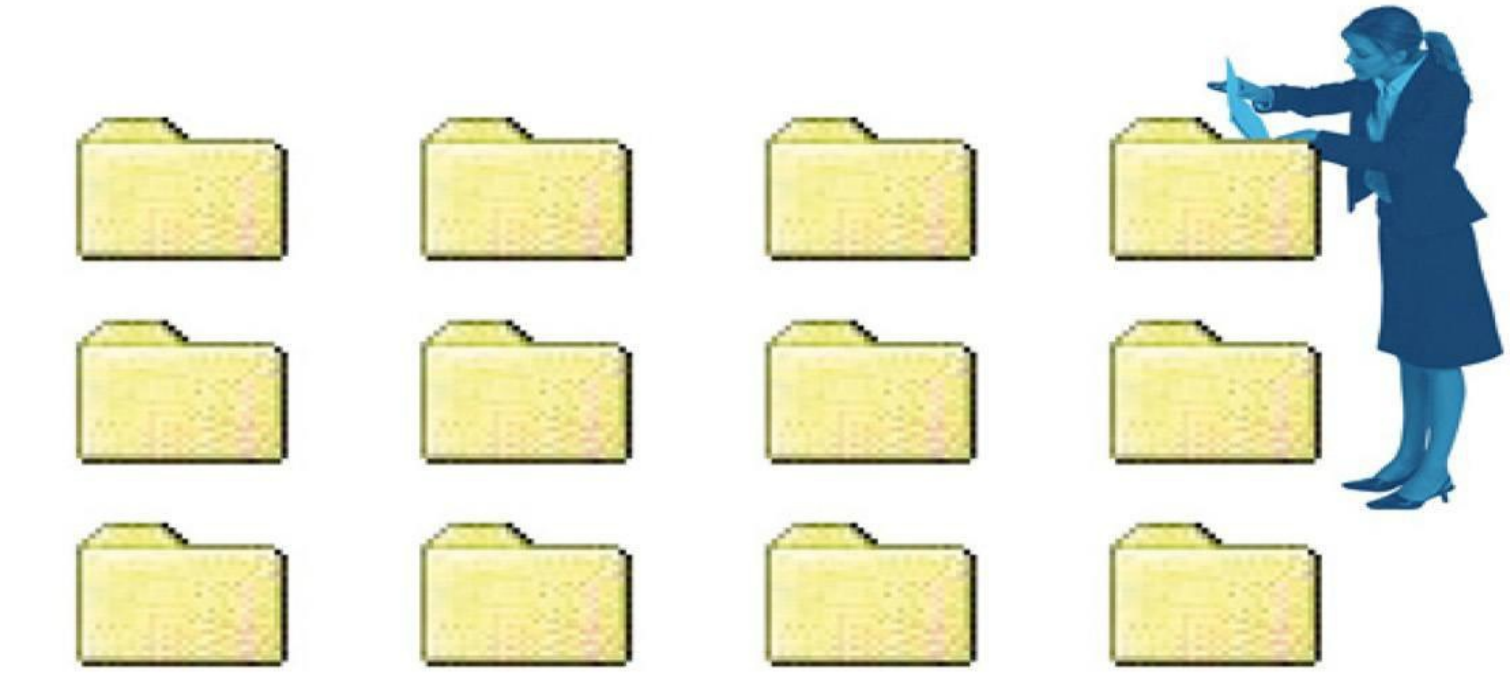
Big data

La revolución de los datos masivos

VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER

TURNER NOEMA



Big data

La revolución de los datos masivos

VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER

TRADUCCIÓN DE ANTONIO IRIARTE

COLECCIÓN NOEMA

Título:
Big Data. La revolución de los datos masivos

Edición original: *Big Data. A Revolution That Will Transform How We Live, Work, and Think*, 2013. Publicado por acuerdo especial con la editorial Houghton Mifflin Harcourt

De esta edición:

Rafael Calvo, 42
28010 Madrid
www.turnerlibros.com
Primera edición: junio de 2013

ISBN: 978-84-15427-81-0

Diseño de la colección:
Enric Satué

Ilustración de cubierta:
Enric Jardí

La editorial agradece todos los comentarios y observaciones:
turner@turnerlibros.com

I
AHORA

En 2009 se descubrió un nuevo virus de la gripe. La nueva cepa, que combinaba elementos de los virus causantes de la gripe aviar y la porcina, recibió el nombre de H1N1 y se expandió rápidamente. En cuestión de semanas, los organismos de sanidad pública de todo el mundo temieron que se produjera una pandemia terrible. Algunos comentaristas alertaron de un brote similar en escala al de la gripe española de 1918, que afectó a quinientos millones de personas y causó decenas de millones de muertes. Además, no había disponible ninguna vacuna contra el nuevo virus. La única esperanza que tenían las autoridades sanitarias públicas era la de ralentizar su propagación. Ahora bien, para hacerlo, antes necesitaban saber dónde se había manifestado ya.

En Estados Unidos, los Centros de Control y Prevención de Enfermedades (CDC) pedían a los médicos que les alertaran ante los casos nuevos de gripe. Aun así, el panorama de la pandemia que salía a la luz llevaba siempre una o dos semanas de retraso. Había gente que podía sentirse enferma durante días antes de acudir al médico. La transmisión de la información a las organizaciones centrales tomaba su tiempo, y los CDC solo tabulaban las cifras una vez por semana. Con una enfermedad que se propaga cada vez más deprisa, un desfase de dos semanas es una eternidad. Este retraso ofuscó por completo a los organismos sanitarios públicos en los momentos más cruciales.

Unas cuantas semanas antes de que el virus H1N1 ocupase los titulares, dio la casualidad de que unos ingenieros de Google[1], el gigante de internet, publicaron un artículo notable en la revista científica *Nature*. Esta pieza causó sensación entre los funcionarios de sanidad y los científicos de la computación, pero, por lo demás, pasó en general inadvertida. Los autores explicaban en ella cómo Google podía "predecir" la propagación de la gripe invernal en Estados Unidos, no solo en todo el ámbito nacional, sino hasta por regiones específicas, e incluso por estados. La compañía lo conseguía estudiando qué buscaba la gente en internet. Dado que Google recibe más de tres mil millones de consultas a diario y las archiva todas, tenía montones de datos con los que trabajar.

Google tomó los cincuenta millones de términos de búsqueda más corrientes empleados por los estadounidenses y comparó esa lista con los datos de los CDC sobre propagación de la gripe estacional entre 2003 y 2008. La intención era identificar a los afectados por el virus de la gripe a través de lo que buscaban en internet. Otros ya habían intentado hacer esto con los términos de búsqueda de internet, pero nadie disponía de tantos datos, capacidad de procesarlos y *know-how* estadístico como Google.

Aunque el personal de Google[2] suponía que las búsquedas podrían centrarse en obtener información sobre la gripe –tecleando frases como "remedios para la tos y la fiebre"–, no era esa la cuestión: como no les constaba, diseñaron un sistema al que no le importaba. Lo único que hacía este sistema era buscar correlaciones entre la frecuencia de ciertas búsquedas de información y la propagación de la gripe a lo largo del tiempo y del espacio. Procesaron un total apabullante de cuatrocientos cincuenta millones de modelos matemáticos diferentes para poner a prueba los términos de búsqueda, comparando sus predicciones con los casos de gripe registrados por los CDC

en 2007 y 2008. Así dieron con un filón: su software halló una combinación de cuarenta y cinco términos de búsqueda que, al usarse conjuntamente en un modelo matemático, presentaba una correlación fuerte entre su predicción y las cifras oficiales de la enfermedad a lo largo del país. Como los CDC, podían decir adónde se había propagado la gripe, pero, a diferencia de los CDC, podían hacerlo en tiempo casi real, no una o dos semanas después.

Así pues, en 2009, cuando estalló la crisis del H1N1, el sistema de Google demostró ser un indicador más útil y oportuno que las estadísticas gubernamentales, con su natural desfase informativo. Y los funcionarios de la sanidad pública consiguieron una herramienta de información incalculable.

Lo asombroso del método de Google es que no conlleva distribuir bastoncitos para hacer frotis bucales, ni ponerse en contacto con las consultas de los médicos. Por el contrario, se basa en los *big data*, los "datos masivos": la capacidad de la sociedad de aprovechar la información de formas novedosas, para obtener percepciones útiles o bienes y servicios de valor significativo. Con ellos, cuando se produzca la próxima pandemia, el mundo dispondrá de una herramienta mejor para predecir, y por ende prevenir, su propagación.

La sanidad pública no es más que una de las áreas en las que los datos masivos están suponiendo un gran cambio. Hay sectores de negocio completos que se están viendo asimismo reconfigurados por los datos masivos. Un buen ejemplo nos lo brinda la compra de billetes de avión.[3]

En 2003, Oren Etzioni tenía que volar de Seattle a Los Ángeles para asistir a la boda de su hermano pequeño. Meses antes del gran día, entró en internet y compró un billete, creyendo que cuanto antes reserves, menos pagas. En el vuelo, la curiosidad pudo más que él, y le preguntó al ocupante del asiento contiguo cuánto había costado su billete, y cuándo lo había comprado. Resultó que el hombre hacía pagado considerablemente menos que Etzioni, aun cuando había comprado el billete mucho más tarde. Furioso, Etzioni le preguntó a otro pasajero, y luego a otro más. La mayor parte habían pagado menos que él.

A la mayoría, la sensación de haber sido traicionados económicamente se nos habría disipado antes de plegar las bandejas y colocar los asientos en posición vertical. Etzioni, sin embargo, es uno de los principales científicos estadounidenses de la computación. Ve el mundo como una serie de problemas de datos masivos: problemas que puede resolver. Y ha estado dominándolos desde el día en que se licenció en Harvard, en 1986, siendo el primer estudiante que se graduaba en ciencias de la computación.

Desde su puesto en la universidad de Washington, Etzioni impulsó un montón de compañías de datos masivos antes incluso de que se diese a conocer el término. Ayudó a crear uno de los primeros buscadores de la red, MetaCrawler, que se lanzó en 1994 y acabó siendo adquirido por InfoSpace, por entonces una firma online importante. Fue cofundador de Netbot, la primera gran página web de comparación de precios, que luego vendió a Excite. Su firma *start up*, o emergente, para extraer sentido de los documentos de texto, llamada ClearForest, fue posteriormente adquirida por Reuters.

Una vez en tierra, Etzioni estaba decidido a encontrar la forma de que la gente pudiese saber si el precio del billete de avión que ve en internet es buen negocio o no. Un asiento en un avión es un producto: cada uno es básicamente indistinguible de los demás en el mismo vuelo. Sin embargo, los precios varían de forma brutal, al estar basados en una multitud de factores que, esencialmente, solo conocen las líneas aéreas.

Etzioni llegó a la conclusión de que no necesitaba descifrar la causa última de esas diferencias. Le bastaba con predecir si el precio mostrado tenía probabilidades de aumentar o disminuir en el futuro. Eso es algo posible, aunque no fácil de hacer. Basta con analizar todas las ventas de billetes de avión para una ruta determinada y examinar los precios pagados en función del número de días que faltan para el viaje.

Si el precio medio de un billete tendiese a disminuir, tendría sentido esperar y comprarlo más adelante. Si el precio medio aumentase habitualmente, el sistema recomendaría comprar el billete de inmediato. En otras palabras, lo que se precisaba era una versión potenciada de la encuesta informal que Etzioni había llevado a cabo a 9.000 metros de altitud. Por descontado, se trataba de otro problema descomunal para la ciencia informática, pero también de uno que podía resolver. Así que se puso a trabajar.

Usando una muestra de doce mil registros de precios de vuelos, recabada a través de una web de viajes a lo largo de un periodo de cuarenta y un días, Etzioni creó un modelo predictivo que ofrecía a sus pasajeros simulados un ahorro estimable. El modelo no ofrecía ninguna explicación del *porqué*, solo del *qué*. Es decir, no conocía ninguna de las variables que intervienen en la fijación de precios de las líneas aéreas, como el número de asientos sin vender, la estacionalidad, o si de alguna forma mágica la pernoctación durante la noche del sábado podría reducir el importe. Basaba su predicción en lo que sí sabía: probabilidades recopiladas de datos acerca de otros vuelos. "Comprar o no comprar, esa es la cuestión", se dijo Etzioni. Por consiguiente, denominó Hamlet a su proyecto.[4]

Ese pequeño proyecto evolucionó hasta convertirse en una empresa *start up* financiada con capital-riesgo y de nombre Farecast. Al predecir si era probable que subiera o bajara el precio de un billete de avión, y cuánto, Farecast les atribuyó a los consumidores el poder de elegir cuándo hacer clic en el botón de "comprar". Los armó con una información a la que nunca antes habían tenido acceso. Ensalzando las virtudes de la transparencia a sus expensas, Farecast incluso puntuaba el grado de confianza que le merecían sus propias predicciones y les brindaba a los usuarios también esa información.

Para funcionar, el sistema necesitaba montones de datos, así que Etzioni intentó mejorarlo haciéndose con una de las bases de datos de reservas de vuelos de la industria aérea. Con esa información, el sistema podía hacer predicciones basadas en todos los asientos de todos los vuelos, en la mayoría de las rutas de la aviación comercial estadounidense, en el transcurso de un año. Farecast estaba procesando ya cerca de doscientos mil millones de registros de precios de vuelos para realizar sus predicciones. Y, con ello, estaba permitiéndoles a los consumidores ahorrarse un buen dinero.

Con su cabello castaño arenoso, sonrisa dentona y belleza de querubín, Etzioni no parecía precisamente la clase de persona que le negaría a las líneas aéreas millones de dólares de ingresos potenciales. Pero, de hecho, se propuso hacer aún más que eso. Llegado el año 2008, estaba planeando aplicar el método a otros bienes, como las habitaciones de hotel, las entradas de conciertos y los coches de segunda mano: cualquier cosa que presentase una diferenciación reducida de producto, un grado elevado de variación en el precio y toneladas de datos. Pero, antes de que pudiera llevar sus planes a la práctica, Microsoft llamó a su puerta, se hizo con Farecast[5] por alrededor de ciento diez millones de dólares, y lo integró en el motor de búsqueda Bing. Para el año 2012, el sistema acertaba el 75 por 100 de las veces y le estaba ahorrando una media de cincuenta dólares por billete a los viajeros.

Farecast es el modelo perfecto de la compañía de *big data*, y un buen ejemplo de hacia dónde se

encamina el mundo. Cinco o diez años antes, Etzioni no podría haber creado la empresa. "Habría sido imposible", afirma. La capacidad de computación y almacenamiento que precisaba resultaba demasiado cara. Aunque los cambios en la tecnología resultaron un factor crucial a la hora de hacerlo posible, algo más importante cambió asimismo, algo sutil: se produjo una modificación en la perspectiva acerca del posible uso de los datos.

Los datos ya no se contemplaban como algo estático o rancio, cuya utilidad desaparecía en cuanto se alcanzaba el objetivo para el que habían sido recopilados, es decir, nada más aterrizar el avión (o, en el caso de Google, una vez procesada la búsqueda en curso). Por el contrario, los datos se convirtieron en una materia prima del negocio, en un factor vital, capaz de crear una nueva forma de valor económico. En la práctica, con la perspectiva adecuada, los datos pueden reutilizarse inteligentemente para convertirse en un manantial de innovación y servicios nuevos. Los datos pueden revelar secretos a quienes tengan la humildad, el deseo y las herramientas para escuchar.

DEJAR HABLAR A LOS DATOS

Los frutos de la sociedad de la información están bien a la vista, con un teléfono móvil en cada bolsillo, un ordenador portátil en cada mochila, y grandes sistemas de tecnología de la información funcionando en las oficinas por todas partes. Menos llamativa resulta la información en sí misma. Medio siglo después de que los ordenadores se propagaran a la mayoría de la población, los datos han empezado a acumularse hasta el punto de que está sucediendo algo nuevo y especial. No solo es que el mundo esté sumergido en más información que en ningún momento anterior, sino que esa información está creciendo más deprisa. El cambio de escala ha conducido a un cambio de estado. El cambio cuantitativo ha llevado a un cambio cualitativo. Fue en ciencias como la astronomía y la genética, que experimentaron por primera vez esa explosión en la década de 2000, donde se acuñó el término *big data*, "datos masivos"[6]. El concepto está trasladándose ahora hacia todas las áreas de la actividad humana.

No existe ninguna definición rigurosa de los datos masivos. En un principio, la idea era que el volumen de información había aumentado tanto que la que se examinaba ya no cabía en la memoria que los ordenadores emplean para procesarla, por lo que los ingenieros necesitaban modernizar las herramientas para poder analizarla. Esees el origen de las nuevas tecnologías de procesamiento, como Map-Reduce, de Google, y su equivalente de código abierto, Hadoop, que surgió de Yahoo. Con ellos se pueden manejar cantidades de datos mucho mayores que antes, y esos datos –esto es lo importante– no precisan ser dispuestos en filas ordenadas ni en las clásicas tabulaciones de una base de datos. Otras tecnologías de procesamiento de datos que prescinden de las jerarquías rígidas y de la homogeneidad de antaño se vislumbran asimismo en el horizonte. Al mismo tiempo, dado que las compañías de internet podían recopilar vastas cantidades de datos y tenían un intenso incentivo financiero por hallarles algún sentido, se convirtieron en las principales usuarias de las tecnologías de procesamiento más recientes, desplazando a compañías de fuera de la red que, en algunos casos, tenían ya décadas de experiencia acumulada.

Una forma de pensar en esta cuestión hoy en día –la que aplicamos en este libro– es la siguiente: los *big data*, los datos masivos, se refieren a cosas que se pueden hacer a gran escala, pero no a una escala inferior, para extraer nuevas percepciones o crear nueva formas de valor, de tal forma que transforman los mercados, las organizaciones, las relaciones entre los ciudadanos y los gobiernos,

etc.

Pero esto no es más que el principio. La era de los datos masivos pone en cuestión la forma en que vivimos e interactuamos con el mundo. Y aun más, la sociedad tendrá que desprenderse de parte de su obsesión por la causalidad a cambio de meras correlaciones: ya no sabremos *por qué*, sino solo *qué*. Esto da al traste con las prácticas establecidas durante siglos y choca con nuestra comprensión más elemental acerca de cómo tomar decisiones y aprehender la realidad.

Los datos masivos señalan el principio de una transformación considerable. Como tantas otras tecnologías nuevas, la de los datos masivos seguramente acabará siendo víctima del conocido *hype cycle* [ciclo de popularidad] de Silicon Valley: después de ser festejada en las portadas de las revistas y en las conferencias del sector, la tendencia se verá arrinconada y muchas de las *start ups* nacidas al socaire del entusiasmo por los datos se vendrán abajo. Pero tanto el encaprichamiento como la condena suponen malinterpretar profundamente la importancia de lo que está ocurriendo. De la misma forma que el telescopio nos permitió vislumbrar el universo y el microscopio nos permitió comprender los gérmenes, las nuevas técnicas de recopilación y análisis de enormes volúmenes de datos nos ayudarán a ver el sentido de nuestro mundo de una forma que apenas intuimos. En este libro no somos tanto los evangelistas de los datos masivos cuanto sus simples mensajeros. Y, una vez más, la verdadera revolución no se cifra en las máquinas que calculan los datos, sino en los datos mismos y en cómo los usamos.

Para apreciar hasta qué punto está ya en marcha la revolución de la información, considérense las tendencias que se manifiestan en todo el espectro de la sociedad. Nuestro universo digital está en expansión constante. Piénsese en la astronomía.[7] Cuando el Sloan Digital Sky Survey arrancó en 2000, solo en las primeras semanas su telescopio de Nuevo México recopiló más datos de los que se habían acumulado en toda la historia de la astronomía. Para 2010, el archivo del proyecto estaba a rebosar, con unos colosales 140 terabytes de información. Sin embargo, un futuro sucesor, el Gran Telescopio Sinóptico de Investigación de Chile, cuya inauguración está prevista para 2016, acopiará esa cantidad de datos cada cinco días.

Similares cantidades astronómicas las tenemos también más a mano. Cuando los científicos descifraron por primera vez el genoma humano en 2003, secuenciar los tres mil millones de pares de bases les exigió una década de trabajo intensivo. Hoy en día, diez años después, un solo laboratorio es capaz de secuenciar esa cantidad de ADN en un día. En el campo de las finanzas, en los mercados de valores de Estados Unidos, a diario cambian de manos siete mil millones de acciones[8], dos terceras partes de las cuales se negocian mediante algoritmos de ordenador basados en modelos matemáticos que procesan montañas de datos para predecir ganancias, al tiempo que intentan reducir los riesgos.

Las compañías de internet se han visto particularmente abrumadas. Google procesa más de 24 petabytes de datos al día, un volumen que representa miles de veces la totalidad del material impreso que guarda la Biblioteca del Congreso de Estados Unidos. A Facebook, una empresa que no existía hace una década, se suben más de diez millones de fotos nuevas cada hora. Sus usuarios hacen clic en el botón de "me gusta" o insertan un comentario casi tres mil millones de veces diarias, dejando un rastro digital que la compañía explota para descubrir sus preferencias. Entretanto, los ochocientos millones de usuarios mensuales del servicio YouTube de Google suben más de una hora de vídeo cada segundo. El número de mensajes de Twitter aumenta alrededor de un 200 por 100 al año, y en

2012 se habían superado los cuatrocientos millones de tuits diarios.

De las ciencias a la asistencia médica, de la banca a internet, los sectores pueden ser muy distintos, pero en conjunto cuentan una historia parecida: la cantidad de datos que hay en el mundo está creciendo deprisa, desbordando no solo nuestras máquinas, sino también nuestra propia imaginación.

Son muchos quienes han intentado determinar la cifra exacta de la cantidad de información que nos rodea, y calcular a qué velocidad crece. Lo conseguido ha sido irregular, porque han medido cosas diferentes. Uno de los estudios más completos es obra de Martin Hilbert, de la Annenberg School de comunicación y periodismo de la universidad del Sur de California. Hilbert se ha esforzado por cifrar todo cuanto ha sido producido, almacenado y comunicado, lo cual comprendería no solo libros, cuadros, correos electrónicos, fotografías, música y vídeo (analógico y digital), sino también videojuegos, llamadas telefónicas, hasta navegadores de vehículos y cartas enviadas por correo postal. También incluyó medios de emisión como la televisión y la radio, basándose en sus cifras de audiencia.

Según el cómputo de Hilbert, en 2007 existían más de 300 exabytes de datos almacenados. Para entender lo que esto representa en términos ligeramente más humanos, piénsese que un largometraje entero en formato digital puede comprimirse en un archivo de 1 gigabyte. Un exabyte son mil millones de gigabytes. En resumidas cuentas: una barbaridad. Lo interesante es que en 2007 solo en torno al 7 por 100 de los datos eran analógicos (papel, libros, copias de fotografías, etcétera); el resto ya eran digitales. Sin embargo, no hace demasiado, el panorama era muy diferente. Pese a que los conceptos de "revolución de la información" y "era digital" existen desde la década de 1960, apenas acaban de convertirse en realidad de acuerdo con ciertas medidas. Todavía en el año 2000, tan solo una cuarta parte de la información almacenada en el mundo era digital: las otras tres cuartas partes estaban en papel, película, discos LP de vinilo, cintas de cassette y similares.

La masa total de la información digital de entonces no era gran cosa, lo que debería inspirar modestia a los que llevan mucho tiempo navegando por la red y comprando libros online. (De hecho, en 1986 cerca del 40 por 100 de la capacidad de computación general del mundo revestía la forma de calculadoras de bolsillo, que representaban más poder de procesamiento que la totalidad de los ordenadores personales del momento). Como los datos digitales se expanden tan deprisa –multiplicándose por algo más de dos cada tres años, según Hilbert–, la situación se invirtió rápidamente. La información analógica, en cambio, apenas crece en absoluto. Así que en 2013 se estima que la cantidad[9] total de información almacenada en el mundo es de alrededor de 1.200 exabytes, de los que menos del 2 por 100 es no digital.

No hay manera fácil de concebir lo que supone esta cantidad de datos. Si estuvieran impresos en libros, cubrirían la superficie entera de Estados Unidos, formando unas cincuenta y dos capas. Si estuvieran grabados en CD-ROMs apilados, tocarían la Luna formando cinco pilas separadas. En el siglo III a. de C., cuando Tolomeo II de Egipto se afanaba por conservar un ejemplar de cada obra escrita, la gran biblioteca de Alejandría representaba la suma de todo el conocimiento del mundo. El diluvio digital que está barriendo ahora el planeta es el equivalente a darle hoy a cada persona de la Tierra trescientas veinte veces la cantidad de información que, se estima, almacenaba la biblioteca de Alejandría.

Las cosas se están acelerando de verdad. La cantidad de información almacenada crece cuatro veces

más deprisa que la economía mundial, mientras que la capacidad de procesamiento de los ordenadores crece nueve veces más deprisa. No tiene nada de raro que la gente se queje de sobrecarga informativa. A todos nos abruman los cambios.

Tómese la perspectiva a largo plazo, comparando el actual diluvio de datos con una revolución de la información anterior, la de la imprenta de Gutenberg, inventada hacia 1439. En los cincuenta años que van de 1453 a 1503, se imprimieron unos ocho millones de libros[10], según la historiadora Elizabeth Eisenstein. Esto se considera más que lo producido por todos los escribas de Europa desde la fundación de Constantinopla, unos mil doscientos años antes. En otras palabras, hicieron falta cincuenta años para que las existencias de información casi se duplicaran en Europa, en contraste con los cerca de tres años que tarda en hacerlo hoy en día.

¿Qué representa este incremento? A Peter Norvig, experto en inteligencia artificial de Google, le gusta pensar al respecto con una analogía gráfica. En primer lugar, nos pide que pensemos en el caballo icónico de las pinturas rupestres de Lascaux, en Francia, que datan del Paleolítico, hace unos diecisiete mil años. A continuación, pensemos en una fotografía de un caballo: o aún mejor, en los garabatos de Picasso, que no son demasiado distintos de las pinturas rupestres. De hecho, cuando se le mostraron a Picasso las imágenes de Lascaux[11], comentó con mordacidad: "No hemos inventado nada".

Las palabras de Picasso eran ciertas desde un punto de vista, pero no desde otro. Recuérdese la fotografía del caballo. Mientras que antes hacía falta mucho tiempo para dibujar un caballo, ahora podía conseguirse una representación de uno, mucho más deprisa, mediante una fotografía. Eso supone un cambio, pero puede que no sea el más esencial, dado que sigue siendo fundamentalmente lo mismo: una imagen de un caballo. Sin embargo, ruega Norvig, considérese ahora la posibilidad de capturar la imagen de un caballo y acelerarla hasta los veinticuatro fotogramas por segundo. El cambio cuantitativo ha producido uno cualitativo. Una película es fundamentalmente diferente de una fotografía estática. Lo mismo ocurre con los datos masivos: al cambiar la cantidad, cambiamos la esencia.

Considérese una analogía procedente de la nanotecnología, donde las cosas se vuelven más pequeñas, no más grandes. El principio que subyace a la nanotecnología es que, cuando se alcanza el nivel molecular, las propiedades físicas pueden alterarse. Conocer esas nuevas características supone que se pueden inventar materiales que hagan cosas antes imposibles. A nanoescala, por ejemplo, se pueden dar metales más flexibles y cerámicas expandibles. A la inversa, cuando aumentamos la escala de los datos con los que trabajamos, podemos hacer cosas nuevas que no eran posibles cuando solo trabajábamos con cantidades más pequeñas.

A veces las restricciones con las que vivimos, y que presumimos idénticas para todo, son, en realidad, únicamente funciones de la escala a la que operamos. Pensemos en una tercera analogía, de nuevo del campo de las ciencias. Para los seres humanos, la ley física más importante de todas es la de la gravedad: impera sobre todo cuanto hacemos. Pero, para los insectos minúsculos, la gravedad es prácticamente inmaterial. Para algunos, como los zapateros de agua, la ley operativa del universo físico es la tensión de la superficie, que les permite cruzar un estanque sin caerse en él.

En la información, como en la física, el tamaño sí importa. Por consiguiente, Google mostró que era capaz de determinar la prevalencia de la gripe casi igual de bien que los datos oficiales basados en las visitas de pacientes al médico. Google puede hacerlo peinando cientos de miles de millones de términos de búsqueda, y puede obtener una respuesta casi en tiempo real, mucho más rápido que las fuentes oficiales. Del mismo modo, el Farecast de Etzioni puede predecir la volatilidad del

precio de un billete de avión, poniendo así un poder económico sustancial en manos de los consumidores. Pero ambos solo pueden hacerlo bien mediante el análisis de cientos de miles de millones de puntos de datos.

Estos dos ejemplos demuestran el valor científico y societario de los datos masivos, y hasta qué punto pueden estos convertirse en una fuente de valor económico. Reflejan dos formas en que el mundo de los datos masivos está a punto de revolucionarlo todo, desde las empresas y las ciencias hasta la atención médica, la administración, la educación, la economía, las humanidades y todos los demás aspectos de la sociedad.

Aunque solo nos hallamos en los albores de la era de los datos masivos, nos apoyamos en ellos a diario. Los filtros de spam están diseñados para adaptarse a medida que cambian las clases de correo electrónico basura: no sería posible programar el software para que supiera bloquear "via6ra" o su infinidad de variantes. Los portales de encuentros emparejan a la gente basándose en la correlación de sus numerosos parámetros con los de anteriores emparejamientos felices. La función de "autocorrección" de los teléfonos inteligentes rastrea nuestras acciones y añade palabras nuevas a su diccionario ortográfico basándose en lo que tecleamos. Sin embargo, esos usos no son más que el principio. Desde los coches capaces de detectar cuándo girar o frenar hasta el ordenador Watson de IBM que derrota a las personas en el concurso televisivo *Jeopardy!*, el enfoque renovará muchos aspectos del mundo en el que vivimos.

Esencialmente, los datos masivos consisten en hacer predicciones. Aunque se los engloba en la ciencia de la computación llamada inteligencia artificial y, más específicamente, en el área llamada aprendi zaje automático o de máquinas, esta caracterización induce a error. El uso de datos masivos no consiste en intentar "enseñar" a un ordenador a "pensar" como un ser humano. Más bien consiste en aplicar las matemáticas a enormes cantidades de datos para poder inferir probabilidades: la de que un mensaje de correo electrónico sea spam; la de que la combinación de letras "lso" corresponda a "los"; la de que la trayectoria y velocidad de una persona que cruza sin mirar suponen que le dará tiempo a atravesar la calle, y el coche autoconducido solo necesitará aminorar ligeramente la marcha. La clave radica en que estos sistemas funcionan bien porque están alimentados con montones de datos sobre los que basar sus predicciones. Es más, los sistemas están diseñados para perfeccionarse solos a lo largo del tiempo, al estar pendientes de detectar las mejores señales y pautas cuando se les suministran más datos.

En el futuro –y antes de lo que pensamos–, muchos aspectos de nuestro mundo que hoy son competencia exclusiva del juicio humano se verán incrementados o sustituidos por sistemas computerizados. No solo conducir un coche o ejercer de casamentero, sino tareas aún más complejas. Al fin y al cabo, Amazon puede recomendar el libro ideal, Google puede indicar la página web más relevante, Facebook conoce nuestros gustos, y LinkedIn adivina a quién conocemos. Las mismas tecnologías se aplicarán al diagnóstico de enfermedades, la recomendación de tratamientos, tal vez incluso a la identificación de "delincuentes" antes de que cometan de hecho un delito. De la misma forma que internet cambió radicalmente el mundo al añadir comunicación a los ordenadores, los datos masivos modificarán diversos aspectos fundamentales de la vida, otorgándole una dimensión cuantitativa que nunca había tenido antes.

MÁS, DE SOBRA, YA BASTA

Los datos masivos serán una fuente de innovación y de nuevo valor económico. Pero hay aún más en juego. El auge de los datos masivos representa tres cambios en la forma de analizar la información que modifican nuestra manera de comprender y organizar la sociedad.

El primer cambio se describe en el capítulo II. En este nuevo mundo podemos analizar muchos más datos. En algunos casos, incluso podemos procesar *todos* los relacionados con un determinado fenómeno. Desde el siglo XIX, la sociedad ha dependido de las muestras cuando ha tenido que hacer frente a cifras elevadas. Sin embargo, la necesidad del muestreo es un síntoma de escasez informativa, un producto de las restricciones naturales sobre la interacción con la información durante la era analógica. Antes de la prevalencia de las tecnologías digitales de alto rendimiento, no veíamos en el muestreo una atadura artificial: normalmente lo dábamos por supuesto sin más. El emplear todos los datos nos permite apreciar detalles que nunca pudimos ver cuando estábamos limitados a las cantidades más pequeñas. Los datos masivos nos ofrecen una vista particularmente despejada de lo granular: subcategorías y submercados que las muestras, sencillamente, no permiten estimar.

El considerar un número ampliamente más vasto de datos nos permite también relajar nuestro anhelo de exactitud, y ese es el segundo cambio, que identificamos en el capítulo III. Se llega así a un término medio: con menos errores de muestreo, podemos asumir más errores de medida. Cuando nuestra capacidad de medición es limitada, solo contamos las cosas más importantes. Lo que conviene es esforzarse por obtener el resultado exacto. De nada sirve vender reses cuando el comprador no está seguro de si en el rebaño hay cien cabezas o solo ochenta. Hasta hace poco, todas nuestras herramientas digitales partían de la premisa de la exactitud: asumíamos que los motores de búsqueda de las bases de datos darían con los archivos que se ajustaban a la perfección a nuestra consulta, igual que una hoja de cálculo tabula los números en una columna.

Esta forma de pensar resultaba de un entorno de "datos escasos": con tan pocas cosas que medir, teníamos que tratar de la forma más precisa posible lo que nos molestábamos en cuantificar. En cierto modo, esto es obvio: al llegar la noche, una tienda pequeña puede contar el dinero que hay en la caja hasta el último céntimo, pero no haríamos lo mismo –de hecho, no podríamos– en el caso del producto interior bruto de un país. Conforme va aumentando la escala, también crece el número de errores.

La exactitud requiere datos cuidadosamente seleccionados. Puede funcionar con cantidades pequeñas y, por descontado, hay situaciones que aún la requieren: uno o bien tiene dinero suficiente en el banco para extender un cheque, o no. Pero en un mundo de datos masivos, a cambio de emplear series de datos mucho más extensas podemos dejar de lado parte de la rígida exactitud.

A menudo, los datos masivos resultan confusos, de calidad variable, y están distribuidos entre innumerables servidores por todo el mundo. Con ellos, muchas veces nos daremos por satisfechos con una idea de la tendencia general, en lugar de conocer un fenómeno hasta el último detalle, céntimo o molécula. No es que renunciemos a la exactitud por entero; solo abandonamos nuestra devoción por ella. Lo que perdemos en exactitud en el nivel micro, lo ganamos en percepción en el nivel macro.

Estos dos cambios conducen a un tercero, que explicamos en el capítulo IV: un alejamiento de la tradicional búsqueda de causalidad. Como seres humanos, hemos sido condicionados para buscar causas, aun cuando la búsqueda de la causalidad resulte a menudo difícil y pueda conducirnos por el camino equivocado. En un mundo de datos masivos, en cambio, no necesitamos concentrarnos en la causalidad; por el contrario, podemos descubrir pautas y correlaciones en los datos que nos ofrezcan

perspectivas nuevas e inapreciables. Puede que las correlaciones no nos digan precisamente *por qué* está ocurriendo algo, pero nos alertan de que *algo* está pasando.

Y en numerosas situaciones, con eso basta. Si millones de registros médicos electrónicos revelan que los enfermos de cáncer que toman determinada combinación de aspirina y zumo de naranja ven remitir su enfermedad, la causa exacta de la mejoría puede resultar menos importante que el hecho de que sobreviven. Del mismo modo, si podemos ahorrarnos dinero sabiendo cuál es el mejor momento de comprar un billete de avión, aunque no comprendamos el método subyacente a la locura de las tarifas aéreas, con eso basta. Los datos masivos tratan del *qué*, no del *porqué*. No siempre necesitamos conocer la causa de un fenómeno; preferentemente, podemos dejar que los datos hablen por sí mismos.

Antes de los datos masivos, nuestro análisis se limitaba habitualmente a someter a prueba un reducido número de hipótesis que definíamos con precisión antes incluso de recopilar los datos. Cuando dejamos que hablen los datos, podemos establecer conexiones que nunca hubiésemos sospechado. En consecuencia, algunos fondos de inversión libre analizan Twitter para predecir la evolución del mercado de valores. Amazon y Netflix basan sus recomendaciones de productos en una miríada de interacciones de los usuarios de sus páginas web. Twitter, LinkedIn y Facebook trazan la "gráfica social" de relaciones de los usuarios para conocer sus preferencias.

Por descontado, los seres humanos llevan milenios analizando datos. La escritura nació en la antigua Mesopotamia porque los burócratas querían un instrumento eficiente para registrar la información y seguirle la pista. Desde tiempos bíblicos, los gobiernos han efectuado censos para recopilar enormes conjuntos de datos sobre sus ciudadanos, e igualmente durante doscientos años los analistas de seguros han hecho grandes acopios de datos acerca de los riesgos que esperan entender; o, por lo menos, evitar.

Sin embargo, en la era analógica la recopilación y el análisis de esos datos resultaba enormemente costosa y consumía mucho tiempo. El hacer nuevas preguntas a menudo suponía recoger los datos de nuevo y empezar el análisis desde el principio.

El gran avance hacia la gestión más eficiente de los datos llegó con el advenimiento de la digitalización: hacer que la información analógica fuese legible por los ordenadores, lo que también la vuelve más fácil y barata de almacenar y procesar. Este progreso mejoró drásticamente la eficiencia. La recopilación y el análisis de información, que en tiempos exigía años, podía ahora hacerse en días, o incluso menos. Pero cambió poco más. Los encargados de los datos muy a menudo estaban versados en el paradigma analógico de asumir que los conjuntos de datos tenían propósitos específicos de los que dependía su valor. Nuestros mismos procesos perpetuaron este prejuicio. Por importante que resultase la digitalización para permitir el cambio a los datos masivos, la mera existencia de ordenadores no los hizo aparecer.

No existe un término adecuado para describir lo que está sucediendo ahora mismo, pero uno que ayuda a enmarcar los cambios es *datificación*, concepto que introducimos en el capítulo v. Datificar se refiere a recopilar información sobre cuanto existe bajo el sol –incluyendo cosas que en modo alguno solíamos considerar información antes, como la localización de una persona, las vibraciones de un motor o la tensión que soporta un puente–, y transformarla a formato de datos para cuantificarla. Esto nos permite usar la información de modos nuevos, como en el análisis predictivo: detectar que un motor es proclive a un fallo mecánico basándonos en el calor o en las vibraciones

que emite. Lo que se consigue así es liberar el valor latente e implícito de la información.

Estamos en plena caza del tesoro, una caza impulsada por las nuevas percepciones que podrían extraerse de los datos y el valor latente que podría liberarse si nos desplazamos desde la causalidad a la correlación. Pero no se trata de un único tesoro. Cada serie de datos probablemente tenga algún valor intrínseco y oculto, aún no desvelado, y ha empezado la carrera para descubrirlos y capturarlos todos.

Los datos masivos alteran la naturaleza de los negocios, los mercados y la sociedad, como describimos en los capítulos VI y VII. En el siglo XX, el valor se desplazó de las infraestructuras físicas, como la tierra y las fábricas, a los intangibles, como las marcas y la propiedad intelectual. Estos se expanden ahora a los datos, que se están convirtiendo en un activo corporativo importante, un factor económico vital, y el fundamento de nuevos modelos económicos. Aunque los datos todavía no se registran en los balances de las empresas, probablemente sea solo cuestión de tiempo.

Aunque hace mucho que existen algunas de las técnicas de procesamiento de datos, antes solo podían permitírselas los organismos de seguridad del estado, los laboratorios de investigación y las mayores compañías del mundo. Al fin y al cabo, Walmart y Capital One fueron pioneros en el empleo de datos masivos en la venta al por menor y en la banca, y con ello cambiaron sus respectivas industrias. Ahora, muchas de estas herramientas se han democratizado (aunque no así los datos).

El efecto sobre los individuos acaso suponga la mayor sorpresa de todas. El conocimiento especialista en áreas específicas importa menos en un mundo en el que la probabilidad y la correlación lo son todo. En la película *Moneyball*, los ojeadores de béisbol se veían desplazados por los estadísticos cuando el instinto visceral cedía el paso al análisis más sofisticado. Igualmente, los especialistas en una materia dada no desaparecerán, pero tendrán que competir con lo que determine el análisis de datos masivos. Ello forzará a ajustarse a las ideas tradicionales acerca de la gestión, la toma de decisiones, los recursos humanos y la educación.

La mayor parte de nuestras instituciones han sido creadas bajo la presunción de que las decisiones humanas se basan en una información contada, exacta y de naturaleza causal. Pero la situación cambia cuando los datos son enormes, pueden procesarse rápidamente y admiten la inexactitud. Es más, debido al vasto tamaño de la información, muy a menudo las decisiones no las tomarán los seres humanos, sino las máquinas. Consideraremos el lado oscuro de los datos masivos en el capítulo VIII.

La sociedad cuenta con milenios de experiencia en lo que a comprender y supervisar el comportamiento humano se refiere, pero, ¿cómo se regula un algoritmo? En los albores de la computación, los legisladores advirtieron que la tecnología podía usarse para socavar la privacidad. Desde entonces, la sociedad ha erigido un conjunto de reglas para proteger la información personal. Sin embargo, en la era de los datos masivos, esas leyes constituyen una línea Maginot en buena medida inútil. La gente comparte gustosamente información online: es una característica central de los servicios en red, no una vulnerabilidad que haya que evitar.

Entretanto, el peligro que se cierne sobre nosotros en tanto que individuos se desplaza del ámbito de lo privado al de la probabilidad: los algoritmos predecirán la probabilidad de que uno sufra un ataque al corazón (y tenga que pagar más por un seguro médico), deje de pagar la hipoteca (y se le niegue un crédito) o cometa un delito (y tal vez sea detenido antes de los hechos). Ello conduce a una

consideración ética del papel del libre albedrío frente a la dictadura de los datos. ¿Debería imponerse la voluntad del individuo a los datos masivos, aun cuando las estadísticas argumenten lo contrario? Igual que la imprenta preparó el terreno para las leyes que garantizaban la libertad de expresión –que no existían antes, al haber tan poca expresión escrita que proteger–, la era de los datos masivos precisará de nuevas reglas para salvaguardar la inviolabilidad del individuo.

Nuestra forma de controlar y manejar los datos tendrá que cambiar de muchas maneras. Estamos entrando en un mundo de constantes predicciones sustentadas por datos, en el que puede que no seamos capaces de explicar las razones de nuestras decisiones. ¿Qué significará que el doctor no pueda justificar una intervención médica sin pedirle al paciente que se pliegue al dictamen de algún tipo de "caja negra", como no tiene más remedio que hacer cuando se basa en un diagnóstico sustentado por datos masivos? ¿Necesitará cambiarse la norma judicial de "causa probable" por la de "causa probabilística"? Y, de ser así, ¿qué implicaciones tendrá esto para la libertad y la dignidad humanas?

Son precisos unos principios nuevos para la era de los datos masivos, y los exponemos en el capítulo IX. Y, aunque estos principios se construyen sobre los valores que se desarrollaron y consagraron en el mundo de los datos escasos, no se trata simplemente de refrescar viejas reglas para las nuevas circunstancias, sino de reconocer la necesidad de crearlas de nuevo y desde cero.

Los beneficios para la sociedad resultarán muy numerosos, conforme los datos masivos se conviertan en parte de la solución de ciertos tico, erradicar las enfermedades y fomentar el buen gobierno y el desarrollo económico. Pero la era de los datos masivos también nos invita a prepararnos mejor para las formas en que el aprovechamiento de las tecnologías cambiará nuestras instituciones y nos cambiará a nosotros.

Los datos masivos suponen un paso importante en el esfuerzo de la humanidad por cuantificar y comprender el mundo. Una inmensa cantidad de cosas que antes nunca pudieron medirse, almacenarse, analizarse y compartirse están convirtiéndose en datos. El aprovechamiento de vastas cantidades de datos en lugar de una pequeña porción, y el hecho de preferir más datos de menor exactitud, abre la puerta a nuevas formas de comprender. Lleva la sociedad al abandono de su tradicional preferencia por la causalidad, y en muchos casos aprovecha los beneficios de la correlación.

El ideal de la identificación de los mecanismos causales no deja de ser una ilusión autocomplaciente: los datos masivos dan al traste con ella. Una vez más nos encontramos en un callejón sin salida en el que "Dios ha muerto". Vale decir, que las certezas en las que creíamos están cambiando una vez más, pero esta vez están siendo reemplazadas, irónicamente, por pruebas más sólidas. ¿Qué papel les queda a la intuición, la fe, la incertidumbre, el obrar en contra de la evidencia, y el aprender de la experiencia? Mientras el mundo se mueve de la causalidad a la correlación, ¿cómo podemos seguir adelante pragmáticamente sin socavar los mismos cimientos de la sociedad, la humanidad y el progreso fundado en la razón? Este libro pretende explicar dónde nos hallamos, explicar cómo llegamos hasta aquí, y ofrecer una guía, de necesidad urgente, sobre los beneficios y peligros que nos acechan.

II
MÁS

Los datos masivos consisten en ver y comprender las relaciones en el seno y entre distintos fragmentos de información que, hasta hace muy poco, nos esforzábamos por captar plenamente. Jeff Jonas[12], el experto en datos masivos de IBM, sostiene que hay que dejar que los datos "le hablen a uno". En cierto modo, esto puede parecer obvio, porque los seres humanos hemos prestado atención a los datos para intentar conocer el mundo desde hace mucho tiempo, bien en el sentido informal, el de las innumerables observaciones que hacemos a diario o, fundamentalmente a lo largo del último par de siglos, en el sentido formal de unidades cuantificadas que pueden manipularse con algoritmos poderosos.

Puede que la era digital haya vuelto el procesamiento de datos más sencillo y más rápido, para calcular millones de números en un latido, pero cuando hablamos de datos que hablan nos referimos a algo más, y distinto. Como se ha señalado en el capítulo I, los datos masivos tienen que ver con tres importantes cambios de mentalidad, que al estar interrelacionados se refuerzan entre sí. El primero es la capacidad de analizar enormes cantidades de información sobre un tema dado, en lugar de verse uno forzado a conformarse con conjuntos más pequeños. El segundo es la disposición a aceptar la imprecisión y el desorden –muy del mundo real– de los datos, en lugar de anhelar la exactitud. El tercer cambio pasa por empezar a respetar las correlaciones, en vez de buscar constantemente la elusiva causalidad. Este capítulo examina el primero de estos cambios: el uso de casi todos los datos en lugar de, únicamente, una pequeña porción de ellos.

Hace tiempo que nos acompaña el reto de procesar de forma precisa grandes montones de datos. Durante la mayor parte de la historia, hemos usado muy poca información porque nuestras herramientas para recogerla, organizarla, almacenarla y analizarla eran muy pobres. Trillábamos la que teníamos hasta la mínima expresión, para examinarla luego con más facilidad. Esta era una forma de autocensura inconsciente: tratábamos la dificultad de interactuar con datos como una realidad desafortunada, en lugar de verla por lo que era, una restricción artificial impuesta por la tecnología de la época. Hoy en día, el entorno técnico ha dado un giro de 179 grados. Aún existe, y siempre lo hará, una limitación sobre cuántos datos podemos manejar, pero es mucho menos estrecha que antes, y lo irá siendo cada vez menos con el tiempo.

De algún modo, aún no hemos apreciado del todo nuestra nueva libertad de recopilar y explotar conjuntos más amplios de datos. La mayor parte de nuestras experiencias, y el diseño de nuestras instituciones, han dado por supuesto que la información disponible es limitada. Contábamos con poder recopilar poca información, así que eso era lo que hacíamos habitualmente, y eso se convirtió en un fin en sí mismo. Hasta desarrollamos técnicas complejas para emplear tan pocos datos como fuese posible. Uno de los fines de la estadística, al fin y al cabo, es confirmar el resultado más rico empleando la menor cantidad posible de datos. De hecho, en nuestras normas, procedimientos y estructuras de incentivos codificamos nuestra práctica de amputar la cantidad de información que empleábamos. Para hacerse cabal idea de lo que implica el cambio a los datos masivos, la historia

empieza mirando atrás.

Ha sido solo hace poco cuando las firmas privadas, y hoy en día incluso los particulares, han sido capaces de recoger y clasificar información a escala masiva. Antes, esa tarea recaía en instituciones más poderosas como la iglesia y el estado, que en muchas sociedades venían a ser una cosa y la misma. El registro contable más antiguo que se conserva es de alrededor del año 5.000 a. de C., cuando los mercaderes sumerios utilizaban pequeñas cuentas de arcilla para representar los bienes en venta. Contar a escala superior, sin embargo, era prerrogativa del estado. A lo largo de los milenios, los gobiernos han tratado de vigilar a sus súbditos recogiendo información.

Tomemos por ejemplo el censo. Se supone que los antiguos egipcios llevaban censos a cabo, igual que los chinos. Se los menciona en el Antiguo Testamento, y el Nuevo Testamento nos cuenta que un censo impuesto por César Augusto –"un edicto de empadronamiento para todo el orbe" (Lucas 2:1)– llevó a José y a María a Belén, donde nació Jesús. El *Domesday Book* de 1086, uno de los tesoros más venerados de Gran Bretaña, fue en su día un registro, sin precedentes y exhaustivo, del pueblo inglés, sus tierras y propiedades. Los comisarios reales recorrieron todo el país recopilando información para el libro, que más tarde recibiría el nombre de *Domesday*, o "Día de cuentas", porque el proceso era como el del juicio final según lo cuenta la Biblia, cuando se someterían a cuenta todos los actos de la vida de una persona.

Elaborar censos resulta al tiempo costoso y lento; el rey Guillermo I de Inglaterra, que encargó el *Domesday Book*, no vivió para verlo terminado. Sin embargo, la única alternativa a tamaña carga era renunciar a recoger la información. E incluso después de tanto tiempo y gasto, la información era solo aproximativa, dado que los funcionarios del censo no podían contar a todo el mundo perfectamente. La misma palabra censo procede del término latino *censere*, que significa "estimar".

Hace más de trescientos años, un mercero británico llamado John Graunt tuvo una idea novedosa. Graunt quería saber la población de Londres en la época de la gran peste. En lugar de contar una a una a todas las personas, desarrolló una aproximación –lo que hoy llamaríamos estadística– que le permitió *inferir* el tamaño de la población. Su planteamiento era tosco, pero sentó la idea de que a partir de una pequeña muestra se podían extrapolar conocimientos útiles acerca de la población general. Pero lo importante es cómo se hace. Graunt, sencillamente, extrapoló hacia arriba a partir de su muestra.

Su sistema recibió grandes parabienes, aun cuando se supiera más tarde que sus cifras eran razonables por pura chiripa. Durante generaciones, el muestreo siguió presentando tremendos fallos. Así pues, en el caso de los censos y similares empresas de datos masivos, se impuso el enfoque de la fuerza bruta: tratar de contar todos los números.

Como los censos eran tan complejos y costosos, y requerían tanto tiempo, se llevaban a cabo a intervalos largos. Los antiguos romanos, que durante mucho tiempo presumieron de una población de cientos de miles de habitantes, realizaban un censo cada cinco años. La constitución de Estados Unidos[13] dispuso que se hiciera uno cada década, conforme el país crecía y empezaba a medirse por millones. Pero para finales del siglo XIX, hasta eso empezaba a resultar problemático. Los datos sobrepasaban la capacidad de absorción de la oficina del censo.

El censo de 1880 precisó un pasmoso plazo de ocho años para llegar a su conclusión. La información se quedó obsoleta antes incluso de estar disponible. Aun peor, los funcionarios estimaron que el censo de 1890 habría requerido trece años enteros para su tabulación: una situación ridícula, además de anticonstitucional. Sin embargo, como el prorrateo de los impuestos y de la representación parlamentaria en el congreso se basaba en la población, resultaba esencial no solo

conseguir cuantificarla con precisión, sino a tiempo.

El problema al que se enfrentó la oficina del censo estadounidense es similar a la lucha de los científicos y los hombres de negocios en los albores del nuevo milenio, cuando quedó claro que los datos los desbordaban: la cantidad de información recogida había anegado literalmente las herramientas empleadas para procesarla, y se precisaban técnicas nuevas. En la década de 1880, la situación era tan abrumadora que la oficina del censo firmó un contrato con Herman Hollerith, un inventor estadounidense, para aplicar tarjetas perforadas y máquinas tabuladoras de su invención al censo de 1890.

Con gran esfuerzo, Hollerith logró reducir el plazo de tabulación de ocho años a algo menos de uno. Fue una hazaña asombrosa, que señaló el principio del procesamiento automatizado de datos (y estableció los fundamentos de lo que después sería IBM). Pero, como método para adquirir y analizar datos masivos, seguía siendo muy costoso. Al fin y al cabo, todas las personas de Estados Unidos tenían que rellenar un impreso, cuya información había de ser trasladada a una tarjeta perforada, que se empleaba para la tabulación. Con esos métodos tan onerosos, resultaba difícil imaginar un censo con periodicidad inferior a la década, aun cuando el desfase le resultara tan perjudicial a una nación que estaba creciendo a pasos agigantados.

En ello radicaba la tensión: ¿debían usarse todos los datos, o solo unos pocos? Conseguir todos los datos acerca de lo que se está midiendo, sea lo que fuere, es sin duda el método más sensato. Lo que ocurre es que no siempre resulta práctico cuando la escala es vasta. Pero, ¿cómo escoger una muestra? Algunos sostuvieron que construir a propósito una muestra que fuera representativa del conjunto sería el modo más adecuado de seguir adelante. Ahora bien, en 1934, Jerzy Neyman[14], un estadístico polaco, demostró de forma tajante que eso conduce a errores enormes. La clave para evitarlos es apostar por la aleatoriedad al escoger a quién muestrear.

Los estadísticos han demostrado que la precisión de la muestra mejora acusadamente con la aleatoriedad, no con el mayor tamaño de la muestra. En realidad, aunque pueda parecer sorprendente, una muestra aleatoria de 1.100 observaciones individuales sobre una pregunta binaria (sí o no, con aproximadamente las mismas probabilidades de darse) es notablemente representativa de toda la población. En 19 de cada 20 casos, presenta un margen de error inferior al 3 por 100, tanto si el tamaño de la población total es de cien mil como si es de cien millones. La razón resulta algo complicada de explicar en términos matemáticos, pero en resumen lo que ocurre es que, superado cierto punto, al principio, conforme las cifras van haciéndose mayores, la cantidad marginal de informaciones nuevas que se consigue de cada observación es cada vez menor.

El hecho de que la aleatoriedad se impusiera al tamaño de la muestra supuso una revelación sorprendente. Allanó el camino para un nuevo enfoque de la recolección de información. Los datos que usan muestras aleatorias podían recopilarse a bajo coste y, sin embargo, extrapolarse para el conjunto con gran exactitud. En consecuencia, los gobiernos podían acometer versiones reducidas del censo empleando muestras aleatorias cada año en vez de una sola cada diez. Y eso fue lo que hicieron. La oficina del censo estadounidense, por ejemplo, lleva a cabo cada año más de doscientos estudios económicos y demográficos basados en el muestreo, por añadidura al censo, además del censo decenal que pretende contabilizar a todo el mundo. El muestreo venía a solucionar el problema de la sobrecarga informativa de la era anterior, cuando la recopilación y el análisis de los datos resultaban en verdad muy difíciles de hacer.

La aplicación de este nuevo método se extendió rápidamente más allá del ámbito del sector público y de los censos. En esencia, el muestreo aleatorio reduce el problema de los datos masivos a

unas dimensiones más manejables. En el terreno de los negocios, se utilizó para asegurar la calidad de las manufacturas, al hacer que las mejoras resultaran más fáciles y menos costosas. Originalmente, el control de calidad exhaustivo exigía examinar todos y cada uno de los productos que salían de la cadena de montaje; ahora, bastaría con unas pruebas sobre una muestra aleatoria de un grupo de productos. Del mismo modo, el nuevo método introdujo las encuestas a consumidores en la venta al por menor y las encuestas sobre intenciones de voto en la política. Y así, transformó una buena parte de lo que antes llamábamos humanidades en *ciencias* sociales.

El muestreo aleatorio ha constituido un tremendo éxito, y es el espinazo de la medición a escala moderna. Pero no deja de ser un atajo, una alternativa de segundo orden a recopilar y analizar el conjunto entero de datos. Trae consigo una serie de debilidades inherentes. Su exactitud depende de que se haya garantizado la aleatoriedad al recopilar los datos de la muestra, pero el logro de esa aleatoriedad resulta peliagudo. Se producen sesgos sistemáticos en la forma de recopilar los datos que pueden hacer que los resultados extrapolados sean muy incorrectos.

Las encuestas electorales efectuadas por teléfono fijo dan fe, por ejemplo, de algunos de estos problemas. La muestra está sesgada en contra de la gente que solo usa teléfonos móviles[15] (que suelen ser más jóvenes y más progresistas), como ha señalado el estadístico Nate Silver. Esto se ha constatado en pronósticos electorales erróneos. En la elección presidencial de 2008 entre Barack Obama y John McCain, las principales empresas de sondeos electorales de Gallup, Pew y ABC/Washington Post hallaron diferencias de entre uno y tres puntos porcentuales al efectuar las encuestas, con y sin ajuste a los usuarios de teléfono móvil: un margen excesivo, considerando lo ajustado de la contienda.

De forma aún más preocupante, el muestreo aleatorio no resulta sencillo de extrapolar para incluir subcategorías, por lo que al parcelar los resultados en subgrupos cada vez menores aumenta la posibilidad de llegar a predicciones erróneas. Es fácil comprender por qué: supongamos que se le pregunta a una muestra aleatoria de mil personas por su intención de voto en las siguientes elecciones. Si la muestra es lo suficientemente aleatoria, existen posibilidades de que los pareceres de toda la población estén recogidos con un margen de error del 3 por 100 en las opiniones de la muestra. Pero, ¿qué ocurre si más o menos 3 por 100 no es lo bastante preciso? ¿O si después se quiere dividir el grupo en subgrupos más pequeños, por sexo, localidad, o nivel de renta?

¿Y qué pasa si se desea combinar esos subgrupos para determinar un nicho de la población? En una muestra global de mil personas, un subgrupo como el de "mujeres ricas votantes del nordeste" tendrá menos de cien miembros. Usar solo unas pocas docenas de observaciones para pronosticar las intenciones de voto de *todas* las mujeres pudientes en el nordeste resultará impreciso, aun con una aleatoriedad cuasi perfecta. Y estos pequeños sesgos en la muestra global harán que los errores de los subgrupos sean más pronunciados.

Por consiguiente, el muestreo deja de ser útil en cuanto se quiere ahondar más, para escrutar minuciosamente alguna subcategoría de datos que nos llame la atención. Lo que funciona en el nivel macro se viene abajo en el micro. El muestreo es como una copia fotográfica analógica. A cierta distancia, se ve muy bien, pero cuando se mira más de cerca, enfocando algún detalle particular, se vuelve borrosa.

El muestreo requiere, además, una planificación y ejecución cuidadosas. Normalmente no se les puede "pedir" a los datos de la muestra cuestiones nuevas que no se hayan contemplado desde el principio. Así pues, aunque como atajo resulta útil, el coste de oportunidad es precisamente el de que, al final, solo es un atajo. Y siendo una muestra en lugar de un todo, el conjunto de datos carece

de la extensibilidad o maleabilidad que serían necesarias para que los mismos datos pudieran ser analizados otra vez con un propósito enteramente distinto de aquel para el que fueron recopilados en origen.

Considérese el caso del análisis del ADN. El coste de secuenciar el genoma de un individuo era de cerca de mil dólares en 2012, lo que lo acercaba más a una técnica de consumo masivo que puede llevarse a cabo a gran escala. En consecuencia, está surgiendo una nueva industria de secuenciación de genes individuales. Desde 2007, la empresa *start up* de Silicon Valley 23andMe se ha dedicado a analizar el ADN de quien lo solicita por apenas un par de cientos de dólares. Su técnica permite revelar en el código genético ciertos rasgos, como el de ser más susceptible a ciertas enfermedades: por ejemplo, el cáncer de pecho o las afecciones cardíacas. Al agregar la información sobre el ADN y la salud de sus clientes, 23andMe espera descubrir cosas nuevas que de otro modo no podrían ser advertidas.

Pero hay un pero. La compañía no secuencia nada más que una pequeña porción del código genético de una persona: lo que ya sabe que son marcadores de determinadas debilidades genéticas. Entretanto, miles de millones de pares base de ADN permanecen sin secuenciar. Así pues, 23andMe solo puede dar respuesta a las preguntas acerca de los marcadores que toma en cuenta. Cada vez que se descubre un marcador nuevo, el ADN de una persona (o, con mayor precisión, la parte relevante del mismo) ha de ser secuenciada de nuevo. Trabajar con un subconjunto, en lugar del todo, implica un coste: la compañía puede encontrar lo que busca más deprisa y de forma más barata, pero no puede contestar a interrogantes que no hubiese contemplado de antemano.

El legendario director general de Apple, Steve Jobs[16], adoptó un enfoque completamente diferente en su lucha contra el cáncer. Se convirtió en una de las primeras personas del mundo en secuenciar todo su ADN, al igual que el de su tumor. Y pagó por ello una suma de seis dígitos: muchos cientos de veces la tarifa de 23andMe. A cambio, no recibió una muestra, un mero juego de marcadores, sino un archivo de datos con sus códigos genéticos completos.

Al prescribir la medicación para un enfermo de cáncer cualquiera, los médicos tienen que confiar en que el ADN del paciente sea lo bastante similar al de quienes hayan participado en las pruebas del fármaco para que este dé resultado. Sin embargo, el equipo médico de Steve Jobs podía elegir unas terapias en función de su específica constitución genética. Cuando un tratamiento perdía efectividad, porque el cáncer había mutado y proseguía su ataque, los médicos podían cambiar de fármaco: "saltar de una hoja de lirio a otra –como lo describió Jobs–. O bien seré uno de los primeros en vencer a un cáncer como este, o seré uno de los últimos en morir de él", bromeó. Si bien, por desgracia, su predicción no se cumplió, el método –disponer de todos los datos, no solo de unos cuantos– le prolongó la vida varios años.

DE ALGUNOS A TODOS

El muestreo es producto de una época de restricciones en el procesamiento de datos, cuando podíamos medir el mundo pero carecíamos de las herramientas para analizar lo recogido. En consecuencia, es asimismo un vestigio de ese tiempo. Las deficiencias al contar y al tabular ya no existen hasta el mismo punto. Los sensores, los GPS de los teléfonos móviles, los clics en la red y en Twitter recopilan datos de forma pasiva; los ordenadores procesan la información con mayor facilidad cada vez.

El concepto del muestreo no tiene ya el mismo sentido cuando resulta posible explotar grandes cantidades de datos. El instrumental técnico para manejar información ha cambiado drásticamente, pero nuestros métodos y mentalidades se van adaptando más despacio.

Además, el muestreo acarrea un coste del que se es consciente desde hace mucho, y que se ha dejado de lado. Se pierde detalle. En algunos casos, no queda más remedio que proceder por muestreo; en muchas áreas, sin embargo, se está produciendo un cambio de orientación, de la recogida de algunos datos a la recopilación de todos los posibles, y cuando es factible, de absolutamente todos: *N = todo*.

Como se ha visto, usar N = todo implica que podemos ahondar considerablemente en los datos; las muestras no permiten hacerlo igual de bien. En segundo lugar, recuérdese que en nuestro anterior ejemplo de muestreo, al extrapolar a la población entera teníamos un margen de error de solo el 3 por 100. En algunas situaciones, ese margen de error es estupendo, pero se pierden los detalles, la granularidad, la capacidad de estudiar de cerca determinados subgrupos. Una distribución normal, en fin, no es más que normal. A menudo, las cosas verdaderamente interesantes de la vida aparecen en lugares que las muestras no consiguen captar por completo.

Por consiguiente, Google Flu Trends[17], el indicador de tendencias de la gripe de Google, no se basa en una pequeña muestra probabilística, sino que utiliza miles de millones de búsquedas en internet en Estados Unidos. Usar todos estos datos en lugar de una muestra perfecciona el análisis hasta el extremo de poder predecir la propagación de la gripe a una ciudad determinada en lugar de a un estado o a la nación entera. Oren Etzioni de Farecast usó al principio doce mil puntos de datos, una muestra apenas, y le funcionó bien. Ahora bien, conforme fue añadiendo más datos, la calidad de las predicciones mejoró. Con el tiempo, Farecast utilizaba los registros de vuelos nacionales en la mayoría de las rutas durante todo un año. “Se trata de datos temporales: sigues recopilándolos a lo largo del tiempo, y así vas adquiriendo mayor percepción de los patrones”, afirma Etzioni.

En consecuencia, a menudo nos dará mejor resultado dejar de lado el atajo del muestreo aleatorio y tender a recopilar datos más exhaustivos. Para hacerlo se precisa una amplia capacidad de procesamiento y almacenaje, y herramientas de tecnología punta para analizarlo todo. También se necesitan formas sencillas y baratas de recopilar los datos. Hasta ahora, cada una de estos procesos suponía un problema económico, pero hoy en día el coste y complejidad de todas las piezas del rompecabezas han disminuido drásticamente. Lo que antes no estaba al alcance más que de las mayores empresas, hoy resulta posible para la mayoría.

El empleo de la totalidad de los datos hace posible advertir conexiones y detalles que de otro modo quedan oscurecidos en la vastedad de la información. Por ejemplo, la detección de los fraudes con tarjeta de crédito funciona mediante la búsqueda de anomalías, y la mejor forma de hallarlas es procesar todos los datos en lugar de solo una muestra. Los valores atípicos ofrecen la información más interesante, y solo se los puede identificar en comparación con la masa de transacciones normales. He aquí un problema de *big data*. Como las transacciones de tarjeta de crédito se producen instantáneamente, el análisis debería realizarse también en tiempo real.

Xoom es una firma especializada en transferencias internacionales de dinero, y la respaldan nombres importantes en el área de los datos masivos. Analiza todos los datos asociados con las transacciones que trata. El sistema hizo sonar la alarma en 2011 cuando advirtió un número ligeramente superior a la media de operaciones con tarjeta Discover con origen en Nueva Jersey. “Vimos un patrón donde no debería haber ninguno”, explicaba el director general de Xoom, John Kunze.[18] Tomadas una a una, todas las transacciones parecían legítimas, pero resultaron ser obra de

un grupo de delincuentes. La única forma de detectar la anomalía era examinar todos los datos: una muestra podría no haberlo advertido.

Emplear todos los datos no tiene por qué ser una tarea enorme. Los datos masivos no son necesariamente grandes en términos absolutos, aunque a menudo lo sean. Google Flu Trends afina sus predicciones basándose en cientos de millones de ejercicios de modelización matemática que emplean miles de millones de puntos de datos. La secuencia completa de un genoma humano representa tres mil millones de pares base. Sin embargo, no es solo el valor absoluto de puntos de datos, el tamaño del conjunto de datos, lo que hace que estos sean ejemplos de datos masivos. Lo que los convierte en casos de datos masivos es el hecho de que, en lugar de usar el atajo de una muestra aleatoria, tanto Flu Trends como los médicos de Steve Jobs hicieron uso, en lo posible, del conjunto íntegro de datos.

El descubrimiento de combates amañados en el deporte nacional de Japón, el sumo, es un buen ejemplo de por qué utilizar N = todo no tiene por qué significar "grande". Que hay combates trucados ha sido una acusación constante en el deporte de los emperadores, siempre enérgicamente negada. Steven Levitt, economista de la universidad de Chicago, buscó indicios de corrupción en los registros de todos los combates a lo largo de más de una década. En un artículo delicioso aparecido en el *The American Economic Review*, y luego recogido en el libro *Freakonomics*, un colega y él describieron la utilidad del examen de tantos datos.

Analizaron once años de encuentros de sumo, más de 64.000 combates, y encontraron un filón. En efecto, se amañaban combates, pero no donde la mayoría de la gente sospechaba. En lugar de en los encuentros que puntúan para el campeonato, que pueden estar amañados o no, los datos revelaron que ocurría algo raro en los combates finales de los torneos, mucho menos populares. Al parecer, hay poco en juego, ya que los luchadores no tienen posibilidad de obtener un título.

Pero una peculiaridad del sumo es que los luchadores necesitan obtener una mayoría de victorias en los torneos de 15 combates para poder mantener su rango y su nivel de ingresos. Esto conduce a veces a asimetrías de interés, cuando un luchador con un palmarés de 7-7 hace frente a un oponente con uno de 8-6 o aún mejor. El resultado del combate supone mucho para el primer luchador y prácticamente nada para el segundo. En tales casos, reveló el procesamiento de los datos, es muy probable que venza el luchador que necesita la victoria.

¿Pudiera ser que los sujetos que precisan ganar peleen con más determinación? Tal vez; sin embargo, los datos sugieren que ocurre algo más. Los luchadores que más se juegan ganan alrededor de un 25 por 100 más a menudo de lo normal. Es difícil atribuir una discrepancia tan grande únicamente a la adrenalina. Al analizar los datos con más detenimiento, se vio que en el enfrentamiento siguiente de esos mismos dos luchadores, el perdedor del encuentro anterior tenía muchas más probabilidades de vencer que cuando disputaban los combates finales. Así pues, la primera victoria parece ser un "obsequio" de un oponente al otro, dado que en el mundo del sumo lo que se siembra se cosecha.

Esta información siempre estuvo a plena vista. Pero el muestreo probabilístico de los encuentros podría no haber logrado revelarla. Aun cuando se basa en estadísticas elementales, sin saber qué es lo que hay que buscar no se hubiera podido saber qué tamaño necesitaba la muestra. Por el contrario, Levitt y su colega la sacaron a la luz usando un conjunto de datos mucho mayor, esforzándose por examinar el universo entero de los encuentros. Una investigación que usa datos masivos es casi como ir de pesca: al empezar no solo no está claro si alguien va a pescar algo: es que no se sabe *qué* puede pescar uno.

El conjunto de datos no necesita medir terabytes. En el caso del sumo, el conjunto íntegro de datos contenía menos bits que la típica foto digital que se hace hoy en día. Pero desde el punto de vista del análisis con datos masivos, examinaba más que una muestra aleatoria. Cuando hablamos de datos masivos, nos referimos al tamaño no tanto en términos absolutos como relativos: relativos al conjunto exhaustivo de datos.

Durante mucho tiempo, el muestreo aleatorio resultó un buen atajo. Hizo posible el análisis de problemas con un elevado número de datos en la era predigital. Pero, igual que sucede cuando se guarda una imagen o una canción digitales en un fichero más pequeño, al muestrear se pierde información. Disponer del conjunto de datos completo (o prácticamente completo) ofrece mucha más libertad para explorar, para estudiar los datos desde diferentes perspectivas, o para examinar más de cerca determinados aspectos.

Una analogía adecuada la brinda la cámara Lytro, que no captura un único plano de luz, como las cámaras convencionales, sino haces de todo el campo luminoso, unos once millones.[19] El fotógrafo puede decidir más tarde qué elemento del archivo digital desea enfocar. No es necesario enfocar para hacer la foto, ya que el recoger toda la información de entrada hace posible hacerlo a posteriori. Como se incluyen haces de todo el campo de luz, están todos los datos: N = todo. En consecuencia, la información es más "reutilizable" que la de las fotografías corrientes, en las que el fotógrafo tiene que decidir qué quiere enfocar antes de apretar el botón.

Igualmente, ya que los datos masivos se basan en toda la información, o por lo menos en toda la posible, nos permiten examinar detalles o explorar nuevos análisis sin correr el riesgo de que se vuelvan borrosos. Podemos someter a prueba nuevas hipótesis a muchos niveles distintos de granularidad. Esta cualidad es la que nos permite detectar el amaño de combates en el sumo, seguir la propagación del virus de la gripe región a región, y luchar contra el cáncer centrándonos en una porción precisa del ADN del paciente. Nos permite trabajar con un nivel asombroso de claridad.

Por descontado, no siempre es necesario utilizar todos los datos en lugar de una muestra. Seguimos viviendo en un mundo de recursos limitados. Pero usar toda la información que tengamos a mano sí tiene sentido cada vez en más casos; y el hacerlo es hoy factible, cuando antes no lo era.

Una de las áreas que se está viendo más acusadamente transformada por N = todo es la de la ciencias sociales.[20] Estas ciencias han perdido su monopolio sobre la interpretación de los datos sociales empíricos, mientras el análisis de datos masivos sustituye a los expertos del pasado. Las disciplinas de las ciencias sociales dependían fundamentalmente de estudios de muestras y cuestionarios. Pero cuando los datos se recogen de forma pasiva, mientras la gente sigue haciendo lo que hace de todas maneras en condiciones normales, los antiguos sesgos asociados con el muestreo y los cuestionarios desaparecen. Hoy en día, podemos recopilar información que antes no estaba a nuestro alcance, sean relaciones reveladas por llamadas de teléfono móvil o sentimientos expresados mediante tuits. Y, aún más importante: desaparece la necesidad de elaborar muestras.

Albert-László Barabási, una de las principales autoridades mundiales en la ciencia de la teoría de redes, quiso estudiar las interacciones de las personas a escala de toda la población. Para ello, él y sus colegas examinaron los registros anónimos de las llamadas de telefonía móvil a través de un operador inalámbrico que atendía a cerca de una quinta parte de la población de un país europeo sin identificar: todos los registros de un periodo de cuatro meses. Fue el primer análisis de redes a nivel societario usando un conjunto de datos que respetaba el espíritu de N = todo. El actuar a una escala tan grande, examinando todas las llamadas entre millones de personas a lo largo del tiempo, dio pie a nuevas percepciones que, probablemente, no habrían salido a la luz de ninguna otra manera.

Curiosamente, y en contraste con estudios similares, el equipo descubrió que si uno retira de la red a personas con muchos vínculos en el seno de su comunidad, la red social resultante se degrada, pero no falla. Por otra parte, cuando se retira de la red a personas con vínculos al margen de su comunidad inmediata, la red social se desintegra repentinamente, como si su estructura se hubiese venido abajo. Fue un resultado importante, pero un tanto inesperado. ¿Quién habría pensado que las personas con muchos amigos a su alrededor resultan harto menos importantes para la estabilidad de la red que aquellas con vínculos con gente más distante? Esto sugiere que existe una prima a la diversidad en el seno de un grupo, y en la sociedad en general.

Tendemos a pensar en el muestreo estadístico como una especie de fundamento inmutable, como los principios de la geometría o las leyes de la gravedad. Sin embargo, el concepto tiene menos de un siglo de vida, y fue desarrollado para resolver un problema particular en un momento dado, bajo restricciones tecnológicas específicas. Esas restricciones ya no existen con el mismo alcance. Echar mano de una muestra aleatoria en la era de los datos masivos es como aferrarse a una fusta de caballo en la era del motor de explosión. Todavía podemos recurrir al muestreo en algunos contextos, pero no tiene por qué ser –y de hecho, no será– la forma predominante que emplearemos para el análisis de grandes conjuntos de datos. Cada vez más, podremos ir a por todos.

III
CONFUSIÓN

El uso de todos los datos disponibles resulta factible cada vez en más contextos, pero implica un coste. El incremento de la cantidad le franquea la puerta a la inexactitud. Por descontado, siempre se han deslizado cifras erróneas y fragmentos corrompidos en los conjuntos de datos, pero la clave estaba en tratarlos como problemas e intentar deshacerse de ellos, en parte porque podíamos. Lo que nunca quisimos fue considerarlos inevitables y aprender a vivir con ellos. Este es uno de los cambios fundamentales de pasar a los datos masivos desde los escasos.

En un mundo de datos escasos, reducir los errores y garantizar la buena calidad era un impulso natural y esencial. Puesto que solo recogíamos un poco de información, nos asegurábamos de que esas cifras que nos molestábamos en recopilar fueran lo más exactas posible. Generaciones de científicos optimizaron sus instrumentos para hacer sus medidas cada vez más y más precisas, ya fuese para determinar la posición de los cuerpos celestes o el tamaño de los objetos en la lente de un microscopio. En un mundo de muestreo, la obsesión con la exactitud se hizo aún más crucial. Analizar solo un número limitado de puntos de datos implica que los errores pueden verse ampliados, reduciendo potencialmente la exactitud de los resultados totales.

Durante la mayor parte de la historia, los mayores logros de la humanidad han surgido de dominar el mundo midiéndolo. La búsqueda de la exactitud se inició en Europa a mediados del siglo XIII, cuando los astrónomos y los sabios se encargaron de la cuantificación precisa del tiempo y del espacio: de "la medida de la realidad", en palabras del historiador Alfred Crosby.[21]

La creencia implícita rezaba que quien podía medir un fenómeno, podía entenderlo. Más adelante, la medición se vinculó al método científico de la observación y la explicación: la capacidad de cuantificar, registrar y presentar resultados reproducibles. "Medir es saber", pronunció lord Kelvin.[22] Se convirtió en un respaldo de autoridad. "El conocimiento es poder", enseñaba Francis Bacon. Paralelamente, los matemáticos, y lo que más adelante llegarían a ser analistas de seguros y contables, desarrollaron métodos que hicieron posible recopilar, registrar y gestionar con exactitud los datos.

Llegado el siglo xix, Francia –por entonces la principal nación científica del planeta– había desarrollado un sistema de unidades de medida, definidas con precisión, para capturar el espacio, el tiempo y demás, y había empezado a lograr que otras naciones adoptasen los mismos estándares. Este proceso llegó hasta el punto de establecer en tratados internacionales unos prototipos de unidades de medida de aceptación universal que sirviesen de patrón. Fue el culmen de la edad de la medida. Apenas medio siglo después, en la década de 1920, los descubrimientos de la mecánica cuántica destruyeron para siempre el sueño de la medición exhaustiva y perfecta. Y aun así, al margen de un círculo relativamente pequeño de físicos, la mentalidad que inspiró el impulso humano de medir sin fallos persistió entre los ingenieros y los científicos. En el ámbito de los negocios incluso se expandió, cuando las ciencias racionales (estadística, matemáticas) empezaron a ejercer influencia sobre todas las áreas del comercio.

Sin embargo, en numerosas situaciones nuevas que están surgiendo hoy en día, tolerar la imprecisión –la confusión– puede resultar un rasgo positivo, no una deficiencia. Es una especie de término medio. A cambio de tolerar la relajación del número de errores permisibles, se pueden obtener muchos más datos. No se trata solo de que "más es mejor que algo", sino de que, en la práctica, a veces "más es mejor que mejor".

Son varias las clases de confusión a las que hay que enfrentarse. El término "confusión" puede referirse al mero hecho de que la probabilidad de error aumenta a medida que se añaden más puntos de datos. Por consiguiente, multiplicar por mil el número de mediciones de la tensión de un puente incrementa la posibilidad de que algunas puedan ser erróneas. Pero puede aumentarse asimismo la confusión al combinar diferentes tipos de información de fuentes distintas, que no siempre están perfectamente alineadas. Por ejemplo, usar un software de reconocimiento de voz para caracterizar las quejas recibidas en un centro de llamadas, y comparar esos datos con el tiempo que precisan los operadores para gestionar las llamadas, puede permitir obtener una foto imperfecta, si bien útil, de la situación. La confusión también puede referirse a la disparidad de formatos, por la que los datos necesitan ser "limpiados" antes de su procesamiento. Hay innumerables formas de referirse a IBM[23], apunta el experto en datos masivos DJ Patil, desde IBM a International Business Machines, pasando por T. J. Watson Labs. La confusión puede surgir al extraer o procesar los datos, ya que con ello los estamos transformando, convirtiéndolos en algo distinto, igual que cuando llevamos a cabo análisis de sentimientos en los mensajes de Twitter para predecir los resultados de taquilla en Hollywood. La confusión misma es confusa.

Supongamos que necesitamos medir la temperatura en un viñedo. Si no tenemos nada más que un sensor para toda la parcela, debemos asegurarnos de que sus mediciones son exactas y de que está funcionando todo el tiempo: no se permiten confusiones. Pero, si disponemos de un sensor para cada una de las cientos de vides, podremos usar sensores más baratos y menos sofisticados (mientras no introduzcan un sesgo sistemático). Existe la posibilidad de que, en algunos puntos, unos pocos sensores indiquen datos incorrectos, dando lugar a un conjunto de datos menos exacto, o más "confuso", que el que nos brindaría un único sensor preciso. Cualquier medición aislada puede ser incorrecta, pero la agregación de tantas mediciones ofrecerá una imagen mucho más exhaustiva, porque este conjunto, al consistir en más puntos de datos, es más valioso, lo que probablemente compensa su confusión.

Supongamos ahora que incrementamos la frecuencia de las mediciones del sensor. Si efectuamos una medición por minuto, podremos estar aceptablemente seguros de que la secuencia con la que lleguen los datos será perfectamente cronológica. Pero si cambiamos a diez o cien mediciones por segundo, la exactitud de la secuencia puede tornarse menos segura. Mientras la información recorre una red, un registro dado puede demorarse y llegar fuera de secuencia, o puede sencillamente perderse en la riada. La información será un poco menos precisa, pero su gran volumen hace que valga la pena renunciar a la exactitud estricta.

En el primer ejemplo, sacrificamos la precisión de cada punto de datos en aras de la amplitud, y recibimos a cambio un grado de detalle que no habríamos podido ver de otro modo. En el segundo caso, renunciamos a la exactitud por la frecuencia, y a cambio percibimos un cambio que de otra forma se nos habría escapado. Aunque puede que consigamos superar los errores si dedicamos a ello recursos suficientes –al fin y al cabo, en la Bolsa de Nueva York[24], donde la secuencia correcta tiene mucha importancia, tienen lugar treinta mil transacciones por segundo–, muchas veces resulta más provechoso tolerar el error que esforzarse en prevenirlo.

Por ejemplo, podemos aceptar cierto grado de confusión a cambio de una escala mayor. Como dice la consultoría tecnológica Forrester: "A veces, 2 más 2 puede ser igual a 3,9, y con eso basta". Por supuesto, los datos no pueden ser completamente incorrectos, pero estamos dispuestos a sacrificar un poco de exactitud si a cambio descubrimos la tendencia general. Los datos masivos convierten los cálculos aritméticos en algo más probabilístico que preciso. Este es un cambio al que va a costar mucho acostumbrarse, y trae sus propios problemas, que consideraremos más adelante en este libro. Por ahora, basta simplemente con apuntar que muchas veces necesitaremos tolerar la confusión cuando incrementemos la escala.

Puede advertirse un cambio similar, comparando la importancia de tener más datos en relación con otras mejoras, en el campo de la computación. Todo el mundo sabe cuánto ha aumentado a lo largo de los años la capacidad de procesamiento, como predijo la ley de Moore, que estipula que el número de transistores en un chip viene a duplicarse cada dos años. Este perfeccionamiento continuo ha hecho que los ordenadores sean más rápidos y con más memoria. Pero no todo el mundo sabe que la prestación de los algoritmos que impulsan muchos de nuestros sistemas también ha aumentado: en muchas áreas, más que la mejora de los procesadores según la ley de Moore. Muchos de los beneficios para la sociedad que traen los datos masivos, sin embargo, se producen no tanto por los chips más rápidos o los mejores algoritmos[25], sino porque hay más datos.

Por ejemplo, los algoritmos de ajedrez han cambiado muy poco en las últimas décadas, puesto que las reglas del ajedrez son sobradamente conocidas y muy rígidas. La razón de que los programas informáticos de ajedrez jueguen mejor hoy en día que antes se debe, entre otras cosas, a que juegan mejor la fase final de la partida. Y lo hacen sencillamente porque a los sistemas se les han suministrado más datos. De hecho, las fases finales, en las que solo quedan en el tablero seis piezas o menos, han sido analizadas por completo y todas las jugadas posibles (N = todo) han sido representadas en una tabla masiva que, una vez descomprimida, ocupa más de un terabyte de datos. Eso permite que los programas de ajedrez jueguen sin fallos las fases finales de las partidas de ajedrez, que son las cruciales. Ningún ser humano podrá superar nunca al sistema.

El hecho de que más datos es preferible a mejores algoritmos ha quedado demostrado de forma contundente en el campo del procesamiento del lenguaje natural: la forma en la que los ordenadores aprenden a analizar las palabras según las usamos en el habla cotidiana. Hacia el año 2000, dos investigadores de Microsoft, Michele Banko y Eric Brill, estaban buscando un método para mejorar el corrector gramatical que incorpora el programa Word de la compañía. No estaban muy seguros de a qué resultaría más útil dedicar su esfuerzo: si a mejorar los algoritmos existentes, a hallar nuevas técnicas, o a añadir características más sofisticadas. Antes de elegir una de esas vías, decidieron comprobar qué ocurriría si introducían muchos más datos en los métodos existentes. La mayoría de los algoritmos de aprendizaje de máquinas se basan en un corpus textual de un total de un millón de palabras o menos. Banko y Brill tomaron cuatro algoritmos corrientes y les introdujeron más datos en hasta tres órdenes de magnitud: diez millones de palabras, luego cien millones, y, por último, mil millones de palabras.

Los resultados fueron asombrosos. Conforme iban entrando más datos, el rendimiento de los cuatro tipos de algoritmos mejoró drásticamente. De hecho, un algoritmo simple, que era el que peor funcionaba con medio millón de palabras, operaba mejor que ningún otro en cuanto procesaba mil millones de palabras. Su grado de precisión pasó del 75 por 100 a más del 95 por 100. A la inversa, el algoritmo que mejor funcionaba con pocos datos fue el que peor resultado dio con grandes cantidades, aunque, al igual que los demás, mejoró mucho, pasando de alrededor del 86 por 100 a

casi el 94 por 100 de exactitud. "Estos resultados nos llevan a pensar que quizá debamos reconsiderar la disyuntiva entre gastar tiempo y dinero en el desarrollo de algoritmos frente a gastarlo en el desarrollo del corpus", escribieron Banko y Brill en uno de sus artículos científicos sobre el tema.

Así que más cantidad es mejor que menos, y a veces más cantidad es mejor que más inteligencia. ¿Qué pasa entonces con la confusión? Pocos años después de que Banko y Brill recogieran todos esos datos, los investigadores de la empresa rival, Google, pensaban en algo similar, pero a una escala aún mayor. En lugar de probar los algoritmos con mil millones de palabras, emplearon un billón. Google no lo hizo para desarrollar un corrector gramatical, sino para cascar una nuez aún más compleja: la traducción de idiomas.

La llamada traducción automática ha sido un objetivo de los pioneros de la informática desde el alba de los ordenadores, en la década de 1940, cuando los aparatos estaban hechos de lámparas de vacío y ocupaban una habitación entera. La idea cobró particular urgencia durante la Guerra Fría, cuando Estados Unidos capturaba ingentes cantidades de material escrito y hablado en ruso pero carecía de la fuerza laboral para traducirlo rápidamente.

Al principio, los informáticos optaron por una combinación de reglas gramaticales y un diccionario bilingüe. Un ordenador de IBM tradujo sesenta frases rusas al inglés en 1954, usando doscientos cincuenta pares de palabras[26] en el vocabulario del ordenador y seis reglas de gramática. Los resultados fueron muy prometedores. Se introdujo *"Mi pyeryedayem mislyi posryedstvom ryechyi"* en el ordenador IBM 701 por medio de tarjetas perforadas, y salió: "Transmitimos pensamientos por medio del habla". Las sesenta frases se tradujeron sin problemas, según una nota de prensa de la empresa, celebrando su logro. El director del programa de investigación, Leon Dostert, de la universidad de Georgetown, pronosticó que la traducción automática sería un "hecho acabado" en un plazo de "cinco, puede que tres años".

Pero el éxito inicial resultó ser un espejismo. Ya en 1966 un comité de expertos en traducción automática tuvo que reconocer su fracaso. El problema era más arduo de lo que habían pensado. Enseñar a un ordenador a traducir tiene que ver con enseñarle no solo las reglas, sino también las excepciones. La traducción no consiste únicamente en memorizar y recordar; también se trata de elegir las palabras correctas de entre muchas opciones. ¿Es *bonjour* realmente "buenos días"? ¿O es "buen día", "hola", o "qué tal"? La respuesta: depende…

A finales de la década de 1980, los investigadores de IBM dieron con una idea novedosa. En vez de tratar de introducir en un ordenador las reglas lingüísticas explícitas junto con un diccionario, decidieron permitir que el computador emplease la probabilidad estadística para calcular qué palabra o frase de un idioma dado era la más adecuada en otro. En la década de 1990, el programa Candide de IBM[27] utilizó el equivalente a diez años de transcripciones de sesiones del Parlamento de Canadá publicadas en francés y en inglés: unos tres millones de pares de frases. Al tratarse de documentos oficiales, las traducciones eran de altísima calidad y, para los estándares de la época, la cantidad de datos era enorme. La traducción estadística automática, como llegó a ser conocida la técnica, convirtió hábilmente el desafío de la traducción en un gran problema matemático. Y pareció dar resultado. De repente, la traducción por ordenador mejoró mucho. Tras el éxito que supuso ese salto conceptual, sin embargo, IBM solo logró pequeñas mejoras, pese a invertir montones de dinero. Y a la larga, acabó cerrando el grifo.

Pero menos de una década después, en 2006, Google se lanzó a traducir, dentro de su objetivo de "organizar la información del mundo y hacerla universalmente accesible y útil". En lugar de páginas

de texto bien traducidas en dos idiomas, Google utilizó un conjunto de datos más vasto, pero también mucho más confuso: todo el contenido global de internet. Su sistema absorbió todas las traducciones que pudo encontrar, para entrenar al ordenador. Así, entraron páginas web corporativas en múltiples idiomas, traducciones idénticas de documentos oficiales e informes de organismos intergubernamentales como las Naciones Unidas y la Unión Europea. Se incluyeron hasta traducciones de libros del proyecto de escaneo de libros de Google. Mientras que Candide había usado tres millones de frases cuidadosamente traducidas, el sistema de Google aprovechó miles de millones[28] de páginas de traducciones de calidad muy variable, según el director de Google Translate, Franz Josef Och, una de las autoridades punteras en este campo. Su corpus de un billón de palabras representaba noventa y cinco mil millones de frases en inglés, aunque fueran de dudosa calidad.

Pese a lo confuso de la información que se le aportó, el servicio de Google es el que mejor funciona. Sus traducciones son más precisas que las de los demás sistemas (aun cuando siguen siendo muy imperfectas). Y es mucho, muchísimo más rico. A mediados de 2012, su base de datos cubría más de sesenta idiomas. Incluso podía aceptar entradas de voz en catorce idiomas para efectuar traducciones fluidas. Como trata el lenguaje sencillamente como un conjunto de datos confusos con los que estimar probabilidades, puede incluso traducir entre idiomas para los que existen escasas traducciones directas que añadirle, por ejemplo, el hindi y el catalán. En esos casos, recurre al inglés como puente. Y es mucho más flexible que otras aproximaciones, puesto que puede añadir y retirar palabras conforme vayan introduciéndose y cayendo en desuso.

La razón por la que el sistema de traducción de Google funciona bien no es porque disponga de un algoritmo más inteligente. Funciona bien porque sus creadores, como hicieron Banko y Brill en Microsoft, lo abastecieron de más datos, y no solo de alta calidad. Google fue capaz de usar un conjunto de datos *decenas de miles* de veces mayor que el del Candide de IBM porque aceptó la confusión. El corpus de un billón de palabras que Google dio a conocer en 2006 se recopiló a partir de todo el aluvión de contenido de internet; "datos salvajes", por así decir. Ese fue el "conjunto de datos de entrenamiento" mediante el cual el sistema pudo calcular la probabilidad, por ejemplo, de que una palabra siguiese a otra en inglés. Era muy distinto de su abuelo en este campo, el célebre corpus de Brown[29] de la década de 1960, que suponía un total de un millón de palabras inglesas. El usar el conjunto de datos más amplio permitió grandes avances en el procesamiento de lenguajes naturales, sobre el que se basan los sistemas para tareas como el reconocimiento de voz y la traducción por ordenador. "Los modelos simples y con un montón de datos vencen a los sistemas más elaborados basados en menos datos", escribió Peter Norvig, gurú de la inteligencia artificial de Google, junto con unos colaboradores en un artículo titulado "La efectividad irrazonable de los datos".

Como explicaron Norvig y sus coautores, la clave estaba en la confusión: "De alguna forma, este corpus supone un paso atrás respecto al corpus de Brown: procede de páginas web sin depurar, por lo que contiene oraciones truncadas, errores ortográficos, gramaticales y de todo tipo. No se ha anotado, etiquetado ni corregido cuidadosamente a mano las distintas partes de la oración. Aun así, el hecho de ser un millón de veces más amplio que el corpus de Brown compensa esas desventajas".

MÁS ES MEJOR QUE MEJOR

A los analistas convencionales de muestras, que se han pasado la vida centrados en prevenir y erradicar la imprecisión, les resulta difícil aceptarla. Se esfuerzan con ahínco en reducir las tasas de error al recoger sus muestras, y en someterlas a prueba para detectar sesgos potenciales antes de anunciar sus hallazgos. Recurren a múltiples estrategias de reducción de error, entre ellas la de asegurarse de que sus muestras las han recogido, siguiendo un protocolo preciso, unos expertos especialmente formados con ese fin. Tales estrategias resultan caras de implementar, incluso para un número limitado de puntos de datos, y casi imposibles para datos masivos. No solo resultarían demasiado costosas, sino que no habría forma de garantizar los estándares exigentes de recogida de datos a tamaña escala. Ni siquiera eliminando la interacción humana se resolvería el problema.

Entrar en un mundo de datos masivos requerirá que cambiemos nuestra forma de pensar acerca de los méritos de la exactitud. Aplicar la mentalidad de medición clásica al mundo digital y conectado del siglo xxi supone cometer un error de bulto. Como ya se ha mencionado, la obsesión con la exactitud es un resabio de la era analógica privada de información. Cuando los datos eran escasos, cada punto de datos resultaba crucial, por lo que se ponía gran cuidado en evitar que viniera uno a sesgar el análisis.

Hoy en día no vivimos ya esa situación de carestía de información. Al tratar con conjuntos de datos cada vez más amplios, que captan no solo un pequeño fragmento del fenómeno en cuestión, sino muchas más partes del mismo, ya no necesitamos preocuparnos tanto por unos puntos de datos individuales que puedan sesgar el análisis global. Más que aspirar a erradicar todo atisbo de inexactitud a un coste cada vez más elevado, calculamos con la confusión en mente.

Considérese la forma en que los sensores se van abriendo camino en las fábricas. En la refinería Cherry Point de BP en Blaine (Washington) se han instalado sensores inalámbricos por toda la planta, formando una red invisible que recaba grandes cantidades de datos en tiempo real. El ambiente de calor intenso y la maquinaria eléctrica a veces distorsionan las mediciones, y producen datos confusos. Sin embargo, la enorme cantidad de información que generan los sensores, tanto los cableados como los inalámbricos, compensa sobradamente esos escollos. Basta con incrementar la frecuencia y el número de localizaciones de las medidas de los sensores para que el resultado merezca la pena. Al medir la tensión de las tuberías a todas horas, en vez de a intervalos precisos, BP[30] descubrió que algunas clases de crudo de petróleo son más corrosivas que otras, una cualidad que no podía detectar, ni por consiguiente solucionar, cuando su conjunto de datos era más pequeño.

Cuando la cantidad de datos es enormemente mayor y de un tipo nuevo, la exactitud ya no sigue siendo el objetivo en algunos casos, siempre y cuando podamos captar la tendencia general. Pasar a una escala grande cambia no solo las expectativas de precisión, sino también la capacidad práctica de alcanzar la exactitud. Aunque en un primer momento pueda parecer contraintuitivo, tratar los datos como algo imperfecto e impreciso nos permite afinar en los pronósticos, y así comprender mejor nuestro mundo.

Merece la pena señalar que la confusión no es algo inherente a los datos masivos. Se trata, por el contrario, de una función de la imperfección de las herramientas que usamos para medir, registrar y analizar la información. Si la tecnología llegara a ser perfecta, el problema de la inexactitud desaparecería. Pero mientras siga siendo imperfecta, la confusión es una realidad práctica con la que tenemos que contar. Y es probable que nos acompañe durante mucho tiempo. En muchos aspectos, los esfuerzos agónicos por incrementar la exactitud no tendrán sentido económico, puesto que nos importará más el valor de disponer de conjuntos de datos mucho mayores. Al igual que los estadísticos de otras épocas dejaron de lado su interés por las muestras de mayor tamaño en aras de

más aleatoriedad, podemos vivir con un poco de imprecisión a cambio de más datos.

El Proyecto de los Mil Millones de Precios[31] constituye un ejemplo curioso de esto. Todos los meses, la oficina de estadísticas laborales de Estados Unidos publica el índice de precios al consumo, o CPI, que se emplea para calcular la tasa de inflación. Este dato resulta crucial para los inversores y las empresas. La reserva federal lo tiene en cuenta para decidir si sube o baja los tipos de interés. Las empresas basan los sueldos en la inflación. El gobierno federal la utiliza para indexar pagos, como las pensiones de la seguridad social, y el interés que paga sobre determinados bonos.

Para obtener el porcentaje, la oficina de estadísticas laborales utiliza a cientos de sus empleados que llaman, envían faxes, y visitan tiendas y oficinas en noventa ciudades por toda la nación, informando luego de unos ochenta mil precios de todas las cosas, desde los tomates hasta las carreras de taxi. Elaborarlo cuesta alrededor de 250 millones de dólares al año. Por ese precio, los datos son nítidos, limpios y ordenados. Pero, para cuando se difunden las cifras, ya tienen unas semanas. Como demostró la crisis financiera de 2008, unas pocas semanas pueden suponer un desfase terriblemente largo. Los que toman las decisiones necesitan un acceso más rápido a las cifras de la inflación para poder reaccionar mejor a ellas, pero con métodos convencionales, centrados en la precisión del muestreo y de la fijación de precios, no lo tienen.

Dos economistas del Massachusetts Institute of Technology (MIT), Alberto Cavallo y Roberto Rigobon, se enfrentaron a este problema hallando una alternativa basada en datos masivos que seguía un rumbo mucho menos preciso. Empleando programas de búsqueda en la red, recopilaron medio millón de precios de productos vendidos en su país a diario. La información es confusa, y no todos los puntos de datos recogidos son fácilmente comparables. Sin embargo, al comparar la recolección de datos masivos con un análisis inteligente, el proyecto fue capaz de detectar un giro deflacionario en los precios inmediatamente después de que Lehman Brothers se declarase en quiebra en septiembre de 2008, mientras que quienes se basaban en los datos oficiales del CPI tuvieron que aguardar hasta noviembre para observarlo.

El proyecto del MIT ha dado lugar a una empresa comercial llamada PriceStats, a la que recurren la banca y otras instituciones antes de tomar decisiones económicas. PriceStats recopila millones de productos vendidos a diario por cientos de minoristas en más de setenta países. Por supuesto, las cifras requieren una interpretación muy cuidadosa, pero son mejores que las estadísticas oficiales a la hora de indicar las tendencias de la inflación. Como hay más precios y las cifras están disponibles en tiempo real, les ofrecen a los que tienen que tomar decisiones una ventaja significativa. (El método también sirve como comprobación externa creíble de los organismos estadísticos nacionales.[32] Por ejemplo, *The Economist* desconfía del método que emplea Argentina para calcular la inflación, así que se basa en las cifras de PriceStats para ello).

LA CONFUSIÓN EN ACCIÓN

En numerosas áreas de la tecnología y la sociedad, nos estamos inclinando a favor de lo más abundante y confuso, antes que menos y exacto. Pensemos en el caso de la categorización de contenidos. A lo largo de los siglos, los seres humanos hemos desarrollado taxonomías e índices con el fin de poder almacenar y recuperar material. Estos sistemas jerárquicos siempre han sido imperfectos, como podrá atestiguar con pesar cualquiera que esté familiarizado con el sistema de catalogación de una biblioteca; pero, cuando los datos eran escasos, funcionaban aceptablemente

bien. Sin embargo, si se aumenta la escala en muchos órdenes de magnitud, estos sistemas, que parten del supuesto de la perfecta colocación de todo lo que hay en ellos, se vienen abajo. Por ejemplo, en 2011, la web Flickr, dedicada a compartir fotos, contenía más de seis mil millones de fotos de más de setenta y cinco millones de usuarios. Habría sido inútil intentar etiquetar cada foto de acuerdo con unas categorías preestablecidas. ¿Realmente habría existido una titulada "Gatos que se parecen a Hitler"?

En vez de eso, las taxonomías estrictas están dejando el sitio a unos mecanismos más imprecisos pero eminentemente más flexibles y adaptables a un mundo que evoluciona y cambia. Cuando subimos fotos a Flickr[33], las "etiquetamos". Es decir, que les asignamos un número cualquiera de etiquetas textuales, y las usamos para organizar el material y buscar en él. Las etiquetas las crean y colocan los usuarios de forma *ad hoc:* no hay categorías estándar predefinidas, ninguna taxonomía preexistente a la que haya que ceñirse. Más bien, cualquiera puede añadir nuevas etiquetas, *tags*, con solo escribirlas. El etiquetado se ha impuesto como el estándar *de facto* para la clasificación en internet, y se emplea en las redes sociales como Twitter, en blogs, etc. Hace que sea más navegable el vasto contenido de la red, especialmente por lo que se refiere a contenidos como imágenes, vídeos y música que no están basados en texto, para los que las búsquedas por palabras no dan resultado.

Por supuesto, puede que algunas etiquetas estén mal escritas, y esa clase de errores introduce imprecisión: no en los datos mismos, pero sí en cómo se organizan. Esto puede molestar a quienes estaban acostumbrados a la exactitud. Ahora bien, a cambio de cierta imprecisión en la forma de organizar nuestras colecciones de fotos, adquirimos un universo de etiquetas mucho más rico, y, por extensión, un acceso más profundo y amplio a nuestras fotos. Podemos combinar las etiquetas de búsqueda para filtrar las fotos de maneras que antes no eran posibles. La imprecisión inherente al etiquetado implica aceptar el desorden natural del mundo. Es un antídoto para sistemas más precisos que intentan imponer una esterilidad falaz sobre el tumulto de la vida real, fingiendo que todo cuanto hay bajo el sol puede disponerse en unas filas y columnas ordenadas. Hay más cosas en el cielo y en la tierra de las que sueña esa filosofía.

Muchas de las páginas web más populares de la red hacen gala de su admiración por la imprecisión en lugar de aspirar a la exactitud. Si se fija uno en un icono de Twitter o en un botón "Me gusta" de Facebook en cualquier página web, verá que muestran el número de personas que han hecho clic en ellos. Cuando la cifra es pequeña, se ven todos los clics, "63", por ejemplo. Pero conforme van aumentando, el número que se muestra es una aproximación, como "4K". No es que el sistema no conozca el total exacto: es que, conforme va aumentando la escala, mostrar la cifra exacta resulta menos importante. Además, las cantidades pueden cambiar tan deprisa que una cifra dada podría quedar desfasada nada más aparecer. Igualmente, el correo electrónico Gmail de Google presenta la hora de llegada de los mensajes más recientes con toda exactitud, como "hace 11 minutos", pero despacha las duraciones más largas con un displicente "hace 2 horas", como también hacen Facebook y otros.

La industria de la inteligencia empresarial y el software analítico se levanta de antiguo sobre la promesa a los clientes de "una única versión de la verdad": el popular cliché alrededor del año 2000 en labios de los vendedores de tecnología de estas áreas. Los ejecutivos usaban la frase sin intención irónica. Algunos lo siguen haciendo. Con ella, lo que quieren decir es que cualquiera que acceda a los sistemas de tecnología de la información de una empresa puede disponer de los mismos datos; que el equipo de márketing y el de ventas no tienen que pelearse por quién tiene las cifras de ventas o de clientes correctas antes de que empiece la reunión. Sus intereses podrían estar más en línea si los

hechos fueran coherentes, suele pensarse.

Pero la idea de "una única versión de la realidad" está por cambiar de partido. Estamos empezando a comprender no solo que a lo mejor es imposible que exista una única versión de la realidad, sino también que perseguirla es una pérdida de tiempo. Para acceder a los beneficios de la explotación de los datos a escala, tenemos que aceptar que la imprecisión es lo normal y esperable, no algo que debamos tratar de eliminar.

Incluso estamos empezando a ver cómo el espíritu de la inexactitud invade una de las áreas más intolerantes con la imprecisión: el diseño de bases de datos. Los motores tradicionales de bases de datos requerían que los datos fuesen muy precisos y estructurados. Los datos no se almacenaban sin más: se partían en "archivos" que contenían campos. Cada campo incluía información de un tipo y longitud determinados. Por ejemplo, si un campo numérico tenía siete dígitos de longitud, no podía archivarse en él una suma de diez millones o más. Si uno quería introducir "No disponible" en un campo para números telefónicos, no se podía hacer. Para dar acomodo a estas entradas había que alterar la estructura de la base de datos. Aún seguimos peleando contra estas restricciones cuando el software de nuestros ordenadores y teléfonos inteligentes no acepta los datos que deseamos introducir.

También los índices tradicionales estaban predefinidos, y eso limitaba lo que uno podía buscar. Para añadir un índice nuevo había que crearlo desde cero, lo que requería tiempo. Las bases de datos clásicas, las llamadas relacionales, están pensadas para un mundo en el que los datos son escasos, por lo que pueden seleccionarse con mucho cuidado. Es ese mundo las preguntas para las que uno busca respuesta tienen que estar claras de entrada, y la base de datos está diseñada para darles respuesta –y solo a ellas– de forma eficiente.

Pero este concepto del almacenamiento y análisis está cada vez más reñido con la realidad. Ahora disponemos de grandes cantidades de datos, de clase y calidad variables. Raras veces encajan en alguna de las categorías definidas con precisión que se conocen de antemano. Y las preguntas que queremos hacer a menudo surgen solo cuando recogemos los datos y empezamos a trabajar con ellos.

Estas realidades han dado pie a diseños novedosos de bases de datos que rompen con los principios de antaño: principios de archivos y campos predefinidos que reflejan nítidas jerarquías de información. El lenguaje más corriente para acceder a las bases de datos ha sido desde hace mucho tiempo el SQL, *Structured Query Language* o "Lenguaje de Pregunta Estructurado". Su mismo nombre evoca rigidez. Pero el gran cambio en años recientes se ha producido hacia algo llamado "no SQL", que no requiere una estructura de archivo predeterminada para operar. El no SQL acepta datos de clases y dimensiones variables y permite efectuar búsquedas en ellos. A cambio de tolerar el desorden estructural, estos diseños de bases de datos requieren más capacidad de procesamiento y almacenaje. Pero es un compromiso que podemos aceptar a la vista de la caída en picado de esos costes.

Pat Helland[34], una de las primeras autoridades mundiales en diseño de bases de datos, describe este cambio fundamental en un artículo titulado "Cuando se tienen demasiados datos, 'Con eso basta' es bastante". Después de identificar algunos de los principios nucleares del diseño tradicional que se han visto erosionados por los datos desestructurados de proveniencia y exactitud diversas, expone las consecuencias: "Ya no podemos fingir que vivimos en un mundo limpio". El procesamiento de datos masivos ocasiona una inevitable pérdida de información –Helland lo llama "con pérdida"–, pero lo compensa con un resultado rápido. "No pasa nada si obtenemos respuestas con pérdida; muchas veces, eso es lo que necesita el negocio", concluye Helland.

El diseño tradicional de bases de datos promete suministrar resultados coherentes en el tiempo. Por ejemplo, si uno pide su saldo bancario, espera recibir la cantidad exacta. Y si vuelve a preguntar unos segundos más tarde, espera que el sistema le ofrezca el mismo resultado, suponiendo que todo siga igual. Sin embargo, conforme aumentan la cantidad de datos recogidos y el número de usuarios que acceden al sistema, resulta más difícil mantener esa coherencia.

Los grandes conjuntos de datos no existen en un único lugar: tienden a estar repartidos entre múltiples discos duros y ordenadores. Para garantizar fiabilidad y rapidez, un registro puede estar archivado en o dos o tres emplazamientos distintos. Cuando se actualiza el registro en un lugar, los datos en las demás localizaciones dejan de ser correctos hasta que se actualizan a su vez. Así pues, los sistemas tradicionales experimentarían un retraso hasta que se completaran todas las actualizaciones, y eso no resulta práctico cuando los datos están ampliamente distribuidos, y el servidor recibe el bombardeo de decenas de miles de consultas cada segundo. Por el contrario, aceptar el desorden es una especie de solución.

Este cambio lo ejemplifica la popularidad de Hadoop, un rival de fuente abierta del sistema MapReduce de Google, que es muy bueno a la hora de procesar grandes cantidades de datos. Esto lo consigue fragmentando los datos en porciones más pequeñas, y repartiéndolas por otros ordenadores. Cuenta con que el hardware pueda fallar, así que la redundancia ya va incorporada. También da por supuesto que los datos no están limpios ni ordenados; de hecho, asume que la cantidad de datos es demasiado enorme para poder limpiarlos antes de procesarlos. Mientras que el análisis de datos típico requiere una operación llamada "extraer, transferir y cargar" [*extract, transfer, and load*, o ETL], para desplazar los datos al lugar donde serán analizados, Hadoop prescinde de tantas finezas. En cambio, da por supuesto que la cantidad de información es tan increíblemente enorme que resulta imposible desplazarla, y que ha de ser analizada ahí donde está.

La producción de Hadoop no es tan precisa como la de las bases de datos relacionales: no es de fiar para lanzar un cohete al espacio ni para certificar los movimientos de una cuenta bancaria. Sin embargo, para muchas tareas de importancia menos crucial, donde no se requiere una respuesta ultraprecisa, cumple su cometido mucho más rápido que las demás acciones. Piénsese en tareas como segmentar una lista de clientes para hacer a algunos de ellos destinatarios de una campaña específica de márketing. Usando Hadoop[35], la compañía de tarjetas de crédito VISA fue capaz de reducir el tiempo de procesamiento de dos años enteros de registros de prueba, unas 73.000 millones de transacciones, de un mes a trece minutos escasos. Esa suerte de aceleración del procesamiento es la que transforma los negocios.

La experiencia de ZestFinance, una compañía fundada por el antiguo director de sistemas informáticos de Google, Douglas Merrill, refuerza el mismo argumento. Su tecnología ayuda a los prestamistas a decidir si deben ofrecer o no préstamos relativamente pequeños y a corto plazo a personas que aparentemente tienen poco crédito. Sin embargo, mientras que la valoración crediticia habitual se basa solo en un puñado de señales fuertes, como anteriores atrasos en los pagos, ZestFinance analiza una cantidad enorme de variables "más débiles". En 2012, podía presumir de una tasa de impago inferior en un tercio a la media del sector. Ahora bien, la única manera de hacer que funcione el sistema es incorporando la confusión.

"Una de las cosas interesantes –dice Merrill– es que nadie rellena todos los campos: siempre falta un buen montón de datos". La matriz de la información recopilada por ZestFinance es increíblemente escasa: un archivo de base de datos repleto de celdas vacías. Así que la compañía "imputa" los datos que faltan. Por ejemplo, alrededor del 10 por 100 de los clientes de ZestFinance

están listados como muertos, pero resulta que devuelven el préstamo igual. "Así que, obviamente, cuando toca prepararse para el apocalipsis zombi, la mayor parte de la gente asume que no se pagará ninguna deuda. Ahora bien, de acuerdo con nuestros datos, parece que los zombis liquidan sus deudas", añade Merrill con ironía.

A cambio de vivir con el desorden, obtenemos servicios tremendamente valiosos que resultarían imposibles a esa escala y alcance con los métodos e instrumentos tradicionales. Según algunas estimaciones, solo el 5 por 100 de todos los datos digitales están "estructurados", es decir, en una forma que encaja limpiamente en una base de datos[36] tradicional. Si no se acepta la confusión, el 95 por 100 restante de datos sin estructurar, como páginas web y vídeos, permanecen en la oscuridad. Tolerando la imprecisión, abrimos una ventana a un universo de perspectivas por explotar.

Nuestra sociedad ha aceptado dos acuerdos implícitos que han llegado a integrarse de tal manera en nuestra forma de actuar que ya ni siquiera los reconocemos como tales, sino como el estado natural de las cosas. En primer lugar, presuponemos que no podemos utilizar muchos más datos, así que no lo hacemos. Pero esa restricción resulta cada vez menos importante, y hay muchas ventajas en usar algo que se aproxime a N = todo.

El segundo acuerdo tiene que ver con la calidad de la información. Resultaba racional dar prioridad a la exactitud en una época de pocos datos, porque cuando recogíamos información limitada su precisión tenía que ser la mayor posible. En muchos casos, puede que esto todavía importe. Pero, para muchas otras cosas, la exactitud rigurosa resulta menos importante que obtener una percepción rápida de su contorno general o de su progreso en el tiempo.

Cómo pensamos en usar la totalidad de la información en vez de pequeños fragmentos, y cómo lleguemos a apreciar el descuido en lugar de la precisión, tendrán profundas consecuencias en nuestra interacción con el mundo. Conforme las técnicas de datos masivos vayan convirtiéndose en parte habitual de la vida cotidiana, nosotros en tanto que sociedad podremos empezar a esforzarnos en comprender el mundo desde una perspectiva más amplia e integral que antes, una especie de N = todo mental. Y podremos tolerar lo borroso y lo ambiguo en áreas en las que solíamos exigir claridad y certeza, aun cuando se tratase de una claridad falsa y de una certeza imperfecta. Podremos aceptarlo, siempre y cuando obtengamos a cambio un sentido más completo de la realidad: es el equivalente de un cuadro impresionista, en el que cada pincelada resulta confusa si se la examina de cerca, pero basta con apartarse uno del cuadro para contemplar una imagen majestuosa.

El enfoque de los datos masivos, que pone el acento en los conjuntos de datos de gran extensión y en la confusión, nos ayuda a acercarnos a la realidad más que nuestra antigua dependencia de los datos escasos y la exactitud. El atractivo de términos como "algunos" y "ciertos" resulta comprensible. Puede que nuestra comprensión del mundo fuese incompleta y ocasionalmente errónea cuando las posibilidades de análisis eran limitadas, pero existía una confortable seguridad en ello, una estabilidad tranquilizadora. Además, como estábamos constreñidos en cuanto a los datos que podíamos recopilar y examinar, no hacíamos frente a la misma compulsión por conseguirlo todo, verlo todo desde todos los ángulos posibles. En los estrechos confines de los datos escasos, podíamos enorgullecernos de nuestra precisión… aun cuando, al medir las minucias hasta el infinito, los árboles no nos dejaran ver el bosque.

En última instancia, los datos masivos pueden exigirnos cambiar, sentirnos más cómodos con el desorden y la imprecisión. Las estructuras de exactitud que parecen proporcionarnos coordenadas en la vida –que la pieza redonda va en el agujero circular; que hay una única respuesta a cada pregunta– son más maleables de lo que admitimos; y sin embargo, reconocer esta plasticidad y aceptarla nos

acerca más a la realidad.

Por muy radical que sea la transformación que suponen estas modificaciones en la forma de pensar, conducen además a un tercer cambio con potencial para tirar abajo una convención social aún más fundamental: la idea de comprender las razones que hay detrás de cuanto sucede. Por el contrario, como explica el capítulo siguiente, con encontrar asociaciones entre los datos y actuar en función de ellas a menudo puede ser más que suficiente.

V
DATIFICACIÓN

Matthew Fontaine Maury[56] era, en 1839, un prometedor oficial de la armada estadounidense, de camino a su nuevo destino en el bergantín *Consort*, cuando su diligencia se salió repentinamente del camino, volcó y lo lanzó por los aires. Maury cayó mal, fracturándose el fémur y dislocándose la rodilla. Un médico de la vecindad le volvió a poner la articulación en su sitio, pero la fractura no cerró bien y hubo que volver a romper el hueso unos días después. Las lesiones dejaron a Maury parcialmente tullido e incapacitado para navegar, con solo treinta y tres años. Después de casi tres años recuperándose, la armada lo mandó a un despacho, para dirigir el Depósito de Cartas de Navegación e Instrumentos, de nombre poco inspirado.

Pero aquel resultó ser el destino perfecto para él. De joven, a Maury lo dejaba perplejo que los barcos navegaran en zigzag por el mar en vez de seguir una ruta más directa. Cuando le preguntaba a los capitanes al respecto, le contestaban que era mucho mejor seguir un curso familiar que arriesgarse por uno menos conocido y con posibles peligros ocultos. Contemplaban el océano como un reino impredecible en el que los marinos se enfrentaban a lo inesperado con cada viento y cada ola.

Pero Maury sabía por sus viajes que esto no era enteramente cierto. Él advertía patrones en todas partes. Durante una escala prolongada en Valparaíso (Chile), fue testigo de cómo los vientos funcionaban como un reloj. Un vendaval a última hora de la tarde cesaba bruscamente al ponerse el sol y se convertía en una suave brisa, como si alguien hubiese cerrado un grifo. En otro viaje atravesó las cálidas aguas azules de la corriente del Golfo al paso de esta entre los muros oscuros de las aguas del Atlántico, tan distinguible y fija como si fuera el río Mississippi. De hecho, los portugueses habían navegado por el Atlántico durante siglos apoyándose en los vientos uniformes de levante y poniente conocidos como “alisios”.

Cada vez que el guardiamarina Maury llegaba a un puerto nuevo, se dedicaba a buscar a antiguos capitanes retirados para beneficiarse de su conocimiento, basado en experiencias transmitidas a lo largo de las generaciones. Aprendió acerca de mareas, vientos y corrientes marinas que funcionaban con regularidad y que no aparecían en los libros y mapas que la Armada facilitaba a sus marinos. Al contrario, estos libros se basaban en cartas que a veces tenían cien años, muchas de ellas con omisiones importantes o muy inexactas. En su nuevo puesto como superintendente del Depósito de Cartas de Navegación e Instrumentos, Maury se propuso arreglar esto.

Al asumir el puesto, hizo inventariar los barómetros, brújulas, sextantes y cronómetros de la colección del depósito. Asimismo levantó acta de los innumerables libros náuticos, mapas y cartas que almacenaba. Halló cajas mohosas llenas de antiguos cuadernos de bitácora de remotos viajes de capitanes de la Armada. Sus antecesores en el puesto los habían considerado basura. Con sus ocasionales versos jocosos o dibujos en los márgenes, a veces aquellos cuadernos parecían más una escapatoria del tedio de la travesía que un registro de la posición de los navíos.

Pero conforme Maury iba desempolvando los libros manchados de agua salada y revisando su

contenido, empezó a emocionarse mucho. Ahí estaba la información que necesitaba: anotaciones acerca del viento, el agua y el tiempo en lugares específicos y en fechas concretas. Aunque algunos de los cuadernos carecían de interés, muchos rebosaban de informaciones útiles. Si se recopilaban todas –comprendió Maury–, resultaría posible crear una forma enteramente nueva de carta de navegación. Maury y sus doce "computadores" –así se llamaba el puesto de quienes calculaban datos– iniciaron el proceso laborioso de extractar y tabular la información encerrada en aquellos cuadernos de bitácora medio podridos.

Maury agregó los datos y dividió todo el Atlántico en bloques de cinco grados de longitud y latitud. Luego, anotó la temperatura, la velocidad y la dirección del viento y del oleaje en cada segmento, así como el mes, puesto que esas condiciones variaban según la época del año. Una vez combinados, los datos revelaron patrones y apuntaron unas rutas más eficientes.

Los consejos seculares de los antiguos navegantes a veces mandaban a los barcos directamente a zonas de calma chicha o los arrojaban contra vientos y corrientes contrarios. En una ruta normal, de Nueva York a Río de Janeiro, hacía mucho que los marinos tendían a combatir los elementos en lugar de apoyarse en ellos. A los capitanes estadounidenses se les había enseñado a evitar los peligros de la ruta directa rumbo sur hacia Río. Así pues, sus barcos seguían un oscilante rumbo sudeste, antes de virar hacia el sudoeste, tras cruzar el ecuador. La distancia navegada a menudo representaba tres travesías completas del Atlántico. Esa ruta tortuosa no tenía ningún sentido. Funcionaba mejor un rumbo aproximadamente recto en dirección sur.

Para mejorar la exactitud, Maury necesitaba más información. Entonces creó un impreso estándar para registrar los datos de los barcos y consiguió que todos los buques de la Armada estadounidense lo usaran y lo entregaran al volver a puerto. Los barcos mercantes se mostraron ansiosos por hacerse con las cartas de Maury, quien, por su parte, insistía en que a cambio le entregasen también sus cuadernos de bitácora (a modo de precoz red social viral). "Cada barco que navega en alta mar –proclamó– puede ser considerado de ahora en adelante como un observatorio flotante, un templo de la ciencia". Para perfeccionar las cartas náuticas, buscó otros puntos de datos (igual que Google perfeccionó el algoritmo de PageRank para incluir más señales). Logró que los capitanes arrojasen al mar, cada cierta distancia, unas botellas con notas indicando el día, la posición, el viento y la corriente dominante, y que recogieran las botellas de este tipo que se topasen. Muchos barcos lucían una enseña especial para indicar que colaboraban en el intercambio de información (presagiando los iconos de compartir enlaces que hoy figuran en algunas páginas web).

A partir de los datos, se revelaron unos caminos marinos naturales, en los que los vientos y las corrientes eran particularmente favorables. Las cartas náuticas de Maury redujeron la duración de los viajes largos normalmente en una tercera parte, ahorrándoles buen dinero a los comerciantes. "Hasta que seguí su trabajo, había cruzado el océano a ciegas", le escribió un capitán agradecido. Hasta los viejos lobos de mar que rechazaban las cartas novedosas y seguían confiando en los métodos tradicionales o en su intuición cumplían un propósito útil: si sus travesías exigían más tiempo o acababan en desastre, demostraban la utilidad del sistema de Maury[57]. Para 1855, cuando publicó su obra magistral *The Physical Geography of the Sea*, Maury había trazado 1,2 millones de puntos de datos. "De esta manera, el joven marino, en vez de abrirse camino a tientas hasta que lo alumbraran las luces de la experiencia […] encontrará aquí, de una sola vez, que ya dispone de la experiencia de un millar de navegantes para guiarlo", escribió.

Su trabajo resultó esencial para tender el primer cable telegráfico transoceánico. Además, después de una trágica colisión en alta mar, puso a punto rápidamente el sistema de rutas de

navegación que hoy es de uso corriente. Hasta aplicó su método a la astronomía: cuando el planeta Neptuno fue descubierto en 1846, Maury tuvo la brillante idea de peinar los archivos en busca de referencias equivocadas al mismo como una estrella, lo que permitió calcular su órbita.

Maury ha sido ampliamente ignorado en los libros de historia estadounidense, tal vez porque presentó la dimisión de la Armada de la Unión durante la guerra de Secesión, y ejerció de espía para la confederación en Inglaterra. Pero unos años antes, cuando llegó a Europa buscando recabar apoyo internacional para sus cartas, cuatro países lo hicieron caballero y recibió medallas de oro de otros ocho, entre ellos la Santa Sede. Al alumbrar el siglo XXI, las cartas de navegación publicadas por la Armada estadounidense todavía llevaban su nombre.

El comandante Maury, el "explorador de los mares", fue de los primeros en darse cuenta de que existe un valor especial en un cuerpo de datos enorme, que falta en cantidades más pequeñas: un principio central del enfoque de datos masivos. De forma aún más esencial, comprendió que los mohosos cuadernos de bitácora de la Armada constituían en realidad "datos"[58] que podían extraerse y tabularse. Y con ello se convirtió en uno de los pioneros de la datificación, de desenterrar datos del material al que nadie concedía el menor valor. Al igual que Oren Etzioni de Farecast, que usó las informaciones antiguas sobre precios del transporte aéreo para crear un negocio lucrativo, o que los ingenieros de Google, que aplicaron búsquedas antiguas a la comprensión de las epidemias de gripe, Maury tomó una información generada para un propósito y la convirtió en algo distinto.

Su método, similar en líneas generales a las técnicas de datos masivos de hoy, resulta asombroso, teniendo en cuenta que lo elaboró con papel y lápiz. Su historia pone de relieve en qué grado el empleo de datos masivos precede a la digitalización. Hoy en día tendemos a fusionar los dos, pero es importante mantenerlos separados. Para adquirir una comprensión plena de cómo se extraen datos de los lugares más insospechados, veamos ahora un ejemplo más moderno.

Shigeomi Koshimizu, un profesor del Instituto Avanzado de Tecnología Industrial de Japón, sito en Tokio, se dedica al arte y la ciencia de analizar el trasero de los demás. Pocos pensarían que la forma de sentarse de una persona constituye información, pero puede serlo. Cuando alguien está sentado, el contorno del cuerpo, la postura y la distribución del peso pueden cuantificarse y tabularse. Koshimizu y su equipo de ingenieros convirtieron los traseros en datos, midiendo con sensores la presión en 360 puntos diferentes del asiento de un coche, e indexando cada punto en una escala de 0 a 256. El resultado es un código digital único para cada individuo. Durante una prueba, el sistema se mostró capaz de distinguir entre un grupo de personas con un 98 por 100 de acierto.

Esta investigación no es ninguna necedad. Se trataba de desarrollar la tecnología para un sistema antirrobo en vehículos. Un coche equipado con él podría detectar si estaba al volante alguien distinto del conductor autorizado, y exigir una contraseña para poder seguir conduciendo, o incluso detener el motor. La transformación de posiciones en datos crea un servicio viable y un negocio potencialmente lucrativo. Y su utilidad puede ir más allá de impedir el robo de vehículos. Por ejemplo, los datos agregados podrían brindar pistas acerca de la relación entre la postura de los conductores al volante y la seguridad vial si, por ejemplo, se dan una serie de cambios reveladores en la posición del conductor antes de un accidente. El sistema también podría ser capaz de detectar cuándo un conductor se vence ligeramente hacia un lado por el cansancio, y enviar una alerta o aplicar automáticamente los frenos. Y podría no solo prevenir el robo de un coche, sino identificar al ladrón por donde la espalda pierde su nombre.

El profesor Koshimizu tomó algo que nunca había sido tratado como datos –ni siquiera se había pensado que pudiera tener calidad informativa– y lo transformó en un formato cuantificado numéricamente. De modo similar, el comodoro Maury tomó un material que parecía de escasa utilidad y le extrajo información, convirtiéndolo en datos eminentemente útiles. Y con ello, consiguió que se utilizara la información de forma novedosa y que cobrara un valor único.

La palabra latina *data* significa "dado", en el sentido de "hecho". Este término se convirtió en el título de una obra clásica de Euclides, en la que explica la geometría a partir de lo que se sabe, o se puede mostrar que se sabe. Hoy en día, por datos se entiende una descripción de algo que permite ser registrado, analizado y reorganizado. Aún no existe un término adecuado para la clase de transformaciones producidas por el comodoro Maury y el profesor Koshimizu. Así pues, vamos a llamarlas *datificación*. "Datificar" un fenómeno es plasmarlo en un formato cuantificado para que pueda ser tabulado y analizado.

Ahora bien, esto es algo muy diferente de la digitalización, o proceso por el que se convierte la información analógica en los unos y ceros del código binario para que los ordenadores puedan manejarla. La digitalización no fue lo primero que hicimos con los ordenadores. La era inicial de la revolución informática fue computacional, como sugiere la etimología de la palabra. Usamos máquinas para efectuar cálculos que los métodos anteriores hacían más despacio, como las tablas de trayectorias de misiles, los censos y las predicciones meteorológicas. Lo de tomar contenido analógico y digitalizarlo vino después. De ahí que, en 1995, cuando Nicholas Negroponte, del laboratorio de medios del MIT, publicó su sobresaliente libro titulado *Ser digital*, uno de sus principales temas era el paso de los átomos a los bits. En la década de 1990, fundamentalmente nos dedicamos a digitalizar textos. Más recientemente, puesto que la capacidad de almacenaje, la potencia de procesamiento y el ancho de banda han aumentado, lo hemos hecho también con otros formatos de contenido, como las imágenes, los vídeos y la música.

Hoy en día, los tecnólogos comparten la creencia implícita de que el linaje de los datos masivos se remonta a la revolución del silicio. Esto, simplemente, no es así. Es cierto que los modernos sistemas de tecnología de la información (TI) ciertamente han hecho posibles los datos masivos, pero, en esencia, el paso a los datos masivos es una continuación de esa antigua misión humana que es medir, registrar y analizar el mundo. La revolución de la TI es evidente en todo lo que nos rodea, pero el énfasis se ha puesto fundamentalmente en la T, la tecnología. Es hora de volver la vista para fijarnos en la I, la información.

Para poder capturar información cuantificable, para datificar, necesitamos saber cómo medirla y cómo registrar lo que medimos. Esto requiere instrumentos adecuados. También precisa del deseo de cuantificar y registrar. Ambos son requisitos para la datificación, y los bloques de construcción necesarios para ese menester los habíamos desarrollado muchos siglos antes de que amaneciese la era digital.

CUANTIFICAR EL MUNDO[59]

La capacidad de archivar información es una de las líneas que separan las sociedades primitivas de las avanzadas. La contabilidad y las medidas de longitud y peso básicas se hallan entre las herramientas conceptuales más antiguas de las civilizaciones tempranas. Para el tercer milenio a. de C., la idea de la información registrada había progresado significativamente en el valle del Indo, en

Egipto y en Mesopotamia. Aumentó la exactitud, al igual que el empleo de las mediciones en la vida cotidiana. Aparte, la evolución de la escritura en Mesopotamia aportó un método preciso para llevar la cuenta de la producción y de las transacciones de negocios. El lenguaje escrito les permitió a las primeras civilizaciones medir la realidad, tomar nota de ella y recuperarla más tarde. A la par, los actos de medir y registrar facilitaron la creación de los datos. Fueron los cimientos primeros de la datificación.

Así se hizo posible copiar la actividad humana. Los edificios, por ejemplo, podían ser reproducidos a partir de los registros de sus dimensiones y de los materiales empleados. Asimismo, se daba pie a la experimentación: un arquitecto o un constructor podían alterar determinadas dimensiones manteniendo las demás intactas, creando así un nuevo diseño… que luego podía ser archivado a su vez. Las transacciones comerciales podían anotarse, de forma que constara cuánto grano se había obtenido en tal cosecha o campo (y cuánto se llevaría el estado en impuestos). La cuantificación permitió predecir y, por consiguiente, planificar, aun cuando fuese algo tan pedestre como suponer que la cosecha del año venidero sería igual de abundante que la de los anteriores. Ahora, las partes de una transacción podían saber cuánto se adeudaban entre sí. Sin medir y archivar no podría haber dinero, porque no habría habido datos en qué basarlo.

A lo largo de los siglos, la medición se extendió de la longitud y el peso al área, el volumen y el tiempo. A principios del primer milenio de nuestra era, las características principales de la medición ya estaban implantadas en Occidente. Ahora bien, la forma de medir de las primeras civilizaciones tenía una desventaja significativa: no estaba optimizada para los cálculos, ni siquiera para los relativamente sencillos. El sistema de cómputo de los numerales romanos no se ajustaba bien al análisis numérico. Sin un sistema de numeración "posicional" de base diez, ni decimales, la multiplicación y la división de números grandes resultaba una tarea complicada incluso para los expertos, y las simples sumas y restas se les resistían a la mayoría de los demás.

En la India, alrededor del siglo I, se desarrolló un sistema alternativo de números. De ahí llegó a Persia, donde fue mejorado, y luego pasó a los árabes, que lo refinaron considerablemente. Es la base de los números arábigos que usamos hoy. Puede que las Cruzadas asolasen las tierras invadidas por los europeos, pero durante ellas el conocimiento emigró del este al oeste, y quizá el trasplante más relevante fue el de los guarismos arábigos. El papa Silvestre II, que los había estudiado, abogó por su uso a finales del primer milenio. Llegado el siglo XII, varios textos árabes que describían el sistema se tradujeron al latín y se difundieron por toda Europa. Y así se produjo el despegue de las matemáticas.

Pero antes incluso de que los números arábigos llegasen a Europa, el cálculo se había visto mejorado mediante el empleo de tableros de recuento, es decir, unas tablas lisas sobre las que se colocaban fichas que representaban cantidades. Al mover las fichas a determinadas áreas del tablero, se sumaba o restaba. El método, sin embargo, presentaba serios inconvenientes. Resultaba difícil calcular números muy grandes y muy pequeños al mismo tiempo. Peor aún: los números sobre los tableros eran transitorios. Un movimiento en falso o un golpe por descuido podían cambiar un dígito y llevar a resultados incorrectos. Los tableros de recuento eran tolerables para hacer cálculos, pero no servían para registrar nada. La única forma de registrar y archivar los números que aparecían en el tablero pasaba por convertirlos de nuevo en ineficientes números romanos. (Los europeos nunca se vieron expuestos al ábaco[60] oriental; retrospectivamente, quizá fuese una ventaja, porque ese instrumento podría haber prolongado el uso de los números romanos en Occidente).

Las matemáticas dieron un nuevo sentido a los datos: ahora estos podían ser *analizados*, no solo

registrados y recuperados. La adopción generalizada de la numeración arábiga en Europa tardó cientos de años, desde su introducción en el siglo XII hasta finales del siglo XVI. Para entonces, los matemáticos se jactaban de poder calcular seis veces más deprisa con guarismos arábigos que con tableros de recuento. Lo que finalmente ayudó a consolidar los números arábigos fue la evolución de otra herramienta de datificación: la contabilidad por partida doble.

La escritura la inventaron los contables en el tercer milenio a. de C. Aunque la contabilidad evolucionó a lo largo de los siglos siguientes, en líneas generales siguió siendo un sistema para registrar una transacción económica en un soporte determinado. Lo que no hacía era mostrar de forma sencilla, en cualquier momento, lo que más les importaba a los contables y a sus jefes, los mercaderes: si una cuenta determinada o una operación comercial completa eran rentables o no. Eso empezó a cambiar en el siglo xiv, cuando los contables italianos registraron por primera vez las transacciones empleando dos asientos: uno para los créditos y otro para los débitos, de forma que las cuentas en conjunto quedaran equilibradas. La belleza de este sistema radicaba en que hacía fácil advertir las pérdidas y las ganancias. Y así, de repente, los datos mudos empezaron a hablar.

Hoy en día, solo le damos valor a la contabilidad por partida doble por sus consecuencias para la contabilidad y las finanzas, en general, pero su aparición representó un hito en el empleo de los datos. Permitió registrar la información bajo la forma de "categorías" que vinculaban las cuentas. Funcionaba mediante una serie de reglas sobre cómo anotar los datos: uno de los primeros ejemplos del registro normalizado de información. Un contable cualquiera podía mirar los libros de otro y comprenderlos. Estaba organizada de forma que un tipo específico de consulta de datos –el cálculo de pérdidas o ganancias para cada cuenta– resultase rápido y sencillo. Y ofrecía además un rastro de transacciones que se podía seguir, de forma que los datos resultaban más fáciles de localizar. Los maniáticos de la tecnología pueden apreciarlo bien hoy en día: el diseño incorporaba de fábrica la "corrección de errores". Si uno de los lados del libro mayor parecía estar mal, se podía comprobar la entrada correspondiente.

Con todo, igual que los números arábigos, la contabilidad de doble asiento no fue un éxito inmediato. Doscientos años después de que se inventara, aún hicieron falta un matemático y una familia de mercaderes para cambiar la historia de la datificación.

El matemático fue un monje franciscano llamado Luca Pacioli, que en 1494 publicó un manual, dirigido al lector profano, acerca de las matemáticas y su aplicación al comercio. La obra fue un gran éxito y se convirtió, de hecho, en el libro de texto de matemáticas de su época. Fue asimismo el primero que empleó de forma sistemática la numeración arábiga, y su popularidad facilitó la adopción de esta en Europa. Su mayor y más duradera contribución, sin embargo, fue la sección dedicada a la teneduría de libros, en la que Pacioli explicaba claramente el sistema de contabilidad por partida doble. A lo largo de las décadas siguientes, esta parte sobre contabilidad se publicó por separado en seis idiomas, y durante siglos siguió siendo el texto de referencia sobre la materia.

En cuanto a la familia de mercaderes, se trataba de los Médici, famosos comerciantes y mecenas. En el siglo XVI, los Médici se convirtieron en los banqueros más influyentes de Europa, en no poca medida porque empleaban un método superior de registro de datos, el sistema de doble asiento. Juntos, el libro de texto de Pacioli y el éxito de los Médici al aplicarlo, sellaron la victoria de la contabilidad por partida doble y, por extensión, establecieron el uso de los números arábigos en Occidente.

En paralelo a los avances en el registro de datos, las formas de medir el mundo –en tiempo, distancia, área, volumen y peso– siguieron ganando cada vez mayor precisión. El celo por entender

la naturaleza a través de la cuantificación caracterizó a la ciencia del siglo XIX, a medida que los investigadores iban inventando nuevas herramientas y unidades para medir y registrar corrientes eléctricas, presión del aire, temperatura, frecuencia del sonido, y demás. Fue una era en la que absolutamente todo tenía que ser definido, diferenciado y explicado. La fascinación llegó hasta el extremo de medir el cráneo de la gente como aproximación a su capacidad mental. Afortunadamente, la pseudociencia de la frenología ha desaparecido casi por completo, pero el deseo de cuantificar no ha hecho sino intensificarse.

La medición y el registro de la realidad prosperaron debido a la combinación de herramientas disponibles y la mentalidad receptiva. Esa mezcla es la tierra fértil en la que ha arraigado la datificación moderna. Los ingredientes para datificar estaban preparados, aunque en un mundo analógico la cosa resultaba aún costosa y consumía mucho tiempo. En muchos casos, exigía una paciencia, al parecer, infinita, o cuando menos la dedicación de toda una vida, como las fastidiosas observaciones nocturnas de estrellas y planetas que hacía Tycho Brahe a principios del siglo XVI. En los contados casos en que triunfó la datificación durante la era analógica, como el de las cartas náuticas del comodoro Maury, se debió a una afortunada serie de coincidencias: Maury, por ejemplo, se vio confinado a un trabajo de despacho, pero con acceso a todo un tesoro en forma de de cuadernos de bitácora. Sin embargo, en los casos en que la datificación sí tuvo éxito, aumentó enormemente el valor de la información subyacente, y se lograron unas percepciones extraordinarias.

El advenimiento de las computadoras trajo consigo equipos de medida y almacenaje que hicieron sumamente más eficiente el proceso de datificación. También facilitó en gran medida el análisis matemático de los datos, permitiendo descubrir su valor oculto. En resumen, la digitalización propulsa la datificación, pero no la sustituye. El acto de digitalizar –convertir información analógica a un formato legible por el ordenador– no datifica por sí mismo.

CUANDO LAS PALABRAS SE CONVIERTEN EN DATOS

La diferencia entre la digitalización y la datificación se torna evidente cuando se examina un terreno en el que se han producido ambas, y se comparan sus consecuencias. Pensemos en los libros. En 2004, Google anunció un proyecto de increíble osadía: iba a hacerse con todas las páginas de cuantos libros pudiera, y –en la medida posible en el marco de las leyes sobre propiedad intelectual– permitir a cualquier persona del mundo acceder a esos libros por internet, y realizar búsquedas en ellos gratis. Para lograr esta hazaña, la compañía se asoció con algunas de las mayores y más prestigiosas bibliotecas universitarias del mundo y puso a punto unas máquinas de escanear que pasaban automáticamente las páginas, de manera que el escaneado de millones de libros fuese al tiempo factible y económicamente viable.

Primero, Google *digitalizó* el texto: todas y cada una de las páginas fueron escaneadas y guardadas en archivos de imagen digital de alta resolución que se almacenaron en los servidores de la empresa. Cada página había sido transformada en una copia digital, que podría ser fácilmente recuperada a través de la red por cualquier persona. Sin embargo, para recuperar esa información hacía falta, o bien saber qué libro la contenía, o bien leer mucho hasta dar con el pasaje correcto. Uno no podía buscar unas palabras determinadas en el texto, ni analizarlo, porque el texto no había sido datificado. Google disponía solo de unas imágenes que los seres humanos podían convertir en información útil únicamente leyéndolas.

Aunque esto habría supuesto una gran herramienta de todas maneras –una biblioteca de Alejandría digital, más exhaustiva que ninguna otra antes–, Google quería más. La compañía comprendía que la información encerraba un valor que solo se haría evidente una vez datificado. Así que Google empleó un programa de reconocimiento óptico de caracteres que podía tomar una imagen digital e identificar las letras, palabras y párrafos. El resultado fue un texto datificado en lugar de una imagen digitalizada de una página.

Ahora, la información de la página era utilizable no solo por lectores humanos, sino también por los ordenadores, que podían procesarla; y por los algoritmos, que podían analizarla. La datificación hizo que pudiera indexarse el texto y que, por consiguiente, pudieran hacerse búsquedas en él. Y permitió un flujo inacabable de análisis textual: ahora podemos descubrir cuándo se utilizaron por primera vez determinadas palabras o frases, o cuándo se volvieron populares, conocimientos que arrojan nueva luz sobre la diseminación de las ideas y la evolución del pensamiento humano a través de los siglos, y en muchos idiomas.

El lector puede hacer la prueba por sí mismo. El Ngram Viewer de Google (http://books.google.com/ngrams) generará una gráfica del uso de palabras o frases a lo largo del tiempo, empleando el índice íntegro de Google Books como fuente de datos. En cuestión de segundos, descubrimos que hasta 1900 el término "causalidad" se usaba con mayor frecuencia que el de "correlación", pero luego se invirtió la ratio. Podemos comparar estilos de escritura y dilucidar ciertas disputas de autoría. La datificación hace asimismo mucho más fácil detectar el plagio en obras científicas; a raíz de ello, una serie de políticos europeos, entre ellos un ministro de defensa alemán, se han visto forzados a dimitir.

Se estima que se han publicado unos 130 millones de libros individuales desde la invención de la imprenta a mediados del siglo xv. En 2012, siete años después de iniciar Google su proyecto bibliográfico, había escaneado más de veinte millones de títulos, más del 15 por 100 del legado escrito de la humanidad, es decir, una porción considerable. Esto ha originado una nueva disciplina académica llamada "culturonomía": la lexicología informática que intenta comprender el comportamiento humano y las tendencias culturales mediante el análisis cuantitativo de textos.

En un estudio, varios investigadores de Harvard[61] revisaron millones de libros (que representaban más de 500.000 millones de palabras) y descubrieron que menos de la mitad de las palabras inglesas que aparecen en los libros están recogidas en los diccionarios. Al contrario, escribieron, la mayor abundancia de palabras "consiste en 'materia oscura' léxica sin documentar en las obras de referencia estándar". Es más, al analizar algorítmicamente las referencias al pintor Marc Chagall, cuyas obras fueron prohibidas en la Alemania nazi por su origen judío, los investigadores demostraron que la supresión o censura de una idea o persona deja "huellas dactilares cuantificables". Las palabras son como fósiles incrustados en las páginas en vez de en la roca sedimentaria; quienes se dedican a la culturonomía pueden excavarlas como si fuesen arqueólogos. Por descontado, ese conjunto de datos trae consigo una cantidad astronómica de prejuicios implícitos: ¿constituyen acaso los libros de las bibliotecas un reflejo fiel del mundo real, o simplemente del mundo que les gusta a los autores y a los bibliotecarios? Aún y con todo, la culturonomía nos ha facilitado una lente enteramente nueva con la que intentar entendernos a nosotros mismos.

La transformación de las palabras en datos da rienda suelta a numerosos usos. Ciertamente, los datos pueden ser usados por los seres humanos para la lectura y por las máquinas para el análisis. Pero, como modelo perfecto de compañía de datos masivos, Google sabe que la información tiene

múltiples objetivos potenciales que pueden justificar su recopilación y datificación. Así que, astutamente, utilizó el texto datificado de su proyecto de escaneo de libros para mejorar su servicio de traducción automática. Como se ha explicado en el capítulo III, el sistema identificaba qué libros eran traducciones, y analizaba qué palabras y frases usaban los traductores como alternativas entre un idioma y otro. Sabiéndolo, podía tratar luego la traducción como si fuese un gigantesco problema de matemáticas, con el ordenador calculando probabilidades para determinar qué palabra es la mejor sustituta de otra entre dos idiomas.

Por supuesto, Google no ha sido la única organización que ha soñado con llevar la riqueza del legado escrito del mundo a la era de los ordenadores, ni mucho menos la primera en intentarlo. Ya en 1971, el proyecto Gutenberg, una iniciativa de voluntarios para poner online obras de dominio público, aspiraba a hacer disponibles los textos para su lectura, pero no valoraba los usos secundarios de tratar las palabras como datos. Se trataba de leer, no de reutilizar. Igualmente, los editores llevan años experimentando con versiones electrónicas de sus libros: también ellos consideraban que el valor fundamental de los libros era el contenido, no los datos, porque ese es su modelo de negocio. Así pues, nunca usaron, ni permitieron a otros usar, los datos inherentes al texto del libro. Nunca vieron esa necesidad, ni apreciaron el potencial que tenía.

Muchas compañías rivalizan ahora para conquistar el mercado del libro electrónico. Amazon, con sus lectores electrónicos Kindle, parece haber tomado una buena delantera. En este área, sin embargo, las estrategias de Amazon y Google difieren considerablemente.

Amazon también ha datificado libros, pero, a diferencia de Google, no ha sabido explotar los posibles usos nuevos del texto en tanto que datos. Jeff Bezos, fundador y director general de la empresa, convenció a centenares de editoriales para que publicasen sus títulos en formato Kindle. Los libros Kindle no están hechos de imágenes de páginas. Si lo estuvieran, no sería posible cambiar el tamaño de la fuente, ni mostrar la página tanto en una pantalla a color como en una en blanco y negro. El texto no solo es digital: está datificado. En realidad, Amazon ha hecho con millones de libros nuevos lo que Google está intentando conseguir esforzadamente con muchos antiguos.

Sin embargo, aparte del brillante servicio de "palabras estadísticamente significativas", Amazon –que emplea algoritmos para hallar vínculos entre los temas de los libros que de otro modo podrían no resultar aparentes–, no ha utilizado su riqueza en palabras para el análisis de datos masivos. Considera que su negocio de libros se basa en el contenido que leen los seres humanos antes que en el análisis del texto datificado. Y, para ser del todo justos, probablemente tenga que hacer frente a restricciones de las editoriales conservadoras sobre el uso que puede dar a la información contenida en sus libros. Google, como chico malo de los datos masivos, dispuesto a superar todos los límites, no se siente constreñido de esa manera: su pan lo gana con los clics de los usuarios, no por su acceso a los catálogos de las editoriales. Quizá no resulte injusto decir que, por lo menos por ahora, Amazon entiende el valor de digitalizar el contenido, mientras que Google comprende el de datificarlo.

CUANDO LA LOCALIZACIÓN SE CONVIERTE EN DATOS

Uno de los elementos de información más básicos que hay en el mundo es, por así decir, el mundo mismo. Sin embargo, durante la mayor parte de la historia, el área espacial nunca se cuantificó ni se usó en forma de datos. Por supuesto, la geolocalización de la naturaleza, los objetos y las personas constituye información. La montaña está allí; la persona, aquí. Pero, para resultar lo más útil posible,

esa información necesita ser transformada en datos. Datificar una localización exige ciertos requisitos. Necesitamos un método para medir cada centímetro cuadrado del área terrestre. Necesitamos un procedimiento estandarizado para anotar las mediciones. Necesitamos un instrumento para monitorizar y registrar los datos. Cuantificación, estandarización, recopilación. Solo entonces podremos archivar y analizar la localización no como un sitio *per se*, sino en forma de datos.

En Occidente, la cuantificación de la localización empezó con los griegos. Hacia el 200 a. de C., Eratóstenes inventó un sistema de retícula similar al de latitud y longitud para demarcar una localización. Sin embargo, como tantas otras buenas ideas de la Antigüedad, la práctica se perdió con el tiempo. Mil quinientos años después, alrededor del año 1400, un ejemplar de la *Geographia* de Ptolomeo llegó a Florencia desde Constantinopla, justo cuando el Renacimiento y el comercio marítimo estaban avivando el interés en la ciencia y la sabiduría de los antiguos. El tratado causó sensación, y sus viejas lecciones se aplicaron a la resolución de los modernos desafíos de la navegación. Desde entonces, los mapas siempre han incluido longitud, latitud y escala. El sistema fue mejorado posteriormente, en 1570, por el cartógrafo flamenco Gerardus Mercator, permitiéndoles a los marinos trazar un rumbo recto en un mundo esférico.

Aunque para entonces existiese una forma de registrar la localización, no había ningún formato aceptado universalmente para compartir esa información. Se precisaba un sistema de identificación común, de la misma manera que internet se beneficiaría de los nombres de dominios para hacer que cosas como el correo electrónico funcionaran en todo el mundo. La normalización de longitud y latitud tomó mucho tiempo. Quedó finalmente consagrada en 1884 en la Conferencia Internacional del Meridiano en Washington, donde veinticinco naciones eligieron Greenwich, en Inglaterra, como meridiano principal y punto cero de la longitud (los franceses, que se consideraban los líderes de los estándares internacionales, se abstuvieron). En la década de 1940 se creó el sistema de coordenadas Universal Transversal de Mercator (UTM), que dividió el mundo en sesenta zonas para mayor exactitud.

La localización geoespacial ya podía ser identificada, registrada, contramarcada, analizada y comunicada en un formato numérico estándar. La posición podía ser datificada, pero debido al elevado coste de medir y registrar la información en un entorno analógico, raras veces se hacía. La datificación tendría que esperar a que se inventasen herramientas para medir localizaciones de forma asequible. Hasta la década de 1970, la única forma de determinar la localización física pasaba por el empleo de puntos de referencia, las constelaciones astronómicas, a estima, o la tecnología limitada de radioposición.

En 1978 tuvo lugar un gran cambio, al lanzarse el primero de los veinticuatro satélites que forman el Sistema de Posicionamiento Global (GPS). Los receptores en tierra pueden triangular su posición anotando las diferencias en el tiempo que tardan en recibir una señal de los satélites que giran a 20.278 km por encima de sus cabezas. Desarrollado por el departamento de defensa de Estados Unidos, el sistema se abrió por primera vez a usos no militares en la década de 1980, llegó a estar plenamente operativo en los 90 y su precisión se aquilató para aplicaciones comerciales una década más tarde. Con un margen de error de solo un metro, el GPS señaló el momento en que un método de medir la localización, el sueño de navegantes, cartógrafos y matemáticos desde la Antigüedad, se unía por fin a un medio técnico de lograrlo rápidamente, de forma (relativamente) barata, y sin necesidad de conocimientos especializados.

Pero la información ha de generarse de hecho. Nada les impedía a Eratóstenes y a Mercator estimar su paradero a cada minuto del día, si lo hubiesen deseado. Pero, aunque era factible,

resultaba impracticable. Del mismo modo, los primeros receptores de GPS eran complejos y costosos, adecuados para un submarino, pero no para todo el mundo y en todo momento. Todo esto cambiaría gracias a la ubicuidad de los chips baratos incorporados a los artilugios electrónicos. El coste de un módulo GPS cayó de cientos de dólares en los 90 a cosa de un dólar hoy en día para grandes cantidades. Normalmente, bastan unos segundos para que el GPS fije una localización, y las coordenadas estén normalizadas. Así que 37°14'06" N, 115°48'40" W solo puede significar que uno se halla en la supersecreta base militar estadounidense sita en un remoto lugar de Nevada conocido como "Área 51", donde (quizá) se retiene a seres alienígenas.

Hoy en día, el GPS no es más que un sistema entre muchos de determinar la localización. Hay sistemas de satélites rivales en curso de instalación en China y Europa. Se puede alcanzar incluso mayor exactitud triangulando entre torres de telefonía móvil o *routers* wifi para determinar la posición basándose en la intensidad de la señal, ya que los GPS no funcionan en el interior ni entre edificios altos. Eso ayuda a explicar por qué algunas empresas como Google, Apple y Microsoft han establecido sus propios sistemas de geolocalización para complementar el GPS. Los vehículos del Street View de Google recogían información de *routers* wifi mientras sacaban fotos, y el iPhone era un teléfono "espía" que recopilaba datos sobre localización y wifi y la remitía a Apple sin que los usuarios fuesen conscientes de ello. (Los teléfonos Android de Google y el sistema operativo de los móviles de Microsoft también recogían esta clase de datos).

Ahora no solo se puede rastrear a la gente, sino a los objetos. Con la instalación de módulos inalámbricos[62] en los vehículos, la datificación de la localización transformará el concepto de los seguros. Los datos ofrecen una vista pormenorizada de los tiempos, localizaciones y distancias de conducción real que permiten un precio mejor en función del riesgo. En Estados Unidos y Gran Bretaña, los conductores pueden ajustar el precio del seguro del coche dependiendo de a dónde y cuándo conducen efectivamente, en lugar de pagar una póliza anual basada en su edad, sexo e historial. Esta aproximación al precio de los seguros crea incentivos al buen comportamiento. Transforma la naturaleza misma del seguro: antes se basaba en el riesgo agrupado y ahora se basa en la actuación individual. Seguir la pista de los individuos por medio de sus vehículos también altera la naturaleza de los costes fijos, como las carreteras y demás infraestructuras, al vincular el uso de esos recursos a los conductores y otras personas que los "consumen". Esto era imposible de hacer antes de poderse expresar la geolocalización en forma de datos, de manera continua, para todos y para todo; así es el mundo hacia el que nos encaminamos.

La empresa de mensajería UPS, por ejemplo, utiliza datos de geolocalización de múltiples maneras. Sus vehículos están equipados con sensores, módulos inalámbricos y GPS, de modo que el cuartel general puede predecir las averías del motor, como vimos en el capítulo anterior. Es más, ello permite a la compañía conocer el paradero de sus furgones en caso de retrasos, monitorizar a los empleados y analizar sus itinerarios para optimizar los trayectos. La ruta más eficiente se determina en parte merced a los datos de entregas anteriores, igual que las cartas de Maury se basaron en travesías marítimas previas.

Este programa analítico ha surtido extraordinarios efectos. En 2011, UPS eliminó la impresionante cantidad de 48 millones de kilómetros de las rutas de sus conductores, ahorrando así más de 11,3 millones de litros de combustible y 30.000 toneladas de emisiones de dióxido de carbono, según Jack Levis, director de gestión de procesos de la compañía. Asimismo mejoró la seguridad y la eficiencia: el algoritmo recopila rutas que no obliguen a pasar por cruces con tráfico, porque tienden a causar accidentes, pérdidas de tiempo y más consumo de combustible, ya que los furgones a

menudo tienen que esperar con el motor al ralentí antes de poder girar.

"La predicción nos trajo el conocimiento –afirma Levis, de UPS[63]–, pero detrás del conocimiento hay algo más: sabiduría y clarividencia. En algún momento futuro, el sistema será tan listo que será capaz de predecir los problemas y corregirlos antes de que el usuario caiga siquiera en la cuenta de que algo va mal".

La localización datificada a lo largo del tiempo se está aplicando de forma particularmente notable a las personas. Durante años, los operadores de telefonía móvil han recogido y analizado información para mejorar el nivel de servicio de sus redes. Estos datos están siendo usados cada vez más para otros fines y recopilados por terceros para nuevos servicios. Algunas aplicaciones para teléfono inteligente, por ejemplo, recogen información acerca de localizaciones con independencia de que la propia app disponga de alguna función basada en la localización. En otros casos, la misma razón de ser de una app es crear negocio a partir del conocimiento de las localizaciones de los usuarios. Un ejemplo lo brinda Foursquare, que le permite a la gente "registrarse" en sus lugares favoritos. Esta app deriva sus ingresos de loa programas de fidelidad, las recomendaciones de restaurantes y otros servicios relacionados con la localización.

La capacidad de recopilar los datos de geolocalización de los usuarios está convirtiéndose en algo extremadamente valioso. A escala individual permite la publicidad personalizada allí donde esté o se prevea vaya a estar una persona. Es más, la información puede ser agregada para revelar tendencias. Por ejemplo, acumular datos de localización permite a las empresas detectar atascos de tráfico sin necesidad de ver los coches: la información la proporcionan el número y velocidad de los teléfonos que se desplazan por una carretera. La compañía AirSage procesa a diario 15.000 millones de registros de geolocalización de los desplazamientos de millones de usuarios de telefonía móvil, para crear informes en tiempo real acerca del tráfico en más de cien ciudades de todo Estados Unidos. Otras dos empresas de geolocalización, Sense Networks y Skyhook, usan datos de localización para determinar cuáles son las zonas de la ciudad con la vida nocturna más animada, o para estimar cuántos asistentes ha habido en una manifestación.

Sin embargo, puede que los usos no comerciales de la geolocalización acaben siendo los más importantes. Sandy Pentland[64], director del Laboratorio de Dinámica Humana del MIT, y Nathan Eagle fueron ambos pioneros de lo que llaman *reality mining*, "minería de la realidad". Se refiere a procesar enormes cantidades de datos procedentes de teléfonos móviles para extraer inferencias y predicciones sobre el comportamiento humano. En uno de sus estudios, el análisis de los movimientos y los patrones de llamadas les permitió identificar a personas que habían contraído la gripe antes de que ellas mismas supiesen que estaban enfermas. En caso de una epidemia mortal de gripe, esta capacidad podría salvar millones de vidas, al permitirles a los funcionarios de la sanidad pública saber cuáles son las áreas más afectadas en todo momento. Ahora bien, en manos irresponsables, el poder del *reality mining* podría tener consecuencias terribles, como veremos más adelante.

Eagle, fundador de la *start up* de datos inalámbricos Jana, ha usado datos agregados de telefonía móvil de más de doscientos operadores en más de cien países –unos 3.500 millones de personas en América Latina, África y Europa– para dar respuesta a preguntas cruciales para los ejecutivos de márketing, como cuántas veces a la semana hace la colada una familia. También ha usado los datos masivos para contemplar cómo prosperan las ciudades, por ejemplo. Un colega y él combinaron datos de localización de suscriptores de móviles de prepago en África con la cantidad de dinero que gastaban cuando recargaban sus cuentas. El valor muestra una correlación fuerte con el nivel de

ingresos: las personas ricas compran más minutos de una sola vez. Ahora bien, uno de los hallazgos inesperados de Eagle es que los suburbios, en vez de ser solo centros de pobreza, también actúan como trampolines económicos. La cosa está en que estos usos indirectos de los datos de localización nada tienen que ver con las rutas de las comunicaciones móviles, que era el propósito para el que se generó inicialmente la información. Al contrario, una vez se ha datificado la localización, surgen usos nuevos y se puede crear un valor nuevo.

CUANDO LAS INTERACCIONES SE CONVIERTEN EN DATOS

Las próximas fronteras de la datificación son más personales: nuestras relaciones, experiencias y estados de ánimo. La idea de la datificación constituye el espinazo de muchas de las compañías de medios sociales de la red. Las plataformas de redes sociales no nos ofrecen meramente una forma de localizar y mantener el contacto con amigos y colegas: también toman elementos intangibles de nuestra vida diaria y los transforman en datos que pueden usarse para hacer cosas nuevas. Facebook datificó las relaciones; siempre existieron y constituyeron información, pero nunca fueron definidas formalmente como datos hasta la "gráfica social" de Facebook. Twitter permitió la datificación de los sentimientos al crear una forma fácil de que la gente anotase y compartiese sus pensamientos inconexos, que previamente se perdían en las brumas del tiempo. LinkedIn datificó nuestras experiencias profesionales pretéritas, igual que Maury transformó los antiguos cuadernos de bitácora, convirtiendo esa información en predicciones acerca de nuestro presente y futuro: a quién conocemos, o qué trabajo puede interesarnos.

Estos usos de los datos se hallan aún en estado embrionario. En el caso de Facebook, la firma ha sabido mostrarse paciente y astuta, consciente de que revelar demasiado pronto demasiadas finalidades nuevas para los datos de sus usuarios podría espantarlos. Además, la empresa todavía está ajustando su modelo de negocio (y su política de privacidad) a la cantidad y clase de recogida de datos que desea desarrollar. De ahí que buena parte de las críticas que ha recibido se centren más en qué información es capaz de recopilar que en lo que ha hecho en realidad con esos datos. En 2012, Facebook[65] tenía alrededor de mil millones de usuarios, interconectados mediante cien mil millones de amistades. La gráfica social resultante representa más del 10 por 100 de la población total del mundo, datificada y a disposición de una sola compañía.

Los usos potenciales son extraordinarios. Una serie de empresas de nueva creación han estudiado adaptar la gráfica social para utilizarla como señales que permitan establecer valoraciones crediticias. La idea es que "Dios los cría y ellos se juntan": las personas prudentes hacen amistad con gente de mentalidad parecida, mientras que los derrochadores incurren juntos en impago. Si sale bien, Facebook podría convertirse en el próximo FICO, el organismo de calificación crediticia. Los ricos conjuntos de datos de las firmas de medios sociales bien podrían constituir la base de unos negocios nuevos que vayan más allá de compartir superficialmente fotos, actualizaciones de estado y "me gustas".

También Twitter ha visto cómo se usaban sus datos de manera interesante. Para algunas personas, los cuatrocientos millones de sucintos tuits que enviaron cada día de 2012 más de ciento cuarenta millones de usuarios al mes parecen poco más que parloteo irreflexivo. Y de hecho, a menudo eso es exactamente lo que son. Sin embargo, la compañía permite la datificación de pensamientos, estados de ánimo e interacciones de la gente, que anteriormente nunca habían podido ser aprehendidos.

Twitter ha llegado a acuerdos con dos empresas, DataSift y Gnip, para comercializar el acceso a los datos. (Si bien todos los tuits son públicos, el acceso al *firehose*, el "grifo de datos" de Twitter tiene un coste). Muchas empresas analizan los tuits, recurriendo a veces a una técnica llamada análisis de sentimientos, para almacenar comentarios de clientes o valorar el impacto de las campañas de márketing.

Dos fondos de inversión, Derwent Capital de Londres y MarketPsych de California, empezaron a analizar el texto datificado de los tuits como indicios para la inversión en el mercado de valores. (Sus estrategias comerciales reales fueron mantenidas en secreto; en lugar de invertir en firmas a las que se daba mucho bombo, puede que apostaran en su contra). Ambos fondos venden ahora la información a los inversores. En el caso de MarketPsych, se asoció con Thomson Reuters para ofrecer no menos de 18.864 índices para ciento diecinueve países, actualizados cada minuto, sobre estados emocionales como el optimismo, la melancolía, la alegría, el miedo, la cólera, y hasta temas como la innovación, el litigio y el conflicto. Los datos no los utilizan tanto las personas cuanto los ordenadores: los cerebritos matemáticos de Wall Street, conocidos como *quants*, insertan los datos en sus modelos algorítmicos para buscar correlaciones inadvertidas que puedan traducirse en beneficios. La propia frecuencia de los tuits sobre un tema determinado puede servir para predecir varias cosas, como los ingresos en taquilla de un filme de Hollywood[66], según Bernardo Huberman, uno de los padres del análisis de redes sociales. Huberman y un colega de HP desarrollaron un modelo que escrutaba la tasa de aparición de nuevos tuits. Una vez listo, fueron capaces de pronosticar el éxito de una película mejor que otros modelos de predicción ya habituales.

Pero las posibilidades no acaban ahí. Los mensajes de Twitter están limitados a unos escasos 140 caracteres, pero los metadatos –es decir, la "información acerca de la información"– que lleva asociados cada tuit son muy ricos. Incluyen 33 elementos específicos. Algunos no parecen demasiado útiles, como el fondo de escritorio de la página del usuario de Twitter, o el programa que emplea para acceder al servicio, pero otros resultan extremadamente interesantes, como el idioma de los usuarios, su geolocalización, y la cantidad y los nombres de las personas a los que siguen o que los siguen a ellos. En un estudio acerca de esos metadatos, mencionado en *Science* en 2011, el análisis de 509 millones de tuits enviados a lo largo de más de dos años por 2,4 millones de personas de ochenta y cuatro países mostró que los estados de ánimo de la gente siguen patrones diarios y semanales similares en todas las culturas del mundo; algo que había sido imposible advertir anteriormente. Los estados de ánimo han quedado datificados.

La datificación no solo tiene que ver con expresar las actitudes y estados de ánimo en una forma analizable, sino también con el comportamiento humano. Este resulta difícil de seguir de otro modo, especialmente en el contexto de la comunidad más amplia y de los subgrupos que contiene. Mediante el análisis de tuits, el biólogo Marcel Salathé, de la universidad estatal de Pensilvania, y el ingeniero de programas Shashank Khandelwal descubrieron que la actitud de las personas ante las vacunaciones era igual a la probabilidad de que se pusieran la vacuna de la gripe. Aún más importante: su estudio utilizaba los metadatos de quién estaba conectado con quién en Twitter[67] para ir todavía un paso más allá. Se dieron cuenta de que podrían existir subgrupos de personas sin vacunar. Lo que hace destacable esta investigación es que, mientras que otros estudios, como Google Flu Trends, empleaban datos agregados para considerar el estado de salud de los individuos, el análisis de sentimientos de Salathé predijo en la práctica los *comportamientos* relacionados con la salud.

Estos primeros hallazgos indican hacia dónde se dirigirá seguramente la datificación. Al igual que

Google, una bandada de redes de medios sociales como Facebook, Twitter, LinkedIn, Foursquare y otras están apostadas encima de un enorme cofre del tesoro lleno de información datificada que, una vez sometida a análisis, arrojará luz sobre la dinámica social a todos los niveles, desde el individuo hasta la sociedad en su conjunto.

LA DATIFICACIÓN DE TODO

Con un poco de imaginación, una plétora de cosas pueden expresarse en forma de datos, sorprendiéndonos de paso. Siguiendo el espíritu del trabajo del profesor Koshimizu sobre los traseros en Tokio, IBM obtuvo en 2012 una patente en Estados Unidos sobre "Seguridad de las oficinas mediante tecnología de computación basada en la superficie". Eso es jerga de abogado de la propiedad intelectual para referirse a un revestimiento del suelo sensible al tacto, algo en cierto modo parecido a una pantalla gigante de teléfono inteligente. Sus usos potenciales son muy numerosos. Sería capaz de identificar los objetos que tuviera encima. En su forma básica, sabría cuándo encender las luces de una habitación, o cuándo abrir las puertas al entrar una persona. Aún más importante, sin embargo, es que podría identificar a los individuos por su peso, su postura o su forma de caminar. Podría saber si alguien se ha caído y no se ha vuelto a levantar, detalle importante en el caso de las personas de edad. Los minoristas podrían conocer el flujo de clientes por sus tiendas. Cuando se datifica el suelo, no hay techo para los usos posibles.

Datificar todo lo posible no es tan disparatado como parece. Piénsese en el movimiento del "ser cuantificado"[68]. Se refiere a un grupo variopinto de fanáticos de la forma física, maniáticos de la medicina y yonquis tecnológicos que miden cada elemento de sus cuerpos y vidas para vivir mejor o, cuando menos, para aprender cosas nuevas que no podrían haber sabido antes de forma enumerativa. El número de "automedidores" es pequeño por el momento, pero va creciendo.

Gracias a los teléfonos inteligentes y a la tecnología de computación barata, nunca ha sido tan fácil la datificación de los actos más esenciales de la vida. Un montón de *start ups* permiten que la gente registre sus patrones de sueño mediante la medición de sus ondas mentales durante la noche. Una firma, Zeo, ya ha creado la mayor base de datos del mundo sobre la actividad del sueño y ha descubierto diferencias en las cantidades de sueño REM que experimentan los hombres y las mujeres. La firma Asthmapolis ha incorporado un sensor con GPS a un inhalador para el asma; al agregar las informaciones, la compañía consigue discernir los factores ambientales que disparan los ataques de asma, como la proximidad a determinados cultivos, por ejemplo.

Las firmas Fitbit y Jawbone facilitan que la gente mida su actividad física y su sueño. Otra empresa, Basis, permite que los usuarios de su brazalete monitoricen sus constantes vitales, entre ellas el ritmo cardíaco y la conductividad de la piel, que son medidas del estrés. Obtener los datos se está volviendo más fácil y menos invasivo que nunca. En 2009 le fue concedida a Apple una patente para recopilar datos sobre oxigenación de la sangre, ritmo cardíaco y temperatura corporal a través de sus auriculares de audio.

La datificación tiene mucho que enseñarnos acerca de cómo funciona nuestro cuerpo. Unos investigadores del colegio universitario de Gjøvik, en Noruega, y la empresa Derawi Biometrics[69] han desarrollado una app para teléfonos inteligentes que analiza el paso de un individuo al andar y usa esa información como sistema de seguridad para desbloquear el teléfono. Mientras tanto, dos profesores del Instituto de Investigación Tecnológica de Georgia, Robert Delano y Brian Parise,

están poniendo a punto otra app llamada iTrem que utiliza el acelerómetro del teléfono para monitorizar los tremores corporales de una persona en busca de Parkinson y otras enfermedades neurológicas. La app es una bendición tanto para los médicos como para los pacientes. A estos les permite ahorrarse costosos tests en la consulta del médico; a los profesionales de la medicina les permite monitorizar a distancia la incapacidad de las personas y su respuesta a los tratamientos. Según unos investigadores de Kyoto, un teléfono inteligente es solamente un poquito menos eficaz a la hora de medir los tremores que el acelerómetro triaxial que usa un equipo médico especializado, por lo que puede utilizarse con confianza. Una vez más, un poco de desorden es preferible a la exactitud.

En la mayoría de estos casos, estamos capturando información y dándole forma de datos que permiten reutilizarla. Esto puede ocurrir casi en cualquier lugar y prácticamente con cualquier cosa. Green-Goose, una firma *start up* de San Francisco, vende diminutos sensores de movimiento que pueden colocarse en objetos para controlar cuánto se usan. Poner el sensor en un paquete de hilo dental, una regadera o un paquete de arena para gatos hace posible datificar la higiene dental y el cuidado de las plantas y las mascotas. El entusiasmo por el "internet de las cosas" –implantar chips, sensores y módulos de comunicación en objetos cotidianos– tiene que ver en parte con el *networking* o las redes de contactos, pero casi tanto o más con datificar todo cuanto nos rodea.

Una vez que se ha datificado el mundo, los usos potenciales de la información no tienen más límite que el ingenio personal. Maury datificó las travesías de navegantes anteriores mediante una concienzuda tabulación manual, y con ello sacó a la luz unas perspectivas extraordinarias y valiosas. Hoy en día disponemos de las herramientas (estadísticas y algoritmos) y del equipo necesario (procesadores y almacenamiento digitales) para llevar a cabo unas tareas similares mucho más deprisa, a escala, y en muchos contextos diferentes. En la era de los datos masivos, hasta los traseros tienen su utilidad.

Nos hallamos inmersos en un gran proyecto de infraestructura que, de alguna manera, rivaliza con los del pasado: de los acueductos romanos a la *Encyclopédie* de la Ilustración. No llegamos a advertirlo con claridad porque el proyecto de hoy es demasiado nuevo, porque estamos en mitad de él, y porque, a diferencia del agua que fluye por los acueductos, el producto de nuestras labores es intangible. El proyecto es la datificación. Como aquellos otros avances infraestructurales, traerá consigo cambios fundamentales en la sociedad.

Los acueductos hicieron posible el crecimiento de las ciudades; la imprenta facilitó la Ilustración; los periódicos permitieron el auge del estado-nación[70]. Pero estas infraestructuras estaban enfocadas hacia flujos: de agua, de conocimiento. Otro tanto cabe decir del teléfono y de internet. En cambio, la datificación representa un enriquecimiento esencial de la comprensión humana. Con la ayuda de los datos masivos, nuestro mundo dejará de parecernos una sucesión de acontecimientos que explicamos como fenómenos naturales o sociales: veremos un universo constituido esencialmente por información.

Durante más de un siglo, los físicos han venido sugiriendo algo parecido: que el fundamento de cuanto existe no son los átomos, sino la información. Esto, reconozcámoslo, puede sonar esotérico. Mediante la datificación, sin embargo, en muchos casos podemos ahora capturar y calcular a una escala mucho más amplia los aspectos físicos e intangibles de la existencia, y actuar sobre ellos.

Ver el mundo como información, como océanos de datos que pueden explorarse cada vez más lejos y más hondo, nos ofrece un nuevo panorama de la realidad. Es una perspectiva mental que puede penetrar todas las áreas de la vida. Hoy formamos una sociedad aritmética porque presumimos

que el mundo se puede comprender mediante los números y las matemáticas. Y damos por supuesto que el conocimiento se puede transmitir a través del tiempo y del espacio porque el concepto de la escritura está muy arraigado. Puede que el día de mañana las generaciones siguientes tengan una "conciencia de datos masivos": la presunción de que hay un componente cuantitativo en todo cuanto hacemos, y de que los datos son indispensables para que la sociedad aprenda. La noción de transformar las innumerables dimensiones de la realidad en datos probablemente le parezca novedosa por ahora a la mayoría de la gente. Pero en el futuro, seguramente la trataremos como algo dado (lo cual, de forma agradable, nos retrotrae al origen mismo del término "dato").

Con el tiempo, puede que el impacto de la datificación deje pequeño el de los acueductos y los periódicos, rivalizando acaso con el de la imprenta e internet al facilitarnos las herramientas para cartografiar el mundo mediante datos. Por el momento, sin embargo, los usuarios más avanzados de la datificación se hallan en el mundo de los negocios, donde los datos masivos se están usando para crear nuevas formas de valor: es el tema del capítulo siguiente.