pasaba a través de Suiza). La mayoría de ellos implicaban viajes secundarios en los que se pasaba por el mismo lugar dos veces, algo muy práctico para quien se aloja por la noche en una posada y pasa el día de visita por la zona. Pero había uno que no hacía paradas repetidas. Una solución moderna al mismo problema demuestra que la respuesta del manual es bastante buena, como puede verse en la imagen.

Este dilema se vino a conocer como el problema del viajante, o TSP, por sus siglas en inglés, pero luego se cambió a problema de la persona viajante, para evitar un uso sexista del lenguaje (tiene las mismas siglas en inglés, lo que es muy práctico). Es un ejemplo fundacional de la rama de las matemáticas conocida como optimización combinatoria. Lo que quiere decir «encontrar la mejor opción entre un conjunto de posibilidades que es demasiado grande como para comprobarlas una a una». Es curioso que el nombre TSP no parece haberse empleado de manera explícita en ninguna publicación referente a este problema antes de 1984, aunque su uso fuese habitual mucho antes en discusiones informales entre matemáticos.



Recorrido (1285 km) por 45 ciudades alemanas tomado del manual de 1832, representado por las rectas continuas (en negrita y delgadas). Las líneas continuas en negrita gruesa y discontinuas representan el recorrido más corto (1248 km) encontrado con métodos modernos.

Para ser una cuestión con un origen tan pragmático como este, el TSP ha llevado a la comunidad matemática a profundidades muy hondas, incluido el problema del milenio «¿P ≠ NP?», cuyo premio de un millón de dólares sigue a la espera de un ganador. Este dilema se pregunta, en un sentido técnico preciso, si dado un interrogante para el que puede *verificarse* de manera

eficaz una respuesta propuesta (una estimación aproximada, si se quiere), es posible *encontrar* siempre la solución con eficacia. La mayor parte de los matemáticos y de los informáticos teóricos creen que la respuesta es que no. Sin duda, comprobar cualquier estimación concreta puede hacerse con mucha más rapidez que dar con la solución correcta. Después de todo, si alguien muestra un rompecabezas de 500 piezas resuelto, lo más habitual es que un simple vistazo permita saber si lo ha hecho bien. En cambio, resolverlo desde el principio es harina de otro costal. Por desgracia, los rompecabezas no proporcionan una respuesta. Constituyen una metáfora útil, pero no tienen validez desde el punto de vista técnico. De modo que ahora mismo nadie puede demostrar que la creencia de que P es diferente de NP es verdadera, ni lo contrario. Ese es el motivo por el que lograrlo se paga con un estupendo millón de dólares[22]. Volveré sobre $P \neq NP$ más adelante, pero veamos antes los primeros avances en la resolución del TSP.

*

La época de los viajantes de comercio pasó hace mucho y la de las personas viajantes de comercio, menos machistas, le pisó los talones. En la era de internet, es muy raro que las empresas manden a alguien de ciudad en ciudad con una maleta llena de muestras para vender sus productos. Como es habitual (eficacia irrazonable), este cambio cultural no ha hecho que el TSP quede obsoleto. Con el crecimiento exponencial de las compras online, la necesidad de maneras eficaces de determinar rutas y horarios ha cobrado incluso más importancia para todo, desde paquetes o pedidos del supermercado hasta pizzas. Tal vez deberíamos cambiar el nombre del TSP a «problema de la compra del súper»: ¿cuál es el recorrido óptimo para la furgoneta de reparto?

La portabilidad de las matemáticas también entra en juego. Las aplicaciones del TSP no se limitan a recorridos entre ciudades o a lo largo de las calles de la ciudad. En la pared de nuestro salón cuelga un gran cuadrado de tela negra bordado en azul con un elegante patrón de espirales ribeteadas de lentejuelas basado en la conocida serie de Fibonacci. El diseñador lo ha llamado las «lentejuelas de Fibonacci»[23]. Lo hizo empleando una máquina controlada por ordenador capaz de bordar cualquier diseño hasta el tamaño de un edredón. La aguja que cose los hilos está unida a una varilla y puede desplazarse a lo largo de esta. A su vez, la varilla es capaz de moverse en perpendicular a su longitud. Es posible llevar la aguja a donde se quiera

mediante la combinación de ambos movimientos. Por razones prácticas (aprovechamiento del tiempo, desgaste de la máquina o ruido) no es deseable que la aguja salte de un lado a otro por todas partes, por lo que la distancia total de desplazamiento debe reducirse al mínimo. Es un problema muy parecido al TSP. El linaje de estas máquinas se remonta a los albores del diseño gráfico por ordenador y a un dispositivo conocido como plotter XY, que movía un rotulador de la misma manera.

Dilemas como este abundan en ciencia. Érase una vez que los astrónomos más destacados poseían sus propios telescopios o los compartían con unos pocos colegas. Podía variarse la dirección de los instrumentos con facilidad para apuntar a nuevos cuerpos celestes, así que improvisar era sencillo. Ya no es así, porque los telescopios que emplean los astrónomos son enormes, tienen un coste desorbitado y funcionan por internet. Se tarda mucho en dirigirlos hacia objetos nuevos y los instrumentos no pueden emplearse para hacer observaciones mientras se desplazan. Si los objetivos se recorren en el orden equivocado, se pierde mucho tiempo al mover el telescopio una gran distancia y luego volver casi al punto de partida. Al secuenciar ADN, las secuencias fragmentarias de las bases del ácido tienen que unirse de manera correcta entre sí y debe optimizarse el orden en el que se hace esto para evitar desperdiciar el tiempo de cálculo del ordenador.

Otras aplicaciones van desde trazar rutas para aeronaves de manera eficaz hasta el diseño y la fabricación de microchips y placas de circuitos impresos en informática. Se han empleado soluciones aproximadas del TSP para encontrar recorridos eficaces para el servicio de comidas a domicilio para personas dependientes y para optimizar el reparto de sangre en hospitales. Incluso una versión del TSP apareció en la «guerra de las galaxias», mejor llamada la hipotética Iniciativa de Defensa Estratégica del presidente Reagan, según la cual un potente láser que orbitase la Tierra se habría dirigido contra una serie de misiles nucleares atacantes.

*

Karl Menger, algunas de cuyas aportaciones se consideran ahora como precursoras de los fractales, parece haber sido el primer matemático que escribió sobre el TSP, en 1930. Llegó hasta el problema desde una perspectiva muy diferente. Había estado estudiando las longitudes de curvas desde el punto de vista de las matemáticas puras. En esa época, ese parámetro se definía como el valor más grande obtenido al sumar las longitudes de

cualquier aproximación poligonal a la curva cuyos vértices fuesen un conjunto finito de puntos en la curva, recorridos en el mismo orden en el que se sitúan sobre ella. Menger demostró que se obtiene el mismo resultado si se sustituye cada polígono por un conjunto finito de puntos sobre la curva y se calcula la distancia total mínima a lo largo de *cualquier* polígono con esos vértices en cualquier orden que se desee. La relación con el problema del viajante estriba en que el trayecto más corto de Menger es el que resuelve el TSP si se consideran los vértices del polígono como ciudades. Lo denominó «problema del mensajero», porque decía que era válido tanto para carteros como para agentes comerciales itinerantes y escribió:

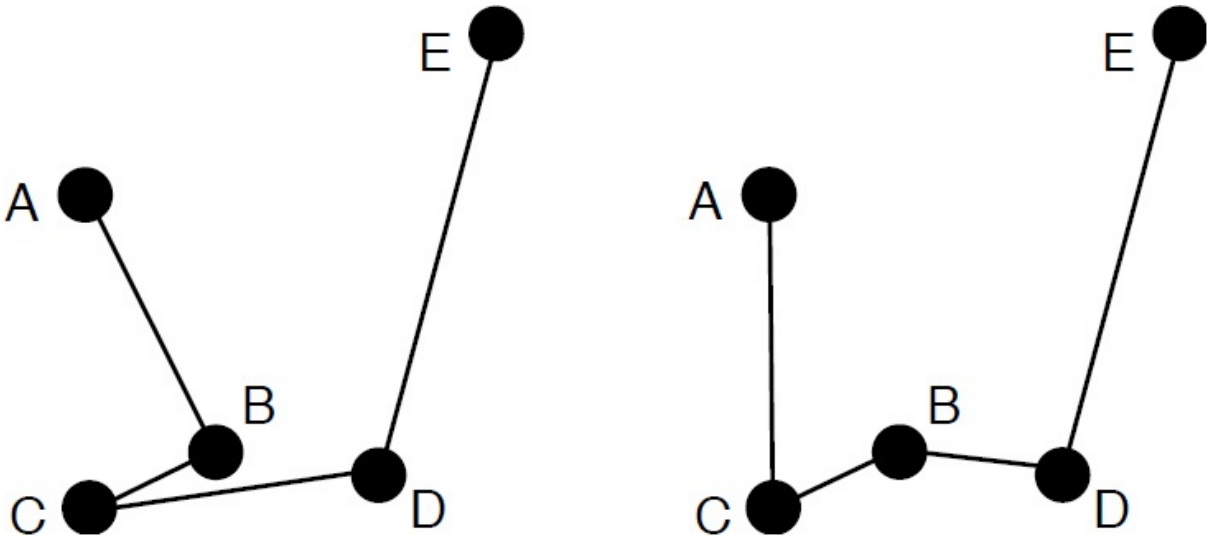
> Este problema puede resolverse mediante un número finito de intentos. No se conocen reglas que reduzcan el número de intentos por debajo de la cantidad de permutaciones de los puntos dados. La norma según la cual se debe ir desde el punto inicial al más cercano, luego al más próximo a este, etcétera, no conduce en general a la ruta más corta.

Esta cita demuestra que comprendía dos características principales del problema. La primera, es que *existe* un algoritmo para encontrar la respuesta. Solo hay que comprobar todos los recorridos uno tras otro, calcular sus longitudes y ver cuál es el más corto. El número total de trayectos posibles es precisamente la cantidad de permutaciones de los puntos, que es finita. En la época en que escribía, no se conocía otro algoritmo mejor, pero no es factible comprobar todas las posibilidades para más de una docena de ciudades o así porque hay demasiados recorridos. En segundo lugar, sabía que el método «evidente» (ir de un punto cualquiera al siguiente más cercano) no suele funcionar. Los expertos llaman a este método «heurística del vecino más cercano». El dibujo muestra un motivo por el que falla.

Menger fue profesor invitado en la Universidad de Harvard durante seis meses, entre 1930 y 1931, y el gran topólogo Hassler Whitney asistió a sus clases e hizo algunas sugerencias sobre la cuestión. Un año después, Whitney pronunció una conferencia en la que mencionó el descubrimiento del recorrido más corto entre los 48 estados que había entonces en Estados Unidos. El nombre «problema de los 48 estados» estuvo en circulación durante un tiempo y nadie parece saber con certeza quién acuñó la denominación más pegadiza de «problema del viajante». La primera referencia impresa conocida en la que se emplea el término TSP es un informe de Julia Robinson de 1949.

Menger prosiguió su trabajo en el TSP y en problemas relacionados. En 1940, Laszlo Fejes Tóth investigó lo que viene a ser la misma cuestión: encontrar el camino más corto a través de *n* puntos en el cuadrado unidad. En 1951, Samuel Verblunsky demostró que la respuesta tiene una longitud menor que $2 + \sqrt{2} \cdot 8n$. Varios matemáticos demostraron teoremas un poco mejores según los cuales la longitud mínima sobre *n* puntos en una región fija no es mayor que alguna constante multiplicada por la raíz cuadrada de *n*, para constantes que se hacían más pequeñas cada vez.

A finales de la década de 1940, una de las instituciones más destacadas en investigación operativa era la Corporación RAND de Santa Mónica, en California. Sus investigadores habían hecho mucho trabajo en una cuestión relacionada, el problema del transporte, y George Dantzig y Tjalling Koopmans sugirieron que sus aportaciones a lo que se conoce hoy en día como programación lineal podían ser relevantes al TSP. La programación lineal es un potente marco práctico para muchos problemas de optimización combinatoria. Constituye un método para maximizar alguna combinación lineal de variables, sometida a desigualdades que afirman que ciertas combinaciones lineales más deben ser positivas o negativas. Dantzig inventó el primer algoritmo práctico, el método símplex, que todavía se emplea mucho hoy en día. Las desigualdades definen un poliedro convexo multidimensional y el algoritmo mueve un punto a lo largo de las aristas que aumentan la cantidad que queremos maximizar hasta que se queda atascado.



Una manera en la que falla la heurística del vecino más cercano. En el recorrido de la izquierda, se parte de A y se va siempre a la ciudad más cercana de aquellas que todavía no se han visitado, por lo que se pasa por ABCDE en orden. No obstante, el recorrido de la derecha, que visita ACBDE, es más corto.

Los investigadores de RAND Dantzig, Delbert Fulkerson y Selmer Johnson hicieron el primer progreso significativo de verdad en el TSP mediante el

empleo del método de programación lineal de Dantzig. Lo adaptaron para aplicarlo a este problema e introdujeron nuevos métodos sistemáticos, en concreto, el uso de «planos de corte». La conclusión fue un límite inferior para la longitud óptima del recorrido. Si puede encontrarse un trayecto cuya longitud sea tan solo un poco mayor que la ideal es que se anda cerca de la respuesta, en cuyo caso la astucia instintiva puede en ocasiones rematar la tarea. Dantzig, Fulkerson y Johnson emplearon estas ideas para obtener la primera solución del TSP con un número razonable de poblaciones, en concreto, el recorrido más corto por 49 ciudades (una en cada uno de los 48 estados de Estados Unidos más Washington D. C.). Es probable que este fuese el problema que Whitney mencionó en la década de 1930 y es la versión exacta a la que hizo referencia Robinson en 1949.

*

En 1956, el pionero de la investigación de operaciones Merrill Flood argumentó que es presumible que el TSP sea complejo. Esto plantea una duda fundamental: ¿hasta qué punto? Para responderla, debemos volver a considerar P y NP, esas medidas de complejidad computacional que valen un millón de dólares. Parece muy probable que Flood tuviera razón.

Los matemáticos siempre han prestado atención a que los métodos para resolver problemas sean eficientes, aunque, si no hay más remedio, son de la opinión de que cualquier método es mejor que ninguno. Para propósitos estrictamente teóricos, el mero hecho de ser capaces de demostrar que *existe* una solución a un interrogante puede ser un adelanto enorme. ¿Por qué? Porque si no es posible tener la certeza de que haya una, buscarla podría ser una pérdida de tiempo.

Mi ejemplo favorito en este sentido es el que yo llamo «la tienda de Madre Gnat». Bebé Gnat flota sobre el suelo a 30 centímetros (metros, millas… cualquier cosa mayor que cero) de altura. Madre Gnat quiere hacer una tienda que tenga su base en el suelo y que cubra a Bebé Gnat y quiere emplear la menor cantidad de tela posible. ¿Qué tienda tiene el área más pequeña? Si representamos a Bebé Gnat por un único punto, la respuesta es que no existe tal cosa. Se puede hacer una que sea cónica, alta y delgada con cualquier superficie mayor que cero, pero si tiene un área superficial nula es una línea, no una tienda. Dada cualquiera de ellas, hay otra que cumple los requisitos empleando la mitad de tela. Así que no puede haber un área que sea la más pequeña.

Respecto al TSP, existe sin duda alguna una solución para cualquier conjunto finito de ciudades dispuestas como se quiera, porque el número de recorridos posibles es limitado. Esto garantiza que no se pierde el tiempo cuando se busca el trayecto más corto, pero no dice cuál es este. Si se busca desenterrar un tesoro escondido, no sirve de nada saber que está sin duda *en alguna parte* porque no es factible excavar todo el planeta.

El científico de la computación Donald Knuth señaló, hace mucho, que en informática hace falta algo más que una prueba de que existe una respuesta. Se necesita saber cuánto va a costar encontrarla. No en euros y céntimos, sino en esfuerzo de cálculo. La rama de las matemáticas que aborda este problema se llama teoría de la complejidad computacional. En muy poco tiempo ha pasado de ser unas pocas ideas sencillas a un conjunto sofisticado de teoremas y métodos. Sin embargo, hay una distinción básica que refleja en cierta medida, en términos muy simples, la diferencia entre una solución eficiente y otra que no lo es.

La cuestión principal es esta: ¿a qué velocidad crece el tiempo de ejecución (medido como el número de pasos en el cálculo) en cualquier método para calcular la respuesta a un interrogante en comparación con el tamaño de los datos necesarios para formular el problema en primer lugar? En concreto, si hacen falta n dígitos binarios para plantearlo, ¿cómo depende el tiempo de ejecución de n? Para algoritmos eficientes, este parámetro tiende a crecer como una potencia de n, digamos n^2 o n^3. Se dice que estos se ejecutan en tiempo polinómico y se simbolizan como clase P. Por el contrario, los que no son eficientes crecen mucho más rápido, a menudo en tiempo exponencial, como 2^n o 10^n. El algoritmo que prueba todos los recorridos en el TSP es de este tipo. ¡Se ejecuta en tiempo factorial $n!$, que crece a mayor velocidad que cualquier exponencial. Hay una zona gris entre medias en la que el tiempo de ejecución es superior a cualquier tiempo polinómico, pero inferior a uno exponencial. A veces estos algoritmos son eficientes, en otras ocasiones no. Para los propósitos de la presente consideración podemos adoptar un punto de vista muy estricto y arrojarlos a todos ellos a un cubo de basura marcado «no P».

Este no es equivalente a NP.

Estas siglas inducen a confusión porque representan una idea más sutil todavía: tiempo polinómico no determinista. Esta se refiere al tiempo de ejecución de un algoritmo que puede decidir si cualquier solución concreta propuesta es correcta. Recordemos que un número es *primo* si solo es divisible por 1 y por sí mismo. De modo que 2, 3, 5, 7, 11, 13 etcétera, son

primos. El resto son compuestos. Es decir, 26 es compuesto porque es igual a 2×13 y los números 2 y 13 son los factores primos de 26. Supongamos que se quiere encontrar un factor primo de una cifra con doscientos dígitos decimales. Nos pasamos un año entero a la búsqueda de uno, sin encontrarlo, y en nuestra desesperación consultamos el oráculo de Delfos. Nos dice que la respuesta es cierto número muy grande. No tenemos ni idea de cómo ha llegado a este resultado (después de todo, es un oráculo con milagrosas dotes de adivinación), pero podemos sentarnos y echar cuentas para ver si el número que nos ha dicho divide de verdad la enorme cifra que hemos tomado en consideración. Un cálculo así es muchísimo más fácil que encontrar el factor primo en sí mismo.

Supongamos que cada vez que el oráculo ofrece una solución podemos comprobar si es correcta mediante el empleo de un algoritmo de tiempo polinómico (P). En ese caso, el problema en sí mismo es de clase NP (polinómico no determinista). La tarea del oráculo es mucho más difícil que la nuestra, pero siempre podemos decidir si nos ha dado la respuesta acertada.

Es razonable que comprobar una solución propuesta sea mucho más sencillo que encontrarla. Desde un primer momento, ver si el tesoro está enterrado en un punto marcado con una X es más fácil que descubrir dónde está la X. Para poner un ejemplo matemático, casi todo el mundo piensa que encontrar los factores primos de un número es mucho más difícil que comprobar si un primo dado *es* un factor. La prueba más evidente es que se conocen algoritmos que pueden verificar a gran velocidad cualquier factor propuesto, pero no dar con él. Si $P = NP$, entonces dado cualquier problema que tenga una solución rápida y *comprobable*, también sería posible *encontrar* la respuesta con rapidez. Esto suena demasiado bien como para ser cierto y la experiencia de los matemáticos a la hora de resolver problemas es la contraria. Así que casi todo el mundo cree que $P \neq NP$.

No obstante, todos los intentos de demostrar esto, o lo contrario, han fracasado. Puede verificarse que un problema es NP escribiendo un algoritmo explícito y calculando su tiempo de ejecución, pero para comprobar que *no* está en P hay que tomar en consideración *todos los algoritmos posibles* para resolver el cálculo y demostrar que ninguno de ellos está en la clase P. ¿Cómo se hace esto? Nadie lo sabe.

Un hecho curioso que surge a partir de estos intentos es que una cantidad enorme de problemas candidatos están en la misma situación. Todos son NP. Es más, si puede demostrarse que uno cualquiera de ellos no pertenece a P, entonces ninguno pertenece a P. Uno para todos y todos para uno. Se dice que

problemas como este son NP-completos. Una categoría relacionada y más amplia es NP-complejo. Esta consiste en algoritmos que pueden simular la solución de *cualquier* problema NP en tiempo polinómico. Si resulta que este algoritmo tiene un tiempo de ejecución polinómico, esto demuestra de manera automática que lo mismo es cierto para cualquier problema NP. En 1979, Michael Garey y David Johnson demostraron que el TSP es NP-complejo[24]. Si se asume que $P \neq NP$, esto implica que todo algoritmo que lo resuelve tiene un tiempo de ejecución mayor que cualquier polinomio.

Flood tenía razón.

*

Eso no es excusa para rendirse, porque hay al menos dos salidas posibles.

Una, que pasaré a explorar enseguida, se basa en la experiencia de problemas en la práctica. Si un problema es no P, entonces no es factible resolverlo en el peor de los casos. Pero a menudo, el peor de los casos resulta ser muy retorcido y no es característico de los ejemplos que se encuentran en el mundo real. Así que los matemáticos que trabajaban en investigación de operaciones se propusieron ver cuántas ciudades podían gestionar en situaciones reales. Y resultó que las variaciones en el método de programación lineal propuestas por Dantzig, Fulkerson y Johnson se desempeñaban a menudo extraordinariamente bien.

En 1980, el récord estaba en 318 ciudades. En 1987, ya eran 2392 y para 1994, se había elevado el listón hasta las 7397, un resultado que consumió cerca de tres años de tiempo de CPU en una red de ordenadores muy potentes. En 2001, se obtuvo una solución exacta para 15 112 ciudades alemanas mediante el empleo de una red de 110 procesadores. Se habrían necesitado más de veinte años en un ordenador normal de sobremesa. En 2004, se resolvió el TSP para la totalidad de las 24 978 poblaciones que hay en Suecia y en 2005, el Concorde TSP Solver halló la solución al problema para un recorrido de todos y cada uno de los 33 810 puntos que había sobre una placa de circuito impreso. Batir récords no es el único motivo para estas investigaciones. Los métodos empleados para ello funcionan de hecho con mucha rapidez para problemas más pequeños. Es habitual que se puedan resolver hasta cien ciudades en unos pocos minutos y llegar al millar en unas cuantas horas en un procesador normal de sobremesa.

La otra opción es conformarse con menos: una solución que no esté muy alejada de la óptima, pero que sea más fácil de encontrar. En algunos casos,

esto puede lograrse mediante el empleo de un descubrimiento sorprendente hecho en 1890 en una rama de las matemáticas tan novedosa que muchas de las figuras destacadas en la época no pensaban que tuviese valor alguno y, a menudo, no creían en las respuestas que investigadores más visionarios encontraban poco a poco. Lo que es peor, los problemas que estos últimos abordaban parecían ser «matemáticas por sí mismas», sin relación perceptible con ninguna otra cosa en el mundo real. En general, se consideraba que sus resultados eran muy artificiosos y las nuevas formas geométricas que construían se calificaban de «patológicas». Muchos pensaban que, incluso si esas conclusiones eran correctas, no aportaban ni un ápice a favor de la causa de las matemáticas y que solo ponían obstáculos insensatos al progreso, en una orgía autocomplaciente de veleidades lógicas.

*

Un método para encontrar buenas soluciones, aunque no sean óptimas, al TSP surgió a partir de uno de estos obstáculos insensatos. Durante unas pocas décadas en torno a 1900, las matemáticas experimentaron una transición. Se había agotado el espíritu intrépido inicial de avances temerarios que pasaban por alto los detalles incómodos y su despreocupación por problemas básicos tales como «¿de qué estamos hablando en realidad aquí?» o «¿de verdad es esto tan evidente como pensamos?» había sembrado confusión y perplejidad donde debía haber claridad y comprensión. Las preocupaciones en torno a áreas avanzadas como el cálculo, en la que los matemáticos habían diseminado procesos infinitos a diestro y siniestro con alegre indolencia, empezaban poco a poco a desandar el camino desde los misterios para iniciados hasta los lugares comunes. En lugar de albergar dudas acerca de las integrales de funciones matemáticas complicadas, como los logaritmos complejos, muchos se preguntaban qué era una función. En vez de definir una curva continua como «aquella que puede dibujarse libremente con la mano», se buscaba un rigor mayor, que se echaba en falta. Incluso la naturaleza de algo tan básico y evidente como un número se demostraba esquiva. No solo en el caso de desarrollos novedosos, tales como los complejos, sino también para los archiconocidos números naturales 1, 2, 3. Las matemáticas convencionales habían proseguido su avance bajo la asunción táctica de que los problemas de este tipo acabarían por resolverse y de que todo se arreglaría. El estatus lógico de los cimientos podía dejarse sin peligro en manos de quisquillosos y pedantes. Y aun así… empezó a cristalizar un

sentimiento generalizado de que este enfoque despreocupado de la disciplina no podía continuar mucho tiempo más.

Las cosas empezaron a torcerse de verdad cuando los audaces métodos anteriores empezaron a arrojar respuestas que se contradecían entre sí. Teoremas en cuya veracidad se confiaba desde hacía mucho resultaron ser falsos en circunstancias excepcionales, por lo general bastante extrañas. Una integral calculada de dos maneras diferentes daba dos resultados distintos. Una serie que se creía convergente para todos los valores de la variable divergía en ocasiones. No era tan catastrófico como descubrir que 2 + 2 a veces da 5, pero hizo que algunas personas se preguntasen qué eran en realidad 2 y 5, por no hablar de + e =.

De modo que, sin dejarse desanimar por una mayoría de agoreros (o al menos no tanto como para cambiar de opinión), unos pocos quisquillosos escarbaron túneles por todo el edificio de las matemáticas, desde las elevadas alturas hasta los sótanos, en busca de terreno firme sobre el que empezar a renovar por completo la construcción.

Como en todas las reformas, el resultado final fue diferente del punto de partida de una manera sutil pero inquietante. La noción de curva en el plano, que estaba presente desde la época de la Grecia antigua, reveló complejidades ocultas. Los ejemplos tradicionales (los círculos, las elipses y las parábolas de Euclides y Eratóstenes, la cuadratriz que los griegos empleaban para la trisección del ángulo y la cuadratura del círculo, la lemniscata en forma de ocho del neoplatónico Proclo, los óvalos de Giovanni Domenico Cassini, las cicloides y su progenie más compleja, tales como las hipocicloides y epicicloides de Ole Rømer) conservaban su propia fascinación y habían conducido a avances notables. Sin embargo, del mismo modo que los animales domesticados constituyen una imagen engañosa de la vida en las selvas tropicales y en las extensiones desérticas de la Tierra, estas curvas eran demasiado mansas como para ser representativas de las bestias salvajes que merodean por la jungla matemática. Como ejemplos de la complejidad potencial de las curvas continuas, eran demasiado sencillas y se comportaban demasiado bien.

Una de las características más fundamentales de las curvas, tan evidente que nadie se atrevía a ponerla en duda, es que son *delgadas*. Como escribió Euclides en sus *Elementos*, «una línea es una longitud sin anchura». El área de una línea (solo de esta, no de aquello que encierra) es, desde luego, cero. Sin embargo, en 1890, Giuseppe Peano describió una construcción de una curva continua que recubre por completo el interior de un cuadrado[25]. No se

limita a ir de un lado a otro dentro del área en un garabateo complicado que se acerca a cualquier punto, sino que atraviesa *todos* los puntos en la superficie de manera exacta. Desde luego, la curva de Peano «no tiene anchura», en el sentido de que se construye al trazar una línea con un lápiz cuyo extremo es un único punto geométrico, pero da vueltas de una manera muy enrevesada y pasa varias veces por las regiones que ya ha visitado antes. Peano se dio cuenta de que si se hace ondulada hasta lo infinito, de forma controlada y cuidadosa, acabará por recubrir toda la superficie. En particular, el área de la curva es la misma que la del cuadrado, de modo que no es cero.

Este descubrimiento supuso una conmoción de la intuición inocente. En esa época, las curvas de este tipo se calificaron de «patológicas» y muchos matemáticos reaccionaron ante ellas del modo en que habitualmente se hace ante lo enfermizo: con miedo y asco. Más adelante, el gremio se acostumbró a ellas y asumió las profundas lecciones topológicas que imparten. Hoy en día, la curva de Peano se entiende como un ejemplo temprano de geometría de fractales y se comprende que estos no son en modo alguno inusuales ni patológicos. Son lugares comunes, incluso en las matemáticas, y en el mundo real proporcionan modelos excelentes de estructuras naturales muy complejas, tales como nubes, montañas y costas.

Los pioneros de esta nueva era de las matemáticas escrutaron los conceptos intuitivos antiguos, como continuidad y dimensión, y empezaron a plantear preguntas de difícil respuesta. En lugar de asumir que podían salirse con la suya con las triquiñuelas tradicionales empleadas en ramas más sencillas de la disciplina, estos precursores se preguntaron si esos trucos funcionaban con una generalidad suficiente y, de ser así, *por qué* lo hacían. O, si no sucedía tal cosa, qué era lo que fallaba. Este enfoque escéptico molestó a muchos matemáticos convencionales, que lo percibieron como una actitud negativa sin razón de ser. En 1893, Charles Hermite escribió a su amigo Thomas Stieltjes: «Me aparto con terror y horror de este lamentable flagelo de funciones continuas sin derivaciones».

Los tradicionalistas estaban mucho más interesados en ampliar las fronteras y en asumir que todo en el jardín lógico era encantador. Sin embargo, el nuevo escepticismo, con su vendaval de desafíos estrafalarios a la intuición, era una reacción necesaria frente a la inocencia. Al llegar la década de 1930, el valor de este enfoque más riguroso se hacía evidente. Treinta años después era dominante casi por completo. Podría escribirse todo un libro acerca de este periodo en el desarrollo de nuestra disciplina y eso es lo que

han hecho algunos autores. Aquí voy a centrarme en un tema derivado: las curvas continuas y el concepto de dimensión.

*

Es probable que el concepto de curva se remonte al momento en el que algún humano primitivo arrastró por primera vez la punta de un palo a través de una superficie de arena o barro y descubrió que dejaba una marca. Empezó a adquirir su forma actual cuando la manera lógica de abordar la geometría echó a volar en la antigua Grecia y Euclides afirmó que un punto tiene solo posición y que una línea carece de anchura. Una curva es una línea que no necesita ser recta. El ejemplo más sencillo es una circunferencia o un arco de la misma. Los griegos desarrollaron y analizaron varias de ellas, las ya mencionadas: elipse, cuadratriz, cicloide, etcétera. Solo tomaron en consideración ejemplos concretos, pero la idea general era «más o menos evidente».

Tras la introducción del cálculo, cobraron relevancia dos propiedades de las curvas. Una de ellas era la continuidad: una curva es continua si no tiene interrupciones. La otra, menos evidente, era la suavidad: una curva es suave si no tiene picos abruptos. El cálculo integral funciona bien para las curvas continuas y el diferencial lo hace para las suaves (esta descripción es *muy* chapucera para no interrumpir el ritmo de la narración, pero puedo garantizar que se aproxima más a la verdad que a las noticias falsas). Por supuesto, no era tan sencillo: había que definir «interrupción» y «pico» y hacerlo *con precisión*. De manera más sutil, cualquier definición propuesta tenía que ser adecuada para el estudio matemático y estar planteada en los términos de la disciplina. Había que poder *emplearla*. Los detalles todavía desconciertan a los estudiantes universitarios la primera vez que se topan con ellos, así que los voy a evitar aquí.

El segundo concepto fundamental es el de dimensión. Todos hemos aprendido que el espacio tiene tres dimensiones, el plano, dos y la línea, una. No se llega a esta noción a partir de una definición del término «dimensión» y un posterior recuento de cuántas hay en el espacio o en un plano. No exactamente. En su lugar, decimos que el espacio tiene tres porque es posible indicar cualquier posición con el empleo de solo tres números. Se designa algún punto concreto como origen y se eligen tres direcciones: de norte a sur, de este a oeste y de arriba abajo. Después solo hay que medir la distancia desde el origen hasta la posición a indicar en cada una de estas direcciones.

Esto da tres números (las *coordenadas* relativas a esas direcciones elegidas) y cada posición en el espacio se corresponde con uno, y solo uno, de estos tríos de números. De manera similar, un plano tiene dos dimensiones porque es posible ahorrarse una de estas coordenadas (digamos la de arriba abajo) y una línea tiene una dimensión.

Todo parece bastante sencillo hasta que se empieza a reflexionar sobre ello. El párrafo anterior asume que el plano en cuestión es horizontal. Por eso es posible prescindir de la coordenada arriba abajo. Pero ¿qué pasa si está en una pendiente? Entonces la dirección perpendicular cobra importancia. No obstante, resulta que el número que indica la vertical viene siempre determinado por los otros dos (siempre que se sepa la inclinación de la pendiente). De modo que lo importante no es la cantidad de direcciones a lo largo de las cuales se midan las coordenadas, sino el número de ellas que son *independientes*. Es decir, que no son combinaciones de otras direcciones.

Ahora todo se complica un poco más porque no basta con contar cuántas coordenadas hay. Se trata más bien de determinar el menor número que hay que utilizar de ellas. Y esto plantea otra pregunta más profunda: ¿cómo se sabe que dos *es* de verdad la cantidad más pequeña que hay que emplear para un plano? Podría ser verdad (y si no lo es, se necesita una definición mejor), pero no es evidente. Con esto se abre la veda. ¿Cómo se sabe que tres es la cantidad más pequeña que hay que emplear para el espacio? ¿Cómo se sabe que *cualquier* elección de direcciones independientes siempre da tres números? Por lo que a esto respecta, ¿hasta qué punto se tiene la certeza de que tres coordenadas son suficientes?

Esta última pregunta es en realidad competencia de la física experimental y conduce, a través de Einstein y su teoría general de la relatividad, a inferir que el espacio físico no es, de hecho, el espacio plano tridimensional de Euclides, sino una versión curva. O, si la teoría de cuerdas está en lo cierto, que es un espacio-tiempo que tiene diez u once dimensiones, todas las cuales, excepto cuatro, son demasiado pequeñas para percibirlas o son inaccesibles. La primera y la segunda pregunta pueden resolverse de manera satisfactoria, aunque no sean triviales, con solo definir el espacio euclídeo tridimensional en términos de un sistema de coordenadas con tres números y pasando luego de cinco a seis semanas en un curso universitario de espacios vectoriales, en los que es posible cualquier número de coordenadas, para demostrar que las dimensiones de un espacio vectorial son únicas.

Un aspecto inherente al enfoque del espacio vectorial es la noción de que un sistema de coordenadas se basa en líneas rectas y que el espacio es plano.

De hecho, se conoce también como «álgebra lineal». ¿Qué pasa si nos marcamos un Einstein y permitimos que el sistema de coordenadas no sea recto? Bueno, si se dobla de una manera suave (lo que se conoce tradicionalmente como «coordenadas curvilíneas») no hay problema. Pero en 1890, el matemático italiano Giuseppe Peano descubrió que si se tuerce de una manera brusca (tanto que ya no sea suave, aunque no deje de ser continuo), entonces un espacio de dos dimensiones puede tener un sistema de coordenadas con *un* solo número. Lo mismo vale si es tridimensional. En esta configuración más general y flexible, de repente «el» número de dimensiones se hace cambiante.

Una manera de reaccionar ante este extraño descubrimiento es desdeñarlo. No hay duda de que se deben emplear coordenadas suaves, o algo así. Pero resultó que era mucho más útil y creativo, y también más *divertido*, aceptar lo inaudito y ver lo que podía ocurrir. Los críticos tradicionalistas eran bastante puritanos y no querían que las nuevas generaciones se lo pasasen bien.

*

Vamos al grano. Lo que Peano había descubierto (o construido) era una curva continua que pasaba a través de cada punto en un cuadrado. No solo en su perímetro. Eso es fácil. También por todos los que hay en su interior. Y debe pasar de manera exacta por cada uno de ellos, no solo aproximarse mucho.

Supongamos que existe una curva así. Entonces no es más que algún tipo de línea ondulante, que tiene su propio sistema de coordenadas intrínseco (qué distancia se debe recorrer sobre ella). Eso es un solo número, de modo que la curva es unidimensional. No obstante, si pasa a través de todos los puntos de una superficie cuadrada, que tiene dos dimensiones, se habrá conseguido especificar cada punto de ese cuadrado con una única coordenada que varía de manera continua. ¡De modo que la superficie es en realidad unidimensional!

Intento evitar por norma los signos de exclamación cuando escribo, pero este descubrimiento los merece. Es una locura. También es verdad.

Peano había encontrado el primer ejemplo de lo que ahora se conoce como una curva que «recubre el espacio». Su existencia se basa en la distinción, sutil pero crucial, entre curvas suaves y continuas. Estas últimas pueden ser ondulantes. Las suaves… no. Por lo menos no *así* de ondulantes.

Peano tenía la actitud adecuada para construir su curva. Le gustaban las disquisiciones sobre detalles lógicos. También fue el primero en escribir

axiomas precisos para el sistema de los números naturales (una lista sencilla de propiedades que especifican ese sistema con precisión). No construyó su curva solo por diversión: daba los toques finales a la obra de otro predecesor de inclinaciones parecidas, que también tenía un profundo interés en la naturaleza de los números naturales y en lo que significa contar. Su nombre era Georg Cantor y lo que le interesaba de verdad era el infinito. La mayoría de los matemáticos destacados de la época repudiaban las ideas radicales y brillantes de Cantor, lo que le llevó a la desesperación. Es probable que el rechazo no fuese la causa de su posterior enfermedad mental, como se ha sugerido a veces, pero no cabe duda de que no le hizo ningún bien. Entre los pocos matemáticos reconocidos que supieron apreciar lo que Cantor intentaba hacer se encontraba uno que escaló las cumbres de la disciplina hasta sus cimas más elevadas: David Hilbert. Puede que Hilbert fuera el matemático más destacado de su época y, más tarde en su vida, se convirtió en uno de esos pioneros de la lógica matemática y de los fundamentos de la disciplina. Tal vez reconoció en Cantor a un alma gemela.

En cualquier caso, todo empezó con Cantor y con su introducción de los cardinales transfinitos (cómo contar cuántos elementos hay en un conjunto infinito). Su demostración de que algunos infinitos son mayores que otros es famosa. Más en concreto, no existe una correspondencia biyectiva entre los números enteros y los reales. Buscaba un cardinal transfinito que fuese mayor que el de estos últimos y durante un tiempo estuvo convencido de que el cardinal del plano debía ser más grande que el de la línea. En 1874, escribió a Richard Dedekind:

> ¿Es posible hacer corresponder unívocamente una superficie (digamos un cuadrado incluyendo su frontera) con una línea (digamos un segmento de recta incluyendo sus puntos extremos), de manera unívoca tal que a cada punto de la superficie le corresponda un punto de la línea, e inversamente a cada punto de la línea un punto de la superficie?
>
> En este momento tengo la impresión de que la respuesta a esta pregunta (si bien uno se ve aquí tan inclinado al *no*, que podría parecerle casi superflua la demostración) ofrece graves dificultades.

Tres años más tarde escribió de nuevo para decir que estaba equivocado. Muy equivocado. Había encontrado una correspondencia biyectiva entre el intervalo unidad y el espacio de n dimensiones para cualquier n finito. Es decir, una manera de hacer corresponder los miembros de ambos conjuntos

entre sí de forma que cada elemento de uno de ellos se correspondiera exactamente con uno solo del otro conjunto. «Lo veo», escribió Cantor, «pero no lo creo».

La idea principal es sencilla: dados dos puntos en el intervalo unidad (entre 0 y 1) es posible escribirlos en decimales como:

$$x = 0{,}x_1x_2x_3x_4\ldots$$

$$y = 0{,}y_1y_2y_3y_4\ldots$$

y hacer que esto corresponda a un punto en el intervalo unidad cuya expansión decimal sea

$$0{,}x_1y_1x_2y_2x_3y_3x_4y_4\ldots$$

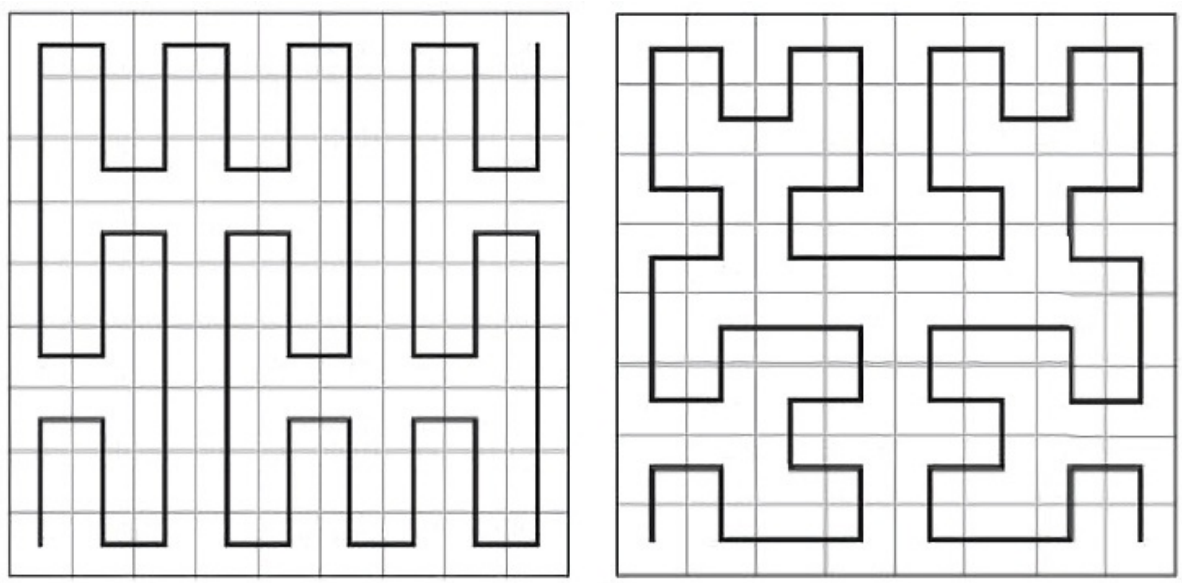
al intercalar las posiciones decimales, del mismo modo que se hace con los naipes de las dos mitades de un mazo de cartas[26]. La principal diferencia es que la baraja de Cantor es infinita. Ahora bien, cuando se mezclan entre sí dos mazos infinitos se obtiene *una* baraja infinita. Así es como Cantor consigue reunir dos coordenadas en una. Para gestionar tres dimensiones no hay más que usar tres barajas, etcétera.

Cantor publicó algunos de estos resultados en 1878. Investigó los conjuntos numerables, que pueden ponerse en correspondencia biyectiva con los números naturales, así como aquellos conjuntos que tienen este tipo de relación entre sí. También se dio cuenta de que su correspondencia entre el intervalo unidad y el cuadrado unidad no conserva las dimensiones (se pasa de dos a una) y, lo que es más importante para nuestra historia, insistió en que la correspondencia que había construido no es continua. Es decir, puntos muy próximos entre sí en el intervalo unidad no tienen por qué corresponder a puntos muy próximos en el cuadrado unidad.

Las ideas de Cantor eran muy controvertidas y algunos matemáticos eminentes las consideraron un sinsentido, con toda probabilidad porque eran tan originales que requerían imaginación y una mente abierta para apreciarlas. Otros, sobre todo Hilbert, declararon que la nueva rama de la disciplina que había inaugurado Cantor era un «paraíso». El reconocimiento completo de la importancia de su obra solo llegó tras su fallecimiento.

*

En 1879, Eugen Netto[27] dio respuesta a una pregunta evidente al demostrar que no existe una correspondencia biyectiva *continua* entre el intervalo unidad y la superficie cuadrada unidad, algo que es más difícil de lo que parece a primera vista. El logro más significativo se produjo en 1890, cuando Peano soltó el zorro en el gallinero con su curva que recubría el espacio y demostró que nuestra imagen mental por defecto de una curva continua puede ser engañosa en extremo.



Izquierda: una etapa temprana en una interpretación geométrica de la curva que recubre el espacio de Peano. *Derecha*: una etapa temprana en la construcción de la curva que recubre el espacio de Hilbert.

No hay ilustraciones en el artículo de Peano. Define la curva mediante el empleo de expansiones ternarias de puntos en el intervalo unidad y su construcción es similar a la geometría que se muestra, arriba, en la imagen de la izquierda[28]. En 1891, Hilbert publicó otro ejemplo de curva que recubre el espacio, con un trazo similar a la figura de la derecha. Ambas construcciones son bastante complejas. Las ilustraciones muestran etapas tempranas de un proceso iterativo en el que se sustituyen de manera repetida polígonos sencillos por otros más complicados. Desde entonces se han descubierto muchas otras curvas que recubren el espacio.

Este tipo de curvas tienen aplicaciones en informática, tales como el almacenamiento y la recuperación de datos multidimensionales[29]. La idea básica es que es posible recorrer una matriz de varias dimensiones al seguir una aproximación a una curva que recubre el espacio, lo que reduce los problemas al caso unidimensional. Otra aplicación proporciona una solución rápida, aunque no sea exacta, al problema de la persona viajante. La idea es ejecutar una aproximación finita a una curva que recubre el espacio a través

de la región que contiene las ciudades, ordenar estas a lo largo de la línea y después visitarlas en ese orden mediante la trayectoria de menor longitud que las une en cada paso. Con esto se obtiene un recorrido que no suele ser más de un 25 % mayor que el caso óptimo[30].

¿Qué otras formas pueden recubrir una curva? Es posible extender la construcción de Hilbert a tres dimensiones y obtener una curva que recubra el cubo unidad, mientras que otras también pueden recubrir hipercubos de una cantidad cualquiera de dimensiones. El último grito es un teorema demostrado por Hans Hahn y Stefan Mazurkiewicz, que caracteriza por completo los espacios topológicos que puede recubrir una curva[31]. Resulta que son casi cualesquiera, siempre que sean compactos (con extensión finita) y cumplan unas cuantas condiciones técnicas que descartan espacios inverosímiles.

*

La persona viajante puede tener todavía la última palabra. En 1992, Sanjeev Arora y colaboradores[32] descubrieron que la clase de complejidad NP («comprobable con facilidad») tiene una curiosa propiedad que arroja dudas sobre cualquier expectativa de hallar algoritmos de clase P («calculable con facilidad») que den buenas soluciones aproximadas. Demostraron que si $P \neq NP$ y el tamaño del problema es superior a cierto umbral, entonces calcular una buena aproximación a la respuesta no es más fácil que encontrar esta en sí misma. La única alternativa a esta conclusión sería que $P = NP$, lo que valdría un millón de dólares, pero de momento tiene que permanecer como hipótesis.

Su aportación está relacionada con una idea notable de verdad: las demostraciones holográficas. Las demostraciones son la esencia de las auténticas matemáticas. En la mayoría de las ramas de la ciencia las teorías pueden contrastarse con la realidad, mediante observaciones o experimentos. Ese lujo no está disponible en matemáticas, aunque todavía hay una manera de verificar sus resultados. En primer lugar, estos deben venir respaldados por una demostración lógica. Después, hay que comprobar esta para garantizar que no contiene errores y que no se ha pasado nada por alto. Es difícil alcanzar este ideal y, de hecho, no es lo que hacen los investigadores en realidad, aunque ese sea su objetivo. Todo aquello que no supera esta prueba se califica de manera inmediata como «erróneo», aunque todavía puede ser útil como paso hacia una demostración mejor que sea correcta. De modo que, desde los tiempos de Euclides hasta la actualidad, los matemáticos han

gastado mucho tiempo en repasar con cuidado las demostraciones, tanto las propias como las de los demás, renglón a renglón, en busca de elementos con los que estuviesen de acuerdo y de otros para los que no les salieran las cuentas.

En los últimos años ha surgido una manera diferente de hacer las comprobaciones: emplear ordenadores. Esto exige volver a escribir las demostraciones en un lenguaje que los ordenadores puedan procesar mediante algoritmos. Funciona y ha conseguido éxitos espectaculares con algunos de los problemas más difíciles que se conocen, aunque hasta el momento no ha desplazado a otros métodos más tradicionales. Un efecto secundario de esta idea es un enfoque renovado en la forma de presentar las demostraciones de manera asequible para los ordenadores, que a menudo es diferente por completo de cualquier cosa que pueda digerir un humano. Los procesadores no se quejan si se les dice que repitan la misma operación millones de veces ni que comparen dos filas de miles de dígitos binarios para asegurarse de que son idénticas. Se ponen manos a la obra sin más.

A los matemáticos humanos les gustan más las demostraciones que cuentan una historia, con principio, trama y final evidentes y con una narración atractiva que los lleve desde el punto de partida, las hipótesis del teorema, hasta su conclusión. El relato es más importante que las disquisiciones lógicas. El objetivo es ser claro, conciso y por encima de todo, *convincente*. Hay que tener en cuenta que la dificultad de convencer a un matemático es notoria.

A los informáticos teóricos que estudian las demostraciones comprobables por ordenador se les ocurrió un enfoque totalmente diferente: las demostraciones interactivas. En lugar de presentarlas como un relato, escritas por un investigador y leídas por otro, esta nueva perspectiva las convierte en un enfrentamiento. Un matemático, con el tradicional nombre de Pat, quiere convencer a Vanna de que su demostración es correcta y Vanna quiere persuadirle de que no lo es. Intercambian preguntas y respuestas entre sí hasta que uno de ellos admite la derrota (Pat Sajak y Vanna White eran personajes famosos en la televisión estadounidense que salían en el programa *La ruleta de la fortuna*). Es como una partida de ajedrez, en la que Pat anuncia «jaque mate en cuatro jugadas». Vanna no está de acuerdo, así que Pat hace un movimiento. Vanna replica con un «¿qué pasa si hago esto?» y Pat mueve de nuevo. Este tira y afloja sigue hasta que Vanna pierde. Ahora ella empieza a dar marcha atrás. «Supongamos que mi última jugada hubiese sido *esta* en lugar de la anterior». Pat hace un movimiento diferente y ¡jaque mate! Se

continúa así hasta que se agotan todas las posibles respuestas de Vanna ante las jugadas de Pat y este gana, o bien hasta que él se ve obligado a admitir que en realidad no era jaque mate en cuatro jugadas. Según mi experiencia, justo esto es lo que hacen los matemáticos en la vida real cuando trabajan juntos para resolver un problema en una investigación y los ánimos pueden llegar a calentarse. La versión en forma de relato de la solución a la que se llega en última instancia es la que se presenta en un seminario.

Laszlo Babai y otros se valieron de este tipo de técnica de demostración argumentativa para llegar a la noción de demostración holográfica mediante el empleo de herramientas matemáticas tales como polinomios sobre campos finitos y códigos de corrección de errores[33]. Una vez establecidos estos métodos, se dieron cuenta de que los ordenadores pueden aprovechar una característica que se evita en aras de la claridad y la concisión: la redundancia. Resulta que una demostración lógica puede volverse a escribir en una forma que la hace mucho más larga, pero que implica también que si hay un error aparece casi por todas partes. Cada paso del argumento se extiende por toda la demostración en muchas copias relacionadas casi idénticas. En cierto modo es como un holograma, en el que se transforma una imagen de manera que puede reconstruirse a partir de cualquier porción pequeña de los datos. En esos casos, la demostración puede comprobarse con solo tomar una pequeña muestra aleatoria. Cualquier error aparecerá casi con total seguridad en ella. Si se hace así, se tiene una demostración holográfica. El teorema acerca de la inexistencia de soluciones aproximadas de clase P es una consecuencia.

*

Volvamos al artículo de la paloma de Gibson, Wilkinson y Kelly en *Animal Cognition*. Parten de la afirmación de que el TSP se había empleado hacía poco para examinar aspectos del conocimiento humano y animal, sobre todo de la capacidad de planificar acciones antes de emprenderlas. No obstante, no estaba claro si esta habilidad estaba limitada a los primates. ¿Pueden hacerlo también otros animales o solo siguen reglas rígidas, desarrolladas al evolucionar? Los investigadores decidieron emplear palomas en pruebas de laboratorio que las situaban ante TSP sencillos que tenían dos o tres comederos de destino. Las aves empezaban en una ubicación, visitaban cada comedero en algún orden y continuaban hasta un destino final. El equipo concluyó que «las palomas sopesaban con detenimiento la proximidad de la

siguiente ubicación, pero parecían planificar varios pasos por adelantado cuando aumentaba el coste del desplazamiento a causa del comportamiento ineficaz. Los resultados proporcionan una evidencia clara y firme de que los animales, aparte de los primates, son capaces de planificar recorridos de desplazamiento sofisticados».

En una entrevista, los investigadores explicaban la relación con la paloma conductora de autobús. Sugerían que el humano podría tener dos razones para poner pegas: la evidente de la seguridad y la preocupación de que el ave fuera incapaz de seguir una ruta para recoger a los pasajeros de manera eficaz conforme el autobús atravesaba la ciudad. Como afirma el título del artículo, el equipo llegaba a la conclusión a partir de sus experimentos de que la segunda preocupación no estaba justificada.

¡Deja que la paloma conduzca el autobús!

*

Si los gobiernos mundiales y los fabricantes de coches se salen con la suya, pronto ni el humano ni la paloma conducirán nada. Será el autobús el que conduzca el autobús en su lugar. Nos adentramos en una era feliz de vehículos autónomos.

O tal vez no.

El aspecto más difícil de lograr en los vehículos sin conductor es garantizar que interpretan su entorno de manera correcta. Dotarles de sus propios «ojos» es fácil, porque en la actualidad se fabrican miles de millones de pequeñas cámaras de alta resolución. Pero la visión necesita un cerebro tanto como unos ojos, de modo que coches, camiones y autobuses se equipan con *software* de visión por ordenador. Así pueden saber qué es lo que ven y reaccionar de manera acorde.

Según los fabricantes, una ventaja potencial de los vehículos autónomos es la seguridad. Los conductores humanos cometen errores y causan accidentes. Un procesador no se distrae y, con bastante investigación y desarrollo, un ordenador debería ser más seguro al volante que cualquier persona. Otro aspecto positivo es que no hay que pagar al autobús para que se conduzca a sí mismo. Una gran desventaja, aparte de hacer que los conductores pierdan su empleo, es que esta tecnología se halla todavía en pañales y que los sistemas disponibles en la actualidad no están a la altura de las expectativas que generan. Ya han fallecido algunos peatones y conductores de prueba en accidentes y, sin embargo, en la actualidad se

siguen ensayando vehículos completamente autónomos en las calles de las ciudades de varios países. El argumento es que deben probarse en el mundo real y que, en última instancia, salvarán más vidas de las que pueden costar. La facilidad con la que los reguladores se han dejado convencer por este seductor razonamiento es notable. Si alguien sugiriese probar un nuevo medicamento en pacientes aleatorios, sin su conocimiento ni consentimiento, alegando que esto salvará a más personas de las que va a matar, se armaría un escándalo. De hecho, sería ilegal en casi todos los países y es sin duda contrario a la ética.

A este respecto, la tecnología principal tras la visión por ordenador es el campo, aún más de moda, del aprendizaje autónomo. Una red de aprendizaje profundo, que ajusta sus fuerzas de conexión de modo que identifica figuras de manera correcta, se entrena con una cantidad enorme de imágenes hasta que alcanza un nivel aceptable de precisión. Este procedimiento ha tenido un éxito enorme en un amplio rango de aplicaciones. No obstante, en 2013 se hizo evidente que se prestaba demasiada atención a los éxitos del aprendizaje autónomo y muy poca a sus fallos potenciales. Un problema grave son los «ejemplos contradictorios», que son representaciones modificadas de manera deliberada para que un humano acierte mientras que un ordenador comete fallos espectaculares.

Estas dos imágenes, que difieren solo en unos pocos píxeles, se mostraron a la red Inception V3, que clasificó la de la izquierda como un gato y la de la derecha como guacamole.

La fotografía muestra dos imágenes de un gato. Obvio. Solo se diferencian en unos pocos píxeles y son idénticas a nuestros ojos. Una red neuronal estándar, entrenada con cantidades enormes de figuras de gatos y de lo que no son

gatos, identifica de manera correcta la imagen de la izquierda como tal. Sin embargo, insiste en que la de la derecha es guacamole, la salsa mexicana verde hecha de aguacate. Es más, alcanza una certeza del 99 % de que es guacamole, en comparación con solo un 88 % para el gato. Como dice el dicho, un ordenador es un aparato para cometer millones de errores muy rápido.

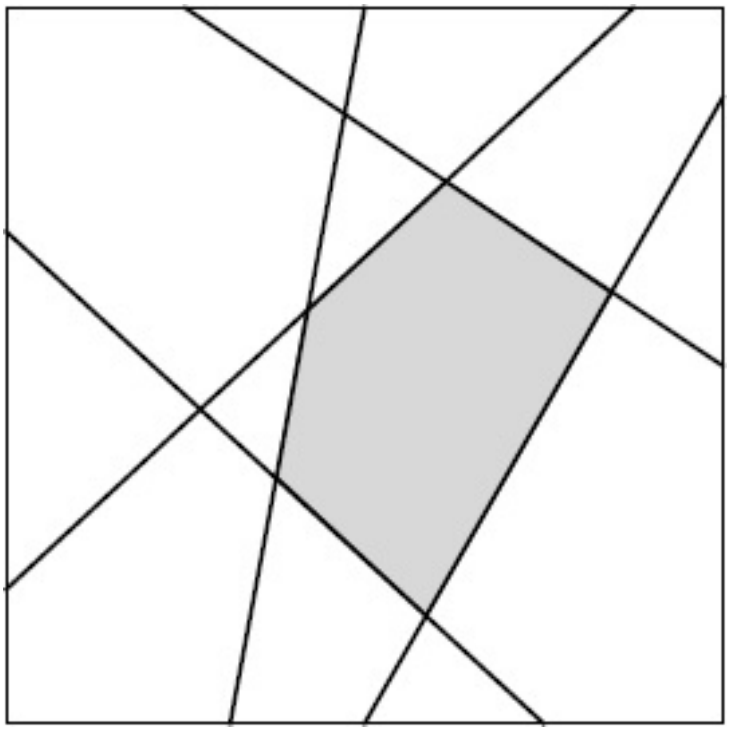
Se dice que imágenes de este tipo son «contradictorias» porque surgen cuando alguien intenta engañar al sistema de manera deliberada. En la práctica, la red identificará un gato en la mayoría de las fotografías como estas. Christian Szegedy y sus colaboradores se dieron cuenta en 2013 de que existían estas imágenes[34]. En 2018, Adi Shamir y sus colegas[35] explicaron por qué es posible producir ejemplos contradictorios en sistemas de aprendizaje profundo, por qué son inevitables y por qué basta con cambiar unos pocos píxeles para despistar a la red neuronal.

La causa fundamental de esta susceptibilidad a errores graves es la extensión. La manera habitual de medir cómo son de diferentes dos cadenas de código binario es calcular su distancia de Hamming, el número de bits que es necesario cambiar para convertir una en la otra. Por ejemplo, este parámetro para 10001101001 y 10101001111 es cuatro, con los bits diferentes resaltados en negrita en 10**1**01**0**01**11**1. Una fotografía está representada en el ordenador por una cadena binaria muy larga. Si la imagen ocupa 1 MB (megabyte), su longitud es 2^{23}, o cerca de 8 millones de bits. De modo que el espacio de las imágenes tiene una extensión de 8 millones sobre el campo finito que consiste en 0 y 1. Contiene $2^{8\,388\,608}$ puntos diferentes.

El algoritmo de reconocimiento de figuras incorporado a una red neuronal entrenada tiene que situar cada imagen de este espacio en un número mucho más pequeño de categorías. En el caso más sencillo, esto se reduce a dividir el espacio de las imágenes en regiones mediante la creación de hiperplanos, un procedimiento que se ilustra en la fotografía para dos dimensiones. Esto reparte el espacio en numerosas celdas, una para cada categoría. Si se cambia la imagen a una que esté a una distancia de Hamming de, pongamos, 40, entonces solo hay que sustituir 40 elementos en la cadena. El ojo percibe 8 millones de bits, por lo que la diferencia supone apenas el 0,0005 % del total, muy por debajo del umbral a partir del que un humano notaría alguna diferencia significativa. No obstante, el número de imágenes dentro de esta distancia de Hamming es de 2^{50}, en torno a los mil billones. Esta cantidad es mucho mayor que el total de las categorías que puede distinguir el sistema de

visión por ordenador. De modo que no resulta sorprendente que un cambio tan pequeño en la imagen pueda llevar a la red a malinterpretarla.

Para el análisis matemático es conveniente representar las cadenas de bits no sobre un campo finito sino como números reales. Por ejemplo, un solo byte de 8 bits, digamos 10001101, puede considerarse como el número real con expansión binaria 0,10001101. Ahora el espacio de todas las imágenes de 1 MB se convierte en un espacio vectorial real de extensión un millón. Con esta modificación, Shamir y sus colegas demostraron algo mucho más determinante. Dada una fotografía en una celda de la disposición de hiperplanos y dada una segunda celda, ¿cuántos bits hay que cambiar en la imagen para llevarla hasta esta última? Su análisis demuestra que, por ejemplo, si el espacio de imágenes está dividido en un millón de categorías mediante 20 hiperplanos, entonces solo hay que cambiar dos coordenadas para mover un punto dado a cualquier celda, siempre que la extensión del espacio de imágenes sea mayor que 250. En general, si se ha entrenado a la red para que distinga un número dado de categorías, la cantidad de coordenadas que debe cambiarse para mover una imagen dada a *cualquier* celda es más o menos el mismo que el número de estas.



División del espacio de la imagen por hiperplanos. En este caso, hay dos dimensiones y cinco hiperplanos (aquí, rectas) dividen el espacio en trece celdas. Una se muestra sombreada.

Comprobaron su teorema en un sistema comercial de reconocimiento de números. En este caso hay diez categorías, que son los dígitos de 0 a 9.

Produjeron imágenes contradictorias para persuadir al sistema de que reconociese el 7 como cualquiera de las diez posibilidades, de 0 a 9. Solo había que cambiar 11 bits para lograrlo y lo mismo vale para cualquier dígito que no sea el 7.

¿Deberíamos preocuparnos? Las imágenes «naturales», del tipo que encontrarán de manera habitual los vehículos autónomos, no se construyen de forma deliberada para engañar al sistema. No obstante, el coche percibirá a su alrededor en torno al medio millón de imágenes al día y basta con una interpretación errónea para que se produzca un accidente. La amenaza principal es que gamberros o terroristas puedan modificar fácilmente las señales de tráfico con solo añadir pequeños trozos de cinta negra o blanca y engañar así al ordenador para que piense que una señal de stop es en realidad un límite de velocidad de 90 kilómetros por hora. Todo lo cual contribuye a la impresión de que la introducción de los coches autónomos se hace de manera apresurada, indebida e insegura a causa de presiones comerciales. Si alguien no está de acuerdo, me permito repetir: *nunca* se introduciría un fármaco o procedimiento médico nuevos de forma tan chapucera. Sobre todo si hay buenas razones para sospechar que puede ser peligroso.

¡No dejes que el autobús conduzca el autobús!

4
Los riñones de Königsberg

> Agregada a esa rama de la geometría que se relaciona con magnitudes existe otra rama, la cual Leibniz mencionó primero llamándola *geometría de la posición*… Por esto, cuando se mencionó un problema que parecía geométrico pero que no requería la medición de distancias, yo no tuve dudas de que este problema estaba relacionado con la geometría de la posición. Por tanto, he decidido dar aquí el método que he encontrado para resolver esta clase de problemas.
>
> LEONHARD EULER,
> *Solutio problematis ad geometriam situs pertinentis*, 1736

A lo largo de la mayor parte en la historia de la humanidad, los órganos con los que se nacía eran aquellos con los que se fallecía y, a menudo, también *de* los que se moría. Si dejaba de funcionar el corazón, o el hígado, o los pulmones, o los intestinos, o el estómago, o los riñones, también lo hacía la persona. Podían amputarse unas pocas partes del cuerpo, sobre todo brazos y piernas, y si el paciente sobrevivía a la experiencia, podía llevar algún tipo de vida. El descubrimiento de la anestesia y de la esterilización de los quirófanos hizo que las operaciones fuesen menos dolorosas, al menos mientras se realizaban y el paciente estaba inconsciente, y aumentó mucho las posibilidades de sobrevivir. Con la aparición de los antibióticos, a menudo era posible curar infecciones que antes eran mortales.

Estos milagros de la medicina moderna suelen darse por sentado, pero hicieron posible, por primera vez, que doctores y cirujanos *curasen* enfermedades. Se ha conseguido echar a perder la mayoría de las ventajas de los antibióticos, al dárselos en grandes cantidades a los animales de granja. No para curar enfermedades como tal, sino para hacerles engordar más y más rápido. También por los millones de personas que dejan de tomar estos medicamentos en cuanto se sienten mejor, en lugar de completar el tratamiento como les indicó el doctor. Ambas prácticas son innecesarias por completo y han fomentado el desarrollo de la resistencia bacteriana ante estos fármacos. Ahora los científicos investigan de manera frenética para

desarrollar la próxima generación de antibióticos. Si lo consiguen, espero que se tenga el sentido común de no echarlos a perder también.

De la misma manera, se ha hecho realidad otra panacea de los cirujanos de épocas anteriores: los trasplantes de órganos. Hasta el momento, parece que eso no se ha conseguido arruinar todavía. Si las circunstancias son favorables, es posible obtener un corazón nuevo, o un pulmón nuevo, o un riñón nuevo. Incluso una cara nueva. Puede que algún día haya cerdos bondadosos que produzcan órganos de repuesto, aunque no sea de manera voluntaria.

En 1907, el investigador médico estadounidense Simon Flexner especulaba sobre el futuro de la medicina y sugería que iba a ser posible sustituir órganos enfermos mediante cirugía por otros sanos de otra persona. En concreto, mencionaba las arterias, el corazón, el estómago y los riñones. El primer trasplante de riñón lo realizó en 1933 el cirujano ucranio Yuri Vorony, quien extrajo el órgano de un donante que había fallecido seis horas antes y lo implantó en el muslo de su paciente. Este murió dos días después porque no era del grupo sanguíneo correcto y rechazó el trasplante. El mayor obstáculo para que el procedimiento salga bien es el sistema inmune, que reconoce que el nuevo órgano no es parte del cuerpo del paciente y lo ataca. El primer caso con éxito fue el realizado por Richard Lawler en 1950. El órgano donado duró diez meses antes de ser rechazado, pero para ese momento los propios riñones de la paciente ya se habían recuperado lo suficiente como para que pudiese vivir cinco años más.

Una persona normal tiene dos riñones, y se las apaña muy bien con solo uno de ellos, de modo que pueden realizarse operaciones a partir de donantes vivos, lo que simplifica mucho el proceso. Se trata del órgano más fácil de trasplantar. Es sencillo asegurarse de que el tipo de tejido del donante es compatible con el del paciente, con lo que se impide el rechazo, y si algo sale mal hay máquinas de diálisis disponibles para realizar las tareas del riñón. Hasta que se empezaron a emplear medicamentos para frenar el rechazo, lo que ocurrió a partir de 1964, no había trasplantes de riñones de personas fallecidas (al menos en Estados Unidos y en Reino Unido). Pero había muchos donados por voluntarios vivos.

En la mayoría de los casos, estos eran familiares cercanos del receptor. De este modo aumentaba la posibilidad de que los tejidos fuesen compatibles, pero la razón principal era que muy pocas personas estaban dispuestas a sacrificar un riñón por un desconocido. Después de todo, cuando se tiene uno de repuesto es posible llevar una vida normal si el otro deja de funcionar. Una

posibilidad que se pierde si ya se ha donado a alguien anónimo. En cambio, si lo va a recibir la madre, el hermano o la hija, las ventajas superan a los inconvenientes, sobre todo si se sabe que va a morir en caso de negarse. No es algo tan personal cuando se trata de un desconocido y es menos probable que se acepte el riesgo.

Algunos países ofrecían un incentivo: dinero. Se podía retribuir a alguien anónimo para que donase un riñón al familiar de quien hacía el pago. El peligro de permitir este tipo de transacciones es bastante evidente: sobornar a personas pobres para que entreguen un órgano a un rico desconocido, por ejemplo. En Reino Unido, se prohibió donar un riñón a cualquiera que no fuese un familiar cercano. En 2004 y 2006, se aprobaron leyes para eliminar este obstáculo, pero se añadieron garantías para impedir los abusos. Una de ellas es que no puede haber intercambio monetario.

La modificación de la ley permitió que surgiesen nuevas estrategias para emparejar donantes con receptores y propició que se tratara a muchos más pacientes. También hizo surgir un conjunto significativo de problemas matemáticos: cómo emplear estas estrategias de manera eficaz. De hecho, ya existían herramientas potentes para resolver este dilema. Lo más relevante es que todo empezó hace casi 300 años con un pequeño rompecabezas absurdo.

*

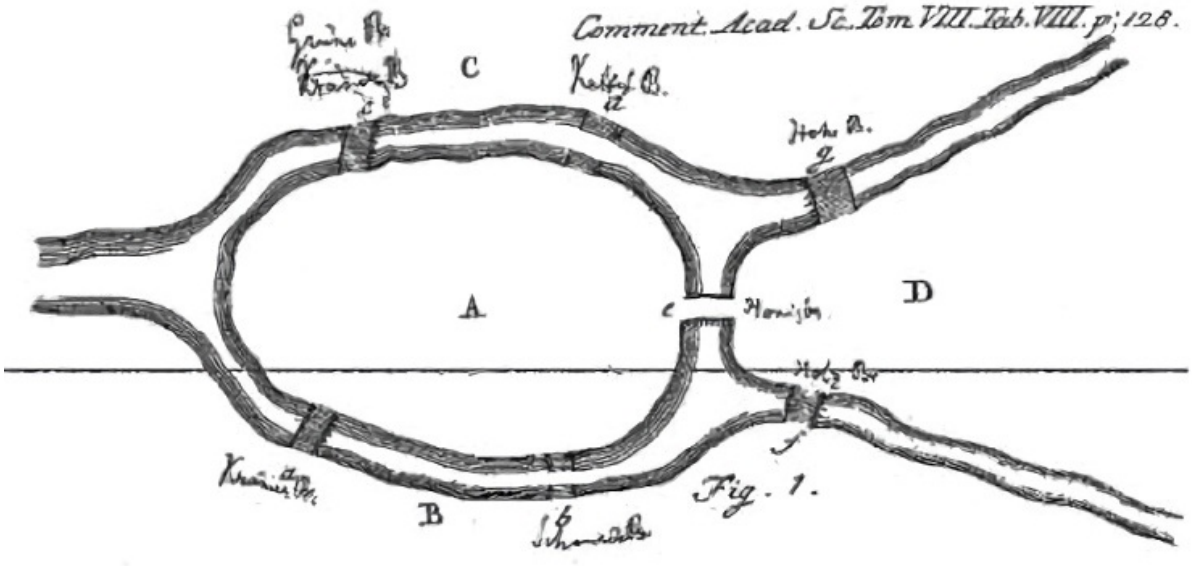
Es una historia muy conocida, pero la voy a contar de todos modos. Y esto por dos motivos: porque pone las matemáticas en contexto y porque es muy común contarla mal. Sin duda, yo lo he hecho en el pasado.

Kaliningrado, que hoy en día está en Rusia, se conocía antes como Königsberg. En el siglo XVIII era parte de Prusia. El río Pregel atraviesa la ciudad y forma dos islas, Kneiphof y Lomse. Había siete puentes. Cada orilla del río estaba unida a Kneiphof por dos de ellos y a Lomse, por uno. Para terminar, el séptimo unía las dos islas entre sí. El trazado actual es bastante diferente. La ciudad fue bombardeada en la segunda guerra mundial y los puentes marcados como b y d en la imagen fueron destruidos. El a y el c fueron demolidos y sustituidos para construir una carretera nueva. Junto con los tres puentes originales restantes, uno de los cuales se reconstruyó en 1935, hay en la actualidad cinco puentes en sus ubicaciones primigenias.

Cuenta la leyenda que los nobles ciudadanos de Königsberg se preguntaban desde hacía mucho tiempo si era posible dar un paseo por la ciudad en el que se cruzase cada puente solo una vez. Era un rompecabezas

sencillo e inocente, del tipo de los que pueden encontrarse hoy en día en la página de pasatiempos del periódico o en su contrapartida digital. Las pruebas con diferentes recorridos no conducen a una solución. Inténtenlo. No obstante, otros problemas parecidos sí que tienen una, aunque a veces sea difícil de encontrar.

Es más, el número de trayectos que pueden hacerse es infinito, aunque solo sea porque las formas que hay de ir de un lado a otro, o de adelante hacia atrás, cuando se pasea a lo largo de un recorrido, son ilimitadas. Así que no puede encontrarse una solución, ni demostrar que no hay una, tomando en consideración todos los caminos posibles.



Esquema de Euler de los siete puentes de Königsberg.

Se puede resolver el rompecabezas con facilidad mediante una trampa de algún tipo. Por ejemplo, se podría caminar por un puente, darse la vuelta, salir de él sin haber llegado a pisar el otro extremo y pretender que se ha «cruzado». La condición «cruzado» debe definirse de manera explícita para no permitir algo así. Del mismo modo, «pasear» quiere decir que no puede hacerse una parte del recorrido a nado, en bote, en globo aerostático ni hacer autostop para que nos lleve el Doctor Who en su TARDIS. Ni remontar el curso del río para encontrar un puente que no esté en el diagrama de Euler. Los entusiastas de los rompecabezas saben que, aunque amañar así el pasatiempo puede ser divertido, e incluso exigir mucho ingenio, no deja de ser hacer trampas. No voy a detallar cada una de las condiciones necesarias para impedir este tipo de amaños. Me interesa mucho más la manera en que puede demostrarse que el rompecabezas, reformulado de manera adecuada

como un problema matemático, es imposible de resolver *a no ser* que se hagan trampas. Estas tienen que ver con la forma en que se reformula el problema, no con su solución ni con la demostración de que esta no existe una vez que ha *sido* planteado.

En esto llegó Euler, el matemático más destacado de su época. Se dedicó a casi todas las ramas de las matemáticas que había (y a otras más que no existían hasta que él las inauguró) y aplicó la disciplina a una variedad enorme de problemas de la vida real. Su obra abarca desde volúmenes eruditos en áreas fundamentales de las matemáticas puras y de la física matemática hasta curiosidades y rarezas que atrajeron su interés. A principios del siglo XVIII, aplicó su ingenio al rompecabezas de los puentes de Königsberg. Lo formuló como un problema con lenguaje técnico preciso y encontró una demostración de que, tal y como está planteado, no puede existir un paseo así. Ni siquiera aunque no sea circular, sino que termine en algún lugar diferente del punto de partida.

Euler se había mudado a San Petersburgo, en Rusia, en 1727, cuando el país estaba regido por la emperatriz Catalina I, para convertirse en matemático de la corte. El marido de la soberana, el zar Pedro I, había fundado la Academia de Ciencias de San Petersburgo (*Academia Scientiarum Imperialis Petropolitanae*) entre 1724 y 1725, pero falleció antes de haberle dado forma definitiva. Euler presentó su trabajo a la Academia en 1735 y lo publicó un año después. Como matemático que era, y puede decirse que el más prolífico de la historia, le sacó el jugo al rompecabezas tanto como pudo: encontró condiciones necesarias y suficientes para que existiese una solución, no solo para el caso de los puentes de Königsberg, sino para cualquier problema de un tipo parecido. Puede haber 50 000 pasarelas que unan 40 000 islas en una disposición compleja en extremo y el teorema de Euler seguirá diciendo si existe una solución. Si se considera la demostración con detenimiento, esta indica incluso cómo encontrar una respuesta (después de darle vueltas un rato). La argumentación de Euler era un poco incompleta y pasaron casi 150 años antes de que alguien resolviera todos los detalles, a pesar de que no son tan difíciles.

Llegados a este punto, muchos libros de teoría de grafos afirman que Euler demostró que el rompecabezas no tiene solución al reducirlo a una cuestión en apariencia más sencilla sobre grafos. En este sentido, un grafo es un conjunto de puntos (denominados nodos o vértices) unidos por líneas (llamadas aristas) que forman una especie de red[36]. La reformulación mediante estos instrumentos convierte el problema de los puentes de

Königsberg en el de trazar un recorrido a lo largo de un grafo particular que emplea cada arista una única vez. Sin duda, esta es la manera en que se aborda la cuestión hoy en día, pero no es lo que hizo Euler. Así es la historia. A los historiadores de las matemáticas les encanta decir lo que ocurrió en realidad, en lugar de lo que cuenta la historia conocida. La verdad es que Euler desentrañó el enredo de manera simbólica[37].

Etiquetó cada zona de tierra firme (isla u orilla de río) y cada puente con una letra. Empleó mayúsculas A, B, C y D para las primeras y minúsculas a, b, c, d, e, f y g para los segundos. Cada uno de estos une dos regiones diferentes. Por ejemplo, f une A con D. Se empieza un paseo en una zona cualquiera y se describe el recorrido al anotar en un listado, en orden, las regiones por las que se pasa y los puentes que se cruzan, hasta acabar en la última franja de tierra firme que se visita. Durante buena parte de su artículo, Euler hace esto de manera verbal y en gran medida se limitó a trabajar con la secuencia de zonas. No importa qué puente se use para ir de A hasta B, solo que el número de veces que aparece AB sea el mismo que el de caminos que las unen. De manera alternativa, es posible emplear la secuencia de puentes, siempre que se deje claro dónde se empieza y se cuente el número de veces que se visita una zona dada. Se podría decir que es más sencillo hacerlo así. Hacia el final del artículo emplea ambos símbolos y pone un ejemplo con la secuencia

EaFbBcFdAeFfCgAhCiDkAmEnApBoElD

que corresponde a un trazado más complicado[38].

En esta formulación, el recorrido preciso que sigue el caminante, en cada franja de tierra o a lo largo de cada camino que las conecta, es irrelevante. Lo único de lo que hay que estar pendiente es de la secuencia de paso por las zonas y de cruce por los puentes. Cruzar un puente se define como que «las dos letras mayúsculas a cada lado sean diferentes». Esto descarta entrar en uno de ellos y salir de nuevo por el mismo extremo. Una solución es una secuencia de mayúsculas y minúsculas alternas, entre A–D y a–g, en la que cada minúscula aparece una y solo una vez y en la que las mayúsculas que hay antes y después de cada letra que corresponde a un puente sean las de las dos regiones que une este.

Es posible hacer un listado de estas uniones para cada minúscula:

a une A con B

b une A con B
c une A con C
d une A con C
e une A con D
f une B con D
g une C con D

Supongamos que se parte de la zona B. Hay tres puentes que la unen con otras regiones: a, b o f. Supongamos que se escoge f. Entonces la secuencia empieza por Bf. La franja de tierra en el otro extremo de f es D, con lo que ahora se tiene BfD. Hay dos puentes que unen D a las otras zonas y que no se han utilizado: e y g (no se puede volver a usar f). Si se prueba con g, el paseo es ahora BfDg. El otro extremo de g es C, lo que da BfDgC. Ahora, los puentes c y d son la única manera de seguir, pues no está permitido volver sobre g. Se intenta el puente c, lo que lleva a BfDgCc y luego a BfDgCcA. Desde la zona A, hay cuatro puentes posibles: a, b, d y e (ya se ha empleado c).

¿Puede, a continuación, cruzarse d? No, porque eso da BfDgCcAd y luego BfDgCcAdC. Ya se han empleado los tres puentes unidos a C, es decir, c, d y g. Pero no se ha resuelto el rompecabezas, porque no se ha cruzado b. Olvidémonos de d. Por razones similares no se puede salir a través de e, porque eso lleva a D y a quedarse atascado. Es más, b se ha quedado sin cruzar de nuevo. ¿Qué pasa con a? Esto da BfDgCcAaB y la única salida que no se ha usado es a través de b, lo que da BfDgCcAaBbA. Ahora solo d o e son caminos válidos. El primero lleva a BfDgCcAaBbAdC sin que haya más opciones, pero no se ha cruzado e. El segundo lleva a BfDgCcAaBbAeD sin otras salidas, pero no se ha cruzado d.

Vale, esa sucesión de opciones no funciona, pero se podrían haber elegido otras diferentes al principio. Pueden desarrollarse de manera sistemática todas las secuencias admisibles… y resulta que deben descartarse sin excepción. Siempre se quedan atascadas en algún momento, sin salida de la zona en la que se encuentran, pero con un puente al menos que todavía no se ha cruzado. La lista de secuencias posibles es finita y lo bastante pequeña como para escribirla entera. Quien lo desee lo puede intentar.

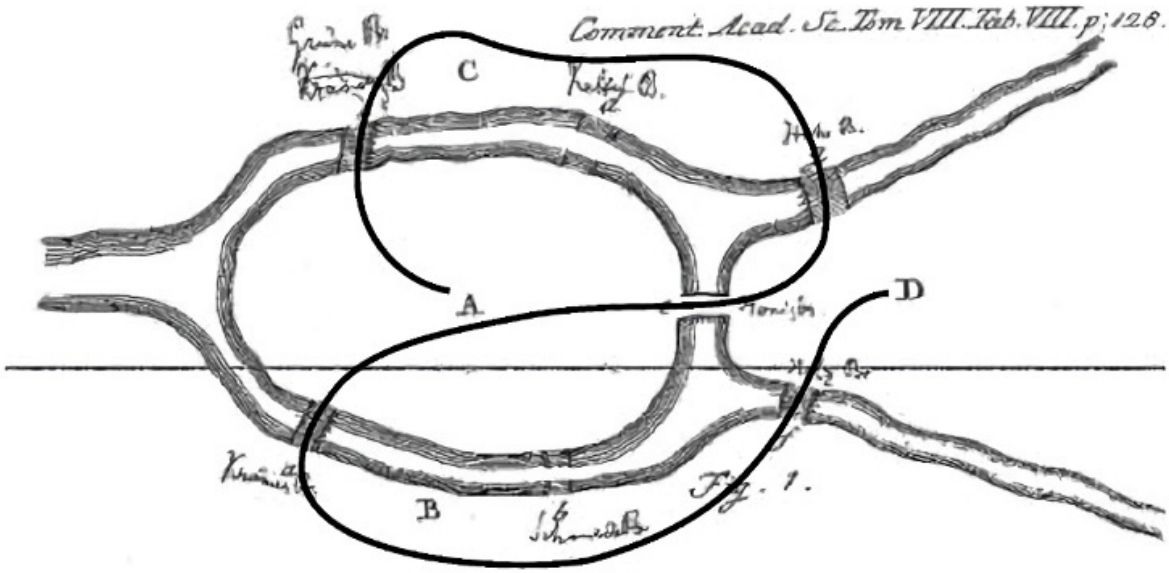
Supongamos que se ha hecho así y se ha demostrado que este rompecabezas en concreto no tiene solución. Esto podía ser suficiente para los ciudadanos de Königsberg, pero no lo era para Euler. En primer lugar, no está claro *por qué* siempre se llega a una situación sin salida. Después, la respuesta

no dice cuándo pueden resolverse, o no, otros rompecabezas del mismo tipo. De modo que Euler se planteó la pregunta más importante que se hacen los matemáticos siempre que alguien resuelve un problema: «Sí, pero ¿por qué ha funcionado?». Acompañada por la siguiente en importancia: «¿Puede hacerse mejor?».

Euler le dio alguna vuelta más e hizo tres observaciones básicas:

- Si hay una solución, cada zona debe estar conectada a todas las demás por *alguna* secuencia de puentes. Por ejemplo, si hubiese dos islas más E y F conectadas entre sí por una o más pasarelas nuevas h, i, j… sin otras que conecten esas islas con el resto de las zonas, entonces la única manera de cruzarlas sería ir adelante y atrás entre E y F. De modo que no se podría llegar a ninguno de los demás puentes.
- Si se asume que la condición anterior de «conectividad» es correcta, entonces, excepto por las dos zonas al principio y al final del paseo, siempre que se entre en una franja de tierra se debe salir de esta de nuevo a través de un puente diferente.
- Siempre que se haga así, dos puentes conectados a esa zona dejan de estar disponibles.

Por lo tanto, al pasear, se utilizan los puentes por parejas. Esta es la noción fundamental. Si una zona se encuentra en el extremo de un número par de puentes, pueden usarse todos sin quedar atrapado en ella. Si se encuentra en el extremo de un número impar de puentes, pueden emplearse todos excepto uno sin quedar atascado. Pero hay que cruzar ese puente en algún momento. Y al hacerlo, no hay salida.



Recorrido abierto que emplea los cinco puentes que quedan en pie.

Quedar atrapado en mitad de un hipotético recorrido es catastrófico. No obstante, no es un inconveniente si se ha llegado al final del paseo. Invertir este y, de hecho, caminar hacia atrás, tampoco es un problema al iniciar el recorrido. Este razonamiento implica que si existe una solución, debe haber como mucho dos zonas al final de un número impar de puentes. Para el caso de Königsberg:

A está unida por 5 puentes
B está unida por 3 puentes
C está unida por 3 puentes
D está unida por 3 puentes

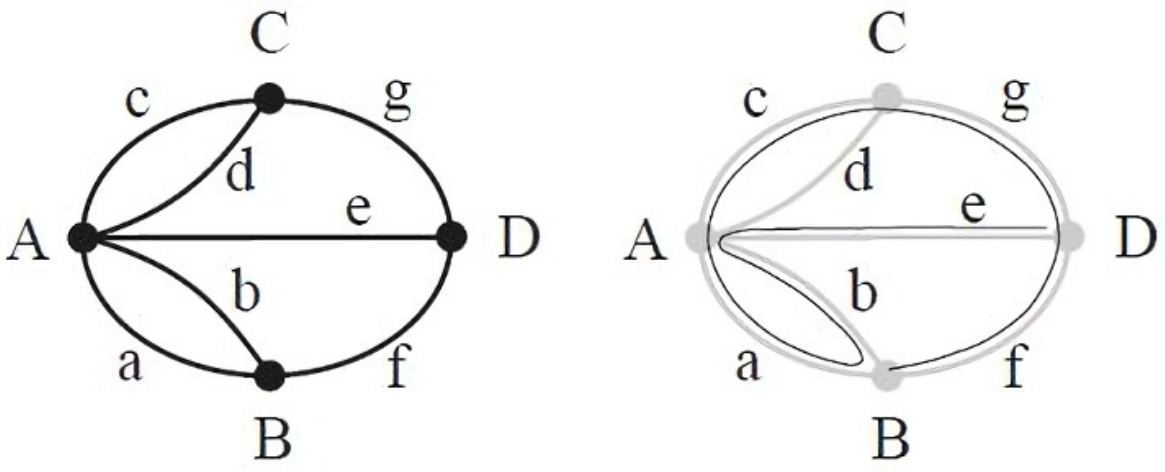
Por lo tanto, el número de zonas al final de un número impar de puentes es cuatro, que es mayor que dos. Así que no existe recorrido alguno.

Euler afirmó también, sin demostrarlo, que la misma condición par/impar es suficiente para que exista una solución. Esto es un poco más difícil y no voy a considerarlo. Lo demostró Carl Hierholzer justo antes de su muerte en 1871 y se publicó de manera póstuma en 1873. Además, Euler observó que si se desea un circuito cerrado, que termine donde empezó, entonces una condición necesaria y suficiente es que todas las zonas se encuentren al final de un número par de puentes[39].

Si se emplean solo los cinco puentes que, de algún modo, sobreviven en la actualidad, tanto B como C están unidas por dos de ellos. De este modo, este problema revisado debe tener una solución, pero solo para un circuito cerrado. Los puntos de los extremos deben estar en A y en D, dado que estas zonas siguen unidas por un número impar de puentes. La imagen muestra una solución así. Hay otras. ¿Quién puede encontrarlas todas?

Euler expresó todo lo anterior en términos de secuencias de símbolos, como BfDgCcAaBbAeD. Cierto tiempo después, alguien se dio cuenta de que es posible dar a todo esto una interpretación visual. No está claro quién fue exactamente, porque la idea estaba en el aire a mediados del siglo XIX, pero James Joseph Sylvester introdujo el término «grafo» en 1878. Se dibuja una imagen con cuatro puntos A-D y siete líneas a-f. Se hace que cada una de ellas conecte las dos zonas en los extremos del puente correspondiente. Con esto se simplifica el mapa de islas y puentes, como en la imagen de la izquierda. La secuencia de símbolos que se acaba de mencionar corresponde

al recorrido en la imagen de la derecha, que empieza en B y acaba en D, donde queda atascado.



Izquierda: grafo que muestra las conexiones de los puentes de Königsberg. *Derecha*: intento sencillo de recorrido. Se omite el puente d.

Esta simplificación visual es el *grafo* de los puentes de Königsberg. En esta representación, no importa dónde se sitúen los cuatro puntos (aunque deberían permanecer separados para impedir que se confundan unos con otros) ni la forma precisa de las líneas. Lo único relevante son los vértices que conecta una arista dada. En este marco gráfico, la demostración de Euler resulta muy natural. Cualquier recorrido que llegue a una zona a través de un puente debe salir cruzando otro, a no ser que se trate del final de un circuito cerrado. De manera similar, cualquier recorrido que deje una zona a través de un puente debe haber entrado en ella cruzando otro, a no ser que se trate del inicio de un circuito cerrado. De modo que los puentes van en parejas excepto en los dos extremos del trayecto. Por lo tanto, todas las regiones que no son la primera ni la última reciben un número par de puentes. Si los extremos reciben una cantidad impar, solo es posible un recorrido abierto. De forma alternativa, el comienzo y el final pueden estar en la misma zona, de modo que pueden unirse entre sí sin emplear más puentes, lo que da lugar a un circuito cerrado. Ahora todas las regiones reciben un número par de puentes.

Al resolver esta clase única de problemas, Euler consiguió fundar dos ramas importantes de las matemáticas. Una es la teoría de grafos, que estudia puntos unidos por líneas. Parece sencilla, incluso infantil. Lo es. Al mismo tiempo, es profunda, útil y complicada, como se verá. La otra es la topología, en ocasiones denominada «geometría de la lámina de goma», en la que las formas pueden modificarse de manera continua sin que se considere que son diferentes en lo esencial. Aquí los perfiles de las líneas y las ubicaciones de

los puntos pueden deformarse tanto como se quiera, siempre que no cambie el modo en el que están conectados (el requisito de la continuidad) y el grafo resultante es el mismo en esencia. El mismo, en el sentido de que contiene la misma información acerca de qué está conectado a qué.

Me parece sorprendente que un rompecabezas sencillo pueda llevar a innovaciones tan significativas. Eficacia irrazonable, sin duda. Esto contiene también una lección importante, que a menudo el mundo exterior pasa por alto. No hay que subestimar las matemáticas que *parecen* sencillas, más como un juego de niños que algo serio. Lo que cuenta no es lo simple que sea el juguete, sino lo que se hace con él. De hecho, un objetivo principal de las buenas matemáticas es lograr que todo sea tan sencillo como sea posible. Habrá quien se ría. Está justificado si se tiene en cuenta lo complicada que parece buena parte de la disciplina. Debo añadir la advertencia que se atribuye a Einstein: tan sencillo como sea posible, *pero no más*. Reducir islas a puntos y puentes a líneas no modifica el rompecabezas, pero elimina información irrelevante. ¿Qué tiempo hace? ¿Hay barro? ¿El puente es de metal o de madera? Estas cosas son importantes para un paseo dominical o para construir una pasarela. Pero si se quiere dar respuesta al enigma que intrigaba a los nobles ciudadanos de Königsberg, se trata de elementos innecesarios.

*

¿Qué tienen que ver los puentes de Königsberg con los trasplantes de riñones? Directamente, no mucho. De manera indirecta, el artículo de Euler sentó las bases para el desarrollo de la teoría de grafos, que ha permitido la aparición de potentes métodos para emparejar receptores con donantes, incluso cuando la mayoría de estos solo están dispuestos a donar sus órganos a parientes cercanos[40]. Cuando entró en vigor el Decreto de Tejidos Humanos de Reino Unido en 2004, se pudieron donar riñones de manera legal a personas que no fueran familiares.

Un problema importante es emparejar los receptores con los donantes porque, incluso cuando alguno de estos se presenta voluntario, sus tipos de tejido y sangre pueden no ser adecuados para quien requiere el trasplante. Supongamos que el tío Fred necesita un riñón y que su hijo William está dispuesto a donar uno, aunque no a un desconocido. Por desgracia, el tipo de tejido del órgano de William no es el correcto. Antes de 2004 no habría habido nada más que hablar y Fred estaría condenado desde entonces a realizar tratamientos periódicos con una máquina de diálisis. Lo mismo habría

ocurrido con muchos otros receptores potenciales cuyo tipo de tejido sí correspondía al de William. Supongamos ahora que John Smith, que no es pariente de Fred ni de William, tiene el mismo problema: su hermana Emily necesita un riñón nuevo y él está dispuesto a donar uno, pero, una vez más, no a un desconocido. Como en el caso anterior, su tipo de tejido es diferente del de Emily. Así que nadie puede recibir un trasplante.

No obstante, supongamos que el tipo de tejido de John corresponde al de Fred y el de William, al de Emily. Después de 2004, esta situación reúne las condiciones para una donación cruzada legal. Los cirujanos implicados pueden discutirlo y sugerir que John permita que su órgano se trasplante a Fred a condición de que Emily reciba el de William. Es mucho más probable que ambos acepten este arreglo, porque su pariente recibe un riñón nuevo, mientras que cada uno de ellos hace la donación que siempre estuvo dispuesto a hacer si era para el bien de su familiar. El hecho concreto de quién recibe el órgano de quién no tiene mucha importancia ni para los donantes ni para los receptores, si bien es fundamental para emparejar los tipos de tejido.

Gracias a las comunicaciones modernas, los cirujanos pueden identificar cuándo se producen coincidencias así al mantener un registro de donantes y receptores potenciales junto con sus tipos de tejido. Cuando el conjunto de pacientes y de voluntarios es pequeño, es poco probable que surjan intercambios oportunos como este, pero se vuelven mucho más viables al aumentar las cantidades. El número de receptores potenciales es bastante elevado: en 2017 había en Reino Unido más de 5000 personas en la lista de espera para recibir un riñón nuevo. Podría provenir de un donante fallecido o vivo, pero el número de estos es menor (en torno a 2000 en la misma fecha), lo que da lugar a un tiempo de espera habitual de más de dos años para un adulto y de nueve meses para un niño.

Una manera de garantizar que se benefician más pacientes todavía y de que reciben un tratamiento más rápido, es establecer cadenas de donaciones cruzadas más elaboradas. Además, ahora la ley permite este sistema. Supongamos que Amelia, Bernard, Carol y Deirdre necesitan un trasplante. Cada uno de ellos tiene un voluntario en espera, dispuesto en principio a hacer la donación, pero solo a esa persona. Supongamos que estos son Albert, Beryl, Charlie y Diana. El sistema se pone en marcha con una donante altruista, Zoe, que está dispuesta a donar un riñón a quien sea. Supongamos que los tipos de tejidos permiten una cadena como la que sigue:

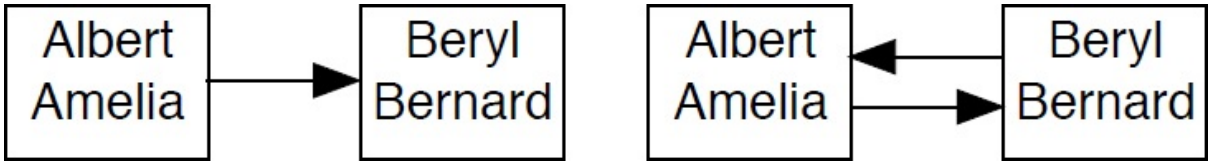
Zoe dona a Amelia.
El donante de Amelia, Albert, acepta donar a Bernard.

La donante de Bernard, Beryl, acepta donar a Carol.
El donante de Carol, Charlie, acepta donar a Deirdre.
La donante de Deirdre, Diana, acepta donar a la lista de espera.

En conjunto, todo el mundo queda satisfecho. Amelia, Bernard, Carol y Deirdre reciben riñones nuevos. Albert, Beryl, Charlie y Diana donan uno de los suyos, no a su pariente, sino como parte de una cadena que *beneficia* a su familiar. A menudo estarán de acuerdo con hacerlo así, lo que permite que estos arreglos sean posibles. De hecho, si no lo están, su pariente no recibirá un riñón en esta ocasión. Zoe también está satisfecha porque su donación altruista beneficia a alguien y no le importa a quién sea. En este caso, es a Amelia. Por último, un riñón adicional va a la lista de espera, lo que siempre es útil.

Si en vez de esto Zoe hubiese hecho la donación a la lista de espera, la única manera de que Amelia, Bernard, Carol y Deirdre hubiesen recibido trasplantes habría sido a través de la lista. Al no hacerlo así, liberan cuatro riñones más. Es lo que se conoce como cadena de donaciones de efecto dominó. Zoe tumba una ficha y todas las demás en la serie caen a su vez. Lo abreviaremos a *cadena*.

Lo importante aquí no son los nombres, sino los tipos de tejido. Amelia puede ser cualquiera que tenga el mismo tipo de tejido que Zoe; Bernard, que el *donante* de Amelia; Carol, que el donante de Bernard, etcétera. Cadenas así son comunes para números razonables de receptores y donantes y los cirujanos pueden identificarlas. No obstante, esto lleva mucho tiempo incluso cuando delegan la tarea y cada riñón es un tesoro, así que tiene sentido diseñar las cadenas de la mejor manera posible. Es complicado, porque puede haber muchas en potencia a un mismo tiempo. En este caso, los cirujanos pueden seguir adelante con el procedimiento de manera simultánea, a no ser que dos cadenas contengan al mismo donante y requieran que el trasplante se haga a dos personas diferentes. Entonces se rompe una de las cadenas.



Dos tipos de donación cruzada.

Optimizar el diseño de los procesos… Hum. Suena a matemáticas. Si puede formularse el problema en términos matemáticos y aplicar las técnicas adecuadas, tal vez se consiga resolver. Es más, la solución no tiene por qué

ser perfecta. Basta con que sea mejor que cualquier otra que se pueda lograr a ojo de buen cubero. David Manlove descubrió una manera de convertir el problema de la donación cruzada de riñones en un asunto de grafos. El teorema de Euler no ayuda a resolverlo porque su función era fundar toda la rama. A lo largo de los años que han pasado desde entonces, los matemáticos han desarrollado la disciplina y han inventado muchas técnicas nuevas de teoría de grafos. Dado que estos son objetos discretos («en realidad» no son más que una lista de nodos, de aristas y de información sobre qué arista une qué nodos), se adaptan de manera ideal a su manejo por ordenador. Se han desarrollado potentes algoritmos para analizar grafos y extraer su estructura útil. Entre estos hay algunos que pueden encontrar el reparto óptimo de donantes con respecto a los pacientes para conjuntos de tamaño realista. Estos métodos aplicados por ordenador se utilizan de manera rutinaria en la actualidad en Reino Unido.

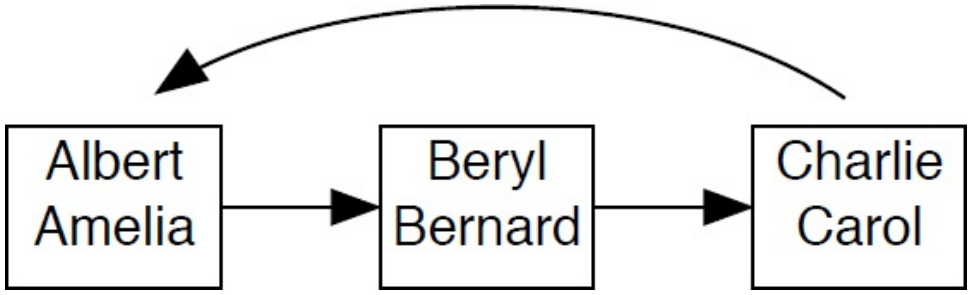
*

Las parejas compatibles de donantes y receptores son sencillas: solo hay que intercambiar sus riñones. Esto exige que haya dos cirujanos operando al mismo tiempo, uno a cada persona. De este modo, es posible dejar de lado las parejas compatibles cuando se buscan cadenas y centrarse en las que no lo son. Estas *parejas* constituyen los nodos del grafo.

Por ejemplo, supongamos que Albert está dispuesto a hacer una donación a Amelia, pero que su tipo de tejido coincide con el de Bernard. Puede representarse esta situación por la imagen de la izquierda. El nombre del donante está arriba y el de su pariente incompatible debajo. La flecha quiere decir que «el donante en la base es compatible con el receptor en la punta». Esta figura constituye un tipo especial de grafo en el que las aristas tienen sentidos definidos. A diferencia de los puentes de Königsberg, estas aristas son unidireccionales: los matemáticos las denominan dirigidas y el grafo resultante es un grafo dirigido, abreviado como dígrafo. En un dibujo, las aristas dirigidas se representan mediante flechas.

Si además resulta que Beryl es compatible con Amelia, las reglas indican que hay que dibujar otra flecha en el sentido opuesto. Con esto se crea una conexión bidireccional, como en la imagen de la derecha. Esta ilustra el tipo más sencillo de donación cruzada, lo que en teoría de grafos se denomina un ciclo C_2. Los cirujanos pueden sugerir que Albert done su órgano a Bernard, a condición de que el de Beryl vaya a Amelia. Si todas las partes se ponen de

acuerdo, entonces Amelia y Bernard reciben los trasplantes, mientras que Albert y Beryl hacen una donación cada uno. Aunque su pariente no recibe *su* riñón, recibe *un* riñón. Ambos receptores resultan beneficiados y ambos donantes cumplen su objetivo, de modo que la mayoría de los voluntarios potenciales están dispuestos a aceptar este tipo de intercambio.



Una donación cruzada de riñones tipo C_3.

El tipo siguiente en orden de complicación es un ciclo C_3. Ahora hay una tercera pareja, con el donante Charlie y la receptora Carol. Supongamos que

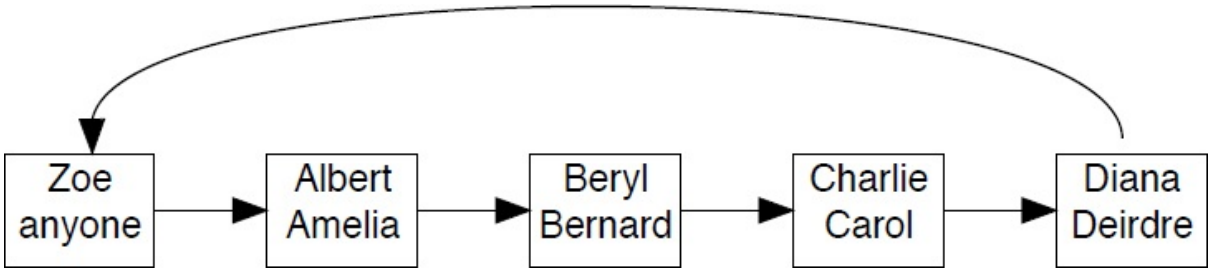
Albert es compatible con Bernard.
Beryl es compatible con Carol.
Charlie es compatible con Amelia.

Entonces los cirujanos pueden disponer que el riñón de Albert vaya a Bernard, el de Beryl, a Carol y el de Charlie, a Amelia. De nuevo, la mayoría de los voluntarios lo aceptarían.

Los donantes altruistas como Zoe tienen que considerarse de una manera un tanto diferente porque no están emparejados con ninguna persona en concreto. Hay que hacer unos cuantos trucos matemáticos. Se crea el nodo correspondiente al emparejar a Zoe con un receptor denominado «alguien», que se considera compatible con todos los donantes no altruistas. En la práctica, este paciente comodín representa a todas las personas en la lista de espera. Se asume que cada una de ellas es compatible con *algún* donante no altruista, algo razonable porque la lista de espera es grande. A continuación, se traza una flecha desde

nodo Z = (Zoe, alguien)

hasta cualquier nodo cuyo receptor sea compatible con Zoe. En la cadena de efecto

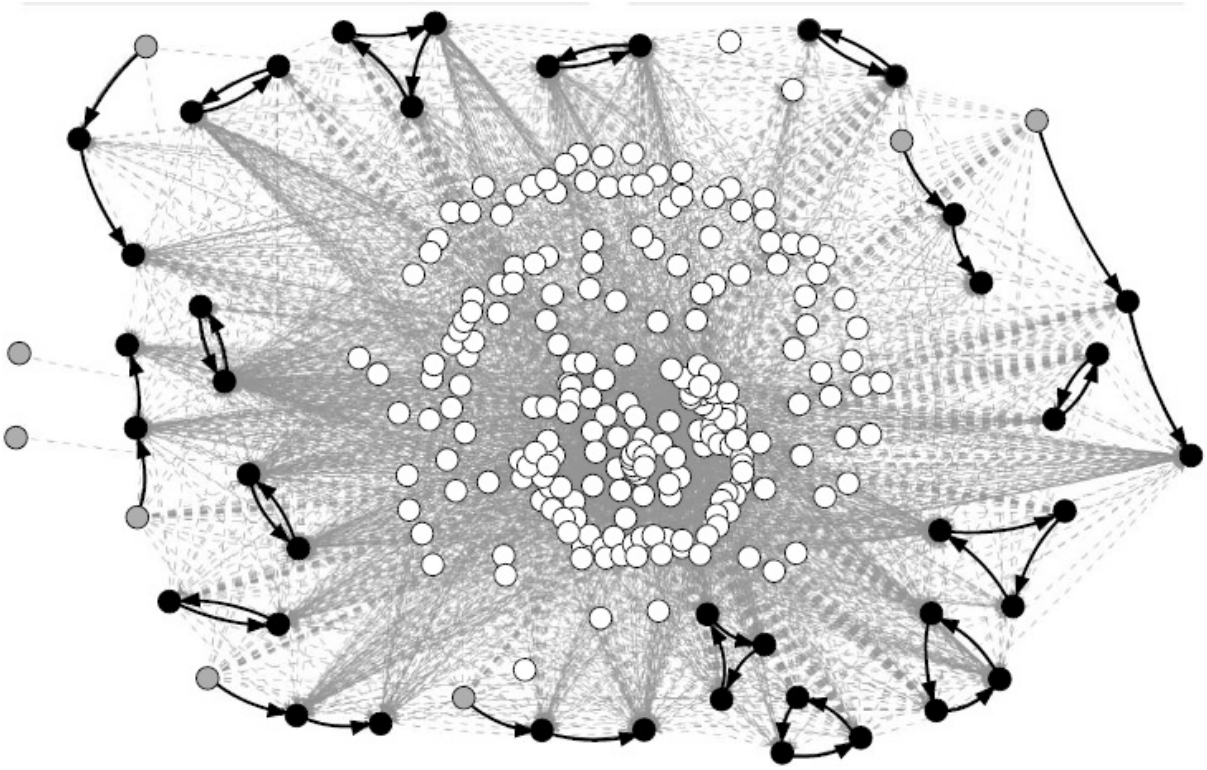


Esta cadena es demasiado larga como para resultar práctica.

Cadenas como esta no son prácticas. Esta necesitaría a diez cirujanos operando de manera simultánea. Y tiene que ser básicamente a la vez porque de otro modo alguien, pongamos por caso Charlie, podría cambiar de repente de opinión y negarse a hacer la donación a Deirdre, cuando ya Carol ha recibido un riñón de Beryl. Las personas no siempre cumplen con lo acordado cuando les resulta ventajoso no hacerlo, por mucho que hayan firmado documentos legales. Si se lo proponen, encontrarán algún modo de zafarse. Fingirán una enfermedad o algo. Se romperán una pierna.

Por este motivo, las donaciones cruzadas se limitan en la actualidad a cuatro situaciones: los ciclos C_2 y C_3, que ya se han descrito, y los ciclos correspondientes en los que participa un donante altruista, que se denominan cadenas cortas y largas. Una cadena corta incluye solo a Zoe, Albert, Amelia y a alguien de la lista de espera. Una larga incluiría, además, a Beryl y a Bernard. Un *intercambio* es cualquiera de estas cuatro situaciones.

Cabe destacar el truco de prestidigitador. He mostrado esta cadena como un ciclo de cinco nodos. Pero cuando se realiza como un intercambio, no lo es en realidad, porque Zoe no tiene en mente a ningún receptor concreto. Está dispuesta a hacer una donación a quien sea, pero es Diana quien la *hace a alguien* (es decir, a la lista de espera) al final de la cadena. Esa persona cualquiera que recibe el órgano de Zoe en realidad es Amelia, que no es la misma a la que Diana acaba por donar el riñón. Las matemáticas se encargan de este aspecto porque se infiere quién es ese «alguien» en cada caso a partir de la estructura del análisis del dígrafo.



Dígrafo de donaciones cruzadas de riñones empleado en julio de 2015 y una solución óptima (líneas negras continuas). Puntos blancos = donantes y receptores sin pareja. Puntos grises = donantes altruistas. Puntos negros = receptores/donantes emparejados.
Muggleton (redibujada).

Los dígrafos anteriores muestran ciclos y cadenas individuales que emplean cantidades reducidas de nodos y aristas dirigidas. En la vida real hay muchas parejas, un número más bien pequeño de donantes altruistas y un volumen enorme de flechas. Esto ocurre porque el dígrafo debe tener una de estas entre dos nodos *cualesquiera* X e Y tales que el donante de X sea compatible con el receptor de Y, y el mismo voluntario puede serlo con muchos pacientes. Por ejemplo, en octubre de 2017 había registrados en total 266 parejas no altruistas y 9 donantes altruistas, con 5964 flechas. La imagen ilustra una complejidad similar en otra fecha. El reto matemático es encontrar en el dígrafo no ya un intercambio, sino el conjunto óptimo de estos.

*

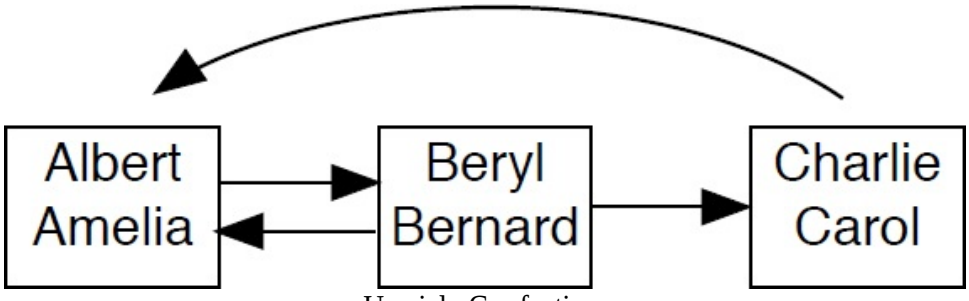
Para resolver este problema de manera matemática, es necesario precisar la noción de «óptimo». No se trata solo del conjunto de intercambios que incluye al mayor número de personas. Hay otros aspectos que tener en cuenta, tales como el coste y la tasa probable de éxito. En este sentido, la opinión y la experiencia de los médicos son determinantes. El departamento Sangre y Trasplantes del Servicio Nacional de Salud (NHSBT, por sus siglas en inglés)

de Reino Unido ha desarrollado un sistema de puntuación estándar para cuantificar las ventajas probables de cualquier trasplante concreto. Este toma en consideración factores tales como el tiempo que ya ha esperado el paciente, los niveles de incompatibilidad del tipo del tejido y la diferencia de edad entre el donante y el receptor. Se emplea un análisis estadístico y se combinan estos factores en una única puntuación numérica, una cantidad real que se denomina «peso», y que se calcula para cada flecha en el dígrafo, es decir, para cada trasplante potencial.

Una condición es evidente: dos donaciones cruzadas diferentes no pueden incluir ningún nodo en común porque es imposible dar el mismo riñón a dos personas distintas. En sentido matemático, esto significa que los ciclos que componen el conjunto no pueden solaparse. El resto de las condiciones son más sutiles. Algunos ciclos C_3 muestran una característica adicional muy útil: una flecha más en el sentido inverso entre dos nodos. La imagen muestra un caso en el que, además de las flechas del ejemplo anterior, hay otra que une a Beryl con Amelia. Es decir, Beryl es compatible con Amelia tanto como con Carol. Si en un momento posterior, Charlie se retira del acuerdo de intercambio, Carol puede eliminarse también. Queda entonces un ciclo C_2 en el que Albert hace una donación a Bernard y Beryl a Amelia y este proceso puede seguir adelante. Desde un punto de vista matemático, este trío de nodos forma un ciclo C_3 y además dos de los vértices forman otro, tipo C_2. La flecha adicional se denomina «arco invertido». Los ciclos C_3 con arcos invertidos y cualquier ciclo C_2 se denominan ciclos C_2 *efectivos*.

El Grupo Asesor de Riñón del NHSBT ha definido un conjunto de donaciones cruzadas de riñón como óptimo si:

1. Maximiza el número de ciclos C_2 efectivos.
2. Contiene tantos ciclos como sea posible, sujeto a (1).
3. Emplea tan pocos ciclos C_3 como sea posible, sujeto a (1) y (2).
4. Maximiza el número de arcos invertidos, sujeto a (1) hasta (3).
5. Maximiza el peso total de los ciclos, sujeto a (1) hasta (4).



Un ciclo C_2 efectivo.

Tras esta definición se oculta la intuición de que ciertas características son prioritarias. Una vez que se han satisfecho estas, es posible tomar otras de menor relevancia en consideración, de manera sucesiva. Por ejemplo, la condición (1) garantiza que incluir donaciones cruzadas de tres direcciones no reduce el número de las bidireccionales que podrían hacerse de otro modo. Esto tiene sus ventajas. Una de ellas es la simplicidad y otra la posibilidad de seguir adelante con el ciclo C_2 si alguien se echa atrás. La condición (5) quiere decir que después de tomar las decisiones principales en (1) hasta (4) y solo entonces, el conjunto de intercambios debería ser tan eficaz y tener tanta probabilidad de salir con éxito como fuese posible.

El problema matemático es encontrar un conjunto óptimo de donaciones cruzadas conforme a estos criterios. Con solo pensar un poco y hacer unos cuantos cálculos a vuelapluma, se ve que no es factible comprobar todos los conjuntos de intercambios posibles. Hay demasiadas combinaciones. Digamos que son 250 nodos y 5000 aristas. Como promedio, inciden 20 de ellas en cada vértice. Para dar una cifra aproximada puede asumirse que este se halla en la punta de 10 flechas y en la base de otras 10. Supongamos que se quiere hacer un listado de todos los ciclos C_2 posibles. Se escoge un nodo y se siguen las 10 flechas que salen de él. Cada una de ellas termina en un vértice diferente, cada uno a su vez con sus 10 flechas salientes propias. Hay un ciclo C_2 si el último nodo es el mismo que el primero. Eso significa que son 100 combinaciones que verificar. Encontrar un ciclo C_3 requiere $100 \times 10 = 1000$ comprobaciones de este tipo (1100 por nodo). Con 250 de estos, hay que verificar 275 000 casos, si se dejan de lado los atajos que pueden reducir algo el total, pero que no cambien el orden de magnitud general.

Sin embargo, todo lo que se ha conseguido hasta el momento es un listado de ciclos C_2 y C_3 potenciales. Un intercambio es un *conjunto* de estos y el número de conjuntos crece de manera exponencial con el de ciclos. En octubre de 2017, el dígrafo tenía 381 ciclos de tipo C_2 y 3815 del C_3. La cantidad de conjuntos solo para C_2 es 2^{381}, una cifra de 115 dígitos. Para C_3, tiene 1149 dígitos. Y ni siquiera se ha comprobado que los conjuntos no se solapen.

No hace falta decir que el problema no se resuelve así. Pero queda claro que hay que inventar unos métodos muy potentes para poder hacerlo. Solo voy a indicar algunas de las ideas necesarias. Puede pensarse en esta cuestión como en una especie exagerada de problema de la persona viajante: es un asunto de optimización combinatoria con condiciones bastante diferentes, aunque algunas consideraciones son similares. La más importante es el

tiempo que se tarda en calcular una solución óptima. Es posible analizar esta estrategia desde el punto de vista de la complejidad computacional, como en el capítulo 3.

Si las donaciones cruzadas solo incluyen ciclos C_2, puede obtenerse un conjunto óptimo de intercambios en tiempo polinómico, clase de complejidad P, mediante el empleo de métodos estándar para maximización de emparejamientos ponderados en un grafo. Cuando hay ciclos C_3, incluso sin donantes altruistas, el problema de optimización es NP-complejo. En todo caso, Manlove y colaboradores diseñaron un algoritmo viable basado en programación lineal, como la que ya se ha visto en el capítulo 3. Su algoritmo, UKLKSS, reformula el problema de optimización de modo que puede resolverse mediante el empleo de una secuencia de cálculos de programación lineal. El resultado de cada uno alimenta al siguiente como una restricción adicional, de modo que se optimiza la condición (1). Para ello se emplea un método denominado algoritmo de Edmond en la versión de Silvio Micali y Vijay Vazirani. El algoritmo de Edmond encuentra la correspondencia máxima en un grafo en un tiempo proporcional al número de aristas multiplicado por la raíz cuadrada del número de nodos. Una correspondencia asocia parejas de vértices en los extremos de una arista común y el problema es formar tantos pares como sea posible sin emplear dos aristas que se encuentren en un vértice común.

Una vez optimizada la condición (1), se alimenta esta solución al cálculo de la (2), para lo que se emplea un algoritmo denominado solucionador de programación en enteros COIN-Cbc, que es parte de la colección de algoritmos del proyecto Computational Infrastructure for Operations Research, y así en adelante.

A finales de 2017, estos métodos de teoría de grafos habían identificado un total de 1278 trasplantes potenciales, pero solo se habían llevado a cabo 760 debido a problemas prácticos de todo tipo que pueden surgir en las etapas finales de evaluación: descubrir que los tipos de tejido no son tan compatibles como se creía o que los donantes o los receptores estaban demasiado enfermos como para someterse a la operación. No obstante, el empleo sistemático de algoritmos de teoría de grafos para organizar de manera eficaz los trasplantes cruzados de riñón es una mejora importante respecto a los métodos anteriores. También señala el camino para mejoras futuras, porque ahora es posible mantener sanos los riñones fuera del cuerpo durante periodos más largos, de modo que no es necesario que las operaciones en cadena se

realicen el mismo día. Esto hace que sea factible plantearse cadenas más largas, lo que supone a su vez nuevos problemas matemáticos.

No voy a dar crédito a Euler por ser un visionario. No tenía ni la menor idea de que su brillante solución a un rompecabezas irrelevante iba a resultar útil alguna vez en medicina. Desde luego no en trasplantes de órganos, en una época en que la cirugía era una dolorosa forma de carnicería. Pero sí quiero dárselo por percatarse, incluso en esa época temprana, de que el rompecabezas permitía entrever algo mucho más profundo. Lo dijo así, de manera explícita. Basta con ver la cita al inicio de este capítulo. Euler menciona en repetidas ocasiones la «geometría de la posición» como el contexto relevante. El término que empleó en latín fue *analysis situs*. Lo atribuyó a Leibniz al igual que, por extensión, el hecho de darse cuenta de que una rama así podía ser importante. No cabe duda de que le intriga la noción de una forma de geometría que no trata de las formas euclidianas habituales. No la rechaza porque sea poco ortodoxa, al contrario. No se deja constreñir por la tradición. Y le complace aportar su propio granito de arena al desarrollo de una geometría así. Le divierte.

El sueño de Leibniz dio fruto en el siglo XX, con algunos avances significativos en el XIX. En la actualidad se le llama topología y mostraré algunos de sus usos novedosos en el capítulo 13. La teoría de grafos todavía mantiene alguna relación con ella, pero en buena medida se ha desarrollado en una línea diferente. Nociones tales como el peso de una arista son numéricas y no topológicas. Pero la idea de que pueden emplearse grafos para modelar sistemas de interacciones complicados y para resolver cuestiones de optimización se remonta a Euler y a su forma de abordar este nuevo tipo de problema que *le llamó la atención* y de desarrollar su propia manera de hacerlo. En San Petersburgo, Rusia, bajo la emperatriz Catalina I, hace casi tres siglos. Cualquiera que reciba un trasplante de riñón, en Reino Unido o en cualquier otro país que use técnicas de teoría de grafos para distribuir los órganos de manera más eficaz, debería estar encantado de que lo hiciese.

5
Con seguridad en el ciberespacio

> Nadie ha descubierto todavía ninguna aplicación militar de la teoría de los números y de la relatividad, y no parece probable que alguien lo haga en muchos años[41].
>
> GODFREY HAROLD HARDY,
> *Apología de un matemático*, 1940

Pierre de Fermat es famoso por su «último teorema», según el cual, si *n* es al menos 3, la suma de dos números naturales elevados a *n* no puede ser otro con la misma potencia *n*. Andrew Wiles acabó hallando una demostración técnica y moderna en 1995, unos 385 años después de que se enunciase el teorema[42]. Fermat era abogado en el Parlamento de Toulouse, pero pasaba la mayor parte de sus días dedicado a las matemáticas. Tenía un amigo que se llamaba Frénicle de Bessy, un matemático parisino más conocido por elaborar un catálogo con los 880 cuadrados mágicos de orden cuatro que hay. Ambos se escribían con frecuencia y el 18 de octubre de 1640, Fermat envió una carta a Bessy en la que le decía (en francés) que «todo número primo divide… una de las potencias menos uno de cualquier progresión en la que el exponente de esta potencia dada es un factor del número primo dado menos uno».

En lenguaje algebraico, Fermat estaba afirmando que si p es primo y a es cualquier número, entonces $a^{p-1} - 1$ es divisible por p (sin resto). Así, por ejemplo, dado que 17 es primo, asegura que todos los números

$$1^{16} - 1,\ 2^{16} - 1,\ 3^{16} - 1 \ldots\ 16^{16} - 1,\ 18^{16} - 1 \ldots$$

son múltiplos exactos de 17. Es obvio que hay que dejar fuera a $17^{16} - 1$, que no puede satisfacer la relación porque es una unidad menor que un múltiplo, en concreto que 17^{16}. Fermat sabía que se necesitaba esta condición adicional, pero no lo decía en la carta. Comprobemos un caso:

$$16^{16} - 1 = 18\,446\,744\,073\,709\,551\,615$$

y este número dividido por 17 es

$$1\,085\,102\,592\,571\,150\,095$$

con exactitud. ¡Quién lo iba a decir!

Esta curiosidad se conoce en la actualidad como el pequeño teorema de Fermat, para distinguirlo del último (o grande). Fermat fue uno de los pioneros de la teoría de números, que estudia las propiedades profundas de los números naturales. En su época, y durante los tres siglos posteriores, esta era la forma más pura de las matemáticas puras. No tenía aplicaciones importantes y no daba la impresión de que las fuese a tener nunca. Desde luego, Godfrey Harold Hardy, uno de los matemáticos teóricos más destacados de Gran Bretaña, era de esa opinión y lo dijo así en su pequeña joya *Apología de un matemático*, publicada en 1940. Una de sus ramas favoritas de la disciplina era la teoría de números y junto con Edward Maitland Wright publicó en 1938 un texto clásico, *Una introducción a la teoría de números*. En él figura el pequeño teorema de Fermat, como teorema 71 en el capítulo VI. De hecho, todo ese apartado trata de sus consecuencias.

Las opiniones políticas y matemáticas de Hardy estaban teñidas de las actitudes prevalentes en los niveles más altos del mundo académico y ahora resultan un tanto afectadas. Sin embargo, su escrito es elegante y ofrece una comprensión valiosa de las actitudes académicas de su época. Algunas son incluso relevantes hoy en día, si bien otras se han visto superadas en buena medida por los acontecimientos. Hardy escribió que: «Es una experiencia melancólica para un matemático profesional encontrarse a sí mismo escribiendo sobre matemáticas. La función de un matemático es hacer algo, es probar nuevos teoremas, es contribuir a las matemáticas y no hablar de lo que él u otros matemáticos han hecho». A la porra la «divulgación», tan valorada en el mundo académico actual (aunque la misma actitud elitista hacia la comunicación prevalecía en ese ámbito hace cuarenta años).

Un motivo por el que Hardy se sentía en la necesidad de justificar su profesión era que, desde su punto de vista, el tipo de matemáticas a las que había dedicado su vida no tenían ninguna aplicación útil y era poco probable que encontrasen alguna. No salía a cuenta. Su interés en la disciplina era meramente intelectual: la satisfacción de resolver problemas difíciles y de hacer avanzar el conocimiento humano abstracto. No estaba muy preocupado

por su utilidad, pero se sentía un poco culpable por ello. Lo que sí le inquietaba, como pacifista de toda la vida, era que las matemáticas no se empleasen para fines bélicos. La segunda guerra mundial estaba en su apogeo y, a través de los tiempos, algunas ramas de las matemáticas habían tenido usos militares serios. Se dice que Arquímedes empleó su conocimiento de las parábolas para concentrar la luz del sol sobre los barcos enemigos y prenderles fuego, al igual que la ley de la palanca para diseñar una enorme garra que podía levantar los navíos y sacarlos del agua. La balística indica cómo apuntar la artillería, desde balas de cañón macizas hasta explosivas. Los misiles y los drones dependen de sofisticadas matemáticas, tales como la teoría del control. Pero Hardy estaba seguro de que su adorada teoría de números no tendría jamás un uso militar (al menos, no durante mucho tiempo) y estaba orgulloso de ello.

*

Hardy escribía en una época en la que un profesor titular de Cambridge pasaba unas cuatro horas al día investigando, tal vez alguna hora suelta dando clase y el resto del tiempo descansando para recargar sus pilas intelectuales. Veía el críquet y leía el periódico. Es de suponer que no se le ocurrió que incluso un investigador destacado en matemáticas podía emplear también todo ese tiempo libre para informar a los no especialistas acerca de lo que estaban haciendo los expertos. De ese modo, podría contribuir a las matemáticas *y* escribir acerca de ello. Que es lo que hacemos muchos de nosotros en la profesión hoy en día.

El argumento general de Hardy, de que una buena parte de las matemáticas «puras» no tienen aplicación directa y de que es probable que no la tengan nunca, es válido a menudo[43]. Sin embargo, como era de prever en cierto modo, desde el momento en que puso ejemplos concretos de temas sin utilidad, corría el riesgo de escoger justo los erróneos. Cuando dijo que era poco probable que la teoría de números y la relatividad sirviesen a ningún propósito militar durante muchos años, se equivocaba por completo (aunque hay que reconocerle que no descartó del todo usos así). El problema más difícil es decidir, por adelantado, qué ideas van a encontrar aplicación y cuáles no. Quien lo sepa resolver puede hacerse de oro. Sin embargo, son precisamente las ramas que *no* parecen ser aplicables a nada las que de repente pueden elevarse hasta la primera línea de la industria, del comercio y, desgraciadamente, de la guerra. Así ocurrió con la teoría de números. Más en

concreto, con el pequeño teorema de Fermat, que forma ahora la base de lo que se cree son códigos indescifrables.

La ironía es que dos años *antes* de que Hardy escribiera su apología, el director del MI6 había adquirido Bletchley Park, donde se iba a alojar la Escuela Gubernamental de Códigos y Cifrado (GC&CS, por sus siglas en inglés), el centro secreto de los aliados para descifrar claves durante la segunda guerra mundial. Es bien sabido que fue aquí donde equipos de criptógrafos lograron desentrañar el código Enigma que los alemanes empleaban durante la guerra, junto con otros muchos sistemas codificados del Eje. El integrante más conocido de Bletchley Park, Alan Turing, empezó su formación en 1938 y llegó a la residencia el día en que se declaró la guerra. Los criptógrafos del centro se sirvieron del ingenio y de las matemáticas para descifrar los códigos alemanes y las nociones de la teoría de números figuraban entre sus métodos. Cuarenta años después estaba en marcha una revolución en criptografía, basada con firmeza en la teoría de números y con importantes aplicaciones militares, así como civiles. Al poco se hizo imprescindible para el funcionamiento de internet. Hoy en día, dependemos de ella, en buena medida sin darnos cuenta de su existencia.

También la relatividad adquirió usos militares y civiles. Tuvo un efecto periférico en el Proyecto Manhattan para el desarrollo de una bomba atómica, como queda reflejado en el popular mito de que fue la famosa ecuación de Einstein $E = mc^2$ la que convenció a los físicos de que pequeñas porciones de materia encierran cantidades enormes de energía. En buena medida, fue una explicación *a posteriori* a la que se recurrió tras los ataques sobre Hiroshima y Nagasaki, para proporcionar a la población una manera fácil de entender cómo eran posibles estas armas. Incluso puede que su objetivo fuera desviar la atención pública del verdadero secreto: entender la física de las reacciones nucleares. Más recientemente y de manera más concreta, la precisión del Sistema de Posicionamiento Global, GPS, de navegación por satélite (capítulo 11), depende tanto de la relatividad especial como de la general para calcular las posiciones de forma correcta. Su desarrollo fue financiado por el Ejército de Estados Unidos, para cuyo uso se reservó en un principio.

Ejército dos, Hardy cero.

No culpo a Hardy. No tenía ni idea de lo que sucedía en Bletchley Park y apenas podía prever el rápido desarrollo de la informática ni de las comunicaciones digitales. En esencia, «digital» quiere decir trabajar con números naturales y eso es de lo que trata la teoría de números. De repente, los resultados obtenidos por generaciones de matemáticos teóricos

estimulados por pura curiosidad intelectual podían explotarse para obtener tecnologías innovadoras. Hoy en día, hay una cantidad enorme de matemáticas (no solo teoría de números, sino de todo tipo, desde combinatoria hasta álgebra abstracta y análisis funcional) incorporadas en los dispositivos electrónicos que una cuarta parte de la humanidad lleva en el bolsillo cada día. La confidencialidad de las transacciones online, sean de individuos, empresas o servicios de seguridad y militares, está garantizada por ingeniosas transformaciones matemáticas que hunden sus raíces en la adorada teoría de números de Hardy. Nada de esto habría sorprendido a Turing, que estaba tan adelantado a su tiempo que ya pensaba con seriedad en la inteligencia artificial en 1950. Pero Turing era un visionario. En ese entonces, ni siquiera se trataba de ciencia ficción. Era pura fantasía.

*

Un código es un método para convertir un mensaje escrito en lenguaje ordinario, un texto sin formato, en otro cifrado que parece no tener sentido. Es habitual que la conversión dependa de una clave, una pieza crucial de información que se mantiene en secreto. Por ejemplo, se dice que Julio César empleaba un código en el que cada letra del alfabeto se desplazaba tres posiciones. Aquí la clave es «tres». Este tipo de códigos de sustitución, en el que las letras del abecedario se transforman en otras de forma fija, pueden descifrarse con facilidad con un suministro adecuado de mensajes cifrados. Basta con saber las frecuencias con las que aparecen los elementos del alfabeto en el texto sin formato. Con eso ya puede hacerse un intento bastante bueno de adivinar el código. Al principio habrá unos pocos errores, pero si un segmento parece descifrarse como JULFO CÉSAR, no hace falta ser un genio para darse cuenta de que la F debería ser una I.

Aunque puede ser sencillo e inseguro, el código de César es un buen ejemplo de un principio general que, hasta hace poco, subyacía a la práctica totalidad de los sistemas de cifrado: es simétrico, lo que quiere decir que tanto el emisor como el receptor emplean en esencia la misma clave. Digo «en esencia» porque lo hacen de manera diferente: Julio desplaza el alfabeto tres espacios hacia delante, mientras que el destinatario lo hace tres espacios hacia atrás. No obstante, si se conoce la manera en que se ha empleado la clave para cifrar el mensaje, es fácil invertir el proceso y usar la misma para descifrarlo. Incluso algunos códigos muy sofisticados y seguros son simétricos. La

seguridad exige que *la* clave se mantenga en secreto para todo el mundo excepto para el emisor y el receptor.

Como dijo Benjamin Franklin, «tres pueden guardar un secreto si dos de ellos están muertos». En un código simétrico, hay al menos dos personas que necesitan conocer la clave, lo que en opinión de Franklin es una de más. En algún momento en 1944 o 1945, alguien (tal vez Claude Shannon, inventor de la teoría de la información) en los Laboratorios Bell de Estados Unidos, sugirió proteger las comunicaciones de voz frente a las escuchas no autorizadas al añadir ruido aleatorio a la señal y luego sustraerlo de nuevo cuando se recibe. Este método también es simétrico porque la clave es el ruido aleatorio y la sustracción invierte la adición. En 1970, James Ellis, un ingeniero en el Cuartel General de Comunicaciones del Gobierno de Reino Unido (GCHQ, por sus siglas en inglés), el sucesor de la GC&CS, se preguntó si el ruido podría generarse de manera matemática. De ser así, era al menos concebible que esto pudiera hacerse, no por la mera adición de señales, sino mediante algún proceso matemático que fuese muy difícil de invertir, incluso aunque se supiese cuál era. Por supuesto, el receptor tenía que poder hacerlo, pero esto se podía lograr empleando una *segunda* clave que solo conociese el destinatario.

Ellis denominó a esta idea «encriptación no secreta», aunque el término que se emplea en la actualidad es «criptografía de clave pública». Estas expresiones quieren decir que la regla para cifrar un mensaje puede desvelarse al público en general, porque sin el conocimiento de la segunda clave nadie puede descubrir cómo invertir el procedimiento y descifrar el texto. El único problema era que Ellis no fue capaz de concebir un método de encriptación adecuado. Buscaba lo que ahora se conoce como una función trampa: fácil de calcular pero difícil de invertir, algo así como caer en un trampilla. Sin embargo, como siempre, debía haber una segunda clave secreta que permitiese al receptor legítimo revertir el proceso con la misma facilidad, como una escalera oculta que se pudiese usar para salir de la trampa.

En esto llegó Clifford Cocks, un matemático británico que también trabajaba en el GCHQ. En septiembre de 1973, tuvo una idea genial. Consiguió hacer realidad el sueño de Ellis mediante el empleo de las propiedades de los números primos para crear una función trampa. Desde el punto de vista matemático, multiplicar dos o más de ellos entre sí es fácil. Puede hacerse a mano para dos primos de 50 dígitos, lo que da un resultado con 99 o 100 dígitos. A la inversa, partir de una cifra de estas dimensiones y encontrar sus factores primos, es mucho más difícil. El método tradicional de

«probar todas las posibilidades una tras otra» no es factible porque son demasiadas. Cocks concibió una función trampa basada en el producto de dos primos grandes, el resultado de multiplicarlos entre sí. El código resultante es tan seguro que esta multiplicación puede hacerse pública, si bien no los factores en sí mismos. Descifrarla requiere conocer los dos primos *por separado* y esa es la segunda clave secreta. A no ser que se conozcan ambos números, no hay manera de resolverlo. No basta con conocer solo su producto. Por ejemplo, supongamos que digo que he encontrado dos primos cuya multiplicación es

1 192 344 277 257 254 936 928 421 267 205 031 305 805 339 598 743 208 059
530 638 398 522 646 841 344 407 246 985 523 336 728 666 069

¿Es posible encontrar los números originales[44]? Un superordenador que sea rápido de verdad puede hacerlo, pero a un portátil le costaría. Si hubiese más dígitos, incluso el superordenador se quedaría atascado.

Sea como sea, Cocks tenía formación en teoría de números y concibió una manera de emplear un par de primos como estos para crear una función trampa. Explicaré cómo lo hizo un poco más adelante, cuando haya expuesto los conceptos necesarios. Era tan sencillo que al principio *ni siquiera lo puso por escrito*. Más adelante, incluyó los detalles en un informe a sus superiores. Sin embargo, a nadie se le ocurría una manera de emplear este método con los ordenadores rudimentarios de la época, así que se archivó. También se compartió con la Agencia de Seguridad Nacional de Estados Unidos. Ambas organizaciones veían el potencial militar, porque incluso aunque los cálculos fuesen lentos, podía emplearse este sistema de criptografía para enviar de manera electrónica a alguien la clave de otro código completamente diferente. De hecho, esta es la forma principal en la que este tipo de cifrado se emplea en la actualidad, tanto en aplicaciones civiles como militares.

Los burócratas británicos tienen un historial largo y poco distinguido de no darse cuenta de que tienen entre manos una gallina de los huevos de oro: la penicilina, el motor a reacción o el análisis del ADN. Sin embargo, en este caso pueden excusarse con la ley de la propiedad intelectual: para poder patentar algo hay que revelar lo que es. De un modo u otro, la idea revolucionaria de Cocks se archivó, un poco como la escena final de *En busca del arca perdida*, cuando la caja que contiene el Arca de la Alianza se esconde en las profundidades de un almacén gubernamental enorme y anónimo, lleno hasta los topes de cajas idénticas entre sí.

Mientras tanto, al llegar 1977, salió a la luz un método idéntico, redescubierto de manera independiente y publicado con rapidez por tres matemáticos estadounidenses: Ronald Rivest, Adi Shamir y Leonard Adleman. En la actualidad se denomina sistema criptográfico RSA en su honor. Por último, en 1997, los servicios de seguridad británicos desclasificaron la obra de Cocks, por lo que ahora se sabe que él lo descubrió primero.

*

La teoría de números se vuelve criptografía tan pronto como se comprende que cualquier mensaje puede representarse mediante un número. En el código de César, esta cifra es la posición de una letra en el alfabeto, que los matemáticos prefieren dar de 0 a 26 en lugar de 1 a 27 por motivos de conveniencia algebraica. De modo que la A es 0, la B es 1 y así hasta Z = 26. Los números que caigan fuera de este intervalo pueden convertirse a otra cantidad que sí esté dentro mediante suma o resta de múltiplos de 27. Este arreglo organiza las 27 letras en torno a un círculo, de modo que después de la Z se vuelve a la A. El código de César puede entonces reducirse a una sencilla regla matemática, de hecho, a la fórmula:

$$n \rightarrow n + 3$$

El proceso inverso tiene un aspecto muy similar:

$$n \leftarrow n + 3$$
o
$$n \rightarrow n - 3$$

Esto es lo que hace que el código sea simétrico.

Pueden inventarse códigos nuevos al cambiar las reglas, mediante modificaciones en las expresiones matemáticas. Solo se necesitan una manera sencilla de convertir un mensaje en un número y dos fórmulas: una para pasar del texto sin formato a texto cifrado y otra para recuperar el primero. Cada una de estas tiene que ser la inversa de la otra.

Hay muchas maneras de convertir un mensaje sin formato en un número. Una sencilla es emplear una cifra de 0 a 26 para cada letra y ponerlas todas juntas, con la salvedad de escribir el intervalo 0-9 como 00-09. De este modo,

JULIO se convierte en 0921110815 (recuérdese que A = 00). Tal vez se necesiten números adicionales para los espacios, los signos de puntuación o lo que sea. Una regla que convierte una cantidad en otra se denomina una función en teoría de números.

Disponer los números en torno a un círculo es un truco habitual de los teóricos de números, denominado aritmética modular. Escojamos una cantidad, en este caso 27. Ahora puede hacerse como si el 27 fuese equivalente al 0, de modo que las únicas cifras que se necesitan son las que van del 0 al 26. En 1801, Carl Friedrich Gauss señaló en sus famosas *Disquisitiones arithmeticae* que en un sistema así se pueden sumar, restar y multiplicar números y obedecer todas las reglas habituales del álgebra sin salir del intervalo elegido de 0 a 26. Basta con hacer las operaciones habituales con las cifras ordinarias y luego tomar el resto de dividir la solución por 27. De modo que, por ejemplo, $31 \times 7 = 217$, que es $8 \times 27 + 1$. El resto es 1, así que $31 \times 7 = 1$ en esta versión poco común de la aritmética.

Esta misma idea funciona igual si se sustituye 27 por cualquier otra cantidad. Esta se denomina *módulo* y puede escribirse (mod 27) para enfatizar lo que se ha hecho. Con más propiedad, el cálculo anterior es $31 \times 7 = 1 \pmod{27}$.

¿Qué ocurre con la división? Si se divide por 7 y no se presta mucha atención a lo que eso significa, resulta

$$31 = 1/7 \pmod{27}$$

de modo que dividir por 7 es lo mismo que multiplicar por 31. Es posible establecer ahora una nueva regla del código:

$$n \rightarrow 31n \pmod{27}$$

cuya inversa es

$$n \leftarrow 7n \pmod{27}$$

Esta regla desordena el alfabeto de manera considerable, que queda

AEIMPTXBFJNQUYCGKÑRVZDHLOSW

No deja de ser un código de sustitución a nivel de letras individuales, así que puede descifrarse con facilidad, pero demuestra que es posible cambiar la fórmula. También ilustra el empleo de la aritmética modular, la clave para áreas muy amplias de la teoría de números.

No obstante, la división puede ser más complicada. Dado que $3 \times 9 = 27 = 0 \pmod{27}$, no se puede dividir por 3. De otro modo, se deduciría que $9 = 0/3 = 0 \pmod{27}$, lo que es erróneo. La regla general es que puede dividirse por cualquier número que no comparta factores primos con el módulo. Así que 0 queda descartado. Pero eso no es ninguna sorpresa: tampoco es posible dividir números naturales ordinarios por 0. Si el módulo es un primo, puede dividirse por cualquier cantidad menor que el módulo, excepto 0.

La ventaja de la aritmética modular es que proporciona una estructura algebraica a la lista de «palabras» del texto sin formato. Esto hace posible una gran variedad de reglas para transformar el mensaje inicial en otro cifrado y a la inversa. Lo que hizo Cocks, y Rivest, Shamir y Adleman después de él, fue elegir una fórmula muy ingeniosa.

Cifrar una a una las letras de un mensaje y emplear el mismo número para cada una no resulta particularmente seguro: sea cual sea la regla, se trata de un código de sustitución. Pero si se divide el texto en bloques, pongamos de diez caracteres de longitud, o más bien en torno a la centena como es el caso en la actualidad, y se convierte cada uno de ellos en una cifra, se tiene un código de sustitución por bloques. Si estos son lo bastante largos, ya no se repetirán con un patrón de frecuencia distintivo, por lo que no será posible descifrar el mensaje mediante la observación de los números que aparecen más a menudo.

*

Cocks y RSA derivaron sus reglas del hermoso teorema que Fermat descubrió en 1640, que dice cómo se comportan las *potencias* de los números en la aritmética modular. En notación moderna, Fermat le dijo a su amigo de Bessy que, si n es primo, entonces

$$a^n = a \pmod{n}$$

o de manera equivalente

$$a^{n-1} = 1 \pmod{n}$$

para cualquier cantidad *a*. «Le mandaría la demostración, si no temiese que fuera demasiado larga», escribió. Euler aportó la demostración que faltaba en 1736 y en 1763 publicó un teorema más general que es válido cuando el módulo no es primo. Ahora, *a* y *n* no deben tener factores comunes y la potencia $n - 1$ en la segunda versión de la fórmula se sustituye por la función «indicatriz» de Euler, $\varphi(n)$. No hace falta ahora saber cuál es[45], pero sí que si $n = pq$ es el producto de dos números primos *p* y *q*, entonces $\varphi(n) = (p - 1)(q - 1)$.

El sistema criptográfico RSA procede de este modo:

- Se eligen dos números primos *p* y *q* grandes.
- Se calcula el producto $n = pq$.
- Se calcula $\varphi(n) = (p - 1)(q - 1)$ y se guarda el resultado en secreto.
- Se escoge un número *e* que no tenga factores primos en común con $\varphi(n)$.
- Se calcula *d*, tal que $de = 1 \pmod{\varphi(n)}$.
- El número *e* puede hacerse público (esto aporta muy poca información útil acerca de $\varphi(n)$, por cierto).
- Se guarda *d* en secreto (esto es fundamental).
- Sea *r* un mensaje de texto sin formato, cifrado como un número de módulo *n*.
- Se convierte *r* al texto cifrado $r^e \pmod{n}$ (esta regla también puede hacerse pública).
- Para descifrar r^e, se eleva a la potencia $d \pmod{n}$ (sin olvidar que *d* es secreto). Esto da $(r^e)d$, que es equivalente a r^{ed}, que es igual a *r* por el teorema de Euler.

Aquí, la regla de cifrado es «tómese la potencia de índice *e*»:

$$r \rightarrow r^e$$

y la de descifrado es «tómese la potencia de índice *d*»:

$$s \rightarrow s^d$$

Algunos trucos matemáticos, en los que no voy a entrar, hacen que sea posible realizar todos estos pasos con rapidez (en los ordenadores de hoy en día), siempre que se conozcan *p* y *q por separado*. Para remate, si no se sabe

cuáles son estos, tener *n* y *e* no ayuda mucho a calcular *d*, algo necesario para descifrar el mensaje. En resumen, hace falta encontrar los factores primos *p* y *q* de *n*, algo que, como se ha visto, es (en apariencia) mucho más difícil que multiplicar *p* y *q* para obtener *n*.

En otras palabras, «elevar a la potencia *e*» es la función trampa necesaria.

En la actualidad, todo esto puede hacerse en un minuto más o menos en un ordenador portátil para, pongamos, números primos *p* y *q* de 100 dígitos. Una característica conveniente del sistema RSA es que, aunque los procesadores se hagan más potentes, bastará con tomar *p* y *q* mayores. Sirve el mismo método.

Una desventaja es que el RSA, si bien es práctico del todo, es demasiado lento como para emplearlo de manera rutinaria en el contenido completo de cada mensaje. La principal aplicación en la vida real es emplearlo como una forma segura de transmitir una clave secreta para algún sistema de cifrado completamente diferente, uno que sea mucho más rápido de aplicar y que sea seguro siempre que nadie conozca la clave. De modo que el sistema RSA resuelve el problema de compartir esta última, algo que ha sido un dolor de cabeza para la criptografía desde sus inicios. Un motivo por el que se pudo descifrar Enigma fue porque ciertos ajustes de la máquina de encriptado se distribuían a los operarios al principio de cada jornada de manera poco segura. Otra aplicación común es verificar una firma electrónica, es decir, un mensaje de código que establece la identidad del emisor.

El jefe de Cocks, Ralph Benjamin, director científico, de ingeniería y director superintendente de GCHQ, estaba muy atento y se dio cuenta de esta posibilidad. Escribió en un informe: «Consideré que era más importante para uso militar. En una situación bélica fluida pueden surgir amenazas u oportunidades imprevistas. Si es posible compartir la clave de manera rápida y electrónica, se dispone de una ventaja fundamental sobre el adversario». Sin embargo, los ordenadores del momento no eran capaces de realizar la tarea y el Gobierno británico dejó pasar lo que, *a posteriori*, era una gran oportunidad.

*

Pocas veces hay técnicas matemáticas «listas para usar» en la resolución de problemas prácticos. Como todo lo demás, en general necesitan adaptarse y modificarse para superar varias dificultades. Lo mismo es válido para el sistema RSA: no es tan sencillo como lo acabo de describir. De hecho, surgen

una serie de interrogantes teóricos fascinantes para los matemáticos tan pronto como se deja de admirar la idea y se reflexiona sobre lo que puede salir mal.

No es difícil demostrar que calcular $\varphi(n)$ sin conocer sus factores primos p y q es tan complejo como encontrar los propios p y q. De hecho, parece que esa es la única manera de hacerlo. De modo que la pregunta principal es: ¿qué dificultad tiene la factorización? La mayoría de los matemáticos piensan que esta es extrema en un sentido técnico: cualquier algoritmo de factorización tiene un tiempo de ejecución que crece de manera explosiva con la cantidad de dígitos en el producto pq. (A propósito, el motivo por el que se emplean solo dos primos, en lugar de, pongamos, tres, es que este es el caso más difícil. Cuantos más factores tenga un número, más fácil es dar con uno de ellos. Si se divide por este, la cantidad resultante es mucho menor, de modo que es más fácil encontrar también los demás). No obstante, nadie puede *demostrar* en la actualidad que la factorización en primos sea compleja. No se tiene ni idea de cómo empezar a hacerlo. De modo que la seguridad del método RSA descansa en una conjetura sin demostración.

El resto de interrogantes y dificultades se refieren a detalles más rebuscados del método. Una mala elección de los números puede hacer que el sistema RSA sea vulnerable ante ataques astutos. Por ejemplo, si e es demasiado pequeño, entonces puede determinarse el mensaje r al calcular la raíz de índice e del texto cifrado re, considerado como un número ordinario, es decir, no de módulo n. Otro fallo potencial surge si se envía el mismo mensaje a e receptores empleando la misma potencia e, incluso si p y q son diferentes para cada uno de ellos. Es posible aplicar un resultado encantador, denominado teorema chino del resto, para desvelar el texto sin formato.

Tal y como se ha descrito, el sistema RSA es, además, inseguro desde el punto de vista semántico, lo que quiere decir que en principio puede descubrirse la clave mediante la codificación de muchos textos sin formato diferentes e intentando emparejar los resultados con el cifrado que se quiere descodificar. En resumidas cuentas, por prueba y error. Tal vez esto no sea factible para mensajes largos, pero puede llegar a serlo si se envían muchos cortos. Para evitarlo, se modifica el método RSA y se añaden dígitos adicionales al mensaje, conforme a algún esquema concreto pero aleatorio. Esto hace que el texto sin formato sea más largo y evita que se envíe lo mismo muchas veces.

Otro método para descifrar un código RSA aprovecha no un fallo matemático, sino una característica física del ordenador. En 1995, el

emprendedor en criptografía Paul Kocher observó que si el descifrador sabe lo suficiente acerca del hardware que se emplea y puede medir cuánto se tarda en descodificar varios mensajes, entonces es capaz de deducir con facilidad la clave secreta *d*. Dan Boneh y David Brumley demostraron una versión práctica de este ataque en 2003 para mensajes enviados a través de una red convencional mediante el protocolo estándar SSL (*Secure Sockets Layer*).

La existencia de métodos matemáticos que pueden *en ocasiones* factorizar un número grande muy rápido implica que los primos *p* y *q* deben escogerse para satisfacer algunas condiciones restrictivas. No deberían estar uno cerca del otro, o sería válido un método de factorización que se remonta a Fermat. En 2012, un grupo dirigido por Arjen Lenstra intentó hacer esto con millones de claves públicas extraídas de internet y fueron capaces de descifrar una de cada 500.

Un ordenador cuántico que funcionase en la práctica cambiaría las reglas del juego. Estas máquinas, que todavía están en pañales, emplean bits cuánticos en lugar de los habituales dígitos binarios 0 y 1 y en principio pueden realizar cálculos gigantescos, tales como factorizar números enormes, con velocidades sin precedente. Dejaré una consideración más detallada para más adelante en este capítulo.

*

El sistema RSA es solo uno más de una serie de códigos basados en la teoría de números o en su pariente cercana, la combinatoria, un método para contar de cuántas maneras puede lograrse alguna disposición sin hacer un listado de todas las posibilidades. Para mayor convencimiento de que el manantial matemático todavía no se ha agotado en lo que se refiere a la criptografía, describiré un sistema de cifrado alternativo, que aprovecha una de las ramas más profundas y excitantes de la teoría de números actual. Trata de «curvas elípticas», que, entre otras cosas, son cruciales para la épica demostración de Andrew Wiles del último teorema de Fermat.

La teoría de números ha avanzado mucho desde la época de Fermat y Euler. También lo ha hecho el álgebra, cuyo énfasis se ha desplazado desde la representación mediante signos de números desconocidos hasta las propiedades genéricas de sistemas simbólicos definidos por reglas concretas. Estos dos campos de investigación se solapan de manera sustancial. Algunas ideas fascinantes sobre códigos secretos han surgido a partir de una combinación de dos ramas especiales del álgebra y de la teoría de números:

cuerpos finitos y curvas elípticas. Para comprender lo que implican, primero hace falta saber lo que son.

Ya se ha visto que en la aritmética en un módulo cualquiera es posible sumar, restar y multiplicar «números» al mismo tiempo que se obedecen las reglas algebraicas habituales. Para evitar perder el hilo, no dije cuáles eran estas reglas, pero los ejemplos habituales son la propiedad conmutativa $ab = ba$ y la asociativa $(ab)c = a(bc)$. Ambas son válidas para la multiplicación y hay otras similares para la suma. También se cumple la propiedad distributiva $a(b + c) = ab + ac$ y hay reglas sencillas que se aplican a 0 y a 1, tales como $0 + a = a$ y $1a = a$. Cualquier sistema que obedezca estas leyes se denomina *anillo*. Si también es posible la división (excepto por 0) y se cumplen las propiedades habituales, se obtiene un *cuerpo*. Estos nombres son los tradicionales, importados del alemán, y en resumidas cuentas se refieren a «algún tipo de conjunto de cosas que obedece las reglas especificadas». Los números enteros módulo 27 forman un anillo, conocido como $\mathbb{Z}_{27}$. Ya se ha visto que hay problemas cuando se divide por 3, así que no es un cuerpo. Dije (sin explicar por qué) que los enteros en módulo un número primo no presentan tales inconvenientes, de modo que $\mathbb{Z}_2$, $\mathbb{Z}_3$, $\mathbb{Z}_5$, $\mathbb{Z}_7$ y demás (números enteros módulo 2, 3, 5 y 7) son todos cuerpos.

Los números naturales ordinarios no se acaban nunca: forman un conjunto infinito. Por el contrario, sistemas como $\mathbb{Z}_{27}$ y $\mathbb{Z}_7$ son finitos. El primero comprende solo las cantidades del 0 al 26 y el segundo del 0 al 6. El primero es un anillo finito y el segundo es un cuerpo finito. En realidad, es extraordinario que haya sistemas acotados que puedan obedecer tantas reglas del álgebra sin inconsistencias lógicas. Los conjuntos limitados de números, si no son demasiado grandes, se adaptan muy bien a los cálculos por ordenador porque pueden realizarse estos de manera *exacta*. Por lo tanto, no es sorprendente que haya una variedad de códigos basados en cuerpos finitos. No solo criptográficos, para mantener el secreto, sino otros para detectar y corregir fallos, o para garantizar que los mensajes se reciben sin errores debidos a «ruido» aleatorio, tal como interferencias eléctricas. Hay toda una rama novedosa de las matemáticas, la teoría de códigos, que aborda estas cuestiones.

Los cuerpos finitos más sencillos son $\mathbb{Z}_p$, los enteros módulo un primo p. Fermat ya sabía que forman un cuerpo (aunque no con esa denominación). El revolucionario francés Évariste Galois, muerto en un trágico duelo a la edad de veinte años, demostró que estos no son los únicos. De hecho, encontró todos los que hay: uno por cada potencia de primo p^n y contiene p^n

«números» diferentes de manera exacta (atención: si n es mayor que 1, este cuerpo no son los números enteros módulo p^n). De modo que hay infinitos cuerpos con 2, 3, 4, 5, 7, 8, 9, 11, 13, 16, 17, 19, 23, 25… elementos, pero no con 1, 6, 10, 12, 14, 15, 18, 20, 21, 22, 24… Un teorema muy curioso.

Las curvas elípticas (que están relacionadas solo de manera muy indirecta con las elipses) tuvieron su origen en una rama diferente: la teoría clásica de números. En torno al año 250, el matemático de la Grecia antigua Diofanto de Alejandría escribió un texto acerca de la solución de ecuaciones algebraicas mediante el empleo de números naturales (o racionales). Por ejemplo, el famoso triángulo con lados de longitud 3, 4 y 5 tiene un ángulo recto gracias a Pitágoras, porque $3^2 + 4^2 = 5^2$. Por lo tanto, estas cantidades son soluciones de la ecuación pitagórica $x^2 + y^2 = z^2$. Uno de los teoremas de Diofanto demuestra cómo encontrar todas las soluciones de esta expresión en fracciones y, en particular, en naturales. Esta rama general, resolver ecuaciones en números racionales, pasó a conocerse como *ecuaciones diofánticas*. La restricción a los racionales cambia las reglas del juego. Por ejemplo, $x^2 = 2$ puede resolverse en números reales, pero no en racionales.

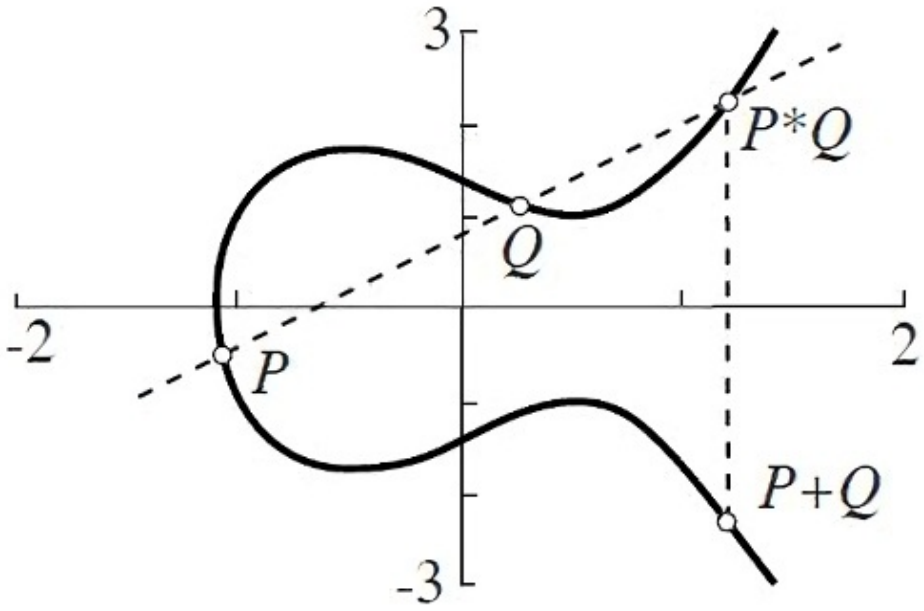
Uno de los problemas de Diofanto es: «Descomponer un número dado en dos números cuyo producto sea un cubo menos su lado». Si la cantidad original es a, se descompone en Y y $a - Y$, y se quiere resolver

$$Y(a - Y) = X^3 - X$$

Diofanto examinó el caso cuando $a = 6$. Un cambio adecuado de variables (restar 9, cambiar Y a $y + 3$ y X a $-x$) convierte esta ecuación en

$$y^2 = x^3 - x + 9$$

Obtuvo entonces la solución $X = 17/9$, $Y = 26/27$.



Para «sumar» dos puntos *P* y *Q* en una curva elíptica, se unen entre sí con una recta que corta la curva en un tercer punto *P*Q*. Después se proyecta en el eje x para obtener *P* + *Q*.

Cabe destacar que aparecieron ecuaciones similares en geometría cuando los matemáticos intentaron emplear el análisis (cálculo avanzado) para obtener la longitud de arco de un segmento de elipse. De hecho, así es como surgió el nombre de «curva elíptica». Sabían cómo resolver el problema análogo para una circunferencia mediante el empleo del cálculo. La cuestión en ese caso se reduce a encontrar la integral de una función que incluye la raíz cuadrada de un polinomio cuadrático y esto puede hacerse mediante razones trigonométricas (inversas). El mismo método aplicado a una elipse lleva a la integral de una función que incluye la raíz cuadrada de un polinomio cúbico y, tras algunos experimentos infructuosos, se hizo evidente que se necesitaba una nueva clase de funciones. Resultó que estas eran bastante bonitas, aunque complicadas, y se denominaron curvas elípticas por su conexión con la longitud del arco de una elipse. La raíz cuadrada de un polinomio cúbico es una solución *y* de la ecuación

$$y^2 = x^3 + ax + b$$

(cualquier término x^2 a la derecha puede transformarse en cero). En geometría de coordenadas esta expresión define una curva en el plano, de modo que esta (y su versión algebraica como ecuación) pasó a llamarse «curva elíptica».

Cuando los coeficientes son enteros, puede considerarse la ecuación en aritmética modular, pongamos en $\mathbb{Z}_7$. Cada solución en números enteros

ordinarios lleva a otra en aritmética de módulo 7. Dado que este sistema es finito, es posible resolverla por prueba y error. Para la ecuación diofántica $y^2 = x^3 - x + 9$ se descubre enseguida que las únicas soluciones (mod 7) son:

$$x = 2,\ y = 2,\ x = 2,\ y = 5,$$
$$x = 3,\ y = 1,\ x = 3,\ y = 6,$$
$$x = 4,\ y = 3,\ x = 4,\ y = 4.$$

Estos resultados tienen implicaciones para cualquier solución en números enteros ordinarios: debe reducirse en módulo 7 a una de estas seis combinaciones. Lo mismo es válido para racionales, siempre que el denominador no sea un múltiplo de 7 (estos están prohibidos porque en $\mathbb{Z}_7$ una cantidad así se convierte en 0). Si se cambia 7 por otra cifra, se obtiene más información acerca de la forma de cualquier solución racional hipotética.

Ahora se tienen curvas (ecuaciones) elípticas sobre anillos y cuerpos finitos. La representación geométrica de una curva no es aplicable en realidad, dado que solo se dispone de un conjunto finito de puntos, pero es conveniente emplear el mismo nombre. La imagen muestra una forma habitual y una característica adicional, conocida por Fermat y Euler, que intrigó a los matemáticos de principios del siglo XX. Dadas dos soluciones, es posible «sumarlas» para obtener otra, según se indica. Si son números racionales, también lo es la suma. No es solo una oferta «3 × 2», sino «muchísimas más gratis × 2», porque puede repetirse la operación. En ocasiones, esto lleva de vuelta al punto de partida, pero la mayoría de las veces genera una cantidad infinita de soluciones diferentes. De hecho, estas tienen una estructura algebraica elegante: forman un grupo, el grupo de Mordell-Weil de la curva elíptica. Louis Mordell demostró sus propiedades fundamentales y André Weil las generalizó, mientras que «grupo» quiere decir que la suma obedece a una lista breve de reglas sencillas. Es conmutativo, lo que significa que $P + Q = Q + P$, algo evidente a partir de la imagen dado que la recta que pasa a través de P y Q es también la que pasa por Q y P. La existencia de una estructura de grupo como esta es poco habitual. La mayoría de las ecuaciones diofánticas son menos agradecidas. Muchas no tienen solución en absoluto, algunas solo unas pocas y es difícil saber por adelantado de qué tipo es la que se tiene entre manos. De hecho, las curvas elípticas son objeto de intensas investigaciones, por este y por otros motivos. Andrew Wiles demostró una conjetura profunda acerca de curvas elípticas como un paso crucial en su demostración del último teorema de Fermat.

*

La estructura de grupo de una curva elíptica es de interés también para los criptógrafos. Es habitual considerarla como una forma de «sumar» soluciones, aunque la fórmula es mucho más complicada porque es conmutativa, y el símbolo + se ha popularizado en la teoría de grupos conmutativos. En particular, si se tiene una solución (x, y), que puede entenderse como un punto P en un plano, entonces es posible generar otras $P + P$, $P + P + P$ y así en adelante. Resulta natural denominarlas $2P$, $3P$, etcétera.

En 1985, Neal Koblitz y Victor Miller se dieron cuenta de manera independiente de que puede emplearse la ley de grupo de una curva elíptica para crear un código. La idea es trabajar en un cuerpo finito cualquiera con una cantidad elevada de elementos. Para cifrar P, se calcula kP para algún número entero k muy grande, algo fácil de hacer con un ordenador, y el resultado se designa por Q. Para invertir este proceso se parte de Q y se obtiene P (mediante división, de hecho, por k). Debido a la complejidad de la fórmula de grupo, este cálculo inverso es muy difícil, con lo que se ha inventado un nuevo tipo de función trampa y por tanto un sistema criptográfico de clave pública. Se conoce como criptografía de curva elíptica, o ECC, por sus siglas en inglés. Al igual que pueden emplearse muchos números primos diferentes para aplicar el método RSA, es posible tomar una gran cantidad de curvas elípticas distintas para el sistema ECC, sobre muchos cuerpos finitos, con varias opciones de P y del multiplicador k. De nuevo, hay una clave secreta que permite un descifrado rápido.

La ventaja resulta estar en que un grupo más pequeño genera un código que es igual de seguro que uno del sistema RSA, basado en números primos mucho mayores. De modo que el obtenido por curvas elípticas es más eficaz. Cifrar y descifrar un mensaje es más rápido y sencillo siempre que se conozca una clave secreta. Y es igual de difícil de descodificar si no se dispone de esta. En 2005, la Agencia de Seguridad Nacional de Estados Unidos recomendó que la investigación en criptografía de clave pública se trasladase a la nueva rama de las curvas elípticas.

Al igual que para el método RSA, no hay una demostración rigurosa de que el sistema ECC sea seguro. El arsenal de ataques posibles es parecido también al de aquel.

En este momento, hay un gran interés en las criptomonedas, que son sistemas monetarios que no están controlados por bancos convencionales, si bien estos también se han mostrado interesados. No se les escapa una cuando

se trata de hacer dinero. La más conocida de ellas es Bitcoin. Su seguridad está garantizada por una técnica denominada cadena de bloques (*blockchain*), que es un registro cifrado de todas las transacciones que implican a esa «moneda» en concreto. Se crean nuevos bitcoines mediante «minería», que en esencia significa realizar un gran número de cálculos que por otro lado carecen de sentido. Esta actividad consume cantidades enormes de electricidad sin otro propósito práctico que enriquecer a unos pocos individuos. En Islandia, donde la luz es muy barata gracias a la producción geotérmica a partir de vapor subterráneo, la minería de bitcoines emplea más energía que los hogares de toda la población juntos. Es difícil saber cómo puede contribuir esta actividad a combatir el calentamiento global y la emergencia climática, pero bueno.

Bitcoin y muchas otras criptomonedas emplean una curva elíptica particular, conocida por el pegadizo nombre de secp256k1. Su ecuación, $y^2 = x^3 + 7$, es mucho más llamativa y esta parece ser la principal razón por la que se escogió. El cifrado a través de secp256k1 se basa en un punto en la curva dado por

$x = 55066263022277343669578718895168534326250603453777594175500$
187360389116729240
$y = 32670510020758816978083085130507043184471273380659243924$
$32759389043357573374824 24$

Esto ilustra los números enteros gigantescos implicados en la aplicación práctica del sistema ECC.

*

He dicho varias veces que la seguridad del sistema RSA se basa en la presunción no demostrada de que la factorización en primos es muy difícil. Incluso si eso fuese cierto, y es muy probable que lo sea, podría haber otras formas de comprometer la invulnerabilidad del código y lo mismo puede decirse de todos los métodos criptográficos de clave pública clásicos. Una manera potencial de que esto ocurra es que alguien invente un procesador que sea mucho más rápido que cualquiera de los disponibles en la actualidad. Hoy en día, este nuevo tipo de amenaza a la seguridad se yergue en el horizonte: los ordenadores cuánticos.

Un sistema físico clásico existe en un estado concreto. Una moneda sobre una mesa muestra la cara o la cruz. Un interruptor está encendido o apagado.

Un dígito binario (o «bit») en la memoria de un ordenador es 0 o 1. Pero los sistemas cuánticos no son así. Un objeto cuántico es una onda y las ondas pueden ponerse unas encima de otras, lo que de manera técnica se conoce como una superposición. El estado de una superposición es una combinación de los de sus componentes. El conocido (de hecho, famoso) caso del gato de Schrödinger es un ejemplo gráfico: mediante alguna diablura hecha con un átomo radiactivo, un frasco de gas venenoso y un felino en una caja hermética, el estado cuántico del desdichado animal es una superposición de «vivo» y «muerto». Un gato clásico debe estar en uno u otro, pero un minino cuántico puede estar en los dos al mismo tiempo.

Hasta que se abre la caja.

Entonces la función de onda del animal «colapsa» a solo uno de los estados clásicos. O está vivo o está muerto. La curiosidad (abrir la caja) mató al gato. O no.

No quiero entrar en el controvertido y a menudo acalorado debate acerca de si los estados cuánticos de los felinos se comportan en realidad de este modo[46]. Lo que importa aquí es que la física matemática funciona a las mil maravillas para objetos más sencillos y que se emplea ya para fabricar procesadores cuánticos rudimentarios. En lugar del bit, un 0 o un 1, está el bit cuántico, o cúbit, que es 0 y 1 al mismo tiempo. Un ordenador clásico, del tipo de los que tiene todo el mundo en sus escritorios, mochilas o bolsillos, gestiona la información como una secuencia de 0 y 1. En realidad, emplea efectos cuánticos para hacerlo, hasta tal punto es pequeño el tamaño de los circuitos actuales, pero la conclusión es que los cálculos corresponden a la física clásica. Los ingenieros que construyen ordenadores convencionales se esfuerzan mucho para garantizar que un 0 no deja de ser un 0, un 1 no deja de ser un 1 y que no se mezclan, como el agua y el aceite. El gato clásico debe estar vivo o muerto. De modo que una cadena de (pongamos) 8 bits puede almacenar una única secuencia, como 01101101 o 10000110.

En un ordenador cuántico sucede exactamente lo contrario. Un registro de 8 cúbits es capaz de retener ambas secuencias al mismo tiempo, junto con todas las otras 254 combinaciones posibles de 8 bits. Es más, puede hacer sumas con las 256 variaciones *de manera simultánea*. Es como tener 256 ordenadores en lugar de uno. Cuanto más largas sean las secuencias, mayor es el número de posibilidades que aprovechan. Una cadena de 100 bits puede almacenar una única secuencia de 100 bits. Un registro de 100 cúbits es capaz de retener y manejar todas las 10^{30} secuencias posibles de 100 bits. Se trata de «procesamiento en paralelo» a una escala enorme y es por lo que muchas

personas están entusiasmadas con la computación cuántica. En lugar de hacer los 10^{30} cálculos uno a uno, pueden hacerse *todos al mismo tiempo*.

En principio.

En la década de 1980, Paul Benioff propuso un modelo cuántico de la máquina de Turing, la formulación teórica de la informática clásica. Poco después, el físico Richard Feynman y el matemático Yuri Manin señalaron que un ordenador de este tipo podría ser capaz de realizar cantidades enormes de cálculos en paralelo. Un avance fundamental en el aspecto teórico se produjo en 1994, cuando Peter Shor inventó un algoritmo cuántico muy rápido para hallar los factores primos de números grandes. Esto demuestra que el sistema de criptografía RSA es vulnerable en potencia ante ataques de un enemigo que utilice un ordenador cuántico. Pero, lo que es más importante, demuestra que un algoritmo así puede superar con mucho el rendimiento de uno clásico para un problema razonable y no rebuscado.

En la práctica, los obstáculos para la construcción de un procesador de este tipo que funcione son enormes. Alteraciones minúsculas debidas a fuentes externas, o incluso la vibración molecular que se conoce como calor, pueden hacer que un estado superpuesto deje de ser «coherente», es decir, que se deshaga, muy rápido. Para paliar este problema, en la actualidad la máquina debe enfriarse hasta muy cerca del cero absoluto, a − 273 °C, lo que requiere un suministro de helio-3 que es un producto secundario de las reacciones nucleares muy escaso. Ni siquiera esto puede impedir que se produzca la decoherencia, solo la retrasa. De modo que todos los cálculos deben hacer gala de un sistema de corrección de errores que identifique las alteraciones debidas a fuentes externas y devuelva los estados de los cúbits a los que deberían ser. El teorema del umbral cuántico afirma que esta técnica funciona siempre que se puedan rectificar los fallos más rápido de lo que la decoherencia los produce. Como estimación aproximada se considera que la tasa de error para cada puerta lógica debe ser como mucho de un 1 por 1000.

La corrección de errores también tiene un coste: requiere más cúbits. Por ejemplo, para factorizar mediante el algoritmo de Shor un número que pueda almacenarse en n bits cuánticos, el cálculo se ejecuta en un tiempo que es más o menos proporcional a una cantidad entre n y n^2. Con la rectificación de los fallos, que en la práctica es imprescindible, se convierte en algo cercano a n^3. Para un número de 1000 cúbits, este proceso multiplica el tiempo de ejecución por mil.

Hasta hace muy poco, nadie había construido un ordenador cuántico que tuviese más que unos pocos cúbits. En 1998, Jonathan A. Jones y Michele

Mosca usaron un dispositivo de dos bits cuánticos para resolver el problema de Deutsch, que surge a raíz de un trabajo de David Deutsch y Richard Jozsa en 1992. Es un algoritmo cuántico que se ejecuta exponencialmente más rápido que cualquier otro convencional, y alcanza siempre una solución que es correcta en todas las ocasiones. El problema que resuelve es este. Hay un dispositivo hipotético, un *oráculo,* que implementa alguna función booleana: una que convierte cualquier cadena de bits de *n* dígitos en un 0 o en un 1. Desde el punto de vista matemático, el oráculo *es* esa función. Se sabe también que la función booleana toma, o bien un valor de 0 para todas las entradas, o de 1 para todas las entradas, o de 0 justo para la mitad de las cadenas de bits y de 1 para la otra mitad. La cuestión es determinar cuál de estos tres casos es el que se produce, mediante la aplicación de la función a cadenas de bits y la observación del resultado. El problema de Deutsch es artificial de manera deliberada, la demostración de un concepto, más que algo práctico. Su mérito reside en que trae a colación un asunto concreto en el que un algoritmo cuántico sobrepasa con toda probabilidad a cualquier otro convencional. Desde el punto de vista técnico, demuestra que la clase de complejidad EQP (soluciones exactas de tiempo polinómico en un ordenador cuántico) es diferente de la clase P (soluciones exactas de tiempo polinómico en un ordenador clásico).

El año 1998 también fue testigo de una máquina de tres cúbits y en 2000 aparecieron otras con cinco y siete. En 2001, Lieven Vandersypen y sus colaboradores[47] implementaron el algoritmo de Shor empleando como bits cuánticos siete núcleos de espín 1/2 en una molécula sintetizada de manera específica, que podían manipular con técnicas de resonancia magnética nuclear de estado líquido a temperatura ambiente, para encontrar los factores primos del número entero 15. La mayoría de las personas pueden hacerlo de cabeza, pero fue una importante demostración conceptual. Para 2006, los investigadores habían alcanzado 12 cúbits, mientras que una empresa llamada D-Wave afirmaba haber llegado a 28 en 2007.

*

Mientras sucedía esto, los investigadores aumentaron mucho el tiempo que podía conservarse un estado cuántico antes de perder la coherencia. En 2008, se almacenó un cúbit durante más de un segundo en un núcleo atómico. Para 2015, la duración era de seis horas. Es difícil comparar estos tiempos porque dispositivos diferentes emplearon métodos cuánticos distintos, pero el

progreso era impresionante. En 2011, D-Wave anunció que había construido un ordenador cuántico disponible comercialmente, el D-Wave One, con un procesador de 128 cúbits. Para 2015, la empresa afirmó haber superado los 1000 cúbits.

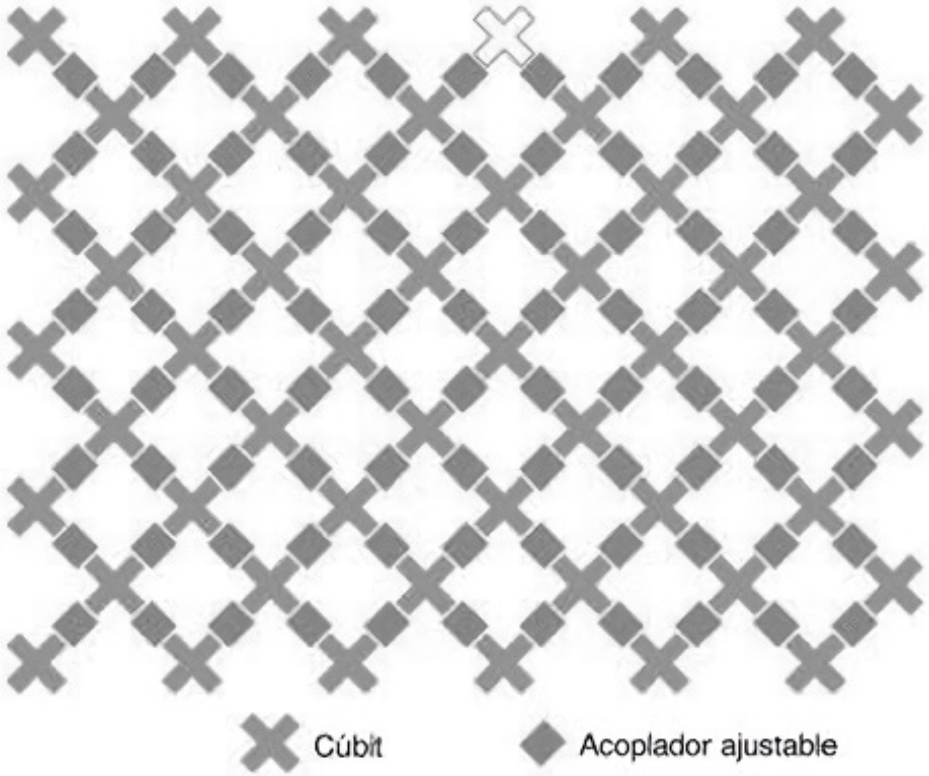
La primera reacción ante estos anuncios fue el escepticismo. La arquitectura del dispositivo no era la habitual y hubo quien cuestionó si era un ordenador cuántico de verdad y no uno clásico alambicado que empleaba artilugios relacionados con la cuántica. En las pruebas superó el rendimiento de los aparatos comerciales en tareas útiles, pero había sido diseñado de manera específica para estas, mientras que los ordenadores clásicos contra los que competía no. Sus ventajas parecieron desaparecer cuando los dispositivos convencionales se diseñaron de manera específica para las tareas en cuestión. La controversia continúa, pero las máquinas de D-Wave están en uso y funcionan bien.

Un objetivo principal de la investigación es la supremacía cuántica: construir un dispositivo cuántico que supere el rendimiento de los mejores ordenadores clásicos en al menos un cálculo. En 2019, un equipo de Google AI publicó un artículo en *Nature* con el título «*Quantum supremacy using a programmable superconducting processor*»[48]. Anunciaron que habían construido un procesador cuántico llamado Sycamore con 54 cúbits, pero que uno había fallado y reducido el número a 53. Lo habían empleado para resolver en 200 segundos un problema en el que en un ordenador convencional habría tardado 10 000 años.

Enseguida se puso en duda esta afirmación a partir de dos cuestiones. Una era que el cálculo se podría haber hecho con toda probabilidad en menos tiempo en un procesador clásico. La otra era que el problema que había resuelto Sycamore era bastante rebuscado: tomar muestras de la salida de un circuito cuántico pseudoaleatorio. El diseño del circuito conecta componentes de manera aleatoria con el fin de calcular la distribución de probabilidad de una muestra de las posibles salidas. Resulta que algunas de estas son mucho más probables que otras, de modo que esta distribución es muy complicada y no es uniforme. Un cálculo clásico crece de manera exponencial con el número de cúbits. Sea como sea, el equipo tuvo éxito en su objetivo principal: demostrar que no hay obstáculos prácticos a la creación de un ordenador cuántico que pueda superar a los convencionales en *algo*.

Una duda que surge enseguida es: ¿cómo se sabe si la respuesta es correcta? Es imposible esperar 10 000 años para que un ordenador clásico resuelva el problema y no puede darse la solución por válida sin más antes de

comprobarla. El equipo resolvió esto mediante el empleo de un método llamado comparación de entropía cruzada, que compara las probabilidades de cadenas de bits concretas con las teóricas calculadas en un ordenador clásico. Esto proporciona una medida de la probabilidad de que el resultado sea correcto. La conclusión fue que este era preciso dentro de un intervalo del 0,2 % con una probabilidad muy alta («5 sigma»).



Arquitectura del procesador cuántico Sycamore.

A pesar de todo este progreso, la mayoría de los expertos piensa que aún falta mucho para tener un ordenador cuántico práctico. No pocos siguen sin estar convencidos de que se pueda lograr alguna vez. El físico Mijail Diakonov escribió:

> El número de parámetros continuos que describen el estado de tal ordenador cuántico útil, en cualquier momento dado, debe ser de… en torno a 10^{300}… ¿Podremos aprender alguna vez a controlar los más de 10^{300} parámetros variables de forma continua que definen el estado cuántico de un sistema así? Mi respuesta es sencilla. *No, nunca.*

*

Diakonov podría estar en lo cierto, pero otros discrepan. De cualquier modo, la mera posibilidad de que alguien (con toda probabilidad un equipo de investigación enorme financiado por un gobierno o una gran empresa) *pudiera* construir un ordenador cuántico es suficiente para quitar el sueño a los servicios de seguridad de muchos países y al sector financiero. Las fuerzas enemigas serían capaces de descifrar los mensajes del ejército y los criminales podrían acabar con el comercio y la banca por internet. De modo que los teóricos han prestado atención al aspecto que podría tener la criptografía en un mundo poscuántico, para intentar adelantarse a la jugada y restablecer unas comunicaciones seguras.

La buena noticia es que un ordenador cuántico también puede volver indescifrable aquello que es capaz de descifrar. Esto exige métodos de criptografía nuevos que empleen la computación cuántica para crear códigos novedosos que ni siquiera un procesador de este tipo pueda descifrar, lo que implicaría una manera diferente de pensar acerca de las matemáticas subyacentes. Una característica interesante es que buena parte de estas emplearían todavía la teoría de números, aunque de una cosecha más moderna que la de Fermat.

La potencial aparición inminente de los ordenadores cuánticos ha desatado una ola de investigación para concebir métodos de encriptación que no puedan descifrar estos aparatos. Hace poco, el Instituto Nacional de Estándares y Tecnología (NIST, por sus siglas en inglés) puso en marcha un programa de criptografía poscuántica dirigido a identificar sistemas criptográficos clásicos que se encuentren en peligro y a encontrar nuevas maneras de combatir sus vulnerabilidades. En 2003, John Proos y Christof Zalka[49] estimaron la seguridad del método RSA y del de curvas elípticas frente a un ordenador cuántico que ejecutase el algoritmo de Shor. Martin Roetteler y colaboradores[50] actualizaron sus resultados en 2017. Demostraron que para una curva elíptica sobre un cuerpo finito con q elementos, donde q es aproximadamente 2^n, el sistema RSA es vulnerable ante un ordenador cuántico con $9n + 2\log_2 n + 10$ cúbits y un circuito de como máximo $448n^3 \log_2 n + 4090n^3$ puertas de Toffoli. Una puerta de Toffoli es un tipo especial de circuito lógico, a partir del cual es posible construir otros circuitos para llevar a cabo cualesquiera funciones lógicas. Lo que, es más, es reversible: se puede ir hacia atrás a partir de la salida para deducir las entradas. El estándar actual para el sistema RSA es emplear un número con 2048 bits, es decir, con unas 616 posiciones decimales. El equipo estimó que

el método RSA de 2048 bits sería vulnerable ante un ordenador cuántico con $n = 256$ y la criptografía de curvas elípticas lo sería para $n = 384$.

Identificar vulnerabilidades está muy bien, pero el gran interrogante es qué hacer para protegerse frente a ellas. Esto requiere métodos de criptografía completamente nuevos. La idea general es la misma de siempre: basar el método de cifrado en problemas matemáticos difíciles con algún tipo de puerta trasera sencilla. Pero ahora «complejo» quiere decir «complejo para un ordenador cuántico». Hasta el momento se han identificado cuatro clases principales de problemas de este tipo:

- Códigos de corrección de errores lineales aleatorios.
- Resolver sistemas de ecuaciones no lineales sobre cuerpos finitos grandes.
- Encontrar vectores cortos en redes de muchas dimensiones.
- Encontrar recorridos entre vértices aleatorios de grafos de aspecto aleatorio.

Echemos un vistazo rápido al cuarto, que implica las ideas más novedosas y algunas matemáticas muy avanzadas.

A efectos prácticos se trabaja con un grafo que tiene cerca de 10^{75} vértices y una cantidad igual de grande de aristas. El código depende de encontrar un recorrido entre dos nodos concretos. Se trata de un tipo de TSP y su complejidad es comparable. Para crear una puerta trasera, el grafo debe tener alguna estructura oculta que haga fácil la solución. La idea central es emplear unas criaturas de nombre adorable: grafos de isogenias supersingulares, SIG, por sus siglas en inglés. Estos se definen mediante el empleo de curvas elípticas con propiedades especiales, que se denominan supersingulares. Los vértices corresponden a *todas* las curvas elípticas supersingulares sobre la clausura algebraica de un cuerpo finito con p elementos. Hay cerca de $p/12$ curvas así.

Una isogenia entre dos curvas elípticas es un mapa polinómico desde una hasta la otra que conserva la estructura del grupo de Mordell-Weil y se emplean para definir las aristas del grafo. Para ello, se toma un segundo número primo q. Las aristas corresponden a las isogenias de grado q entre las dos curvas elípticas relativas a los extremos de aquellas. De cada vértice emanan exactamente $q + 1$ aristas. Estos grafos son «expansores», lo que quiere decir que los recorridos aleatorios que empiezan en cualquier nodo divergen con rapidez conforme avanza el trayecto, al menos para un gran número de pasos.

Un grafo expansor puede emplearse para crear una función resumen, que es una función booleana que convierte cadenas de *n* bits en otras de *m* bits, donde *m* es mucho menor que *n*. Alice puede emplear una función resumen para convencer a Bob de que conoce alguna cadena particular de *n* bits, que también es conocida para Bob, *sin tener que divulgar cuál es*. Es decir, ella obtiene el valor resumen de esa cadena, que es mucho más corto, y se la envía a Bob. Él calcula el valor resumen de la suya y los compara.

Se requieren dos condiciones para que este método sea seguro. Una es una condición trampa denominada resistencia a preimágenes: no es factible desde un punto de vista computacional invertir la función resumen y calcular una cadena de *n* bits que tenga ese valor resumen. Habrá muchas en general, pero la cuestión es que en la práctica no se pueda encontrar *ninguna*. La otra condición deseable es una función resumen resistente a colisiones, lo que quiere decir que no es factible desde un punto de vista computacional intentar encontrar dos cadenas de *n* bits con el mismo valor resumen de *m* bits. Lo que quiere decir esto es que si una oyente no autorizada, Eva, escucha la conversación, el valor resumen que manda Alice no la ayuda a averiguar la cadena de *n* bits original.

Dados dos números primos *p* y *q*, con algunas condiciones técnicas más es posible aprovechar esta idea mediante la construcción del SIG correspondiente y emplear sus propiedades expansoras para definir una función resumen resistente a preimágenes y a colisiones. Esta puede entonces emplearse para crear un código muy seguro. Descifrarlo requiere calcular un montón de isogenias entre curvas elípticas. El tiempo de ejecución del mejor algoritmo cuántico para un cálculo así es $p^{1/4}$. Si *p* y *q* son lo bastante grandes (las matemáticas nos dicen cuánto), se tiene un sistema criptográfico que no puede descifrar ni un ordenador cuántico.

Todo es muy técnico. No pretendo que se entiendan los detalles. Para empezar, ni siquiera he hablado de la mayoría de ellos. Pero espero que haya calado la idea de que las matemáticas muy avanzadas y abstractas, que tienen que ver con geometría algebraica sobre cuerpos finitos, podrían ser justo lo que se necesita para proteger nuestras comunicaciones personales, comerciales y militares frente a receptores no autorizados armados con ordenadores cuánticos, que en la actualidad son hipotéticos pero que pronto podrían ser muy reales.

La adorada teoría de números de Hardy ha resultado ser mucho más útil de lo que él se imaginaba. Pero algunas de sus aplicaciones actuales le habrían decepcionado. Tal vez debamos pedirle disculpas.

6
El plano de los números

> El espíritu divino se manifestó sublimemente en esta maravilla del análisis, en este portento del mundo de las ideas, este anfibio entre el ser y el no ser, que llamamos raíz imaginaria de la unidad negativa.
>
> GOTTFRIED WILHELM LEIBNIZ,
> *Acta eruditorum*, 1702

> En este momento, nos hallamos en medio de una segunda revolución cuántica. La primera nos aportó reglas nuevas que gobiernan la realidad física. La segunda revolución cuántica tomará esas reglas y las empleará para desarrollar tecnologías novedosas.
>
> JONATHAN DOWLING y GERARD MILBURN,
> *Philosophical Transactions of the Royal Society*, 2003

Ha habido mucha actividad en nuestra parte de la ciudad de Coventry en los últimos meses. Hay furgonetas blancas aparcadas por todas partes en el arcén, a menudo acompañadas de camiones cargados de palas y carretillas. Excavadoras pequeñas sobre orugas patrullan las calles y abren zanjas a lo largo de las aceras, a través de las calzadas y de los jardines. El asfalto recién puesto serpentea sobre el paisaje como el rastro baboso de unos caracoles del tamaño de perros. Han aparecido hombres con chalecos reflectantes y, en apariencia, se los han tragado de nuevo unos agujeros en el suelo, allí donde se han abierto las tapas de los registros. Rollos de cable decoran los arriates y se apoyan contra los setos, a la espera de ser absorbidos por las canalizaciones. Operarios perplejos, sentados bajo toldos en la lluvia, manipulan miles de cables con códigos de colores dentro de grandes cajas metálicas.

Las furgonetas llevan un mensaje en el lateral que explica toda esta actividad: LA BANDA ANCHA DE FIBRA ÓPTICA SUPERRÁPIDA HA LLEGADO A SU ZONA.

Esta maravilla de las comunicaciones modernas se instaló en los centros de las ciudades de Reino Unido hace años, pero nuestra casa está donde el

aire da la vuelta. Una empresa se negó una vez a venir por la lejanía, la friolera de más de cinco kilómetros y medio. A decir verdad, el límite municipal se encuentra a solo unos pocos cientos de metros de distancia. Pero tender los cables es más caro y la densidad de población es menor porque una vez que se cruza esta división, los terrenos son en su mayoría campos de cultivo. No es fácil expandirse en esta dirección a bajo coste. Sencillamente, no hemos sido una propuesta rentable. Sin embargo, al fin, después de que el gobierno empezase a presionar a las empresas de telecomunicaciones, se ha puesto toda la carne en el asador para hacer llegar conexiones de fibra óptica a todas las zonas urbanas y a la mayoría de las rurales. En lugar de contemplar con desesperación cómo las zonas con mayor densidad de población recibían mejoras de manera repetida, con servicios cada vez más rápidos, ya que eran más rentables, por fin el resto del país se pone a la altura. O, por lo menos, no se queda todavía más atrás.

En una época en la que prácticamente cualquier actividad se ha desplazado a internet, la banda ancha rápida ha pasado de ser un lujo a algo básico. Puede que no sea tan fundamental como el agua o la electricidad, pero por lo menos es tan necesaria como el teléfono. La electrónica avanzada que impulsa la revolución informática y las comunicaciones globales rápidas ha convertido la década de 2020 en un mundo que treinta años antes habría parecido de ciencia ficción desde todo punto de vista. Y no ha hecho más que empezar. El aumento de la oferta crea un crecimiento explosivo en la demanda. Los días en que los cables telefónicos estaban hechos de cobre y portaban conversaciones se han quedado atrás con rapidez. E incluso estos solo podían funcionar, en los últimos años, gracias a astutos trucos electrónicos y matemáticos para aumentar su capacidad. Hoy en día, los cables de comunicación transportan muchísimos más datos que conversaciones. Por eso la fibra óptica ha ganado relevancia.

Dentro de unas pocas décadas, esta se habrá quedado tan pasada de moda como el carruaje de caballos. Los adelantos futuros, que permitirán transmitir cantidades mucho más grandes de datos a velocidades vertiginosas, están a punto de caramelo. Algunos ya existen. La física clásica de la electricidad y el magnetismo sigue siendo fundamental, pero los ingenieros electrónicos se encomiendan cada vez más al extraño mundo de los cuantos de energía para construir la próxima generación de dispositivos de comunicación. Subyacente tanto a la física clásica como a la mecánica cuántica, en las que se basan todos estos desarrollos, se encuentra uno de los inventos matemáticos más curiosos que jamás se han hecho. Se puede remontar a la Grecia clásica, consiguió un

punto de apoyo tenue durante el Renacimiento italiano y floreció por completo en el siglo XIX, cuando conquistó con rapidez la mayor parte de las matemáticas. Se empleaba con generosidad mucho antes de que nadie comprendiese en realidad de qué se trata.

Lo he llamado invento, en vez de descubrimiento, porque no fue una noción derivada del mundo natural. Si se encontraba «ahí fuera», a la espera de que alguien diese con ello, entonces ese «fuera» era un lugar muy extraño, el reino de la imaginación humana y de las imposiciones de la lógica y de la estructura. Era un nuevo tipo de ente, tanto que se denominó «imaginario». Ese nombre se mantiene en vigor hoy en día y los números imaginarios no dejan de resultar extraños desde todo punto de vista para la mayoría de las personas, incluso aunque sus vidas dependen cada vez más de ellos.

Se habrá oído hablar de la recta de los números.

Pues aquí está el plano de los números.

*

Para entender cómo se produjo este extraño desarrollo y por qué, en primer lugar, hay que fijarse en los tipos de números tradicionales. Los números son tan habituales, tan familiares, que es fácil pasar por alto sus sutilezas. Se sabe que dos más dos son cuatro y que cinco por seis son treinta. Pero ¿qué *son* «dos», «cuatro», «cinco», «seis» y «treinta»? No son las palabras en sí, porque diferentes idiomas emplean términos distintos para las mismas cantidades. Tampoco son los símbolos 2, 4, 5, 6 ni 30, porque diferentes culturas emplean signos distintos. En la notación binaria que se emplea en informática estos números están representados por 10, 100, 101, 110 y 11110. De todos modos, ¿qué es un símbolo?

Todo era mucho más sencillo cuando los números se concebían como descripciones directas de la naturaleza. Si alguien tenía diez ovejas, «diez» era una afirmación de la cantidad de ovejas que poseía. Si vendía cuatro de ellas, le quedaban seis. En esencia, los números eran instrumentos de contabilidad. Pero conforme los matemáticos empezaron a utilizarlos de maneras cada vez menos comprensibles, este enfoque pragmático empezó a ponerse en duda. Si no se sabe lo que son los números, ¿cómo puede tenerse la certeza de que los cálculos no se van a contradecir nunca entre sí? Si una granjera cuenta dos veces el mismo rebaño de ovejas, ¿debe obtener el mismo resultado de manera necesaria? Y ya que estamos, ¿qué significa «contar»?

En el siglo XIX, este tipo de cuestiones puntillosas llegaron al paroxismo, porque los matemáticos ya habían ampliado en varias ocasiones el concepto de número. Cada nueva versión incorporaba a las anteriores, pero su relación con la realidad se volvía cada vez más indirecta. Los primeros en aparecer en escena fueron los números «naturales» o «enteros positivos», empleados para contar: 1, 2, 3… Más tarde, las fracciones, como 1/2, 2/3 o 3/4. El cero se coló en algún momento dado. Hasta ese punto, la relación con la realidad era bastante directa: es posible tomar dos naranjas y otras tres naranjas y contarlas para comprobar que en total hay cinco naranjas. Con la ayuda de un cuchillo de cocina puede mostrarse media naranja. ¿Cero? Una mano vacía.

Incluso aquí hay dificultades. Media fruta no es en sentido estricto un *número* de frutas. Ni siquiera es una fruta en absoluto, solo una parte de ella. Hay muchas formas de partir una naranja por la mitad y no todas tienen una apariencia similar. Es más sencillo con trozos de cuerda, siempre que se corten de la manera evidente y no se hagan tonterías como dividirlos a lo largo. Ahora todo es sencillo de nuevo. Un pedazo de cuerda tiene una longitud que es la mitad de otro si dos copias del primero, unidas por sus extremos, miden lo mismo que el segundo. Las fracciones funcionan mejor cuando se *miden* cosas. Los griegos de la antigüedad pensaban que las medidas eran más fáciles de manejar que los símbolos numéricos, así que Euclides puso esta idea del revés. En lugar de emplear un número para medir la longitud de una recta, empleaba la recta para representar el número.

El paso siguiente, las cantidades negativas, es más complicado porque es imposible mostrar a nadie menos cuatro naranjas. Es más sencillo cuando se emplea dinero, en cuyo caso un número negativo puede entenderse como una deuda. Todo esto se sabía en China alrededor del año 200 d. C. y la primera fuente conocida es el *Jiuzhang suanshu* (*Los nueve capítulos sobre arte matemático*), pero no cabe duda de que la noción es anterior. Cuando se asocian los números a medidas, emergen de manera natural otras interpretaciones. Por ejemplo, es posible entender una temperatura negativa como aquella que está por debajo de cero, mientras que una positiva es la que se encuentra por encima de este valor. En algunas circunstancias, una medida positiva queda a la derecha de un punto determinado y una negativa está a la izquierda, etcétera. Lo negativo es lo opuesto de lo positivo.

Hoy en día, los matemáticos arman mucho revuelo en torno a las distinciones entre estos tipos de sistemas numéricos, pero para los usuarios corrientes todos son variaciones de un mismo tema: los *números*. Es aceptable mantener esta interpretación más bien inocente porque en estos conjuntos sin

excepción se cumplen las mismas reglas de la aritmética y porque cada nuevo tipo de número no hace sino extender el sistema anterior sin cambiar lo que ya se sabe. La ventaja de hacerlo así es que con cada ampliación es posible realizar operaciones que antes eran inviables. No puede dividirse dos entre tres con números naturales, pero sí con fracciones. No puede restarse cinco de tres con números naturales, pero sí con negativos. Todo lo cual hace que las matemáticas sean *más sencillas*, porque ya no hay que preocuparse acerca de si ciertas operaciones aritméticas están permitidas.

*

Las fracciones pueden dividir las cosas en porciones tan pequeñas como se quiera. Es posible seccionar metros en milímetros, de una milésima de tamaño, o en micrómetros, de una millonésima, o en nanómetros, de una mil millonésima, etcétera. Se agotan los nombres mucho antes que los ceros. A efectos prácticos, siempre hay pequeños errores de medición, así que las fracciones son todo lo que se necesita. De hecho, es posible apañarse solo con quebrados cuyo denominador sea una potencia de diez (véase cualquier calculadora electrónica). Pero para propósitos teóricos cruciales y para mantener las matemáticas ordenadas y concisas, las fracciones eran insuficientes.

La secta de los pitagóricos, en la Grecia clásica, creía que el universo funciona a base de números, un punto de vista que todavía prevalece en la física avanzada, si bien de una manera más sofisticada. Solo aceptaban los naturales y las fracciones positivas. De modo que su sistema de creencias se vio sacudido hasta la médula cuando uno de ellos descubrió que la longitud de la diagonal de un cuadrado no es una fracción exacta de la longitud de su lado. Este descubrimiento llevó hasta los denominados números «irracionales», en este caso a la raíz cuadrada de dos. Tras un complejo desarrollo histórico, que abarca desde China en el siglo IV a. C. hasta Simon Stevin en 1585, tales números vinieron a representarse como decimales:

$$\sqrt{2} = 1{,}414213562373095048\ldots$$

Dado que esta cantidad es irracional, debe continuar *para siempre*, sin acabar nunca en cero. Ni siquiera puede repetir el mismo conjunto de dígitos una y otra vez, como 1/3, que en decimales es 0,33333333333… Es un «decimal infinito». No puede escribirse nunca por completo, pero desde un punto de

vista conceptual es posible pretender que sí, al poder anotar en principio tantos dígitos como se quiera.

A pesar de la necesidad de recurrir a un proceso infinito, los decimales de este tipo tienen propiedades matemáticas muy convenientes y en particular proporcionan representaciones *exactas* de longitudes geométricas tales como $\sqrt{2}$, que de otro modo no tendrían valor numérico en absoluto. Los decimales infinitos acabaron por denominarse números «reales», porque eran medidas (idealizadas) de cantidades reales tales como longitud, área, volumen o peso. Cada dígito sucesivo representa un múltiplo de un tamaño básico que se divide por diez en cada paso. Es posible imaginar que este procedimiento prosigue de manera indefinida, con subdivisiones que se hacen cada vez más pequeñas. De este modo puede representarse la cantidad en cuestión con una precisión arbitrariamente alta. La física real no es así en el nivel atómico y es probable que ni siquiera lo sea el propio espacio, pero los números reales representan la realidad muy bien para muchos propósitos.

*

Desde un punto de vista histórico, los nuevos tipos de números se enfrentaban en general a reticencias cuando se proponían por primera vez. Después, conforme su utilidad se hacía evidente y su empleo se generalizaba, el público se acostumbraba a la noción. Tras una generación, la mayor parte de la resistencia había desaparecido. Si se crece empleando algo de manera regular, parece natural por completo. Los filósofos podían discutir si cero es un número, como lo hacen todavía, pero el resto de las personas lo empleaban cuando lo necesitaban y dejaron de preguntarse qué era. Incluso los matemáticos lo hicieron así, aunque con sentimientos ocasionales de culpa. En cierto sentido, la terminología deja al descubierto la jugada: las nuevas incorporaciones son negativas e irracionales.

No obstante, incluso entre los matemáticos, algunas innovaciones causaron dolores de cabeza que persistieron durante siglos. Lo que *de verdad* dejó suelto al zorro en el gallinero fue la introducción de los denominados números «imaginarios». Incluso el nombre (que todavía se emplea, aunque solo por motivos históricos) sugiere un cierto asombro, un indicio de que estos eran, de alguna forma, indignos. De nuevo, el problema subyacente estaba en las raíces cuadradas.

Una vez que se ha ampliado el sistema numérico para incluir los decimales infinitos, cualquier cantidad positiva tiene una raíz cuadrada. De

hecho, tiene dos: una positiva y otra negativa. Por ejemplo, las de 25 son +5 y −5. Este curioso hecho es una consecuencia de la regla «menos por menos es igual a más», algo que a menudo desconcierta a quienes lo ven por primera vez. Hay quien no lo acepta nunca. Sin embargo, no es más que un corolario del principio según el cual los números negativos deberían obedecer las mismas reglas aritméticas que los positivos. Es algo razonable, pero implica que *las cantidades negativas no tienen raíces cuadradas*. Por ejemplo, −25 carece de ellas. Parece injusto, ya que su primo +25 tiene dos. Por ello, los matemáticos empezaron a especular sobre un nuevo ámbito numérico, en el que las cantidades negativas *sí tuvieran* raíces cuadradas. También asumieron de manera tácita que en este reino ampliado seguirían siendo válidas las reglas habituales de la aritmética y del álgebra. Entonces se hizo evidente que solo se necesitaba un número radicalmente nuevo: una raíz cuadrada de −1. Si se le da a este engendro recién concebido el símbolo i, que en la actualidad emplea todo el mundo excepto los ingenieros (ellos usan j), su característica principal es que

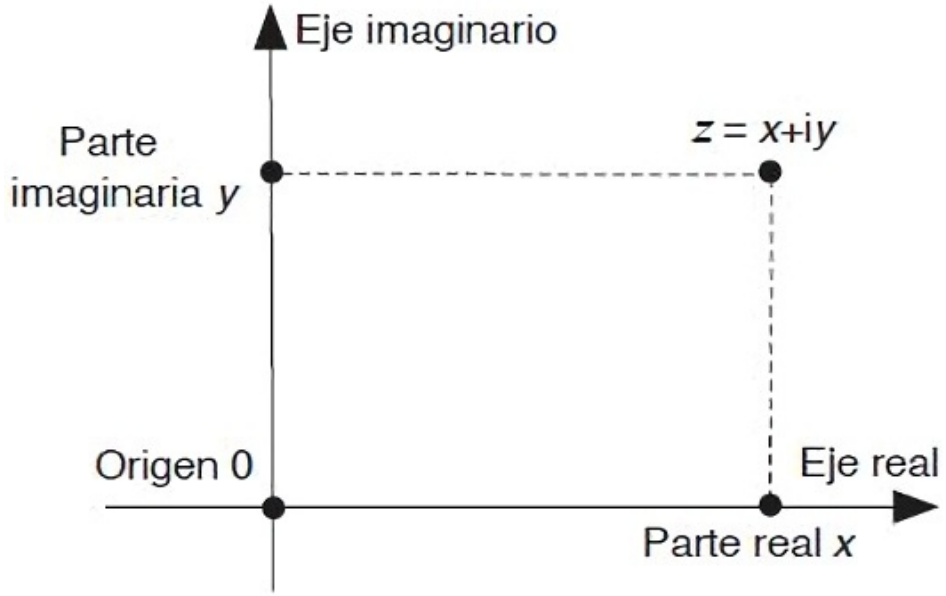
$$i^2 = -1$$

Ahora reina la justicia y todos los números, positivos o negativos, tienen dos raíces cuadradas[51]. Excepto por el cero, porque $-0 = +0$, pero este es a menudo excepcional, así que nadie se preocupa por eso[52].

La noción de que una cantidad negativa puede tener una raíz cuadrada significativa se remonta al matemático e ingeniero de la Grecia clásica Herón de Alejandría, aunque los primeros pasos para dotar de sentido a la idea se dieron milenio y medio más tarde en el Renacimiento italiano. Girolamo Cardano mencionó la posibilidad en su obra *Ars magna* (uno de los primeros textos de álgebra) en 1545, aunque la desestimó como carente de interés. Un avance importante se produjo en 1572, cuando el experto en álgebra italiano Rafael Bombelli escribió reglas para realizar cálculos con una hipotética raíz cuadrada de menos uno y obtuvo las soluciones en números *reales* de una ecuación cúbica al emplear una fórmula que sumaba dos «números» que no podían ser cantidades reales de ninguna manera concebible. Las partes «imposibles» se cancelaban entre sí de manera muy oportuna y quedaba la respuesta correcta (y real). Esta audaz gesta de arcana prestidigitación llamó la atención de los matemáticos, ya que esas soluciones se podían comprobar de forma directa y eran válidas.

Para dorar la píldora, se dijo que los recién llegados eran «imaginarios», a diferencia de los «reales» tradicionales que podían emplearse para medir objetos existentes. Esta terminología otorgaba a estos últimos números un estatuto especial que no merecían y mezclaba un concepto matemático con el modo habitual en que se empleaba. Como se verá, los imaginarios tienen usos e interpretaciones perfectamente razonables, pero no como medida de cantidades físicas convencionales, como la longitud o la masa. Bombelli fue el primero en demostrar que los números imaginarios (sean lo que sean los condenados) podían emplearse para resolver problemas reales por completo. Era como si un extraño utensilio de carpintería, que ni siquiera existe, pudiese cogerse de algún modo y utilizarse para fabricar una silla del todo normal. Por supuesto, era una herramienta conceptual, pero incluso así, el procedimiento era sorprendente. Más sorprendente todavía era la evidencia de que funcionaba.

De forma milagrosa, no dejó de hacerlo en un arsenal siempre creciente de aplicaciones. Al llegar el siglo XVIII, los matemáticos usaban estos nuevos números con libertad. Euler introdujo el símbolo convencional i para la raíz cuadrada de menos uno en 1777. La combinación de reales e imaginarios condujo a un sistema hermoso y autoconsistente conocido como números complejos (en el sentido de «compuestos de varias partes», no de «complicados»). Desde el punto de vista algebraico, tienen la apariencia de $a + b$i, donde a y b son reales. Pueden sumarse, restarse, multiplicarse, dividirse, es posible sacar sus raíces cuadradas, cúbicas, etcétera, sin salir nunca del sistema que forman.



El plano de los números complejos.

Su principal defecto es que es difícil encontrar una interpretación en el mundo existente, o al menos eso es lo que pensaban todos los entendidos en esa época. No está claro el aspecto que tiene una medida de, digamos, 3 + 2i. Se sucedieron los debates semifilosóficos acerca de la legitimidad de los números complejos, hasta que los matemáticos descubrieron cómo emplearlos para resolver problemas de física matemática. Dado que los resultados podían comprobarse por otros medios y que siempre parecían ser correctos, la polémica pasó a segundo plano ante la fiebre por aprovechar estas técnicas nuevas y poderosas.

*

Durante mucho tiempo, los matemáticos intentaron justificar los números imaginarios mediante el recurso a un vago pero genérico «principio de permanencia», que en esencia afirmaba que cualquier regla algebraica que fuese válida para los números reales debía serlo también, de manera automática, para los complejos. En lo que supone un triunfo de la voluntad frente a la lógica, la principal prueba a favor de esta afirmación era que, en la práctica, el uso de estos últimos conduce a soluciones correctas. En resumen, que funcionaban porque sí y la demostración era que… funcionaban.

Solo mucho más tarde solucionaron los matemáticos la manera de representar números complejos. De hecho, al igual que las cantidades

negativas, tienen varias interpretaciones diferentes en el «mundo real». Veremos en breve que en ingeniería eléctrica combinan la amplitud (el tamaño máximo) de una señal oscilatoria con su fase en un único paquete compacto y conveniente. Lo mismo ocurre en mecánica cuántica. De manera más prosaica, igual que los números reales representan puntos en una recta, los complejos lo hacen en un plano. Así de simple. Y, al igual que muchas ideas sencillas, se pasó por alto durante siglos.

El primer indicio de este avance pudo verse en 1685 en el *Álgebra* de John Wallis. Este autor extendió la representación convencional de los reales como una línea a los complejos. Supongamos un número $a + b\mathrm{i}$. La «parte real» a no es más que una cantidad convencional de ese tipo, así que puede ubicarse en la recta habitual, concebida como una línea fija en un plano. La componente restante $b\mathrm{i}$ es un número imaginario, de modo que no corresponde a ningún punto de esa representación. No obstante, el coeficiente b es real, así que puede trazarse una línea de longitud b en ese plano en un ángulo de 90° con respecto a la recta anterior. De este modo se obtiene un punto en el plano que se identifica con $a + b\mathrm{i}$. Hoy en día resulta evidente de manera inmediata que esto representa a ese número como el punto del plano con coordenadas (a, b), pero en su época, la propuesta de Wallis cayó en oídos sordos. El reconocimiento histórico suele atribuirse a Jean-Robert Argand, quien lo publicó en 1806, aunque se le había adelantado en 1797 un agrimensor danés poco conocido, Caspar Wessel. No obstante, el artículo de Wessel estaba escrito en danés y pasó desapercibido hasta que vio la luz una traducción al francés, un siglo más tarde. Ambos propusieron construcciones geométricas de estilo euclídeo que mostraban cómo sumar y multiplicar dos números complejos cualesquiera.

Por último, en 1837, el matemático irlandés William Rowan Hamilton señaló de manera explícita que puede representarse un complejo como un par de números reales (las coordenadas de un punto en el plano):

número complejo = (primer número real, segundo número real)

Entonces escribió de nuevo las construcciones geométricas como dos fórmulas para sumar y multiplicar esos pares. Las mostraré aquí porque son bastante sencillas y elegantes:

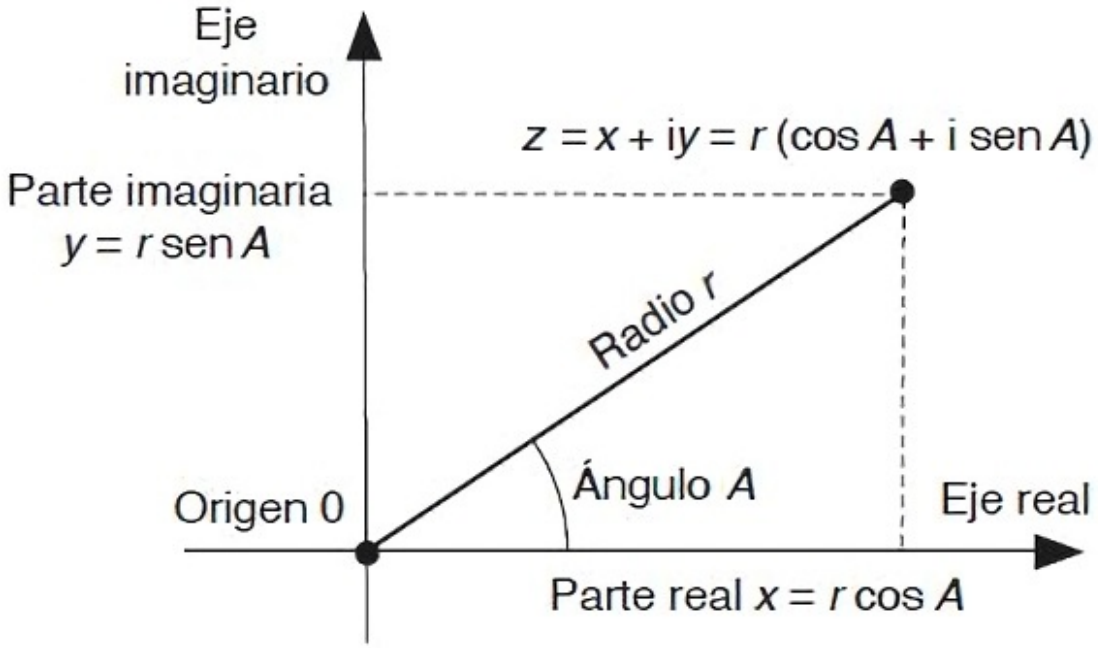
$$(a, b) + (c, d) = (a + c, b + d)$$
$$(a, b) \times (c, d) = (ac - bd, ad + bc)$$

Esto puede parecer un poco críptico, pero funciona de maravilla. Los números con la forma (a, 0) se comportan como reales y el misterioso i es el par (0, 1). Esta es la sugerencia de Wallis de que los imaginarios forman un ángulo recto con los reales cuando se escriben como coordenadas. La fórmula de Hamilton dice que

$$i^2 = (0, 1) \times (0, 1) = (-1, 0)$$

algo que ya se ha identificado con el número real −1. ¡Conseguido! Por supuesto, luego resultó que Gauss ya había mencionado la misma idea en una carta a Wolfgang Bolyai en 1831 pero que no había publicado.

Algo que tal vez Gauss no acabó de entender, pero Hamilton sí, era que ambas fórmulas también permiten demostrar que los números complejos obedecen todas las reglas habituales del álgebra, anteriormente asociada solo con los números reales. Leyes como la propiedad conmutativa $xy = yx$ y la propiedad asociativa $(xy)z = x(yz)$, que la mayoría de las personas dan por sentadas cuando abordan por primera vez el álgebra. Para demostrar que también son válidas para los números complejos, basta con sustituir los símbolos por pares de números reales, aplicar las fórmulas de Hamilton y comprobar que ambos términos de la igualdad resultan en el mismo par, solo mediante el empleo de las reglas algebraicas que obedecen los reales. Fácil. Resulta irónico que para cuando Gauss y Hamilton resolvieron la lógica subyacente en términos de pares de números «reales» ordinarios, los matemáticos habían hecho tanto uso de los complejos que en buena medida habían perdido el interés por darles un significado lógico concreto.



Geometría de coordenadas cartesianas y polares del plano de los complejos. Aquí cos y sen se refieren a las funciones trigonométricas coseno y seno (de hecho, la imagen las define).

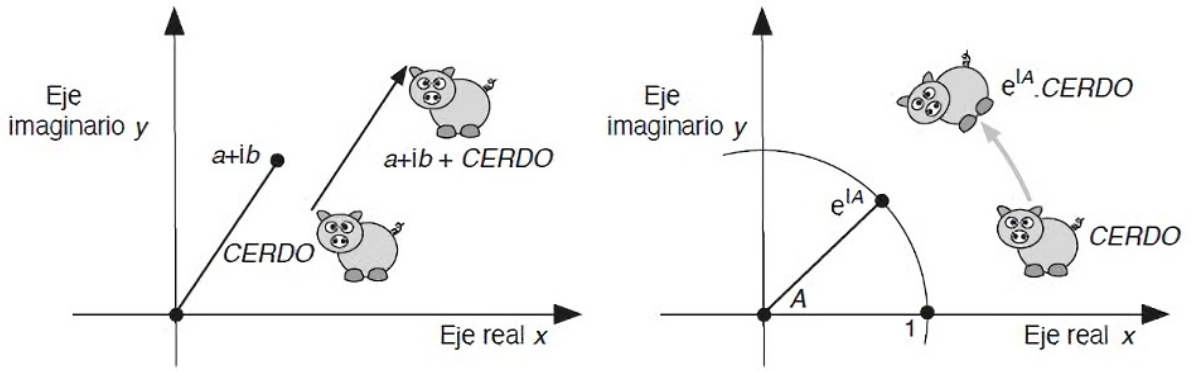
Entre esas aplicaciones sobresalían problemas de física, tales como los campos magnéticos y eléctricos, la gravedad y el flujo de fluidos. Cabe destacar que algunas ecuaciones básicas del análisis complejo (cálculo con funciones complejas) eran la correspondencia exacta de las que se empleaban de manera habitual en la física matemática. De modo que se podían resolver las ecuaciones físicas al hacer el cálculo con números complejos. La limitación principal era que estos reposan en un plano, de modo que la física también debía tener lugar en un espacio así o ser equivalente a un problema en él.

*

Los números complejos otorgan al plano una estructura algebraica sistemática que se adapta de manera perfecta a la geometría y por lo tanto también al movimiento. Es posible entender el resto de este capítulo como un ensayo bidimensional de problemas similares en geometría tridimensional, que es de lo que trata el siguiente. Habrá unas cuantas fórmulas (*es* álgebra, al fin y al cabo), pero no estoy seguro de cómo pueden evitarse sin que todo parezca bastante impreciso.

Cuando se representa un número complejo z en la forma $z = x + iy$, con x e y reales, la geometría subyacente es el sistema cartesiano de coordenadas, así llamado por René Descartes, con dos ejes perpendiculares entre sí: la parte

real *x* (eje horizontal) y la parte imaginaria *y* (eje vertical). No obstante, hay otro sistema de coordenadas importante en el plano, denominadas polares, que representa un punto como un par (*r*, *A*), donde *r* es un número real positivo y *A* es un ángulo. Ambas representaciones están relacionadas de manera estrecha: *r* es la distancia desde el origen 0 hasta *z* y *A* es el ángulo entre el eje horizontal y la recta que une el origen con *z*.



Desplazamiento (*izquierda*) y rotación (*derecha*) del conjunto de puntos CERDO mediante el empleo de números complejos.

Las coordenadas cartesianas son ideales para describir el movimiento de los objetos sin rotación. Si un punto $x + iy$ se desplaza *a* unidades en horizontal y *b* unidades en vertical, se mueve hasta $(x + iy) + (a + ib)$. Si se extiende esta idea a un conjunto de puntos, con un listado de valores para *x* e *y*, entonces todo él se mueve *a* unidades en horizontal y *b* unidades en vertical cuando se suma un número complejo $a + ib$ fijo a cada elemento del conjunto. Es más, este movimiento es *rígido*: todo el objeto se mueve sin cambiar su forma ni su tamaño.

Otro tipo importante de movimiento rígido es la rotación. De nuevo, el objeto no cambia de forma ni de tamaño, pero sí de orientación y gira un cierto ángulo en torno a un punto central. Una observación fundamental en este sentido es que multiplicar por i rota 90° los elementos en torno a un centro en el origen. Por eso el eje *y*, que representa la parte «imaginaria» *y* de *z*, es perpendicular al *x*, que corresponde a la componente «real» *x*. (A pesar del nombre, la parte imaginaria es un número real: solo se hace imaginaria cuando se la multiplica por i para obtener iy).

Si se quiere rotar 90° un conjunto de puntos se multiplica cada uno de estos por i. De manera más genérica, si lo que se quiere es que roten un ángulo cualquiera *A*, un poco de trigonometría revela que hay que multiplicarlos todos por el número complejo

$$\cos A + \mathrm{i}\,\mathrm{sen}\,A$$

Euler descubrió una hermosa y notable relación entre esta expresión y el análogo complejo de la función exponencial e^x, donde $e = 2{,}71828\ldots$ es la base de los logaritmos naturales. Es posible definir la función exponencial e^z de un número complejo z de modo que tenga las mismas propiedades básicas que su análoga real y que coincida con esta cuando z tiene parte imaginaria nula. Resulta que

$$e^{iA} = \cos A + \mathrm{i}\,\mathrm{sen}\,A$$

Una manera elegante de ver el motivo por el que sucede esto es emplear ecuaciones diferenciales. Lo he puesto en las notas**[53]** porque es demasiado técnico.

En la representación en coordenadas polares de un número complejo, resulta que estas se identifican con el punto

$$r\,(\cos A + \mathrm{i}\,\mathrm{sen}\,A) = re^{iA}$$

que es una fórmula sencilla y compacta.

La belleza de los números complejos, en lo que respecta a la geometría, reside en que tienen dos sistemas de coordenadas naturales de manera simultánea: cartesianas y polares. Desplazar un objeto tiene una fórmula sencilla en las primeras, pero es un lío en polares. Rotar un objeto tiene una fórmula sencilla en las segundas, pero es un lío en cartesianas. Si se emplean números complejos es posible elegir la representación más adecuada para nuestros propósitos.

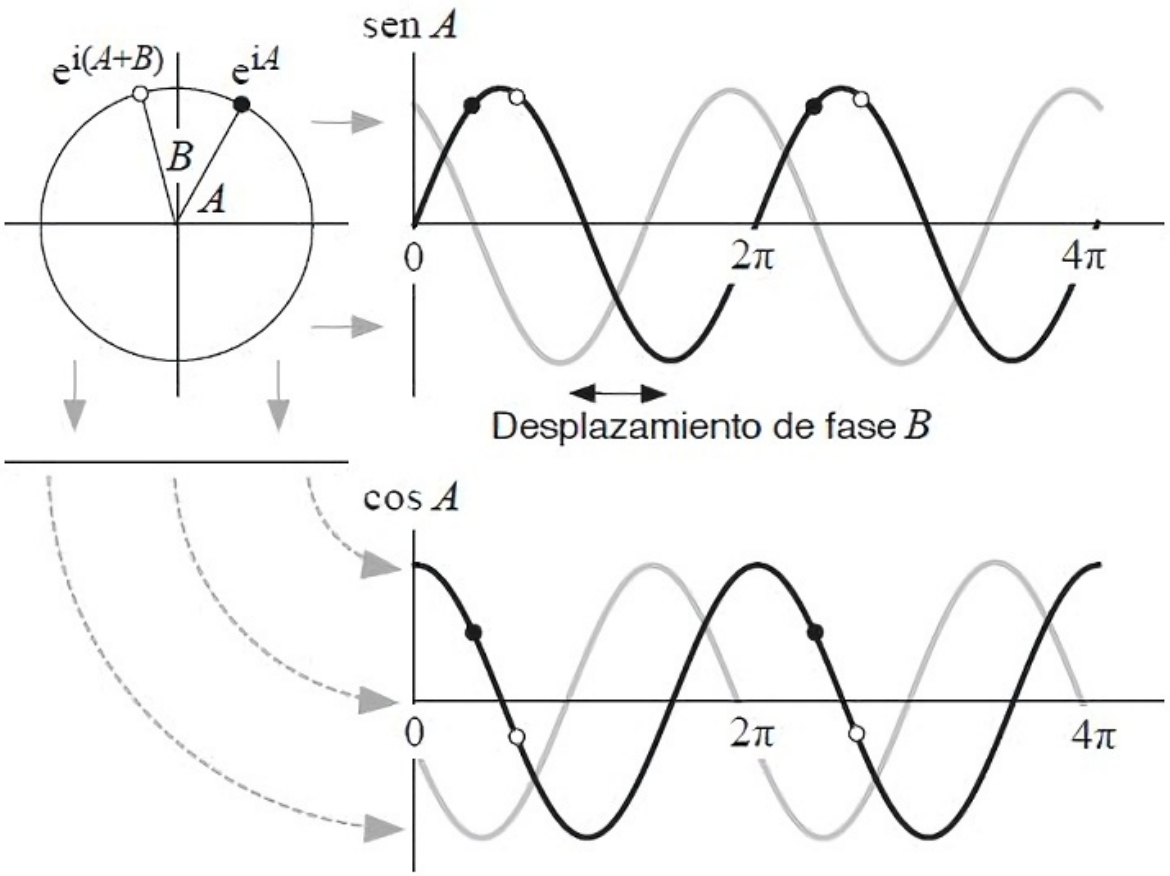
Estas características geométricas del álgebra compleja podrían aprovecharse en gráficos de ordenador bidimensionales, pero resulta que el plano es sencillo y que a las máquinas no les preocupan las fórmulas complicadas, así que no se adelanta mucho. En el capítulo 7 se verá que un truco parecido funciona a las mil maravillas para los gráficos tridimensionales. No obstante, de momento hace falta rematar la historia de los números complejos y tomar en consideración algunas aplicaciones útiles de verdad.

*

Los matemáticos han llegado de manera gradual a la conclusión de que, a pesar de carecer de una interpretación física evidente, los números complejos son a menudo más sencillos que los reales y arrojan luz sobre algunas características de estos que serían intrigantes de otro modo. Por ejemplo, como ya observaron Cardano y Bombelli, las ecuaciones cuadráticas tienen, o bien dos soluciones reales, o bien ninguna, y las cúbicas tienen, o bien una solución real, o bien tres. El caso es mucho más sencillo en complejos: las ecuaciones cuadráticas siempre tienen dos soluciones complejas y las cúbicas siempre tienen tres. Por lo que a esto respecta, las ecuaciones de décimo grado tienen diez soluciones complejas pero pueden tener diez, ocho, seis, cuatro, dos o ninguna solución real. En 1799, Gauss demostró algo que se sospechaba desde hacía mucho, que ya había conjeturado Paul Roth nada menos que en 1608 y que se vino a conocer como el teorema fundamental del álgebra: una ecuación polinómica de grado *n* tiene *n* soluciones complejas. Todas las funciones habituales del análisis, tales como la exponencial, el seno, el coseno, etcétera, tiene análogos complejos naturales y, en general, sus propiedades se hacen más sencillas cuando se consideran desde esta perspectiva.

Una consecuencia práctica es que los números complejos se han convertido en una técnica habitual en la ingeniería electrónica, sobre todo porque proporcionan una manera sencilla y elegante de manejar corrientes alternas. La corriente eléctrica es un flujo de electrones, partículas subatómicas con carga. En la directa, como la que produce, por ejemplo, una batería, todos los electrones fluyen en la misma dirección. En la alterna, muy empleada en la red comercial porque es más segura, los electrones oscilan hacia delante y hacia atrás. La gráfica del voltaje (y de la intensidad) tiene la apariencia de una función coseno de la trigonometría.

Una manera sencilla de producir una curva así surge cuando se piensa en un punto en el borde de una rueda que gira. Por simplicidad, supongamos que esta última tiene radio 1. Si se atiende a la proyección horizontal del punto en rotación, se mueve de lado a lado y alcanza valores de +1 y −1 en sus extremos. Si la rueda gira a una velocidad constante, la gráfica de esta distancia horizontal es una función coseno y la gráfica de la distancia vertical es una función seno (las curvas en negrita en la imagen).



La rotación en el plano complejo se proyecta para producir oscilaciones periódicas. Sumar *B* al ángulo *A* desplaza las gráficas hacia la izquierda: un desplazamiento de fase.

La posición del punto que se mueve es el par de cantidades reales (cos *A*, sen *A*), donde *A* es el ángulo entre aquel y el eje horizontal. Si se emplea el truco de Hamilton, es posible interpretar esto como el complejo cos *A* + i sen *A*. Al variar el ángulo, este número recorre una vez tras otra el círculo unidad en el plano complejo. Si *A* se mide en radianes, completa una vuelta cada vez que el valor aumenta desde 0 hasta 2π. Después hace otra cuando *A* pasa de 2π a 4π, etcétera, de modo que el movimiento es periódico con periodo 2π.

La fórmula de Euler implica que, conforme *A* recorre los números reales, el valor correspondiente de e^{iA} da vueltas y más vueltas a lo largo del círculo unidad a velocidad constante. Esta relación proporciona un modo de convertir cualquier afirmación acerca de una función oscilante, con forma parecida al seno o al coseno, en una exponencial compleja. Desde un punto de vista matemático, esta es más sencilla y manejable. Es más, el ángulo *A* tiene una interpretación física natural como la fase de la oscilación, lo que quiere decir que sumar un ángulo constante *B* para cambiar *A* desplaza las funciones seno y coseno en una cantidad correspondiente (las curvas en gris en la imagen).

Todavía mejor, las ecuaciones diferenciales básicas del voltaje y la intensidad en circuitos pueden convertirse sin cambios a sus análogas complejas correspondientes. La oscilación física pasa a ser la parte real de una exponencial compleja y son válidos los mismos métodos para corrientes alternas y directas. Es como si el comportamiento real tuviese un compañero imaginario secreto y los dos juntos fuesen más sencillos que cualquiera de ellos por sí solo. Los ingenieros electrónicos emplean este truco matemático de manera rutinaria para simplificar sus cálculos, incluso cuando utilizan ordenadores.

*

En esta aplicación a la electrónica, los números complejos se hacen aparecer como si fuesen conejos matemáticos que salen de la chistera de un mago y que además resulta que hacen más fácil la vida de los ingenieros. Pero hay un contexto notable en el que son imprescindibles y en el que tienen un significado físico. Se trata de la mecánica cuántica.

Wigner hizo de este ejemplo de irrazonable eficacia el elemento central de su conferencia:

> No olvidemos que el espacio de Hilbert de la mecánica cuántica es el espacio de Hilbert complejo… Es seguro que para la mente despreocupada los números complejos están lejos de lo natural y lo sencillo, y no pueden resultar sugeridos por las observaciones físicas. Más aún, el uso de números complejos no es en este caso un truco de cálculo de la matemática aplicada, sino que está muy cerca de ser una necesidad en la formulación de las leyes de la mecánica cuántica.

Tampoco ahorró esfuerzos para enfatizar lo que quería decir con «irrazonable»:

> Nada en nuestra experiencia, ciertamente, sugiere la introducción de tales cantidades. En realidad, si a un matemático se le pide que justifique su interés en los números complejos, indicará con cierta indignación los muchos y bellos teoremas de la teoría de ecuaciones, de las series de potencias y de las funciones analíticas en general, que deben su origen a la introducción de los números complejos… Es difícil evitar la impresión de que aquí nos enfrentamos a un milagro, completamente

> comparable… a los dos milagros de la existencia de leyes de la naturaleza y de la capacidad de la mente humana para adivinarlas.

La mecánica cuántica cobró forma en torno a 1900 para explicar el extraño comportamiento de la materia en pequeña escala, que los físicos experimentales habían empezado a descubrir, y creció con rapidez hasta ser la teoría física de mayor éxito que la humanidad ha desarrollado hasta el momento. Al nivel de las moléculas, de los átomos y, sobre todo, de las partículas subatómicas que se unen para formar estos últimos, la materia se comporta de maneras sorprendentes y desconcertantes. Tanto, que no está claro en absoluto que la palabra «materia» siga siendo válida. Las ondas, como la luz, se comportan en ocasiones como partículas: los fotones. Las partículas, como los electrones, se comportan en ocasiones como ondas.

Esta dualidad entre ondas y partículas se resolvió al fin con la introducción de ecuaciones matemáticas que rigen tanto unas como otras, aunque incluso hoy en día hay muchos aspectos que siguen siendo inexplicables. Al hacerlo así, la forma en que ambas se representan en las matemáticas sufrió, como diría Shakespeare, «la transformación del mar, en algo rico y extraño». Hasta ese momento, los físicos describían el estado de una partícula material en términos de una lista breve de cantidades: masa, tamaño, posición, velocidad, carga eléctrica, etcétera. En mecánica cuántica, el estado de cualquier sistema viene descrito por una onda, más en concreto, por su función de onda. Como su propio nombre indica, esta es una función matemática con propiedades similares a las de una onda.

Una función es una regla o proceso matemático que transforma alguna cantidad en otra de una forma concreta. De manera más general, es capaz de transformar un listado de números en una sola cifra o incluso en otro listado. De manera todavía más general, puede operar no solo sobre números, sino sobre conjuntos de objetos matemáticos de cualquier tipo. Por ejemplo, la función «área» actúa sobre el conjunto de todos los triángulos y cuando se aplica a cualquiera de ellos en concreto, su resultado es el área de ese triángulo.

La función de onda de un sistema cuántico opera sobre el listado de mediciones posibles que se podrían realizar en el sistema, tales como sus coordenadas de posición o de velocidad. En mecánica clásica, una cantidad finita de estos números determina, por lo normal, el estado del sistema, pero en mecánica cuántica este listado podría incluir infinitas variables. Estas se toman del denominado espacio de Hilbert, que es (a menudo) un espacio de

infinitas dimensiones con una noción bien definida de distancia entre dos cualesquiera de sus miembros[54]. El resultado de la función de onda es un único número para cada elemento en el espacio de Hilbert, pero esa cantidad no es real: es compleja.

En la mecánica clásica, un observable (algo que puede medirse) asocia un número a cada estado posible del sistema. Por ejemplo, cuando se observa la distancia de la Tierra a la Luna se obtiene un único valor y la medición es una función definida sobre el espacio de todas las configuraciones posibles que pueden tomar en principio la Tierra y la Luna. En mecánica cuántica, los observables son *operadores*. Un operador toma un elemento del espacio de Hilbert de estados y lo convierte en un número complejo. Además, debe obedecer una lista breve de reglas matemáticas. Una es la linealidad. Supongamos que hay dos estados x e y, y que los resultados del operador L para ellos son los valores $L(x)$ y $L(y)$. En teoría cuántica, los estados pueden superponerse (sumarse entre sí) para dar $x + y$. La linealidad quiere decir que el resultado de L en este caso debe ser $L(x) + L(y)$. El listado completo de propiedades exigidas conforma un operador denominado hermítico, que se comporta de manera adecuada con relación a las distancias en el espacio de Hilbert.

Los físicos elijen estos espacios y operadores de varias maneras para modelar sistemas cuánticos concretos. Si están interesados en los estados de posición y momento para una sola partícula, el espacio de Hilbert comprende todas las funciones «de cuadrado integrable» y tiene infinitas dimensiones. Si lo que les interesa es el espín de un único electrón, el espacio de Hilbert tiene dos dimensiones y comprende los denominados «espinores». Un ejemplo es la ecuación de Schrödinger, que tiene la siguiente apariencia:

$$\mathrm{i}\hbar\frac{\mathrm{d}}{\mathrm{d}t}\left|\Psi(t)\right\rangle = \hat{H}\left|\,(t)\right\rangle$$

No hace falta entender las matemáticas, pero echemos un vistazo a los símbolos. Sobre todo al que va delante del todo, que en buena medida pone las cartas sobre la mesa: es i, la raíz cuadrada de menos uno. He aquí la ecuación básica de la mecánica cuántica y lo primero que aparece es el número imaginario i.

El siguiente, $\hbar$, es una cantidad llamada constante de Planck reducida y es minúscula: en torno a 10^{-34} julios por segundo. Es la razón de que haya cuantos (saltos pequeños pero discontinuos en los valores que pueden asumir varios observables) en la mecánica cuántica. Después viene una fracción, d/dt.

La *t* representa el tiempo y las dos letras d indican que se debe encontrar una tasa de cambio, como en el cálculo, así que es una ecuación diferencial. El grupo de símbolos es la función de onda, que especifica el estado cuántico del sistema en un instante *t*, y es aquello cuya tasa de cambio se quiere conocer. Por último, $\hat{H}$ es el denominado hamiltoniano: en esencia, la energía.

La interpretación habitual de la función de onda es que no representa un estado individual, sino la *probabilidad* de que una observación encuentre al sistema en ese estado. No obstante, las probabilidades son números reales entre 0 y 1, mientras que los resultados de la función de onda son complejos de cualquier tamaño. Por eso los físicos se centran en la amplitud (lo que los matemáticos llaman el módulo) de este número, que es lo alejado que se encuentra del origen (la *r* en coordenadas polares). Conciben esa cantidad como una probabilidad relativa, de modo que si un estado tiene una amplitud 10 y otro tiene 20, entonces el segundo es dos veces más probable que el primero.

El módulo dice a qué distancia del origen se encuentra un complejo, pero no la dirección en la que hay que ir para alcanzarlo. Esta viene especificada por otro número real, el ángulo *A* en coordenadas polares. Los matemáticos lo denominan argumento, pero los físicos lo llaman fase (la longitud que hay que recorrer a lo largo del círculo unidad). De modo que la función de onda compleja tiene una amplitud, que cuantifica la probabilidad relativa de que ocurra esa observación, y una fase, que no modifica la otra coordenada y que es casi imposible de medir. Sin embargo, sí afecta al modo en que se superponen los estados individuales y por tanto también a la probabilidad de que ocurra cada uno de ellos, pero en la práctica está oculta a la observación experimental.

Todo esto quiere decir que un número real por sí solo es el tipo erróneo de número para cuantificar un estado. Ni siquiera se puede *formular* la mecánica cuántica en términos de cantidades reales convencionales.

*

Si la pregunta es «¿qué usos prácticos tienen los números complejos?», entonces puede señalarse la miríada de aplicaciones de la mecánica cuántica, con la certeza de saber que también deben serlo de los complejos. Hasta hace poco, la mayoría de las respuestas habrían remitido a experimentos de laboratorio, de las fronteras más remotas de la física, pero no al tipo de cosas que pueden encontrarse en la cocina o en el cuarto de estar. La electrónica

moderna ha cambiado eso y muchos de los dispositivos favoritos funcionan por motivos mecánico-cuánticos. Los ingenieros necesitan comprender estos temas con mucha profundidad y detalle, pero el resto de las personas pueden repantingarse y admirar sus creaciones. O, de vez en cuando, maldecirles cuando algún tecnicismo poco claro en el modo en que se ha configurado el maldito chisme no deja hacer lo que uno quiere.

Mi banda ancha de fibra óptica recién instalada es un ejemplo. Parece un cable convencional, pero es parte de un sistema de transmisiones que ya depende de este tipo de tecnologías. Sin embargo, la parte cuántica no está en la fibra en sí. Se halla en los dispositivos a lo largo de su trayectoria que crean los pulsos luminosos con los que funciona todo el conjunto. Por supuesto, la luz es en realidad cuántica en todo caso, pero estos dispositivos se han *diseñado* empleando una mecánica de este tipo y no funcionarían sin ella.

El término «fibra» se refiere a un cable de muchas fibras, las cuales son hilos finos de vidrio que transmiten luz. Están diseñadas de modo que esta rebota en las paredes en vez de escapar, así que los cables pueden doblarse para salvar esquinas y la luz permanece en su interior. La información está codificada en el haz luminoso como una serie de pulsos definidos. La industria de las telecomunicaciones introdujo este modelo porque combina varias ventajas. Las fibras disponibles en la actualidad son muy transparentes, de modo que transmiten la luz a lo largo de grandes distancias sin degradar la señal. Los pulsos luminosos son capaces de portar mucha más información que los cables telefónicos de cobre tradicionales. Este ancho de banda mayor es lo que le proporciona su «velocidad» superior: no se trata de lo rápido que se desplaza el pulso, sino de cuántos de estos, cuánta información, puede portarse. Los cables de fibra son más ligeros que los de cobre, así que son más sencillos de transportar e instalar y menos proclives a las interferencias eléctricas.

Un sistema de comunicaciones ópticas tiene cuatro componentes principales: un emisor (una fuente de luz); un cable para portar la señal; una serie de repetidores que recogen esta última antes de que se haya degradado demasiado, la limpian y la reenvían; y, por supuesto, un receptor (detector). Solo voy a considerar uno, el emisor. Debe ser un dispositivo capaz de producir luz y que pueda controlarse de modo que esta surja como una serie de pulsos discretos que pueden encenderse (1) o apagarse (0) para codificar un mensaje en binario. La conmutación debe ser extremadamente rápida y todo tiene que ser muy preciso. En particular, la longitud de onda («color») de

la luz debería tener un valor concreto. Por último, los pulsos deben mantener su forma, de modo que el receptor pueda reconocerlos.

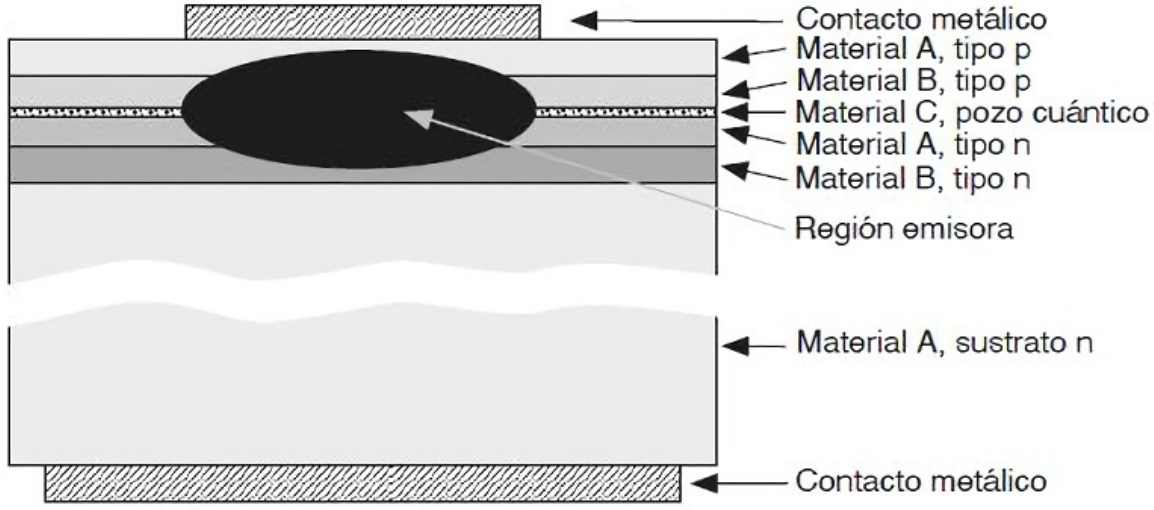
El medio ideal (de hecho, el único) para este fin es el láser, porque emite un haz potente de luz coherente de una longitud de onda concreta. «Coherente» quiere decir que todas las ondas en el haz están en fase entre sí, de modo que no se cancelan unas con otras. El láser consigue esto al hacer que la luz (en forma de fotones) rebote entre una pareja de espejos, lo que desencadena una cascada cada vez mayor de fotones en un ciclo de retroalimentación positiva. Cuando el haz se ha hecho lo bastante fuerte se le permite escapar.

Los primeros láseres eran grandes y aparatosos, pero hoy en día la mayoría de los de pequeño tamaño se fabrican mediante los mismos procesos generales que producen los circuitos microscópicos en los chips (circuitos semiconductores integrados) informáticos. Durante los últimos treinta años, casi todos los dispositivos de este tipo empleados en las tecnologías de consumo y corporativas (tales como los reproductores Blu-ray, que son posibles gracias a láseres que producen luz azul) han sido aparatos con heteroestructura de confinamiento separado (SCH, por sus siglas en inglés). Estos representan una mejora de los modelos de punto cuántico, que son una especie de emparedado en el que la capa intermedia actúa como un pozo cuántico. De este modo, se crean funciones de onda que tienen la apariencia de una serie de escalones en lugar de una curva, por lo que los niveles energéticos están cuantificados (son nítidos y separados, en vez de ser borrosos y unidos entre sí). Estos niveles pueden ajustarse, mediante un diseño adecuado del pozo cuántico, para emitir luz de la frecuencia correcta para el funcionamiento del láser.

Los dispositivos SCH añaden dos capas adicionales en las partes superior e inferior del emparedado, con un índice de refracción menor que las tres intermedias, lo que confina la luz dentro de la cavidad del láser. Es razonable suponer que es imposible diseñar un dispositivo cuántico de este tipo sin aplicar una buena cantidad de mecánica cuántica. De modo que incluso las fibras ópticas de la década de 1990 ya empleaban componentes cuánticos, algo que es más cierto si cabe hoy en día.

En el futuro, es probable que un amplio arsenal de dispositivos cuánticos novedosos transforme nuestras vidas. El principio de incertidumbre de Heisenberg de la mecánica cuántica afirma que hay ciertos observables que no pueden medirse de manera exacta al mismo tiempo. Por ejemplo, si se conoce con precisión dónde se encuentra una partícula, no se puede tener

certeza acerca de la velocidad a la que se desplaza. Es posible emplear esta característica para detectar si alguien no autorizado accede a mensajes secretos. Cuando Eva la Cotilla observa a hurtadillas el estado cuántico de una señal emitida (digamos el espín de un fotón), este cambia y ella no puede controlar cómo lo hace. Es como si hubiera una alarma incorporada en el código que suena cada vez que Eva intenta leerlo.



Estructura esquemática de un láser SCH. Los términos «tipo n» y «tipo p» se refieren, respectivamente, a semiconductores en los que la carga es transportada por electrones o por «huecos», en los que faltan electrones.

Una manera de implementar esta idea es emplear fotónica cuántica, las propiedades mecánico-cuánticas de los fotones. Otra es manipular el espín de las partículas, la rama en desarrollo de la *espintrónica*. Tales dispositivos son capaces de transportar más información que una señal convencional al codificar datos adicionales en el espín de las partículas y no solo en el hecho de que estén o no presentes. De modo que mi banda ancha superrápida puede verse superada pronto por la fibra espintrónica superultrarrápida. Hasta que alguna mente brillante invente la holografía sensorial hiperdefinida en seis dimensiones y haga que todo ese ancho de banda adicional se quede corto.

7
Papá, ¿puedes multiplicar tripletes?

¿Las olas del océano golpean con gentileza contra tu barco en *ASC: Black Flag*?

Mates.

¿Esas balas que vuelan por encima de tu cabeza en *Call of Duty: Ghosts*?

Mates.

¿Que Sonic es capaz de correr muy deprisa y Mario de saltar?

Mates.

¿Derrapar en esa esquina en *Need for Speed* a 140 kilómetros por hora?

Mates.

¿Bajar la ladera haciendo *snowboard* en *SSX*?

Mates.

¿Ese cohete que despega en *Kerbal Space Program*?

Mates.

Página web de *Forbes*,
«This is the Math behind Super Mario»

La aldea tiene un aire medieval, con cabañas de techo de paja, carretas de caballos en un camino de tierra, campos con cultivos y ovejas. La cinta delgada de un riachuelo fluye entre las edificaciones apiñadas y brilla dorado bajo el sol de poniente. Se ve la escena desde arriba, como si se contemplase desde una aeronave. La panorámica gira y se balancea, conforme el vehículo desciende y tuerce. Pero no se trata de ningún aeroplano: el cambio a un punto de vista en el suelo revela el perfil inconfundible de un dragón. Que se acerca. Vuelta a la perspectiva del ojo de la criatura, que ahora baja en un picado pronunciado y roza los tejados al pasar. Se ve una llamarada que se extiende al frente y la paja se incendia…

Podría ser una película o un videojuego. Ambos pueden ser casi indistinguibles en la actualidad. De cualquier modo, es un triunfo de las imágenes generadas por computadora o IGC.

¿Son matemáticas?

Sin duda.

Entonces, deben ser muy recientes.

No del todo. La *aplicación* es nueva y algunas de las matemáticas son a la vez nuevas y sofisticadas, pero la parte en la que estoy pensando tiene unos 175 años de antigüedad. Y es algo que nunca se tuvo la intención de emplear para hacer imágenes por ordenador: No había ninguno en esa época.

Sí que *estaba* pensada para abordar un problema más general, independiente de cualquier hardware: la geometría en el espacio tridimensional. Desde una perspectiva actual, el potencial para relacionarla con las imágenes por computadora es evidente. Pero no tenía el aspecto de geometría. Parecía álgebra. Excepto por el hecho de que violaba una de las reglas básicas del álgebra, de modo que ni siquiera era eso. La trajo a la vida el prodigio irlandés de las matemáticas, sir William Rowan Hamilton, que bautizó a su criatura como *cuaternión*. Resulta irónico que esto no fuera lo que buscaba en realidad y hay un buen motivo para ello.

Lo que buscaba no existe.

*

Hoy en día, hay en el planeta más ordenadores que personas. La población de seres humanos está por encima de los 7600 millones. Solo portátiles, hay más de 2000 millones. Y hay casi 9000 millones de teléfonos inteligentes y tabletas, que tienen a menudo más capacidad de cálculo que el mejor superordenador que se podía encontrar en el mercado en la década de 1980[55]. Si se consideran los minúsculos procesadores que los fabricantes se apresuran a meter con calzador en cualquier lavavajillas, tostadora, nevera, lavadora y gatera del planeta, el número de ordenadores supera al de humanos cuatro a uno.

Es difícil darse cuenta de que no fue siempre así. El ritmo de las innovaciones ha sido frenético. Los primeros ordenadores domésticos (Apple II, TRS-80, Commodore PET) llegaron al mercado en 1977, hace apenas algo más de 40 años. Casi desde el principio, uno de sus usos principales fue jugar. Los gráficos eran rudimentarios y los juegos muy sencillos. Algunos consistían solo en mensajes de texto: «Estás en una mazmorra de pasadizos enrevesados, todos diferentes». Al que seguía una frase todavía más siniestra: «Estás en una mazmorra de pasadizos enrevesados, todos iguales».

Conforme los procesadores se hacían más rápidos, las memorias se volvían casi infinitas y los precios se desplomaban, las imágenes generadas por computadora eran cada vez más convincentes, hasta el punto en que

empezaron a dominar la industria cinematográfica. El primer largometraje animado producido por completo por ordenador fue *Toy Story* en 1995, aunque ejemplos más breves se remontan a una década antes. Los efectos especiales se han vuelto ahora muy realistas y su uso es tan generalizado que apenas se nota que están ahí. Cuando Peter Jackson rodó la trilogía de *El señor de los anillos* no se preocupó por la iluminación. Esta se arreglaba después, en la posproducción por ordenador.

Los gráficos de alta calidad y movimiento rápido son tan habituales que apenas nadie se pregunta de dónde sale todo esto. ¿Cuándo apareció el primer videojuego? Treinta años antes que los ordenadores domésticos. En 1947, los pioneros de la televisión Thomas Goldsmith Jr. y Estle Ray Mann registraron una patente de un «aparato de diversión con tubo de rayos catódicos». Un tubo de rayos catódicos es una botella de vidrio corta y gruesa que tiene una base amplia que se curva con suavidad (la pantalla) y un cuello estrecho. Un dispositivo en el cuello dispara electrones sobre la pantalla y unos electroimanes controlan la dirección de este haz, que barre la pantalla en una serie de pasadas horizontales, como un ojo humano que lee una página de texto. Allí donde las partículas cargadas golpean la parte frontal del tubo, producen una fluorescencia en un revestimiento especial que crea un punto brillante de luz. La mayoría de los aparatos de televisión empleaban este sistema, hasta que aparecieron los televisores de pantalla plana en el mercado, en torno a 1997. El juego de Goldsmith y Mann se inspiraba en las pantallas de radar de la segunda guerra mundial. El punto de luz representaba un misil y el jugador intentaba que alcanzase los objetivos, que estaban dibujados en papel y pegados a la pantalla.

Para 1952, la unidad central EDSAC había conquistado las vertiginosas cumbres del tres en raya. El gran éxito de ventas fue *Pong*, un videojuego para máquinas recreativas fabricado por Atari que era una versión bidimensional y simplificada del tenis de mesa, con una bola que rebotaba en dos palas, una para cada jugador. Para los estándares actuales, los gráficos eran muy básicos: dos rectángulos que se podían desplazar para las palas y un cuadrado móvil para la bola. Casi no había acción, pero hasta que se pudo disponer de tecnología más avanzada, *Pong* fue la vanguardia de los videojuegos.

No hace falta insistir en que Hamilton no podía albergar la intención de que su criatura matemática se emplease de este modo. Esa idea tardó otros 142 años en dar fruto. Pero es fácil ver *a posteriori* que esta posibilidad era inherente al tipo de problema que su descubrimiento pretendía resolver. Hay

matemáticas de muchos estilos. Quienes se dedican a ellas pueden ser resolutivos, determinados a dar respuesta a una cuestión concreta, sea en el mundo real o en el mental de las matemáticas puras. Pueden ser teóricos, organizadores de incontables teoremas especiales en el seno de un marco unificado. Pueden ser inconformistas, vagabundos errantes de un campo al otro, que trabajan en lo que les llama la atención. O pueden ser fabricantes de herramientas y crear nuevos instrumentos que resulten útiles cuando se abordan cuestiones que todavía no se han planteado: un método en busca de una aplicación.

Buena parte de la reputación de Hamilton descansa en su trabajo como teórico, aunque los cuaterniones son un ejemplo de su pericia como fabricante de herramientas. Los inventó para proporcionar una estructura algebraica a los cálculos sistemáticos acerca de la geometría del espacio tridimensional.

*

Hamilton nació en Dublín, Irlanda, en 1805 y era el cuarto de nueve hijos. Su madre fue Sarah Hutton y su padre era Archibald Hamilton, un abogado. Cuando William tenía tres años le enviaron a vivir con su tío James, que dirigía una escuela. Demostró tener un talento precoz para los idiomas y parece que también fue autodidacta en buena medida en matemáticas. Esta fue la disciplina que estudió en el Trinity College de Dublín a partir de los 18 años y en la que obtuvo unas calificaciones estratosféricas. John Brinkley, obispo de Cloyne, afirmó que «este joven, no diré qué *va a ser*, sino que *es*, el matemático más destacado de su época». Estaba en lo cierto y en 1837, cuando todavía no se había licenciado, Hamilton obtuvo la cátedra Andrews de astronomía y fue nombrado astrónomo real de Irlanda. Pasó el resto de su vida profesional en el observatorio Dunsink, cerca de Dublín.

Sus contribuciones más célebres se produjeron en óptica y en dinámica, sobre todo por el descubrimiento de una relación notable entre estas dos ramas separadas de la física matemática. Hamilton las reformuló en términos de un concepto matemático común: la función principal. Hoy en día se denomina «hamiltoniano» y ha llevado a avances destacados en ambos campos. Más tarde, resultó que era justo lo que se necesitaba para la muy extraña teoría de la mecánica cuántica, cuando esta apenas había echado los dientes.

Ya se ha hablado de Hamilton brevemente en el capítulo anterior. En 1833 resolvió un dilema cuasi filosófico que tenía siglos de antigüedad y despojó a

los números complejos de su halo de misterio, los expuso como impostores y dejó claro que su aparente novedad era el resultado de un ingenioso disfraz, mientras que su verdadera naturaleza era casi trivial. Un número complejo, dijo Hamilton, no es ni más ni menos que un par ordenado de cantidades reales, pertrechado con una lista concreta de reglas para sumar y multiplicar pares. También se ha dicho que esta solución del rompecabezas llegó demasiado tarde para impresionar a nadie y que cuando a Gauss se le había ocurrido la misma idea, ni siquiera se había tomado la molestia de publicarla. En cualquier caso, la forma de pensar de Hamilton en torno a los números complejos demostró ser de gran valor porque le llevó a la creación de los cuaterniones.

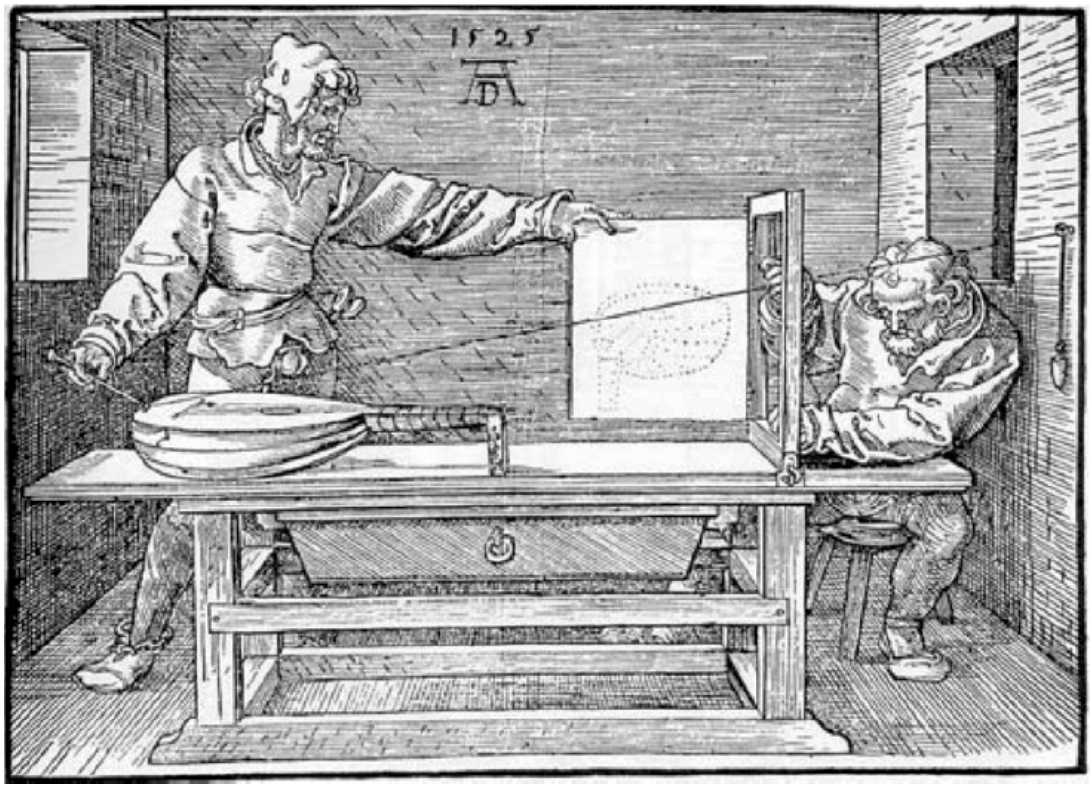
Por estos y otros adelantos matemáticos, se le nombró caballero en 1835. Los cuaterniones llegaron más tarde y cuando lo hicieron, pocas personas, aparte del mismo Hamilton y de unos cuantos discípulos, apreciaron su importancia. Creo que mientras estuvo vivo, la mayoría de los matemáticos y de los físicos veían su promoción entusiasta de la causa de los cuaterniones como una especie de obsesión. No del todo una chifladura, pero algo peligrosamente cercano. Se equivocaban. Su nueva invención desató una revolución y llevó las matemáticas a un territorio virgen e inexplorado. Puede entenderse por qué la mayoría no percibía su potencial, pero Hamilton sabía que tenía algo importante entre manos. Su territorio virgen ofrece todavía hoy nuevas y tentadoras ideas.

*

Algunas preguntas por las que se preocupan muy pocos jugadores o espectadores de cine son: ¿cómo funcionan los gráficos? ¿Cómo se crean estas ilusiones? ¿Qué las hace tan convincentes? Vale, no hace falta saberlo para pasarlo bien al jugar el videojuego o al ver la película. Pero el desarrollo histórico, las técnicas que tuvieron que inventarse para hacerlo posible y las empresas que se especializan en IGC y que desarrollan juegos, necesitan a muchas personas muy formadas que sepan cómo funcionan los diferentes trucos, en un detalle técnico considerable, y que tengan el dominio y la creatividad suficientes para inventar otros nuevos. No es un sector en el que se puedan quedar dormidos en los laureles.

Los principios geométricos básicos se conocen desde hace al menos 600 años. Durante el Renacimiento italiano, pintores destacados empezaron a comprender la geometría del dibujo en perspectiva. Estas técnicas permitieron

a los artistas crear imágenes realistas de un mundo tridimensional sobre un lienzo de dos dimensiones. El ojo humano hace algo muy parecido, con la retina en el lugar de la tela. Es difícil dar una descripción completa, pero en pocas palabras, el pintor *proyecta* una escena real sobre un lienzo plano al trazar una recta desde cada punto de la escena hasta otro que representa el ojo del espectador y hace una marca en el lugar en el que esta línea corta la superficie. El fantástico grabado de Alberto Durero titulado *Hombre dibujando un laúd* es una representación muy gráfica de este procedimiento.



Hombre dibujando un laúd, de Alberto Durero. Ilustra una proyección desde el espacio tridimensional a un lienzo de dos dimensiones.

Esta descripción geométrica puede convertirse en una fórmula matemática sencilla que transforma las tres coordenadas de un punto en el espacio en las dos que tiene la imagen correspondiente sobre la tela. Para aplicarla solo hay que conocer las posiciones del lienzo y del ojo del espectador con relación a la escena. Por motivos prácticos no se aplica esta transformación, denominada una proyección, a *todos* los puntos del objeto, sino a suficientes para obtener una buena aproximación. Este aspecto es visible en el grabado, que muestra un conjunto de puntos con forma de laúd y no el perfil completo del mismo. Esta colección de puntos se puede «cubrir» después con detalles más pequeños, tales como la paja de un techo, las ondas en el riachuelo y, por

supuesto, sus colores, mediante el empleo de técnicas en las que no voy a entrar porque necesitaría otro libro.

Esto es lo que ocurre en esencia cuando se muestra la aldea desde el punto de vista del dragón. El ordenador ya tiene coordenadas representativas de todas las características importantes del poblado almacenadas en la memoria. La retina del dragón hace el papel del lienzo. Si se sabe dónde está la bestia y con qué ángulo mira, es posible emplear la fórmula para calcular lo que vería. Así se obtiene un fotograma de la película, que muestra la aldea en un instante concreto. En el siguiente plano, esta sigue en el mismo sitio, pero el dragón y su retina se han movido. Se determina a dónde han ido, se repite el cálculo y se obtiene otra imagen más. Si se sigue la trayectoria de la criatura por el cielo, es posible construir fotograma a fotograma la película de lo que ve.

Esta descripción no es literal, por supuesto, solo es la idea principal subyacente. Hay trucos especiales para hacer que los cálculos sean más eficaces y se ahorre tiempo de procesado en el ordenador. Por sencillez, se pasarán por alto.

El mismo tipo de cálculo es válido para las secuencias del dragón que llega, visto desde el suelo. Ahora se necesita otro conjunto de puntos para especificar dónde se halla en su vuelo y la pantalla sobre la que se proyecta todo está en el suelo y no en la bestia. Para ser más concretos, limitémonos al punto de vista del dragón. En este caso, el ojo está fijo y es la *aldea* la que parece moverse. Conforme la criatura se abalanza hacia el suelo, todo en el poblado parece hacerse más grande, inclinarse y girar a imitación de sus propios movimientos. Cuando se eleva hacia las nubes, la aldea se reduce. La perspectiva no puede dejar de ser convincente en ningún momento de la secuencia y la clave matemática para lograrlo es considerar la aldea como un objeto *rígido* (y bastante complejo). Para hacerse una idea de lo que esto implica se puede pretender ser el dragón, sostener algún objeto delante de los ojos y moverlo hacia delante y hacia atrás, y darle vueltas en un sentido y en otro.

Ahora todo se representa en el «sistema de referencia del dragón», que es fijo *con relación a la criatura*. La aldea se mueve como un cuerpo rígido, lo que en sentido matemático quiere decir que la distancia entre dos puntos cualesquiera no cambia. Sin embargo, el conjunto total puede desplazarse por el espacio. Hay dos tipos básicos de movimiento: de traslación y de rotación. En el primero, el cuerpo se desplaza en alguna dirección sin girar ni desviarse. En el segundo, rota en torno a una línea fija, el eje, y todos los puntos se

desplazan el mismo ángulo en un plano que es perpendicular a esta recta. El eje puede ser cualquier línea en el espacio y el ángulo tener cualquier tamaño.

Todo movimiento rígido es una combinación de una traslación y de una rotación (aunque la primera puede ser de una distancia nula y la segunda podría ser de un ángulo nulo, en cuyo caso estas transformaciones no tendrían efecto alguno). En realidad, esto no es cierto del todo: otro movimiento rígido posible es la reflexión, que se comporta como un espejo. Pero no hay manera de obtener reflexiones mediante movimientos continuos, así que pueden pasarse por alto.

Ya se ha dado el paso principal para convertir dragones voladores en matemáticas. Lo que es necesario comprender es cómo cambian las coordenadas de un punto en el espacio cuando se aplica una traslación o una rotación. Una vez que se ha conseguido esto, es posible emplear la fórmula habitual para proyectar el resultado sobre una pantalla plana. Al final, las traslaciones son fáciles. La pesadilla son las rotaciones.

*

Todo es mucho más sencillo en dos dimensiones, en un plano. Euclides formalizó la geometría plana en torno al año 300 a. C. No obstante, no lo hizo mediante el empleo de movimientos rígidos. En su lugar, utilizó triángulos congruentes, que son aquellos de tamaño y forma idénticos, pero en posiciones diferentes. Al llegar el siglo XIX, los matemáticos habían aprendido a interpretar estos pares de triángulos como movimientos rígidos, transformaciones del plano que mueven el primero de ellos a la posición del segundo. Georg Bernhard Riemann *definió* la geometría en términos de tipos específicos de transformaciones.

Por una ruta muy diferente, los matemáticos también habían ideado maneras eficaces de calcular movimientos rígidos en el plano. Fue un inesperado efecto secundario de un nuevo desarrollo en el álgebra, que ya se ha visto en el capítulo anterior: los números complejos. Para trasladar (deslizar) una forma, tal como el CERDO ilustrado en la página 150, se suma una cantidad compleja fija a cada punto del conjunto. Para hacerla rotar un ángulo A, se multiplica cada uno de ellos por e^{iA}. La guinda en el pastel es que los números complejos son ideales para resolver las ecuaciones diferenciales de la física… pero solo en el espacio bidimensional.

Todo eso llevó a Hamilton a concebir una idea que se convirtió en una obsesión. Dado que los complejos son tan eficaces en la física en dos

dimensiones, debería haber unos números «supercomplejos» análogos que hicieran lo mismo en tres dimensiones. Si fuese capaz de encontrar un nuevo sistema de elementos como esos, se abrirían de par en par las puertas a toda la física de lo existente. Incluso era evidente por dónde empezar. Dado que los números complejos son *pares* de cantidades reales, estos supercomplejos hipotéticos deberían ser *tríos* de cantidades reales. Una para cada dimensión. La fórmula para sumar estos tríos (o tripletes, como Hamilton los llamaba a menudo) era evidente: bastaba con sumar las componentes correspondientes. Ya están resueltas las traslaciones. Ahora todo lo que tenía que hacer era descubrir cómo multiplicarlos. Pero todo lo que intentaba fracasaba y para 1842 estaba tan obsesionado con este obstáculo que incluso sus hijos se dieron cuenta. Cada día le preguntaban: «Papá, ¿puedes multiplicar tripletes?». Y cada día, Hamilton sacudía la cabeza con amargura. Sumarlos o restarlos, sí, pero no multiplicarlos, de ninguna manera.

A menudo es difícil saber la fecha exacta en la que se ha producido un gran avance matemático porque hay una «prehistoria», larga y confusa, durante la cual los investigadores se fueron acercando poco a poco hacia el descubrimiento final. Pero en ocasiones se conocen el sitio y el momento exactos. En este caso, la fecha crucial fue el lunes 16 de octubre de 1843 y el lugar fue Dublín. Incluso puede hacerse una conjetura bastante acertada de la hora, porque Hamilton, que entonces era presidente de la Real Academia de Irlanda, paseaba con su esposa a lo largo de un sendero junto a un canal, de camino a una reunión del Consejo de la Academia. Mientras descansaba un momento en el puente de Brougham Bridge, cruzó su mente la solución al problema que le acongojaba desde hacía años y la grabó en la mampostería con su cortaplumas:

$$i^2 = j^2 = k^2 = ijk = -1$$

La inscripción se ha borrado desde entonces, pero cada año un grupo de científicos y matemáticos recorren el «paseo Hamilton» para mantener vivo el recuerdo.

Sin más explicaciones, no hay esperanza alguna de entender esta inscripción. Incluso *con* ellas, puede parecer un despropósito inútil a primera vista, pero eso ocurre a menudo con los grandes avances matemáticos. Se tarda en asimilarlos. Si el descubrimiento hubiesen sido los números complejos, Hamilton habría grabado una regla muy sencilla: $i^2 = -1$. Esta ecuación encierra la clave de todo el sistema de los complejos. Todo lo demás

se sigue a partir de aquí si se insiste en que sigan siendo válidas las reglas habituales de la aritmética. Añádanse j y k, junto con i, y la fórmula de Hamilton define un sistema más extenso de números o, si se prefiere, de objetos parecidos a números. Los denominó *cuaterniones* porque tienen cuatro componentes, cada una de ellas una cantidad real convencional. Estos elementos integrantes son un real tradicional, un múltiplo real de un número denominado i, que se comporta de la misma manera que la unidad imaginaria habitual que lleva ese símbolo, y dos nuevas componentes: un múltiplo real de un número denominado j y un múltiplo real de un número denominado k. Un cuaternión típico es por lo tanto una combinación $a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$, en la que a, b, c y d son cuatro cantidades reales ordinarias. O, para eliminar cualquier misterio, un cuarteto (a, b, c, d) de números reales que obedece a un listado breve de reglas aritméticas.

Al día siguiente de su pequeño acto de vandalismo, Hamilton escribió a su amigo, el matemático John Graves: «Allí caí en la cuenta de que debemos admitir, en cierto sentido, una *cuarta dimensión* del espacio para poder calcular con tríos». En una carta a su padre, escribió: «Pareció cerrarse un circuito eléctrico y saltó una chispa». Tenía más razón de lo que entonces podía saber, porque hoy en día su descubrimiento desempeña un papel vital en miles de millones de circuitos eléctricos que producen trillones de pequeñas chispas. Se los conoce como PlayStation 4, Nintendo Switch y Xbox y se emplean para jugar videojuegos tales como *Minecraft*, *Grand Theft Auto* y *Call of Duty*.

Ahora se comprende por qué Hamilton tuvo tantos problemas al intentar multiplicar tríos. No puede hacerse. Había dado por sentado que debían cumplirse todas las reglas habituales del álgebra y, en particular, que se podría dividir por cualquier cantidad distinta de cero. Pero daba igual la fórmula que probase, no conseguía obedecer todas las condiciones necesarias. Más tarde, los especialistas en álgebra demostraron que este requisito es lógicamente contradictorio. Si se quieren cumplir todas las reglas, no es posible ir más allá de los números complejos. Se queda uno atrapado en las dos dimensiones. Si se juguetea con las fórmulas de Hamilton y se asume que se cumple la propiedad asociativa, enseguida se descubre que él ya había descartado una de estas reglas, la propiedad conmutativa de la multiplicación. Por ejemplo, sus fórmulas implican que $\mathrm{ij} = \mathrm{k}$, mientras que $\mathrm{ji} = -\mathrm{k}$.

Hamilton tuvo la intuición de abandonar esta condición pese a que era problemático, como poco. Sin embargo, ahora se sabe que ni siquiera así es posible construir un sistema numérico autocontenido de tríos. Un hermoso

teorema de Adolf Hurwitz, publicado de manera póstuma en 1923, afirma que los números reales, los complejos y los cuaterniones son las «únicas álgebras de división normadas». Es decir, el truco solo funciona con una, dos o cuatro componentes reales, *pero no con tres*. De estos, solo los reales y los complejos obedecen la propiedad conmutativa. Si se debilita la propiedad asociativa también puede obtenerse un sistema de ocho componentes, denominados octoniones o números de Cayley. La siguiente cantidad natural de componentes sería 16, pero ahora falla incluso la forma débil de la asociatividad. Eso es todo. No hay otra posibilidad en ese sentido. Es una de esas curiosidades extrañas que a veces tienen las matemáticas: en este contexto, el siguiente elemento de la secuencia 1, 2, 4, 8… no existe.

De modo que el pobre sir William pasó años de infructuoso esfuerzo en busca de un logro imposible. Al final, su avance se basó en el abandono de *dos* principios fundamentales: que la multiplicación debería ser conmutativa y que el sistema numérico «correcto» para la física tridimensional debería tener tres componentes. Merece un profundo reconocimiento por darse cuenta de que para progresar tenía que abandonar ambos.

*

El nombre que dio Hamilton a su nuevo sistema, cuaterniones, es un reflejo de su relación con las cuatro dimensiones. Promovió su empleo en muchas ramas de las matemáticas y de la física y demostró que un tipo especial de cuaternión, la «parte vectorial» $b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$, es capaz de representar el espacio tridimensional de una forma elegante. No obstante, se pasaron de moda cuando apareció un sistema más sencillo, el álgebra vectorial. Retuvieron su interés en las matemáticas puras y en la física teórica, pero no estuvieron a la altura de las expectativas que se había formado su creador respecto a sus usos prácticos. Eso fue así, claro, hasta que aparecieron los videojuegos y las IGC en la industria cinematográfica.

La relación con los cuaterniones surge porque los objetos de IGC tienen que rotar en el espacio tridimensional. El mejor método para conseguirlo se basa en los cuaterniones de Hamilton. Estos proporcionan una herramienta algebraica sencilla para calcular los efectos de las rotaciones con rapidez y precisión. Hamilton estaría sorprendido porque las películas no existían en su época. Las matemáticas antiguas pueden adquirir usos radicalmente nuevos.

La propuesta de emplear cuaterniones en los gráficos por ordenador aparece en un artículo de 1985 de Ken Shoemake, «Animating rotation with

quaternion curves»[56]. El texto empieza con la afirmación de que «los cuerpos sólidos giran y ruedan a través del espacio. En las animaciones por ordenador, también lo hacen las cámaras. La mejor manera de describir las rotaciones de estos objetos es emplear un sistema de cuatro coordenadas, los cuaterniones». Shoemake prosigue con la afirmación de que estos tienen la ventaja principal de permitir «intermediaciones» sin saltos; es decir, la interpolación de imágenes entre dos puntos finales dados.

Antes de entrar en detalle, merece la pena considerar unas pocas características de la animación por ordenador que motivan este enfoque. Así se hace mucho más sencillo este análisis y también se emplean muchas otras técnicas. Una película o una imagen animada en una pantalla es en realidad una serie de planos fijos mostrados en rápida sucesión para crear la ilusión del movimiento. En la época temprana de la animación (piense en los dibujos de Walt Disney) los artistas dibujaban cada una de estas imágenes fijas como una obra de arte individual. Crear movimientos realistas (tanto como lo puede ser un ratón que habla) requería una gran habilidad. Podían emplearse varios trucos para simplificar el proceso, tales como tener un único fondo que permanecía igual a lo largo de una secuencia y superponer los objetos que cambiaban.

Este método es muy laborioso y poco práctico para batallas trepidantes en el espacio o para cualquier otra animación de alta calidad. Piénsese en alguien que tiene que animar una escena de una película o de un videojuego en la que interactúan varias naves espaciales. Un artista gráfico ha diseñado (en un ordenador) cada una de ellas por anticipado. Han quedado representadas por conjuntos fijos de puntos en el espacio, unidos entre sí para formar una red de triángulos minúsculos. A su vez, estos pueden representarse mediante listados adecuados de números: las coordenadas de los puntos y la información de cuál está conectado con cuál. Un programa informático es capaz de «renderizar» este conjunto de números (y otros datos, tales como el color) para crear una imagen bidimensional de la nave. Esta muestra el aspecto que va a tener cuando se sitúe en alguna posición de referencia y se vea desde alguna ubicación concreta.

Para hacer que la nave se mueva, el animador cambia esas cantidades de una manera adecuada. Por ejemplo, para llevarla a una nueva ubicación se añade un trío fijo de números (el vector de desplazamiento) a todos los puntos, mientras que las uniones entre estos siguen como antes. Después se renderiza este nuevo listado para obtener la siguiente imagen fija, etcétera. Sumar un vector es sencillo y rápido, pero los objetos también pueden rotar

en el espacio. Son capaces de girar en torno a cualquier eje y es posible que este cambie conforme se desplaza el objeto. Las rotaciones también cambian los listados de números, pero de maneras más complicadas.

Muy a menudo, el animador sabe dónde empieza el movimiento (digamos que en el suelo) y a dónde tiene que llegar (en línea recta, de cara a una luna distante). Es fundamental que la posición sea precisa en la pantalla bidimensional, porque eso es lo que ve el espectador. Debe tener una apariencia artística o emocionante adecuada. De modo que estas dos ubicaciones, principio y fin, están representadas por dos listados de números calculados con cuidado. Si el movimiento preciso entre los puntos es menos crucial, es posible dar instrucciones al ordenador para que haga interpolaciones entre los extremos. Es decir, se combinan los dos listados mediante una regla matemática que representa la transición de uno a otro. Por ejemplo, promediar cada pareja de coordenadas correspondientes proporciona un objeto a medio camino entre el principio y el final. No obstante, eso es demasiado burdo para ser aceptable. Es habitual que distorsione la forma de la nave.

El truco es emplear movimientos rígidos en el espacio para hacer la interpolación. Podría empezarse por trasladar la nave al punto medio y rotarla 45°. Si se hace esto de nuevo, estará en la ubicación final adecuada y habrá girado 90°. Para una ilusión de desplazamiento continuo, puede trasladarse 1/90 parte de la diferencia en posiciones y rotar 1° cada vez de manera repetida. En la práctica se emplean pasos mucho más pequeños.

De un modo más abstracto, podemos pensar este procedimiento en términos del «espacio de configuración» de todos los movimientos rígidos. Cada elemento de este espacio corresponde a un movimiento rígido único y los puntos cercanos representan movimientos parecidos. De modo que una secuencia de movimientos, cada uno próximo al anterior, corresponde a una secuencia de puntos, cada uno cercano al anterior. Unir estos entre sí en orden produce una trayectoria poligonal en el espacio de configuración. Si los pasos se hacen muy pequeños, se obtiene una trayectoria continua. Así que ahora el problema de intermediar desde una imagen inicial a otra final se ha reformulado como el de encontrar una trayectoria a través del espacio de configuración. Si se quiere que las transiciones no sean abruptas, esta debería ser suave, sin curvas cerradas. Hay buenos métodos para hacer que un polígono sea más suave.

Este espacio de configuración tiene seis «dimensiones» (es decir, coordenadas necesarias para definir un punto en él). Hay tres para las

traslaciones: una para cada una de las direcciones norte-sur, este-oeste y arriba-abajo. Después, se necesitan dos más para definir la posición de un eje de rotación y una última para el ángulo de la rotación. De modo que lo que empezó como un problema sobre el movimiento sin saltos de un objeto en tres dimensiones, se ha convertido en el de mover un punto a lo largo de una trayectoria suave en seis dimensiones. Este problema de la animación reformulado puede abordarse con el empleo de técnicas de la geometría multidimensional para diseñar trayectorias adecuadas.

*

En las matemáticas aplicadas, la manera tradicional de tratar las rotaciones de un objeto rígido se remonta a Euler. En 1752, demostró que todo movimiento rígido que no refleja el objeto es, o bien una traslación o bien una rotación en torno a algún eje[57]. No obstante, para los cálculos él combinaba tres giros en torno a tres ejes en la representación de coordenadas habitual del espacio, un método que ahora se denomina ángulos de Euler. Como ejemplo, Shoemake consideraba la orientación de una aeronave, que en aeronáutica viene especificada por tres ángulos:

- Guiñada, en torno a un eje vertical, que da la dirección de la aeronave en un plano horizontal.
- Cabeceo, rotación en torno a un eje horizontal que atraviesa las alas.
- Alabeo, rotación en torno a la recta que va del morro a la cola.

El primer problema con este tipo de representación es que el orden en el que se aplican las componentes es vital. Las rotaciones no son conmutativas. El segundo es que la elección de los ejes no es única y que diferentes campos de aplicación emplean opciones diferentes. Un tercero es que las fórmulas para combinar dos rotaciones sucesivas, expresadas en ángulos de Euler, es complicada en extremo. Estas características no causan demasiados problemas en aplicaciones aeronáuticas básicas, que tratan en buena medida de las fuerzas que actúan sobre la aeronave cuando está en una orientación dada, pero resultan incómodas para la animación por ordenador, en la que los objetos experimentan secuencias completas de movimientos.

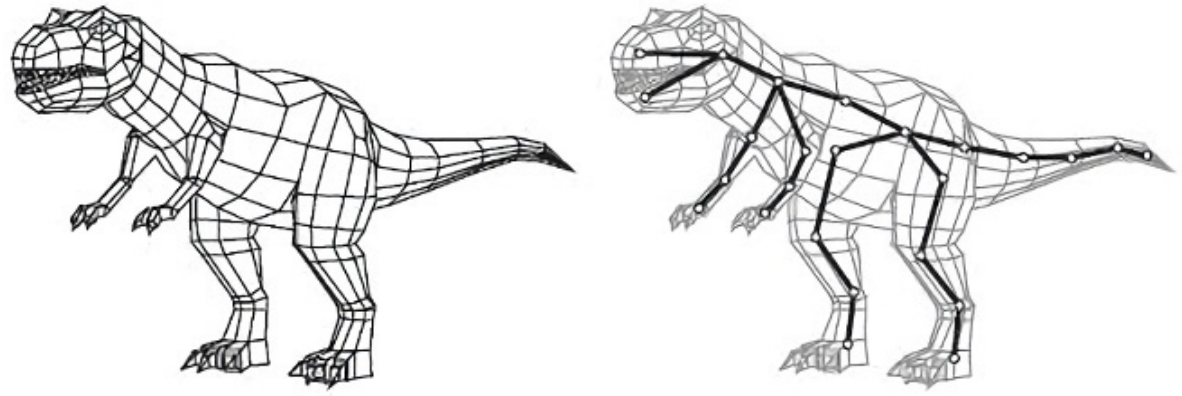
Shoemake afirmaba que los cuaterniones, aunque eran menos directos, proporcionaban una forma de especificar las rotaciones que es mucho más conveniente para los animadores, sobre todo en lo que se refiere a la intermediación. Un cuaternión $a + b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$ se divide en una parte escalar a

y en otra vectorial $v = b\mathrm{i} + c\mathrm{j} + d\mathrm{k}$. Para rotar un vector v en un cuaternión q, se multiplica v por q^{-1} a la izquierda y por q a la derecha para obtener $q^{-1}vq$. Sea lo que sea q, el resultado es de nuevo un vector con una parte escalar nula. Cabe destacar que las reglas de Hamilton para multiplicar cuaterniones demuestran que *cualquier* rotación corresponde a uno solo de estos elementos. La parte escalar es el coseno de la mitad del ángulo en que gira el objeto y la parte vectorial apunta a lo largo del eje de rotación y tiene una longitud equivalente al seno de la mitad de ese ángulo. De modo que el cuaternión encierra con acierto toda la geometría de la rotación, con el pequeño inconveniente de que las fórmulas naturales funcionan con la mitad del ángulo, no con el ángulo de manera directa[58].

Los cuaterniones evitan las distorsiones que pueden acumularse si se rota un objeto muchas veces, como debe hacerse a menudo. Los ordenadores son capaces de realizar cálculos exactos con números naturales, pero es imposible representar los reales con precisión absoluta, de modo que se acumulan pequeños errores. Con los métodos habituales para representar transformaciones, el objeto que se manipula cambia de forma ligeramente, algo que no escapa a la vista. Por el contrario, si se toma un cuaternión y se cambian los números un poco, el resultado no deja de ser un cuaternión ni de representar un giro porque *cada* uno de ellos corresponde a alguna rotación. Tan solo que es un poco diferente de la exacta. La vista es menos sensible a estos errores y pueden compensarse con facilidad si se hacen demasiado grandes.

*

Los cuaterniones son una forma de crear movimientos realistas en tres dimensiones, pero lo que he descrito hasta el momento sirve para objetos rígidos. Tal vez naves espaciales, pero no dragones. Estos son *flexibles*. De modo que, ¿cómo se hace un dragón realista en IGC? Hay un método común que es válido no solo para dragones, sino casi para cualquier cosa y lo explicaré para el caso de un dinosaurio porque tengo imágenes adecuadas. Este enfoque reduce el movimiento de un objeto flexible al de un conjunto de objetos rígidos unidos. Se emplea el método que se quiera para estos últimos, con ajustes adicionales para unirlos entre sí de manera correcta. En particular, si se utilizan cuaterniones para rotarlos y trasladarlos, los mismos métodos pueden adaptarse para que funcionen con un dinosaurio flexible.



Izquierda: *Tyrannosaurus rex* tosco en malla de polígonos. *Derecha*: malla unida a un esqueleto rudimentario.

El primer paso es crear un modelo digital tridimensional del dinosaurio, en el que su superficie es una complicada malla de polígonos planos: triángulos, rectángulos y cuadriláteros menos regulares. El *software* empleado para esta tarea muestra la forma de manera geométrica y es posible moverla, rotarla, ampliar los detalles, etcétera y cada movimiento se muestra en la pantalla del ordenador. No obstante, lo que manipula el programa no es la geometría en sí misma, sino un listado de coordenadas numéricas de los puntos en los que se unen los polígonos. De hecho, las matemáticas que emplea el *software* para ayudar a dibujar el dinosaurio son casi las mismas que se usan para animar el resultado. La principal diferencia es que en esta etapa el animal está fijo y es el punto de vista el que se rota y se traslada. En la animación, el punto de vista puede estar fijo mientras se mueve el dinosaurio o cambiar también, como en el caso del dragón que desciende.

Ya se ha conseguido un dinosaurio tosco y rígido. ¿Cómo se hace para que se mueva? No es necesario hacer como los artistas en los días de Mickey Mouse: volver a dibujar la imagen con el personaje en una posición un poco diferente y repetirlo cientos de veces. Es deseable que el ordenador haga todo el trabajo repetitivo. De modo que el dinosaurio se reduce a un esqueleto rudimentario, un pequeño número de varillas rígidas («huesos») conectadas en sus extremos que se hacen pasar a través del cuerpo, los miembros, la cola y la cabeza. No es un diseño correcto en sentido anatómico, sino una estructura que permite doblar las partes principales del animal. También se representa este esqueleto como un listado de coordenadas de los dos extremos de cada hueso.

Una manera muy eficaz de conseguir animaciones realistas, en especial de personas o de criaturas humanoides, es la captura de movimiento. Un actor realiza los movimientos requeridos delante de una cámara o de varias para

obtener datos en tres dimensiones. Se pegan puntos blancos a partes importantes de su cuerpo, tales como pies, rodillas, caderas y codos y el ordenador analiza un vídeo del actor para extraer la forma en que se mueven los puntos. Los datos de estos últimos se emplean para animar el esqueleto. Así es como se hizo la animación de Gollum en la trilogía de *El señor de los anillos*. Por supuesto, si se quieren movimientos impropios de humanos (pero realistas), el actor tiene que moverse de forma convenientemente extraña.

Sea como sea que se anima el esqueleto, una vez que el resultado es satisfactorio hay que «cubrirlo» con la malla. Es decir, se combinan los dos listados de coordenadas y se especifican las uniones adicionales entre las posiciones de los huesos y de las porciones de malla que los rodean. Después, durante una buena parte del proceso, se deja de lado la cobertura y se anima el esqueleto. Es aquí donde sale rentable el trabajo que ya se ha hecho con los movimientos rígidos, porque cada hueso es rígido y el fin es moverlo en tres dimensiones. También se deben imponer límites al movimiento, de modo que el esqueleto permanezca unido. Si se mueve un hueso, también deben hacerlo algunos extremos de los otros que están unidos a él, así que se desplazan las coordenadas de esos puntos a las posiciones adecuadas. Después se mueven también esos huesos de forma rígida, lo que por supuesto afecta a los que están unidos a *estos*… y un elemento tras otro se consigue que todo el esqueleto se doble un poco. Puede hacerse que mueva los pies para que camine, que agite la cola arriba y abajo o a los lados, que abra sus feroces mandíbulas… pero todo en el esqueleto. Esto es más sencillo, rápido y *económico* porque este armazón tiene muchas menos piezas.

Cuando se está satisfecho de que los huesos se muevan como se desea, puede resultar útil volver a cubrirlos con la malla, empezando por el primer fotograma de la secuencia. Entonces el *software* de animación hace que la cobertura siga el movimiento del esqueleto a lo largo de planos sucesivos sin que haya que hacer más trabajo, aparte de un clic o dos del ratón. Al hacerlo así, es posible comprobar que la animación no deja de parecer realista cuando el dinosaurio sigue a su esqueleto.

Ahora puede uno ponerse creativo. Desplazar la posición de la «cámara», el punto de vista empleado por el *software*, ampliar para obtener un primer plano, ver al dinosaurio que corre en la distancia… lo que sea. Es posible dar forma a otras criaturas, como una manada de herbívoros que huyen del monstruoso tiranosaurio. Una vez más, esto se hace en un primer momento con esqueletos que luego se cubren con mallas. Cada criatura puede animarse por separado y luego se juntan todas para crear una escena de caza.

Dado que los esqueletos no son más que monigotes de rayas, tal vez no se haya hecho nada hasta este momento para impedir que dos animales ocupen el mismo espacio. Hay más apaños de *software* que avisan de cualquier choque de este tipo. Cuando se cubren las estructuras con las mallas, los polígonos que se hallan al frente se solapan con los de atrás y, dado que los dinosaurios no son transparentes, hay que eliminar cualquier región que debiera estar oculta. Todo esto se hace mediante cálculos sencillos en geometría de coordenadas, aunque son muchos. No fue algo factible hasta que los ordenadores se volvieron rápidos de verdad. Hoy en día es rutinario.

Todavía hay más trabajo que hacer, porque un dinosaurio con la apariencia de un montón de polígonos no resulta muy impresionante que digamos. Hay que cubrir la malla con patrones de piel realistas, después resolver la información del color y tal vez crear texturas creíbles (el pelaje tiene un aspecto muy diferente a las escamas). Cada paso requiere un *software* diferente que aplica técnicas matemáticas distintas. Este paso se denomina renderización y monta la imagen final que aparece en la pantalla cuando se ve la película. Pero por debajo de todo ello hay miles de millones de cálculos que mueven puntos y aristas de manera rígida.

Además, estos métodos matemáticos tienen otra ventaja. Es posible decidir en cualquier etapa que hay algo que no es correcto y cambiarlo. Si se quiere un dinosaurio verde en lugar de uno marrón, no hay que dibujarlo todo de nuevo. Se emplea el mismo esqueleto y la misma malla, los mismos movimientos y la misma textura de la piel, pero se cambia el color.

Cuando se anima una película o un videojuego, hay equipos de expertos que emplean un conjunto de paquetes de *software* estándar que la industria ha desarrollado para llevar a cabo estos procesos. Para dar una idea de la complejidad de estas actividades, echaré un vistazo a algunas de las empresas y suites de programas que se emplearon para hacer la película *Avatar*.

La mayor parte de la animación la hizo la empresa Weta Digital en Nueva Zelanda, conocida por su trabajo en *El señor de los anillos* y *El hobbit*. Industrial Light & Magic, fundada por George Lucas en 1975 para los efectos especiales de la primera película de *La guerra de las galaxias*, creó 180 secuencias de este tipo, sobre todo la aeronave en la batalla final. El resto, en Reino Unido, Canadá y Estados Unidos, añadió detalles especiales imprescindibles, tales como pantallas en salas de control y mensajes de alerta en visores, para simular la tecnología del futuro. El programa Autodesk Maya se encargó de la mayoría de estas escenas y se empleó Luxology Modo para diseñar modelos, en particular el Escorpión. Con Houdini se crearon los

planos de la Puerta del Infierno y los interiores. Las criaturas extraterrestres se diseñaron con ZBrush. Autodesk Smoke realizó la corrección de color, Massive simuló la vegetación alienígena y Mudbox se encargó de las montañas flotantes. El arte conceptual inicial y las texturas se crearon con Adobe Photoshop. En conjunto, participaron una docena de empresas y emplearon 22 herramientas de *software* diferentes, además de innumerables complementos desarrollados de manera específica.

*

En la actualidad, se añaden unas matemáticas muy sofisticadas al cóctel de la animación de IGC. El propósito, siempre, es hacer el trabajo de los animadores tan sencillo como sea posible, obtener resultados realistas y reducir el coste y el tiempo. Hay que tenerlo todo, ahora y por poco dinero.

Supongamos, por ejemplo, que el estudio cinematográfico tiene una librería con animaciones de un dinosaurio que realiza varias secuencias de movimiento. En una, galopa hacia delante y recorre un «ciclo de marcha», una porción de un movimiento que se repite de forma periódica. En otra, salta en el aire y se cae. Se quiere crear una escena en la que corre tras un pequeño herbívoro y se abalanza sobre él. Una manera eficaz de empezar sería combinar entre sí una docena de ciclos de marcha para el galope y luego añadir el salto al final. Por supuesto, después se apañaría todo para que no fuese evidente que se repite la misma animación doce veces, pero este es un buen comienzo.

Tiene sentido combinar las secuencias entre sí al nivel del esqueleto. Todo lo demás, como cubrir con mallas y añadir colores y texturas, puede hacerse después. Se dan los pasos obvios: se unen doce copias del ciclo de marcha a una secuencia de un salto y se mira a ver cómo queda.

Fatal.

Las partes están bien por separado, pero no se ajustan de manera continua. El resultado va a saltos y es poco convincente.

Hasta hace poco, el único recurso hubiese sido modificar las uniones a mano y, tal vez, interpolar algunas partes adicionales al movimiento. Incluso así, habría sido complicado. Pero algunos desarrollos recientes en las técnicas matemáticas prometen resolver este problema de un modo mucho mejor. La idea es emplear métodos de alisado para rellenar los huecos y eliminar las transiciones abruptas. El paso principal es encontrar una buena manera de hacerlo para un solo hueso del esqueleto o, de modo más general, para una

única curva. Una vez que se ha resuelto ese problema, es posible combinar toda la estructura a partir de los componentes individuales.

La rama de las matemáticas que se ensaya en la actualidad se denomina teoría de formas. De modo que empecemos por la pregunta obvia: ¿qué es una forma?

En geometría ordinaria es posible encontrar muchas formas habituales: triángulo, cuadrado, paralelogramo o circunferencia. Cuando estas se interpretan en geometría de coordenadas, se convierten en ecuaciones. Por ejemplo, en el plano, los puntos (x, y) en una circunferencia unidad son precisamente los que cumplen la ecuación $x^2 + y^2 = 1$. Otra manera muy conveniente de representar esta figura es emplear lo que se llama un parámetro. Es una variable auxiliar, digamos t, que puede interpretarse como el tiempo, junto con fórmulas que describen el modo en que x e y dependen de t. Si este recorre algún intervalo numérico, cada valor de t arroja dos coordenadas $x(t)$ e $y(t)$. Cuando las fórmulas son correctas, estos puntos definen la circunferencia.

Las expresiones paramétricas habituales para un círculo son trigonométricas:

$$x(t) = \cos t,\ y(t) = \operatorname{sen} t$$

No obstante, también es posible cambiar la manera en que aparece el parámetro en las fórmulas sin dejar de obtener una circunferencia. Por ejemplo, si se pasa de t a t^3, entonces

$$x(t) = \cos t^3,\ y(t) = \operatorname{sen} t^3$$

también determina una circunferencia, que es la misma. Este efecto se produce porque el parámetro tiempo aporta más información que solo el modo en que varían x e y. En la primera expresión, el punto se mueve a velocidad constante conforme varía t. En la segunda no.

La teoría de formas es una manera de sortear esta falta de unicidad. Una forma es una curva, considerada como un objeto que no depende de fórmula paramétrica alguna en particular. De modo que dos curvas paramétricas definen la misma forma si puede cambiarse el parámetro y convertir una fórmula en la otra, como pasar de t a t^3. A lo largo del siglo pasado, los matemáticos han ideado un método estándar para conseguir este tipo de cosas.

Es poco probable que se le haya ocurrido a alguien más porque requiere un punto de vista más bien abstracto.

El primer paso es considerar, no solo una curva paramétrica, sino el «espacio» de *todas* las *posibles*. Entonces se dice que dos «puntos» en este espacio (es decir, dos curvas paramétricas) son equivalentes si puede pasarse de una a otra mediante un cambio de parámetro. Después se define una «forma» como toda una clase de equivalencia de curvas, el conjunto de todas las que son equivalentes a una dada.

Es una versión más general del truco que se emplea para hacer aritmética en un módulo. Por ejemplo, en los enteros módulo 5, el «espacio» son todos los enteros y dos de ellos son equivalentes si su diferencia es un múltiplo de 5. Hay cinco clases de equivalencia:

Todos los múltiplos de 5.
Todos los múltiplos de 5 a los que se suma 1.
Todos los múltiplos de 5 a los que se suma 2.
Todos los múltiplos de 5 a los que se suma 3.
Todos los múltiplos de 5 a los que se suma 4.

¿Por qué no seguir? Porque un múltiplo de 5 al que se suma 5 no es más que otro múltiplo de 5 un poco mayor.

En este caso, el conjunto de clases de equivalencia, que se denota por $\mathbb{Z}_5$, tiene mucha estructura útil. De hecho, en el capítulo 5 se mostró como buena parte de la teoría de números básica depende de esta estructura. Se dice que $\mathbb{Z}_5$ es el «espacio cociente» de los enteros módulo 5. Es lo que se obtiene cuando se pretende que los números que difieren en cinco unidades son idénticos.

Algo parecido ocurre para obtener el espacio de formas. En lugar de enteros, está el espacio de todas las curvas paramétricas. En lugar de cambiar números por múltiplos de 5, se cambia la fórmula paramétrica. De modo que se acaba con un «espacio cociente» formado todos los cambios de las curvas paramétricas módulo en el parámetro. Es probable que eso parezca carente de sentido, pero es un truco habitual cuyo valor se ha puesto en evidencia a lo largo de mucho tiempo. Un motivo por el que el espacio cociente es valioso, es porque supone la descripción natural de los objetos que interesan aquí. Otro es porque de manera habitual hereda una estructura interesante del espacio original.

Para el espacio de formas, el principal elemento de interés de la estructura es una medida de la distancia entre dos de ellas. Tómese una circunferencia y

defórmese un poco: se obtiene una curva cerrada, que es parecida a la circunferencia pero que es diferente. Si se deforma mucho se obtiene una curva cerrada que es, de manera intuitiva, *más* diferente, «más alejada». Esta intuición puede precisarse y es posible demostrar que el espacio de formas tiene un concepto razonable y natural de distancia: una métrica.

Una vez que un espacio tiene una métrica se pueden hacer todo tipo de cosas útiles. En particular, discernir los cambios continuos de los discontinuos e incluso subir la apuesta para distinguir los suaves de los que no lo son. Y, al fin, se vuelve al problema de combinar las secuencias de animación entre sí. Como mínimo, esta métrica en el espacio de formas permite detectar discontinuidades o falta de suavidad *en el ordenador*, al resolver los cálculos, en vez de hacerlo a ojo. Pero hay más.

Las matemáticas tienen muchas técnicas de alisado que son capaces de transformar una función discontinua en otra continua, o una que no es suave en otra que sí lo es. Se ha descubierto que se pueden aplicar estos métodos al espacio de formas. De modo que el ordenador consigue modificar de manera automática una secuencia de movimientos combinada que presente un salto repentino para eliminar la discontinuidad con solo hacer los cálculos adecuados. No es fácil, pero puede hacerse y de manera lo bastante eficaz como para ahorrar costes. Solo calcular la distancia entre dos curvas emplea métodos de optimización parecidos a los que se vieron en el problema de la persona viajante. Alisar una secuencia implica resolver una ecuación diferencial bastante parecida a la de Fourier para el flujo de calor, que se verá en los capítulos 9 y 10. Ahora puede persuadirse a toda una sucesión animada de curvas para que «fluyan» hacia una sucesión animada diferente, lo que elimina las discontinuidades (algo parecido de nuevo a un flujo de calor que alisa una onda cuadrada[59]).

Otras formulaciones abstractas parecidas hacen posible también convertir animaciones en otras similares, pero diferentes. Una secuencia que muestra un dinosaurio que camina puede apañarse para hacer que el animal corra. No se trata solo de acelerar la acción, porque la manera en que corre un animal es diferente, a ojos vistas, a cómo anda. Esta metodología está aún en pañales, pero sugiere con firmeza que un poco de pensamiento matemático de alto nivel podría dar grandes dividendos en películas de animación futuras.

Estas son solo algunas de las maneras en que las matemáticas contribuyen a la animación. Otras crean versiones simplificadas de procesos físicos para imitar las olas en el océano, los remolinos de nieve, las nubes y las montañas. El objetivo es obtener resultados realistas al mismo tiempo que los cálculos se

mantienen todo lo sencillos que sea posible. En la actualidad hay teorías matemáticas muy amplias acerca de la representación de caras de personas. En *Rogue One*, parte de la serie de *La guerra de las galaxias*, los actores Peter Cushing (que había fallecido en 1994) y Carrie Fisher (fallecida en 2016) fueron recreados de manera digital al cubrir las caras de unos dobles con las suyas. No era demasiado convincente y los seguidores protestaron con vehemencia. En *Los últimos Jedi* usaron un método mejor: se escogieron escenas descartadas de Fisher de películas anteriores, se unieron entre sí y se adaptó el guion para que sirviera. No obstante, aún se necesitaron muchas IGC para cambiar su vestuario, por consistencia. De hecho, casi todo, excepto su cara, se recreó de manera digital: su cabeza, su peinado, su cuerpo y sus ropas**[60]**.

Ya se emplean las mismas técnicas para crear vídeos ultrafalsos como propaganda política. Se graba a alguien haciendo comentarios racistas o machistas o cuando parece estar borracho. Después se cubre su cara con la del oponente y se lanza a las redes sociales. Incluso cuando se detecta la falsificación ya se ha ganado mucho, porque los rumores se extienden más rápido que la verdad. Las matemáticas y la tecnología que depende de ellas pueden emplearse para el bien tanto como para el mal. Lo que importa es cómo se usen.

8
¡Boing!

> Un muelle es un objeto elástico que recupera su forma original cuando se le libera, tras ser comprimido o estirado. Se emplea para almacenar energía mecánica al ejercer una tensión constante o absorber movimiento. Se utilizan muelles en casi todos los sectores industriales, desde la fabricación de automóviles y la construcción hasta el mobiliario.
>
> Confederation of British Industry,
> *Product factsheet: springs in Europe*

Hace poco, compramos un colchón nuevo. Elegimos uno que tiene 5900 muelles. El diagrama de su sección transversal en la tienda mostraba filas apretadas y muy juntas de resortes poco enroscados, con otra capa encima de otros más pequeños. Los colchones caros de verdad tenían otros 2000 muelles más *dentro* de la capa principal. La tecnología actual ha adelantado mucho desde la época en que los jergones tenían solo 200 resortes, más bien grandes y no muy cómodos.

Un muelle es uno de esos mecanismos que están por todas partes pero a los que casi nunca se presta atención, hasta que fallan. Hay válvulas de resorte en los motores de los coches, muelles largos y delgados en los bolígrafos retráctiles y otros de muchas formas y tamaños diferentes en los teclados de los ordenadores, en tostadoras, en los pomos de las puertas, en relojes, en camas elásticas, en sofás y en reproductores de Blu-ray. Nadie se fija en ellos porque están ocultos dentro de nuestros aparatos y muebles, donde ni se ven ni, por lo tanto, se les presta atención. Son un gran negocio.

¿Alguien sabe cómo se fabrican? Desde luego, yo no me enteré hasta 1992, cuando sonó el teléfono de mi oficina.

«¿Hola? Mi nombre es Len Reynolds. Soy un ingeniero de la Asociación de Fabricantes e Investigadores de Resortes de Sheffield. He leído su libro sobre teoría del caos y en él menciona un método para encontrar la forma de un atractor caótico a partir de observaciones. Creo que podría ayudar a resolver un problema que hemos tenido en la industria de fabricación de

muelles durante los últimos 25 años. Lo he probado en algunos datos de muestra con mi ZX81».

El Sinclair ZX81 fue uno de los primeros ordenadores domésticos en el mercado. Empleaba un televisor como monitor y un casete para almacenar el *software*. Estaba hecho de plástico, tenía un tamaño parecido al de un libro y una magnífica memoria de 1 kB. Era posible conectar otros 16 kB en la parte de atrás, siempre que se tomasen precauciones para que no se cayesen. Yo construí una estructura de madera para sostener la ampliación de RAM en su sitio. Otras personas emplearon Blu-Tack.

No era lo que se dice tecnología informática puntera, pero los resultados preliminares de Len eran lo bastante prometedores como para merecer una beca de 90 000 libras esterlinas (unos 150 000 dólares estadounidenses de la época) del Ministerio de Comercio e Industria, con fondos equivalentes (en especie, no en efectivo) de un consorcio de fabricantes de muelles y alambres. Este dinero sirvió para pagar una investigación de tres años para mejorar las pruebas de control de calidad de la materia prima para resortes, lo que a su vez llevó a otros dos proyectos por un plazo de cinco años. En un momento dado, se estimó que el resultado podía ahorrar a las industrias de muelles y alambres unos 18 millones de libras esterlinas (30 millones de dólares) al año.

No es exagerado decir que hay miles de aplicaciones como esta, de las matemáticas a los problemas industriales, que ocurren todo el tiempo y que en buena medida pasan desapercibidas. Muchas son secretos comerciales, protegidas por acuerdos de confidencialidad. De vez en cuando, organizaciones de Reino Unido, tales como el Consejo para la Investigación de las Ciencias Físicas y de la Ingeniería o el Instituto de Matemáticas y sus Aplicaciones, publican casos prácticos breves de unos pocos de estos proyectos y lo mismo ocurre en Estados Unidos y en todas partes. Sin estas investigaciones, y muchos otros usos dirigidos de las matemáticas por empresas grandes y pequeñas de todo el mundo, no existiría ninguno de los aparatos y dispositivos que se emplean a diario. Sin embargo, es un mundo oculto y pocas personas sospechan siquiera de su existencia.

En este capítulo, sacaré a la luz los tres proyectos en los que he participado. No porque tengan una importancia especial, sino porque sé lo que implicaron. Las ideas fundamentales se publicaron, sobre todo en revistas especializadas del sector, y son de dominio público. Mi objetivo es demostrar que la manera en que las matemáticas se emplean en la industria es a menudo indirecta y sorprendente, con una dosis de casualidad.

Como la llamada telefónica de Len.

*

El problema que traía en vilo a la industria del alambre y de los muelles desde hacía un cuarto de siglo era sencillo y básico. Los resortes (producidos por fabricantes de resortes) consumen alambre (producido por fabricantes de alambre) y lo hacen pasar por máquinas enrolladoras para dar forma al producto. La mayor parte de la materia prima responde a la perfección y da lugar a muelles en el intervalo correcto de tamaño y elasticidad. Pero de vez en cuando un lote se resiste a enrollarse de la manera correcta, incluso en manos de un operario muy hábil. Los métodos de control de calidad habituales de principios de la década de 1990 no podían discernir el alambre bueno del malo. Ambos superaban las mismas pruebas de composición química, resistencia a la tensión, etcétera. En el aspecto visual, parecían idénticos. Pero cuando se alimentaba a una máquina enrolladora con un buen alambre se obtenía el muelle que se deseaba, mientras que si se empleaba uno malo, lo que salía era algo que parecía un resorte pero del tamaño equivocado o, en el peor de los casos, era un enredo sin solución.

Probar a enrollar el alambre no era una prueba eficaz ni determinante. Si era malo, podía tener ocupada una máquina muy cara durante un par de días, hasta que el operario se convenciera de que era imposible hacer muelles con ese lote. Por desgracia, dado que la materia prima había superado los controles habituales, su fabricante podía decir, con razón, que no había ningún problema con ella: algo debía haber ido mal con el ajuste de la máquina enrolladora. Ambos sectores se quejaban del cruce de acusaciones mutuas resultante, ambos querían una manera fiable de decidir quién estaba en lo cierto y ambos estaban resueltos a descubrir que no era culpa suya. Tenían la disposición adecuada, pero necesitaban una prueba objetiva.

Cuando empezamos con el proyecto, uno de los primeros pasos fue llevar a los matemáticos a una empresa fabricante de muelles y mostrarles la manera en que el alambre se transforma en un resorte. Todo es cuestión de geometría.

Los muelles más comunes son los de compresión. Si se empuja para acercar sus extremos, estos oponen una fuerza. El diseño más sencillo es una hélice geométrica, como una escalera de caracol. Imaginemos un punto que gira y gira en torno a una circunferencia a velocidad constante. Ahora desplacémoslo en perpendicular con respecto a ese plano también a velocidad constante. La curva que describe en el espacio es una hélice. Por motivos prácticos, los muelles helicoidales se cierran a menudo en ambos extremos, como si el punto móvil diese vueltas primero en el plano antes de empezar a

moverse en perpendicular y después dejase de desplazarse en esa dirección para la última vuelta. Esto protege al resorte para que sus extremos no se enganchen en cosas y también a las personas para que no sean esas cosas a las que se enganchan los extremos.

Desde el punto de vista matemático, una hélice se caracteriza por dos propiedades: su curvatura y su torsión. La primera mide lo cerradas o abiertas que son sus curvas y la segunda mide cuánto se aleja del plano, determinada en la dirección en la que gira (por supuesto, hay una definición técnica, pero no nos enredemos en la geometría diferencial de curvas espaciales). Para una hélice, ambas cantidades son constantes. De modo que, si se mira desde el lado, las vueltas están separadas de manera uniforme e inclinadas el mismo ángulo (lo que se deriva de la velocidad constante a lo largo de su eje). Cuando se mira desde un extremo, todas las vueltas están alineadas entre sí y forman una circunferencia, por el movimiento uniforme de rotación. Una circunferencia pequeña corresponde a una curvatura elevada, una grande a una curvatura baja; una hélice que sube de manera pronunciada corresponde a una torsión elevada y una que lo hace poco a poco a una torsión baja.

Una máquina enrolladora encarna mecánicamente estas propiedades de forma maravillosa y sencilla. Toma la materia prima de una gran bobina poco apretada, denominada carrete, y la alimenta a través de una pequeña herramienta que no es más que una pieza rígida de metal. Con esto dobla el alambre en una dirección y de manera simultánea lo empuja un poco en perpendicular. Al doblarlo se crea la curvatura y al empujarlo, la torsión. Conforme el alambre pasa a través de la máquina, esta produce una vuelta tras otra de la hélice. Cuando esta es lo bastante larga, otra herramienta la corta y queda lista para dar forma al siguiente muelle. Aparatos adicionales reducen la torsión a cero cerca de cada extremo para que esas vueltas sean planas y cerradas. El proceso es rápido, de varias unidades por segundo. Un fabricante hacía resortes minúsculos a partir de un material especial a un ritmo de 18 por segundo en cada máquina enrolladora.

Las empresas de alambre y de muelles suelen ser bastante pequeñas y desde un punto de vista técnico, son *pymes*. Están atrapadas entre proveedores muy grandes, como British Steel, y clientes enormes, tales como fabricantes de vehículos y de colchones, de modo que sus márgenes de beneficio están limitados por ambos lados. Tienen que ser muy eficientes para poder sobrevivir. Ninguna de ellas por sí sola puede asumir el coste de tener su propio departamento de investigación, así que la Asociación de Fabricantes e Investigadores de Resortes (SRAMA, por sus siglas en inglés, desde entonces

bajo el nuevo nombre de Instituto de Tecnología de Resortes) es una especie de organismo conjunto de investigación y desarrollo, una iniciativa en colaboración financiada por sus integrantes. Len y sus colegas de SRAMA ya habían conseguido algunos avances en el problema del enrollado, basándose en el análisis de lo que salía mal. La curvatura y la torsión al avanzar el proceso dependen de las propiedades materiales del alambre, tales como su plasticidad (lo fácil o difícil que es doblarlo). Cuando se ha formado una hélice correcta y regular, estas propiedades son uniformes a lo largo del carrete. Cuando no se consigue, es que no lo son. De modo que parecía probable que una capacidad de enrollado pobre se debiera a una variación irregular de estas propiedades materiales a lo largo del alambre. La cuestión entonces era saber cómo detectar estas alteraciones.

La respuesta estribaba en forzar el alambre a dar vueltas al enrollarlo en torno a una barra de metal, de un modo parecido a como lo hacen los espaguetis en un tenedor. Entonces es posible medir la separación entre vueltas sucesivas. Si son todas bastante parecidas, el lote es bueno. Si cada una es de su padre y de su madre, es malo. Excepto que a veces pueden variar mucho y el alambre sigue haciendo resortes a pesar de todo. Tal vez no de forma tan precisa como un lote bueno de verdad, pero lo suficiente para algunas aplicaciones. De modo que el meollo del asunto era saber cómo cuantificar (poner un número) al punto hasta el que un alambre «es de su padre y de su madre».

Los ingenieros de SRAMA aplicaron todas las herramientas estadísticas habituales a su listado de mediciones, pero ninguna estaba relacionada de cerca con la capacidad de formar muelles. En ese momento apareció mi libro sobre la teoría del caos.

*

La teoría del caos, que es un nombre inventado por los medios de comunicación, es más conocida entre los matemáticos como una parte de la teoría más amplia de la dinámica no lineal, que trata de la evolución de sistemas cuando su comportamiento a lo largo del tiempo está regido por una regla matemática concreta. Se mide el estado del sistema *en este momento*, se aplica la regla y se obtiene otro correspondiente a un poco después. Y se hace de nuevo. Conforme pasa el tiempo, es posible calcular el estado tan lejos en el futuro como se quiera. Esta técnica es la parte dinámica. *Grosso modo*, «no lineal» quiere decir que la regla no se limita a hacer que el estado posterior

sea proporcional al actual ni a la diferencia entre este y otra configuración de referencia. Para un tiempo que varía de manera continua, la regla viene determinada por una ecuación diferencial que relaciona la tasa de cambio de las variables del sistema con sus valores actuales.

También hay una versión discreta en la que el tiempo transcurre paso a paso, descrita por una relación de recurrencia: el estado después de una unidad de tiempo es lo que le ocurre al actual cuando se aplica la regla. Es la versión discreta la que resuelve el problema del enrollado. Por suerte, es la más fácil de entender. Funciona así:

estado en el momento 0 →
estado en el momento 1 →
estado en el momento 2 → …

donde la flecha quiere decir «se aplica la regla». Por ejemplo, si esta es «multiplicar por dos» y se parte de un estado inicial igual a 1, entonces los pasos sucesivos dan lugar a la secuencia 1, 2, 4, 8… el doble cada vez. Esta regla es lineal porque el resultado es proporcional al estado inicial. Otra del tipo «elevar al cuadrado y restar 3» es no lineal y en este caso da lugar a la secuencia

$$1 \rightarrow -2 \rightarrow 1 \rightarrow -2 \rightarrow \ldots$$

que repite los mismos dos números una vez tras otra. Se trata de una dinámica «periódica», muy parecida al ciclo de las estaciones, por decir algo. El comportamiento futuro es predecible por completo a partir del estado inicial: tan solo alterna entre 1 y −2.

Por otro lado, si la regla es «elevar al cuadrado y restar 4», se obtiene

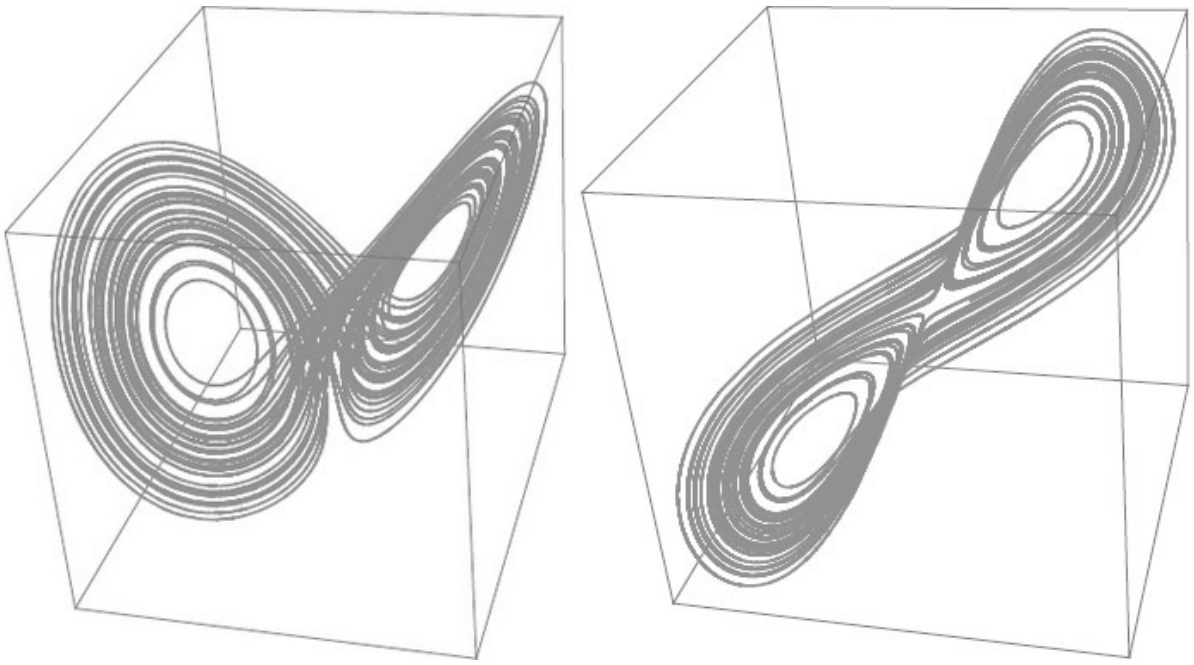
$$1 \rightarrow -3 \rightarrow 5 \rightarrow 21 \rightarrow 437 \rightarrow \ldots$$

y los números no dejan de crecer, excepto por el segundo. La secuencia todavía es predecible: solo hay que aplicar la regla sin cesar. Dado que es determinista (no tiene características aleatorias) cada valor sucesivo está determinado en exclusiva por el anterior, de modo que *todo el futuro* es predecible por completo.

Lo mismo ocurre con versiones de tiempo continuo, aunque la posibilidad de hacer predicciones no es tan evidente en ese caso. Una secuencia de números de este tipo se denomina serie temporal.

Siguiendo los pasos de Galileo Galilei y de Newton, los matemáticos y los científicos descubrieron innumerables reglas de este tipo, tales como la de Galileo para la posición de un cuerpo que cae sometido a la fuerza de la gravedad y la ley de Newton de la gravitación universal. Este proceso llevó a la creencia de que todos los sistemas mecánicos obedecen reglas deterministas y que, por lo tanto, son predecibles. No obstante, el gran matemático francés Henri Poincaré descubrió una inconsistencia en este argumento, que publicó en 1890. La ley de Newton de la gravitación universal determina que dos cuerpos celestes, tales como una estrella y un planeta, se mueven en órbitas elípticas en torno a su centro de masas común, que en este caso se encuentra por lo normal dentro de la estrella. El movimiento es periódico y el periodo es el tiempo que se tarda en recorrer una vez la órbita y volver al punto de partida. Poincaré investigó lo que ocurre cuando hay tres cuerpos (Sol, planeta, satélite) y descubrió que en algunos casos el movimiento es irregular en extremo. Matemáticos posteriores, que retomaron este descubrimiento mucho después, se dieron cuenta de que este tipo de irregularidad hace que no se pueda predecir la evolución de un sistema así. La inconsistencia en la «demostración» de la predictibilidad consiste en que esta solo se conserva si puede medirse el estado inicial y hacer todos los cálculos con una precisión absoluta, correcta hasta una cantidad infinita de decimales. De otro modo, cualquier discrepancia minúscula podría crecer con velocidad exponencial hasta saturar por completo el valor verdadero.

Se trata del caos, o con mayor propiedad, del caos determinista. Incluso cuando se conocen las reglas y estas no tienen características aleatorias, es posible que el futuro no sea predecible en la práctica, aunque sí lo sea en la teoría. De hecho, el comportamiento puede ser tan irregular que parezca aleatorio. En un sistema que lo sea de verdad, el estado actual no proporciona información alguna acerca del siguiente. Pero en uno caótico hay patrones sutiles ocultos tras el caos. Estos son geométricos y es posible hacerlos visibles al trazar las soluciones de las ecuaciones del modelo como curvas en un espacio cuyas coordenadas sean las variables de estado. En algunas ocasiones, si se espera un poco, estas representaciones empiezan a dibujar una forma geométrica compleja. Si las curvas que parten de diferentes puntos iniciales trazan todas el mismo dibujo, este se denomina atractor. El atractor caracteriza los patrones ocultos en el comportamiento caótico.



Izquierda: el atractor de Lorenz. *Derecha*: una reconstrucción de su topología a partir de una única variable.

Un ejemplo habitual son las ecuaciones de Lorenz, un sistema dinámico de tiempo continuo que es un modelo de un gas convectivo, tal como el aire caliente en la atmósfera. Estas ecuaciones tienen tres variables. En un gráfico que represente su evolución y que emplee un sistema de coordenadas tridimensional, las curvas de la solución acaban todas por desplazarse a lo largo de una forma que tiene un cierto parecido con un antifaz: el atractor de Lorenz. El caos resulta porque, aunque las curvas de la solución se desplazan alrededor (bueno, muy cerca) de este atractor, cada una lo hace de manera muy diferente. Una podría, pongamos por caso, dar seis vueltas en torno a la circunferencia de la izquierda y luego otras siete en torno a la de la derecha. Otra cercana podría dar ocho vueltas alrededor de la izquierda y luego tres alrededor de la derecha, etcétera. De modo que los futuros predichos para estas soluciones son muy diferentes, incluso aunque partan de valores muy parecidos de las variables.

No obstante, las predicciones a corto plazo son más fiables. En primer lugar, dos curvas cercanas permanecen próximas entre sí. Solo empiezan a divergir más tarde. De modo que un sistema caótico es predecible a corto plazo, a diferencia de uno aleatorio de verdad, que no es predecible en absoluto. Este es uno de los patrones ocultos que distinguen el caos determinista de la aleatoriedad.

Cuando se trabaja con un modelo matemático concreto, se conocen todas las variables y puede emplearse un ordenador para calcular cómo cambian estas. Es posible hacer visible el atractor al representar estos cambios en las

coordenadas. Cuando se observa un sistema real que podría ser caótico, no siempre se dispone de estos lujos. En el peor de los casos, tal vez solo se sea capaz de medir una de las variables. Dado que no se conocen las demás, no puede dibujarse el atractor.

Y aquí entra en juego la idea de Len. Los matemáticos han concebido métodos astutos para «reconstruir» un atractor a partir de las mediciones de una única variable. El más sencillo es el de Packard-Takens o de ventana deslizante, desarrollado por Norman Packard y Floris Takens. Introduce nuevas variables «falsas» al medir una sola, pero en momentos diferentes. De modo que en lugar de las tres variables originales en un mismo instante, se toma solo una dentro de una ventana de tres unidades de tiempo de duración. Entonces se desliza la ventana un paso y se hace lo mismo y el proceso se repite muchas veces. La imagen de la derecha muestra cómo funciona este proceso para el atractor de Lorenz. No es *idéntica* a la de la izquierda, pero a no ser que se haga una elección muy mala de las unidades de tiempo, las dos representaciones tienen la misma topología: el atractor reconstruido es una versión distorsionada de forma continua del real. Aquí, ambas imágenes tienen la apariencia de un antifaz, con dos agujeros para los ojos, aunque una es una versión retorcida de la otra.

Esta técnica proporciona una representación cualitativa del atractor que dice el *tipo* de caos que se debe esperar. Así que Len, que se preguntaba si el mismo truco podría funcionar con sus datos de los muelles, hizo un gráfico bidimensional en el que trataba las separaciones sucesivas entre vueltas como una serie temporal y les aplicaba una reconstrucción de ventana deslizante. Sin embargo, no obtuvo una forma geométrica nítida como un antifaz, sino una nube difusa de puntos. Esto indicaba que la secuencia de las separaciones podría no ser caótica en el sentido técnico en que emplean el término los matemáticos.

Entonces, ¿el método era inútil?

En absoluto.

Lo que llamó la atención de Len fue la *forma* global de esa nube difusa. Se habían hecho pruebas meticulosas a las muestras de alambre en una máquina enrolladora, así que sabía cuáles eran buenas, malas o indiferentes. ¿Podría la nube de puntos reconstruida indicar cuál era cuál? Aparentemente sí. Cuando el alambre era muy bueno, se enrollaba con facilidad y daba como resultado muelles muy precisos, la nube era pequeña y más o menos circular. Si el material era aceptable, se enrollaba con cierta facilidad pero producía resortes de tamaños más variables, la distribución de puntos era mayor

aunque todavía aproximadamente circular. Por el contrario, cuando era malo e imposible de enrollar para formar los muelles, la nube era larga y delgada, como un puro.

Si se repetía el mismo patrón en otras muestras, se podría prescindir de las pruebas, lentas y caras, en una máquina enrolladora y utilizar la forma y el tamaño de la nube difusa para caracterizar a los alambres buenos, indiferentes y malos. Con esto quedaría resuelto el problema práctico de encontrar una prueba eficaz y barata de la capacidad de enrollarse. En realidad, no *importa* si la separación entre las vueltas es aleatoria, caótica o tiene un poco de ambas. No hace falta saber con exactitud cómo varían las propiedades materiales a lo largo del alambre, ni siquiera cuáles son estas. Desde luego, no hay que hacer cálculos muy complicados en teoría de la elasticidad, verificados con experimentos igual de complejos, para comprender cómo redundan estas variaciones en una capacidad de enrollarse buena o mala. Todo lo que se necesita saber es cómo distingue el gráfico de ventana deslizante el buen alambre del malo y esto puede comprobarse al probarlo con más muestras de material y compararlas con su comportamiento en una máquina enrolladora.

El motivo por el que las herramientas estadísticas habituales de los datos, tales como el valor medio (promedio) y la varianza (dispersión) no resultaban de ayuda es ahora evidente. Esas medidas pasan por alto el orden en el que se han tomado los datos y cómo cada separación de las vueltas se relaciona con la anterior. El valor medio y la varianza no cambian si se mezclan los datos, pero la forma de la nube de puntos puede verse alterada de manera dramática. Y es muy probable que esa sea la clave para hacer buenos muelles.

Para investigar esta idea construimos una máquina de control de calidad, FRACMAT, que enrollaba unas vueltas de prueba en torno a una barra de metal, las escaneaba con un micrómetro láser para medir las separaciones sucesivas, introducía estos datos en un ordenador, les aplicaba una reconstrucción de ventana deslizante para obtener una nube de puntos, estimaba la elipse que mejor se ajustaba a ella, para ver si era circular o tenía forma de puro y estimar su tamaño, y calculaba lo buena o mala que era la muestra de alambre. Era una aplicación práctica de la *teoría* del caos, del método de reconstrucción, a un problema que era probable que ni siquiera fuese caótico en sentido técnico. Resulta adecuado que la financiación del ministerio no fuese para investigación, sino para transferencia tecnológica: transferimos el método de reconstrucción de las matemáticas de la dinámica caótica a las series temporales de observaciones de un sistema del mundo real

que es muy posible que no fuese caótico. Que es ni más ni menos lo que les dijimos que íbamos a hacer.

*

«Caos» es algo más que otra manera de decir «aleatorio». El caos es predecible a corto plazo. Si se lanza un dado, el resultado de la tirada no dice nada sobre lo que ocurrirá a continuación. Salga lo que salga esta vez, todos los números 1, 2, 3, 4, 5, 6 son igual de probables en la próxima ocasión. Siempre que el dado no esté trucado para que la probabilidad de algunos sea mayor, claro. El caos es diferente. Si los dados fuesen caóticos habría patrones. Tal vez un 1 en una tirada solo podría ir seguido por un 2 o un 5, mientras que después de un 2 solo podría salir un 4 o un 6, etcétera. Podría predecirse hasta cierto punto el próximo resultado, aunque el de la quinta o sexta tirada tras la actual pudiese ser cualquier cosa. La incertidumbre en la predicción es mayor cuanto más adelante en el tiempo se quiera hacer esta.

El segundo proyecto, DYNACON, surgió a partir del primero, cuando nos dimos cuenta de que se podía aprovechar esta predictibilidad a corto plazo del caos para controlar una máquina enrolladora. Si de algún modo fuésemos capaces de medir las longitudes de los resortes conforme se fabricaban y de detectar tendencias en las gráficas que sugiriesen que la máquina se comportaba de manera caótica de verdad, tal vez fuese posible ver venir los muelles malos y ajustar el proceso para compensarlo. Los fabricantes ya habían encontrado maneras de medir la longitud de los resortes conforme se producían, para separar los imprecisos en un contenedor aparte. Pero queríamos más. No solo retirar los malos conforme se fabricaban, sino evitar por completo que salieran. No a la perfección, pero sí lo suficiente como para evitar desperdiciar montones de alambre.

La mayoría de los matemáticos buscan la precisión. Un número es (o no) igual a 2. Pertenece (o no) al conjunto de los primos. Pero el mundo real es a menudo mucho más difuso. Una medición puede ser cercana a 2 pero no igual del todo; es más, si se mide la misma cantidad de nuevo, tal vez el resultado sea un poco diferente. Aunque un número no puede ser «casi primo», desde luego que puede ser «casi entero». Una descripción así es razonable para cantidades como 1,99 o 2,01, por ejemplo. En 1965, Lotfi Zadeh y Dieter Klaua formularon, de manera independiente, una descripción matemática precisa de este tipo de indefinición, conocida como teoría de conjuntos difusos, junto con un concepto relacionado de lógica difusa.

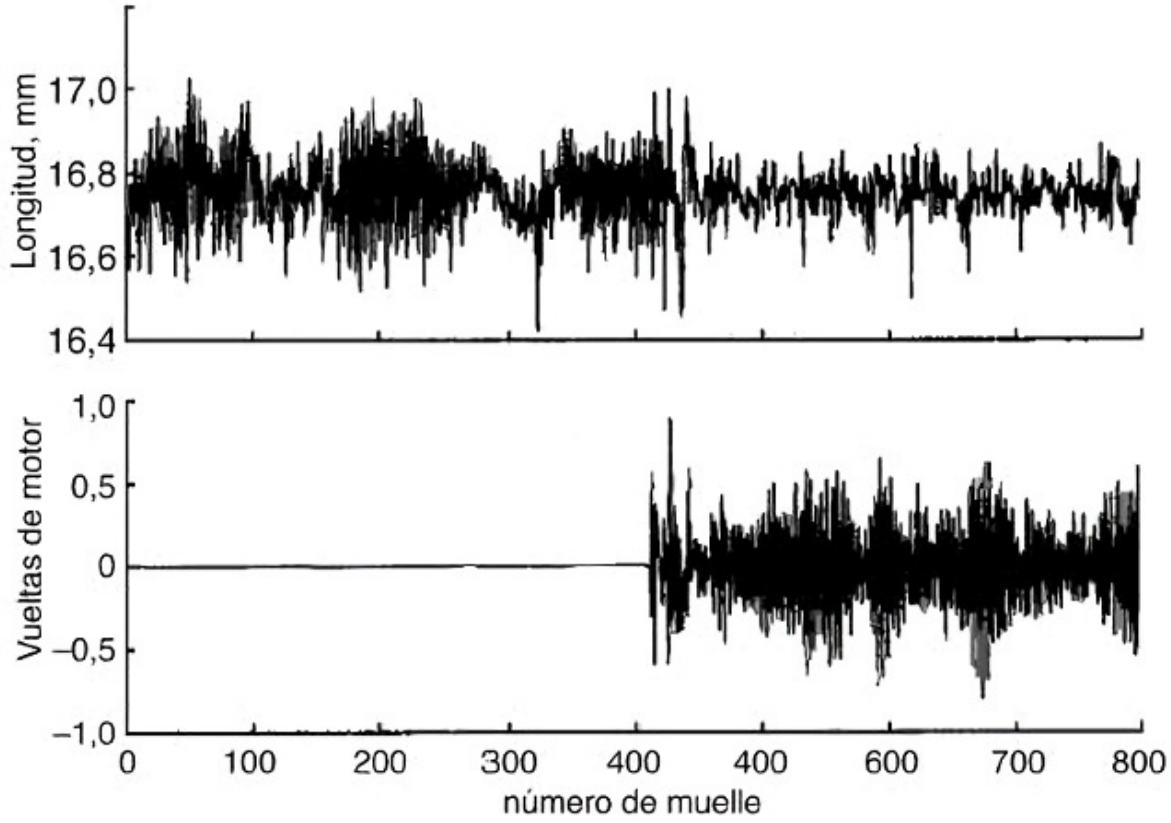
En teoría de conjuntos convencional, un objeto (tal como un número) pertenece a un conjunto concreto o no lo hace. En cambio, en su contrapartida difusa hay una medida numérica precisa del *grado* en que pertenece. De modo que el 2 podría estar a medias en el conjunto, o solo en un tercio. Si la medida es 1, el número es un elemento con certeza y si es 0, es seguro que no lo es. Si solo hay 0 y 1, se obtiene la teoría de conjuntos convencional. Si se permite cualquier valor entre 0 y 1, el grado de pertenencia difusa representa la zona gris que hay entre esos dos extremos.

Algunos matemáticos destacados se apresuraron a desdeñar la idea y, o bien decían que la teoría de conjuntos difusos no es más que teoría de probabilidades disfrazada, o bien que la lógica de la mayor parte de las personas ya es bastante difusa sin pretender que los matemáticos sigan por la misma senda. Me resulta desconcertante la motivación que tienen algunos académicos para desdeñar de manera tan inmediata las ideas nuevas, sobre todo cuando sus motivos para hacerlo no tienen razón de ser. Nadie proponía *sustituir* la lógica convencional por una difusa. Solo se ofrecía como un arma nueva en el arsenal. Aunque los conjuntos difusos se parecen de manera superficial a la teoría de probabilidades, las reglas son diferentes y también lo es su interpretación. Si un número pertenece a un conjunto con una probabilidad de 1/2 y se adopta un enfoque frecuentista, se afirma que si se repite el experimento muchas veces, este número aparecerá en los resultados cerca de la mitad de las veces. Si el enfoque es bayesiano, la confianza de que pertenezca al conjunto es del 50 %. Pero en la teoría de conjuntos difusos no existe el elemento suerte. El número está con certeza *en* el conjunto, pero el grado en el que pertenece a él no es 1. Es *exactamente* 1/2. Respecto a la broma sobre la lógica deficiente, cabe decir que la lógica difusa tiene reglas precisas y que cualquier argumento que la emplea es correcto o no, dependiendo de si ha obedecido esas reglas. Supongo que la palabra «difusa» llevó a algunas personas a asumir, sin molestarse en comprobarlo, que las reglas en sí mismas eran maleables y estaban mal definidas. En absoluto.

Un tema diferente, que sin duda hizo mucho por enturbiar el debate, es hasta qué punto los conjuntos difusos y la lógica difusa contribuyen de un modo significativo a las matemáticas. Es fácil diseñar sistemas formales extensos que son poco menos que cajones de sastre pretenciosos de fórmulas carentes de contenido («sinsentidos abstractos»). Sospecho que era muy tentador ver la criatura de Zadeh bajo esta luz, sobre todo desde el momento en que los aspectos básicos no eran profundos ni difíciles. Bien, puede ser que el movimiento solo se demuestre andando, pero hay varias maneras de

evaluar si una contribución matemática es valiosa y su profundidad intelectual es solo una de ellas. Otra, bastante relevante para este libro, es su *utilidad*. Y muchas ideas matemáticas casi triviales han resultado ser útiles en extremo. Por ejemplo, la notación decimal. Es brillante, novedosa, inteligente y rompedora, pero no profunda. Los niños son capaces de entenderla.

Es probable que la lógica difusa y la teoría de conjuntos difusos no superen el criterio de la profundidad, al menos si se las compara con la hipótesis de Riemann o con el último teorema de Fermat. Pero lo cierto es que han demostrado ser muy útiles. Cobran relevancia cada vez que no se está seguro del todo de la precisión de la información que se observa. Hoy en día, las matemáticas difusas se emplean con profusión en campos tan diversos como la lingüística, la toma de decisiones, el análisis de datos y la bioinformática. Se usan cuando resuelven la papeleta mejor que cualquiera alternativa y pueden dejarse de lado sin problemas cuando no lo hacen.



Efecto de poner en marcha el controlador autoajustable difuso. El número de muelles va de izquierda a derecha. *Arriba*: longitudes medidas de los muelles. *Abajo*: actividad del controlador, medida por el número de veces que gira el motor del controlador. Los muelles 1 a 400 no tienen control y la variabilidad en altura es grande. Los resortes 401 a 800 se enrollaron con el controlador encendido. La variabilidad es menor a ojos vistas.

No quiero entrar en los pormenores de la teoría de conjuntos difusos, que en realidad no son necesarios para apreciar nuestro segundo proyecto. Ensayamos varios métodos para predecir cuándo iba la máquina enrolladora a fabricar muelles defectuosos y ajustarla de manera acorde. Uno de ellos se conoce en la industria como un modelo de controlador difuso de tipo Takagi-Sugeno, por los ingenieros Tomahiro Takagi y Michio Sugeno[61].

Este implementa, en el formalismo preciso de las matemáticas difusas, unos sistemas de reglas que son difusas en sí mismas. En este caso, las reglas toman la forma «si la medición (necesariamente imprecisa) de la longitud del muelle actual es *X*, entonces haz *Y* para ajustar la máquina enrolladora». También se tomaban en cuenta los ajustes previos, junto con una estimación de las perturbaciones causadas por las propiedades materiales variables del alambre, el desgaste de la máquina herramienta, etcétera. Todos los datos son difusos y por lo tanto también lo son las acciones que se realizan. El formalismo matemático gestiona esto de manera automática para ajustar la máquina enrolladora sobre la marcha.

Para nuestro proyecto de tiras de metal ensayamos tres métodos diferentes. En primer lugar, hicimos funcionar la máquina con el sistema de control apagado para establecer una línea de referencia a partir de la cual evaluar la eficacia de los demás métodos. Los datos obtenidos también ayudaron a estimar varios parámetros en los modelos matemáticos. A continuación, pusimos la máquina a trabajar con un controlador integral, que emplea una fórmula matemática fija para predecir el cambio necesario en los ajustes entre una vuelta y la siguiente. Por último, empleamos un control autoajustable difuso, que adapta sus propias reglas sobre la marcha según las longitudes observadas de los muelles. Cuando lo hicimos así con alambre de acero inoxidable, la desviación estándar de las longitudes de los resortes (una medida de lo variables que resultan) fue de 0,077 sin control, de 0,065 con control integral y de 0,039 con autoajuste difuso. De modo que el método de lógica difusa era el que funcionaba mejor y reducía la variabilidad a la mitad.

*

Otro principio básico en matemáticas es que una vez que has encontrado algo que funciona, lo exprimes al máximo. Una idea que ha demostrado su valor puede a menudo aprovecharse en circunstancias parecidas, pero diferentes. Para nuestro tercer proyecto, también parte de DYNACON, volvimos a FRACMAT, modificando el dispositivo de prueba para adaptarlo a un sector

industrial que es parecido a la fabricación de resortes pero que emplea tiras metálicas en vez de alambre.

Es casi seguro que todo el mundo tiene en su casa algún producto hecho con tiras metálicas. En Reino Unido, todos los enchufes contienen un fusible sujeto mediante pinzas de cobre. Estas se fabrican a partir de un gran carrete de tira de cobre, delgada y estrecha. Una máquina alimenta la materia prima a través de una serie de herramientas dispuestas en círculo, orientadas todas hacia el centro, que es por donde pasa la tira. Cada una de ellas hace una doblez, con un ángulo concreto y en una posición determinada, perfora un orificio o realiza cualquier otra operación necesaria. Por último, una herramienta de corte separa la pinza terminada que cae a un contenedor. Una máquina normal puede hacer diez o más pinzas por segundo.

El mismo proceso se emplea para fabricar una enorme cantidad de pequeños objetos metálicos. Una empresa de Reino Unido se ha especializado en hacer las pinzas que sujetan los soportes de los techos técnicos y produce cientos de miles al día. Del mismo modo que los fabricantes de resorts tienen problemas para saber si el alambre se va a enrollar bien, los fabricantes de pinzas tienen problemas para saber si una muestra dada de tira se va a doblar de la forma que se espera. El origen del problema es parecido: propiedades materiales variables, tales como la plasticidad, a lo largo de la tira. De modo que parecía razonable probar con el mismo método de reconstrucción de ventana deslizante en este caso.

No obstante, no tiene sentido forzar una tira de metal para que se enrolle. Tiene la forma equivocada para hacerlo con facilidad, y dar vueltas tiene poco que ver con el modo de fabricar las pinzas. La cantidad crucial aquí es cuánto se dobla el metal cuando se aplica una fuerza dada. De modo que tras pensarlo mucho, volvimos a diseñar la máquina de prueba y se nos ocurrió algo mucho más sencillo: hacer pasar la tira entre tres rodillos y que el de la mitad la obligase a doblarse. Dejábamos que este rodillo se moviese un poco, sobre un resorte fuerte, y medíamos cuánto se desplazaba cuando la tira pasaba por debajo de él. El metal se doblaba y luego se aplanaba de nuevo y se podía medir la fuerza necesaria para hacerlo así. Si la plasticidad variaba con la longitud, también lo hacía esta fuerza.

En lugar de mediciones discretas de separaciones de vueltas en el alambre, tomadas mediante un micrómetro láser, ahora teníamos mediciones continuas de fuerzas. La máquina también media la fricción superficial, que resultó tener un efecto importante sobre la calidad. Sin embargo, el análisis de los datos es muy parecido. Esta máquina de prueba es más pequeña que la

FRACMAT, más sencilla de fabricar y tiene la ventaja de que no es destructiva: la tira vuelve a su estado inicial y podría emplearse en la fabricación si se quisiera.

*

¿Qué aprendimos?

Es probable que le hayamos ahorrado a las empresas de alambre y muelles bastante dinero, así que aprendimos que este tipo de análisis de datos matemáticos tiene un valor contante y sonante. Hasta cierto punto, la mera existencia de la FRACMAT convenció a los fabricantes de alambre de mejorar sus procesos de producción, lo que a su vez fue una ayuda para los de resortes. Las máquinas siguen en uso y el Instituto funciona todavía como un recurso colectivo para muchas empresas pequeñas y hace las pruebas por ellas.

Aprendimos que las reconstrucciones de ventana deslizante pueden ser útiles incluso cuando no se sabe si los datos han sido generados por alguna dinámica caótica, precisa desde el punto de vista matemático, nítida y correcta. ¿Varían de manera caótica las propiedades materiales del alambre, en el sentido técnico? No lo sabemos. No nos *hizo falta* saberlo para crear el nuevo procedimiento de prueba y la máquina. Los métodos matemáticos no se limitan al contexto concreto para el que fueron desarrollados en primer lugar. Son portátiles.

Aprendimos que, en ocasiones, cuando se intenta transferir un truco que funciona a un contexto nuevo (control), no siempre sirve. En ese caso, hay que buscar otros métodos que sí lo hagan (lógica difusa).

Aprendimos que a veces este tipo de transferencia funciona bien de verdad. Incluso mejor, en algunos aspectos, que el primer intento. Nuestra máquina para tiras de metal sirve también para el alambre y no es destructiva.

Sobre todo, aprendimos que cuando un equipo de personas con conocimientos y experiencia muy diferentes unen sus fuerzas frente a un problema común, pueden resolverlo de maneras que ningún integrante del equipo habría sido capaz de idear por sí solo. Conforme la humanidad sigue su andadura en el siglo XXI y se enfrenta a problemas nuevos e interrelacionados a todos los niveles, desde el social al tecnológico, esa es una lección muy importante.

9
Confíe en mí, soy una transformada

Un paciente visitó a un doctor por primera vez.
—¿A quién vio antes de venir a la consulta? —le preguntó el médico.
—Al farmacéutico del pueblo.
—¿Y qué tontería le aconsejó ese cretino?
—Que viniera a verle.

Autor desconocido

La manera en la que el autor llega a sus ecuaciones no está exenta de dificultades y su análisis para integrarlas deja aún algo que desear, sea respecto a la generalidad, sea incluso del lado del rigor.

Informe sobre la presentación de Joseph Fourier al Premio de Matemáticas de 1811 del Instituto de París

En la actualidad, una visita al hospital a menudo implica un escáner; los hay de muchos tipos: resonancias magnéticas, tomografías por emisión de positrones, ultrasonidos… Algunos muestran imágenes que se mueven en tiempo real, otros emplean trucos informáticos (es decir, matemáticos) para proporcionar instantáneas tridimensionales. La característica más importante de estas maravillas de la tecnología es que muestran lo que ocurre *dentro* del cuerpo. Hasta hace no mucho, algo así se habría considerado magia. Todavía lo parece.

En épocas antiguas, lo que en este caso quiere decir todas antes de 1895, los médicos tenían que recurrir a sus propios sentidos para investigar las dolencias de sus pacientes. Podían palpar el cuerpo para hacerse una idea del tamaño, la forma y la posición de algunos órganos internos. Podían escuchar el latido del corazón y tomar el pulso. Podían tomar la temperatura y oler, tocar y probar los fluidos corporales. Pero la única manera en la que podían descubrir qué aspecto tenía en realidad el interior del cuerpo humano era cortar y abrirlo. En ocasiones ni siquiera eso, porque a menudo las autoridades religiosas prohibían la disección, aun cuando esta era muy habitual en el campo de batalla, si bien no para fines médicos. Esas mismas

autoridades solían aprobar este tipo de disección, siempre que se hiciese sobre personas de otras creencias.

El 22 de diciembre de 1895 se inició una nueva era, cuando el profesor de física alemán Wilhelm Röntgen fotografió la mano de su esposa y obtuvo una imagen que mostraba los huesos de sus dedos. Como casi todas las instantáneas de la época, era en blanco y negro y más bien borrosa, pero la capacidad de ver dentro de un cuerpo vivo no tenía precedentes. Sin embargo, a ella no le hizo mucha gracia. Al ver la radiografía de una parte de su esqueleto, dijo: «He visto mi muerte».

El descubrimiento de Röntgen fue pura casualidad. En 1785, un actuario llamado William Morgan realizó algunos experimentos en los que hacía pasar una corriente eléctrica a través de un vacío parcial en un tubo de vidrio. De este modo se producía un brillo débil, que se veía mejor en la oscuridad y presentó sus resultados a la Real Sociedad de Londres. Para 1869, los físicos que hacían experimentos en el campo de los tubos de descarga, por entonces de moda, se dieron cuenta de un tipo nuevo y extraño de radiación al que llamaron rayos catódicos porque los emitía el cátodo (electrodo negativo) del tubo. En 1893, Fernando Sanford, un profesor de física, publicó un artículo sobre «fotografía eléctrica». Construyó un tubo con una lámina delgada de aluminio en un extremo y recortó un agujero en ella. Cuando conectaba la corriente, fuese lo que fuese que causaba el débil brillo atravesaba el orificio, alcanzaba una placa fotográfica y reproducía la forma del agujero. Se informó de su descubrimiento en la prensa. El titular del *San Francisco Examiner* decía: «Sin lente ni luz. Fotografías tomadas con una placa y un objeto en la oscuridad». Era fascinante, desconcertante y en apariencia no tenía utilidad, pero los físicos estaban intrigados y no dejaron de intentar entender lo que ocurría.

Radiografía de Röntgen de la mano de su esposa.

Röntgen se dio cuenta de que el extraño brillo era alguna forma de radiación, parecida a la luz, pero invisible. La denominó rayos X, donde, como era costumbre, la «X» indicaba que su naturaleza era desconocida. Parece ser (no se puede tener certeza porque sus diarios no se han conservado) que descubrió de manera accidental que estos rayos podían pasar a través del cartón. Esto le llevó a preguntarse enseguida qué otras cosas podrían atravesar. No una lámina delgada de aluminio, parecía ser, ya que en las fotografías solo se mostraba el orificio. Libros sí; artículos científicos, sí; la mano de su esposa, sí. Los rayos X suponían una ventana sin precedentes al interior de un cuerpo humano vivo. Röntgen comprendió de inmediato su potencial para la medicina y los medios de comunicación no tardaron en darle publicidad. En 1896, la revista *Science* incluía 23 artículos sobre rayos X, el tema de más de 1000 escritos científicos ese mismo año.

Pronto se descubrió que, aunque esta radiación no producía daños aparentes, la exposición repetida o de larga duración podía causar quemaduras

en la piel y caída del cabello. En uno de estos casos, se llevó a un niño que había recibido un disparo en la cabeza a un laboratorio en la Universidad Vanderbilt y John Daniel tomó una radiografía con un tiempo de exposición de una hora. Tres semanas más tarde, observó una zona sin pelo en la cabeza del niño, en el lugar en el que había colocado el tubo de rayos X. A pesar de esta evidencia, muchos médicos siguieron convencidos de que el procedimiento era seguro y culpaban de estos daños a una exposición a radiación ultravioleta o al ozono, hasta que la radióloga estadunidense Elizabeth Fleischman falleció por complicaciones debidas a los rayos X en 1905. Los usos en medicina no cesaron, pero con mayor precaución, y mejores placas fotográficas redujeron el tiempo de exposición. Hoy en día, se reconoce que, pese a la utilidad de las radiografías, la dosis total de radiación debe reducirse al mínimo absoluto. Se tardó un poco. En la década de 1950, cuando yo tenía unos diez años, recuerdo que había zapaterías con máquinas de rayos X para probarse los zapatos y ver cómo se ajustaban a la forma del pie.

Las radiografías adolecían de una serie de defectos. Eran en blanco y negro: zonas negras donde no penetraban los rayos, blancas donde sí lo hacían y tonos de gris para las intermedias. O, lo que era más habitual, al revés, ya que es más fácil emplear un negativo fotográfico. Los huesos se mostraban con nitidez, pero los tejidos blandos eran invisibles en su mayor parte. Sin embargo, la dificultad más importante era que estaban en dos dimensiones. En efecto, reducían a un plano la disposición interna y superponían las imágenes de todos los órganos situados entre la fuente de rayos X y la placa fotográfica. Por supuesto, era posible tomar más radiografías desde otras direcciones, pero interpretar los resultados requería destreza y experiencia y las exposiciones adicionales aumentaban la dosis de radiación.

¿No sería estupendo que hubiese alguna manera de tomar imágenes del interior del cuerpo en tres dimensiones?

*

Da la casualidad de que los matemáticos ya habían hecho algunos descubrimientos fundamentales justo sobre ese tema y demostrado que si se toman montones de fotografías «aplanadas» bidimensionales desde muchas direcciones diferentes es posible deducir la estructura tridimensional de la fuente de las imágenes. No obstante, su motivación no eran los rayos X ni la medicina. Tan solo proseguían la investigación de un método que se había

inventado en primer lugar para resolver problemas sobre olas y flujos de calor.

La historia completa tiene un elenco plagado de estrellas, entre ellas Galileo, que dejaba caer bolas rodando por un plano inclinado y observaba patrones matemáticos de una sencillez deliciosa en la distancia que recorrían en un tiempo dado, y Newton, que descubrió patrones profundos en los movimientos de los planetas. Fue este último quien derivó ambos comportamientos a partir de las ecuaciones del movimiento de un sistema de cuerpos sometidos a fuerzas. En su monumental obra *Philosophiae naturalis principia mathematica* (*Principios Matemáticos de la Filosofía Natural*, habitualmente conocidos como *Principia*), Newton eligió explicar sus ideas mediante el empleo de la geometría clásica, pero su formulación matemática «más nítida» se produjo a raíz de otro de sus descubrimientos, el cálculo, que también fue descubierto de manera independiente por Gottfried Wilhelm Leibniz. Newton se dio cuenta de que las leyes fundamentales de la naturaleza, reinterpretadas de este modo, pueden expresarse mediante ecuaciones diferenciales acerca del *ritmo* al que cambian cantidades importantes a lo largo del tiempo. Así, la velocidad es el ritmo del cambio de posición y la aceleración el del cambio de velocidad.

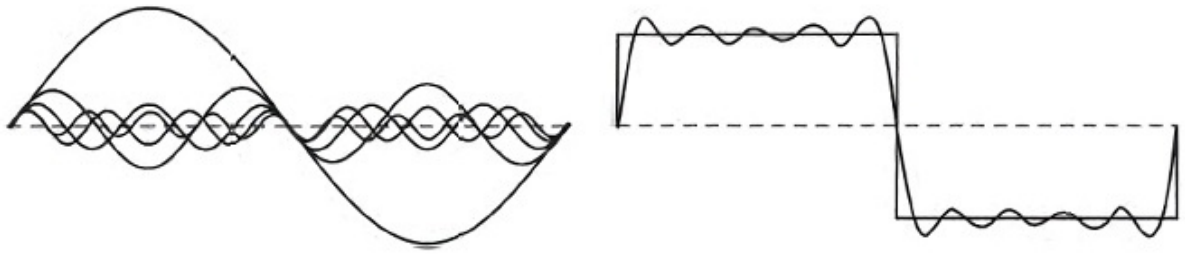
Los patrones de Galileo presentan su forma más sencilla cuando se expresan en términos de la aceleración: una bola que rueda en un plano inclinado se mueve con una aceleración constante. Por lo tanto, su velocidad aumenta a ritmo constante y crece de manera lineal. Su posición se debe al incremento continuo de su velocidad, lo que implica que si parte del reposo en un instante 0, su posición es proporcional al *cuadrado* del tiempo transcurrido. Newton combinó esta idea con otra ley sencilla, según la cual la fuerza de la gravedad actúa de manera inversa al cuadrado de la distancia, y dedujo que los planetas se desplazan en órbitas elípticas, con lo que explicaba las deducciones empíricas anteriores realizadas por Johannes Kepler.

Los matemáticos de la Europa continental aprovecharon estos descubrimientos y aplicaron ecuaciones diferenciales a un conjunto enorme de fenómenos físicos. Las ondas en el agua y las sonoras se rigen por la ecuación de ondas, mientras que la electricidad y el magnetismo tienen también las suyas propias, del mismo modo que la atracción gravitatoria. Muchas de estas expresiones son diferenciales «parciales» y relacionan ritmos de cambio en el espacio con ritmos de cambio a lo largo del tiempo. En 1812, la Academia de Ciencias de Francia anunció que el problema para su premio anual iba a ser el flujo de calor. Los cuerpos calientes se enfrían y la energía

térmica se desplaza a lo largo de los materiales conductores, por eso el mango metálico de una sartén se calienta mucho conforme se cocina el contenido. La Academia quería una descripción matemática del modo en que esto sucede y parecía plausible obtener una ecuación diferencial parcial, porque la distribución de calor cambia tanto en el tiempo como en el espacio.

Joseph Fourier ya había enviado a la Academia un artículo sobre este tema en 1807, pero se negaron a publicarlo. El reto planteado de nuevo lo animó a desarrollar su ecuación diferencial parcial para el flujo de calor y ganó el premio. Su «ecuación del calor» afirma, en términos matemáticos, que la energía térmica en una ubicación dada cambia a lo largo del tiempo al difundirse a regiones vecinas del espacio, como una gota de tinta en el papel secante.

El problema se presentó cuando Fourier intentó resolver su ecuación, partiendo de un caso sencillo: el calor en una barra metálica. Se dio cuenta de que había una solución fácil si la distribución inicial se parece a una curva trigonométrica sinusoide o cosinusoide. Entonces reparó en que podía resolver casos más complicados si combinaba muchas curvas sinusoides y cosinusoides entre sí. Incluso descubrió una fórmula del cálculo para describir de manera exacta cuánto contribuye cada término: multiplicar la expresión de la distribución inicial de calor por el seno o el coseno relevante e integrar. Esto lo llevó a realizar una afirmación audaz: su fórmula, conocida en la actualidad como serie de Fourier, resolvía el problema para *cualquier* distribución inicial de calor. En particular, aseguraba que era válida para casos en que esta última era discontinua, tal como una onda cuadrada: una temperatura constante a lo largo de la mitad de la barra y una diferente a lo largo de la otra mitad.



Cómo obtener una onda cuadrada a partir de senos y cosenos. *Izquierda*: las ondas sinusoides componentes. *Derecha*: la suma de los cinco primeros términos de la serie de Fourier que aproxima una onda cuadrada. Otros adicionales (no mostrados) mejoran la aproximación.

Esta afirmación lo situó en el ojo del huracán de una disputa que llevaba varias décadas encendida. El mismo problema, de hecho, la misma fórmula

integral, había aparecido ya en las investigaciones de Euler y Bernoulli sobre la ecuación de ondas. En ese caso, el ejemplo habitual era una cuerda de violín idealizada. Es imposible ponerla a vibrar si se hace discontinua: se rompe sin más. De modo que la intuición física sugiere que puede haber problemas a la hora de representar funciones discontinuas y la intuición matemática lo secunda al plantear la duda de si las series trigonométricas *convergen*. Es decir, si la suma de un número infinito de curvas sinusoides tiene sentido y, si lo tiene, si el resultado es la onda cuadrada discontinua u otra cosa.

Sin ánimo de ser despectivo, parte del problema radicaba en que Fourier pensaba como un físico, mientras que sus críticos lo hacían como matemáticos. Desde el punto de vista de la física, una onda cuadrada tiene sentido como modelo de calor. La barra metálica se representa de manera idealizada como una recta, que es exactamente lo mismo que hacían Euler y Bernoulli con una cuerda de violín. Si el calor se distribuye de manera uniforme a lo largo de la mitad de esta recta y la otra mitad parte de una temperatura mucho menor, el modelo natural es una onda cuadrada.

Ninguno de estos modelos es preciso del todo como una representación de la realidad, pero la mecánica de la época se refería en exclusiva a objetos idealizados, tales como masas puntuales, colisiones perfectamente elásticas, barras totalmente rígidas de grosor despreciable, etcétera. Una onda cuadrada estaba en buena compañía. Es más, desde un punto de vista matemático, la solución de Fourier predice que la discontinuidad se suaviza de manera inmediata por difusión y pasa a ser una curva pronunciada pero continua que se aplana lentamente, lo que tiene sentido físico y elimina la discontinuidad matemática. Por desgracia, estos argumentos eran demasiado vagos para convencer a los matemáticos, que sabían que las series infinitas pueden ser taimadas y problemáticas. Los directivos de la Academia llegaron a un compromiso: Fourier ganó el premio, pero no publicaron su memoria.

Sin dejarse desanimar, Fourier publicó su obra en 1822 como *Théorie analytique de la chaleur* (*Teoría analítica del calor*). Luego, para acabar de enfadar a todo el mundo, consiguió hacerse nombrar secretario de la Academia y publicar enseguida su memoria original galardonada, sin cambios, en la revista de la Academia. *Touché*.

Resolver los problemas matemáticos que había planteado la afirmación de Fourier costó cerca de un siglo. En general, estaba en lo cierto acerca de muchas cosas, pero equivocado respecto a otros tantos aspectos cruciales. Es verdad que su método funcionaba para la onda cuadrada, con alguna que otra

modificación cuidadosa acerca de lo que sucede exactamente *en* la discontinuidad. Pero desde luego que no lo hacía para distribuciones iniciales más complicadas. Una comprensión completa no se logró hasta que los matemáticos desarrollaron un concepto más general de integral, junto con nociones topológicas que se formulan mejor en términos de teoría de conjuntos.

Mucho antes de que la comunidad matemática resolviera por fin lo que pretendía Fourier, los ingenieros aprovecharon la idea básica y se la apropiaron. Se dieron cuenta de que la esencia de su aportación era una transformada matemática, en la actualidad denominada transformada de Fourier, en la que una señal compleja que varía en el tiempo puede reformularse como una combinación de señales simples con diferentes frecuencias.

La fórmula integral de Fourier dice cómo cambiar el punto de vista desde el dominio temporal al de las frecuencias y viceversa. Cabe destacar que en ambos emplea casi la misma fórmula, lo que establece una «dualidad» entre las dos representaciones.

Esta dualidad significa que al invertirse la transformación se recupera la señal original a partir de las frecuencias que crea, algo así como girar una moneda de la cara a la cruz y luego darle la vuelta una vez más. La ventaja de este procedimiento para la ingeniería es que algunas características que son difíciles de detectar en el dominio temporal son evidentes en el de las frecuencias. También puede suceder al revés, de modo que hay dos métodos muy diferentes para analizar los mismos datos y cada uno de ellos pone el énfasis de manera natural en características particulares que el otro pasa por alto.

Por ejemplo, la respuesta de un edificio alto ante un terremoto parece aleatoria y caótica en el dominio temporal. Pero en el de las frecuencias es posible observar varios picos pronunciados en valores concretos. Estos indican las frecuencias de resonancia en las que el edificio responde de manera violenta al terremoto. Para diseñar la construcción de manera que no se venga abajo si ocurre un seísmo, hace falta suprimir esas frecuencias en particular. Una solución práctica empleada en algunos edificios es apoyarlos en una base de hormigón, muy por debajo de los cimientos, que puede desplazarse hacia los lados. Entonces se «amortiguan» estos movimientos laterales al incorporar pesos enormes o resortes.

Otra aplicación se remonta al descubrimiento de Francis Crick y James Watson de la estructura del ADN. Una evidencia fundamental que confirmó

que estaban en lo cierto fue una imagen de difracción de rayos X de un cristal de ADN. La técnica consiste en hacer pasar un haz de rayos X a través del cristal, lo que provoca que la radiación se desvíe y rebote, un comportamiento que se denomina difracción. Las ondas tienden a acumularse en ciertos ángulos, regidos por la ley de la difracción de Lawrence y de William Bragg, y lo que se muestra en una radiografía es una disposición geométrica compleja de puntos. Este patrón de difracción es en esencia un tipo de transformada de Fourier de las posiciones de los átomos en la molécula de ADN. Al aplicar la transformada inversa (un cálculo informático complicado que es mucho más fácil en la actualidad de lo que era entonces) se deduce la forma de la molécula. Ahora bien, como he dicho, en ocasiones la transformada hace que características estructurales difíciles de detectar en el original se vuelvan evidentes. En este caso, la experiencia de Crick y de Watson con otras imágenes de difracción de rayos X les permitió saber de manera inmediata, sin calcular la transformada inversa, que la molécula era algún tipo de hélice geométrica. Otras ideas añadieron más detalles a esta intuición y llevaron a la famosa doble hélice, que más tarde se confirmó mediante el empleo de la transformada de Fourier.

Estas son solo dos aplicaciones prácticas de la transformada de Fourier y de sus muchos primos. Otras incluyen aumentar la recepción de radio, eliminar el ruido en discos de vinilo antiguos rayados, mejorar el rendimiento y la sensibilidad de los sistemas de sonar empleados por los submarinos y evitar vibraciones indeseadas en vehículos en la etapa de diseño.

Ninguna de las cuales, como es fácil darse cuenta, tiene nada que ver con el flujo de calor. Irrazonable eficacia. Lo importante no es la interpretación física del problema (aunque esta puede muy bien haber tenido influencia sobre la obra original), sino su estructura matemática. Los mismos métodos son válidos para cualquier problema con la misma estructura o con una similar, y es aquí cuando salen a escena los escáneres.

También los matemáticos llegaron a estar intrigados por la transformada de Fourier y la reformularon en el lenguaje de las funciones. Una función es una regla matemática para convertir un número en otro, tal como «elevar al cuadrado» o «sacar la raíz cúbica». Esto incluye todas las tradicionales, tales como polinomios, raíces, la exponencial, el logaritmo y las razones trigonométricas, seno, coseno, tangente, etcétera, pero también es posible dar «reglas» más complicadas que no se expresan como fórmulas, tales como la onda cuadrada que trajo de cabeza a Fourier.

Desde este punto de vista, la transformada de Fourier toma una función de un tipo (la señal original) y la convierte en otra de un tipo diferente (el listado de frecuencias). Hay también una transformada inversa que deshace el efecto de la primera. El aspecto de la dualidad, que la inversa sea casi la misma que la directa, es un añadido elegante. Los contextos correctos son los espacios de funciones con propiedades específicas. Los espacios de Hilbert empleados en teoría cuántica (capítulo 6) son ejemplos de estos, en los que los valores de la función son números complejos, y sus matemáticas están relacionadas de cerca con las de la transformada de Fourier.

Todos los investigadores en matemáticas han adquirido un reflejo muy marcado. Si a alguien se le ocurre algo nuevo que tiene características notables y útiles, enseguida empiezan a preguntarse si hay otras ideas similares que pueden aprovechar el mismo truco en diferentes circunstancias. ¿Hay otras transformadas como la de Fourier? ¿Hay otras dualidades? Los matemáticos teóricos investigan estas cuestiones a su modo, abstracto y general, mientras que los que buscan aplicaciones prácticas (y los ingenieros, los físicos y vaya usted a saber quién más) se preguntan cómo puede utilizarse todo esto. En este caso, el inteligente truco de Fourier inauguró todo un sector de transformadas y dualidades que no se ha acabado de explotar ni siquiera en la actualidad.

*

Entre estas variaciones sobre la melodía de Fourier está la que abrió la puerta a los escáneres médicos modernos. Su inventor fue Johann Radon. Había nacido en 1887 en Tetschen, Bohemia, entonces una región del Imperio austrohúngaro, actualmente conocida como Děčín en la República Checa. Según todas las fuentes, era afable, de buen trato, callado y estudioso. A pesar de ello, no era tímido y no le costaban las relaciones sociales. Como a muchos académicos y profesionales, le encantaba la música. Antes de la aparición de la radio y la televisión, a menudo las personas se reunían en sus casas para pasar la velada. Radon tocaba muy bien el violín y era un cantante excelente. Matemático de formación, trabajó en un primer momento en el cálculo de variaciones, que era el tema de su tesis doctoral, lo que le llevó de manera natural al novedoso campo del análisis funcional, que se desarrollaba con velocidad. Esta rama, iniciada por matemáticos polacos dirigidos por Stefan Banach, interpretó de nuevo las ideas fundamentales del análisis clásico en términos de espacios de funciones de infinitas dimensiones.

En los albores del análisis, los matemáticos se concentraron en calcular cosas como la derivada de una función, su ritmo de cambio, y su integral, el área que hay bajo su gráfica. Conforme progresaba la disciplina, empezó a prestarse atención a las propiedades generales de las operaciones de diferenciación y de integración y a su comportamiento para combinaciones de funciones. Si se suman dos de estas, ¿qué ocurre con sus integrales? Cobraron relevancia algunas características especiales de la función: ¿es continua (sin saltos)? ¿Es diferenciable (varía con suavidad)? ¿Es integrable (tiene sentido hablar de un área)? ¿Cómo están relacionadas estas propiedades entre sí? ¿Cómo se comporta todo si se toma un límite de una secuencia de funciones o la suma de una serie infinita? ¿Qué *tipo* de límite o suma?

Banach y sus colegas formularon estos problemas más generales en términos de «funcionales». Del mismo modo que una función convierte un número en otro, un funcional convierte una función en una cantidad o en otra función. «Integrar» y «diferenciar» son ejemplos. Un truco brillante que descubrieron los matemáticos polacos y otros es que es posible tomar teoremas acerca de funciones de números y convertirlos en teoremas acerca de funcionales. La afirmación resultante puede ser cierta o no: lo divertido es descubrir cuál de estas dos cosas ocurre. La idea se popularizó porque teoremas bastante prosaicos acerca de funciones se convierten en otros, en apariencia mucho más profundos, sobre funcionales, aunque a menudo sigue siendo válida la misma demostración sencilla. Otro truco era dejar de lado todos los aspectos técnicos acerca de cómo integrar fórmulas complicadas de senos, logaritmos, etcétera, y replantearse lo fundamental. ¿De qué trata *en realidad* el análisis? La característica más básica de este campo resultó ser lo cerca que están dos números. Esto se mide con su resta o diferencia, en el orden que sea que la haga positiva. Una función es continua si diferencias pequeñas en los números de entrada dan lugar a diferencias pequeñas en los resultados. Para encontrar la derivada, hay que aumentar la variable una cantidad muy pequeña y ver cómo cambia la función en proporción a ese incremento minúsculo. Para pasar al siguiente nivel del juego, con funcionales, hace falta definir lo que significa que dos *funciones* estén próximas entre sí. Hay muchas maneras de hacerlo. Puede considerarse la resta entre sus valores en algún punto dado y comprobar que es pequeña (para todos los puntos). O verificarse que la integral de la diferencia es pequeña. Cada elección lleva a un «espacio de funciones» diferente, que contiene todas las funciones con las propiedades especificadas, dotado de su propia «métrica» o «norma». En la analogía de los números y de las funciones, este

espacio representa el papel del conjunto de los reales o de los complejos y un funcional es una regla para convertir una función de un espacio de funciones en otra de otro espacio. La transformada de Fourier es un ejemplo de funcional de una importancia particular, que convierte una función en sus coeficientes de la secuencia de Fourier. La transformada inversa funciona al revés: convierte secuencias de números en funciones.

Desde este punto de vista, una buena parte del análisis clásico cobraba sentido de repente como ejemplos del análisis funcional. Es posible concebir las funciones de una o más variables reales o complejas como funcionales más bien simples sobre espacios más bien sencillos (el conjunto de los reales, el conjunto de los complejos o los espacios vectoriales con una cantidad finita de dimensiones formados por secuencias de estos números). Una función de tres variables no es más que un funcional definido en el espacio de todos los tríos de números reales. Otros menos evidentes, tales como la «integral», se definen en (digamos) el espacio de todas las funciones continuas que van del espacio tridimensional a los reales con la métrica «integrar el cuadrado de la diferencia de los valores de las dos funciones en cuestión». La principal diferencia estriba en los *espacios*: los números reales y el espacio tridimensional tienen una cantidad finita de dimensiones, pero estas son infinitas en el que forman todas las funciones continuas. El análisis funcional no es más que el análisis habitual, pero realizado en un espacio de infinitas dimensiones.

Otra gran innovación del periodo se adaptaba también sin fisuras a esta configuración: una teoría novedosa, más general y fácil de manejar de la integración propuesta por Henri Lebesgue, bajo el nombre de «teoría de la medida». Una medida es una cantidad como el área o el volumen, que asigna un número a un conjunto de puntos en algún espacio. El giro novedoso estaba en que este conjunto puede ser complicado en extremo, aunque algunos son *tan* complicados que ni siquiera el concepto de medida de Lebesgue tiene validez en ellos.

El cálculo de variaciones, el tema de la tesis de Radon, *grita* «funcional» a los cuatro vientos desde el momento en que se repara en que trata de encontrar funciones (no números) con propiedades óptimas. De modo que era un paso natural para Radon separarse del cálculo clásico de variaciones y adentrarse en el análisis funcional. Lo hizo con muy buenos resultados y varias ideas y teoremas importantes en teoría de la medida y en el análisis funcional llevan su nombre.

Entre estos se encuentra la transformada de Radon, que descubrió en 1917. Desde el punto de vista del análisis funcional, es un primo hermano matemático de la de Fourier. Se parte de una imagen en el plano, que se concibe como una fotografía en blanco y negro con regiones en las que hay varios tonos de gris. Estos son susceptibles de representarse por un número real, que vaya desde 0 (negro) hasta 1 (blanco). Puede aplastarse la imagen en un plano en cualquier dirección y sumar los números que representan las regiones claras y oscuras, con lo que se obtiene una proyección de la fotografía. La transformada de Radon captura todas estas proyecciones aplastadas en todas direcciones. La idea importante de verdad es la transformada inversa, que permite reconstruir la imagen original a partir de estas proyecciones.

Hasta lo que yo sé, Radon estudió su transformada por motivos matemáticos en exclusiva. Su artículo sobre ella no menciona ninguna aplicación y lo más parecido que hay es una mención breve de relaciones con la física matemática, en concreto con la teoría de potenciales, que es la base común de la electricidad, del magnetismo y de la gravedad. Parecía estar mucho más centrado en las matemáticas y en eventuales generalizaciones. En obras posteriores investigó un análogo tridimensional, en el que una distribución de regiones claras y oscuras en el espacio se aplasta en todos los planos posibles, y también obtuvo una fórmula de reconstrucción para esa operación. Más tarde, otros encontraron generalizaciones a cantidades mayores de dimensiones. Podría ser que Radon estuviese motivado por las radiografías, que realizan justo este tipo de proyección con la distribución de órganos y huesos en el cuerpo humano, e interpretase «claro» y «oscuro» como diferencias en atenuación de los rayos X. Pero se tardaría todavía un siglo hasta que su descubrimiento encontrase aplicación en dispositivos cuyas capacidades de investigar el interior del cuerpo humano parecen casi milagrosas.

*

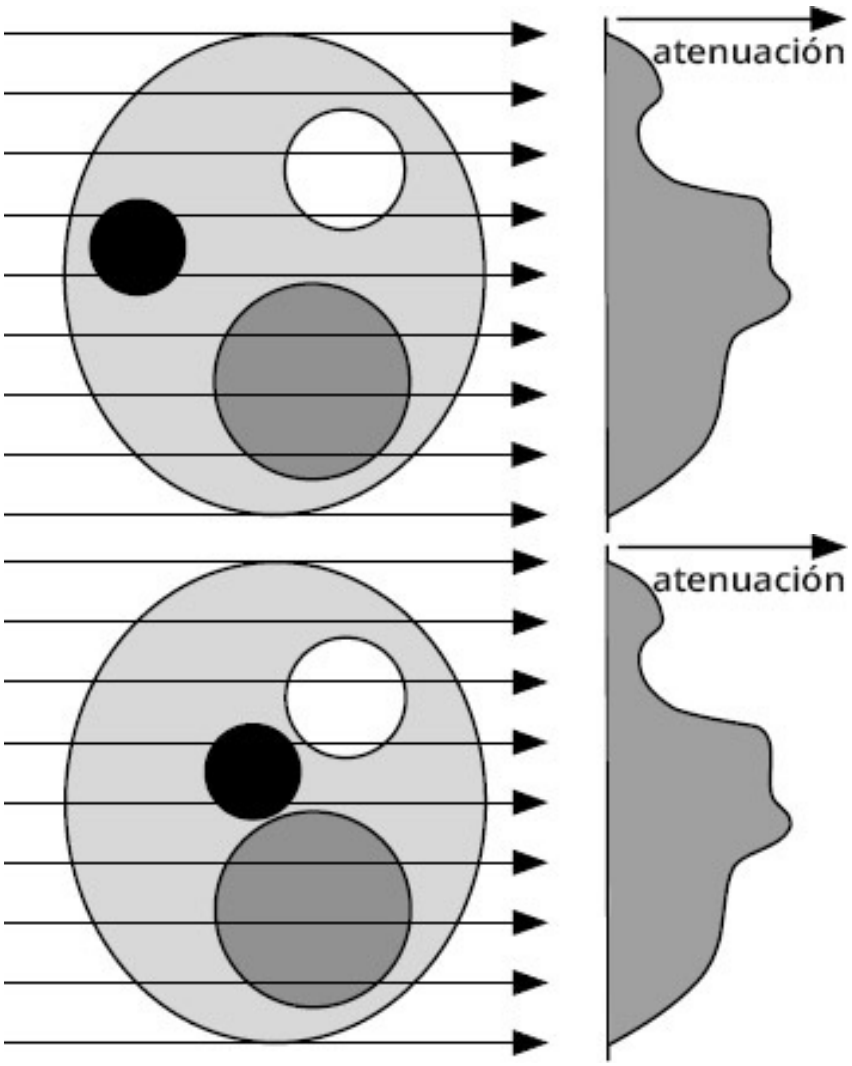
Los escáneres de tomografía axial computarizada (TAC, a menudo denominados en la actualidad escáneres de tomografía computarizada, TC) emplean rayos X para crear imágenes tridimensionales del interior del cuerpo humano. Estas se almacenan en un ordenador y pueden manipularse para mostrar los huesos y los músculos, o para localizar tumores cancerosos. Otros tipos de escáner, tales como los ultrasonidos, se emplean mucho también.

¿Cómo sabe un escáner lo que hay dentro del cuerpo sin tener que cortar para abrirlo? Es sabido que los rayos X atraviesan con facilidad los tejidos blandos, mientras que otros más duros, como el hueso, atenúan más la radiación. Pero una radiografía solo muestra la densidad promedio del tejido visto en una dirección fija. ¿Cómo puede transformarse esto en una imagen tridimensional? Radon empieza su artículo diciendo que ha resuelto este problema:

> Cuando se integra una función de dos variables *x*, *y* (una *función punto f*(*P*) en el plano), sujeta a condiciones de regularidad adecuadas a lo largo de una línea recta arbitraria *g*, entonces se obtienen en la integral los valores *F*(*g*), una *función línea*. En la parte A del presente artículo, el problema que se resuelve es la inversión de esta transformación funcional lineal, es decir, se responde a las siguientes cuestiones: ¿puede cualquier función línea que satisfaga condiciones de regularidad adecuadas considerarse como construida de esta manera? De ser así, ¿se conoce *f* solo a partir de *F* y cómo puede calcularse *f*?

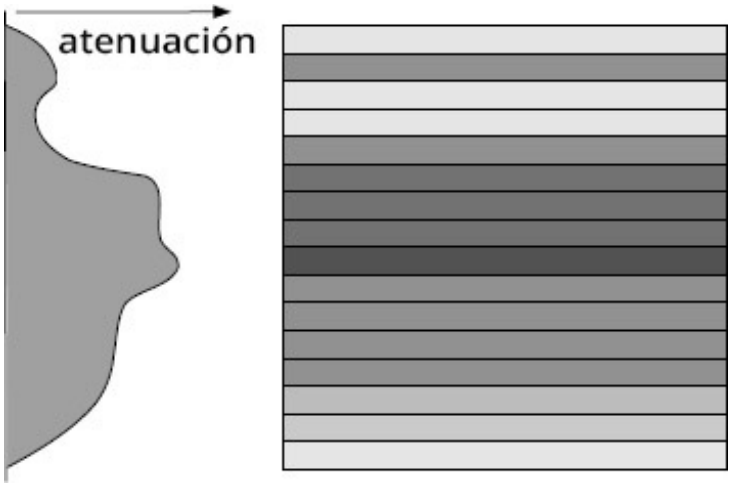
Su respuesta, la transformada de Radon inversa, es una fórmula que reconstruye la disposición interna de los tejidos (más en concreto, su grado de atenuación de los rayos X) a partir del conjunto completo de proyecciones en todas las direcciones.

Para ver cómo funciona, describiré primero lo que una única radiografía (proyección) puede observar. Esta se toma en un corte bidimensional a través del cuerpo. La imagen muestra una vista esquemática de haces paralelos de rayos X que atraviesan un corte de un cuerpo que contiene varios órganos internos de diferente atenuación frente a la radiación. Conforme el haz atraviesa estos órganos, la intensidad de los rayos que pasan al otro lado varía. A mayor atenuación del órgano, a lo largo de esa dirección en concreto, menor será la cantidad de radiación. Puede representarse la variación de la intensidad observada con la posición del haz.



Cuanto mayor es la atenuación en una región, más oscura resulta. *Arriba*: escanear un solo corte del cuerpo desde una única dirección proporciona una gráfica de la atenuación observada ante los rayos X solo en esa dirección. *Abajo*: diferentes disposiciones internas dan como resultado la misma gráfica.

En efecto, una única imagen de este tipo aplasta la distribución de escala de grises dentro del cuerpo a lo largo de la dirección de los haces. Desde un punto de vista técnico, se trata de una proyección de la distribución en esa dirección. Está bastante claro que una sola imagen así no es suficiente para decir con exactitud cómo están dispuestos los órganos. Por ejemplo, si se desplaza el órgano negro en la dirección del haz, no cambia la proyección. No obstante, si se toma otra radiografía del cuerpo desde la vertical, el cambio en posición del disco negro tiene un efecto visible en esa gráfica de atenuación. De manera intuitiva, puede obtenerse incluso más información acerca de las posiciones espaciales de los órganos y tejidos al tomar una serie completa de radiografías, cada una rotada levemente en comparación con la anterior, hasta que se haya observado el cuerpo desde un número grande de direcciones. Pero ¿es suficiente información para encontrar las posiciones con exactitud?



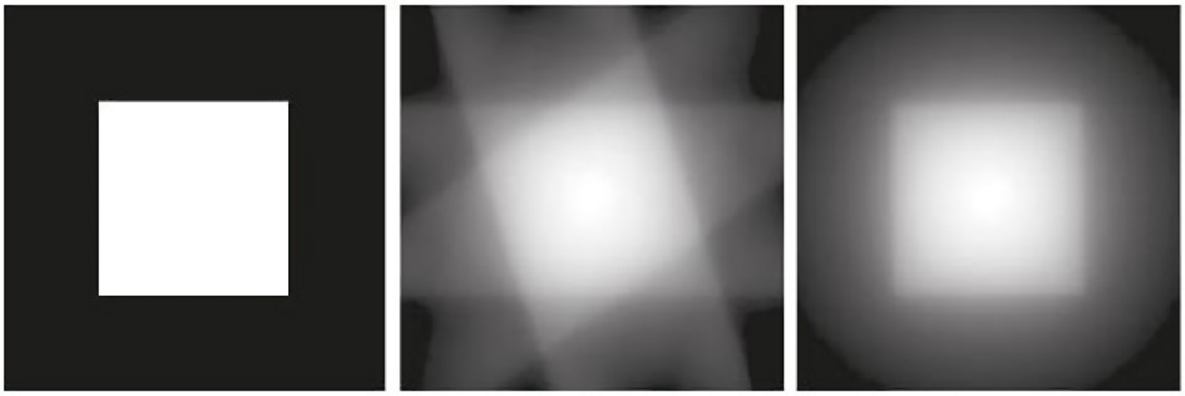
Conversión de la gráfica de atenuación en una serie de filas en escala de grises alineadas con la dirección del haz de rayos X.

Radon demostró que si se conocen las gráficas de atenuación cuando se ve el corte del cuerpo desde todas las direcciones posibles, puede deducirse la distribución en escala de grises bidimensional de los tejidos y de los órganos con precisión. De hecho, hay una manera muy sencilla de lograrlo, denominada retroproyección. Esta difumina la distribución en escala de grises a lo largo de la dirección de la proyección, pero lo hace de manera uniforme. De modo que se obtiene una región cuadrada llena de filas grises. Estas son más oscuras cuanto más alto es el valor correspondiente en la gráfica. De manera intuitiva, se difumina el color gris uniformemente a lo largo de la tira porque no puede afirmarse a partir de la proyección dónde se encuentra ningún órgano interno en concreto.

Es posible hacer esto para todas las direcciones de la serie original de radiografías. La fórmula de inversión de Radon dice que cuando todas estas imágenes en filas se giran en el ángulo correspondiente y se superponen, de modo que se sumen sus valores en la escala de grises en cada punto, el resultado (en una escala adecuada) reconstruye la distribución original de órganos internos. La siguiente imagen muestra cómo funciona esto cuando la forma original es un cuadrado y se reconstruye este por retroproyección desde cinco y desde cien direcciones diferentes. Cuantos más ángulos se usen, mejor será el resultado.

Una vez que se ha reconstruido la distribución de tejidos de un único corte, se desplaza el cuerpo una distancia pequeña y se hace lo mismo de nuevo. Y una vez y otra, hasta que se ha seccionado el cuerpo, de manera conceptual, como un pan de molde. Entonces pueden recomponerse los cortes al «superponerlos» en el ordenador y obtener una descripción completa de la distribución tridimensional de tejido. Este método de detectar estructuras en

tres dimensiones a partir de una serie de cortes bidimensionales se conoce como tomografía y se emplea desde hace mucho en las observaciones por microscopio para ver dentro de objetos sólidos, tales como insectos o plantas. La técnica básica es sumergir el objeto en cera y luego obtener secciones muy finas mediante un dispositivo que es como una cortadora de fiambre en miniatura, llamado microtomo (del griego *mikros*, «pequeño», y *temnein*, «cortar»). Los escáneres TC emplean la misma idea, solo que hacen los cortes con rayos X y trucos matemáticos.



Izquierda: un cuadrado. *Centro*: retroproyección desde cinco direcciones. *Derecha*: retroproyección desde cien direcciones.
Jen Beatty, «The Radon transform and the mathematics of medical imaging», 2012, Tesis de fin de grado, Tesis 646, disponible en: digitalcommons.colby.edu/honorstheses/646.

Después de eso, procesar los datos tridimensionales y proporcionar todo tipo de información relacionada es cosa de técnicas matemáticas rutinarias. Es posible ver el aspecto que tendrían los tejidos a lo largo de un corte completamente diferente, o mostrar solo los tejidos de cierto tipo, o los músculos, órganos y huesos con códigos de colores. Todo lo que se le antoje a uno. Las principales herramientas ahora son métodos de procesamiento de imágenes habituales, que en última instancia dependen de geometría de coordenadas tridimensional.

En la práctica no es tan sencillo. El escáner no toma una cantidad infinita de radiografías desde un continuo de direcciones, sino un número finito grande desde orientaciones discretas separadas muy poco entre sí. Hay que modificar las matemáticas para tener esto en cuenta. Es una gran ayuda filtrar los datos para evitar artefactos en la imagen que son resultado de emplear un conjunto discreto de ángulos de observación. Pero el funcionamiento básico es justo el que había elaborado Radon, más de cincuenta años antes de la invención del primer escáner. El ingeniero en electrónica inglés Godfrey Hounsfield construyó el primer prototipo funcional en 1971. La teoría ya

había sido desarrollada entre 1956 y 1957 por el físico estadounidense, nacido en Sudáfrica, Allan Cormack, quien la publicó entre 1963 y 1964. Cuando lo hizo no estaba al tanto de los resultados de Radon, así que desarrolló lo que necesitaba por sí mismo, aunque después encontró el artículo original, que es más general. El desarrollo de Hounsfield y Cormack de la tomografía axial computarizada les valió el Premio Nobel de Fisiología o Medicina en 1979. Aquella máquina costaba 300 dólares estadounidenses. Hoy en día, un escáner TC comercial cuesta más de 1,5 millones de dólares.

Los escáneres no se emplean solo en medicina. Los egiptólogos los usan en la actualidad de manera rutinaria para descubrir lo que hay dentro de una momia sin deshacer los vendajes. Son capaces de examinar el esqueleto y cualesquiera órganos internos restantes, buscar señales de fracturas o de varias enfermedades y encontrar dónde hay amuletos religiosos escondidos. A menudo, los museos exponen momias virtuales con una pantalla táctil que pueden controlar los visitantes para retirar capas de vendajes de lino, luego la piel y los músculos, hasta que solo quedan los huesos. Todo ello se basa en las matemáticas incorporadas en el ordenador: geometría tridimensional, procesamiento de imágenes y métodos de representación gráfica.

Hay muchos otros tipos de escáneres: de ultrasonido, que emplean ondas sonoras; de tomografía por emisión de positrones (TEP), que detectan las partículas subatómicas emitidas por sustancias radioactivas inyectadas en el cuerpo; e imágenes por resonancia magnética (IRM), que registran los efectos magnéticos en los núcleos de los átomos y que solía denominarse resonancia magnética nuclear (RMN), hasta que al Departamento de Relaciones Públicas empezó a preocuparle que la palabra «nuclear», que se asocia con bombas y centrales energéticas, creara ansiedad en los pacientes. Cada tipo de escáner tiene su propia historia matemática.

10
¡Una sonrisa, por favor!

> La única función de la cámara es no resultar un estorbo para el fotógrafo cuando este hace sus fotos.
>
> KEN ROCKWELL,
> *Por qué tu cámara no importa*

La humanidad sube a internet cerca de un billón de fotografías al año, lo que indica una estimación demasiado optimista del interés que pueden tener los demás en nuestros selfis de vacaciones, en el bebé recién nacido o en objetos varios, algunos innombrables. Es rápido, es fácil y todos los teléfonos son cámaras. Se invierte una gran cantidad de matemáticas en el diseño y la fabricación de estas últimas. Esas lentes de alta precisión en miniatura son milagros de la tecnología e implican algunos elementos de física matemática muy sofisticados acerca de la refracción de la luz en medios sólidos curvos. En este capítulo, me quiero centrar solo en un aspecto de la fotografía actual: la compresión de imágenes. Las cámaras digitales, incorporadas a un teléfono o independientes, guardan imágenes muy detalladas como archivos binarios. Las tarjetas de memoria parecen poder almacenar más información de la que les cabe en realidad. ¿Cómo es capaz un archivo informático pequeño de contener tantas fotografías detalladas?

Lo cierto es que estas fotografías incluyen mucha información redundante que puede eliminarse sin perder definición. Hay técnicas matemáticas que lo hacen posible de manera sistemática y calculada con cuidado. El estándar JPEG en las cámaras digitales automáticas pequeñas, que hasta hace muy poco era el formato de archivo más común y que todavía se usa mucho, emplea *cinco* transformaciones matemáticas separadas, realizadas en sucesión. Estas incluyen análisis de Fourier discreto, álgebra y teoría de códigos. Las transformaciones están incorporadas en el *software* del aparato, que comprime los datos antes de escribirlos en la tarjeta de memoria.

A no ser, claro, que se prefiera el formato RAW, que es, en esencia, lo que la cámara ha captado tal cual es. La capacidad de las tarjetas de memoria crece con tanta rapidez que ya no es imprescindible comprimir los archivos. Pero entonces se acaba por manipular imágenes de 32 MB, cuando solían tener una décima parte de ese tamaño, y se tarda más en subirlos a la nube. Si la molestia merece la pena o no, depende de quién sea el usuario y para qué quiere las fotos. Si es un profesional, es probable que sea imprescindible. Mientras que si es un turista de los de cámara automática, como yo, le puede caber una imagen buena de verdad de un tigre en un archivo JPEG de 2 MB.

La compresión de imágenes es una parte importante del problema más general de la compresión de datos y no deja de ser de vital importancia a pesar de los enormes avances en la tecnología. Cada vez que el internet de próxima generación se hace diez veces más veloz y adquiere una capacidad mayor, algún genio inventa un nuevo formato (vídeo tridimensional de ultra alta definición, pongamos por caso) que necesita muchísimos más datos que antes, y vuelta a la casilla de salida.

En ocasiones, no hay otra opción más que aprovechar cada byte de capacidad de un canal de señales. El 4 de enero de 2004, en Marte, algo cayó del cielo, chocó contra el suelo y rebotó. De hecho, el *Mars Exploration Rover A*, también conocido como *Spirit*, dio 27 botes, rodeado como estaba por globos inflables como una especie de plástico de burbujas cósmico, un aterrizaje puntero. Después de una comprobación general y varios procedimientos de arranque, se puso en marcha para explorar la superficie de aquel planeta y pronto se le unió su compañero, *Opportunity*. Estos dos vehículos tuvieron un éxito enorme y han enviado a la Tierra cantidades enormes de datos. En ese momento, el matemático Philip Davis señaló que la misión descansaba sobre una cantidad tremenda de matemáticas, pero que «el público casi no es consciente de esto». Resulta que no era solo el público. En 2007, Uffe Jankvist y Bjørn Toldbod, estudiantes de posgrado de matemáticas en Dinamarca, visitaron el Laboratorio de Propulsión a Chorro en Pasadena en una misión de investigación periodística: sacar a la luz las matemáticas ocultas en el programa Mars Rover. Pero les dijeron:

—No tenemos nada de eso. De verdad que no empleamos algebra abstracta, ni teoría de grupos ni nada de eso.

Esto era preocupante, así que uno de los daneses preguntó:

—¿Excepto en el canal de codificación?

—¿Eso emplea álgebra abstracta?

—Los códigos Reed-Solomon se basan en campos de Galois.

—La primera noticia que tengo.

De hecho, las misiones espaciales de la NASA emplean unas matemáticas muy avanzadas para comprimir y codificar datos de una manera que corrija los errores inevitables de transmisión. Es algo necesario cuando el emisor se encuentra a mil millones de kilómetros de la Tierra y tiene la potencia de una bombilla. (Enviar los datos a través de un satélite artificial en la órbita de Marte, como *Mars Odissey* o *Mars Global Surveyor*, también ayuda algo.) La mayor parte de los ingenieros no necesitan saberlo, así que lo ignoran. Es un microcosmos de la incomprensión pública de las matemáticas.

*

Todo lo que hay en un ordenador, sea un correo electrónico, una imagen, un vídeo o un disco de Taylor Swift, se almacena en la memoria como una cadena de dígitos binarios, «bits», 0 y 1. Ocho de ellos forman un byte, y 1 048 576 bytes, un megabyte (MB). Una fotografía habitual de baja resolución ocupa cerca de 2 MB. Aunque todos los contenidos digitales tienen esta forma, las diferentes aplicaciones emplean formatos distintos, de modo que el significado de los datos depende del *software*. Todos los tipos de archivos tienen una estructura matemática oculta y a menudo es más importante la conveniencia del procesado que el tamaño. Dar formato a los datos de manera adecuada puede hacer que sean redundantes, por lo que se utilizan más bits de los que requiere en realidad el contenido de información. Esto brinda la oportunidad de comprimir los datos al eliminar la redundancia.

El idioma escrito (y hablado) es redundante en un grado muy alto. Para demostrarlo, he aquí una frase anterior de este capítulo en la que se ha eliminado cada quinto carácter:

rode_do c_mo e_taba_por _lobo_ inf_able_ com_ una_espe_ie d_ plá_tico_de bu_buja_

Es probable que se pueda adivinar lo que dice con un poco de esfuerzo y atención. La información restante es suficiente para reconstruir la totalidad de la frase original.

Sea como fuere, este libro será mucho más agradable a la vista si no convenzo a la editorial de eliminar cada quinta letra para ahorrar tinta. Las palabras correctas resultan más fáciles de procesar para el cerebro porque es lo que ha aprendido a hacer. No obstante, cuando se quiere transmitir una

cadena de bits a un receptor, una secuencia más corta de 0 y 1 es más eficaz. En los albores de la teoría de la información, pioneros como Claude Shannon se dieron cuenta de que la redundancia permite codificar una señal con menos bits. De hecho, él mismo demostró una fórmula que decía cuánto se puede acortar una señal mediante un código para una cantidad dada de redundancia.

Esta es esencial, porque los datos que no son redundantes no se pueden comprimir sin perder información. La demostración es una sencilla reducción al absurdo. Supongamos, por ejemplo, que alguien está interesado en mensajes que tienen una longitud de diez bits, como 1001110101. Hay exactamente 1024 cadenas como esta. Supongamos que quiere comprimir estos diez bits de datos en ocho, de los cuales hay exactamente 256 combinaciones posibles. De modo que hay cuatro veces más mensajes que secuencias comprimidas. No existe forma alguna de asignar una cadena de ocho bits a cada combinación de diez dígitos y que cada una de estas reciba ocho caracteres diferentes al resto. Si cada mensaje de diez bits tiene la misma probabilidad de aparecer, resulta que no hay modo alguno, por astuto que sea, de superar esta limitación. No obstante, si algunos mensajes son muy comunes y otros muy raros, puede elegirse un código que asigne cadenas cortas (pongamos seis bits) a los más comunes y otras más largas (pongamos doce bits) a los que no lo son tanto. Hay montones de combinaciones posibles de doce dígitos, así que no se van a agotar nunca. Cada vez que aparece un mensaje poco común se suman dos bits a la longitud, pero cada vez que hay uno que sí lo es se reducen cuatro. Con las probabilidades adecuadas, se quitan más bits de los que se añaden.

Ha surgido una rama entera de las matemáticas, la teoría de códigos, en torno a estas técnicas. En general, son mucho más sutiles que la que he explicado y aprovechan de manera habitual características del álgebra abstracta para definir los códigos. Esto no debería resultar sorprendente: ya se ha visto en el capítulo 5 que estos últimos son, en lo fundamental, funciones matemáticas y que las que se toman de la teoría de números son particularmente útiles. El objetivo en ese caso era la confidencialidad, mientras que aquí es la compresión de datos, pero son válidos los mismos argumentos generales. El álgebra se refiere a la *estructura*, al igual que la redundancia.

La compresión de datos, y por lo tanto de imágenes, aprovecha la redundancia para crear códigos que reducen el tamaño de los archivos de algún tipo concreto. En ocasiones, el método de compresión no tiene pérdidas: la información original puede reconstruirse de manera exacta a

partir de la versión comprimida. En otras, se pierden datos y la reconstrucción solo es una aproximación del archivo original. Esto sería un problema para, digamos, el saldo del banco, pero a menudo no lo es para imágenes: el truco está en disponer las cosas de modo que la aproximación tenga la misma apariencia que la imagen original para el ojo humano. En ese caso la información que se pierde de manera irremediable no tiene mayor importancia.

La mayor parte de las imágenes del mundo real son redundantes. Las instantáneas de las vacaciones a menudo contienen grandes superficies de cielo azul, en general del mismo tono celeste, de modo que es posible sustituir un montón de píxeles, que contienen todos el mismo número, por dos pares de coordenadas de las esquinas opuestas de un rectángulo y por un código corto que viene a decir «rellenar *esta* región con *este* tono de azul». Este método no tiene pérdidas. No es el que se utiliza en realidad, pero ilustra el motivo por el que la compresión sin pérdidas es factible.

*

Soy de la vieja escuela. Es decir, uso una cámara con una tecnología que es por lo menos de, *¡halaaaaaa!*, diez años atrás. ¡Vergonzoso! Estoy lo bastante puesto al día en tecnología como para emplear mi teléfono para tomar instantáneas en ocasiones, pero no es algo que me salga de manera natural y en viajes de vacaciones importantes, como un safari fotográfico de tigres en los parques nacionales de India, prefiero llevar una pequeña cámara digital automática. Crea archivos de imagen con nombres como IMG_0209.jpg. La nomenclatura JPG indica que el formato es JPEG, las iniciales de Joint Photographic Experts Group, y se refiere a un sistema de compresión de datos. Es un estándar del sector, aunque ha evolucionado a lo largo de los años y ahora se presenta en varias formas diferentes desde el punto de vista técnico.

El formato JPEG**[62]** emplea al menos cinco pasos diferentes en sucesión, la mayoría de los cuales comprimen los datos a partir del anterior. El primero son los datos originales en bruto y los demás vuelven a codificar la fotografía para mayor compresión. Las imágenes digitales están compuestas de millones de cuadrados minúsculos, denominados píxel, del inglés *picture elements*. Los datos en bruto de la cámara asignan una cadena de bits a cada píxel para indicar tanto el color como el brillo. Ambas cantidades se representan de manera simultánea como proporciones de tres componentes: rojo, verde y

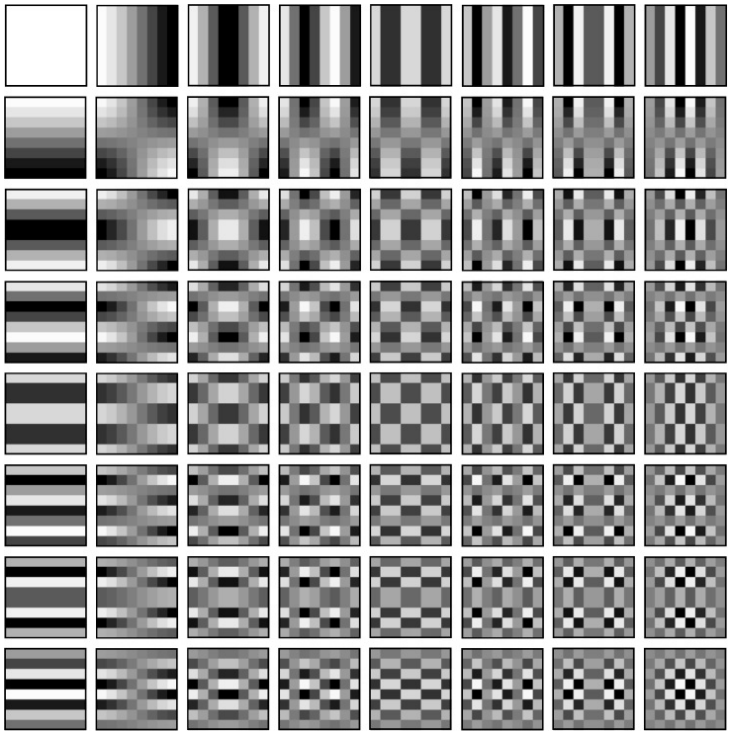
azul. Proporciones bajas de los tres corresponden a colores pálidos y las altas, a intensos. Estos números se convierten en otros tres relacionados que se corresponden mejor con la manera en que el cerebro humano percibe las imágenes. El primero, la luminancia, se refiere al brillo general y se mide con cifras que van desde el negro hasta el blanco, pasando por tonos de gris cada vez más claros. Si se elimina la información de color, queda una imagen a la antigua, en blanco y negro (en realidad con muchos tonos de gris). Los otros dos números, conocidos como crominancia, son las diferencias entre la luminancia y las cantidades de luz azul y roja respectivamente.

De manera simbólica, si R = rojo, G = verde y B = azul, entonces los números iniciales para R, G y B se sustituyen por la luminancia, R + G + B, y las dos crominancias, (R + G + B) − B = R + G y (R + G + B) − R = G + B. Si se conocen R + G + B, R + G y G + B, es posible calcular R, G y B, de modo que no hay pérdidas en este paso.

El segundo sí tiene pérdidas. Recorta los datos de crominancia a valores más pequeños al reducir la resolución. Solo este paso reduce el tamaño del archivo a la mitad. Es aceptable porque, en comparación con lo que «ve» la cámara, el sistema visual humano es más sensible al brillo y menos a las diferencias de color.

El tercer paso es el más matemático. Comprime la información de luminancia mediante el empleo de una versión digital de la transformada de Fourier, que se vio en el capítulo 9 en relación con los escáneres médicos. En ese contexto, la transformada de Fourier original, que convierte las señales en sus frecuencias componentes o viceversa, fue modificada para representar proyecciones de imágenes en escala de grises. En esta ocasión, se representan las propias imágenes en escalas de grises, pero en un formato digital sencillo. La fotografía se divide en bloques minúsculos de píxeles de 8×8, de modo que hay 64 valores diferentes de luminancia posibles, uno para cada píxel. La transformada de coseno discreta, una versión digital de la transformada de Fourier, representa esta región de 8×8 en escala de grises como una superposición de múltiplos de 64 imágenes estándar (ver más adelante). Estos múltiplos son las *amplitudes* de las imágenes correspondientes. Estas tienen la apariencia de rayas y tableros de ajedrez de varias anchuras. De este modo, puede obtenerse cualquier cuadrado de píxeles de 8×8, así que, de nuevo, este paso no tiene pérdidas. En las coordenadas en el bloque, estos cuadrados estándar son versiones discretas de cos (mx) cos (ny) para varios números enteros m y n, dónde x está en horizontal e y en vertical y ambas van de 0 a 7.

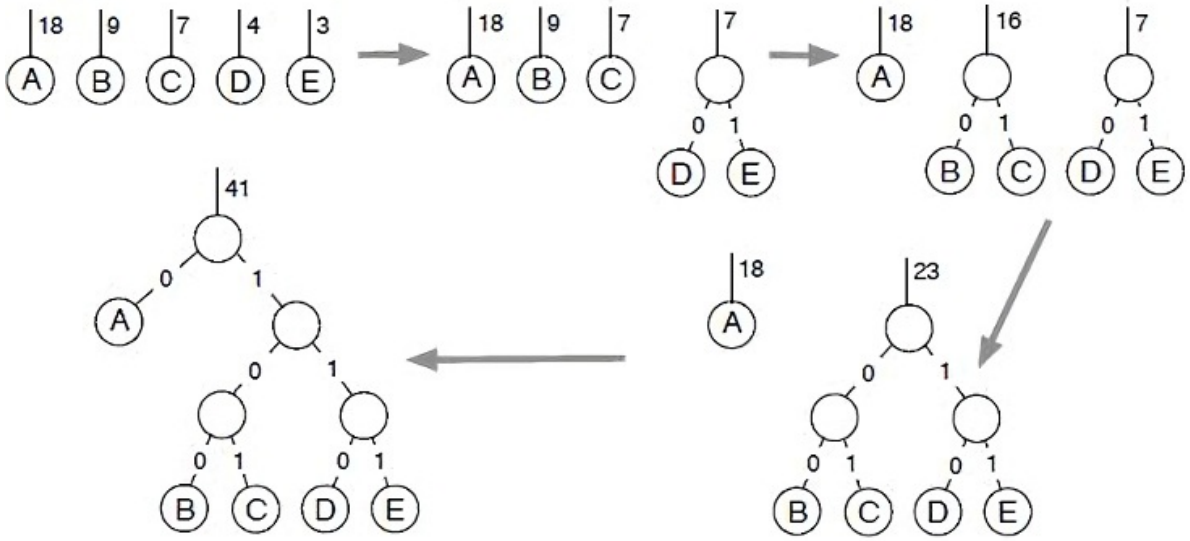
Aunque la transformada de Fourier es discreta no tiene pérdidas, no carece de sentido aplicarla porque hace que el cuarto paso sea posible. De nuevo, este se basa en las limitaciones de sensibilidad de la visión humana, que da lugar a la redundancia. Es fácil darse cuenta si el brillo o el color varían a lo largo de una región grande en una imagen. Pero si lo hacen en zonas pequeñas, el sistema visual los difumina y se percibe solo el promedio. Es por eso que las imágenes impresas son comprensibles, incluso aunque al examinarlas de cerca representen los tonos de gris mediante distribuciones de puntos negros sobre el fondo blanco del papel. Esta característica de la visión humana quiere decir que los patrones de líneas muy delgadas son menos importantes, de modo que sus amplitudes pueden registrarse con menor precisión.



Los 64 patrones básicos de la transformada de coseno discreta.

El quinto paso es un truco técnico llamado «codificación Huffman», para registrar con más eficacia las amplitudes de los 64 patrones básicos. David Huffman inventó este método en 1951 cuando todavía era un estudiante. Se le había asignado la tarea de escribir un trabajo sobre códigos binarios de eficacia óptima, pero era incapaz de demostrar que existiera alguna codificación con esta característica. Estaba a punto de rendirse cuando se le ocurrió un método nuevo y después demostró que era el mejor de los posibles. *Grosso modo*, la solución es codificar un conjunto de símbolos mediante cadenas binarias y usar estas después como un diccionario para convertir un

mensaje a una forma cifrada. Esto debe hacerse de modo que se reduzca al mínimo la longitud total de la información codificada.



Construcción de una codificación de Huffman.

Por ejemplo, los símbolos podrían ser las letras del alfabeto. Hay 27, de modo que bastaría con asignar cadenas de cinco bits, digamos A = 00001, B = 00010, etcétera. Se necesita que sean de cinco porque con cuatro solo pueden formarse 16 combinaciones. Pero esto sería poco eficaz, porque letras muy poco frecuentes, como la Z, emplearían el mismo número de dígitos que las comunes, como la E. Es mejor asignar una cadena corta, como 0 o 1, a E y otras de longitud progresiva al resto de letras conforme disminuye su probabilidad. No obstante, ya que las cadenas de código tienen diferentes longitudes, se necesita información adicional para indicar al receptor dónde separarlas para formar letras individuales. Puede hacerse mediante un prefijo reconocible delante de la secuencia de código, pero esta no debe contener el prefijo: ninguna cadena de código aparece como inicio de otra más larga. Si lo hiciera, no se podría saber dónde termina la primera. Una letra poco común, como Z, necesita muchos más bits, pero como no es habitual, esto se ve más que compensado por la cadena más corta de E. La longitud global de un mensaje característico es menor.

La codificación de Huffman logra este objetivo mediante la formación de un «árbol», una especie de grafo sin bucles cerrados, y es muy común en informática teórica porque representa toda una estrategia de decisiones de sí o no, cada una dependiente de la anterior. Los nodos hoja del árbol son los símbolos A, B, C… y surgen dos ramas de cada uno, correspondientes a los dos bits 0 y 1. Cada uno está etiquetado con un número, denominado peso,

que indica la frecuencia con la que aparece el símbolo correspondiente. El árbol se construye paso a paso al unir las dos hojas menos frecuentes en un nuevo nodo «padre», del que los anteriores pasan a ser «hijos». El peso asignado al padre es la suma de los pesos de los dos nodos hijos. Este procedimiento continúa hasta que se han unido todos los símbolos de este modo. Entonces se lee la cadena de código de una letra a partir de la trayectoria que la lleva a ella.

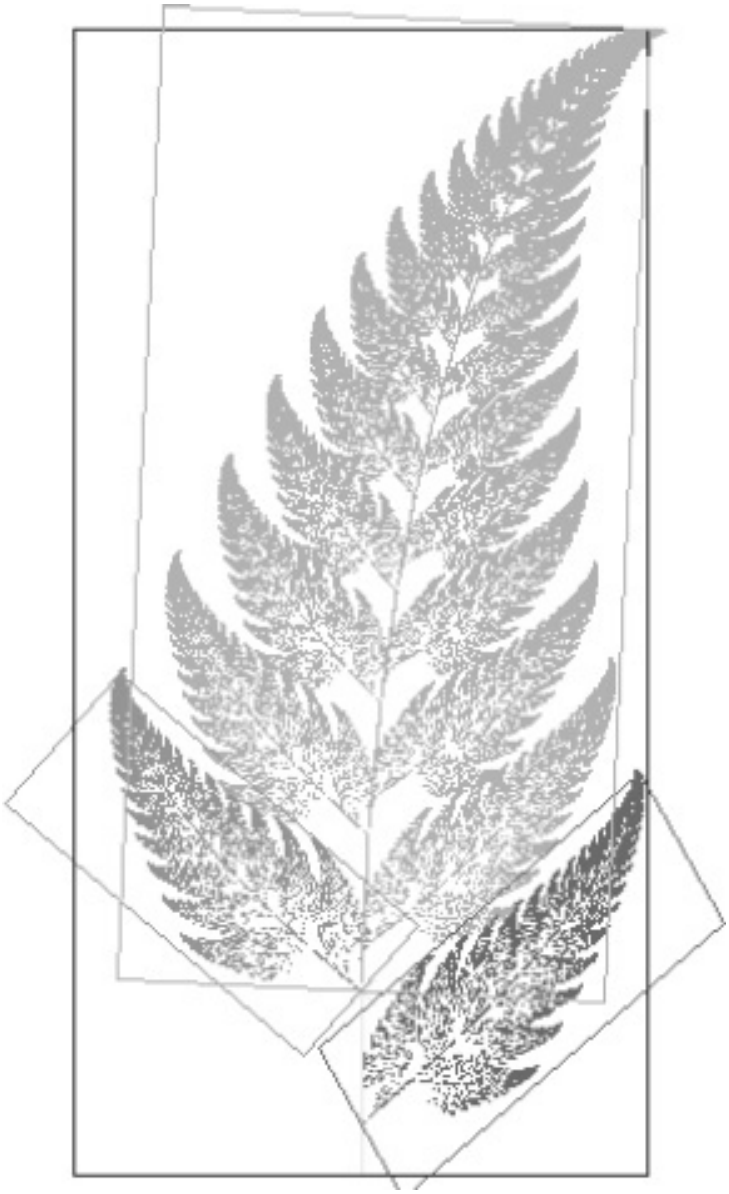
Por ejemplo, en la parte superior izquierda de la imagen se muestran cinco símbolos, A, B, C, D y E y las cantidades 18, 9, 7, 4, 3, que indican lo comunes que son en lengua inglesa. Los dos menos probables son D y E. En la segunda etapa, arriba a la mitad, se unen estos para formar un nodo padre (sin número) con peso 4 + 3 = 7 y los símbolos D y E pasan a ser hijos. Las dos ramas que llevan a ellos se etiquetan 0 y 1. Este proceso se repite hasta que se han unido todos los símbolos (abajo a la izquierda). Ahora es posible leer las cadenas de código con solo seguir las trayectorias a lo largo del árbol. Hasta A se llega por una sola rama etiquetada 0. Se llega a B por la 100, a C por la 101, a D por 110 y a E por 111. Cabe destacar que A, el símbolo más común, tiene una trayectoria corta, mientras que otros menos habituales tienen otras más largas. Si en lugar de esto se hubiese empleado un código de longitud fija, se necesitarían al menos tres bits para cinco símbolos, porque solo hay cuatro combinaciones posibles de dos dígitos. En este caso, las cadenas más largas tienen tres bits, pero la más común solo uno, de modo que en promedio esta codificación es más eficaz. Este procedimiento garantiza que las secuencias no contienen prefijos porque cada trayectoria que lleva a un símbolo se detiene en él. No puede seguir a otro. Es más, al empezar con el menos probable, se asignan las cadenas más cortas a los más habituales. Es una idea muy lúcida, fácil de programar y muy sencilla desde un punto de vista conceptual una vez que se ha entendido.

Cuando la cámara crea un archivo JPEG, es la electrónica incorporada la que hace todos estos cálculos al vuelo, tan pronto como se ha tomado la fotografía. El proceso de compresión tiene pérdidas, pero la mayoría de las personas no se dan cuenta nunca. En cualquier caso, las pantallas de nuestros ordenadores y las impresiones en papel no reproducen los colores ni el brillo con exactitud, a no ser que se hayan calibrado con cuidado. Una comparación directa de la imagen original y de la versión comprimida hace que las diferencias sean más evidentes, pero incluso entonces, se necesita un experto para darse cuenta de estas cuando el tamaño del archivo se ha reducido a un 10 % del original. La mayoría de los mortales solo reparamos en ellas cuando

la reducción llega a cerca del 3 %. De modo que con el formato JPEG es posible almacenar diez veces más imágenes en una tarjeta de memoria dada que con los datos RAW originales. Así es como sucede la magia, con este complicado proceso de cinco pasos, realizado en un abrir y cerrar de ojos entre bambalinas, que emplea por lo menos cinco ramas diferentes de las matemáticas.

*

Otra manera de comprimir imágenes surgió a finales de la década de 1980 a partir de la geometría de fractales. Como se recordará, un fractal es una forma geométrica con estructura detallada en todas las escalas, como la línea de la costa y las nubes. Hay una cantidad asociada a cada uno de ellos llamada dimensión fractal y que es una medida de lo fino o intrincado que es. Lo habitual es que no sea un número natural. Una clase útil de fractales, manejables desde un punto de vista matemático, comprende los que son autosimilares: fracciones pequeñas de ellos, aumentadas de manera adecuada, tienen la apariencia exacta de partes mayores del conjunto. El ejemplo clásico es un helecho, que está compuesto de docenas de ramas más pequeñas, cada una de las cuales parece un helecho en miniatura. Es posible representar los fractales autosimilares mediante un dispositivo matemático denominado sistema iterativo de funciones (SIF). Es un conjunto de reglas que dicen cómo reducir las copias de la forma y mover los rectángulos resultantes de modo que se ajusten unos con otros para reproducir el total. Puede reconstruirse el fractal a partir de estas reglas e incluso hay una fórmula para la dimensión fractal.



Un helecho fractal, hecho a partir de tres copias transformadas de sí mismo.

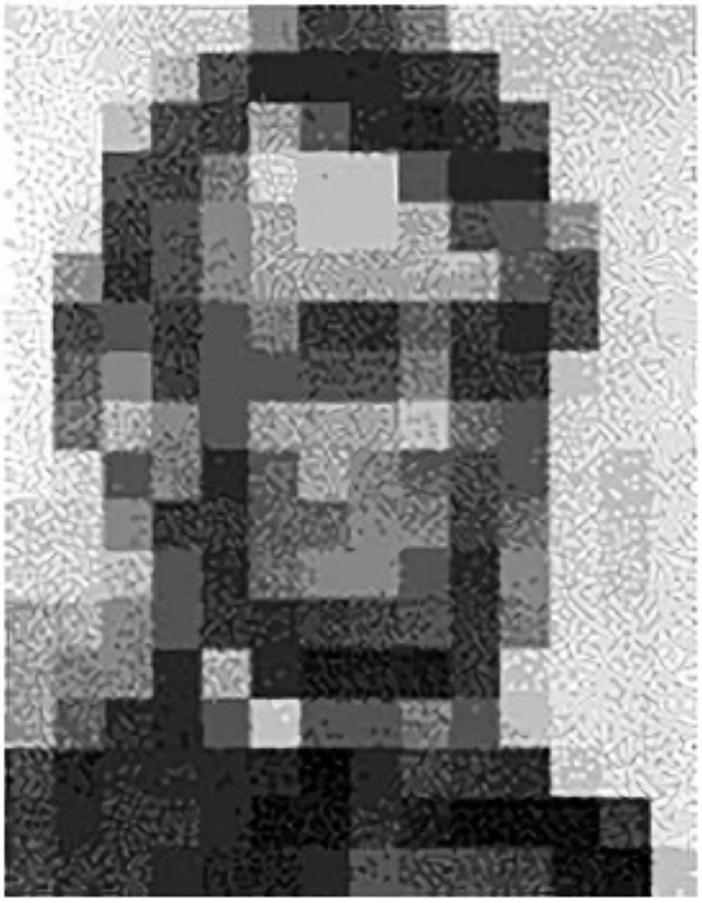
En 1987, Michael Barnsley, un matemático fascinado por estos objetos geométricos, se dio cuenta de que la autosimilitud podía constituir la base de un método de compresión de imágenes. En lugar de emplear grandes cantidades de dígitos para registrar cada detalle minúsculo del helecho, solo se codifica el SIF correspondiente, lo que requiere muchos menos datos. Un *software* puede reconstruir la imagen de la planta a partir del SIF. Junto con Alan Sloan, fundó una empresa, Iterated Systems Inc., que ha registrado más de veinte patentes. En 1992, la empresa consiguió un gran avance: un método automático para encontrar reglas de SIF adecuadas, que busca regiones pequeñas de la imagen que puedan percibirse como versiones reducidas de otras áreas tan solo un poco más grandes. De modo que emplea muchos más rectángulos para cubrir todo el conjunto. No obstante, también es

completamente general y aplicable a cualquier imagen, no solo a las que son autosimilares de manera evidente. La compresión fractal no alcanzó en este terreno el grado de éxito del JPEG por una serie de motivos, pero se empleó en varias aplicaciones prácticas. Es probable que la que más éxito haya tenido sea la enciclopedia digital de Microsoft, *Encarta*, en la que todas las imágenes estaban comprimidas mediante un SIF.

A lo largo de la década de 1990, la empresa hizo arduos intentos para ampliar el método a la compresión de vídeo, pero no cuajó ninguno, sobre todo porque los ordenadores no eran lo bastante rápidos ni tenían suficiente memoria en esa época. Se tardaban quince horas en comprimir un minuto de grabación. Todo eso ha cambiado en la actualidad y se han logrado proporciones de 200:1 en compresión fractal de vídeo, en un minuto más o menos por cada fotograma. No obstante, la potencia mejorada de los ordenadores también hace que sean factibles otros métodos y se ha abandonado la compresión fractal de vídeos por el momento. Pero la idea subyacente fue útil durante un tiempo y no deja de ser una atractiva posibilidad.

*

Los seres humanos tenemos un truco muy extraño para descifrar una imagen de baja calidad: entrecerramos los ojos. Es sorprendente lo a menudo que esto nos ayuda a comprender lo que hay en realidad en la ilustración, sobre todo si está un poco borrosa o si es una imagen informática con píxeles muy gruesos. Hay una conocida fotografía compuesta de 270 cuadrados negros, blancos y grises, que creó Leon Harmon de los Laboratorios Bell en 1973 para un artículo sobre la percepción humana y el reconocimiento de patrones por ordenador. ¿Quién es? En última instancia se hace vagamente reconocible como Abraham Lincoln si se mira con detenimiento, pero si se entrecierran los ojos, *parece* él de verdad.



¿Quién es? ¡Hay que entrecerrar los ojos!

Lo hace todo el mundo, así que se sabe de sobra cómo funciona, aunque parezca una locura. ¿Por qué motivo se consigue mejorar una fotografía de baja calidad al hacer que empeore la visión? La respuesta es psicológica en parte: al entrecerrar los ojos el sistema de procesamiento visual del cerebro pasa a «modo imagen de mala calidad», lo que se presume pone en marcha algoritmos especiales para el tratamiento de la percepción ocular que han evolucionado para gestionar datos deficientes. Pero la otra parte es que, de manera paradójica, entrecerrar los ojos actúa como una especie de etapa de procesado previo que limpia la imagen de un cierto modo muy útil. Por ejemplo, difumina los bordes pixelados de Lincoln, de forma que ya no parece una superposición de cuadrados grises.

Hace unos cuarenta años, los matemáticos empezaron a investigar un equivalente preciso y versátil del recurso humano de entrecerrar los ojos, denominado análisis de ondículas. La técnica sirve para datos numéricos tanto como para imágenes y se introdujo en primer lugar para extraer la estructura de algunas escalas espaciales particulares. Las ondículas permiten ver el bosque y, al mismo tiempo, no reparar en que está compuesto de muchos árboles y arbustos muy retorcidos.

El estímulo original fue teórico en buena medida: eran una técnica estupenda para comprobar teorías científicas sobre cosas como el flujo turbulento de fluidos. Más recientemente, las ondículas han adquirido algunas aplicaciones prosaicas en extremo. En Estados Unidos, el FBI las utiliza para almacenar datos de huellas dactilares por menos dinero y los cuerpos de policía de otros países han seguido su ejemplo. Las ondículas no solo permiten analizar imágenes: también comprimirlas.

En JPEG, las fotografías se comprimen al descartar datos que son menos relevantes para la visión humana. No obstante, pocas veces se representa la información de modo que sea evidente qué bits son menos importantes. Supongamos que alguien quiere mandar un correo electrónico a un amigo con un dibujo en una hoja de papel bastante sucia. Hay muchos puntos negros minúsculos, aparte de la propia imagen. Cuando una persona mira el folio, se da cuenta enseguida de que las marcas son irrelevantes, pero un escáner no lo hace. Se limita a escanear la página línea a línea y representa el dibujo como una cadena larga de señales binarias de blanco y negro. No puede discernir si un punto negro en particular es una parte fundamental de la imagen o algo irrelevante. Algunos podrían ser en realidad la pupila de una vaca en la lejanía o las manchas de la piel de un leopardo.

El principal impedimento es que las señales del escáner no representan los datos de lo que aparece en la hoja de un modo que haga que reconocer y eliminar los elementos no deseados sea fácil. No obstante, hay otras maneras de hacerlo. La transformada de Fourier sustituye una curva por un listado de amplitudes y frecuencias y codifica la misma información de un modo diferente. Y cuando los datos se representan de otra manera, las operaciones que son difíciles o imposibles en un dominio pueden volverse fáciles en el otro. Por ejemplo, se parte de una conversación telefónica, se calcula su transformada de Fourier y se eliminan todas las partes de la señal cuyas componentes tengan frecuencias demasiado altas o bajas para ser captadas por el oído humano. Entonces se aplica la transformada inversa al resultado para recuperar sonidos que, a nuestro juicio, son idénticos a los iniciales. Ahora se pueden mandar más conversaciones por el mismo canal de comunicación. Esto no es factible sobre la señal original sin transformar porque no tiene la «frecuencia» como una característica evidente.

Para ciertos propósitos, la técnica de Fourier tiene un defecto: los senos y cosenos componentes no se acaban nunca. La transformada de Fourier no funciona bien para representar señales compactas. Un pitido breve es una señal sencilla, pero se necesitan cientos de senos y cosenos para reproducir

uno que sea moderadamente convincente. El problema no es reproducir bien su forma, sino conseguir que todo lo demás, aparte de este sonido, sea igual a cero. Hay que anular las colas ondulantes de longitud infinita de todos esos senos y cosenos, lo que se hace añadiendo aún más senos y cosenos de frecuencias elevadas, en un intento desesperado para que se cancelen todos los elementos no deseados. En última instancia, la versión transformada acaba por ser más complicada y necesitar más datos que la señal original.

Izquierda: una curva sinusoide no tiene fin. *Centro*: una ondícula está localizada. *Derecha*: tres generaciones más.

La transformada de ondículas cambia todo eso al emplear pitidos como sus componentes básicos. Esto no es fácil y no puede hacerse con una señal compacta de este tipo, pero para un matemático es evidente por dónde empezar. Se elige alguna forma particular de pitido para que actúe como una ondícula madre. Se generan ondículas hijas (y nietas, bisnietas…, lo que sea) al deslizar la original hacia los lados a varias posiciones y al expandirla o comprimirla mediante un cambio de escala. Para representar una función más general, se suman múltiplos adecuados de estas componentes en escalas diferentes. Del mismo modo, las curvas sinusoides y cosinusoides básicas de Fourier son «sinusoidículas madre» y todos los senos y cosenos del resto de frecuencias son hijas.

Huellas dactilares. *Izquierda*: original. *Derecha*: después de una compresión de datos a 1/26 del tamaño.

Las ondículas están diseñadas para describir datos similares a pitidos de manera eficaz. Es más, debido a que las hijas y las nietas son solo versiones a diferente escala de la madre, es posible centrarse en niveles concretos de detalle. Si se quieren eliminar estructuras de pequeña escala, se quitan todas las ondículas bisnietas en la transformada. Imaginemos representar un leopardo con ondículas: unas pocas grandes para el cuerpo, más pequeñas para los ojos, el morro y las manchas, luego unas minúsculas para los pelos. Para comprimir los datos, pero que no deje de parecer un leopardo, se decide que los pelos individuales no son importantes y se eliminan esas ondículas bisnietas. Quedan las manchas y todavía parece un leopardo. No es posible hacer nada parecido con tanta facilidad (si es que se puede en absoluto) con una transformada de Fourier.

La mayoría de las herramientas matemáticas requeridas para desarrollar las ondículas se conocen en forma abstracta desde hace medio siglo o más, en la rama de Banach del análisis funcional. Cuando aparecieron las ondículas, se descubrió que la poco comprensible maquinaria de esta rama era justo lo que se necesitaba para entenderlas y desarrollarlas para constituir una técnica eficaz. El requisito previo principal para que se pusiera en marcha la máquina del análisis funcional era una forma adecuada de la ondícula madre. Se quiere que todas las hijas sean independientes de esta última desde el punto de vista matemático, sin que se solape la información codificada por la madre ni por las hijas y sin que ninguna parte de estas sea redundante. En la terminología del análisis funcional, la madre y la hija deben ser ortogonales.

A principios de la década de 1980, el geofísico Jean Morlet y el físico matemático Alexander Grossmann desarrollaron una ondícula madre factible. En 1985, el matemático Yves Meyer mejoró la propuesta de Morlet y

Grosssman. Ingrid Daubechies hizo en 1987 el descubrimiento que abrió las puertas al desarrollo del campo. Las ondículas madre previas tenían una conveniente apariencia de pitidos, pero todas tenían una cola matemática minúscula que se agitaba hasta el infinito. Daubechies construyó una sin cola en absoluto: fuera de cierto intervalo, es siempre *exactamente* cero. Su ondícula madre era un pitido genuino, confinado por completo a una región finita del espacio.

*

Las ondículas actúan como una especie de lente de aumento numérica, que se enfoca en características de la información que ocupan escalas espaciales concretas. Es posible emplear esta capacidad para analizar datos y también para comprimirlos. Al manipular la transformada de ondículas, el ordenador «entrecierra sus ojos» sobre la imagen y descarta las escalas de resolución no deseadas. Esto es lo que decidió hacer el FBI en 1993. En ese momento, su base de datos de huellas dactilares contenía 200 millones de entradas, almacenadas como impresiones en tinta sobre fichas de papel, y estaban modernizando sus registros, digitalizando las imágenes y guardando los resultados en un ordenador. Una ventaja evidente es la capacidad de buscar huellas que correspondan a las encontradas en la escena del crimen con rapidez.

Una imagen convencional con suficiente resolución crea un archivo informático que ocupa diez megabytes por cada ficha de huellas dactilares. Por lo tanto, la base de datos del FBI ocupa 2000 terabytes de memoria. Se reciben al menos 30 000 fichas nuevas diarias, de modo que las necesidades de almacenamiento crecen en 2,4 billones de dígitos binarios al día. El FBI necesitaba con desesperación comprimir la información. Intentaron el formato JPEG, pero es inútil para huellas dactilares (a diferencia de las instantáneas de vacaciones) cuando la «relación de compresión» (la proporción del tamaño de los datos originales respecto a los comprimidos) es alta, en torno a 10:1. Luego, las imágenes sin comprimir eran insatisfactorias debido a «artefactos de límite de bloque», en los que la subdivisión en cuadrados de 8×8 dejaba bordes distinguibles. Por supuesto, el método no era de mucha utilidad para el FBI a menos que lograse relaciones de compresión de al menos 10:1. Los artefactos de límite de bloque no son solo un problema estético: limitan seriamente la capacidad de los algoritmos de encontrar huellas correspondientes. Métodos alternativos basados en Fourier también

introducen artefactos inaceptables, todos los cuales pueden atribuirse al problema de las «colas» infinitas de los senos y cosenos de Fourier. De modo que Tom Hopper por el FBI y Jonathan Bradley y Chris Brislawn por el Laboratorio Nacional de Los Álamos, decidieron codificar los registros digitalizados de huellas dactilares con ondículas y emplear un método denominado cuantificación escalar de ondículas, o WSQ, por sus siglas en inglés.

En lugar de eliminar la información redundante al crear artefactos de límite de bloque, la WSQ elimina los detalles pequeños de toda la imagen, tan pequeños que son irrelevantes por lo que respecta a la capacidad del ojo de reconocer la estructura de la huella dactilar. En las pruebas del FBI, tres métodos de ondículas diferentes dieron mejores resultados que otros dos de Fourier, tales como el formato JPEG. En conjunto, la WSQ resultó ser el método más razonable. Proporciona una relación de compresión de al menos 15:1, por lo que reduce el coste de la memoria de almacenamiento un 93 %. La WSQ es ahora el estándar para el intercambio y almacenamiento de imágenes de huellas dactilares. La mayoría de los cuerpos de policía estadounidenses la emplean para este fin con archivos comprimidos a 500 píxeles por pulgada. Para huellas de resolución más alta emplean el formato JPEG[63].

Las ondículas aparecen casi en todas partes. El equipo de Dennis Healy ha aplicado técnicas de mejora de imágenes basadas en ellas a los escáneres de TC, TEP e IRM. También han empleado ondículas para mejorar las estrategias de adquisición de datos de los escáneres en primer lugar. Ronald Coifman y Victor Wickerhauser las han utilizado para eliminar ruido indeseable de grabaciones. Uno de sus triunfos fue una interpretación en la que Johannes Brahms ejecutaba una de sus propias *Danzas húngaras*, grabada originalmente en 1889 en un cilindro de cera que en parte se había derretido. Había sido grabada de nuevo en un disco de 78 rpm. Coifman partió de una emisión radiofónica de esta copia, cuando ya la música era virtualmente inaudible sobre el ruido de fondo. Después de la limpieza con ondículas se podía oír lo que tocaba Brahms. No a la perfección, pero se *podía* oír.

Hace cuarenta años el análisis funcional no era más que otra rama ignota de las matemáticas abstractas, cuyas aplicaciones principales estaban en la física teórica. La aparición de las ondículas lo ha cambiado todo. El análisis funcional proporciona ahora los puntales necesarios para desarrollar nuevos tipos de ondículas con características especiales que las hacen importantes en ciencia y tecnología aplicadas. Tienen un impacto invisible en nuestras vidas

cotidianas: en la lucha contra el crimen, en medicina y en la próxima generación de música digital. El día de mañana, conquistarán el mundo.

11
¿Falta mucho?

> Un viaje de mil millas comienza con un solo paso.
>
> LAO-TSE,
> *El camino de Lao-tse*

Todos los padres que conducen conocen la situación. La familia va a visitar a la abuela, a 500 kilómetros de distancia y seis horas de carretera. Los niños van en el asiento de atrás. Solo ha pasado media hora desde que comenzó el viaje, cuando empiezan a quejarse: «¿Falta mucho?».

Tengo una cuenta pendiente con mis primos transatlánticos, que parecen estar convencidos de que la frase es: «¿Ya llegamos?». Sin duda, lo es en algunas partes de América, pero no debería serlo, porque está claro que esta variante es el resultado de un malentendido. La respuesta a esta segunda versión es siempre evidente: o bien hemos llegado *ya* y es irrelevante, o no lo hemos hecho y no tiene sentido preguntarlo. No, lo que ocurrió fue que, durante un viaje largo cualquiera, cuando los niños estaban inquietos, sus padres les tranquilizaron con amabilidad (o puede ser que solo con hastío). «Ya casi hemos llegado». Aunque faltasen todavía cinco horas para hacerlo. Con eso los mantenían callados durante un rato. Sea como fuere, después de varios viajes, los niños empezaron a dejar caer discretas indirectas, presa de la desesperación más que esperanzados: «¿Ya *casi* llegamos?». Es una pregunta razonable, porque no puede saberse solo con mirar por la ventanilla. A no ser que alguien conozca el paisaje, por supuesto. Tuvimos un gato que lo hacía.

¿Falta mucho? ¿Dónde estamos? Hace dos décadas se necesitaba un mapa, capacidad de entenderlo y alguien en el asiento del copiloto para saberlo. En la actualidad, todo se confía a la magia de la electrónica. Se consulta el navegador por satélite. Es verdad, a veces se acaba en mitad de la nada. Hace poco, un coche se metió en un río, gracias a las indicaciones cortesía del aparato. No se le puede quitar el ojo a la carretera. Aunque incluso eso falla a veces. El año pasado acabamos en los jardines de una casa

de campo mientras buscábamos un hotel rural, porque el navegador no podía distinguir entre una carretera de verdad que parecía la entrada a una finca y la entrada a una finca que parecía una carretera de verdad.

La navegación por satélite parece cosa de magia. Hay una pantalla en el coche que muestra parte de un mapa y que indica con exactitud dónde se encuentra el automóvil. Al conducir, la imagen se desplaza de modo que el símbolo del vehículo está siempre en el lugar correcto. El dispositivo sabe en qué dirección va el coche y el nombre o número de la calle en la que se halla en ese momento. Avisa de los atascos. Conoce el destino, la velocidad y cuándo se supera el límite, dónde están los radares y cuánto falta para llegar. Solo hay que enseñar a los niños a leer estos datos y no necesitarán volver a preguntar nunca.

Como escribió el gran autor de ciencia ficción y futurólogo Arthur C. Clarke: «Cualquier tecnología lo suficientemente avanzada es indistinguible de la magia». Otro escritor del mismo género, Gregory Benford, lo reformuló como: «Cualquier tecnología distinguible de la magia no es lo suficientemente avanzada». La navegación por satélite es lo bastante avanzada, pero no es magia. ¿Cómo funciona?

Conoce el destino porque se le dice. Se pulsan letras y números en la pantalla. Eso es obvio. También es verdad que es la *única* parte evidente. La magia restante se basa en una tecnología compleja: satélites en órbita, muchos; señales de radio; codificaciones; números pseudoaleatorios; procesamiento informático complejo en grandes cantidades. Hay algoritmos para encontrar la ruta más rápida/barata y menos dañina desde el punto de vista del medioambiente. La física fundamental es imprescindible: mecánica orbital basada en la ley de Newton de la gravitación universal, complementada por las teorías especial *y* general de la relatividad de Einstein, que refinan la interpretación clásica. Los satélites dan vueltas en el espacio y transmiten señales temporales. En el otro extremo, en el coche, casi todo ocurre en un minúsculo circuito impreso de ordenador. Más unos cuantos chips de memoria para almacenar el mapa y cosas así.

No se ve ninguna de estas cosas, así que parece cosa de magia.

No hace falta decir que buena parte de esta brujería es matemática y que requiere amplias dosis de cálculos de muchos tipos, por no mencionar las ingentes cantidades de física, química, ciencia de materiales e ingeniería. Algo de tratamiento psiquiátrico sería también una buena idea para algunos usuarios, pero qué se le va a hacer.

Si se pasan por alto la fabricación y el diseño de satélites, junto con la tecnología necesaria para lanzarlos al espacio, un navegador requiere todavía al menos siete ramas de las matemáticas sin las cuales no podría funcionar. Las que tengo en mente son:

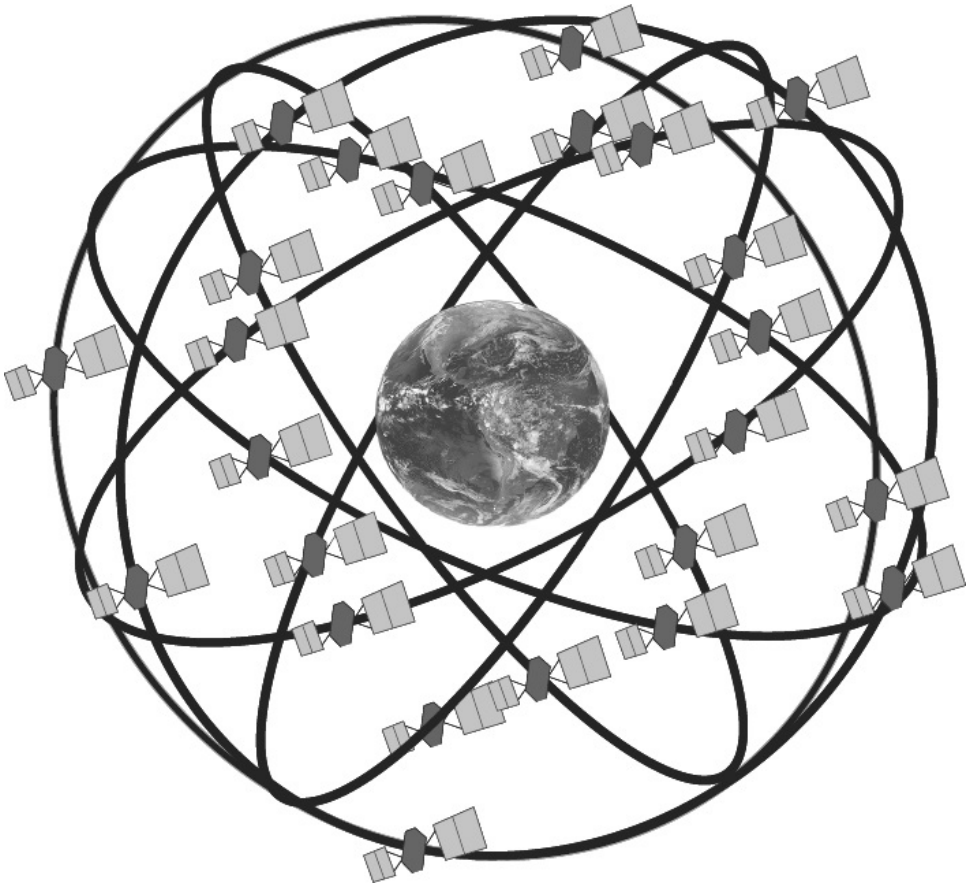
- Calcular las trayectorias de los cohetes de lanzamiento para poner los satélites en órbita.
- Diseñar un conjunto de órbitas que proporcionen buena cobertura: por lo menos tres satélites, y preferiblemente más, deben ser visibles desde cualquier ubicación dada en todo momento.
- Emplear un generador de números pseudoaleatorios para crear señales, lo que hace posible medir con mucha precisión la distancia a la que se encuentra cada satélite.
- Emplear la trigonometría y los datos orbitales para determinar la ubicación del coche.
- Emplear las ecuaciones de la relatividad especial para corregir los cálculos debido al efecto de las altas velocidades de los satélites sobre el paso del tiempo.
- Emplear las ecuaciones de la relatividad general para corregir los cálculos debido al efecto de la gravedad terrestre sobre el paso del tiempo.
- Resolver una variante del problema de la persona viajante para encontrar la mejor ruta conforme al criterio que se haya seleccionado: rápida, corta, respetuosa con el medioambiente.

Consideraré la mayor parte de ellas con más detalle en las siguientes páginas y me concentraré en las más sorprendentes.

*

La navegación por satélite depende de señales temporales de una exactitud increíble producidas por relojes atómicos muy precisos y enviadas desde una cierta cantidad de satélites especiales en órbita. Un cronómetro de cesio, dejado a su suerte, tiene una precisión de 5 partes en 10^{14}, o de cuatro nanosegundos por día. Eso corresponde a un error en la posición del vehículo de cerca de un metro diario. Se compensa esta desviación gradual ajustando los relojes de manera periódica desde la estación de seguimiento en tierra. Hay otras fuentes de errores en estos instrumentos de las que hablaré más adelante.

En la actualidad, hay varios sistemas de navegación por satélite, pero me concentraré en el primero y el de uso más generalizado, el Sistema de Posicionamiento Global o GPS. El proyecto comenzó en 1973 a iniciativa del Departamento de Defensa de Estados Unidos. El núcleo de su estructura es un conjunto de satélites en órbita: al principio eran 24, ahora hay 31. El primer prototipo se lanzó en 1978 y el conjunto completo estaba operativo en 1993. En un primer momento, el servicio estaba limitado a usos militares, pero una orden ejecutiva del presidente Ronald Reagan en 1983 lo puso a disposición de los civiles en un formato de menor resolución. El GPS se halla en un proceso de mejora y varios países disponen ahora de sus propios sistemas de posicionamiento por satélite, empezando por el Sistema Global de Navegación por Satélite (GLONASS) de Rusia, con una precisión de hasta dos metros. En 2018, China inició su Sistema de Navegación por Satélite BeiDou, que debería estar a pleno rendimiento en cualquier momento a partir de ahora. El equivalente de la Unión Europea se llama Galileo. Como el Reino Unido ha dejado la UE no formará parte de él, pero en un triunfo de la ideología sobre el sentido común, el Gobierno británico ha anunciado que desarrollará y pondrá en órbita su propia versión. La India está creando NavInc y Japón construye el Sistema por Satélite Cuasicenital (QZSS), con el que eliminará su dependencia del GPS en 2023.



Constelación original de 24 satélites del GPS, cuatro de ellos en cada una de seis órbitas distintas. Wikipedia.

Desde un punto de vista funcional, la estructura de este último comprende tres «segmentos»: el espacial (los satélites), el de control (estaciones de seguimiento en tierra) y el del usuario (el conductor en el coche). Los primeros envían señales temporales. El segmento de control vigila las órbitas y la precisión de los relojes y si es necesario transmite instrucciones para modificar unas y ajustar los otros. El usuario dispone de un receptor barato de baja potencia, lo bastante pequeño como para caber dentro de un teléfono móvil, que dice a la aplicación dónde se encuentra.

El conjunto de satélites se denomina en general una «constelación», el nombre que reciben desde tiempo inmemorial las disposiciones de estrellas en el cielo nocturno. La constelación original del GPS comprende 24 elementos, cada uno en una órbita aproximadamente circular a 20 200 kilómetros por encima de la Tierra, o 26 600 kilómetros de su centro. Pasaré por alto los satélites adicionales posteriores porque no afectan a la idea principal; tan solo hacen que el sistema sea más confiable y preciso. Hay seis órbitas en planos que cruzan el ecuador en un ángulo de 55°, separadas de manera regular en torno a este. Cada una de ellas está ocupada por cuatro satélites separados entre sí una distancia similar, que se persiguen de manera permanente. El

radio de la órbita se elige mediante el empleo de matemáticas orbitales, de modo que el satélite regresa a la misma posición y completa una vuelta cada 11 horas y 58 minutos. Con esto se garantiza que se halla por encima de (casi) la misma ubicación en la Tierra dos veces al día, aunque se desvía con lentitud.

El siguiente prodigio matemático es la geometría de las órbitas. Esta configuración de la constelación implica que al menos seis de los satélites son visibles en todo momento (es decir, que se pueden recibir sus señales) desde cualquier punto del planeta. Cuáles de ellos, en concreto, depende de dónde esté el receptor y, además, este conjunto cambia con el paso del tiempo, porque la Tierra rota y los satélites giran en sus órbitas.

El GPS está diseñado de modo que los usuarios no necesitan transmitir información alguna. En su lugar, tienen un receptor que recoge las señales temporales de los satélites que son visibles. El aparato procesa los datos temporales para calcular dónde se encuentra. El principio básico es sencillo, así que veámoslo primero. Después señalaré algunos de los ajustes necesarios para que funcione en el mundo real.

Empecemos con un satélite. Envía señales temporales a partir de las cuales el receptor calcula a qué distancia se encuentra del emisor en ese instante. (Más adelante se verá cómo hace esta deducción). Tal vez sean 21 000 kilómetros. Esta información lo coloca en la superficie de una esfera, centrada en el satélite y con un radio de 21 000 kilómetros. Esto no es algo de una utilidad pasmosa en sí mismo, pero hay al menos otros cinco emisores visibles al mismo tiempo. Me permitiré denominarlos satélite 2, satélite 3 y así hasta el satélite 6. Cada uno transmite señales que se reciben de manera simultánea y cada una de estas sitúa al receptor en otra esfera, centrada también en el satélite correspondiente: esferas 2, 3, 4, 5 y 6. La señal del satélite 2, combinada con la del 1, sitúa al receptor en la intersección de las esferas 1 y 2, que es una circunferencia. El satélite 3 aporta otra esfera, que se cruza con la 1 en otra circunferencia. Ambas circunferencias se cortan a su vez en dos puntos, cada uno de los cuales pertenece a las tres esferas. La señal del satélite 4 proporciona la esfera 4, que en general suele discernir cuál de los dos puntos es la ubicación correcta.

En un mundo perfecto se podría parar aquí y los elementos 5 y 6 serían superfluos. En realidad, no es tan sencillo. Todo adolece de errores. La atmósfera de la Tierra puede degradar la señal, tal vez haya interferencias eléctricas, lo que sea. Para empezar, eso implica que la ubicación está *cerca* de la esfera en cuestión, más que *sobre* ella. En lugar de situarse justo en su

superficie, se halla dentro de un cascarón de pared gruesa que la contiene. De modo que cuatro satélites y cuatro señales son suficientes para localizar el receptor hasta un cierto nivel de precisión, pero no a la perfección. Para mejorar esto, el GPS emplea los elementos adicionales. Sus esferas de pared gruesa recortan la región en porciones todavía más pequeñas. Llegados a este punto, las ecuaciones que determinan la ubicación son, casi con certeza, inconsistentes si se ignoran los errores más probables, pero si se toma prestado un viejo truco de la estadística, es posible calcular la mejor estimación de la localización y reducir al mínimo la imprecisión total. Este se denomina el método de los mínimos cuadrados y fue desarrollado por Gauss en 1795.

La conclusión es que el receptor de GPS solo tiene que hacer una serie sistemática de cálculos geométricos relativamente sencillos, lo que lleva a la mejor estimación posible de la ubicación. Al comparar esta con la forma detallada de la Tierra, puede incluso calcular la altura relativa al nivel del mar. En general, las alturas son menos precisas que las posiciones en latitud y longitud.

*

Tal vez «enviar señales temporales» parezca sencillo, pero no lo es. Si se oye un trueno, se sabe que hay una tormenta cerca, pero el ruido en sí mismo no dice a qué distancia se encuentra. Si además se ve el relámpago, que llega antes que el trueno porque la luz viaja más rápido que el sonido, puede emplearse la diferencia temporal entre las dos señales para estimar la distancia a la que estaba la tormenta (a ojo de buen cubero son tres segundos por kilómetro). No obstante, la velocidad del sonido depende del estado de la atmósfera, por lo que esta regla no tiene una precisión absoluta.

El GPS no puede emplear ondas sonoras como una segunda señal, por razones obvias: serían demasiado lentas y, de todos modos, el espacio es un vacío en el que no se transmite el sonido. Pero la idea subyacente de que es posible inferir una diferencia temporal al comparar dos señales distintas pero relacionadas, apunta en la dirección correcta. Cada satélite envía una secuencia de pulsos de 0 y 1 que no contiene repeticiones, a no ser que se espere un tiempo muy largo para que vuelva a empezar toda la cadena. El receptor de GPS es capaz de comparar la sucesión de 0 y 1 que le llega desde el satélite con la misma cadena generada por una fuente local. La primera tiene un retraso temporal porque debe salvar la distancia entre el emisor y el

receptor y es posible calcular su valor al alinear las señales y ver cuánto hay que desplazar una para que coincida con la otra.

Este proceso puede ilustrarse con palabras de este libro en lugar de 0 y 1.

Supongamos que la señal recibida del satélite es:

alinear las señales y ver cuánto hay,

mientras que, al mismo tiempo, la cadena de referencia desde un lugar cercano es:

ver cuánto hay que desplazar una para…

Puede desplazarse la señal local hasta que coincidan las palabras, de este modo:

alinear las señales y ver cuánto hay

ver cuánto hay que desplazar una para…

Ahora se ve que la secuencia del satélite llega cuatro palabras más tarde que la local.

Todo lo que queda por hacer es generar cadenas de bits adecuadas. Una forma sencilla de obtener combinaciones de 0 y 1 con repeticiones muy poco frecuentes es lanzar una moneda millones de veces y apuntar 0 cuando sale cara y 1 cuando sale cruz. Cada bit tiene una probabilidad de 1/2, de modo que una secuencia concreta de digamos 50 dígitos tiene una probabilidad de $1/2^{50}$, que está en torno a uno entre mil billones. En promedio, se repetirá dentro de cerca de mil billones de pasos en la cadena. Cuando se compara una señal así con una versión desplazada una distancia mucho más pequeña, el desplazamiento «correcto», que produce la correspondencia óptima, es único.

No obstante, los ordenadores no son buenos arrojando monedas. Siguen instrucciones concretas y precisamente se trata de que lo hagan de manera exacta y sin errores. Por suerte, hay procesos matemáticos rigurosos que pueden generar cadenas de bits que *parecen* aleatorias, en todos los sentidos estadísticos razonables, incluso aunque el procedimiento en sí sea determinista. Tal método se conoce como generador de números pseudoaleatorios. Este es el tercer ingrediente matemático principal del GPS.

En la práctica, la cadena de bits del generador de números pseudoaleatorios se combina con otros datos que requiere el sistema, en una técnica denominada modulación. El satélite emite su secuencia a un ritmo relativamente bajo de 50 dígitos por segundo. Combina esta señal con una cadena de bits mucho más rápida del generador de números pseudoaleatorios, que tiene un ritmo de más de un millón de *chips* por segundo. Un *chip* es algo muy parecido a un bit pero que toma los valores +1 y −1 en lugar de 0 o 1. Desde el punto de vista físico, es un pulso de onda cuadrada con amplitud +1 o −1. «Modulación» quiere decir que la cadena de datos original se multiplica por el valor del *chip* en cada instante. Dado que, en comparación, la otra secuencia cambia con mucha lentitud, la técnica de «desplazar y buscar la coincidencia» funciona todavía bastante bien, aunque en ocasiones la correspondencia es idéntica y a veces una señal es el negativo de la otra. Mediante el empleo del método estadístico de correlación, no hay más que desplazar las señales hasta que la correlación sea lo bastante alta.

De hecho, el GPS hace lo mismo de nuevo con otro número pseudoaleatorio y modula la señal a un ritmo diez veces más rápido. El más lento se denomina código de adquisición gruesa y es de uso civil. El más rápido, el código preciso, está reservado para el ejército. También está encriptado y tarda siete días en repetirse.

Los generadores de números pseudoaleatorios se basan de manera general en el álgebra abstracta, tal como polinomios sobre campos finitos, o en la teoría de números, tal como enteros en algún módulo. Un ejemplo sencillo de estos últimos es un generador lineal congruencial. Se elige algún módulo m, dos números a y b (mod m) y una cantidad inicial x_1 (mod m). Después se definen otras sucesivas x_2, x_3, x_4, etcétera, mediante la fórmula

$$x_{n+1} = ax_n + b \pmod{m}$$

El efecto de a es multiplicar el número actual x_n por un factor constante a y después b desplaza ese valor una cantidad fija. Esto proporciona el siguiente elemento de la secuencia y se realiza el proceso de nuevo. Por ejemplo, si $m = 17$, $a = 3$, $b = 5$ y $x_1 = 1$, entonces se obtiene la sucesión

1 8 12 7 9 15 16 2 11 4 0 5 3 14 13 10

que se repite de manera indefinida. Esta no contiene patrones evidentes a simple vista. Por supuesto, en la práctica se toma un m mucho mayor. Hay

algunas condiciones matemáticas que garantizan que la secuencia tarda mucho en repetirse y que satisface las pruebas estadísticas razonables de aleatoriedad. Por ejemplo, tras pasar el resultado a binario, cada número (mod *m*) debería aparecer igual cantidad de veces en promedio y lo mismo debería ocurrir para cada cadena de 0 y 1 de una longitud dada hasta algún tamaño razonable.

Los generadores lineales congruenciales son demasiado sencillos para ser seguros y se han desarrollado variantes más complejas. Un ejemplo es el Mersenne *twister*, inventado en 1997 por Makoto Matsumoto. Muchas personas tienen uno, ya que se emplea en docenas de paquetes de *software* comunes, entre ellos, en las hojas de cálculo de Microsoft Excel. Este método combina los números primos, lo que hace que las matemáticas sean más sencillas, con expresiones binarias adecuadas, lo que hace que los cálculos informáticos sean más fáciles. Un primo de Mersenne es un primo de la forma $2^p - 1$ (con p un primo), tal como $31 = 2^5 - 1$ o $131\,071 = 2^{17} - 1$. Son números raros y ni siquiera se sabe si hay una cantidad infinita de ellos. En enero de 2021 se conocen exactamente 51 primos de Mersenne, el mayor de los cuales es $2^{82\,589\,933} - 1$.

Expresados en binario, los dos ejemplos anteriores son

$31 = 11111$
$131\,071 = 11111111111111111$

con cinco y diecisiete 1 repetidos respectivamente. Esto hace que un ordenador digital los pueda emplear en cálculos con facilidad. El Mersenne *twister* se basa en un primo de Mersenne muy grande, habitualmente $2^{19\,937} - 1$, y sustituye los números en la congruencia por matrices sobre el campo con dos elementos 0 y 1. Satisface las pruebas estadísticas para fragmentos de cadena de hasta 623 bits de longitud.

La señal de GPS también incorpora una emisión de frecuencia mucho más baja que aporta información acerca de la órbita del satélite, de las correcciones de su reloj y de otros factores que afectan al estado del sistema. Esto puede parecer muy complicado y lo es, pero la electrónica moderna es capaz de gestionar instrucciones muy complejas sin equivocarse. Hay buenos motivos para esta complejidad. Ayuda a que el receptor no se quede fijado de manera accidental en alguna otra señal aleatoria que flote por ahí en ese momento, porque es muy improbable que una secuencia perdida reproduzca un patrón tan complejo. Cada emisor tiene asignado su propio código pseudoaleatorio particular, de modo que la propia complejidad garantiza que

el receptor no confunde la señal de un satélite con la de otro. Para remate, todos los satélites pueden transmitir en la misma frecuencia sin producir interferencias con los demás, lo que deja más canales libres en el espectro de radio, cada vez más congestionado. Sobre todo, en las operaciones militares, el enemigo no tiene posibilidad alguna de interferir el sistema ni de enviar señales engañosas. Más en general, el Departamento de Defensa de Estados Unidos está a cargo del código pseudoaleatorio, de modo que controla el acceso al GPS.

*

Aparte del desajuste gradual de los relojes atómicos, hay otras fuentes de error temporal, tales como las órbitas, que son ligeramente diferentes en forma y tamaño de las pretendidas en primer lugar. La estación de seguimiento en tierra envía correcciones al satélite, que las transmite a los usuarios y asegura que todo está sincronizado con los relojes de referencia en el Observatorio Naval de Estados Unidos. Pero son los errores relativistas los que requieren mayor intervención matemática. De modo que en lugar de la física newtoniana de la vieja escuela se necesitan las teorías de la relatividad de Einstein[64].

En 1905, Einstein publicó el artículo «Sobre la electrodinámica de cuerpos en movimiento». Examinaba la relación entre la mecánica de Newton y las ecuaciones de Maxwell para el electromagnetismo y encontraba que ambas teorías eran incompatibles. Un problema crucial es que la velocidad con la que se propagan las ondas electromagnéticas (la velocidad de la luz) no solo es constante en un sistema de referencia fijo, sino que también lo es y tiene el mismo valor en un sistema en movimiento. Si se enciende una linterna desde un coche en marcha, los fotones se desplazan a la misma velocidad que lo habrían hecho si el vehículo estuviese parado.

Por el contrario, en la física clásica, a la velocidad de la luz se habría sumado la del coche. Por lo tanto, Einstein proponía modificar las leyes del movimiento de Newton para garantizar que esa primera magnitud se conserva como una constante absoluta, lo que en particular implica que las ecuaciones del movimiento relativo deben modificarse. Por este motivo, la teoría se denominó «relatividad», lo que es un poco engañoso porque el aspecto principal es que la velocidad de la luz *no* es relativa. Einstein pasó muchos años intentando incorporar la gravedad a su esquema, algo que logró por fin

en 1915. Estas dos teorías, relacionadas pero independientes, se conocen como la relatividad especial y general respectivamente.

Este no es un libro de texto sobre relatividad, de modo que solo voy a revisar por encima algunas características destacadas para tener una visión muy general de lo que esta implica. No hay espacio aquí para entrar en matices filosóficos, e incluso aunque lo hubiera sería una digresión, así que ruego que se me excuse si simplifico demasiado.

En la relatividad especial se modifican las ecuaciones del movimiento para asegurar que la velocidad de la luz tiene el mismo valor en cualquier sistema de referencia que se mueva sin aceleración. Esto se logra mediante las transformaciones de Lorenz, fórmulas matemáticas que llevan el nombre del físico holandés Hendrik Lorenz y que describen cómo cambian la posición y el tiempo cuando se comparan diferentes sistemas de referencia. Las predicciones más destacadas son muy extrañas desde un punto de vista newtoniano. Nada puede viajar más rápido que la luz; la longitud de un objeto se reduce conforme su velocidad aumenta y se aproxima a la de la luz, hasta hacerse arbitrariamente pequeña cerca de este valor; mientras esto ocurre, el tiempo subjetivo se ralentiza hasta casi detenerse y la masa aumenta sin límite. *Grosso modo*, a la velocidad de la luz, la longitud de un objeto en la dirección del desplazamiento se reduce a cero, se detiene el tiempo y la masa se hace infinita.

La relatividad general mantiene estos elementos, pero además incorpora la gravedad. No obstante, en lugar de considerarla como una fuerza, tal y como la modeló Newton, la entiende como un efecto de la curvatura del espacio-tiempo, una construcción matemática de cuatro dimensiones que combina las tres del espacio con una temporal. Cerca de cualquier masa, como una estrella, el espacio-tiempo se *curva* y forma una especie de depresión, pero en cuatro dimensiones. Un rayo de luz o una partícula que pase cerca se desvía de una trayectoria recta conforme sigue la curvatura. Esto produce la ilusión de una fuerza de atracción entre la estrella y la partícula.

Ambas teorías se han verificado hasta la saciedad mediante experimentos de gran sensibilidad. A pesar de sus características más bien extrañas, aportan el mejor modelo de la realidad que la física ha descubierto hasta la fecha. Las matemáticas del GPS deben tener en cuenta los efectos relativistas, resultantes tanto de la velocidad del satélite como de la gravedad de la Tierra. De otro modo, el sistema sería inútil. De hecho, su éxito con estas correcciones es una prueba palpable de la validez tanto de la relatividad especial como de la general.

La mayoría de los usuarios de GPS se encuentran, o bien en ubicaciones fijas en la superficie de la Tierra, o bien se mueven con lentitud (digamos que a una velocidad no superior a la de un coche rápido). Por este motivo, los diseñadores decidieron emitir información acerca de las órbitas del satélite mediante el empleo de un sistema de referencia unido de manera rígida a una Tierra que gira y asumir que su velocidad de rotación es constante. La forma de nuestro planeta, denominada un geoide, se aproxima a un elipsoide de revolución ligeramente aplastado.

Cuando se está en el coche y los satélites dan vueltas sobre el vehículo, es evidente que se mueven con relación a este. La relatividad especial predice que se observará que los relojes en órbita se atrasan respecto a uno de referencia en tierra. De hecho, los que están en movimiento pierden unos siete microsegundos al día, debido a la dilatación relativista del tiempo. Además, la fuerza de gravedad aparente a esa altura es menor que al nivel del suelo. Por lo que se refiere a la relatividad general, el espacio-tiempo cerca de los satélites es más plano (menos curvo) que el que hay en la vecindad del coche. Este efecto hace que los relojes en órbita se *adelanten* respecto a uno en tierra. La relatividad general predice que el valor de esta diferencia será de unos 45 microsegundos diarios. Al combinar estos efectos opuestos, un reloj en un satélite se adelantará respecto a uno en tierra cerca de $45 - 7 = 38$ microsegundos al día. Un error así puede percibirse después de 2 minutos y la ubicación se desplazaría respecto a la correcta cerca de 10 kilómetros diarios. Después de un día, el navegador nos situaría en la ciudad equivocada, tras una semana en la provincia errónea y al pasar un mes, en el país incorrecto.

En un primer momento, los ingenieros y científicos que trabajaban en el proyecto del GPS no tenían la certeza de que la relatividad fuese a importar en realidad. Para los estándares humanos, los satélites van muy rápido, pero a paso de tortuga si se les compara con la velocidad de la luz. La gravedad terrestre es minúscula al nivel cósmico. Pero hicieron lo que pudieron para estimar las magnitudes de estos efectos. En 1977, cuando se colocó en órbita el primer prototipo de reloj atómico de cesio, todavía no estaban seguros de la importancia que iban a tener estos efectos o de si iban a ser positivos o negativos. Algunos ni siquiera creían que fuesen necesarias correcciones relativistas en absoluto. De modo que los ingenieros incorporaron un circuito en el reloj que al recibir desde tierra una señal dada podía cambiar su frecuencia para cancelar los efectos relativistas previstos, en caso de que fuese necesario. Durante las primeras tres semanas mantuvieron el circuito apagado y midieron la frecuencia del reloj. Observaron que se adelantaba

442,5 partes por billón, en comparación con otro en tierra. La predicción de la relatividad general era de 446,5 partes por billón. Casi perfecto.

*

Hay muchos otros usos del GPS aparte del evidente de ubicar posiciones (coches, vehículos comerciales, montañeros) y de las aplicaciones militares que llevaron a su creación en primer lugar. Mencionaré unas pocas.

No hace falta saber dónde se está cuando se usa una aplicación para llamar a la grúa porque el coche se ha roto, el GPS lo hace en nuestro lugar. También se emplea para impedir robos de vehículos, para confeccionar mapas y medir terrenos, para vigilar mascotas y cuidar de parientes mayores o para garantizar la seguridad de las obras de arte. Algunos de sus usos principales incluyen la navegación marítima y aérea y el seguimiento de flotas para empresas de transporte. Ahora que la mayor parte de los teléfonos móviles cuentan con receptores de GPS, son capaces de etiquetar las fotografías con la ubicación en la que se han tomado, avisar de dónde se encuentra un dispositivo robado o perdido, y pedir un taxi. Es posible emplear el GPS en conjunto con servicios cartográficos en línea, como Google Maps, de modo que el mapa muestre al usuario de manera automática allí donde se encuentre. Los granjeros pueden manejar tractores sin conductor, los banqueros vigilar transacciones financieras y los viajeros encontrar su equipaje. Los científicos tienen la capacidad de seguir los movimientos de especies amenazadas y de controlar desastres medioambientales, tales como vertidos de crudo.

¿Cómo nos las apañábamos antes sin GPS? Es sorprendente la rapidez con la que unos cuantos bits de magia matemática, que hacen posible una tecnología transformadora (y muy cara), pueden cambiar nuestras vidas.

12
El deshierro del Ártico

> La capa de hielo de Groenlandia se derrite mucho más rápido de lo que se pensaba antes, amenaza con inundaciones a cientos de millones de personas y hace que el impacto irreversible de la emergencia climática sea una realidad mucho más patente. Se pierde hielo en Groenlandia siete veces más rápido que en la década de 1990 y la magnitud y velocidad del deshielo es mucho mayor de lo que se había predicho.
>
> *The Guardian*, diciembre de 2019

No, no es deshielo es *deshierro*. No es un error de imprenta, solo un juego de palabras sin gracia.

El planeta se calienta. Es peligroso y es culpa nuestra. Se sabe porque miles de científicos expertos en el clima, que ejecutan cientos de modelos matemáticos, lo predijeron hace décadas, y observaciones realizadas por meteorólogos igual de competentes han confirmado la mayoría de las conclusiones más importantes. Podría pasarme el resto del libro sacándole los colores a los marrulleros que tergiversan los hechos para convencer al público de que no hay nada de qué preocuparse y comparando sus sandeces con la creciente evidencia de la realidad del cambio climático producido por los seres humanos, al mismo tiempo que explico los muchos pormenores que todavía no son definitivos. Pero como dice Arlo Guthrie hacia la mitad de *Alice's restaurant*, no es eso de lo que he venido a hablar. Hay muchas personas que ya lo hacen mucho mejor de lo que podría hacerlo yo y otras tantas que intentan detenerlas de manera desesperada, no vaya a ser que algunos personajes extremadamente ricos tengan que dejar de destruir el planeta.

El cambio climático es inherentemente estadístico, de modo que se le puede quitar importancia a cualquier suceso concreto y explicarlo como uno de esos fenómenos extraños que ocurren de vez en cuando. Si hay una moneda trucada para que salga cara tres veces de cada cuatro, podrá salir cara o cruz en cualquier tirada individual, como si no hubiese truco. Un solo resultado no basta para descubrir la diferencia. Incluso con una moneda

normal puede salir en ocasiones una serie de tres o cuatro caras consecutivas. No obstante, si en 100 tiradas salen 80 caras y 20 cruces, es bastante evidente que la moneda está trucada.

El clima es parecido. No es lo mismo que el tiempo, que cambia cada hora y cada día. El clima es una media móvil a lo largo de treinta años. A su vez, el clima global hace un promedio de esta para todo el planeta. Se necesitan ingentes cambios a muy largo plazo en una escala planetaria para alterarlo. Sin embargo, se dispone de registros de temperatura de gran calidad y de todo el mundo desde hace aproximadamente 170 años, y 17 de los 18 periodos más cálidos han ocurrido a partir de la década del 2000. No es casualidad.

La naturaleza estadística del clima hace que los negacionistas puedan enfangar las aguas con facilidad. Los climatólogos, que no tienen la capacidad de adelantar el planeta a velocidad rápida, han tenido que depender de modelos matemáticos para echar un vistazo al futuro, estimar la velocidad a la que cambia el clima, calcular los efectos que estas alteraciones podrían tener y examinar lo que puede hacer la humanidad para solventarlo, si consigue ponerse de acuerdo. Los primeros modelos eran bastante rudimentarios, lo que los hacía susceptibles a que cualquiera a quien no le gustasen las predicciones pudiese encontrar objeciones. Sin embargo, echando la vista atrás, resulta que incluso esos modelos predijeron el ritmo de aumento de la temperatura global, junto con muchas otras cosas, con bastante precisión. Se ha mejorado mucho a lo largo de los años y las temperaturas previstas para la actualidad se han correspondido con las reales de manera considerable a lo largo del último medio siglo. La cantidad de hielo que se vaya a derretir como consecuencia está menos clara y parece que se ha subestimado. Los mecanismos que afectan al proceso no se comprenden tan bien y se ha presionado a los científicos durante décadas para que no pareciesen alarmistas.

Hasta el momento, me he enfocado en la manera en que las matemáticas, que funcionan sin alharacas entre bambalinas, afectan a las vidas cotidianas de las personas. He omitido de forma deliberada un montón de aplicaciones importantes en ciencia, sobre todo en la teórica. Pero el cambio climático sí tiene un efecto sobre nuestras vidas diarias. Y si no, que se lo pregunten a los australianos, que tuvieron que enfrentarse a incendios forestales sin precedentes a principios de 2020. Ahí están las olas de calor de récord en todo el globo y las inundaciones, que antes eran seculares y ahora se producen cada cinco o diez años. Ahí están, aunque sea paradójico, las olas ocasionales de tiempo frío en extremo. En cierto modo, parece contradictorio que el

calentamiento global pueda hacer que algunos lugares se vuelvan mucho *más fríos* de lo normal, pero la explicación es sencilla. Este fenómeno se refiere a la cantidad promedio de energía calorífica que entra en la atmósfera, en los océanos y en la tierra. Nadie dijo que todos los lugares se iban a calentar de manera uniforme.

Conforme aumenta la energía calorífica total del planeta, las fluctuaciones en torno al valor promedio se hacen mayores y estas pueden ser más frías de lo normal al igual que más calientes. El caso es que prevalece el aumento de calor en general. Una ola de frío repentina en una ubicación no demuestra que el calentamiento global sea un fraude. Del mismo modo, si el tiempo en una ciudad es diez grados más frío de lo normal, pero es un grado más caliente en otras once localidades, la temperatura global promedio ha *subido*. Si esa misma ciudad está hoy diez grados por debajo de lo normal, pero luego está uno por encima durante once días sueltos, la temperatura media global ha subido de nuevo si todo lo demás permanece igual. De hecho, también habrá aumentado en la ciudad.

El problema es que es fácil darse cuenta de la ola de frío repentina, pero es posible que los efectos que la compensan sean demasiado pequeños como para cobrar consciencia de ellos, o pueden darse de manera muy dispersa u ocurrir en otro lugar. Las muy inusuales olas de frío que se han producido en Europa y en América del Norte en los últimos años han ocurrido porque la corriente en chorro atmosférica empujó aire gélido del Ártico más al sur de lo habitual. De modo que esta masa atmosférica que en condiciones normales habría circulado en torno al casquete polar ártico acabó sobre el océano, Groenlandia, el norte de Canadá y Rusia. ¿Por qué se desplazó esta masa de aire frío hacia el sur? Porque el que había sobre la zona polar era mucho *más caliente* de lo normal y lo desalojó. En conjunto, toda la región afectada se calentó en promedio.

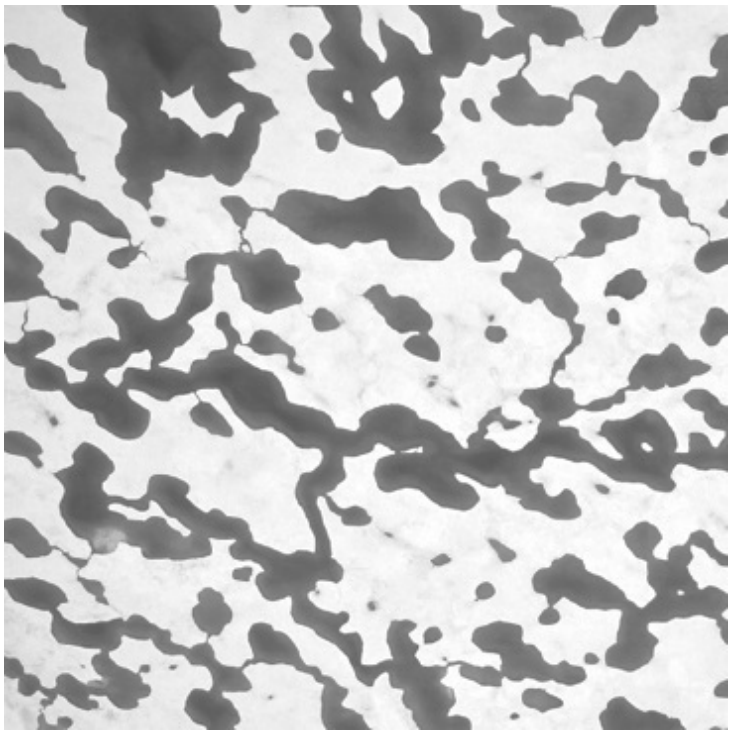
En los modelos del clima hay bastantes matemáticas como para llenar un libro entero, pero no es eso de lo que he venido a hablar. Como Arlo Guthrie, solo estoy preparando el escenario para aquello de lo que sí quiero hablar.

*

El hielo se derrite por todo el planeta. Su cantidad ha aumentado en unos cuantos lugares poco representativos, pero, en todos los demás, merma con rapidez. Los glaciares retroceden y los casquetes polares se reducen en ambos polos. Estos efectos amenazan las fuentes de agua potable de unos 2000

millones de personas y la subida del nivel del mar resultante inundará los hogares de otros 500 millones más, a no ser que evitemos que suceda. De modo que la física y las matemáticas del hielo fundente han cobrado de manera repentina un interés decisivo, a nivel personal, para todo el mundo.

Los científicos saben mucho sobre este asunto. Junto con el agua en ebullición y su conversión en vapor, es un ejemplo clásico de transición de fase, un cambio en el estado de la materia. El agua puede presentarse en una serie de estados y ser sólida, líquida o gaseosa. Aquel en el que se encuentre depende sobre todo de la temperatura y de la presión. Para valores habituales de esta última variable, si el agua está lo bastante fría, es un sólido: hielo. Conforme se calienta y supera el punto de fusión, se vuelve líquida: agua. Si su temperatura aumenta más, hasta el punto de ebullición, se transforma en gas: vapor. En la actualidad, la ciencia conoce 18 fases diferentes del hielo. La última, el «hielo cuadrado», se descubrió en 2014. Tres de estas fases se dan a presión atmosférica normal, el resto requieren valores de esta variable mucho más altos.



Los estanques de fusión oscuros resaltan sobre el hielo ártico blanco. ¿Por qué forman unos patrones tan intrincados?

La mayor parte de lo que se sabe acerca del hielo es el resultado de experimentos de laboratorio en cantidades relativamente pequeñas. Lo que hace falta averiguar con urgencia en la actualidad sobre el hielo fundente se refiere a volúmenes inmensos en el entorno natural. Hay dos maneras

interrelacionadas de descubrirlo: observar y medir lo que ocurre y desarrollar modelos teóricos de la física subyacente. La clave para una comprensión real es unir las dos.

Una de las señales de que el hielo polar, sobre todo el marítimo, se derrite es la formación de estanques de fusión. La superficie helada empieza a fundirse y unos pequeños charcos negruzcos ensucian su blancura prístina o a menudo su no tan inmaculado color gris, debido a depósitos de polvo. Los charcos son agua líquida, que es oscura a diferencia del hielo, de modo que absorbe la luz solar en lugar de reflejarla. La radiación infrarroja en particular los calienta con más rapidez de lo que lo haría si todavía estuviesen helados, así que crecen. Cuando llegan a ser lo bastante grandes, se unen para formar otros mayores, tanto como para calificarse de estanques. Estos son los estanques de fusión y dan lugar a formas intrincadas complejas: puntos unidos por ramas delgadas, que se separan y se extienden como tapices de algún extraño hongo.

La física del crecimiento de los estanques de fusión es una característica crucial del comportamiento del hielo cuando se calienta. Y es exactamente lo que ocurre, sobre todo con el que flota en el Océano Ártico. Lo que vaya a suceder con el hielo marino al calentarse el planeta es una parte fundamental del problema que supone comprender el impacto del cambio climático. Por ello, es natural que los matemáticos investiguen modelos de hielo fundente con la intención de desentrañar algunos de sus secretos. Así lo hacen. No es sorprendente. Lo que sí resulta inesperado es que uno de los modelos que se estudian en la actualidad no tenga nada que ver con el hielo. Trata del magnetismo y se remonta a 1920. Los materiales magnéticos experimentan sus propios tipos de cambios de fase y, en particular, pierden su magnetismo inherente si se calientan demasiado.

Hace tiempo que este modelo en concreto es un ejemplo paradigmático de transición de fase. Fue propuesto por el físico alemán Wilhelm Lenz, así que por supuesto todo el mundo lo llama modelo de Ising. Los matemáticos y los físicos bautizan siempre a las cosas con el nombre de la persona asociada a ellas de manera más estrecha en sus mentes, quien a menudo no es el verdadero inventor. Lenz tenía un estudiante, Ernst Ising, y le propuso un problema para su doctorado: resolver el modelo y demostrar que tiene un cambio de fase magnética. Ising lo resolvió y demostró que no lo tenía. A pesar de todo, sus investigaciones pusieron en marcha toda una rama de la física matemática y ampliaron mucho nuestra comprensión de los imanes.

Y ahora, vamos a por el hielo.

*

Los imanes son tan habituales en la actualidad que casi nadie se pregunta nunca cómo funcionan. Se utilizan para pegar cerdos de plástico en la puerta de la nevera (bueno, en casa lo hacemos), para cerrar las fundas de los móviles y (unos bastante grandes) para detectar el famoso bosón de Higgs, que dota a las partículas subatómicas de masa. Sus usos cotidianos incluyen los discos duros de los ordenadores y los motores eléctricos (del tipo de los que suben o bajan las ventanillas de los coches de manera automática o de los que producen gigawatios de energía eléctrica). A pesar de su ubicuidad, son muy misteriosos. Se atraen o repelen entre sí mediante algún tipo de campo de fuerza invisible. Los imanes de barra más sencillos y comunes tienen dos polos, cada uno cerca de un extremo, denominados norte y sur. Norte y sur se atraen entre sí, mientras que dos polos norte se repelen y lo mismo pasa con los dos polos sur. Así, si se empujan polos iguales de dos imanes pequeños pero potentes para acercarlos, puede notarse cómo se resisten. Por el contrario, si se intentan separar polos opuestos, se nota cómo intentan unirse. Tienen efecto uno sobre otro incluso cuando no se tocan: «acción a distancia». Con los imanes es posible hacer levitar objetos, incluso grandes, como trenes. De un modo misterioso, este campo de fuerza es invisible. No se ve nada.

La humanidad conoce los imanes desde hace al menos 2500 años. Están presentes de forma natural en la magnetita, un mineral que es un óxido de hierro. Un fragmento pequeño de magnetita, conocido como piedra de imán, es capaz de atraer los objetos de hierro y puede convertirse en una brújula si se cuelga de una cuerda o se hace flotar en agua sobre un trozo de madera. Estas piedras se empleaban de manera rutinaria para la navegación desde cerca del siglo XII de nuestra era. Materiales como este, que pueden estar dotados de un campo magnético permanente, se denominan ferromagnéticos. La mayoría son aleaciones de hierro, níquel y/o cobalto. Algunos de ellos mantienen sus propiedades magnéticas de manera casi indefinida, mientras que otros son susceptibles de imantarse temporalmente, pero pierden su fuerza poco después.

Los científicos empezaron a prestar atención en serio a los imanes en 1820, cuando el físico danés Hans Christian Ørsted descubrió una conexión entre el magnetismo y la electricidad. En concreto, que una corriente eléctrica puede crear un campo magnético. William Sturgeon, un científico británico, construyó un electroimán en 1824. La historia del electromagnetismo es

demasiado extensa para describirla en detalle, pero un avance decisivo vino de la mano de los experimentos de Michael Faraday. Estos llevaron a James Clerk Maxwell a formular las ecuaciones matemáticas de los campos eléctrico y magnético y la relación entre ellos. Estas expresiones indican, de manera precisa, que la electricidad en movimiento produce magnetismo y que el magnetismo en movimiento produce electricidad. Entre ambos, generan ondas electromagnéticas que se desplazan a la velocidad de la luz. De hecho, la luz es un fenómeno de este tipo, igual que las ondas de radio, los rayos X y las microondas.

Una característica intrigante de los materiales ferromagnéticos es su respuesta cuando se calientan. Hay una temperatura crítica, denominada temperatura de Curie. Si se calienta uno de estos materiales por encima de este punto, su campo magnético desaparece. No solo eso, sino que el cambio es abrupto. Conforme se aproxima a la temperatura de Curie, el campo magnético empieza a decaer de manera dramática y lo hace más rápido cuanto más se acerca al valor crítico. Los físicos llaman a este tipo de comportamiento una transición de fase de segundo orden. La pregunta crucial es por qué sucede.

Una pista importante vino del descubrimiento del electrón, una partícula subatómica que tiene una carga eléctrica muy pequeña. Una corriente eléctrica es una multitud de estas cargas en movimiento. Los átomos tienen un núcleo, compuesto de protones y neutrones, rodeado por una nube de electrones. Su número y disposición determina las propiedades químicas del átomo. Estas partículas también tienen una propiedad llamada espín. Es cuántica y aunque ellas no rotan en realidad, tiene mucho en común con el momento angular, un nombre matemático elegante para una característica de los cuerpos que rotan en la física clásica. El momento angular indica la potencia de la rotación y la dirección en la que se produce, el eje en torno al cual gira el cuerpo.

Los físicos descubrieron de manera experimental que el espín del electrón lo dota de un campo magnético. Siendo la mecánica cuántica lo que es (en concreto, una cosa muy rara), el espín del electrón medido sobre *cualquier* eje concreto es siempre «arriba» o «abajo». Estos estados se corresponden en general con un imán minúsculo que tiene un polo norte en la parte superior y uno sur en la inferior, o al revés. Antes de medir el espín, este puede ser cualquier combinación de arriba y abajo de manera simultánea, lo que se podría entender en resumidas cuentas como una rotación en torno a un eje diferente, pero cuando se observa sobre el eje seleccionado, resulta que

siempre es arriba; o abajo. Uno de los dos. Esa es la parte más extraña y es diferente por completo de la rotación en la física clásica.

La conexión entre el espín de un electrón y su campo magnético explica en buena medida no solo por qué los imanes pierden su magnetismo si se calientan demasiado, sino cómo sucede. Antes de que un material ferromagnético esté imantado, los espines de sus electrones están alineados de manera aleatoria, de modo que sus minúsculos campos magnéticos tienden a cancelarse. Cuando se imanta el material, bien mediante un electroimán o por acercarse a otro imán permanente, se alinean los espines de sus electrones. Entonces se refuerzan unos a otros y crean un campo magnético detectable a escala macroscópica. Si se deja a sus propios medios, esta disposición se conserva y ya se tiene un imán permanente.

No obstante, si se calienta el material, la energía térmica empieza a agitar los electrones y da la vuelta a algunos de sus espines. Los campos magnéticos que se ejercen en direcciones diferentes se debilitan entre sí, de modo que desciende la fuerza global del campo magnético. Esto explica la pérdida del magnetismo de manera cualitativa, pero no dice nada acerca de por qué hay un cambio de fase tan brusco ni por qué se produce siempre a una temperatura concreta.

En estas llegó Lenz. Se le ocurrió un modelo matemático sencillo: un conjunto de electrones, cada uno de los cuales afecta a sus vecinos según sus espines relativos. En el modelo, cada partícula está ubicada en un punto fijo en el espacio, que suele formar parte de una red regular, como las casillas de un gran tablero de ajedrez. Cada electrón modelo puede estar en uno de dos estados: +1 (espín arriba) o −1 (espín abajo). La red está cubierta en todo momento por un patrón de ±1. En la analogía del tablero, cada casilla es, o bien negra (espín arriba), o bien blanca (espín abajo). Es factible cualquier distribución de cuadrados blancos y negros, por lo menos en principio, porque los estados cuánticos son aleatorios hasta cierto punto, si bien unos patrones son más probables que otros.

Los estudiantes de doctorado son muy útiles para hacer cálculos o experimentos que su director de tesis prefiere evitar, de modo que Lenz le dijo a Ising que resolviera el modelo. Aquí, el significado de «resolver» es muy sutil. No se trata de la dinámica del cambio de espín ni de las secuencias de estados individuales, sino de calcular la distribución de probabilidad de todos los patrones posibles y la dependencia de esta con la temperatura y con cualquier campo magnético externo. Una distribución de probabilidad es un

dispositivo matemático (a menudo una fórmula) que en este caso indica lo probable que es cualquier patrón dado.

El director de tesis ha hablado y si el estudiante quiere conseguir su doctorado, hace lo que le dicen. O, al menos, lo hace lo mejor que puede, porque a veces los problemas que se proponen a los estudiantes son demasiado difíciles. Después de todo, el motivo por el que se les pide que resuelvan el cometido es porque el director de tesis *no conoce la respuesta* y a menudo no tiene ni idea, más allá de una vaga intuición, de lo difícil que va a resultar.

De modo que Ising se puso manos a la obra para resolver el modelo de Lenz.

*

Hay algunos trucos habituales que los directores de tesis conocen y que pueden sugerir a sus estudiantes. Los alumnos brillantes de verdad los descubren por sí mismos, junto con otras ideas que nunca se le han ocurrido al director. Uno de ellos es bastante curioso, pero cierto en general: si se quiere trabajar con un número muy grande, todo es más sencillo si se toma como *infinito*. Por ejemplo, si se quiere entender el modelo de Ising para un tablero de ajedrez grande pero finito, que represente un trozo de material ferromagnético de tamaño realista, es más conveniente, desde un punto de vista matemático, trabajar con un tablero de ajedrez de tamaño infinito. El motivo es que el primero tiene bordes y que estos tienden a complicar los cálculos porque las casillas de los extremos son diferentes de las de la mitad. Esto deshace la simetría del conjunto de electrones y esta es una característica que tiende a hacer que los cálculos sean más sencillos. Un tablero de ajedrez infinito no tiene bordes.

La imagen del tablero corresponde a lo que los matemáticos y los físicos denominan una red bidimensional. La palabra «red» quiere decir que las unidades básicas, las casillas, están dispuestas de un modo muy regular (en este caso, en filas y columnas, todas alineadas a la perfección con sus vecinas). Las redes matemáticas pueden tener cualquier cantidad de dimensiones, mientras que las físicas suelen tener una, dos o tres. El caso más relevante para la física es la red tridimensional: un conjunto infinito de cubos iguales, superpuestos unos a otros de manera ordenada, como cajas idénticas en un almacén. En este caso, los electrones ocupan una región del espacio, del mismo modo que los átomos en un cristal con simetría cúbica, como la sal.

Los matemáticos y los físicos matemáticos prefieren con mucho empezar con un modelo más sencillo, aunque menos realista: una red de una dimensión en la que las ubicaciones de los electrones están dispuestas en una fila en intervalos regulares, como los enteros a lo largo de la recta de los números. No tiene mucho sentido físico, pero sirve para desarrollar ideas en la configuración relevante más sencilla. Conforme aumenta la cantidad de dimensiones, también lo hacen las complicaciones matemáticas. Por ejemplo, hay un tipo de red cristalina en una recta, 17 en el plano y la friolera de 230 en el espacio tridimensional. De modo que Lenz propuso a su estudiante el problema de encontrar cómo se comportan modelos como este y tuvo la sensatez de decirle que se concentrara en la red de una dimensión. El alumno hizo suficientes progresos como para que todos los casos de este tipo se denominen modelos de Ising hoy en día.

Aunque el modelo de Ising se refiere al magnetismo, su estructura y la manera de concebirlo pertenecen a la termodinámica. Esta rama tiene su origen en la física clásica, en la que trata de magnitudes como temperatura y presión en gases. En torno a 1905, cuando los físicos se convencieron por fin de que los átomos existían y se combinaban para formar moléculas, se dieron cuenta de que variables como la temperatura y la presión son promedios estadísticos. Son cantidades «macroscópicas» que pueden medirse con facilidad, debidas a fenómenos que ocurren a una escala «microscópica» mucho más pequeña. Por cierto, no está de más decir que en realidad no son visibles en un microscopio, incluso aunque en la actualidad hay aparatos capaces de captar átomos individuales, si bien solo funcionan cuando estos últimos no se mueven. En un gas, hay volando por ahí una gran cantidad de moléculas, que chocan entre sí y rebotan de manera ocasional. Los rebotes hacen que su movimiento sea aleatorio.

El calor es una forma de energía causada por el movimiento de las moléculas. Cuanto más rápido se mueve, más se calienta el gas y sube la temperatura, que es diferente del calor y es una medida de la calidad del mismo y no de la cantidad. Hay relaciones matemáticas entre las posiciones y velocidades de las moléculas y los promedios termodinámicos. Esta relación es el objeto de una rama denominada mecánica estadística, que pretende calcular variables macroscópicas a partir de las microscópicas, con especial interés en los cambios de fase. Por ejemplo, ¿qué cambia en el comportamiento de las moléculas de agua cuando se derrite el hielo? Y ¿qué tiene que ver la temperatura del material con ello?

*

El problema de Ising era parecido, pero en lugar de moléculas de H_2O y hielo que se vuelve agua al calentarse, analizaba espines de electrones e imanes que pierden su magnetismo cuando se calientan. Lenz había dispuesto su modelo (al que ahora se llama modelo de Ising) para que fuese lo más sencillo posible. Como es habitual en matemáticas, el planteamiento puede ser simple, pero solucionarlo no lo es.

Cabe recordar que «resolver» este problema implica calcular cómo varían con la temperatura las características estadísticas del conjunto de imanes minúsculos. Esto se reduce a encontrar la energía total de la configuración, la cual depende del patrón del magnetismo (el número y disposición de los espines arriba y abajo, de casillas negras y blancas en el tablero). Los sistemas físicos prefieren adoptar estados con la menor energía posible. Este es el motivo por el que, por ejemplo, cae la proverbial manzana de Newton: su energía potencial gravitatoria se reduce conforme cae hacia el suelo. La genialidad de Newton consistió en darse cuenta de que el mismo razonamiento es válido para la Luna, que cae de manera perpetua pero que no alcanza nunca el suelo porque también se desplaza hacia los lados. Hizo los cálculos adecuados y demostró que la misma fuerza gravitatoria explica ambos movimientos de forma cuantitativa.

Sea como sea, todos los imanes diminutos (los electrones con sus direcciones de espín) intentan reducir su energía global todo lo posible. Pero la manera en que lo hacen y el estado que alcanzan dependen de la temperatura del material. A nivel microscópico, el calor es una forma de energía que hace que las moléculas y los electrones se desplacen y agiten de manera aleatoria. Cuanto más caliente está el material, más rápido lo hacen. En un imán, el patrón exacto de espín cambia sin cesar debido a esta agitación fortuita y por eso «resolver» el modelo conduce a una distribución de probabilidad estadística, no a una configuración concreta de espines. No obstante, los patrones más probables parecen todos bastante similares, de modo que cabe preguntarse qué aspecto tiene uno que sea representativo a cualquier temperatura dada.

La parte crucial del modelo de Ising es una regla matemática para describir la interacción de los electrones que especifica la energía de cualquier patrón. Para simplificar, se asume que cada elemento del sistema interactúa solo con sus vecinos inmediatos. En una interacción ferromagnética, la contribución a la energía es negativa cuando los electrones contiguos tienen el

mismo espín. En materiales antiferromagnéticos, es positiva cuando los electrones contiguos tienen el mismo espín. También hay una contribución adicional debida al efecto de un campo magnético externo sobre cada elemento del sistema. En modelos simplificados, todas las fuerzas de interacción entre los electrones contiguos son de la misma dimensión y el campo magnético externo se hace cero.

La clave de las matemáticas en este caso es comprender cómo cambia la energía de un patrón dado cuando el color de una casilla pasa de negro a blanco o viceversa. Es decir, un único electrón en una ubicación arbitraria salta de +1 (negro) a −1 (blanco). Algunos cambios aumentan la energía total y otros la disminuyen. Estos últimos son los más probables. No obstante, no es posible descartar por completo que haya otros que la incrementen, debido al movimiento termal aleatorio. De forma intuitiva, es de esperar que la distribución converja hacia algún estado que tenga la energía más baja. En un material ferromagnético, esto debería hacer que todos los electrones tuvieran el mismo espín, pero no es eso lo que ocurre exactamente en la práctica, porque se tardaría demasiado. En su lugar, a temperaturas moderadas, hay distintas regiones en las que las orientaciones de los componentes están alineadas de manera casi perfecta y se crea un mosaico blanco y negro. A temperaturas más elevadas, los desplazamientos aleatorios superan a las interacciones entre espines contiguos y las regiones se hacen tan pequeñas que no hay relación entre la situación de un electrón y la de sus vecinos. De modo que el patrón es caótico y parece gris, excepto en una escala muy fina de blanco y negro. A bajas temperaturas, las regiones se hacen más grandes, lo que lleva a una configuración más ordenada. Estas distribuciones nunca se quedan fijas del todo porque siempre hay cambios aleatorios. Pero, para una temperatura dada, las características *estadísticas* del patrón permanecen constantes.

Lo que tiene más interés para los físicos es la transición de tener regiones de color separadas, un estado ordenado, a un caos gris y aleatorio. Se trata de un cambio de fase. Los experimentos sobre este fenómeno en materiales ferromagnéticos, que pasan de estar imantados a no estarlo, muestran que por debajo de la temperatura de Curie el patrón magnético presenta regiones. El tamaño de estas difiere de unas a otras, pero se agrupa en torno a un valor típico concreto o «escala de longitud», que disminuye conforme se calienta el material. Por encima del punto de Curie no hay regiones: los dos valores de espín están mezclados. Lo que ocurre *a* la temperatura de Curie es lo que entusiasma a los físicos. En este caso hay regiones de varios tamaños, pero no

hay una escala de longitud dominante. La distribución por zonas forma un fractal, un patrón con estructura detallada en todas las escalas. Una porción aumentada de una parte tiene las mismas características estadísticas que todo el conjunto, de modo que no es posible deducir el tamaño de la región a partir del patrón. Ya no hay una longitud de escala bien definida. No obstante, puede darse una medida numérica del ritmo al que cambia la distribución durante la transición de fase, que se denomina exponente crítico. Los experimentos son capaces de medir esta magnitud con mucha precisión, de modo que proporciona una prueba de gran sensibilidad de las formulaciones teóricas. Un objetivo importante de los investigadores es derivar modelos que proporcionen el exponente crítico correcto.

Las simulaciones por ordenador no logran «resolver» el modelo de Ising con exactitud (no consiguen proporcionar una fórmula de las características estadísticas con una demostración matemática rigurosa que sea correcta). Los sistemas informáticos de álgebra modernos podrían ayudar a los investigadores a encontrar la fórmula, si es que hay una, pero todavía necesitaría demostración. Las simulaciones por ordenador más tradicionales pueden proporcionar una evidencia sólida a favor o en contra de la correspondencia del modelo con la realidad. Pero el Santo Grial para los físicos matemáticos (y para los matemáticos dados a la física, pues es esta última la que motiva el dilema principal, aunque la naturaleza de este sea matemática por completo) es obtener resultados *exactos* acerca de las propiedades estadísticas de los patrones de espín en el modelo de Ising, sobre todo en lo que se refiere al modo en que cambian esas propiedades cuando la temperatura supera el punto de Curie. En particular, los investigadores buscan una demostración de que se produce un cambio de fase en el sistema y pretenden caracterizarlo mediante el exponente crítico y las características fractales de los patrones más probables en el punto de transición.

*

A partir de aquí la cosa se pone más técnica, pero intentaré aportar las ideas principales sin preocuparme de los detalles. Habrá que suspender el escepticismo y dejarse llevar.

El dispositivo matemático más importante en termodinámica es la «función de partición». Esta se obtiene al sumar, para *todos* los estados del sistema, una expresión algebraica particular que depende de la configuración y de la temperatura. Para ser precisos, esta expresión se obtiene para cualquier

estado dado al tomar su energía, convertirla en negativa y dividirla por la temperatura. Se toma la exponencial de esto y se suma cada una de estas expresiones para todos los patrones posibles[65]. La noción física aquí es que los estados con menos energía contribuyen más a la suma, de modo que la función de partición está dominada por (tiene un pico en) el tipo de configuración más probable.

Todas las variables termodinámicas habituales pueden deducirse a partir de la función de partición mediante operaciones adecuadas, de modo que la mejor manera de «resolver» un modelo termodinámico es obtener su función de partición. Ising encontró su solución al derivar una fórmula para la energía libre[66] y deducir otra para la imantación[67]. La expresión que obtuvo es impresionante, pero debió de ser una gran desilusión para él, porque después de todos esos cálculos difíciles, viene a decir que el material no tiene un campo magnético propio en ausencia de uno externo. Lo que es peor, esto es cierto para cualquier temperatura y en todos los casos. De modo que el modelo no predice el cambio de fase ni las propiedades magnéticas espontáneas del material supuestamente ferromagnético.

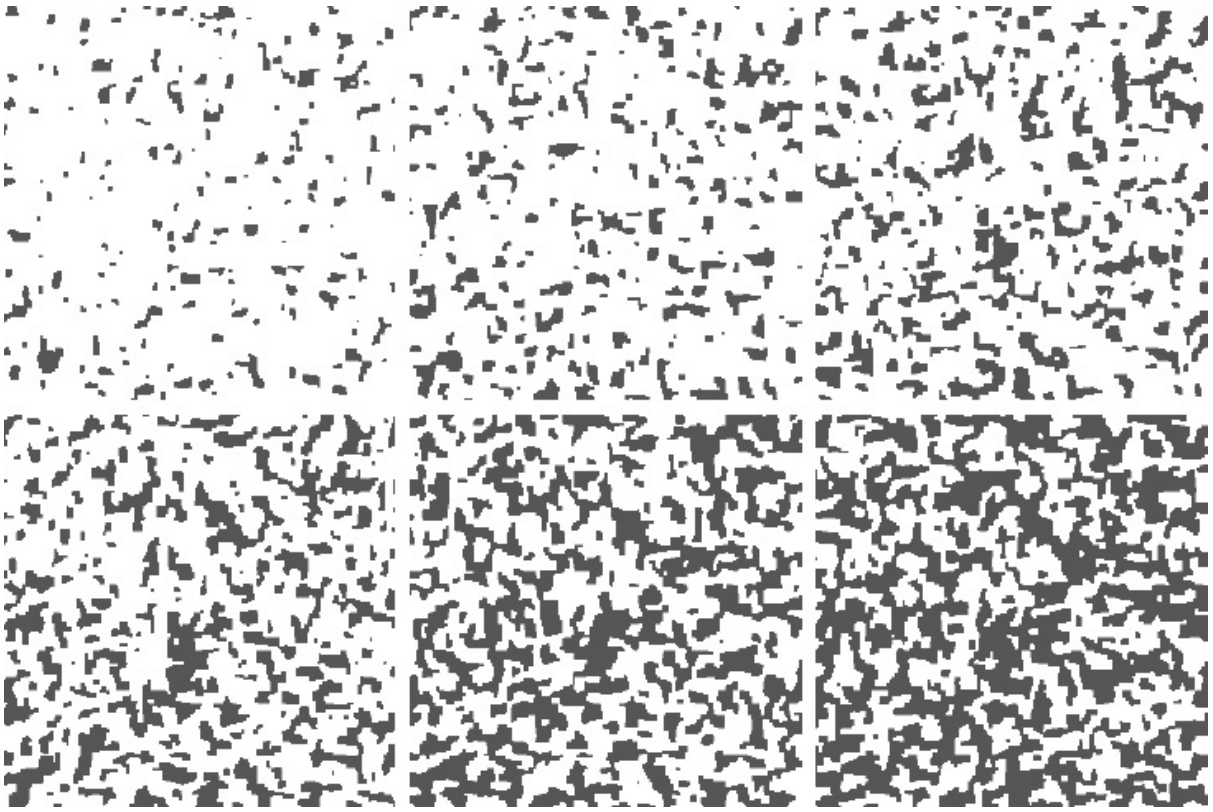
Enseguida se sospechó que la principal razón para este resultado negativo era la simplicidad del modelo. En efecto, se alzó un dedo acusador contra la cantidad de dimensiones de la red. En esencia, una sola de ellas no era bastante para conducir a resultados realistas. Resultaba evidente que el siguiente paso debía ser repetir el proceso para una red bidimensional, pero esto era muy difícil. Los métodos de Ising eran inadecuados. Solo en 1944, después de varios avances que hicieron que los cálculos fuesen más sistemáticos y sencillos, pudo Lars Onsager resolver el problema en dos dimensiones. Fue un *tour de force* matemático, con una respuesta complicada pero explícita. Incluso así, tenía que asumir que no existía campo magnético externo.

La fórmula muestra que ahora *hay* un cambio de fase y lleva a un campo magnético interno distinto de cero por debajo de una temperatura crítica de $2k_B^{-1}J/\log(1+\sqrt{2})$, donde k_B es la constante de Boltzmann de la termodinámica y J es la fuerza de las interacciones entre los espines. Para temperaturas cercanas al punto crítico, el calor específico tiende al infinito, al igual que el logaritmo de la correlación térmica, una característica de los cambios de fase. Contribuciones posteriores también derivaron varios exponentes críticos.

*

¿Qué tiene que ver todo este lío de espines de electrones e imanes con los estanques de fusión en el hielo del mar Ártico? La fusión del hielo es un cambio de fase, pero no se trata de un imán y derretirse no consiste en un cambio de espín. ¿Cómo puede haber una relación útil?

Si las matemáticas permaneciesen ligadas a la interpretación física particular que las ha originado, la respuesta sería «no puede». Sin embargo, no es así. No siempre, por lo menos. Es justo aquí donde entra en juego el misterio de la irrazonable eficacia de las matemáticas y este es el motivo por el que las personas que defienden que la inspiración en la naturaleza explica la *eficacia* se olvidan del calificativo «irrazonable».



Simulación del desarrollo de estanques de fusión basada en el modelo de Ising.
Yi-Ping Ma.

A menudo la primera indicación de que este tipo de portabilidad, en la que una noción matemática salta de un campo de aplicación a otro en apariencia sin relación, es, posiblemente, una inesperada apariencia familiar en una fórmula, un gráfico, un número o una imagen. Es común que este tipo de similitudes resulten no ser más que un juego de imágenes, un accidente o una coincidencia, llena de ruido y furia, que nada significa. Después de todo, la cantidad de gráficos o formas es limitada.

No obstante, a veces y solo a veces, constituye de verdad un indicio de una relación más profunda.

Y así es como se inició la investigación a la que llega por fin este capítulo. Hace cerca de diez años, un matemático llamado Kenneth Golden miraba unas fotografías de hielo marino en el Ártico y se dio cuenta de que tenían un parecido increíble con imágenes de campos de espín de electrones cerca del cambio de fase en el punto de Curie. Se preguntó si el modelo de Ising podría aprovecharse para arrojar luz sobre la manera en que se forman y extienden los estanques de fusión. En el caso del polo los cálculos se aplican a una escala mucho mayor y se sustituye la configuración arriba/abajo de un electrón minúsculo por el estado congelado/derretido de una región superficial del hielo marino de cerca de un metro cuadrado.

Pasó bastante tiempo hasta que esta idea dio frutos en la forma de matemáticas serias, pero cuando lo hizo, llevó a Golden, que trabajaba junto con el científico atmosférico Court Strong, a una descripción novedosa de los efectos del cambio climático sobre el hielo marino. Mostró algunas simulaciones del modelo de Ising a un colega especialista en analizar imágenes de estanques de fusión y este pensó que se trataba de fotografías reales. Un análisis más detallado de las características estadísticas de las figuras (tales como la relación entre las áreas de los estanques y sus perímetros, una medición de lo irregulares que son sus bordes) demostró que los números se correspondían de manera muy estrecha.

La geometría de los estanques de fusión es crucial en la investigación del clima porque influye sobre procesos importantes del hielo marino y de las capas superiores del océano. Estos incluyen la manera en que cambia el albedo de la superficie congelada (la cantidad de luz y calor radiante que refleja) conforme esta se derrite, cómo se fragmentan los témpanos y cómo cambia su tamaño. A su vez, esto afecta a los patrones de luz y sombra bajo el hielo, a la fotosíntesis de las algas y a la ecología de los microbios.

Cualquier modelo aceptable debe estar de acuerdo con dos conjuntos principales de observaciones. En 1998, la expedición SHEBA midió los tamaños de los estanques de fusión mediante fotografías tomadas desde helicópteros. La distribución de probabilidad observada de la superficie de las zonas derretidas es una ley potencial: la probabilidad de encontrar un estanque de área A es aproximadamente proporcional a A^k, donde la constante k está cerca de −1,5 para tamaños entre 10 y 100 metros cuadrados. Este tipo de distribución es indicativa a menudo de una geometría de fractales. Estos mismos datos, junto con las observaciones de la expedición ártica Healey-Oden, HOTRAX, de 2005, revelan un cambio de fase en la geometría fractal de los estanques de fusión conforme crecen y se combinan, al evolucionar

desde formas simples a regiones autosimilares cuyos bordes se comportan como funciones que recubren el espacio. La dimensión fractal de las curvas de las orillas (la relación entre el área y el perímetro) pasa de 1 a cerca de 2 en una superficie crítica de unos 100 metros cuadrados. Esto afecta a la manera en que cambian las anchuras y las profundidades de los estanques, lo que a su vez tiene un efecto sobre la dimensión de la interfaz agua-hielo a través de la que se expanden las zonas derretidas y, en última instancia, sobre la velocidad a la que se funden.

El valor observado del exponente k es $-1{,}58 \pm 0{,}03$, acorde con la estimación de SHEBA de −1,5. El cambio en la dimensión fractal constatado por HOTRAX puede calcularse de forma teórica mediante un modelo de percolación y el tamaño más grande, de cerca de 2, resulta ser $91/48 = 1{,}896$ en esta representación. La simulación numérica del modelo de Ising arroja una dimensión fractal muy próxima a este valor[68].

Una característica interesante de este trabajo es que se manejan escalas de longitud muy pequeñas, de unos pocos metros. La mayor parte de los modelos del clima tienen escalas de varios kilómetros, por lo que este tipo de análisis supone un desarrollo radical y novedoso. Todavía está en pañales y hace falta desarrollar el modelo para incorporar más física del hielo fundente, la absorción y la radiación de la luz solar e incluso el viento. Pero ya sugiere formas nuevas de comparar las observaciones con las formulaciones matemáticas y es un primer paso para explicar por qué los estanques de fusión dibujan unas formas fractales tan intrincadas. También es el primer modelo matemático de la física básica de los estanques de fusión.

El informe de *The Guardian* citado al inicio de este capítulo pintaba un panorama desolador. La reciente aceleración de la pérdida de hielo ártico, constatada a partir de las observaciones, no de modelos matemáticos, implica que la subida del nivel del mar para 2100 será de dos tercios de metro (unos 66 centímetros). Esto son siete centímetros más que la predicción anterior del Grupo Intergubernamental de Expertos sobre el Cambio Climático, IPCC, por sus siglas en inglés. Unos 400 millones de personas estarán en riesgo de sufrir inundaciones cada año, un 10 % más que los 360 millones que el IPCC había previsto antes. La subida del nivel del mar también hace que las marejadas ciclónicas sean más graves, lo que causará más daños en las zonas costeras. En la década de 1990, Groenlandia perdía 33 000 millones de toneladas métricas de hielo anuales. A lo largo de los últimos diez años, esta tasa se ha elevado hasta los 254 000 millones de toneladas anuales. Desde 1992, se han perdido 3,8 billones de toneladas de hielo. Cerca de la mitad de esta pérdida

está causada por glaciares que se mueven con mayor rapidez y que se rompen cuando alcanzan el océano. La otra mitad se debe al derretimiento, que ocurre sobre todo en la superficie. De modo que la física de los estanques de fusión tiene ahora una importancia vital para todo el mundo.

Si la metáfora de Ising se pudiera hacer más precisa, entonces todas las ideas potentes sobre su modelo, adquiridas mediante los esfuerzos denodados de generaciones de físicos matemáticos, podrían aplicarse a los estanques de fusión. En particular, la conexión con la geometría de fractales hace posible que haya ideas nuevas acerca de la compleja geometría de los estanques. Sobre todo, la historia de Ising, del hierro y del deshielo del Ártico es un ejemplo estupendo de la irrazonable eficacia de las matemáticas. ¿Quién podría haberse imaginado, hace un siglo, que el modelo de Lenz del cambio de fase ferromagnético tendría algo que ver con el cambio climático y con la desaparición en curso de los casquetes polares?

13
¡Que alguien llame al topólogo!

> Las características topológicas son robustas. El número de componentes o agujeros no es algo que deba cambiar por un pequeño error en la medición. Esto resulta crucial para las aplicaciones.
>
> ROBERT GHRIST,
> *Elementary Applied Topology*

La topología, un tipo flexible de geometría, era en un principio una parte muy abstracta de las matemáticas puras. La mayoría de las personas que al menos han oído hablar de ella lo siguen pensando así, aunque eso está empezando a cambiar. Parece bastante improbable que pueda existir algo llamado «topología aplicada». Sería como enseñar a un burro a volar: lo digno de mención no sería que el animal lo hiciese bien, sino que lo hiciese en absoluto. Tal vez esta afirmación sea correcta referida a los burros, pero no lo es para la topología. En el siglo XXI, esta se abre paso a zancadas y resuelve problemas importantes del mundo real. Ya hace tiempo que avanzaba en esa dirección sin hacerse notar y ahora ha alcanzado el punto en el que puede considerarse con seguridad como una nueva rama de las matemáticas aplicadas. Y no se trata solo de unas pocas aplicaciones de partes aleatorias de la topología: sus usos prácticos son amplios y generalizados y las herramientas propias de la disciplina involucradas abarcan grandes áreas de esta rama, incluidas las más sofisticadas y abstractas. Trenzas. Complejos de Vietoris-Rips. Campos vectoriales. Homología. Cohomología. Homotopía. Teoría de Morse. Número de Lefschetz. Fibrados. Haces. Categorías. Colímites.

Hay un motivo para esto: *unidad*. La topología en sí misma ha crecido, en poco más de un siglo, de ser un montón de curiosidades triviales a convertirse en un área de investigación y conocimiento integrada por completo. En la actualidad es uno de los pilares fundamentales sobre los que se alza el conjunto de las matemáticas. Y allí donde llegan las matemáticas puras, es

habitual que sigan las aplicadas. En última instancia. (También sucede al revés).

La topología estudia cómo cambian las formas bajo transformaciones continuas y, en particular, qué características permanecen. Ejemplos conocidos de estructuras topológicas son la banda de Möbius, una superficie con una sola cara, y los nudos. Durante cerca de ochenta años, los matemáticos estudiaron este campo por su interés intrínseco y sin pensar en las aplicaciones. La disciplina se volvía más abstracta cada vez y se inventaron estructuras algebraicas difíciles de comprender, las denominadas homología y cohomología, para hacer cosas como contar el número de agujeros en una forma topológica. Todo parecía muy rebuscado y poco claro, sin implicaciones prácticas.

Sin dejarse desanimar, los investigadores siguieron con su trabajo en este campo, debido a su papel central en el pensamiento matemático avanzado. Conforme los ordenadores se hicieron más potentes, se empezaron a buscar maneras de implementar los conceptos topológicos de manera electrónica, que permitiese investigar formas muy complejas. Pero hubo que modificar el enfoque para posibilitar que los ordenadores pudiesen manejar los cálculos. El resultado, la «homología persistente», es un método digital para detectar agujeros.

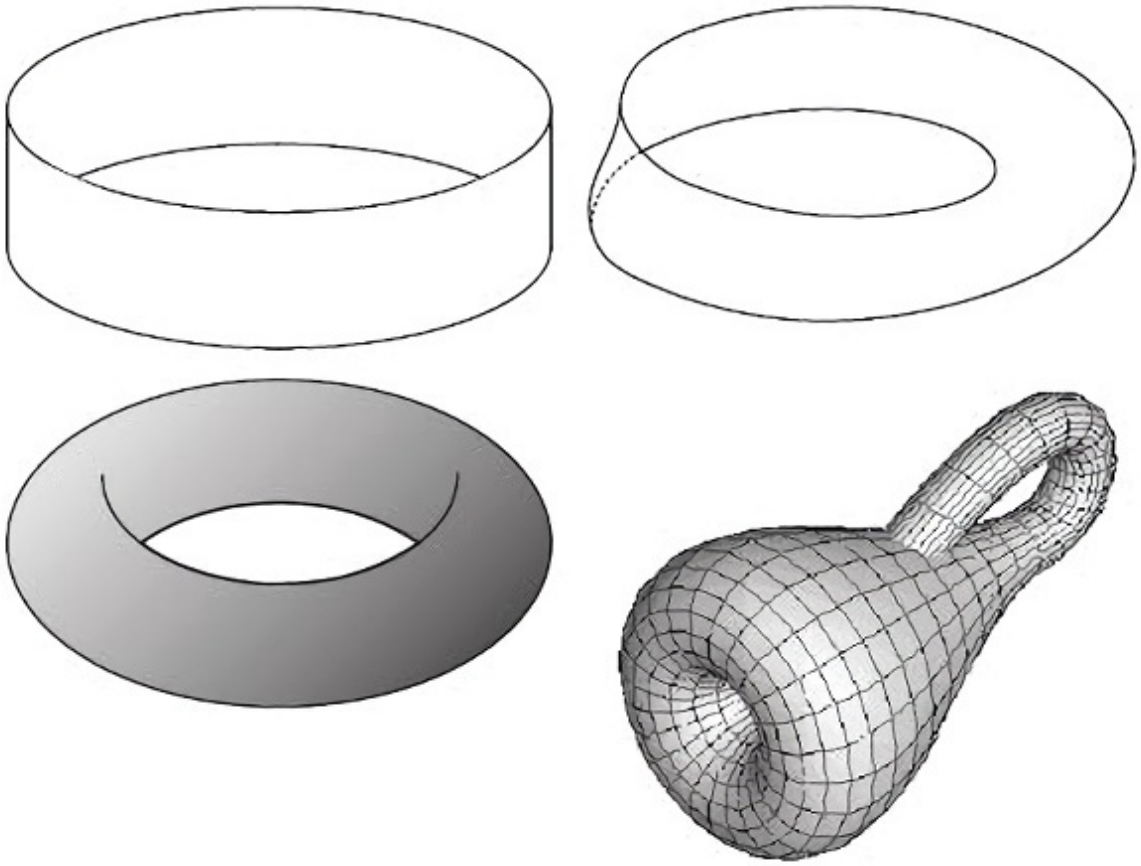
A simple vista, esta tarea parece algo muy alejado del mundo real. Sin embargo, resulta que la topología es ideal para resolver algunos problemas acerca de redes de sensores de seguridad. Imaginemos una instalación gubernamental de alto secreto, rodeada de bosques, que llama la atención de un grupo de terroristas o de ladrones. Para detectar su aproximación, se instalan sensores de movimiento en los árboles. ¿Cuál es la forma más eficaz de hacerlo y cómo puede tenerse la garantía de que no hay agujeros sin cubrir, a través de los cuales podrían pasar los malhechores sin ser detectados?

¿Agujeros? ¡Por supuesto! ¡Que alguien llame al topólogo!

*

Cuando se llega a la topología por primera vez, es habitual aprender algunas formas básicas. Parecen pequeños juguetes, sencillos y extraños. Algunos son caprichosos, otros directamente raros. Pero estos caprichos tienen un objetivo. Como dijo una vez el gran matemático Hilbert: «El arte de hacer las matemáticas consiste en la búsqueda de ese caso especial que contiene todos

los gérmenes de la generalidad». Si se elige bien el juguete, se abren las puertas de toda una nueva disciplina.



Izquierda arriba: cilindro. *Derecha arriba*: banda de Möbius. *Izquierda abajo*: toroide. *Derecha abajo*: botella de Klein.

Las dos primeras formas que aparecen en la imagen pueden hacerse con una tira de papel unida por los extremos. La manera evidente de hacerlo produce una banda cilíndrica. La otra, menos obvia, consiste en girar 180° un extremo antes de unirlo con el otro. Esta es una cinta de Möbius, bautizada así por August Möbius, quien la descubrió en 1858, si bien un estudiante de Gauss, Johann Listing, se había percatado de ella antes. Listing publicó por primera vez el nombre «topología» en 1847, aunque fue Gauss quien le puso en contacto en primer lugar, de manera premonitoria, con esta disciplina naciente.

Un cilindro tiene dos bordes separados, ambos circulares, y dos lados diferentes. Es posible pintar el interior de rojo y el exterior de azul y los colores no se mezclan nunca. En topología, lo que cuentan son las propiedades de la forma que permanecen cuando se modifica esta de manera continua. Se pueden estirar las partes, comprimirlas, darles la vuelta a las cosas, pero no cortar ni romper, a no ser que se vuelva a unir todo más tarde.

La anchura uniforme de la banda cilíndrica de la imagen no es una propiedad topológica: es susceptible de cambiar bajo una deformación continua. Por motivos similares, tampoco lo es la circularidad de los bordes. Pero la condición de borde, que haya dos diferentes y dos lados distintos, sí son propiedades topológicas.

Las formas que se consideran equivalentes cuando se modifican tienen un nombre especial: se denominan espacios topológicos. La definición real es muy abstracta y técnica, de modo que emplearé imágenes más informales. Todo lo que voy a decir puede precisarse y demostrarse con rigor.

Es posible emplear estas propiedades topológicas para comprobar que un cilindro no se puede modificar para convertirlo en una banda de Möbius. Incluso aunque ambos se construyen al pegar los extremos de una tira de papel, son espacios topológicos diferentes. El motivo es que la banda de Möbius solo tiene un borde y un lado. Si se pasa un dedo a lo largo del primero, este se recorre *dos veces* antes de volver al principio y cambia de arriba a abajo debido al giro de 180°. Del mismo modo, si se empieza a pintar la superficie de rojo, se da toda la vuelta hasta descubrir que se está dando color a la parte de atrás de lo que ya se ha pintado, de nuevo por el giro de 180°. De modo que la banda de Möbius tiene propiedades topológicas diferentes comparadas con las del cilindro.

La forma de abajo a la izquierda se parece a un dónut. Los matemáticos la llaman un toroide, nombre con el que designan solo a la superficie, no a la parte interna en la que estaría la masa. Desde este punto de vista, se parece más en realidad a un flotador hinchable. Tiene un agujero. Puede meterse un dedo o, en el caso del flotador, todo el cuerpo. Pero el agujero no está en la propia superficie. Si lo estuviese, el flotador se desinflaría y la persona se hundiría. El agujero consigue estar en un sitio en el que no está la superficie. Vale, es verdad que un instalador de banda ancha dentro de un registro también está en un lugar en el que no está la superficie. Pero el registro tiene bordes, mientras que el toroide consigue tener un agujero incluso cuando carece de ellos. Al igual que el cilindro, tiene dos lados: el que puede verse en la imagen y el que está «dentro».

La forma de abajo a la derecha es menos conocida. Se llama botella de Klein, por el gran matemático alemán Felix Klein y porque parece una botella. Es probable que el nombre sea un juego de palabras en alemán, porque en ese idioma *Fläche* significa «superficie» y *Flasche* es una «botella». Hay un aspecto en el que la imagen es engañosa: la superficie parece pasar a través de sí misma. En la que puede construir un matemático

no lo hace. La intersección consigo misma se debe a que las ilustraciones se dibujan de manera natural como si los objetos estuviesen en el espacio tridimensional. Para obtener una botella de Klein que no cruce su propia superficie hace falta, o bien ir a las cuatro dimensiones, o bien, mejor todavía, seguir la práctica topológica habitual de prescindir por completo de la necesidad de un espacio que rodee el objeto. Entonces es posible ver la botella de Klein como un cilindro cuyos dos extremos circulares se han unido entre sí, pero con uno invertido antes de hacer la unión. Para lograrlo en tres dimensiones hay que meter ese extremo dentro y luego abrirlo de nuevo hacia fuera, pero también puede hacerse de manera conceptual con solo añadir la regla de que cuando se cae por un extremo se acaba en el otro, invirtiendo el sentido de la vuelta. La botella de Klein no tiene bordes, como un toroide, pero además es similar a la banda de Möbius en que solo tiene un lado.

Ya hemos conseguido distinguir estos cuatro espacios topológicos entre sí. O bien tienen un número diferente de bordes, o bien de lados. O diferentes tipos de agujeros, si tan solo se pudiese especificar a lo que se refiere el término «agujero». Esta observación pone sobre la mesa uno de los problemas fundamentales de esta disciplina. ¿Cómo puede saberse si dos espacios topológicos son iguales o diferentes? No basta con fijarse en la forma, porque es susceptible de ser modificada. Como dice el dicho, para un topólogo un dónut es lo mismo que una taza de café. Hay que recurrir a propiedades *topológicas* que distingan los espacios.

Puede ser difícil.

*

La botella de Klein parece el juguete arquetípico de un matemático. Es difícil ver cómo podría ser relevante en el mundo real. Por supuesto, como insistía Hilbert, los juguetes matemáticos son útiles, no por sí mismos, sino por las teorías a que dan pie, de modo que la botella de Klein no necesita justificar su existencia de manera directa. No obstante, sucede que esta superficie extraña se encuentra en la naturaleza. Aparece en el sistema visual de los primates: monos, simios y, por supuesto, nosotros.

Hace más de un siglo, el neurólogo John Hughlings Jackson descubrió que la corteza cerebral humana contiene, de alguna manera, un mapa topográfico de los músculos del cuerpo. La corteza es la superficie sinuosa del cerebro, de modo que todo el mundo lleva un mapa de su musculatura dentro de su cabeza. Tiene sentido, porque este órgano controla las contracciones y

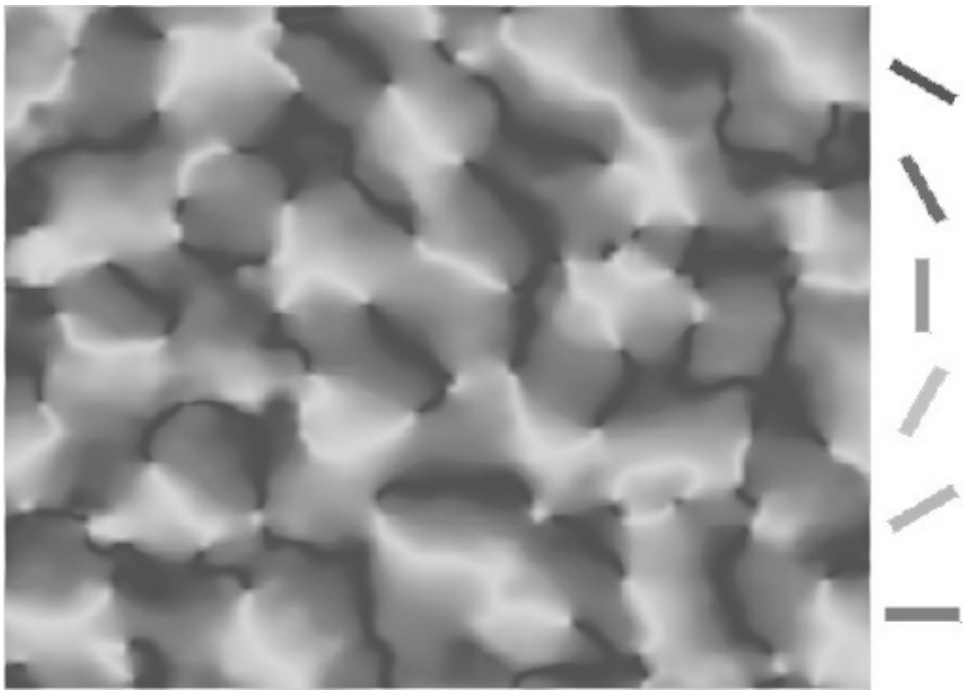
las distensiones de los músculos que son la causa del movimiento. Una gran parte de la corteza se dedica a la visión y ahora se sabe que el córtex visual contiene mapas parecidos que hacen funcionar el proceso de la vista.

Esta no se reduce a un ojo que actúa como una cámara y envía una fotografía al cerebro. Es algo mucho más complicado porque este último tiene que reconocer la imagen tanto como recibirla. Al igual que un objetivo, el globo ocular tiene una lente para enfocar la luz entrante y la retina actúa, en cierta manera, como una película. De hecho, es mucho más parecido al modo en que las cámaras digitales graban las fotografías. La luz incide sobre unos receptores minúsculos en la retina llamados bastoncillos y conos y las conexiones neuronales transmiten las señales resultantes hasta la corteza a lo largo del nervio óptico, un haz de muchas fibras nerviosas. Estos impulsos se procesan a lo largo del trayecto, pero el córtex hace la mayor parte del análisis.

La corteza visual puede concebirse como una serie de capas, unas encima de otras. Cada una desempeña un papel concreto. La capa superior, V1, detecta los límites entre las diferentes partes de la imagen. Este es el primer paso para *segmentar* la señal en sus elementos constituyentes. Los datos resultantes se transmiten a mayor profundidad en el córtex, hacia el interior, son analizados en cada etapa para obtener el siguiente tipo de información estructural, que se transforma para ser enviada al próximo nivel. Por supuesto, esta descripción es una simplificación, como lo son las «capas», y hay un montón de señales que viajan también en el sentido opuesto. Todo el sistema crea en nuestras cabezas una representación tridimensional multicolor del mundo externo, tan viva y detallada que se asume por defecto que *es* el mundo externo. Algo que no es del todo cierto, como demuestran un amplio surtido de ilusiones y ambigüedades ópticas. En cualquier caso, la corteza termina por segmentar en última instancia la imagen en partes que pueden reconocerse como un gato, la tía Vera o lo que sea. Entonces el cerebro es capaz de recuperar información adicional, tal como el nombre del animal o el hecho de que Vera ha ganado la lotería hace poco.

La capa V1 detecta los límites mediante el empleo de regiones de células nerviosas que son sensibles a los bordes que apuntan en direcciones concretas. La ilustración siguiente muestra una porción de V1, obtenida mediante registro de señales ópticas de la corteza visual de un macaco. Los diferentes tonos de gris (colores en el artículo original, de modo que me referiré a ellos así) corresponden a neuronas que se activan cuando reciben datos que indican un borde en esa orientación. Los colores se funden de

manera continua de un tono al siguiente, excepto en ciertos puntos aislados en cuya vecindad están presentes todos ellos, en una configuración parecida a la de un molinillo de papel. Estas ubicaciones son singularidades del campo de orientaciones.



Los colores (en este caso tonos de gris) muestran la orientación que da lugar a la mayor actividad en cada región de la corteza. La orientación percibida cambia con suavidad excepto en singularidades en las que se reúnen los colores.
Blasdel, G. G., «Orientation selectivity, preference, and continuity in monkey striate cortex» en Journal of Neuroscience, 12, 1992, pp. 3139-3161.

Esta disposición se ve limitada por propiedades topológicas del campo de orientaciones. Solo hay dos formas de disponer la serie de colores en torno a una singularidad de modo que cambien de manera continua: o bien seguir la secuencia en el sentido de las agujas del reloj, o bien en el opuesto. La imagen muestra ejemplos de ambos casos. La presencia de singularidades es inevitable porque la corteza tiene que emplear muchos molinillos para detectar una línea completa.

Cabe preguntarse ahora cómo combina el cerebro esta información de las orientaciones con la que se refiere al movimiento del borde. El sentido sobre una recta se suele indicar mediante una flecha (el norte se opone al sur, aunque ambos se encuentren en la misma línea) y tras una rotación de 180° esta apunta al revés. Hay que seguir hasta girar un total de 360° antes de que señale de nuevo al inicio. Los bordes no tienen una flecha que indique el sentido y vuelven a estar en una posición idéntica tras un giro de 180°. La corteza tiene que conseguir que ambas cosas funcionen al mismo tiempo de

alguna manera. Si se dibuja un bucle en torno a una singularidad, las orientaciones varían de manera continua al recorrerlo, pero el campo de direcciones debe saltar de un sentido dado al opuesto (digamos de apuntar al norte a hacerlo al sur) una vez o, de manera más general, un número impar de veces. Estas afirmaciones son de naturaleza topológica y llevaron a Shigeru Tanaka a concluir que los campos respectivos están conectados entre sí con la topología de una botella de Klein[69]. Esta predicción se ha verificado en la actualidad de manera experimental en varios animales, entre ellos monos aotus, gatos y hurones, lo que ha proporcionado pruebas de que la organización de la corteza visual puede ser similar en muchas especies de mamíferos diferentes. No se han realizado experimentos en humanos por motivos éticos, pero somos mamíferos, primates de hecho. De modo que es plausible que, al igual que los macacos, haya botellas de Klein en nuestras cabezas para ayudar a percibir los objetos en movimiento.

Estas ideas no tienen interés solo para los biólogos. La biomímesis es un área en rápida expansión en la que los ingenieros cogen ideas de la naturaleza para mejorar la tecnología, lo que lleva a nuevos materiales y a máquinas novedosas. Por ejemplo, la curiosa estructura del ojo de la langosta ha desempeñado un papel crucial en la invención de los telescopios de rayos X[70]. Para enfocar un haz de este tipo de radiación hay que cambiar su dirección, pero es tan energética que un espejo adecuado solo consigue desviarla un ángulo muy pequeño. La evolución de las langostas resolvió un problema similar para la luz visible hace millones de años y la misma geometría funciona con los rayos X. La recién adquirida comprensión de la capa V1 de la corteza en los mamíferos puede trasladarse a la visión de los ordenadores, con aplicaciones potenciales en cosas tales como vehículos sin conductor e interpretación automática de imágenes de satélite para fines militares o civiles.

*

La pregunta central en la topología es «¿qué forma tiene esto?». Es decir, «¿a qué espacio topológico me enfrento aquí?». Puede parecer una cuestión trivial, pero las matemáticas presentan estos espacios de innumerables maneras (como imágenes, como fórmulas, como soluciones de ecuaciones…), de modo que no siempre es fácil reconocer lo que se tiene entre manos. Por ejemplo, hace falta un topólogo para ver una botella de Klein en la capa V1 de un macaco. Ya se hizo un intento de abordar este problema cuando se

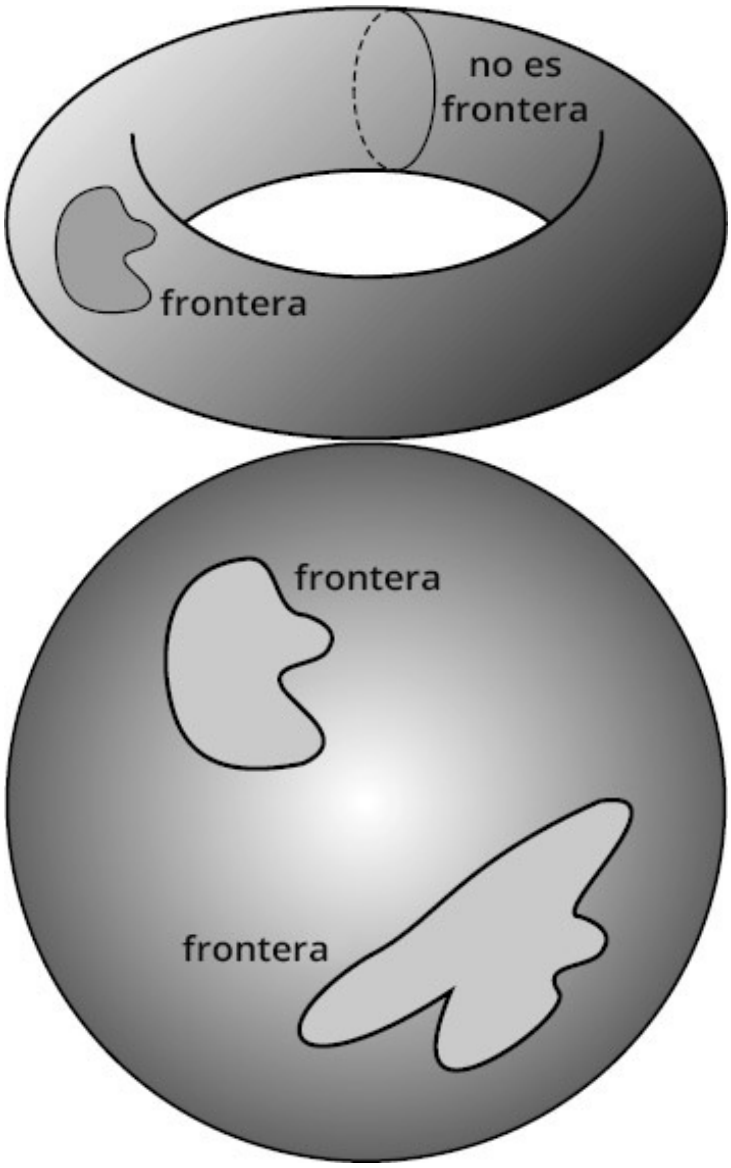
observó que las características topológicas distinguían los cuatro espacios de la ilustración anterior: cilindro, banda de Möbius, toroide y botella de Klein. A finales del siglo XIX y principios del XX, los matemáticos desarrollaron formas sistemáticas de encarar esta cuestión. La idea principal es definir invariantes topológicos: propiedades que pueden calcularse y que son las mismas para espacios equivalentes desde un punto de vista topológico, pero que son diferentes para al menos algunos espacios no equivalentes. Por lo general, no son lo bastante sensibles como para diferenciar todos los casos distintos, pero incluso una clasificación parcial es útil. Si dos formas tienen un invariante diferente de algún tipo, es seguro que tienen topologías dispares. Cuando se consideran los cuatro ejemplos a los que acabo de hacer mención, las propiedades de este tipo son cosas como «¿cuántos bordes tiene?» y «¿cuántos lados hay?».

A lo largo de las décadas, algunos invariantes resultaron ser más útiles que otros y se construyeron algunos de importancia fundamental. El que quiero considerar ahora, en parte porque ha adquirido hace poco algunas aplicaciones destacadas, se denomina homología. De manera resumida, cuenta la cantidad de agujeros de una dimensión dada que tiene un espacio. De hecho, va más allá de contarlos: combina a la vez tanto los agujeros como los no agujeros en un único objeto algebraico denominado grupo de homología.

Hay un espacio topológico muy básico que todavía no he mencionado: la esfera. Al igual que en el caso de un toroide, cuando los matemáticos emplean esta palabra se refieren a la superficie de delgadez infinita de una esfera y no a una maciza, que sería una *bola*. Como el toroide y la botella de Klein, carece de bordes. Se puede demostrar que es diferente desde un punto de vista topológico de esos dos objetos prestando atención a los agujeros o a la falta de ellos.

Empecemos con el toroide. Con solo mirarlo puede verse que tiene un agujero enorme que lo atraviesa por la mitad. Una esfera no tiene ni el más remoto parecido con eso. Pero ¿cómo se define un agujero desde el punto de vista matemático, de manera que no dependa del espacio que lo rodea? La respuesta consiste en fijarse en las curvas cerradas de la superficie. Cada una de ellas sobre una esfera constituye la frontera de una región que, topológicamente hablando, es un disco, el interior de una circunferencia[71]. Demostrarlo es bastante difícil, aunque es factible, así que asumamos que es cierto. En un toroide, algunas curvas cerradas también son fronteras de discos, pero otras no. De hecho, ninguna de ellas que pase «a través» del agujero lo

es. Esto también es muy difícil de demostrar, pero dejémonos llevar de nuevo y démoslo por válido. Queda probado que una esfera es diferente de un toroide en el sentido que se maneja aquí del término, porque «curva cerrada» y «encerrar un disco (topológico)» son propiedades topológicas.

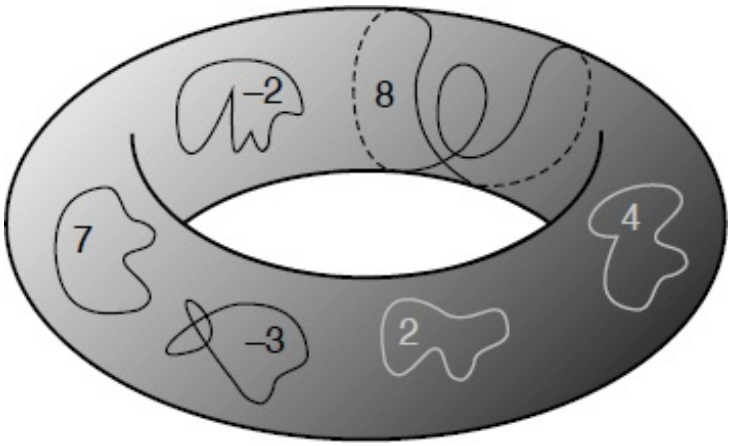


Arriba: en un toroide, algunas curvas cerradas son fronteras, mientras que otras no lo son. *Abajo*: en una esfera, todas las curvas cerradas son fronteras.

Puede hacerse lo mismo para más dimensiones. Por ejemplo, en tres dimensiones es posible sustituir «curva cerrada» por «superficie (topológicamente) esférica» y «encerrar un disco» por «encerrar una bola». Si se encuentra una esfera que no cumple esta condición, el espacio tiene algún tipo de agujero tridimensional. Para ir más allá y dar alguna interpretación de qué tipo de agujero es, los pioneros de la topología descubrieron que es posible sumar y restar esferas o curvas cerradas. Consideraré cómo se hace

para estas últimas en superficies. Es parecido para cantidades mayores de dimensiones, pero más complicado.

En esencia, dos curvas cerradas se suman cuando se dibujan ambas en la misma superficie. Para sumar un conjunto entero de ellas, se trazan todas. Hay algunos refinamientos técnicos: a menudo es útil añadir una flecha para especificar la orientación y puede dibujarse la misma curva muchas veces, o incluso un número negativo de veces. Esto es casi lo mismo que trazar su inversa (idéntica, pero con sentido opuesto) un número positivo de veces, de la manera que explicaré en breve.



Un ciclo en un toroide.

Un conjunto de curvas, etiquetadas con índices que indican cuántas veces hay que trazarlas, se denomina un ciclo. Hay una cantidad infinita de ciclos posibles en una superficie, pero desde el punto de vista topológico, muchos de ellos son equivalentes a otros tantos. Ahora bien, acabo de decir que un ciclo negativo es el mismo que uno positivo con todas las flechas invertidas. Así dicho, esto no es cierto porque «mismo» quiere decir «idéntico» y no es el caso. Pero puede *hacerse* que sea equivalente con la versión topológica del truco que la teoría de números emplea en la aritmética modular. Aquí, aunque 0 y 5 son diferentes, es posible pretender que son semejantes para ciertos fines adecuados y obtener el anillo $\mathbb{Z}_5$ de enteros módulo 5. En teoría de homologías, se hace algo parecido y se considera que cualquier curva cerrada que es frontera de un disco es equivalente a la curva cero (no se traza *ninguna* copia de ella). Una curva así se denomina una frontera y se dice que es homóloga a cero. La misma idea se aplica a los ciclos: un ciclo es homólogo a cero si es una combinación de curvas, cada una de las cuales es una frontera.

Es posible sumar los ciclos C, D para obtener $C + D$, como ya se ha descrito, y también restarlos al invertir la flecha en D para obtener $C - D$, excepto que $C - C$ no tiene por qué ser igual a 0. Esto es incómodo, pero hay

una salida: siempre es *homólogo* a cero. Si se pretende que cualquier cosa homóloga a cero *es* 0, entonces se dispone de un objeto algebraico correcto que se llama grupo de homología de la superficie. En efecto, se realizan operaciones algebraicas sobre ciclos *módulo* (es decir, descartando) las fronteras. Igual que se hace aritmética (mod 5) descartando los múltiplos de 5.

Eso es la homología.

El grupo de homología de una esfera es trivial: todos los ciclos son homólogos a cero y el grupo solo consiste en 0. En cambio, el grupo de homología de un toroide no es trivial: algunos ciclos no son homólogos a cero. Resulta que todos ellos son homólogos a un entero múltiplo del que está marcado como «no frontera» en la imagen, de modo que el grupo de homología del toroide es una forma disfrazada de $\mathbb{Z}$, los números enteros. No voy a hacer los cálculos ni los diagramas, pero el grupo de homología de la botella de Klein es $\mathbb{Z}_2 \times \mathbb{Z}_2$, pares (m, n) de números enteros módulo 2. Así que tiene algún tipo de agujero, pero es de un tipo diferente del que hay en (bueno, no *en*) un toroide.

Hay un motivo por el que he repasado la construcción, más bien complicada, del grupo de homología: para que se tenga una idea de la manera en que los topólogos construyen los invariantes. Pero la única idea con la que hay que quedarse es que todo espacio tiene un grupo de homología, que es un invariante topológico, y que puede emplearse para aprender mucho acerca de la forma que tiene el espacio. Topológicamente hablando.

*

El grupo de homología se remonta a las investigaciones pioneras de Enrico Betti y Poincaré a finales del siglo XIX. Su enfoque era *contar* las características topológicas, tales como los agujeros, pero fue reformulada en lenguaje de la teoría de grupos a finales de la década de 1920 por Leopold Vietoris, Walther Mayer y Emmy Noether y enseguida surgieron generalizaciones de gran envergadura. Lo que he denominado *el* grupo de homología es tan solo el primero de toda una secuencia de tales elementos, que define la estructura algebraica de agujeros de dimensión 1, 2, 3, etcétera. Hay también una noción dual de cohomología y otra asociada de homotopía, que se refiere a la manera en que se deforman las curvas y se unen extremo con extremo, más que a su relación con las fronteras. Poincaré sabía que esa construcción daba como resultado un grupo, que por lo general no es conmutativo. La topología algebraica es en la actualidad una disciplina

enorme y muy técnica y se siguen descubriendo nuevos invariantes topológicos.

Otra área que crece muy rápido es la conocida como topología aplicada. Hay una nueva hornada de matemáticos y científicos que han aprendido topología antes que a caminar, así que les parece mucho menos extraña que a las generaciones anteriores. Hablan con fluidez el idioma de la disciplina y han empezado a encontrar nuevas oportunidades de aplicarla a problemas prácticos. La botella de Klein en la visión es un ejemplo tomado de las fronteras de la biología. En ciencia de materiales y en ingeniería electrónica se encuentran nociones tales como los aislantes topológicos: materiales que pueden pasar de conducir la electricidad a aislar frente a ella al cambiar la topología de sus propiedades eléctricas. Las características topológicas, que se conservan en las deformaciones, son muy estables.

Uno de los conceptos más prometedores en topología aplicada cobró forma cuando los matemáticos teóricos intentaron escribir algoritmos para decirle a un ordenador cómo calcular grupos de homología. Lo consiguieron al reescribir la definición de grupo de un modo más asequible a los cálculos informáticos. Resultó que estas ideas eran un nuevo y potente método para analizar los «macrodatos». Este enfoque, tan de moda en todas las ramas de la ciencia, emplea ordenadores para buscar patrones ocultos en los datos numéricos y, como su nombre sugiere, sus métodos funcionan mejor para cantidades muy grandes de puntos. Por suerte, los sensores y los aparatos electrónicos de hoy en día son increíblemente buenos a la hora de medir, almacenar y manipular volúmenes gigantescos de información. Lo que resulta menos afortunado es que a menudo no se tiene ni idea de qué hacer con ella una vez que se ha recogido, pero es justo aquí donde se encuentra el reto matemático de los macrodatos.

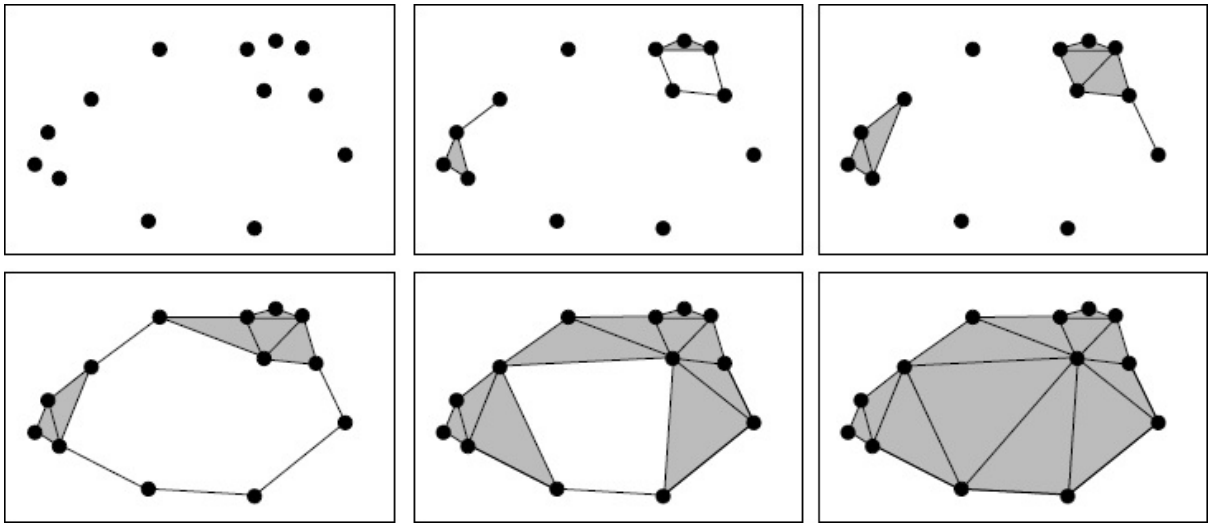
Supongamos que se han medido millones de números y que se representan de manera conceptual como una especie de nube de puntos en un espacio multidimensional de variables. Para extraer patrones significativos de esta aglomeración de datos hace falta encontrar sus características estructurales destacadas. Entre ellas es primordial la *forma* de la nube. No es factible representar esta última sin más en una pantalla y evaluarla a simple vista. Podría mirarse desde el ángulo equivocado, o regiones importantes podrían quedar ocultas por otras, o la cantidad de variables podría ser demasiado grande como para que el sistema visual las procese de manera razonable. Ahora bien, como ya se ha visto, «¿qué forma tiene esto?» es una de las cuestiones fundamentales de la topología. De modo que parece razonable que

los métodos propios de la disciplina resulten de utilidad, pongamos por caso, para distinguir una nube de datos más o menos esférica de una toroidal con un agujero en su interior. Ya se ha visto una versión en miniatura de esto en el proyecto FRACMAT del capítulo 8. En ese caso, lo importante era lo compacta que era la nube de puntos y si era redonda o con forma de puro. Los pormenores topológicos no tenían importancia.

No es posible hacer topología a mano para un millón de puntos: se han de emplear ordenadores. Pero estos no están construidos para hacer análisis de este tipo. De modo que los métodos que los matemáticos teóricos habían desarrollado para cálculos informáticos de grupos de homología se redirigieron al campo de los macrodatos. Y, como siempre, si se empleaban tal como estaban no lograban resolver del todo la tarea asignada. Había que modificarlos para adaptarse a los nuevos requisitos de los macrodatos, el principal de los cuales era que la forma de la nube no es algo bien definido. Depende, en particular, de la escala en la que se observe.

Por ejemplo, imaginemos una manguera de jardín enrollada. Desde una distancia moderada, sus segmentos parecen curvas, objetos de una dimensión desde un punto de vista topológico. Más de cerca, tienen la apariencia de superficies cilíndricas alargadas. Y más de cerca todavía, sus paredes adquieren un grosor; es más, hay un agujero que transcurre a lo largo de la mitad del cilindro. Si se mira desde una cierta distancia pero en una panorámica amplia, resulta que la manguera está enrollada como un muelle comprimido. Si se prescinde del detalle, el rollo de tubo se reduce a… un toroide.

Este tipo de efecto quiere decir que la forma de una nube de datos no es un concepto fijo. De modo que el grupo de homología tampoco es tan buena idea. En su lugar, los matemáticos se preguntaron cómo cambia la topología *percibida* de la nube con la escala de observación.



Unir puntos de datos separados por varias distancias crea una secuencia de triangulaciones, que revela agujeros de varios tamaños. La homología persistente detecta estos efectos.

Si se empieza a partir de una nube y de una escala de longitud seleccionada, es posible crear lo que los topólogos llaman un complejo simplicial al unir dos puntos mediante una arista siempre que estén a una distancia uno del otro menor que la escala de longitud. Después, las aristas que están próximas rodean triángulos y los triángulos que están próximos rodean tetraedros, etcétera. Un tetraedro multidimensional se llama un símplex y un conjunto de estos, unidos entre sí de alguna manera, constituyen un complejo simplicial. Para referirse a esto bastará aquí con un nombre más sencillo: «triangulación». Pero no debe olvidarse que los triángulos pueden ser de cualquier dimensión.

Una vez que se tiene una triangulación, hay reglas matemáticas para calcular la homología. Pero ahora, la triangulación depende de la escala de la observación, de modo que la homología también lo hace. Entonces, la pregunta interesante acerca de la forma se convierte en: ¿cómo cambia la homología de la triangulación al variar la escala? Las características más importantes de la forma deberían ser menos susceptibles ante los cambios que las más fugaces, que dependen de manera sensible de este tipo de variaciones. Así que es posible concentrarse en esos aspectos del grupo de homología que *permanecen* cuando cambia la escala. El aparato resultante no es solo un grupo de homología sino una familia de ellos, uno para cada valor de esta última variable, y se conoce como homología persistente.

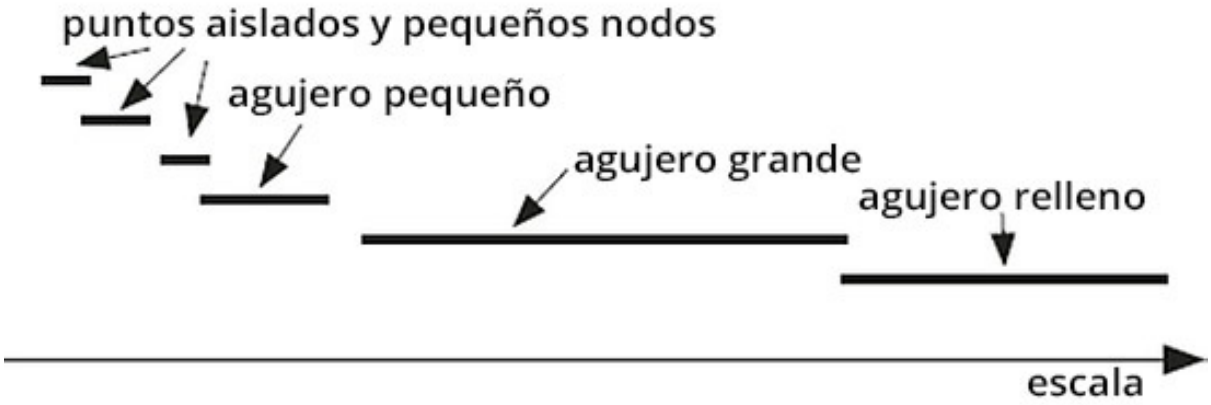


Gráfico de barras de la homología persistente que muestra las estructuras que se mantienen para según qué escalas (esquemático).

En este caso, la secuencia de seis imágenes muestra qué puntos se unen para los diferentes valores. Al aumentar la longitud de escala y contemplar estructuras de menor detalle, una nube inicial de puntos aislados empieza a formar nodos de tamaño reducido, uno de los cuales tiene un pequeño agujero. Este se llena y los nodos crecen. Entonces, estos se unen en un anillo para dejar al descubierto un gran agujero, que empieza a reducirse, aunque sigue siendo grande hasta que la escala es tan amplia que todo se llena. Esta imagen es esquemática y muestra detalles que un algoritmo informático habría eliminado en aras de la claridad. La característica dominante, que aparece en el rango mayor de escalas de longitud, es el gran agujero de la mitad.

Cabe destacar que esta descripción incluye información de distancias tanto como topología. Desde el punto de vista técnico, una transformación topológica no necesita conservar las distancias, pero en el análisis de datos los valores reales de estos son importantes, tanto como las características de la forma en general. Por este motivo, la homología persistente presta atención a las propiedades métricas tanto como a las topológicas. Una forma de presentar la información proporcionada por este método es construir un código de barras que emplea líneas horizontales para representar los intervalos de escalas sobre los que se conservan las características homológicas particulares (tales como agujeros). Por ejemplo, el código de barras para la nube de puntos presentada antes podría tener una apariencia similar a la de la imagen. Este tipo de representación es un resumen esquemático de la manera en que varía la topología con la escala.

*

La homología persistente y sus códigos de barras son muy elegantes, pero ¿para qué valen?

Imaginemos al dueño de un negocio cuyas oficinas están ubicadas en un claro en el bosque. Podrían aproximarse ladrones entre los árboles sin ser vistos. De modo que dispone un conjunto de sensores, cada uno de los cuales es capaz de detectar el movimiento y de comunicarse con los demás en su vecindad, y lo deja encendido por la noche. Si se acerca alguien, con o sin autorización, el sistema dispara una alarma y el personal de seguridad puede ir a investigar. O imaginemos a un general que dirige una base militar en una zona en la que hay grupos terroristas activos. Podría hacer algo parecido, pero con armas.

¿Cómo puede tenerse la certeza de que la zona está cubierta de manera adecuada, sin huecos a través de los cuales se pueda colar un criminal o un terrorista?

Si se emplea un número pequeño de sensores, es posible hacer un mapa de su distribución y evaluar el resultado a simple vista. Con cantidades más grandes, o con limitaciones cambiantes impuestas por el terreno, este enfoque es menos práctico. De modo que hace falta una manera de detectar agujeros en el área cubierta… ¿Detectar *agujeros*? Suena como un encargo para la homología persistente. Y, de hecho, este es uno de los muchos usos que se le dan en la actualidad a esta novedosa idea. Una aplicación similar es la «cobertura de barrera»: determinar si un conjunto de sensores rodea por completo un edificio o complejo importante. A su vez, la «cobertura de barrido» se refiere a sensores móviles, cuya versión doméstica o comercial sirve para aspiradores robóticos autónomos. ¿Limpiarán todo el suelo?

Una aplicación más científica se lleva a cabo en conjunto con el método de la ventana deslizante para reconstruir atractores dinámicos que mencioné en el capítulo 8. La homología persistente tiene la capacidad de detectar cuándo cambia de manera significativa la topología del atractor. En teoría de sistemas dinámicos este efecto se denomina bifurcación y apunta a un cambio importante en la dinámica. Una aplicación destacada es encontrar cómo ha cambiado el clima terrestre a lo largo de millones de años de periodos templados a glaciaciones, incluida la cobertura total del planeta por el hielo durante la etapa conocida como Tierra bola de nieve. Jesse Berwald y sus colegas han demostrado que los códigos de barras a partir de nubes de puntos de ventana deslizante funcionan de maravilla para identificar cambios en el régimen climático global[72]. Otros casos en los que el mismo método tiene validez en sistemas físicos diferentes son las vibraciones en las máquinas

herramienta que se emplean en las fábricas, las cuales pueden dar lugar a fallos y dejar marcas indeseadas en las superficies de los objetos que se elaboran. Firas Khasawneh y Elizabeth Munch han demostrado que mediciones de series temporales de la herramienta de corte son capaces de detectar este tipo de vibración, conocida en el sector como «retemblado»[73]. También hay aplicaciones en las imágenes médicas, tales como la detección de la bifonación en endoscopias de la laringe, por Christopher Tralie y Jos Perea[74]. Este efecto se produce cuando una misma cuerda vocal produce dos frecuencias sonoras a la vez y puede ser síntoma de lesiones o de parálisis. La endoscopia es la introducción de una cámara en el extremo de un cable de fibra óptica que sube por la nariz y baja por la garganta. Saba Emrani y otros[75] han empleado códigos de barras sobre datos de audio para detectar sibilancias en pacientes, un sonido agudo anormal que puede ser síntoma de unas vías aéreas parcialmente bloqueadas o de enfermedades pulmonares, tales como asma y cáncer de pulmón, e insuficiencia cardíaca.

¿Hay algún problema con los datos? ¿Hace falta ayuda urgente?

¡Que alguien llame al topólogo!

14
El zorro y el erizo

> Muchas cosas sabe el zorro, pero el erizo sabe una sola y grande.
>
> Atribuido a ARQUÍLOCO, *c.* 650 a. C.

Cuando investigaba este libro, me topé con la frase: *Πόλλ' οἶδ' ἀλώπηξ, ἀλλ' ἔχῑνος Εν μέγα* . Estudié latín en el colegio, no griego, pero los matemáticos conocemos estos caracteres. Incluso yo puedo reconocer las palabras «*equinos*» y «*mega*» y aventurar una interpretación. Es algo acerca de un erizo grande. La traducción de verdad es: «Muchas cosas sabe el zorro, pero el erizo sabe una sola y grande». Es probable que fuese acuñada por el poeta de la Grecia clásica Arquíloco, aunque no es seguro.

¿Yo debería ser un zorro o un erizo? ¿Debería intentar describir una selección de los innumerables y espectaculares desarrollos en el campo de las matemáticas a lo largo de los últimos cincuenta años y cómo se emplean? ¿O debería centrarme en *una cosa sola y grande*?

Decidí hacer ambas.

Ya se ha superado la parte del zorro con sus trece capítulos. Ahora viene el erizo a resumirlo todo.

Echando la vista atrás a los temas que he tratado, me sorprenden la riqueza y la variedad de las diferentes ramas de las matemáticas que ya se han abierto camino en los sistemas y dispositivos característicos de la vida a principios del siglo XXI. No solo para las personas pudientes en democracias occidentales, aunque es probable que se beneficien más que otras de menos recursos, sino para miles de millones de habitantes de todos los países del mundo. El teléfono móvil ha llevado las comunicaciones modernas a las naciones en desarrollo. Ya son ubicuos y lo han cambiado todo. No siempre para mejor, pero el cambio es un arma de doble filo. Sin las matemáticas y sin tantas y tantas personas formadas para emplearlas a un nivel avanzado, no habría teléfonos móviles.

No se me escapa el gran número de aplicaciones que no he tenido ocasión de mencionar. Las que he descrito no tienen por qué ser las mejores, las más importantes, las más impresionantes ni las más valiosas. Tan solo son un conjunto de ellas que me llamaron la atención porque recurren a unas matemáticas sólidas, nuevas por lo general, y las emplean en un área que resulta sorprendente, dado que ese tipo de cálculos no se inventaron para ese propósito en absoluto. También he buscado la variedad. No creo que hubiese tenido mucho sentido pasar el 90 % del libro tratando de las ecuaciones diferenciales parciales, pongamos por caso, aunque habría sido fácil encontrar material suficiente y justificar su importancia. Me interesaba reflejar la diversidad y el amplio alcance de los usos actuales de mi disciplina, además de demostrar su relevancia para la humanidad en su conjunto.

Para lavar mi conciencia, haré una breve mención a algunos de los cientos de otras aplicaciones de las que podría haber hablado. Estas incluso son solo la punta del iceberg. Mientras investigaba para este libro reuní un fichero, del que he tomado estos ejemplos. No siguen ningún orden en especial.

Predecir los niveles de las inundaciones.
Análisis de datos a gran escala y enfermedad de Lyme.
¿Cuántas veces hay que sacudir la botella para que salga el kétchup?
Cómo optimizar el empleo de la madera en aserraderos.
Cómo aislar mejor una casa o una tubería.
Detectar los prejuicios (de raza, de género) en algoritmos.
Rigidez de bastidores manufacturados, tales como estructuras metálicas para edificios.
Reconocimiento informático de células cancerígenas.
Mejorar la consistencia del espesor en la fabricación de láminas de vidrio.
Producción de dióxido de carbono cuando se fragua el hormigón.
Diseñar sistemas de llaves maestras para edificios de oficinas.
Modelos informáticos de corazones virtuales.
Diseñar construcciones resistentes a los huracanes.
Encontrar relaciones de descendencia entre especies.
Planear el movimiento de robots industriales.
La epidemiología de las enfermedades en el ganado.
Atascos.
Construir una red eléctrica adaptable al tiempo atmosférico.
Mejorar la resistencia de las comunidades frente a marejadas ciclónicas causadas por huracanes.
Cables de comunicación submarinos.
Detectar minas terrestres en países en los que ha terminado un conflicto.

Predecir el movimiento de cenizas volcánicas para ayudar a las líneas aéreas.
Reducir las oscilaciones de voltaje en la red eléctrica.
Mejorar la eficacia de las pruebas víricas durante la pandemia de la COVID-19.

Cada uno de estos temas merecería su propio capítulo. Aportan más ejemplos de la enorme diversidad de maneras en que se emplean las matemáticas para beneficio de todos los habitantes de este planeta.

*

Como demuestran estos ejemplos y los otros que he considerado con más detalle, la gran variedad de aplicaciones de las matemáticas es abrumadora, sobre todo cuando se tiene en cuenta que buena parte de ellas se desarrollaron en primer lugar, o bien con un objetivo diferente en mente, o bien solo porque algún investigador en alguna parte en algún momento pensó que podría ser interesante indagar. Una vez más, esto plantea el profundo problema filosófico que dejó perplejo a Wigner en 1959 y que sigue siendo (para mí, al menos) igual de desconcertante en la actualidad. Incluso más si acaso. Wigner se centraba sobre todo en la irrazonable eficacia de las matemáticas en la física teórica, pero ahora se descubre este mismo efecto en un conjunto de actividades humanas mucho más amplio y cercano. La mayoría de estas tienen una conexión todavía menos evidente con algo matemático.

Al igual que Wigner, no me convence la explicación que ofrecen muchas personas: las matemáticas se derivan del mundo real y por lo tanto deben ser eficaces en este. Como ya he dicho, me parece que no dan en el clavo, aunque consiguen explicar bastante bien la eficacia *razonable*. Las historias que he contado en *¿Para qué sirven las matemáticas?* ilustran algunas de las características que hacen que nuestra disciplina sea útil en áreas que en apariencia no están relacionadas con sus orígenes. El matemático y filósofo Benjamin Peirce definió las matemáticas como «la ciencia que deriva conclusiones necesarias». Si se cumple *esto* y *aquello*, ¿qué va a ocurrir? Esta es una cuestión muy general, común a la mayoría de los problemas que se presentan en el mundo externo. Dado que las matemáticas en la actualidad son muy generales, proporcionan un conjunto de herramientas útiles para responder estas cuestiones y solo están a la espera de que se las ponga a funcionar. No hace falta imaginar todos los usos posibles de un martillo para decidir que merece la pena tener uno. Ser capaz de unir cosas o de separarlas a golpes es una técnica general que con toda probabilidad será de utilidad en

muchas situaciones. Un martillo realiza una tarea, así que podría funcionar bien en otras. Un método matemático que se ha perfeccionado en una aplicación es susceptible a menudo de transferirse, con las modificaciones adecuadas, a otras.

Otra definición de matemáticas que me gusta es la de Lynn Arthur Steen: «La ciencia de forma significante». Las matemáticas tratan de la *estructura*, de cómo apoyarse en esta para comprender un problema. De nuevo, este punto de vista tiene una generalidad considerable y la experiencia demuestra que puede ir al meollo del asunto.

Una tercera definición, propuesta a la desesperada, es aquella según la cual las matemáticas son «lo que hacen los matemáticos». A la cual habría que añadir que un matemático es «alguien que hace matemáticas». Creo que es posible aspirar a algo mejor que esta tautología. ¿Acaso un negocio es «lo que hace un negociante» y un negociante es «alguien que hace negocios»? Sí, pero es más que eso. Lo que convierte a alguien en un emprendedor de éxito no es la actividad comercial en sí misma, sino darse cuenta de una *oportunidad* de negocio que los demás han pasado por alto. Del mismo modo, un matemático es alguien que se da cuenta de una oportunidad de hacer matemáticas que los demás han pasado por alto.

La forma de conseguirlo es pensar de manera matemática.

A lo largo de los siglos, los matemáticos han desarrollado reflejos del pensamiento que se dirigen al núcleo central de las cuestiones. ¿Cuál es el contexto natural del problema? ¿Cuál es el espacio de posibilidades? ¿Cuál es la estructura adecuada en la que expresar las propiedades relevantes? ¿Qué características son esenciales y cuáles son detalles sin importancia o distracciones prescindibles? ¿Cómo se eliminan? ¿Cuál es la estructura natural de lo restante? La comunidad matemática ha perfeccionado estos métodos para innumerables cuestiones difíciles, los ha destilado en teorías elegantes y potentes y ha comprobado estas con problemas del mundo real. Se han vuelto cada vez más generales, interconectadas, potentes y portátiles.

Tal vez la eficacia de las matemáticas no sea tan irrazonable.

Tal vez no sea desconcertante en absoluto.

*

Imaginemos un mundo sin matemáticas.

Ya oigo a muchas personas dando gritos de alegría. Las entiendo, porque no hay ningún motivo por el que algo que me resulta atractivo a mí deba serlo

también para ellas. Pero no me refiero a que puedan evitar tener que aprender la disciplina a un nivel personal. No se trata de ellas.

Supongamos que en el universo, grande, enorme, hay una civilización extraterrestre que consume cantidades ingentes de matemáticas. Lo digo en sentido literal. Algunos físicos afirman que las matemáticas explican el universo con una eficacia irrazonable porque el cosmos *está hecho* de matemáticas. Estas no serían un instrumento humano para entender cosas: serían reales, una sustancia etérea incorporada en todo lo que existe.

Personalmente, pienso que este punto de vista no tiene sentido y que trivializa el dilema filosófico, pero los extraterrestres saben que estoy equivocado. Hace mil millones de años, descubrieron que el universo está hecho en realidad de matemáticas. Y su civilización las consume en cantidades enormes, del mismo modo que lo hacemos los humanos con la mayoría de los recursos de la Tierra. De hecho, han consumido tanta que se habría agotado hace mucho de no haber sido por una sencilla solución. Su tecnología es avanzada en extremo y su actitud es agresiva en la misma medida, de modo que envían flotas de naves espaciales enormes por las estrellas, armadas hasta los dientes, en busca de nuevas formas de vida para servirse de sus matemáticas.

¡Que vienen los matematívoros!

Cuando llegan a un planeta nuevo se *comen* todas sus matemáticas. No solo las ideas, sino a la propia sustancia etérea, y desaparece todo lo que ha dependido alguna vez de la disciplina, al carecer de puntos de apoyo. Los matematívoros prefieren los alimentos más refinados, de modo que empiezan con las matemáticas avanzadas de verdad y las devoran en orden descendente hasta los aspectos más prosaicos. En general, se van cuando llegan a las multiplicaciones largas porque no les gusta el sabor de la aritmética básica, de modo que la civilización del planeta que atacan no desaparece por completo. No obstante, es una pálida sombra de su antiguo esplendor y la galaxia está plagada de planetas cuyos indígenas han vuelto a la Alta Edad Media, sin perspectiva alguna de poder escapar.

Si los matematívoros llegasen mañana, ¿qué perderíamos *nosotros*?

Es probable que no nos diésemos cuenta cuando desapareciesen las matemáticas puras en las fronteras del conocimiento. Aunque es probable que alguna parte de ellas llegue a ser imprescindible dentro de un siglo, no lo es ahora mismo. Pero conforme los matematívoros se abriesen camino hacia abajo desde las altas torres de marfil, empezarían a desaparecer cosas importantes. Las primeras serían los ordenadores, los teléfonos móviles e

internet, los productos más sofisticados desde un punto de vista matemático que hay en el planeta. A continuación, seguiría todo lo que implique vuelos espaciales: satélites meteorológicos, medioambientales y de comunicaciones, los navegadores y la televisión por satélite, la navegación aeronáutica y los observatorios de erupciones solares. Las centrales eléctricas dejarían de funcionar. Se detendrían los robots en las fábricas y las industrias se reducirían a la nada. Volveríamos a las escobas en lugar de los aspiradores. Nada de aviones a reacción: ya no pueden diseñarse sin ordenadores y hace falta la aerodinámica para saber cómo mantenerlos en vuelo. La radio y la televisión se desvanecerían en volutas de humo extraterrestre, porque son tecnologías que dependen de las ecuaciones de Maxwell para la radiación electromagnética, como las ondas de radio. Se desplomarían todos los edificios altos, porque su diseño y construcción se basan en gran medida en métodos informáticos y en la teoría de la elasticidad para garantizar la integridad estructural. Nada de rascacielos, ni de hospitales grandes, ni de estadios deportivos.

La historia va marcha atrás. Ya hemos retrocedido a la vida de hace un siglo y los matematívoros no han hecho más que empezar.

No cabe duda de que algunas pérdidas serían buenas: por ejemplo, las armas nucleares y la mayoría del resto de las aplicaciones militares de las matemáticas, aunque también perderíamos la capacidad de defendernos. La disciplina en sí misma es neutra: si algo es bueno o malo depende de lo que hagan las personas con ello.

Otras pérdidas serían más ambiguas: los bancos cancelarían todas las inversiones en la bolsa, por su incapacidad de predecir su comportamiento y, por lo tanto, de reducir al mínimo la incertidumbre financiera. A los banqueros no les gustan los riesgos excepto aquellos de los que no tienen constancia hasta que colapsa el sistema financiero. Esto reduciría nuestra obsesión autodestructiva con el dinero, pero también impediría que muchos proyectos útiles obtuvieran financiación.

La mayoría de las pérdidas serían malas. La predicción del tiempo volvería a consistir en chuparse un dedo y levantarlo para ver en qué dirección sopla el viento. La medicina perdería sus escáneres y la capacidad de hacer modelos de la difusión de epidemias, aunque conservaría la anestesia y los rayos X. Todo lo que dependa de la estadística sería cosa del pasado. Los médicos ya no podrían evaluar la seguridad y la eficacia de medicamentos y tratamientos nuevos. La agricultura perdería la capacidad de valorar nuevas variedades de plantas y animales. Las industrias ya no podrían

realizar controles de calidad eficaces, de modo que nada de lo que se comprase (del conjunto limitado de bienes que todavía estarían disponibles) sería fiable. Los gobiernos perderían la capacidad de predecir tendencias y demandas futuras. Es posible que no lo hagan de la mejor manera en cualquier caso, pero ahora sería mucho peor. Nuestras comunicaciones volverían a ser primitivas, ni siquiera habría telégrafo. Mandar cartas por mensajero a caballo sería lo más rápido que podríamos conseguir.

Llegados a este punto, sería imposible mantener la población humana actual. No funcionaria ninguno de los astutos trucos que empleamos para obtener más alimentos y para transportar bienes a través de los océanos. Habría que volver a los barcos de vela. Proliferarían las enfermedades y miles de millones de personas morirían de hambre. Habría llegado el final de los tiempos. El apocalipsis estaría a la vuelta de la esquina, conforme los escasos supervivientes se enfrentasen entre sí por lo poco que quedaría de nuestro mundo.

*

Habrá quien piense que este escenario es exagerado. Defenderé con vehemencia que lo único que he exagerado es la metáfora de las matemáticas como una sustancia comestible. Es cierto que dependemos de las matemáticas para casi todo lo que mantiene a nuestro planeta en funcionamiento. Las vidas cotidianas de las personas que piensan que las matemáticas son inútiles se basan, sin que lo sepan, en las actividades de aquellos que saben que esto no es cierto. No es en absoluto culpa suya: estas actividades se desarrollan entre bambalinas, donde nadie, excepto los especialistas, tiene alguna posibilidad de ser consciente de ellas.

No digo que «sin las matemáticas todavía viviríamos en cuevas» porque estoy seguro de que sin ellas habríamos encontrado otras formas de progresar. En absoluto pretendo que haya que atribuir los avances que se han logrado *solo* a las matemáticas. Estas alcanzan su mayor utilidad en conjunto con todo lo demás que la humanidad puede aplicar a los problemas a los que se enfrenta y de cara a los objetivos que se propone. Pero si estamos en el punto en que nos encontramos es porque las matemáticas, junto con todo los demás, nos han traído hasta aquí.

En la actualidad hemos incorporado las matemáticas de manera tan inextricable en nuestras estructuras tecnológicas y sociales que tendríamos graves problemas sin ellas.

En el primer capítulo mencioné seis características de las matemáticas: realidad, belleza, generalidad, portabilidad, unidad y diversidad. Defendí que todas juntas llevan a la utilidad. ¿Se sostiene esta afirmación ahora que se han leído los capítulos del uno al trece?

Muchas de las ideas matemáticas que he considerado tuvieron su origen en el mundo real. Los números, las ecuaciones diferenciales, el TSP, la teoría de grafos, la transformada de Fourier y el modelo de Ising. Las matemáticas encuentran su inspiración en la naturaleza y eso las hace mejores.

Otras ramas de la disciplina surgieron en buena medida por el mero sentido de la belleza que tienen los matemáticos. Los números complejos se inventaron porque no queda bonito que algunos números tengan dos raíces cuadradas y otros ninguna. La aritmética modular, las curvas elípticas y otras partes de la teoría de números se deben a que las personas lo pasaban bien buscando patrones numéricos. La transformada de Radon, una cuestión interesante en la geometría. La topología, que tuvo poca relación con la realidad durante un siglo, es central en el andamiaje matemático porque trata de la continuidad, que es fundamental.

El impulso a generalizar se encuentra en todas partes. Euler no solo resolvió el rompecabezas de los puentes de Königsberg, sino que lo hizo para todos los problemas del mismo tipo y creó una nueva rama de las matemáticas: la teoría de grafos. Los códigos basados en la aritmética modular llevaron a preguntas sobre complejidad computacional y si P = NP. Los números complejos fueron la inspiración de los cuaterniones de Hamilton. El análisis se generalizó al análisis funcional, que sustituye espacios con una cantidad finita de dimensiones por espacios de funciones con infinitas dimensiones y las funciones por funcionales y operadores. Los matemáticos inventaron los espacios de Hilbert de la teoría cuántica mucho antes de que los físicos les encontraran alguna utilidad. La topología empezó a jugar con cosas como las bandas de Möbius y se desarrolló de manera vertiginosa para formar una de las áreas más profundas y abstractas del pensamiento humano. En la actualidad también empieza a dar resultados en la vida cotidiana.

Muchos de los métodos que he descrito son portátiles, de modo que pueden usarse en todas partes sin que importe dónde tuvieron su origen. La teoría de grafos aparece en problemas médicos sobre trasplantes de riñón, en el TSP, en códigos cuánticos (grafos expansores) que pueden proteger nuestros datos frente a ataques con un computador cuántico, o en la capacidad de la navegación por satélite de seleccionar una ruta razonable. La

transformada de Fourier se ideó en primer lugar para estudiar el flujo de calor, pero sus primos incluyen la transformada de Radon, empleada en escáneres médicos, la transformada de coseno discreta, en compresión de imágenes JPEG, y en las ondículas que utiliza el FBI para almacenar de manera eficaz las huellas dactilares.

La unidad de las matemáticas es también un tema recurrente que se repite a lo largo de mis historias. La teoría de grafos confluye con suavidad en la topología. Los complejos aparecen en los problemas de teoría de números. La aritmética modular es la inspiración de la construcción de grupos de homología. La navegación por satélite reúne al menos cinco ramas distintas de las matemáticas en una aplicación, desde los números pseudoaleatorios hasta la relatividad. La dinámica ayuda a poner satélites en órbita y sugiere un nuevo método para el control de calidad del alambre para resortes.

¿Diversidad? Entre todos ellos, los capítulos de este libro cubren docenas de áreas diferentes de las matemáticas, por lo general en combinación. Van de lo numérico a lo geométrico, de los números irracionales a la botella de Klein, desde dividir una tarta de manera justa a los modelos climáticos. Probabilidad (cadenas de Márkov), grafos e investigación de operaciones (métodos Montecarlo) unen sus fuerzas para aumentar las oportunidades de los pacientes de obtener un trasplante de riñón.

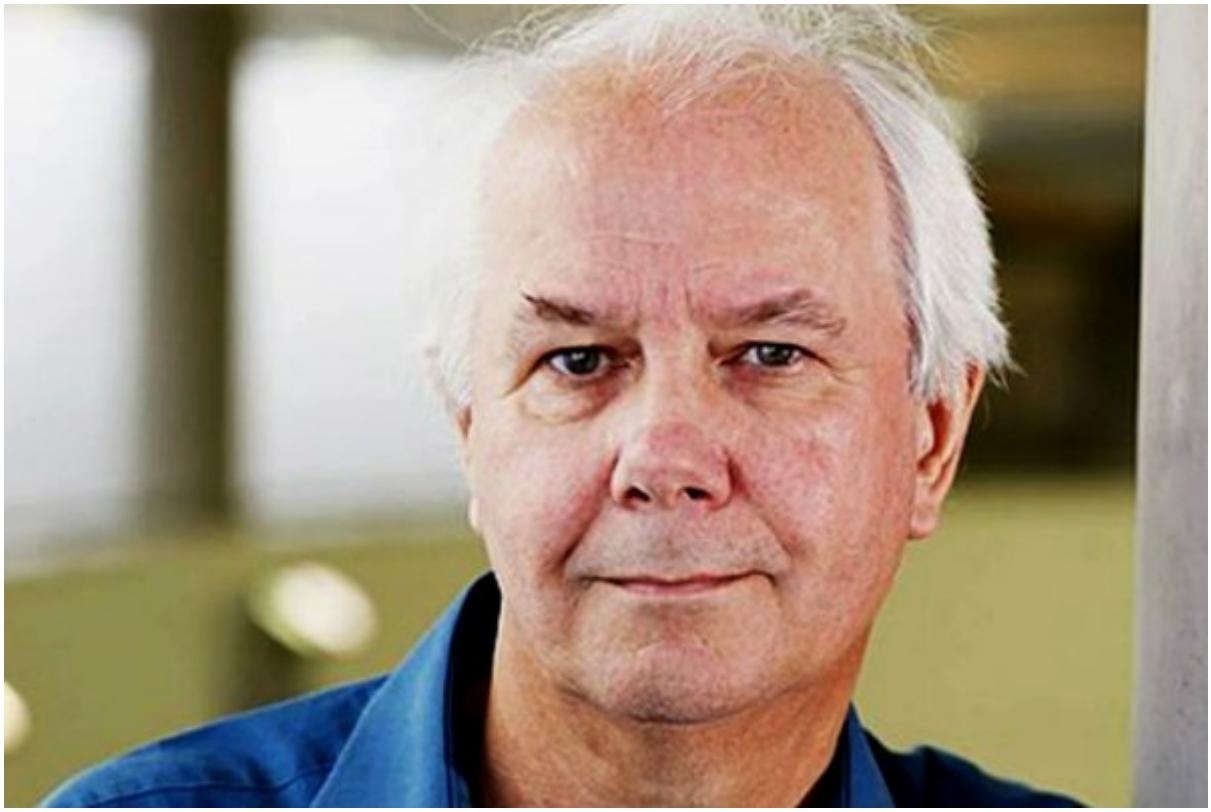
Por lo que se refiere a la utilidad: el conjunto de las aplicaciones es, si acaso, aún más diverso, desde la animación de películas hasta la medicina, desde la fabricación de resortes hasta la fotografía, del comercio online a las rutas aéreas, de los teléfonos móviles a los sensores de seguridad. Las matemáticas están en todas partes. Y solo he mostrado una parte minúscula de lo que hay dando vueltas por el mundo, sin ser visto ni oído. No tengo ni idea de lo que pueden ser la mayoría de las cosas. En cualquier caso, muchas de las mejores ideas son secretos comerciales.

Cuando llegue el momento de la verdad, necesitaremos a todas las personas posibles con una comprensión tan extensa de las matemáticas como se pueda por todos estos motivos. No solo por nuestro propio beneficio personal. Acepto que para la mayoría de nosotros, buena parte de las cosas que nos enseñan en matemáticas no tienen una utilidad directa. Pero eso pasa con todo. Yo estudié historia en el colegio y sin duda me permitió apreciar mejor la cultura en la que vivo, aparte de llenarme la cabeza de propaganda colonialista que hoy en día parece cada vez más llena de prejuicios. Pero no *empleo* la historia en mi trabajo ni en mi vida. La encuentro interesante (más cuanto más mayor me hago), me alegra que haya historiadores que la utilicen

y no se me ocurriría ni por asomo recomendar que no se enseñase. Pero la evidencia es indiscutible: las matemáticas son *imprescindibles* para la forma de vida actual. Es más, es muy difícil predecir qué partes podrían resultar útiles el día de mañana. Spencer, el solador que ponía azulejos en mi baño, no pensaba que π fuese muy útil, hasta que le hizo falta.

Las matemáticas, entendidas de manera adecuada como la disciplina rica y creativa que son en realidad, no como la caricatura grosera que muchas personas imaginan, es uno de los mayores logros de la humanidad. No solo a nivel intelectual, sino práctico también. Sin embargo, la condenamos a la oscuridad. Ha llegado el momento de sacarla a la luz, antes de que los equivalentes en el mundo real de mis matematívoros de ciencia ficción intenten arrebatárnosla.

Sí, el zorro sabe muchas cosas, pero los matemáticos sabemos una sola y grande. Se llama matemáticas y está cambiando la faz de nuestro mundo.



Ian Nicholas Stewart (24 de septiembre de 1945, Inglaterra) es profesor de matemática de la Universidad de Warwick, más conocido como escritor de ciencia ficción y de divulgación científica. Fue el primero en recibir, en 2008, la Medalla Christopher Zeeman, por sus numerosas actividades relacionadas con la divulgación matemática.

Ha publicado más de 140 *papers* científicos, entre ellos una serie de trabajos muy influyentes junto a Jim Collins sobre Osciladores acoplados y Sincronización biológica.

Stewart ha colaborado con el Dr. Jack Cohen y Terry Pratchett en cuatro libros de divulgación científica basados en el Mundodisco de Pratchett. Entre sus numerosos y reconocidos libros de divulgación matemática se encuentran *Historia de las Matemáticas*, *17 ecuaciones que cambiaron el mundo*, *Locos por las matemáticas* y *Cartas a una joven matemática*.

Notas

[1] En 2012, la empresa de auditoría Deloitte realizó un sondeo titulado *Measuring the Economic Benefits of Mathematical Science Research in the UK*. En esa fecha, había 2,8 millones de personas empleadas en profesiones relacionadas con la ciencia de las matemáticas: tanto puras como aplicadas, estadística y ciencias de la computación. Su contribución a la economía de Reino Unido ese año se estimó en 208 000 millones de libras esterlinas (en valor bruto agregado), algo menos de 250 000 millones ajustadas a dinero de 2020, cerca de 290 000 millones de euros. Esas 2,8 millones de personas constituían el 10 % de la población activa del país y suponían el 16 % de su economía. Las áreas principales en que se ocupaban eran la banca, la investigación y el desarrollo industriales, los servicios informáticos, el sector aeroespacial, las farmacéuticas, la arquitectura y la construcción. Los ejemplos mencionados en el informe incluían los teléfonos inteligentes, el pronóstico del tiempo, la sanidad, los efectos especiales en el cine, la mejora del rendimiento deportivo, la seguridad nacional, la gestión de epidemias, la seguridad de los datos en internet y mejorar la eficacia de los procesos industriales. <<

[2] maths.ed.ac.uk/~v1ranick/papers/wigner.pdf. [hay trad. cast. accesible en u-cursos.cl/escverano/2006/3/269/1/material_docente/bajar%3Fid_material%3D118053]. <<

[3] La fórmula es

$$\frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)}$$

donde *x* es el valor de la variable aleatoria, *μ* es la media y σ es la desviación estándar. <<

[4] Vito Volterra era matemático y físico. En 1926, un biólogo marino, Umberto D'ancona, cortejaba a su hija, con la que se casó después. D'ancona había descubierto que la proporción de peces depredadores (tiburones, rayas y pez espada) capturados por los pescadores durante la primera guerra mundial había aumentado, a pesar de que en conjunto se pescaba menos. Volterra desarrolló un modelo sencillo basado en el cálculo sobre los cambios en el tiempo de las poblaciones de depredadores y de presas, que demostraba que el sistema gira en torno a ciclos que se repiten de aumentos explosivos de depredadores y colapsos del número de presas. De manera crucial, la cantidad de depredadores aumenta *en promedio*, más que la de presas. <<

[5] No cabe duda de que Newton también empleó su intuición física y los historiadores afirman que es muy probable que tomara la idea prestada de Robert Hooke. Sin embargo, no tiene sentido darle vueltas al molino. <<

[6] Traducción de Celia Filipetto Isicato, *Mort*, Debolsillo, Barcelona, 2010. <<

[7] theguardian.com/commentisfree/2014/oct/09/virginia-gerrymandering-voting-rights-act-black-voters. <<

[8] El tiempo que se tardaba no era la única pega. En la Convención Constitucional de 1787, de la que surgió el sistema del colegio electoral, aunque bajo un nombre distinto, James Wilson, James Madison y otros eran de la opinión de que el voto popular era mejor. No obstante, decidir quién tenía derecho al sufragio presentaba problemas prácticos y había grandes diferencias de opinión entre los estados del Norte y del Sur. <<

[9] En 1927, E. P. Cox empleó la misma cantidad en paleontología para evaluar la redondez de los granos de arena, lo que ayuda a distinguir los que ha transportado el viento de los que ha traído el agua y aporta pruebas de las condiciones medioambientales en tiempos prehistóricos. Véase Cox, E. P., «A method of assigning numerical and percentage values to the degree of roundness of sand grains», *Journal of Paleontology* I, 1927, pp. 179-183. En 1966, Joseph Schwartzberg propuso emplear la proporción del perímetro de un distrito con respecto a la longitud de la circunferencia de un círculo de área equivalente. Esta se corresponde con la raíz cuadrada de la puntuación de Polsby-Popper, por lo que clasifica los distritos del mismo modo, aunque con valores diferentes. Véase Schwartzberg, J. E., «Reapportionment, gerrymanders and the notion of "compactness"», *Minnesota Law Review*, n.º 50, 1966, pp. 443-566. <<

[10] Al rodear una colina, que es una superficie curva, consiguió encerrar aún más área dentro de su circunferencia. <<

[11] Blåsjö, V., «The isoperimetric problem», *American Mathematical Monthly*, n.º 112, 2005, pp. 526-566. <<

[12] Para un círculo de radio r,

la longitud de la circunferencia (= perímetro) = $2\pi r$

área = πr^2

perímetro^2 = $(2\pi r)^2 = 4\pi^2 r^2 = 4\pi\ (\pi r^2) = 4\pi \times$ el área. <<

[13] Stephanopoulos, N. y McGhee, E., «Partisan gerrymandering and the efficiency gap», *University of Chicago Law Review*, n.º 82, 2015, pp. 831-900. <<

[14] Bernstein, M. y Duchin, M., «A formula goes to court: partisan gerrymandering and the efficiency gap», *Notices of the American Mathematical Society*, n.º 64, 2017, pp. 1020-1024. <<

[15] Barton, T. J., «Improving the efficiency gap», *Math Horizons*, n.º 26.1, 2018, pp. 18-21. <<

[16] A principios de la década de 1960, John Selfridge y John Horton Conway descubrieron de manera independiente un método a prueba de envidia para dividir una tarta entre tres participantes:

1. Alice corta la tarta en tres porciones que ella considera de igual valor.
2. Bob puede dejar pasar su turno, si piensa que dos o más porciones son igual de grandes, o bien recorta la que en su opinión es la más grande para que se cumpla la condición anterior. Lo que ha recortado se denomina «sobras» y se deja a un lado.
3. Charlie, Bob y Alice, por ese orden, eligen el pedazo que les parece más grande o igual en tamaño al más grande. Si Bob no dejó pasar su turno en 2, debe elegir la porción recortada, a no ser que Charlie la haya elegido antes.
4. Si Bob dejó pasar su turno en 2, no hay sobras y se ha acabado el proceso. Si no es así, o Bob o Charlie se ha quedado con la porción recortada. Esta persona será «la que no ha cortado» y la otra será «la que ha cortado». Esta última divide las sobras en tres partes que le parezcan iguales.
5. Los tres participantes eligen una de estas partes en el orden siguiente: «La que no ha cortado», Alice y «la que ha cortado». Ninguno de ellos tiene motivos para envidiar lo que obtienen los demás. Si lo hace, es porque se equivocó de táctica y debería haber elegido de otro modo. Se puede ver una demostración en inglés en: en.wikipedia.org/wiki/Selfridge-Conway_procedure. <<

[17] Brams, S. J. y Taylor, A. D., *The win-win solution: guaranteering fair shares to everybody*, Norton, Nueva York, 1999. <<

[18] Landau, Z., Reid, O. y Yershov, I., «A fair division solution to the problem of redistricting», *Social Choice and Welfare*, n.º 32, 2009, pp. 479-492. <<

[19] Alexeev, B. y Mixon, D. G., «An impossibility theorem for gerrymandering», *American Mathematical Monthly*, n.º 125, 2018, pp. 878-884. <<

[20] Gibson, B., Wilkinson, M. y Kelly, D., «Let the pigeon drive the bus: pigeons can plan future routes in a room», *Animal Cognition*, n.º 15, 2012, pp. 379-391. <<

[21] Mi ejemplo favorito es el de un político que montó un escándalo enorme por el dinero que se estaba malgastando en lo que él llamaba la «teoría de la mentira» (*lie* en inglés), que era de lo que pensaba que se ocupaba esta. No es así. Sophus Lie fue un matemático noruego cuyas contribuciones en grupos continuos de simetrías (grupos de Lie) y sus álgebras asociadas (adivinen el nombre) son fundamentales en buena parte de las matemáticas y más aún de la física. La confusión del político se hizo evidente enseguida… pero él siguió adelante *como si nada*. <<

[22] Por motivos técnicos, mi observación sobre el rompecabezas no resuelve el problema del milenio. Pero si lo hiciera, ¡yo lo dije primero! <<

[23] Juego de palabras intraducible entre *sequins* (lentejuelas) y *sequence* (serie), cuya pronunciación es muy parecida en inglés. (*N. del t.*) <<

[24] Garey, M. R. y Johnson, D. S., *Computers and intractability: a guide to the theory of NP-completeness*, Freeman, San Francisco, 1979. <<

[25] Peano, G., «Sur une courbe qui remplit toute une aire plane», *Mathematische Annalen*, n.º 36, 1890, pp. 157-160. <<

[26] Hay que tomar algunas precauciones porque ciertos números reales no tienen una representación única como decimales. Por ejemplo 0,5000000… = 0,4999999… Pero es un problema fácil de sortear. <<

[27] Netto, E., «Beitrag zur Mannigfaltigkeitslehre», *Journal fur die Reine und Angewandte Mathematik*, n.º 86, 1879, pp. 263-268. <<

[28] Sagan, H. «Some reflections on the emergence of space-filling curves: the way it could have happened and should have happened, but did not happen», *Journal of the Franklin Institute*, n.º 328, 1991, pp. 419-430. Se puede ver una explicación en Jaffer, A., «Peano space-filling curves», people.csail.mit.edu/jaffer/Geometry/PSFC. <<

[29] Lawder, J., *The application of space-filling curves to the storage and retrieval of multi-dimensional data*, tesis doctoral, Birkbeck College, Londres, 1999. <<

[30] Bartholdi, J., *Some combinatorial applications of spacefilling curves* en www2.isye.gatech.edu/~jjb/research/mow/mow. <<

[31] Hahn, H., «Über die allgemeinste ebene Punktmenge, die stetiges Bild einer Strecke ist», *Jahresbericht der Deutschen Mathematiker-Vereinigung*, n.º 23, 1914, pp. 318-322. Hah, H., «Mengentheoretische Charakterisierung der stetigen Kurven», *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften, Wien*, n.º 123, 1914, pp. 2433-2489. Mazurkiewicz, S., «O aritmetzacji kontinuów», *Comptes Rendus de la Societe Scientifique de Varsovie*, n.° 6, 1913, pp. 305-311 y 941-945. <<

[32] Publicado en 1998. Arora, S., Sudan, M., Motwani, M., Lund, C. y Szegedy, M., «Proof verification and the hardness of approximation problems», *Journal of the Association for Computing Machinery*, n.° 45, 1998, pp. 501-555. <<

[33] Babai, L., «Transparent proofs and limits to approximation» en *First European Congress of Mathematics. Progress in Mathematics 3* (Joseph, A., Mignot, F., Murat, F., Prum, B. y Rentschler, R., eds.), Birkhäuser, Basilea, 1994, pp. 31-91. <<

[34] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. y Fergus, R., «Intriguing properties of neural networks» en arXiv:1312.6199, 2013. <<

[35] Shamir, A., Safran, I., Ronen, E. y Dunkelman, O., «A simple explanation for the existence of adversarial examples with small Hamming distance» en arXiv:1901.10861v1 [cs.LG], 2019. <<

[36] No deben confundirse estos con la gráfica de una función, que es una curva que relaciona una variable x con el valor $f(x)$ de la función. Como la parábola para $f(x) = x^2$. <<

[37] Agradezco los amables comentarios de Robin Wilson al señalar el error una vez que lo conté mal en uno de mis libros. <<

[38] Siempre que se sepa en qué zona se empieza, basta con hacer una lista de los símbolos de los puentes en el orden en el que se cruzan. Los puentes consecutivos determinan una zona común a la que ambos están conectados. <<

[39] Es algo bastante fácil de demostrar si se emplea la caracterización de Euler de un recorrido abierto. La idea principal es dividir un circuito cerrado hipotético al eliminar un puente. Así se obtiene un circuito abierto y el puente modificado en primer lugar une los dos extremos. <<

[40] El resto de este capítulo sigue a Manlove, D., «Algorithms for kidney donation», *London Mathematical Society Newsletter*, n.º 475, marzo de 2018, pp. 19-24. <<

[41] Traducción de Jesús Fernández Díez, *Apología de un matemático*, Nivola, Madrid, 1999. <<

[42] No se conoce con certeza la fecha exacta en la que Fermat enunció su último teorema, pero a menudo se considera que fue en 1637. <<

[43] Puede decirse también lo mismo de muchas matemáticas «aplicadas». No obstante, hay una diferencia: la actitud de los matemáticos. Las matemáticas puras están impulsadas por la lógica interna de la disciplina: no solo por una curiosidad animal, sino por la intuición de una estructura y de las partes en las que nuestra comprensión presenta carencias significativas. Las matemáticas aplicadas se ven impulsadas sobre todo por problemas que surgen en el «mundo real», pero tiene mayor predisposición a tolerar atajos sin justificar y aproximaciones en su búsqueda de una solución y esta puede tener aplicaciones prácticas o no. Como ilustra este capítulo, sin embargo, un tema que parece ser inútil por completo en algún momento de la historia puede volverse en vital de repente para aplicaciones prácticas cuando cambian la cultura o la tecnología. Es más, las matemáticas son un conjunto interconectado. Incluso la distinción entre puras y aplicadas es artificial. Un teorema que parece inútil en sí mismo puede inspirar, o incluso implicar, resultados de gran utilidad. <<

[44] La respuesta es:

p = 12 277 385 900 723 407 383 112 254 544 721 901 362 713 421 995 519

q = 97 117 113 276 287 886 345 399 101 127 363 740 261 423 928 273 451

Encontré estos dos números primos mediante prueba y error y los multipliqué entre sí empleando un sistema de álgebra simbólica en un ordenador. Llevó unos pocos minutos, sobre todo para cambiar los dígitos de manera aleatoria hasta que me tropecé con un primo. Entonces le di instrucciones al ordenador para que encontrarse los factores del producto y los buscó durante mucho tiempo sin resultado. <<

[45] Si n es la potencia de un número primo p^k, entonces $\varphi(n) = p^k - p^{k-1}$. Para un producto de potencias de primos, su valor se obtiene de multiplicar estas expresiones entre sí para todas las potencias de primos diferentes en la factorización en primos de n. Por ejemplo, para encontrar $\varphi(675)$, se escribe $675 = 3^3 5^2$. Luego

$\varphi(675) = (3^3 - 3^2)(5^2 - 5) = (18)(20) = 360$. <<

[46] Para ver más detalles acerca de los problemas que esto implica, véase Stewart, Ian, *¿Juega Dios a los dados?*, Crítica, Barcelona, 2007, capítulos 15 y 16. <<

[47] Vandersypen, L. M. K, Steffen, M., Breyta, G., Yannoni, C. S., Sherwood, M. H. y Chuang, I. L., «Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance», *Nature*, n.º 414, 2001, pp. 883-887. <<

[48] Arute, F. *et al*., «Quantum supremacy using a programmable superconducting processor», *Nature*, n.º 574, 2019, pp. 505-510. <<

[49] Proos, J. y Zalka, C., «Shor's discrete logarithm quantum algorithm for elliptic curves», *Quantum Information and Computation*, n.º 3, 2003. <<

[50] Roetteler, M., Naehrig, M., Svore, K. y Lauter. K., «Quantum resource estimates for computing elliptic curve discrete logarithms», en *ASIACRYPT 2017: Advances in Cryptology*, Springer, Nueva York, 2017, pp. 214-270. <<

[51] Por ejemplo, −25 tiene una raíz cuadrada 5i, porque

$(5i)^2 = 5i \times 5i = 5 \times 5 \times i \times i = 25i^2 = 25(-1) = -25$

De hecho, tiene una segunda raíz cuadrada, −5i, por motivos similares. <<

[52] Los expertos en álgebra normalizan la situación al decir que la raíz cuadrada de cero es cero con *multiplicidad* dos. Es decir, el mismo valor ocurre dos veces en un sentido significativo pero técnico. Una expresión como $x^2 - 4$ tiene dos factores, $(x + 2)$ multiplicado por $(x - 2)$, que llevan a dos soluciones respectivas, $x = -2$ y $x = +2$ de la ecuación $x^2 - 4 = 0$. De manera similar, la expresión x^2 tiene dos factores, x por x. Tan solo ocurre que son el mismo. <<

[53] Para un número real c, la función $z(t) = e^{ct}$ obedece a la ecuación diferencial $dz/dt = cz$, con la condición inicial de que $z(0) = 1$. Si se define la función exponencial para un complejo c de modo que se cumpla la misma ecuación, lo que es razonable, y se hace $c = i$, entonces $dz/dt = iz$. Dado que multiplicar por i rota 90° los números complejos, la tangente de $z(t)$ conforme varía t es perpendicular a $z(t)$, de modo que el punto $z(t)$ describe una circunferencia de radio 1 centrada en el origen. Gira a lo largo de esta circunferencia a una velocidad constante de un radián por unidad de tiempo, de modo que en el instante t su posición está en un ángulo de t radianes. Por trigonometría, este punto es $\cos t + i \operatorname{sen} t$. <<

[54] Más en concreto, debe haber un «producto interno» que determine distancias y ángulos. <<

[55] El ordenador más potente en 1988 era el Cray Y-MP, que costaba 20 millones de dólares (más de 50 millones de dólares al valor actual). No lo tendría fácil para ejecutar el sistema operativo de Windows. <<

[56] Shoemake, K., «Animating rotation with quaternion curves», *Computer Graphics*, n.º 19, 1985, pp. 245-254. <<

[57] Euler, L., «Découverte d’un nouveau principe de mécanique» (1752) en *Opera Omnia, Series Secunda* 5, Orel Fusili Turici, Lausana, 1957, pp. 81-108. <<

[58] La propiedad del medio ángulo es importante en mecánica cuántica, en la que una formulación del espín cuántico está basada en los cuaterniones. Si la función de onda de una partícula del tipo conocido como un fermión se rota 360°, su espín pasa a ser el opuesto (esto es algo distinto a rotar la partícula en sí misma). La función de onda debe rotar 720° para que el espín vuelva a su valor original. Los cuaterniones unidad forman una «versión doble» de las rotaciones. <<

[59] Brandt, C., von Tycowicz, C. y Hildebrandt, K., «Geometric flows of curves in shape space for processing motion of deformable objects», *Computer Graphics Forum*, n.º 35, 2016, pp. 295-305. <<

[60] syfy.com/syfy-wire/it-took-more-cgi-than-you-think-to-bring-carrie-fisher-into-the-rise-of-skywalker. <<

[61] Takagi, T. y Sugeno, M., «Fuzzy identification of systems and its application to modeling and control», *IEEE Transactions on Systems, Man, and Cybernetics*, n.º 15, 1985, pp. 116-132. <<

[62] Este es el código JFIF, empleado en la web. El código Exif para cámaras también incluye «metadatos», que describen los ajustes de la cámara, tales como la fecha, la hora y la exposición. <<

[63] Jain, A. y Pankanti, S., «Automated fingerprint identification and imaging systems» en *Advances in Fingerprint Technology* (C. Lee y R. E. Gaensslen, eds.), CRC Press, 2001, pp. 275-326. <<

[64] Ashby, N., «Relativity in the Global Positioning System», *Living Reviews in Relativity*, n.º 6, 2003, 1. Doi: 10.12942/lrr-2003-1. <<

[65] De manera más precisa, $Z = \sum \exp(-\beta H)$, donde la suma se extiende a todas las configuraciones de variables de espín. <<

[66] Definiendo $\beta = 1/k_{\mathrm{B}}T$, donde k_{B} es la constante de Boltzmann, la fórmula es:

$$g(T,H) = -\frac{1}{\beta}\log\left[\mathrm{e}^{\beta J}\cosh(\beta H) + \sqrt{\mathrm{e}^{2\beta J}\cosh^2(\beta H) - 2\mathrm{senh}(2\beta J)}\right]$$

<<

[67] La fórmula es:

$$\frac{\operatorname{senh}(\beta H)}{\sqrt{\operatorname{senh}^2(\beta H) + \exp(-4\beta J)}}$$

donde *H* es la fuerza del campo externo y *J* la de las interacciones entre los espines. En ausencia de un campo externo, $H = 0$, de modo que senh (βH) = 0, por lo que toda la fracción es 0. <<

[68] Ma, Y. P., Sudakov, I., Strong, C. y Golden, K. M., «Ising model for melt ponds on Arctic sea ice», *New Journal of Physics*, n.º 21, 2019, 063029. <<

[69] Tanaka, S., «Topological analysis of point singularities in stimulus preference maps of the primary visual cortex», *Proceedings of the Royal Society of London*, n.º B 261, 1995, pp. 81-88. <<

[70] «Lobster telescope has an eye for X-rays» en www.sciencedaily.com/releases/2006/04/060404194138. <<

[71] Desde un punto de vista técnico, la curva es la *imagen*, bajo un mapeado de un disco a la esfera, de la frontera del disco. La curva puede cruzarse a sí misma y el disco estrujarse. <<

[72] Berwald, J. J., Gidea, M. y Vejdemo-Johansson, M., «Automatic recognition and tagging of topologically different regimes in dynamical systems», *Discontinuity, Nonlinearity and Complexity*, n.º 3, 2014, pp. 413-426. <<

[73] Khasawneh, F. A. y Munch, E., «Chatter detection in turning using persistent homology», *Mechanical Systems and Signal Processing*, n.º 70, 2016, pp. 527-541. <<

[74] Tralie, C. J. y Perea, J. A., «(Quasi) periodicity quantification in video data, using topology», *SIAM Journal on Imaging Science*, n.º 11, 2018, pp. 1049-1077. <<

[75] Emrani, S., Gentimis, T. y Krim, H., «Persistent homology of delay embeddings and its application to wheeze detection», *IEEE Signal Processing Letters*, n.º 21, 2014, pp. 459-463. <<