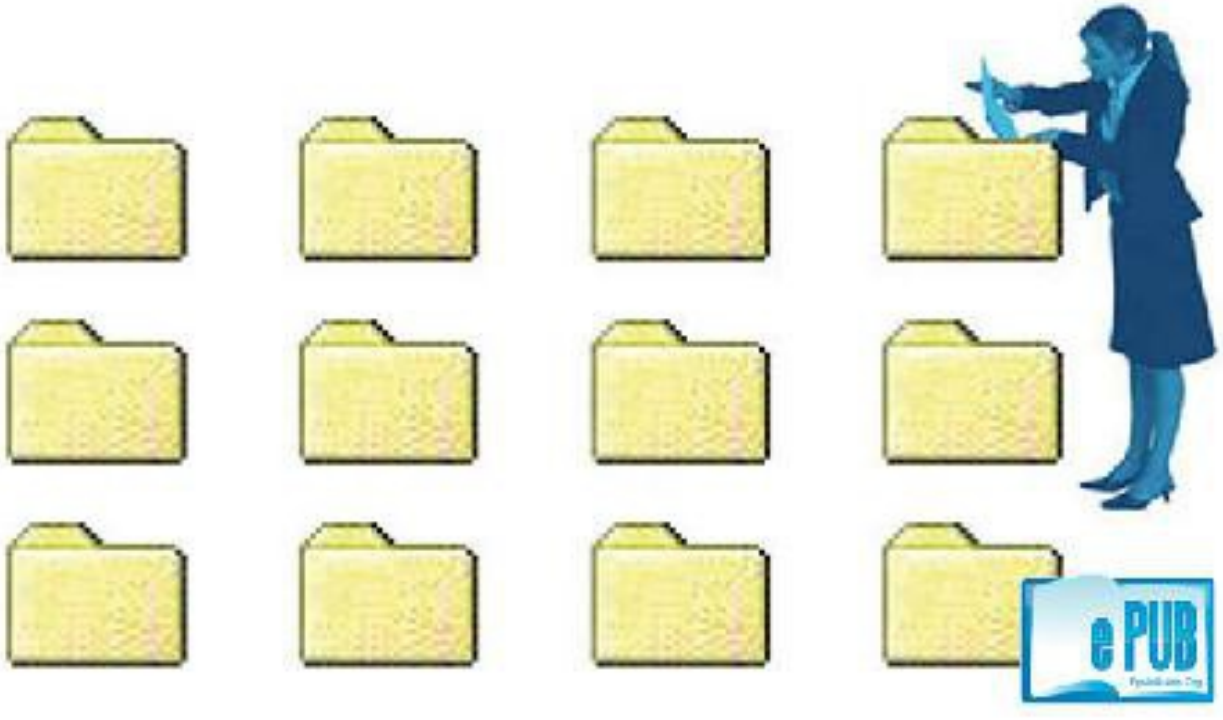
Big data

La revolución de los datos masivos

VIKTOR MAYER-SCHÖNBERGER

KENNETH CUKIER



Un análisis esclarecedor sobre uno de los grandes temas de nuestro tiempo, y sobre el inmenso impacto que tendrá en la economía, la ciencia y la sociedad en general. Los datos masivos representan una revolución que ya está cambiando la forma de hacer negocios, la sanidad, la política, la educación y la innovación. Dos grandes expertos en la materia analizan qué son los datos masivos, cómo nos pueden cambiar la vida, y qué podemos hacer para defendernos de sus riesgos. Un gran ensayo, único en español, pionero en su campo, y que se adelanta a una tendencia que crece a un ritmo frenético.

Viktor Mayer-Schönberger y Kenneth Cukier

Big Data. La revolución de los datos masivos

ePub r1.0
3L1M45145 17.10.15

Título original: *Big Data. A Revolution That Will Transform How We Live, Work, and Think*
Viktor Mayer-Schönberger & Kenneth Cukier, 2013
Traducción: Antonio Iriarte

Editor digital: Titivillus
ePub base r1.2

Para B y v
V.M.S.

Para mis padres

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 sólo 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 sólo 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 ésa 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, sólo 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 9000 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*, sólo 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 sólo 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 aún 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 sólo *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, sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 1200 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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, sólo 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 sólo 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; sólo 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 sólo cuestión de tiempo.

Aunque hace mucho que existen algunas de las técnicas de procesamiento de datos, antes sólo 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 1, 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 sólo 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 5000 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 sólo 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. Aún 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 sólo 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 sólo 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 1100 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 sólo 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 sólo 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 cuasiperfecta. 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, sólo 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 aquél 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 sólo 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 éste 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 sólo 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 sólo 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 sólo una muestra. Los valores atípicos ofrecen la información más interesante, y sólo 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 sólo 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 sólo 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 aquéllas 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 sólo 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 marketing 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 sólo 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 sólo 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 sólo 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 *marketing*. 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 sólo 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, sólo 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.

IV

CORRELACIÓN

Greg Linden tenía veinticuatro años en 1997, cuando se tomó algún tiempo libre en las investigaciones para su doctorado en inteligencia artificial en la universidad de Washington y se puso a trabajar en una *start up* local de internet de nueva creación que vendía libros online. La empresa sólo llevaba dos años abierta, pero estaba haciendo muy buen negocio. “Me encantó la idea de vender libros y conocimiento, y de ayudar a la gente a encontrar cualquier trocito de sabiduría que quisieran disfrutar”, recuerda. La tienda era Amazon.com, y contrató a Linden como ingeniero de programas para que se ocupase de que la página web funcionara sin tropiezos.

Amazon no sólo tenía jóvenes informáticos en la plantilla. Por entonces, también empleaba en torno a una docena de críticos y editores literarios para escribir reseñas y sugerir nuevos títulos. Aunque la historia de Amazon les resulta conocida a muchas personas, son menos las que recuerdan que su contenido era al principio obra humana. Los editores y los críticos evaluaban y escogían los títulos que aparecían en las páginas web de Amazon. Eran responsables de lo que se llamaba “la voz de Amazon”[37], considerada una de las joyas de la corona de la empresa y una de las fuentes de su ventaja competitiva. Por esas fechas, un artículo en *The Wall Street Journal* los celebró como los críticos literarios más influyentes de la nación, ya que impulsaban ventas enormes.

Entonces, Jeff Bezos, fundador y director general de Amazon, empezó a experimentar con una idea poderosa: ¿y si la compañía pudiese recomendar títulos específicos a sus clientes basándose en sus preferencias de compra individuales? Desde su inicio, Amazon había capturado resmas de datos sobre todos sus clientes: qué compraban, qué libros miraban pero no compraban, cuánto tiempo pasaban mirándolos y qué libros compraban al mismo tiempo.

La cantidad de datos era tan considerable que, al principio, Amazon la procesó de la forma convencional: tomando una muestra y analizándola para buscar similitudes entre los clientes. Las recomendaciones resultantes fueron toscas. Si uno compraba un libro sobre Polonia, se veía bombardeado con literatura de Europa del este. Si adquiría uno sobre bebés, le inundaban de libros similares. “Tendían a ofrecerte minúsculas variaciones sobre tu compra anterior, *ad infinitum* —recordaba James Marcus[38], que hizo reseñas de libros para Amazon de 1996 a 2001, en su ensayo *Amazonia*—. Parecía como si hubieras salido de compras con el tonto del pueblo”.

Greg Linden vio una solución. Cayó en la cuenta de que el sistema de recomendaciones no tenía por qué comparar a unas personas con otras, tarea técnicamente compleja. Lo único que necesitaba hacer era hallar asociaciones entre los propios productos. En 1998, Linden y sus colegas presentaron una solicitud de patente sobre filtrado colaborativo “artículo a artículo”, como es conocida la técnica. El cambio de enfoque supuso una gran diferencia.

Como los cálculos podían hacerse por adelantado, las recomendaciones salían a la velocidad de la luz. El método también era versátil, y capaz de funcionar cruzando categorías de productos. Así que cuando Amazon se diversificó y empezó a vender artículos distintos de libros, pudo aconsejar asimismo películas o tostadoras. Y las recomendaciones eran mejores que antes, porque el sistema empleaba todos los datos. “La broma en nuestro grupo era que, si funcionaba a la perfección, Amazon debería mostrarle a uno un solo libro, que sería el siguiente que compraría”, recuerda Linden.

Ahora, la empresa tenía que decidir qué debería aparecer en la web. ¿Contenido generado automáticamente como recomendaciones personales y listas de más vendidos, o reseñas escritas por el equipo editorial de la casa? ¿Lo que los clics decían, o lo que opinaban los críticos? Era un combate de hombres contra ratones.

Cuando Amazon llevó a cabo un test comparando las ventas que conseguían los editores tradicionales con las producidas por textos generados por ordenador, los resultados estaban a años luz. El material derivado de los datos

generaba muchísimas más ventas. Puede que el ordenador no supiese por qué un cliente lector de Ernest Hemingway podría querer comprar también a F. Scott Fitzgerald, pero eso no parecía importar. La caja registradora sonaba. Al final, a los editores se les comunicó el porcentaje exacto de ventas al que Amazon tenía que renunciar cuando presentaba sus reseñas online, y se desmanteló el equipo. “Me entristeció mucho que el equipo editorial desapareciera —recuerda Linden—, pero los datos no mienten, y el coste era muy elevado”.

Hoy en día, se dice que una tercera parte de todas las ventas de Amazon son resultado de sus sistemas de recomendación y personalización. Con estos sistemas, Amazon ha dejado fuera del negocio a numerosos competidores: no sólo a muchas grandes librerías y tiendas de música, sino también a los libreros locales que pensaron que su toque personal los aislaría de los vientos de cambio. En realidad, el trabajo de Linden revolucionó el comercio electrónico, ya que el método ha sido adoptado por casi todo el mundo. Para Netflix, una compañía de alquiler de películas online, las tres cuartas partes de los pedidos nuevos surgen de las recomendaciones[39]. Siguiendo el ejemplo de Amazon, miles de páginas web son capaces de recomendar productos, contenidos, amigos y grupos, sin saber por qué es probable que le interesen a la gente.

Sería agradable saber ése por qué, pero carece de importancia para estimular las ventas. Saber el qué, sin embargo, impulsa los clics. Esta percepción tiene la capacidad de reconfigurar muchas industrias, no sólo el *e-commerce*. Hace mucho tiempo que se les ha dicho a los comerciales de todos los sectores que necesitan comprender qué motiva a los clientes y averiguar así las razones que hay tras sus decisiones. El talento profesional y los años de experiencia siempre han sido muy valorados. El de los datos masivos demuestra que existe otro enfoque, en cierto modo más pragmático. Los innovadores sistemas de recomendación de Amazon sacaron a la luz unas correlaciones valiosas cuyas causas subyacentes no se conocen. Saber el *qué*, no el *porqué*, es más que suficiente.

PREDICCIONES Y PREDILECCIONES

Las correlaciones son útiles cuando los datos escasean, pero es en el contexto de los datos masivos donde realmente destacan. Mediante ellas podemos cobrar nuevas percepciones con mayor facilidad, más rápido y con más claridad que antes.

En esencia, una correlación cuantifica la relación estadística entre dos valores de datos. Una correlación fuerte significa que, cuando cambia uno de los valores de datos, es altamente probable que cambie también el otro. Hemos visto correlaciones fuertes en el caso de Google Flu Trends: cuanta más gente de un lugar geográfico determinado busque unos términos concretos en Google, más vecinos de allí tendrán la gripe. A la inversa, una correlación débil significa que, cuando un valor de datos cambia, apenas le ocurre nada al otro. Por ejemplo, podríamos buscar correlaciones entre la longitud de los cabellos y la felicidad, y descubriríamos que la longitud del pelo no resulta especialmente informativa sobre ese aspecto.

Las correlaciones nos permiten analizar un fenómeno dado, no aclarando cuáles son sus mecanismos internos, sino encontrando alguna aproximación útil. Por descontado, ni siquiera las correlaciones fuertes son siempre perfectas. Es bastante posible que dos cosas se comporten del mismo modo sólo por coincidencia. Puede que, simplemente, nos dejemos "engañar por la aleatoriedad"[40], como dice el empirista Nassim Nicholas Taleb. Con las correlaciones no existe la certeza, sólo la probabilidad. Pero si una correlación es fuerte, la probabilidad de que exista un vínculo es elevada. Numerosos clientes de Amazon pueden dar fe de esto, con sus estanterías cargadas de recomendaciones de la compañía.

Al permitirnos identificar una aproximación realmente buena de un fenómeno, las correlaciones nos ayudan a capturar el presente y predecir el futuro: si *A* tiene lugar a menudo junto con *B*, tendremos que estar pendientes de *B* para predecir que va a ocurrir con *A*. Usar a *B* de aproximación nos ayuda a interpretar lo que probablemente esté ocurriendo con *A*, aunque no podamos medir *A* ni observarlo directamente. Y lo importante es que nos ayuda asimismo a predecir qué podría ocurrir con *A* en el futuro. Por supuesto, las correlaciones no pueden vaticinar el futuro, sólo predecirlo con cierta probabilidad. Pero esa cualidad resulta extremadamente valiosa.

Pensemos en el caso de Walmart, la mayor cadena minorista del mundo, con más de dos millones de empleados y unas ventas anuales que rozan los 450 000 millones de dólares, más que el PIB de las cuatro quintas partes de los países del mundo. Antes de que internet sacase a la luz tantos datos, la compañía tenía tal vez el mayor conjunto de datos del sector privado de Estados Unidos. En la década de 1990, Walmart revolucionó la venta al por menor registrando cada producto como un dato mediante un sistema llamado Retail Link, que permitía a los proveedores monitorizar la tasa y el volumen de ventas y existencias. Establecer esta transparencia permitió a la empresa forzar a sus suministradores a ocuparse ellos mismos del almacenaje. En muchos casos, Walmart no asume la "propiedad" de un producto hasta que éste no llega al punto de venta, con lo que se deshace del riesgo de inventarios y reduce sus costes. Walmart se valió de los datos para convertirse, en la práctica, en la mayor tienda de consignación del mundo.

¿Qué podrían revelar todos esos datos históricos si se analizasen de forma correcta? La cadena trabajó con analistas expertos de Teradata, antes la venerable National Cash Register Company, sacando a la luz unas correlaciones interesantes. En 2004, Walmart echó un vistazo al contenido de sus gigantescas bases de datos de antiguas transacciones: qué artículo había comprado cada cliente y su coste total, qué más había en el carrito de la compra, la hora del día, e incluso el tiempo que hacía. Así, observó que antes de un huracán no sólo aumentaban las ventas de linternas, sino también las de Pop-Tarts, un dulce para el desayuno. Desde entonces, cuando se avecinaba una tormenta, para hacerles la vida más fácil a los clientes con prisa, Walmart colocaba cajas de Pop-Tarts[41] en la parte frontal de las tiendas, junto a los básicos para huracanes, y aumentó mucho sus ventas.

En el pasado, alguien de la dirección habría tenido que tener la inspiración de antemano, para luego reunir la información y someter a prueba la idea. Hoy en día, al disponer de tantísimos datos y de mejores herramientas, las correlaciones salen a la superficie más deprisa y sin coste. (Dicho esto, conviene ser precavido: cuando el número de

puntos de datos crece en orden de magnitud, también se aprecian más correlaciones espurias, fenómenos que parecen estar conectados aun cuando no sea así. Esto obliga a tener especial cuidado, como estamos empezando apenas a advertir).

Mucho antes del advenimiento de los datos masivos, el análisis de correlaciones ya había demostrado su valía. El concepto fue definido en 1888 por *sir* Francis Galton, primo de Charles Darwin, después de observar que la estatura de un hombre guardaba relación con la longitud de sus antebrazos. El cálculo matemático subyacente es relativamente sencillo y robusto, lo cual resulta ser una de sus características esenciales, y ha contribuido a convertirlo en una de las medidas estadísticas de más amplia utilización. Ahora bien, antes de la era de los datos masivos, la utilidad era limitada. Como los datos eran escasos y recopilarlos salía caro, los estadísticos a menudo escogían una aproximación, recogían entonces los datos relevantes y procedían al análisis de correlación para descubrir la calidad de ese sustituto. Pero, ¿cómo se elegía esa aproximación?

Para orientarse, los expertos utilizaron hipótesis basadas en teorías: ideas abstractas acerca de cómo funciona algo. Basándose en esas hipótesis, recopilaron datos y usaron análisis de correlación para comprobar si las aproximaciones resultaban adecuadas. Cuando no lo eran, los investigadores a menudo volvían a probar, testarudamente, por si acaso se hubiesen recogido mal los datos, antes de acabar reconociendo que la hipótesis de partida, o incluso la teoría en la que se apoyaba, era imperfecta y requería modificaciones. El conocimiento progresó mediante este método de ensayo y error, impulsado por hipótesis. Fue un proceso lento, puesto que nuestros prejuicios individuales y colectivos ofuscaban las hipótesis que desarrollábamos, cómo las aplicábamos y, por ende, las aproximaciones que escogíamos. Un proceso engorroso, pero practicable en un mundo de datos escasos.

En la era de los datos masivos, ya no resulta eficiente tomar decisiones acerca de qué variables examinar basándose únicamente en hipótesis. Los conjuntos de datos son excesivos en tamaño y el área a considerar probablemente demasiado compleja. Por suerte, muchas de las limitaciones que nos obligaron a adoptar el enfoque basado en hipótesis ya no existen en la misma medida. Ahora tenemos tantísimos datos a nuestra disposición, y tanta capacidad de procesamiento, que ya no tenemos que escoger laboriosamente una aproximación o un pequeño puñado de ellas y examinarlas una a una. Ahora, un sofisticado análisis computacional puede identificar la aproximación óptima, como hizo para Google Flu Trends después de someter a prueba casi quinientos millones de modelos matemáticos.

Ya no precisamos necesariamente de una hipótesis sustantiva válida sobre un fenómeno para empezar a entender nuestro mundo. Así pues, no tenemos que desarrollar una noción acerca de qué términos busca la gente cuando y donde se propaga la gripe. No necesitamos tener ni la menor idea de cómo fijan las líneas aéreas los precios de sus billetes. No necesitamos preocuparnos de los gustos culinarios de los clientes de Walmart. Por el contrario, podemos someter los datos masivos a análisis de correlación y dejar que éste nos cuente qué búsquedas se relacionan con la gripe, si es probable que se dispare una tarifa aérea, o qué puede apetecerle comer a una familia preocupada durante una tormenta. En lugar del enfoque sustentado por hipótesis podemos emplear uno sustentado por datos. Puede que nuestros resultados sean menos sesgados y más precisos, y es casi seguro que los obtendremos mucho más deprisa.

Las predicciones basadas en correlaciones son el corazón de los datos masivos. Los análisis de correlación se usan con tanta frecuencia hoy en día que, a veces, no valoramos bien el avance que han supuesto. Y sus usos no van a dejar de aumentar.

Por ejemplo, se están empezando a usar las calificaciones financieras del riesgo crediticio para predecir comportamientos personales. La Fair Isaac Corporation, hoy día conocida como FICO[42], inventó las calificaciones de riesgo de crédito a finales de la década de 1950. En 2011, FICO creó la “calificación de adherencia a la medicación”. Para determinar la probabilidad de que una persona tome su medicación, FICO analiza una plétora de variables, incluidas algunas que pueden parecer irrelevantes, como cuánto tiempo lleva viviendo esa persona en el mismo domicilio, si está casada, cuánto tiempo lleva en el mismo trabajo, y si posee un coche. La calificación se ha desarrollado para ayudar a los profesionales de la atención sanitaria a ahorrar dinero indicándoles a qué pacientes deberían dirigir sus recordatorios. No hay relación causal entre poseer un automóvil y tomar antibióticos de acuerdo con la prescripción; el vínculo entre las dos cosas es pura correlación. Pero algunos hallazgos como éstos bastaron para inspirarle al director general de FICO la siguiente declaración orgullosa en 2011: “Sabemos lo que usted va a hacer mañana”.

Otros *brokers* de datos están apuntándose al juego de la correlación, como ha mostrado la serie pionera “Qué saben” en *The Wall Street Journal*. Experian dispone de un producto llamado Income Insight [Percepción de Renta]

que estima el nivel de renta de la gente basándose, entre otras cosas, en su historial crediticio. Este sistema de calificación se desarrolló a partir del análisis de su enorme base de datos de historiales crediticios respecto a datos fiscales anónimos del Servicio de Renta Interna (IRS) de Estados Unidos. A una empresa le supondría un coste de unos diez dólares por unidad confirmar la renta de alguien a partir de sus declaraciones fiscales, mientras que Experian vende su estimación por menos de un dólar. En ejemplos como este, usar una aproximación resulta más rentable que pasar por todo el tráfago de lograr el producto original. Existe otra oficina de crédito, Equifax, que comercializa un "Índice de la capacidad de pago" y un "Índice de gasto discrecional" capaces, según ellos, de percibir lo rellena que tiene alguien la cartera.

Los usos de las correlaciones están ampliándose incluso más. Aviva, una gran aseguradora, ha contemplado la idea de emplear informes crediticios y datos de *marketing* de consumo como aproximación a los análisis de sangre y orina para determinados solicitantes. La idea es identificar a aquéllos con mayor riesgo de padecer enfermedades como hipertensión arterial, diabetes o depresión. El método utiliza datos acerca del estilo de vida del sujeto que incluyen cientos de variables: aficiones, páginas web que visita, horas de televisión que ve, además de estimaciones de nivel de ingresos.

El modelo predictivo de Aviva, desarrollado por Deloitte Consulting, ha demostrado buenos resultados cuando se trata de identificar riesgos de salud. Otras compañías de seguros como Prudential y AIG han examinado iniciativas similares. La ventaja de este método es que permite a la gente solicitar seguros sin tener que entregar muestras de sangre y orina, algo que a nadie le gusta, y que a las aseguradoras les cuesta dinero. Las pruebas de laboratorio salen por unos ciento veinticinco dólares por cabeza, mientras que el enfoque puramente basado en datos se queda en unos cinco dólares.

A algunas personas puede que el método les espante un poco, porque se vale de comportamientos aparentemente sin relación: es como si las empresas tuvieran a un cibersoplón que les informase de cada clic que hacemos en la pantalla. La gente podría pensárselo dos veces antes de visitar páginas web sobre deportes extremos o seguir series de televisión que ensalzan al teleadicto sedentario si creyeran que podrían suponerles primas de seguros más elevadas. Por otra parte, la ventaja es que, al hacer más fácil y más barato contratar seguros, podría subir el número de asegurados, lo que es beneficioso para la sociedad, no sólo para las aseguradoras.

Sin embargo, el paradigma, o tal vez el chivo expiatorio, de las correlaciones de datos masivos son las tiendas de descuento estadounidenses Target, que se apoyan desde hace años en predicciones basadas en correlaciones de datos masivos. En una extraordinaria demostración de periodismo, Charles Duhigg, un corresponsal de negocios de *The New York Times*, explicó cómo Target sabe cuándo está encinta una mujer sin que la futura madre lo haya hecho explícito. Básicamente, su método consiste en recoger todos los datos que puede, y en dejar que las correlaciones saquen sus propias conclusiones.

Saber si una clienta se halla en estado es importante para los comercios, porque el embarazo marca un antes y un después en las parejas, y es probable que sus hábitos de compra cambien. Pueden empezar a ir a otras tiendas y desarrollar nuevas lealtades de marca. Los responsables de *marketing* de Target se dirigieron a su división de análisis para ver si había forma de descubrir el embarazo de sus clientas por medio de sus patrones de compra.

El equipo analítico examinó el historial de compras de las mujeres que se habían apuntado en su registro de regalos para bebés. Se dieron cuenta de que habían comprado montones de cremas sin perfume en torno al tercer mes del embarazo, y que unos meses más tarde tendían a adquirir suplementos nutricionales como magnesio, calcio y cinc. El equipo acabó por identificar alrededor de dos docenas de productos que le permitían calcular una "predicción de embarazo" para cada clienta que pagaba con tarjeta de crédito, o usaba una tarjeta de fidelidad o enviaba cupones de descuento por correo. Las correlaciones incluso permitieron a la marca estimar la fecha de parto con un estrecho margen de error, de forma que podía enviar los cupones más adecuados a cada fase del embarazo.

En su libro *The Power of Habit*, Duhigg cuenta la siguiente anécdota. Un día, un hombre furioso entró en tromba en una tienda Target de las afueras de Minnesota y exigió ver al director.

—¡Mi hija ha recibido esto por correo! —gritó—. Todavía está en el instituto, ¿y ustedes se dedican a mandarle cupones para ropa de bebé y cunas? ¿Intentan animarla a que se quede embarazada?[43]

Cuando el director llamó al hombre unos días después para disculparse, se encontró con una voz más conciliadora.

—He estado hablando con mi hija —dijo el hombre—. Resulta que en mi casa han tenido lugar ciertas actividades de las que yo no estaba del todo informado. Mi hija sale de cuentas en agosto. Soy yo el que les debe una

disculpa.

Encontrar aproximaciones en contextos sociales es una de las aplicaciones de las técnicas de datos masivos. Igualmente poderosas resultan las correlaciones con nuevos tipos de datos para solucionar necesidades cotidianas.

Una de estas técnicas es la llamada analítica predictiva, que está empezando a usarse ampliamente en los negocios para predecir acontecimientos antes de que se produzcan. Puede ser, por ejemplo, un algoritmo que permita detectar una canción de éxito: muy usado en la industria musical para ofrecerles a los sellos discográficos una idea más precisa sobre dónde colocar sus apuestas. La técnica se emplea también para prevenir grandes fallos mecánicos o estructurales: colocar sensores en la maquinaria, motores o infraestructuras, como los puentes, permite monitorizar los patrones de datos que emiten, entre ellos el calor, la vibración, la tensión y el sonido, y detectar cambios que quizá indiquen problemas en el futuro.

El concepto subyacente es que cuando las cosas se vienen abajo, no suelen hacerlo de golpe, sino gradualmente. Pertrechados con datos de los sensores, el análisis de correlaciones y métodos similares pueden identificar los patrones específicos, los signos delatores que se presentan a menudo antes de que algo se estropee: el zumbido o el calentamiento excesivo de un motor, por ejemplo. A partir de ahí, uno sólo necesita estar pendiente de la aparición de ese patrón para saber que algo va mal. Advertir el patrón anormal desde el primer momento permite al sistema emitir un aviso, de forma que se pueda instalar una pieza nueva y resolver el problema antes de que la avería se produzca en realidad. La finalidad es identificar y luego vigilar esa aproximación, y así predecir los hechos venideros.

La compañía de transportes UPS lleva usando la analítica[44] predictiva desde finales de la década de 2000 para monitorizar su flotilla de 60 000 vehículos en Estados Unidos, y así saber cuándo debe llevar a cabo mantenimiento preventivo. Una avería en la carretera puede causar estragos, retrasando las entregas y recogidas. Así que, por precaución, UPS cambiaba ciertas piezas al cabo de sólo dos o tres años. Ahora bien, eso resultaba ineficiente, porque algunas estaban todavía en buen estado. Desde que se pasó a la analítica predictiva, se ha ahorrado millones de dólares al medir y monitorizar las piezas individuales y reemplazarlas sólo cuando era necesario. En un caso, los datos revelaron que todo un grupo de vehículos nuevos tenía una pieza defectuosa que podría haber acarreado serios problemas si no se hubiera detectado antes de que entrase en servicio.

También se fijan sensores en puentes y edificios para poder detectar signos de desgaste. Se emplean asimismo en grandes plantas químicas y refinerías, donde la rotura de una pieza del equipo podría paralizar la producción. El coste de recoger y analizar los datos que indican cuándo actuar preventivamente es inferior al de parar la producción. Téngase en cuenta que la analítica predictiva no explica la causa de un problema; tan sólo indica que existe. Te alertará cuando un motor se recaliente, pero no podrá decirte si se debe a una correa de ventilador deshilachada o a un casquete mal atornillado. Las correlaciones muestran el *qué*, no el *porqué*; pero, como ya se ha visto, a menudo con saber *qué* resulta suficiente.

La misma clase de metodología se usa ya en la atención sanitaria para prevenir averías de la máquina humana. Cuando un hospital le coloca un entramado de tubos, cables e instrumentos a un paciente, se genera un vasto torrente de datos. Sólo el electrocardiograma registra mil medidas por segundo. Sin embargo, sorprendentemente, en la actualidad sólo se usa y conserva una fracción de esos datos. La mayor parte se descarta, aun cuando podrían brindar pistas importantes acerca del estado del paciente y su respuesta a los tratamientos. Si se conservasen los datos y se agregasen a los de otros pacientes, podrían ofrecer perspectivas extraordinarias sobre qué tratamientos tienden a ser efectivos y cuáles no.

Descartar datos quizá resultase apropiado cuando el coste y la complejidad de recopilarlos, almacenarlos y analizarlos era elevado, pero ya no es el caso. La doctora Carolyn McGregor y un equipo de investigadores del Instituto de Tecnología de la universidad de Ontario y de IBM están trabajando junto con una serie de hospitales en el desarrollo de programas que ayuden a los médicos a tomar mejores decisiones diagnósticas en la atención de los bebés prematuros[45]. El *software* capta y procesa los datos de los pacientes en tiempo real, monitorizando dieciséis flujos de datos diferentes, tales como ritmo cardíaco, frecuencia respiratoria, temperatura, tensión y nivel de oxígeno en sangre, que, en conjunto, representan cerca de 1260 puntos de datos por segundo.

El sistema consigue detectar cambios sutiles en el estado de los prematuros, que podrían indicar el principio de una infección veinticuatro horas antes de que se manifiesten los síntomas. “Nosotros no podemos verlo a simple vista, pero un ordenador sí”, explica la doctora McGregor. El sistema no se basa en la causalidad, sino en correlaciones; dice *qué*, no *por qué*. Pero cumple su propósito. El aviso previo les permite a los médicos tratar antes

la infección por medio de intervenciones médicas más livianas, o los alerta antes si un tratamiento parece inefectivo. Ello mejora las perspectivas de los pacientes. Resulta difícil creer que esta técnica no vaya a ser aplicada a muchísimos más pacientes y enfermedades en el futuro. Puede que el algoritmo mismo no tome las decisiones, pero las máquinas hacen lo que saben hacer mejor, y ayudarán a los cuidadores a hacer lo que también ellos saben hacer mejor.

Curiosamente, el análisis de datos masivos de la doctora McGregor fue capaz de identificar correlaciones que, en cierto modo, ponen en cuestión abiertamente la ciencia convencional de los galenos. Descubrió, por ejemplo, que a menudo se detectan signos vitales muy constantes antes de una infección grave. Esto es extraño, puesto que cabría pensar que un deterioro de las condiciones vitales precedería a una infección en toda regla. Podemos imaginar a generaciones de doctores que terminaban su jornada echándole un vistazo al portapapeles al lado de la cuna, comprobando que las constantes vitales del prematuro se estabilizaban, y pensando que podían marcharse a casa sin problemas… para recibir a medianoche una llamada frenética de la maternidad informándolos de que algo había ido trágicamente mal, y su impresión fue errónea.

Los datos de McGregor sugieren que en los prematuros la estabilidad, más que un signo de mejoría, es como la calma que precede a la tempestad; como si el cuerpo del bebé estuviese ordenándole a sus minúsculos órganos que se preparen para hacer frente a la tormenta que se avecina. No podemos estar seguros: lo que los datos indican es una correlación, no una relación de causalidad. Lo que sí sabemos es que hicieron falta métodos estadísticos aplicados a una enorme cantidad de datos para desvelar esta asociación oculta. Que no quede ninguna duda: los datos masivos salvan vidas.

ILUSIONES E ILUMINACIONES

Antes, cuando eran escasos los datos disponibles, tanto las investigaciones causales como los análisis de correlación empezaban con una hipótesis, que se sometía a prueba para desmentirla o confirmarla. Sin embargo, como ambos métodos requerían de una hipótesis de partida, los dos eran igualmente susceptibles al prejuicio y a la intuición errónea. Y además, los datos necesarios no solían estar disponibles. Hoy en día, con tantos datos a nuestro alrededor, y los que quedan por venir, las hipótesis ya no resultan cruciales para el análisis correlacional.

Existe otra diferencia que está empezando a cobrar importancia. Antes del advenimiento de los datos masivos, y con la escasa capacidad de computación de entonces, la mayor parte de los análisis de correlación que usaban grandes conjuntos de datos se veían limitados a la búsqueda de relaciones lineales. En realidad, muchas relaciones son infinitamente más complejas. Con unos análisis más sofisticados, podemos identificar relaciones no lineales entre los datos.

Por poner un ejemplo, durante muchos años los economistas y los especialistas en ciencias políticas creyeron que la felicidad y el nivel de renta estaban directamente correlacionados: si se incrementa su renta, el individuo será más feliz por término medio. Observar los datos en una tabla, sin embargo, revela que la dinámica resulta mucho más compleja. Para niveles de renta por debajo de un determinado umbral, cada incremento de renta se traduce en un crecimiento sustancial de felicidad;[46] pero, por encima de ese límite, ya apenas mejora la felicidad de la persona. Si trasladáramos esto a una gráfica, la línea resultante parecería una curva más que esa línea recta que se deducía del análisis lineal.

El hallazgo fue importante para los políticos. Si habláramos de una relación lineal, tendría sentido elevar la renta de todo el mundo para aumentar la felicidad global. Ahora bien, una vez identificada la naturaleza no lineal de la relación, la recomendación cambió para centrarse en incrementar la renta de los pobres: los datos mostraban que así se conseguirían mejores resultados.

Y la cosa se complica incluso más, cuando la relación de correlación es más multifacética. Por ejemplo, unos investigadores de Harvard y del MIT examinaron la disparidad en la cobertura de la inmunización contra el sarampión: algunos grupos se vacunan y otros no. Al principio, esta disparidad parecía estar correlacionada con las cantidades que cada uno invierte en atención sanitaria. Sin embargo, un examen más minucioso reveló que la correlación no es una línea nítida, sino una curva de forma extraña. Conforme la gente va gastando más dinero en atención sanitaria, la disparidad en la inmunización disminuye (como cabía esperar); pero, cuando gastan aún más, vuelve a aumentar, curiosamente: entre los más acomodados, algunos parecen reacios a ponerse la vacuna del sarampión. Para los funcionarios de la sanidad pública resulta crucial conocer este dato, pero el simple análisis de correlación lineal no lo habría revelado.

Los expertos están empezando a desarrollar los instrumentos necesarios para identificar y comparar las correlaciones no lineales. Al mismo tiempo, a las técnicas de análisis correlacional les está sirviendo de ayuda y de potenciación un conjunto de enfoques y programas nuevos de rápido desarrollo que pueden sacar a la luz unas relaciones no causales en los datos desde muchos ángulos distintos, como pintores cubistas captando la imagen de una mujer desde perspectivas múltiples simultáneamente. Uno de los métodos nuevos más vitales se halla en el floreciente campo del análisis de redes, que hace posible cartografiar, medir y calcular nodos y enlaces para todo: desde los amigos que tiene uno en Facebook hasta quién llama a quién por el teléfono móvil, pasando por qué fallos de qué tribunales citan determinados precedentes. En conjunto, estos instrumentos ayudan a dar respuesta a preguntas empíricas y no causales.

En última instancia, en la era de los datos masivos, estos análisis novedosos conducirán a una oleada de percepciones y predicciones útiles. Veremos vínculos que nunca habíamos advertido antes. Entenderemos complejas dinámicas técnicas y sociales que durante años han escapado a nuestra comprensión, pese a todos nuestros esfuerzos. Pero lo más importante es que estos análisis no causales nos ayudarán a comprender el mundo

preguntando *qué* en lugar de *por qué*.

Al principio, esto puede parecer contradictorio. Al fin y al cabo, como seres humanos, deseamos darle sentido al mundo mediante relaciones causales: queremos creer que todo efecto tiene una causa, si lo analizamos bien. ¿No debería ser ésa nuestra mayor aspiración, la de conocer las razones que subyacen al mundo?

Ciertamente, desde hace siglos existe un debate filosófico acerca de la existencia misma de la causalidad. Si todas las cosas fueran causadas por otras, la lógica dicta entonces que no seríamos libres de decidir nada. No existiría la voluntad humana, puesto que cualquier decisión que tomáramos, cualquier pensamiento que tuviéramos, serían causados por algo que, a su vez, sería el efecto de otra causa, y así sucesivamente. La trayectoria de cualquier vida se vería determinada por la sucesión de unas causas que conducen a unos efectos. De ahí que los filósofos hayan porfiado sobre el papel de la causalidad en nuestro mundo, oponiéndola en ocasiones al libre albedrío. Ese debate abstracto, sin embargo, no es lo que aquí nos interesa.

Cuando afirmamos que los seres humanos ven el mundo a través de las causalidades, nos estamos refiriendo, más bien, a las dos formas fundamentales en que explicamos y entendemos el mundo: mediante la rápida e ilusoria causalidad, o a través de experimentos causales lentos y metódicos. Los datos masivos transformarán el papel de ambas.

Lo primero es nuestro deseo intuitivo de ver conexiones causales. Tenemos el prejuicio de asumir la presencia de causas incluso donde no existe ninguna. Esto no es debido a la cultura, formación ni nivel educativo; más bien, según indican las investigaciones, es que así funciona el conocimiento humano. Cuando vemos ocurrir dos hechos, uno después del otro, nuestras mentes sienten una gran necesidad de contemplarlos en términos causales.

Tómense las tres frases siguientes: “Los padres de Fred llegaron tarde. Los del *catering* iban a llegar enseguida. Fred estaba enfadado”. Al leerlas, intuimos instantáneamente por qué estaba enfadado Fred: no porque fueran a llegar pronto los del *catering*, sino por la tardanza de sus padres. En realidad, no tenemos forma de saberlo por la información que se nos ha suministrado. Aun así, nuestra mente no puede evitar crear lo que asumimos son historias coherentes y causales a partir de los hechos que se nos proporcionan.

Daniel Kahneman[47], profesor de psicología en Princeton y premio Nobel de Economía en 2002, usa este ejemplo para sugerir que tenemos dos formas de pensar: una es rápida y exige poco esfuerzo, dejándonos sacar conclusiones precipitadas en cuestión de segundos; la otra es lenta y trabajosa, y nos obliga a pensar a fondo sobre una cuestión determinada.

La forma rápida de pensar se inclina poderosamente a “ver” vínculos causales aunque no haya ninguno. Está prejuiciada para confirmar nuestro conocimiento y creencias preexistentes. En la antigüedad, esta forma rápida de pensar nos ayudó a sobrevivir en un entorno peligroso, en el que a menudo necesitábamos decidir rápidamente y a partir de una información limitada, aunque muchas veces no se pudiera establecer la verdadera causa de un efecto dado.

Desgraciadamente, argumenta Kahneman, muy a menudo nuestro cerebro es demasiado vago para pensar lenta y metódicamente. Al contrario, dejamos que se imponga el modo rápido. En consecuencia, con frecuencia “vemos” causalidades imaginarias y, así pues, malinterpretamos el mundo en lo fundamental.

Los padres a menudo les dicen a sus hijos que han cogido la gripe por no ponerse gorro o guantes cuando hace frío. Sin embargo, no existe relación causal directa entre el no abrigarse y la gripe. Si vamos a un restaurante y luego nos ponemos malos, tendemos de forma intuitiva a echarle la culpa a lo que ahí comimos (y puede que evitemos volver a ese restaurante en el futuro), aun cuando, posiblemente, la comida nada tuviese que ver con nuestra indisposición. Podíamos haber cogido un virus gástrico de muchas maneras, quizá dándole la mano a una persona infectada. El lado de pensamiento rápido de nuestro cerebro está programado para llegar precipitadamente a cualesquiera conclusiones causales que consiga elaborar. Así pues, a menudo nos hace tomar decisiones equivocadas.

Al contrario de la sabiduría convencional, esta intuición humana de la causalidad no acrecienta nuestra comprensión del mundo. En muchos casos, es poco más que un atajo cognitivo que nos depara una ilusión de percepción, cuando en realidad nos deja en la inopia respecto al mundo que nos rodea. De la misma forma que el muestreo era un atajo que empleábamos porque no podíamos procesar todos los datos, la percepción de la causalidad es un atajo que usa nuestro cerebro para evitar pensar despacio y a fondo.

En la época de datos escasos, demostrar lo equivocadas que estaban las intuiciones causales tomaba mucho tiempo. Esto va a cambiar. En el futuro, las correlaciones de datos masivos se utilizarán para desmentir nuestras

intuiciones causales, mostrando que, a menudo, hay muy poca o ninguna conexión estadística entre el efecto y su supuesta causa. A partir de ahora, nuestro modo de "pensamiento rápido" tendrá que vérselas con la realidad.

Quizá esa lección nos haga pensar con más ahínco (y más despacio) cuando intentemos comprender el mundo. Pero incluso nuestro pensamiento lento —la segunda forma de averiguar causalidades— verá transformado su papel por las correlaciones de datos masivos.

En nuestra vida diaria pensamos tan a menudo en términos causales que podemos llegar a creer que la causalidad es fácil de demostrar. La verdad resulta mucho menos cómoda. A diferencia de las correlaciones, en las que el cálculo matemático es relativamente directo, no existe ninguna forma matemática obvia de "probar" la causalidad. Ni siquiera podemos expresar con facilidad las relaciones causales en forma de ecuaciones normales. De ahí que, incluso pensando despacio y detenidamente, resulte difícil encontrar relaciones causales concluyentes. Como nuestras mentes están acostumbradas a la escasez de información, nos sentimos tentados a razonar con datos limitados, aun cuando muy a menudo están en juego demasiados factores para poder reducir simplemente un efecto a una causa particular.

Pensemos en el caso de la vacuna de la rabia. El 6 de julio de 1885, le presentaron al químico francés Louis Pasteur[48] al niño de nueve años Joseph Meister, que había sido atacado por un perro hidrófobo. Pasteur había inventado la vacunación y había trabajado en una vacuna experimental contra la rabia. Los padres de Meister le suplicaron a Pasteur que usara la vacuna con su hijo. Lo hizo, y Joseph Meister sobrevivió. En los periódicos, se hacían loas a Pasteur por haber salvado al pequeño de una muerte segura y dolorosa.

Pero, ¿era así? Resulta que, por término medio, sólo una de cada siete personas mordidas por un perro rabioso contrae la enfermedad. Aun asumiendo que la vacuna de Pasteur fuese efectiva, sólo habría supuesto una diferencia en uno de cada siete casos. Había una probabilidad de aproximadamente el 85 por 100 de que el niño hubiera sobrevivido en cualquier caso.

En este ejemplo, la administración de la vacuna fue considerada la causa de la curación de Joseph Meister. Ahora bien, hay dos conexiones causales en juego: una, entre la vacuna y el virus de la rabia; la otra, entre que lo muerda a uno un perro rabioso y que desarrolle la enfermedad. Aun cuando la primera sea cierta, la segunda sólo lo es en una minoría de casos.

Los científicos han superado este reto de demostrar la causalidad[49] a través de experimentos en los que la supuesta causa puede ser cuidadosamente introducida o suprimida. Si los efectos corresponden a que la causa fuese aplicada o no, queda sugerida una conexión causal. Cuanto más cuidadosamente se controlen las circunstancias, más elevada será la probabilidad de que el vínculo sea correcto.

Por consiguiente, igual que sucede con las correlaciones, la causalidad rara vez —o ninguna— puede ser demostrada, sólo mostrada con un elevado grado de probabilidad. Sin embargo, a diferencia de las correlaciones, los experimentos para confirmar conexiones causales a menudo no resultan prácticos o suscitan cuestiones éticas peliagudas. ¿Cómo podríamos realizar un experimento causal para identificar la causa de que determinados términos de búsqueda predigan mejor la gripe? Y, en el caso de la vacuna antirrábica, ¿podríamos someter a docenas, tal vez cientos de pacientes a una muerte dolorosa —como integrantes del "grupo de control" al que no se le administra la vacuna—, aun teniendo vacunas para ellos? E incluso cuando los experimentos resultan prácticos, no dejan de ser caros y requieren mucho tiempo.

En cambio, los análisis no causales como las correlaciones suelen ser rápidos y baratos, y para ellos sí disponemos de los métodos matemáticos y estadísticos que permiten analizar las relaciones, y de las herramientas digitales para demostrar su solidez con confianza.

Es más, las correlaciones no sólo son valiosas por sí mismas: también les indican el camino a las investigaciones causales. Al decirnos que dos cosas están potencialmente conectadas, nos permiten seguir investigando para ver si aparece alguna relación causal y, de ser el caso, por qué. Este mecanismo de filtración, barato y expeditivo, reduce el coste de hacer un análisis causal con experimentos controlados especialmente. A través de las correlaciones, podemos alcanzar un atisbo de las variables importantes que luego usamos en los experimentos para investigar la causalidad.

Pero hay que tener cuidado. Las correlaciones son poderosas no sólo porque ofrecen perspectivas, son también porque las que ofrecen están relativamente claras. Esas perspectivas a menudo se oscurecen en cuanto volvemos a introducir la causalidad en la foto. Por ejemplo, Kaggle, una empresa que organiza competiciones empresariales de *data mining* (minería de datos) abiertas a cualquier participante, organizó un concurso en 2012 en torno a la calidad

de los coches usados. Un vendedor de coches de segunda mano suministró datos a los estadísticos participantes para construir un algoritmo que predijera cuáles de los vehículos disponibles en una subasta eran susceptibles de sufrir fallos mecánicos. Un análisis de correlación mostró que los coches pintados de color naranja tenían menos probabilidades de tener defectos: aproximadamente la mitad de la tasa alcanzada por la media de los demás coches.

En el mismo momento de leer esto, ya pensamos en por qué podría ser así. ¿Será que es más probable que las personas propietarias de coches naranjas[50] sean entusiastas de los automóviles y cuiden mejor su vehículo? ¿Será porque un color original significa que el coche se ha construido de forma más cuidadosa y personalizada, también en otros aspectos? ¿O tal vez los coches naranja son más visibles en la carretera, y por ello tienen menos probabilidades de verse envueltos en accidentes, y así están en mejores condiciones cuando se venden de segunda mano?

Nos vemos rápidamente envueltos en una red de hipótesis causales competidoras, pero querer aclarar las cosas por esta vía sólo sirve para oscurecerlas más. Las correlaciones existen: podemos mostrarlas matemáticamente. No resulta fácil hacer lo propio con los vínculos causales. Así pues, haríamos bien en abstenernos de intentar explicar la razón que subyace a las correlaciones: el *porqué* en lugar del *qué*. De otro modo, podríamos acabar aconsejándoles a los propietarios de automóviles que pintaran sus coches viejos de naranja para que los motores fallasen menos, lo que sería una idea ridícula.

Tomando en consideración estos hechos, resulta bastante comprensible que el análisis de correlación y los demás métodos no causales basados en datos concretos sean superiores a la mayoría de las conexiones causales intuidas, resultado de “pensar rápido”. Pero cada vez en más casos, ese análisis resulta también más útil y eficiente que el lento pensamiento causal propio de los experimentos controlados cuidadosamente, y por ello largos y costosos.

Últimamente, los científicos han intentado reducir los costes de los experimentos para investigar causas, por ejemplo, combinando hábilmente diversas encuestas apropiadas para así crear “cuasiexperimentos”. Ello puede hacer más fáciles algunas investigaciones causales, pero la eficiencia de los métodos no causales resulta difícil de superar. Además, los propios datos masivos fomentan las búsquedas causales porque orientan a los expertos hacia causas probables que investigar. En muchos casos, la búsqueda profunda de la causalidad tendrá lugar después de que los datos masivos hayan hecho su trabajo, cuando deseemos específicamente investigar el *porqué*, y no sólo apreciar el *qué*.

La causalidad no va a pasar a la historia, pero está siendo derribada de su pedestal como fuente primaria de sentido. Los datos masivos potencian enormemente los análisis no causales, reemplazando a menudo a las investigaciones causales. El acertijo de las tapas de registro explosivas en Manhattan viene aquí al pelo.

EL HOMBRE CONTRA LAS TAPAS DE REGISTRO

En Nueva York, todos los años, unos cuantos centenares de tapas de registro eléctrico[51] empiezan a echar humo cuando sus entrañas se incendian. A veces, esas tapas de hierro colado, que pueden llegar a pesar casi 140 kilos, saltan por los aires y alcanzan varios pisos de altura antes de estrellarse contra el suelo. Eso no puede ser bueno.

Con Edison, la compañía pública que provee de electricidad a la ciudad, lleva a cabo inspecciones y labores de mantenimiento de los registros todos los años. Antes, básicamente confiaba en la suerte, esperando que una tapa cuya inspección estuviese programada fuese una de las que estaban a punto de estallar: como método, no era mucho mejor que el de darse un paseo sin rumbo Wall Street abajo. En 2007 Con Edison se dirigió a los estadísticos del norte de la ciudad, en la universidad de Columbia[52], con la esperanza de que pudieran usar sus datos históricos, como las listas de problemas previos y de qué infraestructura está conectada con qué, para predecir qué tapas tenían más probabilidades de sufrir problemas, de forma que la compañía supiese donde concentrar sus recursos.

Se trata de un problema de datos masivos complejo. Hay más de 150 000 kilómetros de cables subterráneos en la ciudad de Nueva York, los suficientes para darle tres veces y media la vuelta a la Tierra. Sólo Manhattan cuenta con alrededor de 51 000 pozos y cajas de acometida: parte de esta infraestructura se remonta a los tiempos de Thomas Edison, el homónimo de la compañía. Uno de cada veinte cables ha sido instalado antes de 1930. Aunque la empresa había conservado archivos desde la década de 1880, éstos estaban en un batiburrillo de formatos y nunca se habían pensado para el análisis de datos. Provenían del departamento de contabilidad o de los supervisores de emergencias encargados de redactar a mano los “partes de incidencias”. Decir que los datos eran confusos sería quedarse muy corto. Por dar un solo ejemplo, los estadísticos señalaron que el término “caja de acometida” [*Service Box*], una pieza de lo más corriente, presentaba por lo menos 38 variantes, entre ellas: SB, S, S/B, S.B, S?B, S.B., SBX, S/BX, SB/X, S&BX, S?BX, S BX, S/B/X, S BOX, SVBX, SERV BX, SERV-BOX, SERV/BOX y SERVICE BOX. Hubo que usar un algoritmo de ordenador para identificarlas todas.

“Los datos estaban increíblemente desorganizados —rememora Cynthia Rudin, la estadística y recolectora de datos, actualmente en el MIT, que dirigió el proyecto—. Tengo una copia impresa de las diferentes tablas de cables. Si la desenrollas, ni siquiera puedes sostener el papel en la mano sin que arrastre por el suelo. Y había que conseguir encontrarle sentido a todo eso; excavar para dar con el oro, conseguir como fuera un modelo predictivo realmente bueno”.

Rudin y su equipo tuvieron que usar todos los datos disponibles, no sólo una muestra, puesto que cualquiera de las decenas de miles de tapas podría ser una bomba de relojería. Así que el proyecto requería N = todo. Y, aunque habría estado bien dar con razones causales, se hubiera tardado un siglo, para llegar aun así a un resultado erróneo o incompleto. La mejor forma de cumplir la tarea era encontrar las correlaciones. A Rudin le importaba menos *por qué* que *cuál* —aunque era consciente de que cuando tuvieran que sentarse con los ejecutivos de Con Edison, los estadísticos necesitarían justificar muy bien los *rankings* que presentaran. Puede que las predicciones las hubiese elaborado una máquina, pero los consumidores eran seres humanos, y la gente tiende a querer razones, y a querer entenderlas.

El trabajo de *data mining* sacó a la luz las pepitas de oro que Rudin esperaba hallar. Después de depurar los datos confusos para que los pudiera procesar un ordenador, el equipo se puso manos a la obra con 106 indicadores que predecían una catástrofe en el subsuelo. A continuación, condensaron esa lista hasta dejarla reducida a un puñado de las señales más fuertes. En una prueba de la red eléctrica del Bronx, analizaron todos los datos de que disponían hasta mediados de 2008. Luego usaron esos datos para predecir puntos problemáticos en 2009. Y funcionó. El 10 por 100 de las tapas en cabeza de su lista incluían un asombroso 44 por 100 de las que acabaron protagonizando incidentes considerables.

Al final, los factores fundamentales resultaron ser la antigüedad de los cables y el que las tapas hubiesen experimentado problemas anteriormente. Da la casualidad de que esto fue útil, porque significó que los directivos de

Con Edison pudieron comprender fácilmente la base del *ranking*. Ahora bien, un momento, por favor. ¿Antigüedad y problemas previos? ¿No parece bastante obvio? Bueno, sí y no. Por una parte, como le gusta decir al teórico de las redes Duncan Watts, “Todo resulta obvio una vez que se conoce la respuesta” (es el título de uno de sus libros). Por otra, es importante recordar que en el modelo original había 106 factores de predicción. No era tan evidente cómo ponderarlos, y luego poner en orden varias decenas de miles de tapas, cada una con innumerables variables que, agregadas, daban millones de puntos de datos; datos que ni siquiera estaban en un formato que permitiese su análisis.

El caso de las tapas de registro explosivas pone de relieve el hecho de que los datos se están aplicando a nuevos usos para resolver problemas difíciles del mundo real. Sin embargo, hubo que cambiar de operativa. Tuvimos que emplear todos los datos, tantos como pudiésemos recopilar, y no sólo una porción. Necesitamos aceptar la confusión en vez de considerar la exactitud como una prioridad. Y tuvimos que depositar nuestra confianza en las correlaciones sin conocer del todo las bases causales de las predicciones.

¿EL FIN DE LA TEORÍA?

Los datos masivos transforman nuestra forma de comprender y explorar el mundo. En la era de los datos escasos, nos guiábamos por las hipótesis sobre cómo funcionaba el mundo, que luego intentábamos validar recopilando y analizando datos. En el futuro, nuestro entendimiento será guiado más por la abundancia de datos que por las hipótesis.

Estas hipótesis a menudo derivan de teorías de las ciencias naturales o sociales, lo que a su vez ayuda a explicar y/o a predecir el mundo que nos rodea. Mientras recorremos la transición desde un mundo dirigido por las hipótesis a un mundo dirigido por los datos, puede que nos tiente pensar que tampoco necesitamos ya las teorías.

En 2008, Chris Anderson[53], director de la revista *Wired*, proclamó que "el diluvio de datos vuelve obsoleto el método científico". En el artículo de portada, titulado "La era de los petabytes", proclamó que representaba poco menos que "el final de las teorías". El proceso tradicional del descubrimiento científico —el de la hipótesis puesta a prueba contra la realidad usando un modelo de causalidades subyacentes— está extinguiéndose, afirmaba Anderson, sustituido por el análisis estadístico de correlaciones puras carente de teoría.

Para apoyar su argumento, Anderson describía cómo la física cuántica se ha convertido en un campo casi puramente teórico, porque los experimentos son demasiado caros, demasiado complejos y demasiado grandes para resultar viables. Hay teorías —sugirió— que nada tienen que ver ya con la realidad. Como ejemplos del nuevo método, se refirió al motor de búsqueda de Google y a la secuenciación genética. "Este es un mundo en el que las cantidades masivas de datos y las matemáticas aplicadas sustituyen a cualquier otra herramienta —escribió—. Con suficientes datos, las cifras hablan por sí mismas. Los petabytes nos permiten decir: 'La correlación basta'".

Este artículo desencadenó un debate furibundo e importante, aun cuando Anderson[54] dio rápidamente marcha atrás en su afirmación más osada, pero su argumentación merece ser examinada. En esencia, Anderson sostiene que, hasta hace poco, mientras intentábamos analizar y comprender el mundo que nos rodea, necesitábamos teorías que poner a prueba. Por el contrario, en la era de los datos masivos, según reza su tesis, no necesitamos teorías: podemos limitarnos a examinar los datos. De ser cierto, esto sugeriría que todas las reglas generalizables acerca de cómo funciona el mundo, cómo se comportan los seres humanos, qué compran los consumidores, cuándo se rompen las piezas, etcétera, pueden volverse irrelevantes al imponerse el análisis de los datos masivos.

El "final de la teoría" parece implicar que, mientras que las teorías han existido en campos sustantivos como la física o la química, el análisis de datos masivos no precisa de ningún modelo conceptual. Y esto es ridículo.

El propio enfoque de los datos masivos está basado en la teoría. Por ejemplo, emplea teorías estadísticas y matemáticas y, en ocasiones, recurre también a la teoría de las ciencias informáticas. Sí, éstas no son teorías acerca de las dinámicas causales de un fenómeno determinado como la gravedad, pero no por eso dejan de ser teorías. Y, como ya se ha dicho, los modelos basados en ellas tienen un poder de predicción muy útil. De hecho, los datos masivos pueden ofrecer una perspectiva fresca y enfoques nuevos precisamente porque no están lastradas por el pensamiento convencional ni por el sesgo inherente implícito en las teorías de un campo determinado.

Es más, dado que el análisis de datos masivos está basado en teorías, no podemos huir de ellas: conforman tanto nuestros métodos como nuestros resultados. Todo empieza por cómo seleccionamos los datos. Nuestras decisiones pueden venir dictadas por la conveniencia (¿son fácilmente accesibles los datos?) o por motivos de economía (¿pueden recogerse por poco dinero?). Nuestras elecciones[55] se ven dictadas por la teoría. Lo que escogemos influye en lo que hallamos, como han argumentado danah boyd y Kate Crawford. Después de todo, Google empleó términos de búsqueda, no la longitud del cabello, como aproximación a la gripe. Igualmente, cuando analizamos los datos escogemos instrumentos que se basan en teorías. Y cuando interpretamos los resultados, aplicamos teorías de nuevo. Está claro que la era de los datos masivos no carece de teorías: están por todas partes, con todo lo que ello implica.

Hay que reconocerle a Anderson el mérito de haber planteado las preguntas correctas, y, lo que es más

importante, de haberlo hecho antes que nadie. El enfoque de datos masivos puede no suponer el “fin de la teoría”, pero sí que transforma radicalmente nuestra forma de explicar el mundo. Va a costar mucho acostumbrarse a este cambio. Supone un desafío para muchas instituciones. Sin embargo, el tremendo valor que nos brinda lo convertirá no sólo en un trueque ventajoso, sino inevitable.

Pero antes de que lleguemos a ese punto, merece la pena señalar cómo lo hicimos. A muchas personas en el ámbito de la industria tecnológica les gusta atribuirle el mérito de la transformación a las nuevas herramientas digitales, desde los chips rápidos hasta el *software* eficiente, porque son ellas las creadoras de las herramientas. La brujería técnica tiene su importancia, pero no tanta como uno pudiera pensar. La razón profunda de esta tendencia es que ahora tenemos muchos más datos. Y la razón de que dispongamos de más datos es que estamos plasmando más aspectos de la realidad en formato de datos, que es el tema del siguiente capítulo.

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 sólo 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 aquél 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 ésta 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 sólo 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 éstos podían ser *analizados*, no sólo 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, sólo 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 ésta 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 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 sólo 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 sólo 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 sólo 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 elcapítulo 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 ése 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 sólo 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. Sólo 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 sólo 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 sólo 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 sólo 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 3500 millones de personas en América Latina, África y Europa— para dar respuesta a preguntas cruciales para los ejecutivos de *marketing*, 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 sólo 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 *marketing*.

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 sólo 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 éstos 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.

VI

VALOR

A finales de la década de 1990, la red estaba convirtiéndose a marchas forzadas en un lugar ingobernable, inhóspito y hostil. Los *spambots* inundaban los buzones de entrada de los correos electrónicos y empantanaban los foros online. En el año 2000, Luis von Ahn[71], un joven de veintidós años que acababa de obtener su licenciatura universitaria, tuvo una idea para resolver el problema: obligar a los que se inscribieran a probar que eran humanos. Así que buscó algo que les resultase fácil de hacer a los seres humanos, pero les costara mucho a las máquinas.

Se le ocurrió la idea de presentar letras garabateadas y difíciles de leer durante el proceso de registro. Las personas serían capaces de descifrarlas y reproducirlas en el orden correcto en unos segundos, pero los ordenadores se quedarían atascados. Yahoo implementó su método y redujo el flagelo de los *spambots* en el lapso de veinticuatro horas. Von Ahn llamó a su creación Captcha (por las siglas inglesas de Completely Automated Public Turing Test to Tell Computers and Humans Apart, o test público de Turing completamente automatizado para diferenciar a los seres humanos de las computadoras). Cinco años después, se tecleaban a diario alrededor de 200 millones de Captchas.

Captcha le valió a Von Ahn una fama considerable y un trabajo de profesor de informática en la universidad Carnegie Mellon después de obtener el doctorado. También resultó decisivo a la hora de que se le otorgase, a los 27 años, uno de los prestigiosos premios “genio” de la fundación MacArthur, dotado con medio millón de dólares. Pero cuando cayó en la cuenta de que era responsable de que millones de personas perdiesen montones de tiempo a diario tecleando tediosas letras temblonas —vastas cantidades de información que sencillamente se descartaba después—, no se sintió tan listo.

Buscando formas de darle un uso más productivo a todo ese poder computacional humano, Von Ahn inventó un sucesor, que bautizó apropiadamente como ReCaptcha. En lugar de introducir letras aleatorias, la gente teclea dos palabras procedentes de proyectos de escaneo de textos que el programa de reconocimiento óptico de caracteres de un ordenador no podría entender. La primera palabra sirve para confirmar lo que han introducido otros usuarios, y es, por consiguiente, una señal de que el usuario es humano; la otra es una palabra nueva que precisa desambiguación. Para garantizar la exactitud, el sistema le presenta la misma palabra borrosa a una media de cinco personas distintas, que la deben insertar correctamente antes de darla por válida. Los datos tenían un uso primario —demostrar que el usuario era un ser humano—, pero también tenían un propósito secundario: descifrar palabras que no estuviesen claras en textos digitalizados.

El valor que produce esto es inmenso, si se considera lo que costaría contratar a personas que lo hicieran en su lugar. Con aproximadamente diez segundos por uso, 200 millones de ReCaptchas diarios ascienden a medio millón de horas diarias. El salario mínimo en Estados Unidos era de 7,25 dólares brutos por hora en 2012. Si uno tuviera que dirigirse al mercado para desambiguar las palabras que un ordenador no había conseguido descifrar, costaría alrededor de cuatro millones de dólares diarios, o más de mil millones de dólares al año. En su lugar, Von Ahn diseñó un sistema para hacerlo de forma efectiva, y gratis. Y Google lo pone gratis a disposición de cualquier web que desee usarlo; hoy en día está incorporado en unas 200 000 páginas web, entre ellas Facebook, Twitter y Craigslist.

La historia de ReCaptcha pone de relieve la importancia de la reutilización de los datos. Con el advenimiento de los datos masivos, el valor de los datos está cambiando. En la era digital, ya no son sólo un apoyo de las transacciones, sino que muchas veces se convierten en el bien mismo objeto del intercambio. En un mundo de datos masivos, las cosas cambian de nuevo. El valor de los datos se desplaza de su uso primario a sus potenciales usos futuros. Esto tiene profundas consecuencias: afecta a cómo valoran las empresas los datos que poseen y a quién le

permiten acceder a ellos. Les permite a las compañías, y puede que las obligue, a cambiar sus modelos de negocio. Modifica cómo piensan en los datos las organizaciones y cómo los usan.

La información siempre ha sido esencial para las transacciones de mercado. Los datos permiten descubrir el precio, por ejemplo, que es la señal de cuánto hay que producir. Esta dimensión de los datos ya está bien captada. Ciertas clases de información llevan mucho tiempo negociándose en los mercados: el contenido de libros, artículos, música y películas constituye un ejemplo, igual que la información financiera, o los precios bursátiles. En las últimas décadas, los datos personales se han unido a éstos. En Estados Unidos hay agentes especializados en datos como Acxiom, Experian y Equifax que facturan importes considerables por unos expedientes exhaustivos de información personal sobre cientos de millones de consumidores. Con Facebook, Twitter, LinkedIn y otras plataformas de medios sociales, nuestras conexiones, opiniones y preferencias personales, los patrones de nuestra vida diaria, se han unido al conjunto de información personal que está disponible acerca de nosotros.

En resumen, aunque hace mucho que los datos son valiosos, o bien se consideraban algo secundario, que servía a las operaciones centrales de llevar un negocio, o bien se limitaban a categorías relativamente específicas, como la propiedad intelectual o la información personal. Por el contrario, en la era de los datos masivos, *todos* los datos serán considerados valiosos, en y por sí mismos.

Cuando decimos “todos los datos”, queremos decir incluso los pedacitos de información más brutos, más aparentemente triviales. Piénsese en las mediciones de un sensor de temperatura de la máquina de una fábrica. O en el flujo en tiempo real de coordenadas de GPS, lecturas de acelerómetro y niveles de combustible de un solo vehículo de reparto, o de una flotilla de 60 000. O piénsese en los miles de millones de antiguas búsquedas de información, o en el precio de, prácticamente, cada asiento de cada vuelo comercial en Estados Unidos remontándose años atrás.

Hasta hace muy poco, no había forma fácil de recopilar, almacenar y analizar estos datos, lo que restringía severamente las oportunidades de extraer su valor potencial. En el célebre ejemplo de Adam Smith del fabricante de alfileres[72], con el que ilustraba la división del trabajo en el siglo XVIII, hubiera hecho falta tener a observadores vigilando a todos los trabajadores, no sólo para un estudio determinado, sino todos los días y tomando medidas detalladas, y contando la producción sobre papel grueso con plumas de ave. Incluso medir el tiempo resultaba difícil entonces, ya que aún no se habían generalizado los relojes fiables. Las limitaciones del entorno técnico condicionaban las opiniones de los economistas clásicos acerca de la constitución de la economía, algo de lo que apenas eran conscientes, del mismo modo que el pez no se da cuenta de que está mojado. Por esta razón, cuando consideraban los factores de producción (tierra, trabajo y capital), el papel de la información normalmente brillaba por su ausencia. Aunque el coste de recopilar, almacenar y utilizar datos ha disminuido a lo largo de los últimos dos siglos, hasta hace muy poco aún resultaba relativamente elevado.

Lo que distingue a nuestra época es que muchas de las limitaciones inherentes a la recopilación de datos ya no existen. La tecnología ha llegado a un punto en el que es capaz de capturar y almacenar unas cantidades inmensas de información por poco dinero. Resulta frecuente recopilar datos de forma pasiva, sin mucho esfuerzo o sin que sean conscientes siquiera aquellos que están siendo registrados. Y como el coste del almacenamiento ha caído tanto, es más fácil justificar la conservación de los datos que el deshacerse de ellos. Todo esto hace que haya muchos más datos disponibles y a menor coste que nunca antes. A lo largo del pasado medio siglo, el coste del almacenamiento digital se ha dividido por la mitad cada dos años, mientras que la densidad del almacenaje se ha incrementado cincuenta veces. A la vista de firmas de información como Farecast o Google —en las que los datos brutos entran por un extremo de la cadena de montaje digital, y por el otro sale la información procesada—, los datos están empezando a parecer un nuevo factor de producción.

El valor inmediato de la mayor parte de los datos resulta evidente para quienes los recopilan. De hecho, probablemente los recogen con un objetivo específico en mente. Las tiendas reúnen cifras de ventas para llevar una contabilidad correcta. Las fábricas monitorizan su producción para asegurarse de que esta cumple las normas de calidad. Las páginas web toman nota de cada clic de los usuarios —y a veces, incluso de por dónde pasa el cursor del ratón— para analizar y optimizar el contenido que ofrecen a los visitantes. Estos usos primarios de los datos justifican su recopilación y procesamiento. Cuando Amazon registra no sólo los libros que adquieren sus clientes, sino las páginas web que simplemente miran, sabe que usará esos datos para ofrecer recomendaciones personalizadas. Igualmente, Facebook sigue los cambios en las “actualizaciones de estado” y los “me gusta” de sus usuarios para establecer qué anuncios son los más adecuados en su página y generar ingresos.

A diferencia de las cosas materiales —los alimentos que comemos, una vela que arde—, el valor de los datos no disminuye cuando éstos se usan; pueden volver a procesarse una y otra vez. Los datos constituyen lo que los economistas llaman un bien “no rival”: su uso por una persona no impide que los use otra. Y la información no se desgasta con el uso, como sí lo hacen los bienes materiales. De ahí que Amazon pueda recurrir a datos de transacciones anteriores para hacer recomendaciones a sus clientes; y usarlos repetidamente, no sólo para el cliente que generó los datos, sino para otros muchos también.

Asimismo —lo que es más relevante—, los datos pueden ser explotados con propósitos múltiples. Este punto es importante a la hora de intentar entender cuánto valor tendrá para nosotros la información en la era de los datos masivos. Ya hemos visto concretarse parte de este potencial, cuando Walmart revisó su base de datos de recibos antiguos y advirtió aquella lucrativa correlación entre los huracanes y la venta de Pop-Tarts.

Todo esto sugiere que el valor pleno de los datos es mucho mayor que el que se obtiene de su primer uso. Significa también que las compañías pueden explotar datos de forma efectiva incluso cuando el primer uso, o cada uso subsiguiente, sólo aporten una pequeña cantidad de valor, siempre y cuando utilicen los datos muchas veces.

EL «VALOR DE OPCIÓN» DE LOS DATOS

Para hacerse una idea de lo que implica la reutilización de datos de cara a definir su valor último, piénsese en los coches eléctricos[73]. El que tengan éxito como modo de transporte depende de una vertiginosa variedad de facetas logísticas que tienen todas algo que ver con la duración de la batería. Los conductores necesitan recargar las baterías de sus coches de forma rápida y cómoda, y las compañías eléctricas necesitan garantizar que la energía consumida por estos vehículos no desequilibre la red de suministro. Hoy en día, tenemos una distribución bastante efectiva de las gasolineras, pero todavía no sabemos cuáles serán las necesidades de recarga y la localización de las estaciones de servicio para vehículos eléctricos.

Sorprendentemente, esto no supone tanto un problema de infraestructura cuanto de información. Y los datos masivos forman parte importante de la solución. En 2012, IBM colaboró en una prueba con la Pacific Gas and Electric Company de California y el fabricante de automóviles Honda para la recogida de vastas cantidades de información que respondieran a algunas preguntas fundamentales acerca de cuándo y dónde se surtirían los coches eléctricos, y lo que ello implicaría para el suministro de energía. IBM desarrolló un complejo modelo predictivo basado en numerosos factores: el nivel de la batería del coche, su localización, la hora del día y las tomas disponibles en las estaciones de recarga más próximas. Luego, emparejó esos datos con el consumo actual de la red, por una parte, y por otra con los patrones de consumo histórico de energía. El análisis de los enormes flujos de datos, históricos y en tiempo real, procedentes de múltiples fuentes, permitió a IBM determinar los momentos y lugares óptimos para que los conductores cargaran la batería de su vehículo. También reveló dónde sería mejor construir las estaciones de recarga. A la larga, el sistema necesitará tomar en cuenta las diferencias de precios entre estaciones de recarga próximas. Hasta las previsiones meteorológicas habrán de ser computadas: por ejemplo, si el día es soleado y una estación cercana de recarga alimentada por energía solar está hasta los topes de electricidad, pero se prevé una semana de lluvia durante la cual los paneles solares quedarán inactivos.

El sistema toma información generada para un propósito concreto y la reutiliza para otro: en otras palabras, los datos pasan de unos usos primarios a otros secundarios, lo que los vuelve mucho más valiosos a lo largo del tiempo. El indicador de carga de la batería del coche les dice a los conductores cuándo tienen que “llenarla”. La compañía eléctrica recopila los datos de uso de la red para gestionar su estabilidad. Esos son los usos primarios. Los dos conjuntos de datos alcanzan usos secundarios —y generan nuevo valor— cuando se aplican a un propósito del todo diferente: determinar dónde y cuándo recargar, y dónde construir estaciones de servicio para vehículos eléctricos. Por añadidura, se incorpora información secundaria, como puede ser la localización del coche y el historial de consumo de la red eléctrica. IBM procesa esos datos no una vez, sino una y otra y otra según actualiza de forma continua su perfil del consumo de energía por parte de los vehículos eléctricos y la presión que ejercen sobre la red eléctrica.

El auténtico valor de los datos es como un iceberg que flota en el océano. Sólo se distingue a primera vista una minúscula porción: la mayor parte queda oculta bajo la superficie. Las empresas innovadoras que entienden esto pueden extraer ese valor oculto y obtener beneficios potencialmente enormes. En resumen, el valor de los datos ha de verse a la luz de las muchas formas distintas en que pueden ser empleados en el futuro, no solamente por cómo se usan ahora mismo, como se extrae de muchos de los ejemplos que hemos resaltado antes. Farecast explotó datos acerca de la venta de billetes de avión para predecir el precio futuro del pasaje aéreo. Google volvió a utilizar términos de búsqueda para revelar la prevalencia de la gripe. El doctor McGregor vigiló los indicadores vitales de un niño para predecir el inicio de una infección. Maury le otorgó otra finalidad a unos antiguos cuadernos de bitácora, desvelando las corrientes oceánicas.

Con todo, la importancia de la reutilización de los datos no es apreciada por entero en la sociedad, ni en el ámbito de los negocios. Pocos son los directivos de Con Edison que habrían podido imaginar que la información sobre el cableado y los registros de mantenimiento con siglos de antigüedad podrían prevenir futuros accidentes.

Hizo falta una nueva generación de estadísticos, y una nueva generación de métodos y herramientas, para sacar a la luz el valor de los datos. Muchas compañías tecnológicas y de internet se han mostrado incapaces, hasta hace muy poco, de advertir cuán valiosa puede ser la reutilización de los datos.

Quizá resulte útil considerar los datos de la misma forma que los físicos contemplan la energía "almacenada" o "potencial" que existe en el seno de un objeto, en suspenso. Piénsese en un resorte comprimido o en una pelota que se halla en lo alto de un monte. La energía contenida en esos objetos permanece latente —potencial— hasta que se libera, digamos cuando se dispara el resorte o se le da un empujón a la pelota para que ruede ladera abajo. Ahora, la energía de esos objetos se ha vuelto "cinética", porque están en movimiento, ejerciendo su fuerza sobre otros objetos en el mundo. Después de su uso primario, el valor de los datos aún existe, pero yace durmiente, almacenando su potencial como el resorte o la pelota, hasta que los datos son aplicados a un uso secundario y su poder se ve liberado de nuevo. En la era de los datos masivos, por fin disponemos de la mentalidad, el ingenio y los instrumentos necesarios para extraer el valor oculto de los datos.

En última instancia, el valor de los datos es lo que uno puede ganar de todas las formas posibles en que los emplee. Estos usos potenciales aparentemente infinitos son como las opciones: no en el sentido financiero, sino en el sentido práctico de las opciones que brindan. El valor de los datos es la suma de esas elecciones: su "valor de opción", por así decir. Antes, una vez completado el uso principal de los datos, se pensaba que éstos habían cumplido su propósito, y estábamos listos para eliminarlos, o dejar que se perdiesen. Después de todo, ya se les había extraído el valor principal. En la era de los datos masivos, la información es como una mina de diamantes mágica que sigue produciendo mucho después de haberse agotado su veta principal. Existen tres vías poderosas para desencadenar el valor de opción de los datos: la reutilización básica, la fusión de conjuntos de datos y el hallazgo de combinaciones "dos por uno".

LA REUTILIZACIÓN DE DATOS

Un ejemplo clásico de reutilización innovadora de datos lo ofrecen los términos de búsqueda. A primera vista, la información parece carecer de valor después de haber cumplido su uso primario. La interacción momentánea entre el consumidor y el motor de búsqueda produjo una relación de páginas web y anuncios que cumplieron un propósito específico exclusivo de ese momento. Pero las búsquedas antiguas pueden resultar extraordinariamente valiosas. Una compañía como Hitwise, que se dedica a medir el tráfico en la red, propiedad del agente de datos Experian, permite a sus clientes investigar el tráfico de búsquedas para descubrir las preferencias de los consumidores. Los profesionales de la distribución, por ejemplo, pueden recurrir a Hitwise para hacerse una idea de si el rosa volverá a estar de moda en primavera, o si ha vuelto el negro. Google pone a disposición del público una versión de su sistema analítico de términos de búsqueda para su examen sin restricciones, y ha lanzado un servicio de pronósticos de negocios con el BBVA, el segundo mayor banco de España, para examinar el sector turístico y comercializar indicadores económicos en tiempo real según las búsquedas. El Bank of England usa las búsquedas relacionadas con la propiedad inmobiliaria para conseguir una percepción más ajustada de si los precios de la vivienda están subiendo o bajando.

Las empresas que no supieron apreciar la importancia de reutilizar datos han aprendido la lección por las malas. Por ejemplo, en sus inicios Amazon firmó un acuerdo con AOL para suministrarle la tecnología necesaria a su web de comercio electrónico. A la mayor parte de la gente le pareció un vulgar acuerdo de externalización, pero lo que a Amazon le interesaba de verdad —explica Andreas Weigend, antiguo científico jefe de Amazon— era hacerse con datos acerca de qué miraban y compraban los usuarios de AOL[74], lo que le permitiría mejorar el rendimiento de su motor de recomendaciones. La pobre AOL nunca cayó en la cuenta de esto: sólo veía el valor de los datos en términos de su propósito fundamental, las ventas. La astuta Amazon sabía que podía obtener réditos aplicando esos datos a un uso secundario.

Considérese, si no, el caso de la entrada de Google en el terreno del reconocimiento de voz con su GOOG-411 para búsquedas de listados locales, que estuvo operativo de 2007 a 2010. El gigante de las búsquedas no tenía tecnología propia de reconocimiento de voz, así que tuvo que subcontratarla. Alcanzó un acuerdo con la firma líder en ese campo, Nuance[75], que se entusiasmó al conseguir un cliente tan prestigioso. Pero en Nuance no eran muy listos en lo que a los datos masivos se refiere: el contrato no especificaba a quién le correspondían los registros de

traducción de voz, y Google se los quedó. El análisis de los datos le permite a uno estimar la probabilidad de que un pequeño fragmento de voz digitalizada corresponda a una palabra determinada. Esto resulta esencial para mejorar la tecnología de reconocimiento de voz, o crear un servicio enteramente nuevo. Por aquel entonces, Nuance consideraba que su negocio era licenciar *software*, no procesar datos. En cuanto se dio cuenta de su error, empezó a cerrar acuerdos con operadores de telefonía móvil y fabricantes de microteléfonos para que hicieran uso de su servicio de reconocimiento de voz, para así poder recoger los datos.

El valor presente en la reutilización de datos constituye una buena noticia para aquellas organizaciones que recopilan o controlan grandes conjuntos de datos, pero hacen poco uso de ellos en estos momentos, como pueden ser las empresas convencionales que, esencialmente, operan *offline*. Puede que estén sentadas sobre auténticos géiseres de información sin explotar. Otras compañías pueden haber recogido datos, haberlos empleado una vez (si acaso), y conservado sólo por su bajo coste de almacenamiento en "tumbas de datos", como llaman los especialistas a los lugares donde reside ese tipo de información antigua.

Las compañías tecnológicas y de internet se hallan a la vanguardia de la explotación del diluvio de datos, dado que recopilan muchísima información sólo por estar online, y le llevan ventaja al resto de la industria a la hora de analizarla. Pero todas las empresas están en condiciones de obtener ganancias. La consultora McKinsey & Company cita el caso de una empresa logística —cuya identidad mantiene en el anonimato— que advirtió que en el transcurso de sus entregas de mercancía estaba amasando toneladas de información sobre envíos de productos en todo el mundo. Olfateando una oportunidad, la empresa estableció una división especial para vender esos datos agregados bajo la forma de previsiones económicas y de negocios. En otras palabras, creó una versión *offline* del negocio de antiguas búsquedas de información de Google. O tómese el caso de SWIFT, el sistema bancario mundial de transferencias electrónicas. Observó que los pagos se correlacionan con la actividad económica mundial, así que SWIFT ofrece predicciones sobre la evolución del PIB basadas en los datos de transferencia de fondos que circulan por su red.

Algunas firmas, por su posición en la cadena de valor de la información, pueden hallarse en situación de recopilar enormes cantidades de datos, aun cuando tengan poca necesidad inmediata de ellos o no estén muy versadas en su reutilización. Por ejemplo, los operadores de telefonía móvil reúnen información acerca de la localización de sus clientes con el objeto de encaminar las llamadas. Para esas compañías, tal información sólo tiene usos técnicos limitados. Ahora bien, se torna más valiosa cuando la reutilizan empresas[76] que distribuyen publicidad y promociones personalizadas basadas en la localización. A veces, el valor no surge de puntos de datos individuales, sino de lo que revelan agregados. De ahí que algunas empresas de geolocalización, como AirSage y Sense Networks, que vimos en el capítulo anterior, puedan vender información sobre dónde se congrega la gente una noche de viernes, o la velocidad a la que se mueven los coches en el tráfico. Esta masa de datos puede usarse para determinar el valor de las propiedades inmobiliarias o el precio de las vallas publicitarias.

Hasta la información más corriente puede tener un valor especial si se aplica de la forma correcta. Pensemos nuevamente en los operadores de telefonía móvil: tienen registros sobre dónde y cuándo se conectan los teléfonos a las estaciones base, y con qué intensidad de señal. Los operadores llevan mucho tiempo usando esos datos para ajustar el rendimiento de sus redes, y decidir dónde hay que ampliar o modernizar la infraestructura. Pero los datos tienen otros muchos usos potenciales. Los fabricantes de teléfonos móviles podrían usarlos para averiguar qué influye en la intensidad de la señal, por ejemplo, mejorando así la calidad de recepción de sus dispositivos. Los operadores de telefonía móvil han sido mucho tiempo reacios a vender esa información, por miedo a infringir normas sobre la protección de datos de carácter personal. Sin embargo, están empezando a cambiar de actitud al ver zozobrar sus finanzas, y contemplar sus datos como una fuente potencial de ingresos. En 2012, Telefónica, la gran compañía de telecomunicaciones española e internacional, llegó incluso al extremo de establecer una empresa independiente, llamada Telefónica Digital Insights, para vender datos anónimos y agregados de localización de suscriptores a empresas de distribución y otros sectores.

DATOS RECOMBINANTES

En ocasiones, el valor durmiente sólo puede ser liberado al combinar un conjunto de datos con otro, tal vez muy distinto. Podemos hacer cosas innovadoras al mezclar datos de maneras nuevas. Un ejemplo nos lo brinda un

inteligente estudio publicado en 2011 sobre si los teléfonos móviles incrementan la probabilidad de tener cáncer. Con alrededor de seis mil millones de móviles en el mundo, casi uno por cada ser humano de la Tierra, la cuestión es crucial. Muchos estudios han intentado encontrar una relación, pero se han visto menoscabados por ciertas deficiencias: los tamaños de las muestras eran demasiado pequeños, o los periodos considerados eran demasiado cortos, o se basaban en datos de elaboración propia que estaban llenos de errores. Sin embargo, un equipo de investigadores de la Asociación Danesa contra el Cáncer desarrolló un enfoque interesante basado en datos recopilados previamente.

De los operadores de telefonía móvil[77] se obtuvieron datos sobre todos los abonados desde la introducción de los teléfonos móviles en Dinamarca. El estudio se centró en aquellos que tenían móviles entre 1987 y 1995, con la excepción de los teléfonos de empresa y de otros para los que no se disponía de datos socioeconómicos. El total ascendía a 358 403 personas. Dinamarca disponía, además, de un registro nacional de todos los pacientes de cáncer, que recogía a 10 729 personas que padecieron tumores del sistema nervioso central entre 1990 y 2007, el periodo de seguimiento. Por último, el estudio hizo uso de un registro nacional con información sobre el nivel educativo y la renta disponible de cada ciudadano danés.

Después de combinar los tres conjuntos de datos, los investigadores estudiaron si los usuarios de móvil mostraban una mayor incidencia de casos de cáncer que los no abonados. Entre los primeros, ¿tenían más posibilidades de desarrollar un cáncer quienes habían tenido teléfono móvil durante un tiempo más largo?

A pesar de la escala del estudio, los datos no eran en absoluto desordenados ni imprecisos: los conjuntos de datos se habían establecido bajo estándares de calidad exigentes por razones médicas, comerciales o demográficas. La información no se había recopilado de ninguna forma que pudiera introducir sesgos relacionados con el tema del estudio. En realidad, los datos se habían generado varios años antes, y por razones que nada tenían que ver con este proyecto de investigación. Lo más importante era que el estudio no se basaba en una muestra, sino en algo que se acercaba mucho a N = todo: casi todos los casos de cáncer y prácticamente todos los usuarios de teléfonos móviles, que ascendían a 3,8 millones de años-persona de propiedad de celulares. El hecho de que contuviese casi todos los casos significaba que los investigadores podían buscar subpoblaciones, como la de quienes tenían un nivel de ingresos elevado.

Al final, el grupo no detectó ningún incremento en el riesgo de cáncer asociado con el uso de teléfonos móviles. Por esa razón, los hallazgos apenas tuvieron eco en los medios de comunicación cuando se publicaron en octubre de 2011 en la revista médica inglesa *BMJ*. Pero si se hubiese descubierto tal vínculo, el estudio habría sido noticia de primera página en todo el mundo, y la metodología de los “datos recombinantes” habría experimentado un gran espaldarazo.

Con datos masivos, la suma es más valiosa que sus partes, y cuando recombinamos juntas las sumas de múltiples conjuntos de datos, también esa suma vale más que sus componentes individuales. Hoy en día, los usuarios de internet están familiarizados con los *mashups* [revoltijos] básicos que combinan dos o más fuentes de datos de forma novedosa. Por ejemplo, la página web de propiedad inmobiliaria Zillow superpone información y precios de bienes inmuebles sobre un mapa de vecindarios en Estados Unidos. Asimismo, procesa toneladas de datos, como, por ejemplo, las transacciones recientes en el vecindario y la descripción de las propiedades, para pronosticar el valor de las casas específicas en un área. La presentación visual hace más accesibles los datos. Sin embargo, con datos masivos se puede ir más allá. El estudio danés sobre el cáncer nos ofrece una pista de sus posibilidades.

DATOS EXTENSIBLES

Una forma de hacer posible la reutilización de los datos es incorporar la extensibilidad en su diseño desde el principio, de forma que sean adecuados para usos múltiples. Aunque esto no siempre resulta posible —ya que uno puede identificar los usos posibles sólo mucho después de haber recogido los datos—, hay formas de incentivar los usos múltiples de un mismo conjunto de datos. Por ejemplo, algunos minoristas instalan las cámaras de vigilancia en sus tiendas de forma que no sólo permitan detectar a los ladrones, sino también seguir el flujo de clientes a través del local y tomar nota de dónde se paran más a mirar. Los comercios pueden usar esta última información para diseñar la mejor distribución posible de la tienda, así como para valorar la efectividad de las campañas de *marketing*. Antes de esto, las videocámaras únicamente se usaban por seguridad; ahora se consideran una inversión que puede

incrementar los ingresos.

No resulta sorprendente que una de las mejores empresas a la hora de recopilar datos con miras a su extensibilidad sea Google. Sus controvertidos coches Street View recorrían las ciudades tomando fotos de casas y calles, pero también absorbiendo datos de GPS[78], comprobando información cartográfica e incluso chupando nombres de redes *wifi* (y, de forma tal vez ilegal, el contenido que fluía por las redes inalámbricas abiertas). Un mero trayecto de Google Street View amasaba en cada momento una enormidad de flujos de datos discretos. La extensibilidad aparece porque Google aplicaba los datos no sólo a un uso primario, sino a montones de usos secundarios. Por ejemplo, los datos que recogía de los GPS mejoraron el servicio cartográfico de la empresa, y resultaron indispensables para el funcionamiento de su coche autoconducido.

El coste adicional de recopilar múltiples flujos o muchos más puntos de datos de cada flujo no suele ser muy alto. Así pues, tiene sentido recoger cuantos más datos sea posible, y hacerlos extensibles considerando de partida los potenciales usos secundarios. Ello incrementa el valor de opción de los datos. La cuestión estriba en buscar "dos por uno", en que un único conjunto de datos puede ser usado en múltiples ocasiones si se ha recopilado de una forma determinada. De este modo, los datos pueden hacer doble servicio.

LA DEPRECIACIÓN DEL VALOR DE LOS DATOS

Dado que el coste de archivar los datos digitales ha caído en picado, las empresas tienen una motivación económica fuerte para conservar los datos de cara a volver a usarlos con un fin idéntico o similar. Existe, sin embargo, un límite a su utilidad.

Por ejemplo, cuando empresas como Netflix y Amazon convierten las compras, visitas y reseñas de sus clientes en recomendaciones de nuevos productos, pueden sentirse tentadas a usar los registros muchas veces a lo largo de muchos años. Teniendo esto presente, podría argumentarse que mientras una empresa no se vea constreñida por límites legales y regulatorios, como las leyes de privacidad, debería usar sus archivos digitales siempre, o por lo menos, mientras sea económicamente viable. Sin embargo, la realidad no es tan sencilla.

La mayoría de los datos pierde parte de su utilidad con el paso del tiempo. En tales circunstancias, seguir apoyándose en datos antiguos no sólo deja de aportar valor, sino que acaba por destruir de hecho el valor de los datos más frescos. Piense en un libro que compró en Amazon hace diez años, y que puede que ya no corresponda a sus intereses. Si Amazon emplea esa compra de hace diez años para aconsejarle otros títulos, resulta menos probable que los compre usted, o que le importen siquiera las subsiguientes recomendaciones que le manden. Dado que las recomendaciones de Amazon se basan simultáneamente en información desfasada y datos más recientes, aún valiosos, la presencia de los datos viejos disminuye el valor de los más recientes.

Así que la empresa tiene un tremendo incentivo para usar los datos únicamente mientras sigan siendo productivos. Necesita cuidar de sus hallazgos de forma continua y desechar la información que se ha quedado sin valor. El reto es saber cuáles son esos datos que ya no resultan útiles. Basar esa decisión únicamente en el tiempo transcurrido raras veces resulta adecuado. De ahí que Amazon y otras empresas hayan puesto a punto unos mecanismos sofisticados para ayudarse a separar los datos útiles de los irrelevantes. Por ejemplo, si un cliente mira o compra un libro que le ha sido recomendado en función de una compra anterior, las compañías de comercio electrónico pueden inferir que la adquisición más antigua todavía representa las preferencias actuales del comprador. De ese modo, son capaces de valorar la utilidad de los datos más antiguos, y definir "tasas de depreciación" más precisas para la información.

No todos los datos ven depreciarse su valor al mismo ritmo o de la misma manera. Esto permite explicar por qué algunas firmas creen necesario conservar la información durante el mayor tiempo posible, aun cuando los reguladores o el público prefieran verla eliminada o anonimizada transcurrido cierto plazo. De ahí que Google siempre se haya resistido a obedecer las peticiones de borrar del todo las direcciones de protocolo de internet de los usuarios en las antiguas búsquedas. (Lo que hace, al cabo de nueve meses, es borrar sólo los dígitos finales para anonimizar casi por completo la consulta. Así, puede seguir comparando los datos de un año para otro, como, por ejemplo, las consultas de compras navideñas, pero sólo en el ámbito regional, no descendiendo al detalle del individuo). Además, conocer la localización de los que buscan algo puede ayudar a mejorar la relevancia de los resultados. Por ejemplo, si muchas personas en Nueva York buscan "Turkey" —y hacen clic en páginas web

relacionadas con el país (Turquía), y no con el ave (pavo)—, el algoritmo clasificará en lugares superiores esas páginas para otras consultas neoyorquinas. Aun cuando el valor de los datos disminuya para algunos de sus propósitos, su valor de opción puede seguir siendo elevado.

EL VALOR DE LOS DESECHOS DIGITALES

La reutilización de datos puede cobrar en ocasiones una forma astuta y disimulada. Las empresas de internet pueden capturar datos sobre todas las cosas que hacen los usuarios, y luego tratar cada interacción en particular como una señal que sirva para personalizar la página web, mejorar un servicio o crear un producto digital enteramente nuevo. Tenemos un ejemplo muy brioso de esto en el siguiente relato sobre dos correctores ortográficos.

A lo largo de veinte años, Microsoft desarrolló un sólido corrector ortográfico para su programa Word. El sistema funcionaba comparando el flujo de caracteres que tecleaba el usuario con un diccionario de términos de grafía correcta y actualización frecuente. El diccionario establecía cuáles eran palabras conocidas, y el sistema trataba aquellas variantes no incluidas en el diccionario como erratas, y procedía a corregirlas. Debido al esfuerzo necesario para recopilar y actualizar el diccionario, el corrector ortográfico[79] de Word sólo estaba disponible para los idiomas más corrientes. A la compañía le costó millones de dólares crearlo y mantenerlo.

Considérese ahora el caso de Google. Podría sostenerse que tiene el corrector ortográfico más completo del mundo, y en prácticamente todas las lenguas vivas. El sistema está siendo continuamente mejorado y enriquecido con nuevas palabras: es el resultado incidental de que la gente use el motor de búsqueda a diario. ¿Teclea uno mal “iPad”? Se incorpora. ¿“Obamacare” [el Plan de Sanidad para Estados Unidos propuesto por Obama]? Otro tanto.

Además, Google obtuvo su corrector ortográfico de forma gratuita al parecer, reutilizando las palabras de ortografía errónea que se introducen en su motor de búsqueda con los tres mil millones de consultas que gestiona a diario. Un inteligente bucle de retroalimentación le explica al sistema exactamente qué palabra habían querido escribir en realidad los usuarios. Estos le “indican” explícitamente la respuesta a Google, cuando este plantea la pregunta en la parte superior de la página de resultados. —¿Quería usted decir “epidemiología”?—, al hacer clic encima para iniciar una nueva búsqueda con el término correcto. O si no, la página web a la que los usuarios son dirigidos les muestra la ortografía correcta, puesto que probablemente tenga una correlación más alta con la palabra de ortografía correcta que con la incorrecta. (Esto es más importante de lo que puede parecer: según iba mejorando el corrector ortográfico de Google, la gente dejó de molestarse en escribir correctamente sus consultas, puesto que aun así Google conseguía procesarlas bien).

El sistema de corrección ortográfica de Google muestra que los datos “malos”, “incorrectos” o “defectuosos” pueden seguir siendo muy útiles. Lo interesante es que Google no fue la primera en tener esta idea. Alrededor del año 2000, Yahoo advirtió la posibilidad de crear un corrector ortográfico a partir de las consultas de usuarios mal escritas. Sin embargo, la idea nunca llegó a abrirse camino. Los datos de antiguas consultas eran en general considerados basura. De igual modo, Infoseek y Alta Vista, anteriores motores de búsqueda populares, tuvieron cada uno, en su día, la base de datos más amplia del mundo de palabras mal escritas, pero no supieron valorarlo. Sus sistemas, en un proceso invisible para los usuarios, trataban las erratas como “términos relacionados” y llevaban a cabo una búsqueda. Pero se basaban en diccionarios que le indicaban explícitamente al sistema qué era lo correcto, no en la suma viva y activa de interacciones de los usuarios.

Sólo Google supo ver que los desechos de las interacciones de los usuarios eran en realidad polvo de oro que podía recogerse y fundirse para formar un lingote. Uno de los principales ingenieros de Google estimó que su corrector ortográfico funcionaba mejor que el de Microsoft en por lo menos un orden de magnitud (aunque, al insistirle, admitió que esto no lo había medido de forma fiable). Y se burló de la idea de que resultase “gratis” desarrollarlo. La materia prima —las palabras mal escritas— podían no haber supuesto ningún coste directo, pero Google había gastado probablemente mucho más que Microsoft para desarrollar el sistema, confesó con una amplia sonrisa.

Los diferentes enfoques de las dos compañías son extremadamente reveladores. Microsoft sólo le veía valor al corrector ortográfico para un único propósito: procesar texto. Google, por su parte, entendía su utilidad más profunda. La empresa no sólo usó las erratas para desarrollar el mejor y más actualizado corrector ortográfico del

mundo con el fin de mejorar las búsquedas, sino que aplicó el sistema a otros muchos servicios, como la función "autocompletar" en las búsquedas, Gmail, Google Docs, y hasta en su sistema de traducción.

Ha surgido un término específico para describir el rastro digital que la gente deja tras de sí: "desechos de datos", esos que surgen como subproducto de las acciones y movimientos de la gente en el mundo. En internet, el término se usa para describir las interacciones online de los usuarios: dónde hacen clic, cuánto tiempo miran una página, dónde permanece más tiempo a la espera el cursor del ratón, qué escriben, etc. Muchas compañías diseñan sus sistemas para recoger desechos de datos y reciclarlos de cara a mejorar un sistema preexistente o desarrollar alguno nuevo. Google, el líder indiscutido, aplica a muchos de sus servicios el principio de "aprender de los datos" de forma recurrente. Cada acción que lleva a cabo un usuario es considerada una señal que ha de ser analizada y vuelta a incorporar al sistema.

Por ejemplo, Google es muy consciente de las muchas veces que la gente buscó un término dado así como otros relacionados, y de cuán a menudo hicieron clic en un enlace, para luego regresar a la página de búsqueda, descontentos con lo que habían encontrado, y lanzar otra búsqueda. Sabe si hicieron clic en el octavo enlace de la primera página, o en el primer enlace de la octava página, o si abandonaron la búsqueda del todo. Puede que la compañía no haya sido la primera en tener esta percepción, pero la llevó a la práctica con extraordinaria efectividad.

Esta información es altamente valiosa. Si muchos usuarios tienden a hacer clic en un resultado de búsqueda al pie de la página de resultados, eso da a entender que es más relevante que los que tiene por encima, y el algoritmo de clasificación de Google sabe que ha de situarlo automáticamente en una posición más elevada para las búsquedas subsiguientes. (Lo mismo hace para los anuncios). "Nos gusta aprender de conjuntos de datos amplios y 'ruidosos'", afirma un miembro de Google.

Los desechos de datos constituyen el mecanismo que subyace a muchos servicios como el reconocimiento de voz, los filtros de *spam*, la traducción de idiomas y demás. Cuando los usuarios le indican a un programa de reconocimiento de voz que ha malinterpretado lo que han dicho, en la práctica están "entrenando" al sistema para que funcione mejor.

Muchos negocios están empezando a programar sus sistemas para recoger y usar la información de esta forma. En los primeros tiempos de Facebook, su primer "científico de datos", Jeff Hammerbacher[80] (que se cuenta entre las personas a las que se atribuye la acuñación del término) examinó la rica acumulación de desechos de datos. Él y su equipo descubrieron que un indicador potente de si la gente acometería una acción determinada (subir contenido, hacer clic sobre un icono, etcétera) era si habían visto a sus amigos hacer eso mismo. Así que Facebook rediseñó su sistema para poner mayor énfasis en hacer más visibles las actividades de los amigos, lo que dio lugar a un círculo virtuoso de nuevas contribuciones a la página web.

La idea se está extendiendo mucho más allá del sector de internet, hasta cualquier empresa que disponga de alguna forma de recoger las opiniones de los usuarios. Los libros electrónicos, por ejemplo, capturan cantidades masivas de datos acerca de las preferencias literarias y hábitos de lectura de las personas que los usan: cuánto tardan en leer una página o sección, dónde leen, si pasan la página tras una ojeada o abandonan la lectura del todo. Las máquinas registran todas las veces que los usuarios subrayan un pasaje o toman notas en los márgenes. La capacidad de recopilar esta clase de información transforma la lectura, que durante tanto tiempo fue un acto solitario, en una suerte de experiencia comunitaria.

Una vez agregados, los desechos de datos pueden decirles a los editores y autores cosas que antes nunca pudieron conocer de forma cuantificable: las preferencias, rechazos y patrones de lectura de la gente. Esta información es valiosa desde un punto de vista comercial: cabe imaginar a los fabricantes de libros electrónicos vendiéndosela a las editoriales para mejorar el contenido y estructura de los libros. Por ejemplo, el análisis de Barnes & Noble sobre los datos procedentes de su libro electrónico Nook[81] reveló que la gente tiende a dejar a la mitad los libros largos de no ficción. Ese descubrimiento le inspiró a la compañía la creación de una serie llamada "Nook Snaps": ensayos breves sobre cosas como la salud o los temas de actualidad.

O considérense si no los programas de educación online como Udacity, Coursera y edX. Estos rastrean las interacciones de los estudiantes en la red para ver qué da mejor resultado desde un punto de vista pedagógico. Las dimensiones del alumnado han sido del orden de decenas de miles de estudiantes, produciendo extraordinarias cantidades de datos. Los profesores pueden ahora comprobar si el porcentaje de estudiantes que han vuelto a ver un pasaje de una conferencia es alto, lo que podría sugerir que tal vez no tenían claro cierto punto. Al dar una clase Coursera sobre el aprendizaje automático el profesor de Stanford Andrew Ng observó que alrededor de dos mil

estudiantes contestaron mal a una pregunta determinada de sus tareas para casa; pero todos presentaron exactamente la misma respuesta incorrecta. Claramente, todos estaban cometiendo el mismo error. Ahora bien, ¿cuál era éste?

Investigando un poco, descubrió que estaban invirtiendo dos ecuaciones algebraicas en un algoritmo. Así que ahora, cuando otros estudiantes cometen ese mismo error, el sistema no sólo les comunica simplemente que están equivocados, sino que les da una pista para que verifiquen sus cálculos. El sistema hace uso de los datos masivos, también, al analizar todos los mensajes de foros que han leído los estudiantes, y si acaban sus tareas de forma correcta, para predecir la probabilidad de que un estudiante que ha leído determinado mensaje obtenga resultados correctos, determinando así qué mensajes de los foros resultan de lectura más útil para los estudiantes. Estas son cosas que eran del todo imposibles de saber hasta ahora, y que podrían cambiar para siempre la forma de enseñar y de aprender.

Los desechos de datos pueden suponer una enorme ventaja competitiva para las empresas, y también pueden transformarse en una poderosa barrera para evitar la entrada de rivales. Piénsese: si una compañía recién establecida desarrollase una página web de comercio electrónico, una red social o un motor de búsqueda que fueran mucho mejores que los de los líderes actuales, como Amazon, Google o Facebook, se vería en dificultades para competir no sólo a causa de las economías de escala y los efectos de red o marca, sino porque una parte tan grande del rendimiento de esas firmas punteras se debe a los desechos de datos que recopilan de las interacciones de los consumidores y luego vuelven a incorporar al servicio. ¿Podría una página web nueva de educación online disponer del *know-how* para competir con el que ya ha desarrollado una monstruosa cantidad de datos que le permiten saber qué funciona mejor?

EL VALOR DE LOS DATOS ABIERTOS

Hoy en día, es probable que pensemos que páginas web como Google y Amazon fueron los pioneros del fenómeno de los datos masivos pero, por descontado, fueron los gobiernos los acopiadores originales de información a escala masiva, y todavía pueden competir con cualquier empresa privada por el mero volumen de datos que controlan. La principal diferencia con los tenedores de información del sector privado estriba en que los gobiernos muchas veces pueden forzar a la gente a suministrarles información, en lugar de tener que persuadirlos u ofrecerles algo a cambio. Por consiguiente, los gobiernos seguirán amontonando vastos tesoros de datos.

Las lecciones de los datos masivos son tan aplicables al sector público como a las entidades comerciales: el valor de los datos gubernamentales es latente y precisa de un análisis innovador para salir a la luz. Pese a su ventajosa situación para capturar datos, los gobiernos a menudo se han mostrado ineficaces a la hora de usarlos. Últimamente ha cobrado relieve la idea de que la mejor manera de extraer valor de los datos gubernamentales es facilitarle el acceso a los mismos al sector privado y a la sociedad en general. Un principio subyace a esto: cuando el estado reúne información, lo hace en nombre de sus ciudadanos, por lo que debería proporcionarle acceso a la sociedad (salvo en el contado número de casos en que con ello podría causar perjuicio a la seguridad nacional o vulnerar el derecho a la intimidad de otros).

Esta idea ha dado lugar a incontables iniciativas a favor de los “datos gubernamentales abiertos” por todo el mundo. Argumentando que los gobiernos son tan sólo custodios de la información que reúnen, y que el sector privado y la sociedad serán más innovadores, los defensores de los datos abiertos elevan peticiones a diversos organismos públicos para que difundan públicamente esos datos con fines tanto cívicos como comerciales. Para que esto resulte operativo, por supuesto, los datos tienen que presentarse en un formato estandarizado e interpretable por las máquinas, de forma que se facilite su procesamiento. De otro modo, la información no tendría de pública más que el nombre.

La idea de los datos gubernamentales abiertos recibió un gran impulso cuando el presidente Barack Obama hizo público, durante su primer día completo en el cargo, el 21 de enero de 2009, un memorando presidencial ordenando a los dirigentes de los organismos federales que divulgaran tanta información como les fuera posible. “En caso de duda, ha de prevalecer la apertura”, instruyó. Fue una declaración admirable, particularmente en comparación con la actitud de su predecesor, quien había dado instrucciones de que se hiciera justamente lo contrario. La orden de Obama[82] propició la creación de la página web data.gov, un repositorio de información del gobierno federal de acceso abierto. Esta página creció rápidamente de 47 bases de datos en 2009 a cerca de 450 000, repartidas por 172 organismos, al cumplir su tercer aniversario en julio de 2012.

Hasta en la reticente Gran Bretaña, donde un montón de información gubernamental está bajo llave por ser los derechos de propiedad intelectual de la Corona, y resulta complicado y oneroso licenciar su uso (como es el caso de los códigos postales para compañías de mapas online), se han dado progresos sustanciales. El gobierno del Reino Unido ha promulgado normas para fomentar la información abierta y ha apoyado la creación de un Instituto de Datos Abiertos, codirigido por Tim Berners-Lee, el inventor de la World Wide Web, para promocionar los usos novedosos de los datos abiertos y las formas de liberarlos de las garras del estado.

La Unión Europea también ha anunciado iniciativas de datos abiertos que pronto podrían llegar a ser de ámbito continental. Otros países, como Australia, Brasil, Chile y Kenia, han creado e implementado estrategias de datos abiertos. Un peldaño más abajo, son cada vez más las ciudades y municipios de todo el mundo que han abrazado la causa de los datos abiertos, igual que algunos organismos internacionales como el Banco Mundial, abriendo al público cientos de conjuntos de datos sobre indicadores económicos y sociales que antes eran de acceso restringido.

Paralelamente, se han ido formando en torno a los datos, para intentar sacarles el mayor partido posible, comunidades de diseñadores de webs y pensadores visionarios como Code for America y la fundación Sunlight en Estados Unidos, y la Open Knowledge Foundation en Gran Bretaña.

Un ejemplo precoz de las posibilidades de la información abierta lo ofrece una página web llamada FlyOnTime.us, donde los visitantes pueden averiguar interactivamente (entre otras muchas correlaciones) qué probabilidad existe de que las inclemencias meteorológicas retrasen los vuelos de un aeropuerto determinado. Esta página web combina información del tiempo y de vuelos a partir de datos oficiales de acceso libre y gratuito por internet. Fue desarrollada por unos defensores de los datos abiertos para demostrar la utilidad de la información que amasaba el gobierno federal. Incluso el *software* de la página web es de código abierto, de modo que otros puedan aprender de él y volver a usarlo.

FlyOnTime.us deja que hablen los datos, y a menudo dice cosas sorprendentes. Puede verse que en el caso de los vuelos de Boston al aeropuerto neoyorquino de LaGuardia, los viajeros han de estar preparados para unos retrasos dos veces más largos cuando hay niebla que cuando nieva. Esto, probablemente, no es lo que la mayoría de la gente se habría imaginado mientras se arremolinaban en el área de embarque: la nieve les habría parecido una razón de más peso para el retraso. Pero éste es el tipo de percepción que el uso de datos masivos hace posible, cuando se procesan datos históricos de retrasos de vuelo de la oficina federal de transporte con información aeroportuaria actual procedente de la administración federal de aviación, junto con los pasados partes meteorológicos oficiales y las condiciones reales según el servicio meteorológico nacional. FlyOnTime.us pone de relieve cómo una entidad que no recopila ni controla flujos de información, como un motor de búsqueda o una gran empresa de distribución, puede, sin embargo, obtener y usar datos para crear valor.

VALORAR LO QUE NO TIENE PRECIO

Tanto abiertos al público como guardados bajo llave en cajas fuertes corporativas, el valor de los datos resulta difícil de medir. Piénsese en los acontecimientos del viernes 18 de mayo de 2012. Ese día, el fundador de Facebook, Mark Zuckerberg, de veintiocho años, hizo sonar, de forma simbólica, la campana que abría la sesión del NASDAQ desde la sede de su compañía en Menlo Park, California. La mayor red social del mundo —que se preciaba de tener como miembro a uno de cada diez habitantes del planeta— empezaba así su nueva vida como sociedad anónima. El valor de las acciones subió inmediatamente un 11 por 100, como ocurre con numerosas empresas tecnológicas nuevas en su primer día de cotización. Sin embargo, algo raro pasó a continuación: las acciones de Facebook empezaron a caer. No ayudó que un problema técnico con los ordenadores de NASDAQ obligase a suspender temporalmente la negociación. Pero había un problema mayor en ciernes y, al detectarlo, los suscriptores del capital, encabezados por Morgan Stanley, apuntalaron de hecho la cotización desfalleciente para que se mantuviera por encima de su precio de salida.

La tarde anterior, los bancos de Facebook habían valorado la empresa a razón de 38 dólares por acción, lo que se tradujo en una valoración global de 104 000 millones. (Por comparación, venía a ser aproximadamente la capitalización de mercado de Boeing, General Motors y Dell Computers juntas). ¿Qué valía en realidad Facebook? En su contabilidad financiera de 2011, ya auditada, con la que los inversores le habían tomado el pulso a la compañía, Facebook declaró activos por valor de 6300 millones de dólares. Eso representaba el valor de sus equipos informáticos, mobiliario de oficina y otros bienes físicos. ¿Y qué hay del valor contable atribuido a las vastas cantidades de información que Facebook tenía en su caja fuerte corporativa? Básicamente, nulo. No se incluía, aun cuando la compañía prácticamente no es *nada más que* datos.

La situación se vuelve aún más rara. Doug Laney, vicepresidente de investigación en Gartner, una empresa de análisis de mercados, procesó los datos durante el periodo previo a la oferta inicial al público (OSP) y concluyó que Facebook había reunido 2,1 billones de unidades de “contenido monetizable” entre 2009 y 2011, como “me gustas”, material escrito en *posts* y comentarios. Comparada con su valoración OSP, eso significa que cada unidad, considerada como un punto de datos individual, tenía un valor de unos cinco centavos. Otra manera de verlo es que cada usuario de Facebook valía alrededor de cien dólares, puesto que los usuarios son la fuente de la información recogida por Facebook.

¿Cómo explicar la vasta divergencia entre el valor de Facebook[83] según las normas contables estándar (6300 millones de dólares) y la valoración inicial que le dio el mercado (104 000 millones de dólares)? No hay forma correcta de hacerlo. Al contrario, lo que existe es un acuerdo muy extendido de que el método actual de determinación del valor corporativo a partir del “valor contable” de la compañía (esto es, básicamente, el valor de sus activos físicos) ya no refleja adecuadamente su valor auténtico. De hecho, el desfase entre el valor contable y el “valor de mercado” —lo que valdría la compañía si fuese adquirida directamente en Bolsa— lleva ampliándose varias décadas. En el año 2000, el Senado de Estados Unidos incluso organizó unas audiencias públicas sobre la modernización de las normas de información financiera, cuyo origen se remontaba a la década de 1930, cuando apenas existían negocios basados en la información. La cuestión afecta a bastante más que el balance de una compañía: la incapacidad de estimar adecuadamente el valor de una empresa podría suscitar riesgos para el negocio y volatilidad en el mercado.

La diferencia entre el valor contable y el valor de mercado de una compañía se contabiliza como “activos intangibles”[84]. En Estados Unidos ha pasado de alrededor del 40 por 100 del valor de las empresas cotizadas en Bolsa, a mediados de la década de 1980, a las tres cuartas partes de su valor, al alba del nuevo milenio. Esta es una divergencia de bulto. Se considera que los activos intangibles comprenden marcas, talento y estrategia: cualquier cosa que no sea física ni forme parte del sistema formal financiero y contable. Sin embargo, los activos intangibles están empezando a referirse cada vez más, también, a los datos que las compañías guardan y utilizan.

En última instancia, lo que esto indica es que hoy por hoy no existe ninguna forma obvia de valorar la información. El día que salieron a Bolsa las acciones de Facebook, la diferencia entre sus activos formales y su valor intangible y no registrado era de casi 100 000 millones de dólares, lo cual es ridículo. Sin embargo, desfases como éste tienen que cerrarse, y lo irán haciendo a medida que las empresas encuentren formas de anotar en sus balances el valor de sus activos en datos.

Se han empezado a dar pasitos en esa dirección. Un alto directivo de uno de los mayores operadores inalámbricos de Estados Unidos admitió que su empresa portadora reconocía el inmenso valor de su información, y estudiaba si tratarla como un activo corporativo en términos contables formales. En cuanto llegó a su conocimiento la iniciativa, sin embargo, los abogados de la compañía la pararon en seco. Registrar los datos en los libros de cuentas podría hacer a la empresa legalmente responsable, argumentaron los hombres de leyes, y eso no les parecía, en absoluto, una buena idea.

Entretanto, los inversores también empezarán a fijarse en el valor de opción de los datos. Puede que aumente mucho el valor de las acciones de las firmas que dispongan de datos o puedan recogerlos con facilidad, en tanto que otras, en posiciones menos afortunadas, acaso vean depreciarse su valor de mercado. No hace falta que aparezcan formalmente los datos en el balance para que esto ocurra. Los mercados y los inversores tendrán en cuenta esos activos intangibles[85] al hacer sus valoraciones, si bien con dificultad, como demuestran los altibajos en la cotización de las acciones de Facebook durante sus primeros meses. Pero, según vayan resolviéndose los dilemas contables y las preocupaciones de responsabilidad legal, es casi seguro que el valor de los datos acabará por aparecer en los balances corporativos, convirtiéndose en una nueva clase de activo.

¿Cómo se valorarán los datos? Calcular su valor ya no significará meramente sumar lo que se gana mediante su uso primario. Pero si la mayor parte del valor de la información está latente y surgirá de unos usos secundarios futuros que hoy aún desconocemos, no resulta obvio cómo se podría proceder a estimarlo. Esto se parece a los obstáculos que había para ponerles precio a los derivados financieros antes del desarrollo de la ecuación de Black-Scholes en la década de 1970, o a la dificultad de valorar las patentes, cuando las subastas, intercambios, ventas privadas, acuerdos de licencia y montones de litigios van creando lentamente un mercado para el conocimiento. Desde luego, colgarle una etiqueta con el precio al valor de opción de los datos representa, como mínimo, una espléndida oportunidad para el sector financiero.

Podríamos empezar por examinar las distintas estrategias que aplican los poseedores de datos para extraerles el valor. La posibilidad más obvia es destinarlos al uso propio de la firma. Resulta improbable, sin embargo, que una compañía sea capaz de descubrir el valor latente de todos los datos. De manera más ambiciosa, por tanto, se podría licenciar a terceros el uso de los datos. En la era de los datos masivos, muchos titulares de datos pueden querer optar por un acuerdo que les suponga un porcentaje del valor extraído de los datos antes que un importe fijo. Así lo hacen las editoriales, que pagan a los autores e intérpretes un porcentaje sobre las ventas de libros, música o películas a título de derechos de autor. También sucede algo similar en el campo de la biotecnología, donde quien licencia su propiedad intelectual puede pedir regalías sobre las invenciones subsiguientes surgidas de su tecnología. De esta forma, todas las partes tienen un incentivo para maximizar el valor obtenido de la reutilización de los datos.

Sin embargo, en la medida en que quien compra una licencia puede no lograr extraer el valor de opción pleno, los titulares de los datos no desearán conceder acceso exclusivo a sus tesoros de información. Antes bien, la "promiscuidad de los datos" puede convertirse en la norma. De esa manera, los poseedores de los datos compensan sus apuestas.

Han aparecido una serie de mercados virtuales para experimentar con formas de ponerles precio a los datos. DataMarket, creada en Islandia en 2008, ofrece acceso a conjuntos de datos gratuitos de otras fuentes, como las Naciones Unidas, el Banco Mundial y Eurostat, y obtiene ingresos revendiendo datos de proveedores comerciales como firmas de investigación de mercados. Otras *start ups* han tratado de convertirse en intermediarias de información, plataformas para que terceras partes compartan sus datos de forma gratuita o a cambio de una remuneración. La idea es permitir que cualquiera venda la información de sus bases de datos, igual que eBay aportó una plataforma para que la gente pueda vender las cosas que tiene en el desván. Import.io anima a las firmas a licenciar sus datos, que de otro modo podrían ser "rebañados" en la red y usados de forma gratuita. Factual, fundada por un antiguo miembro de Google, Gil Elbaz, hace disponibles unos conjuntos de datos que se toma la molestia de

recopilar por su cuenta.

Microsoft ha entrado en liza con Windows Azure Marketplace, aspirando a centrarse en datos de alta calidad y a supervisar qué hay en oferta, igual que Apple controla las ofertas de su tienda de apps. Tal como lo ve Microsoft, una directiva de *marketing* que está trabajando con una hoja de cálculo de Excel puede querer cruzar tablas con datos internos de su empresa con predicciones de crecimiento del PIB de una consultoría económica. Así pues, hace clic en ese mismo momento para comprar esos datos, y estos fluyen instantáneamente a las columnas que tiene en pantalla.

Hasta ahora, no es posible saber cómo resultarán los modelos de valoración. Lo que está claro es que están empezando a crearse economías alrededor de los datos, y que muchos nuevos actores van a beneficiarse, mientras que unos cuantos de los antiguos probablemente consigan una sorprendente prórroga vital. “Los datos son una plataforma”, ha dicho Tim O’Reilly[86], editor de tecnología y voz autorizada de Silicon Valley, ya que se trata de bloques de construcción para fabricar nuevos bienes y modelos de negocio.

Lo esencial del valor de los datos es su potencial de reutilización aparentemente ilimitado: su valor de opción. Recopilar la información resulta crucial, pero no lo suficiente, ya que la mayor parte del valor de los datos se halla en su uso, no en su mera posesión. En el siguiente capítulo, examinaremos cómo se están usando los datos en la práctica, y los grandes negocios de *big data* que están emergiendo.

VII

IMPLICACIONES

En 2011, una avispada *start up* de Seattle llamada Decide.com[87] abrió sus puertas online con ambiciones fantásticamente atrevidas: quería convertirse en un motor de predicción de precios para millones y millones de productos de consumo. Pero había planeado comenzar de forma relativamente modesta: con todos los artilugios tecnológicos posibles, de los teléfonos móviles y televisores de pantalla plana a las cámaras digitales. Sus ordenadores se conectaron a los sistemas de alimentación de datos de las páginas web de comercio electrónico, y empezaron a rastrear la red a la caza de cualquier otra información que pudiera haber sobre precios y productos.

Los precios en internet varían constantemente a lo largo del día, actualizándose de forma dinámica merced a innumerables e intrincados factores. Por lo tanto, la compañía tenía que recoger datos sobre precios en todo momento. No se trataba sólo de datos masivos, sino también de "texto masivo", puesto que el sistema tenía que analizar las palabras para identificar cuándo se iba a dejar de fabricar un producto, o a lanzar un modelo nuevo, informaciones que los consumidores deberían conocer y que afectan a los precios.

Un año más tarde,Decide.com analizaba cuatro millones de productos empleando más de veinticinco mil millones de observaciones de precios. Identificó rarezas de la venta minorista que nadie había sido capaz de "ver" antes, como el hecho de que cuando se introducen modelos nuevos los precios de los antiguos pueden subir de forma temporal. La mayoría de la gente compraría el modelo antiguo pensando que tenía que resultar más barato, pero, dependiendo del momento en que hiciesen clic sobre el botón de "comprar", podría acabar costándoles más. Como las tiendas en la red usan cada vez más a menudo unos sistemas automatizados de fijación de precios,Decide.com puede detectar repuntes algorítmicos antinaturales de los precios y prevenir a los consumidores para que aguarden. Las predicciones de la compañía, según sus mediciones internas, aciertan el 77 por 100 de las veces y ofrecen a los compradores un ahorro potencial de alrededor de cien dólares por producto. Tan segura está la empresa que, en aquellos casos en que sus predicciones resultaren incorrectas,Decide.com reembolsará la diferencia a los usuarios de pago del servicio.

A primera vista,Decide.com se parece a muchas empresas emergentes prometedoras que aspiran a explotar la información de forma nueva y a obtener unos beneficios decentes por su esfuerzo. Lo que hace especial aDecide.com no son los datos: la compañía se basa en información que licencia de páginas web de comercio electrónico y que recopila por la red, donde está gratuitamente a disposición del que la quiera. Tampoco es la capacidad técnica: la empresa no hace nada tan complejo que los únicos ingenieros del mundo capaces de llevarlo a cabo sean los que tiene en sus oficinas. Antes bien, aunque la recolección de datos y las capacidades técnicas sean importantes, la esencia de lo que hace especial aDecide.com es la idea: la compañía tiene "mentalidad de datos masivos". Detectó una oportunidad y se dio cuenta de que la explotación de determinados datos podía revelar secretos valiosos. Y si en Decide.com parecen existir ecos de Farecast, la página web de predicción del precio de los pasajes aéreos, es por algo: las dos son creación de Oren Etzioni.

En el capítulo anterior señalamos que los datos se están convirtiendo en una nueva fuente de valor debido, en gran medida, a lo que hemos llamado valor de opción, cuando se dedica a fines novedosos. El énfasis se ponía en las firmas que recogen los datos. Ahora nos vamos a centrar en las empresas que usan los datos, y en cómo encajan en la cadena de valor de la información. Consideraremos lo que esto significa para las organizaciones y los individuos, tanto por lo que se refiere a sus carreras como a sus vidas cotidianas.

Han surgido tres tipos de compañías de datos masivos, que pueden diferenciarse en función del valor que ofrecen. Podemos pensar en ellas como datos, como capacidades o como ideas.

Primero están los datos. Se trata de las compañías que disponen de los datos o, cuando menos, del acceso a ellos.

Ahora bien, tal vez no estén en el negocio para eso. Dicho de otro modo, no necesariamente disponen de las capacidades correctas para extractar el valor de los datos ni para generar ideas creativas sobre qué vale la pena sacar. El mejor ejemplo lo constituye Twitter, que, obviamente, disfruta de un inmenso flujo de datos pasando por sus servidores, pero se dirigió a dos firmas independientes a la hora de licenciarlo a otros para su uso.

En segundo lugar están las capacidades. A menudo son las consultorías, los vendedores de tecnología y los proveedores de analítica quienes tienen el conocimiento especializado y efectúan el trabajo, pero probablemente no dispongan de los datos, ni del ingenio para asignarles los usos más innovadores. En el caso de Walmart y de las Pop-Tarts, por ejemplo, la empresa recurrió a Teradata, unos expertos en analítica de datos, para que la ayudaran a perfilar sus conclusiones.

En tercer lugar se halla la "mentalidad *big data*". En el caso de ciertas empresas, los datos y el *know-how* no son las razones fundamentales de su éxito. Lo que las hace destacar es que sus fundadores y su personal tienen ideas únicas sobre las formas de explotar los datos, para sacar a la luz nuevas formas de valor. Un buen ejemplo lo brinda Pete Warden, el *geek* cofundador de Jetpac, firma que hace recomendaciones de viajes basándose en las fotos que los usuarios suben a su página web.

Hasta ahora, los dos primeros elementos que hemos descrito son los que han atraído más atención: las capacidades, que hoy son escasas, y los datos, que parecen abundantes. En los últimos años ha surgido una nueva profesión, el "científico de datos", que combina las aptitudes del estadístico, del programador de *software*, del diseñador infográfico y del narrador. En lugar de escudriñar por un microscopio para desvelar algún misterio del universo, el científico de datos escruta bases de datos al acecho de descubrimientos. El McKinsey[88] Global Institute efectúa alarmantes predicciones acerca de la escasez de científicos de datos hoy en día y en el futuro (algo que los científicos de datos actuales gustan de citar para sentirse especiales y de paso revalorizar sus salarios).

Hal Varian[89], economista en jefe de Google, tiene fama de decir que el trabajo de estadístico es el "más *sexy*" que existe. "Si quieres tener éxito, tienes que ser escaso y complementario de algo que resulte ubicuo y barato —afirma—. Los datos están tan ampliamente disponibles, y son tan importantes estratégicamente, que lo que escasea es saber cómo extraer conocimiento de ellos. Por eso los estadísticos, los gestores de bases de datos y los profesionales del aprendizaje de máquinas van a encontrarse en una posición fantástica".

Sin embargo, esta forma de centrarse en las capacidades y quitarle importancia a los datos puede acabarse pronto. Según vaya evolucionando la industria, se superará la escasez de personal a medida que se vuelvan más comunes las capacidades que elogia Varian. Es más, existe la creencia errónea de que sólo porque hay tantos datos alrededor, están a disposición gratuita de cualquiera, o su valor es escaso. En realidad, los datos constituyen el ingrediente crucial. Para apreciar por qué, pensemos en las diferentes partes de la cadena de valor de datos masivos, y cómo es previsible que se modifiquen a lo largo del tiempo. Para empezar, vamos a examinar una por una estas categorías: poseedor de datos, especialista en datos y persona con mentalidad de datos masivos.

LA CADENA DE VALOR DE LOS DATOS MASIVOS

La sustancia primaria de los datos masivos es la propia información, por lo que tiene sentido empezar por estudiar a quienes tienen los datos. Pueden no ser los autores del acopio original, pero son los que controlan el acceso a la información y la usan directamente o la licencian a terceros para que extraigan su valor. Por ejemplo, ITA Software, una gran red de reservas de líneas aéreas (por detrás de Amadeus, Travelport y Sabre), proporcionó datos a Farecast para sus predicciones de tarifas aéreas, pero no llevó a cabo el análisis directamente. ¿Por qué no? Tal como lo veía ITA, su negocio consistía en usar los datos para el propósito para el que había sido creada la empresa —vender billetes de avión—, y no para usos secundarios. Como tal, sus competencias centrales eran diferentes. Es más, habría tenido que buscar la forma de hacer el trabajo sin contar con la patente de Etzioni.

La compañía también decidió no explotar los datos debido a la posición que ocupaba en la cadena de valor de la información. "ITA rehuyó los proyectos que implicaran hacer uso comercial de datos relacionados demasiado estrechamente con los ingresos de las líneas aéreas —rememora Carl de Marcken, cofundador de ITA[90] Software y exdirector general de tecnología de la empresa—. ITA tenía acceso especial a esa clase de datos, necesarios para ofrecer los servicios de la compañía, y no podía permitirse ponerlo en riesgo". Por el contrario, se mantuvo con delicadeza a cierta distancia, licenciando los datos pero no usándolos. En consecuencia, ITA ganó poco dinero; la mayor parte del valor secundario de los datos fue a parar a Farecast: a sus clientes, bajo forma de billetes más baratos; a sus empleados y propietarios, por los ingresos que Farecast obtuvo de publicidad, comisiones y, en último término, por la venta de la compañía.

Algunas empresas se han posicionado astutamente en el centro de los flujos de información para ganar escala y capturar valor de los datos. Ese ha sido el caso del sector de las tarjetas de crédito en Estados Unidos. Durante años, el elevado coste que suponía combatir el fraude llevó a muchos bancos pequeños y medianos a no emitir tarjetas de crédito propias, confiando sus operaciones de tarjeta a instituciones financieras de mayor dimensión, que tenían el tamaño y la escala para invertir en la tecnología necesaria. Firmas como Capital One y el MBNA de Bank of America absorbieron el negocio. Hoy en día, los bancos más pequeños lamentan la decisión, porque el haberse desprendido de las operaciones de tarjetas los priva de unos datos sobre patrones de consumo que les permitirían saber más de sus clientes, y venderles así servicios a medida.

En cambio, los grandes bancos y emisores de tarjetas como Visa y MasterCard parecen encontrarse en el punto óptimo de la cadena de valor de la información. Como prestan servicio a numerosos bancos y comercios, ven pasar más transacciones por sus redes y las usan para sacar conclusiones acerca del comportamiento de los consumidores. Su modelo de negocio evoluciona del mero procesamiento de pagos a la recogida de datos. La cuestión que se les plantea entonces es qué hacer con ellos.

MasterCard podría haber licenciado los datos a terceros para que ellos les sacaran el valor, como hizo ITA, pero la compañía prefiere hacer ella misma el análisis. Una división llamada MasterCard Advisors[91] agrega y analiza 65 000 millones de transacciones de 1500 millones de titulares de tarjetas en doscientos diez países, para definir tendencias de negocio y consumo. Luego vende esa información a otros. Entre otras cosas, descubrió que cuando la gente llena de gasolina el depósito del coche alrededor de las cuatro de la tarde, existe la probabilidad de que, a lo largo de la hora siguiente, gasten de treinta y cinco a cincuenta dólares en una tienda de comestibles o en un restaurante. Un publicista podría hacer uso de esa información para imprimir cupones de oferta de los negocios vecinos al dorso de los recibos de la gasolinera alrededor de esa hora del día.

Como empresa intermediaria de los flujos de información, MasterCard se halla en una posición privilegiada para recopilar datos y capturar su valor. Podemos imaginar un futuro en el que las emisoras de tarjetas de crédito renuncien a sus comisiones sobre las transacciones y las procesen gratuitamente a cambio del acceso a más datos, y perciban ingresos de la venta de analíticas cada vez más sofisticadas basadas en éstos.

La segunda categoría agrupa a los especialistas en datos: compañías con la especialización o la tecnología necesarias para desarrollar análisis complejos. MasterCard eligió hacerlo internamente, y hay empresas que oscilan entre una categoría y otra, pero son muchas las que recurren a especialistas. Por ejemplo, la consultora Accenture trabaja con firmas de muchos sectores para desplegar tecnologías avanzadas de sensores inalámbricos y analizar los datos que recopilan. En un proyecto piloto con la ciudad de St. Louis (Missouri), Accenture instaló sensores inalámbricos en el motor de una veintena de autobuses públicos para monitorizarlos, y predecir averías o determinar el momento óptimo para hacer el mantenimiento. Con ello se redujeron los costes en casi un 10 por 100. Un solo hallazgo —que la ciudad podía diferir la sustitución de una pieza, programada cada 320 000 a 400 000 kilómetros, hasta los 450 000 kilómetros— supuso un ahorro superior a los mil dólares por vehículo. Fue el cliente, y no la consultora, quien se benefició del valor de los datos.

En el ámbito de los datos médicos, tenemos otro ejemplo sorprendente de que las firmas tecnológicas externas pueden ofrecer servicios útiles. El MedStar Washington Hospital Center de Washington, en colaboración con Microsoft Research, y empleando el programa Amalga de Microsoft[92], analizó varios años de expedientes médicos anonimizados —estadísticas vitales del paciente, pruebas médicas, diagnósticos, tratamientos y demás— buscando cómo reducir la tasa de reingresos y la de infecciones. Estos son algunos de los aspectos más costosos de la atención sanitaria, por lo que reducirlos implica un ahorro considerable.

La técnica descubrió algunas correlaciones sorprendentes. Uno de los resultados fue una lista de todas las afecciones que incrementaban las posibilidades de que un paciente dado de alta reingresara antes de un mes. Algunas son bien conocidas y no tienen fácil arreglo. Es probable que vuelva un paciente con un fallo cardíaco congestivo, que es una afección difícil de tratar. Pero el sistema detectó asimismo otro indicador de predicción principal inesperado: el estado mental del paciente. La probabilidad de que una persona fuese ingresada de nuevo antes del mes de haber recibido el alta aumentaba considerablemente si el diagnóstico inicial contenía palabras que sugerían alguna aflicción mental, como “depresión”.

Aunque esta correlación no dice nada que permita establecer una causalidad, sugiere, sin embargo, que una intervención posterior al alta para valorar la salud mental de los pacientes podría mejorar también su salud física, reduciendo el número de reingresos y, por tanto, los costes médicos. Este descubrimiento, extraído por un ordenador tras analizar un vasto acopio de datos, es algo que una persona que estudiara los datos podría no haber advertido nunca. Microsoft no controlaba los datos, que eran propiedad del hospital. Y no tuvo ninguna idea asombrosa; no era lo que se necesitaba ahí. En cambio, aportó la herramienta de *software*, el programa Amalga, para detectar la idea.

Las firmas que poseen datos masivos recurren a especialistas para extraer valor de esos datos. Pero, a pesar de los grandes elogios y de darles a los puestos de trabajo unos nombres de tanto relumbrón como “Ninja de datos”, la vida de los expertos técnicos no siempre es tan glamourosa como puede parecer. Trabajan en las minas de diamantes de los datos masivos, y cobran unas nóminas muy satisfactorias, pero le entregan las joyas que extraen a quienes poseen esos datos.

El tercer grupo está compuesto por las compañías y los individuos con mentalidad de datos masivos. En el caso de este grupo, la fuerza radica en que pueden ver las oportunidades antes que los demás, aun cuando carezcan de los datos o de las aptitudes para actuar en función de esas oportunidades. De hecho, tal vez sea porque, como proceden de ámbitos diferentes, sus mentes están libres de barreras imaginarias: pueden ver lo que es posible, en vez de dejarse limitar por el sentido de lo que es factible.

Bradford Cross[93] personifica lo que supone tener una mente de datos masivos. En agosto de 2009, cuando tenía veintipocos años, Cross y otros amigos fundaronFlightCaster.com. Igual que FlyOnTime.us, FlightCaster predecía si un vuelo en el interior de Estados Unidos tenía probabilidades de sufrir retrasos. Para hacer las predicciones, analizaba todos los vuelos de los diez años anteriores, contrastados con datos meteorológicos históricos y contemporáneos.

Lo interesante es que los propios titulares de los datos no podían hacerlo. Ninguno tenía el incentivo —o el mandato regulatorio— para usar los datos de esta manera. De hecho, si las fuentes de los datos —la oficina estadounidense de estadísticas de transporte, la administración federal de la aviación y el servicio nacional de meteorologíase hubiesen atrevido a predecir retrasos en los vuelos comerciales, el Congreso probablemente habría

celebrado sesiones y habría rodado alguna que otra cabeza funcionarial. Y las líneas aéreas no podían hacerlo… o no querían. Se benefician de mantener su mediocre actuación lo más a oscuras posible. En cambio, lograrlo requirió un puñado de ingenieros vestidos con sudaderas. De hecho, las predicciones de FlightCaster eran tan asombrosamente precisas que incluso los empleados de las líneas aéreas empezaron a usarlas: las líneas aéreas no quieren anunciar los retrasos hasta el último minuto, y de ahí que, aunque sean la fuente de información decisiva, nunca sean la primera.

Debido a su mentalidad de datos masivos —su inspirada percepción de que podrían procesarse datos públicos para arrojar respuestas que interesaban a millones de personas—, la empresa FlightCaster de Cross fue una precursora, pero por los pelos. El mismo mes en que se lanzó FlightCaster, los *geeks* de FlyOnTime.us empezaron a reunir datos abiertos para construir su página web. La ventaja de la que disfrutaba FlightCaster pronto se vería recortada. En enero de 2011, Cross y sus socios vendieron la firma a Next Jump, una compañía que gestiona programas de descuentos corporativos usando técnicas de datos masivos.

Acto seguido, Cross puso sus miras en otra industria ya madura, en la que advirtió un nicho donde podría penetrar un innovador externo: la de los medios informativos. Su *start up* Prismatic agrega y clasifica contenidos recogidos por toda la red basándose en el análisis textual, las preferencias del usuario, la popularidad en las redes sociales y la analítica de datos masivos. Lo curioso es que el sistema no establece gran diferencia entre una publicación en el blog de un adolescente, una web de empresa y un artículo en *The Washington Post*: si el contenido es juzgado relevante y popular (en términos de cuántas veces se ve y cuántas se comparte), aparece en la parte de arriba de la pantalla.

Como servicio, Prismatic supone un reconocimiento de la forma en que interactúa con los medios la generación más joven. Para ellos, la fuente de información ha perdido su importancia primordial. Esto supone un humilde aviso a los *popes* de los medios de comunicación generalistas de que el público en conjunto es más entendido que ellos, y que los periodistas con traje de chaqueta tienen que competir con los blogueros en chándal. Sin embargo, el punto crucial es que resulta difícil pensar que Prismatic hubiese podido surgir en el seno de la propia industria de los medios de comunicación, aun cuando recopila montones de información. A los habituales del bar del Club Nacional de la Prensa nunca se les ocurrió reutilizar los datos online acerca del consumo de medios. Tampoco los especialistas en analítica de Armonk (Nueva York), ni los de Bangalore (India), habrían aprovechado la información de esta manera. Tuvo que presentarse Cross, un desconocido de dudosa fama, con los pelos revueltos y voz perezosa, para presumir que, mediante el uso de datos, se le podría decir al mundo que algo merecía su atención más que los editoriales de *The New York Times*.

El concepto de mentalidad de datos masivos y el papel del forastero creativo con una idea brillante no se alejan mucho de lo que sucedió al alba del comercio electrónico, a mediados de la década de 1990, cuando los pioneros no se veían lastrados por el pensamiento atrincherado ni por las restricciones institucionales de las industrias más viejas. Así, fue un cerebrito de un fondo de inversión, y no Barnes & Noble, quien fundó una librería online (Jeff Bezos, de Amazon); un desarrollador de *software*, y no Sotheby's, quien creó una página web de subastas en la red (Pierre Omidyar, de eBay). Hoy en día, los emprendedores con mentalidad de datos masivos muchas veces no disponen de los datos cuando empiezan. Ahora bien, por eso mismo tampoco tienen intereses creados ni desincentivos financieros que pudieran impedirles llevar sus ideas a la práctica.

Como hemos visto, hay casos en los que una firma combina muchas de estas características de datos masivos. Puede que a Etzioni y Cross se les ocurriesen sus ideas rompedoras antes que a otros, pero es que, además, disponían de las aptitudes precisas. Los trabajadores de Teradata y Accenture no se limitan a fichar en el reloj de entrada; también se les reconoce que tienen una gran idea de cuando en cuando. Con todo, los estereotipos resultan útiles para apreciar los papeles que interpretan las distintas firmas. Los pioneros de los datos masivos de hoy suelen tener formaciones y experiencias muy diversas y aplican sus capacidades en una amplia variedad de áreas. Está emergiendo una nueva generación de ángeles inversores y emprendedores, en particular entre antiguos miembros de Google y la así llamada Mafia de PayPal (los antiguos directores de la firma, como Peter Thiel, Reid Hoffman y Max Levchin). Estas personas, junto con un puñado de expertos en informática del ámbito académico, son algunos de los principales apoyos de las actuales *start ups* impregnadas de datos.

La visión creativa de individuos y empresas en la cadena de suministro de los datos masivos nos ayuda a estimar de nuevo el valor de las compañías. Por ejemplo,Salesforce.com puede no ser simplemente una plataforma

conveniente para que las firmas alojen sus aplicaciones corporativas: está asimismo bien situada para sacar a la luz el valor de los datos que fluyen por la parte superior de su infraestructura. Las compañías de telefonía móvil, como vimos en el capítulo anterior, recopilan una monstruosa cantidad de datos, pero muchas veces están cegadas culturalmente a su valor. Podrían, sin embargo, licenciarlos a terceros capaces de extraerles valor nuevo; igual que Twitter decidió ceder los derechos sobre sus datos a dos compañías externas.

Algunas empresas afortunadas se sitúan conscientemente a caballo de los dos mundos: Google recopila datos, como las erratas en las búsquedas de información, tiene la idea brillante de usarlos para crear el que probablemente sea el mejor corrector ortográfico del mundo, y dispone en su propia organización del capital humano necesario para llevar la idea a la práctica. En muchas de sus restantes actividades, Google también se beneficia de la integración vertical en la cadena de valor de los datos masivos, donde ocupa simultáneamente las tres posiciones posibles. Al mismo tiempo, Google pone además a disposición de terceros parte de sus datos mediante interfaces de programación de aplicaciones (API, según sus siglas en inglés), para que puedan ser reutilizados y generen más valor. Un ejemplo lo tenemos en los mapas de Google, que todo el mundo usa gratuitamente a través de la red, desde agencias inmobiliarias a páginas web gubernamentales (si bien las páginas web con visitas muy elevadas tienen que pagar).

También Amazon dispone de la mentalidad, la capacidad técnica y los datos. De hecho, la compañía perfiló su modelo de negocio siguiendo precisamente ese orden, a la inversa de la norma. Inicialmente sólo tenía la idea de su célebre sistema de recomendaciones. En 1997, el prospecto bursátil de la empresa describía el "filtrado colaborativo"[94] antes de que Amazon supiera cómo iba a funcionar en la práctica, o dispusiera de suficientes datos para hacerlo útil.

Tanto Google como Amazon abarcan todas las categorías, pero sus estrategias difieren. Cuando Google se lanza a recoger todo tipo de datos, ya tiene en mente unos usos secundarios. Sus coches Street View, como hemos visto, recopilaban información de los GPS no sólo para su servicio cartográfico, sino también para programar coches autoconducidos. Amazon, sin embargo, se centra más en el uso primario de la información y sólo explota los usos secundarios como una ganancia extra. Su sistema de recomendación, por ejemplo, se basa en los datos de los flujos de clics, pero no ha usado esa información para hacer cosas tan extraordinarias como predecir el estado de la economía o los brotes de gripe.

A pesar de que el lector electrónico Kindle de Amazon es capaz de mostrar si una página determinada ha sido anotada y subrayada a menudo por los usuarios, la firma no vende esa información a los autores ni a las editoriales. A los comerciales les encantaría saber cuáles son los pasajes más populares, para vender mejor los libros. A los autores les podría gustar saber en qué punto de sus excelsos tomos arrojan la toalla la mayoría de los lectores, y mejorar así su obra. Las editoriales podrían detectar temas que presagien el siguiente gran éxito. Pero Amazon parece dejar en barbecho ese campo de datos.

Explotados con astucia, los datos masivos pueden transformar los modelos de negocio de las empresas y las formas de interacción ya asentadas. Un caso curioso es el de un gran fabricante automovilístico europeo, que redefinió su relación comercial con un proveedor de piezas aprovechando datos de uso de los que el fabricante carecía. (Dado que este ejemplo ha llegado a nuestro conocimiento de forma extraoficial, a través de una de las principales empresas que procesaron los datos, lamentamos no poder revelar los nombres de las compañías implicadas).

Hoy en día, los coches están trufados de chips, sensores y *software* que suben datos a los ordenadores del fabricante cuando el vehículo pasa la revisión. Un coche típico de gama media tiene hoy unos cuarenta microprocesadores;[95] toda la electrónica de un coche representa una tercera parte de su coste. Esto hace de los automóviles unos dignos sucesores de los barcos que Maury llamaba "observatorios flotantes". La capacidad de recoger datos sobre cómo responden en la práctica los componentes de un coche en la carretera —y de reincorporar esos datos para luego mejorarlos— está resultando ser una gran ventaja competitiva para las firmas que consiguen esa información.

Trabajando con una compañía analítica independiente, el fabricante de automóviles fue capaz de detectar que el sensor de nivel de combustible producido por un proveedor alemán estaba funcionando fatal: registraba una veintena de alertas falsas por cada una válida. La compañía podía haber facilitado esa información al proveedor, exigiéndole el ajuste. En una era de negocios más caballerosa, probablemente lo hubiera hecho. Ahora bien, el fabricante llevaba una fortuna invertida en su programa analítico, y quiso usar esta información para recuperar parte de su inversión.

La compañía sopesó sus opciones. ¿Debería vender los datos? ¿Cómo se valoraría la información? ¿Y si el proveedor se plantaba y el fabricante se quedaba colgado con una pieza que funcionaba mal? Sabía que, si entregaba la información, también mejorarían las piezas de ese tipo que llevaban los vehículos de la competencia. Asegurarse de que la mejora beneficiara sólo a los vehículos propios pareció una jugada más astuta. Al final, el fabricante automovilístico dio con una idea novedosa: encontró la forma de mejorar la pieza mediante un *software* modificado, patentó la técnica, vendió luego la patente al proveedor… y se embolsó una buena cantidad con todo el proceso.

LOS NUEVOS INTERMEDIARIOS DE DATOS

¿Quién dispone de más valor en la cadena de valor de los datos masivos? Hoy diríamos que quien posee la mentalidad, las ideas innovadoras. Como se vio en la era de las punto com, aquellos que se anticipan pueden prosperar de verdad, pero esa ventaja quizá no dure mucho tiempo. Según vaya avanzando la época de los datos masivos, otros irán adoptando esa mentalidad y la ventaja de los pioneros disminuirá en términos relativos.

Entonces, ¿no es posible que el núcleo del valor se halle realmente en las capacidades? Al fin y al cabo, una mina de oro no vale nada si no se puede extraer el mineral. Sin embargo, la historia de la informática apunta en sentido contrario. Hoy en día, las especializaciones en gestión de bases de datos, ciencia de datos, analítica, algoritmos de aprendizaje de máquinas y cosas similares están muy demandadas. Pero a lo largo del tiempo, conforme los datos masivos se vayan convirtiendo cada vez más en parte de la vida diaria, según vayan perfeccionándose los instrumentos y se vuelvan más fáciles de usar, y más gente adquiera los conocimientos, el valor de las aptitudes también disminuirá en términos relativos. Así sucedió con la capacidad de programar ordenadores, que se volvió más común entre las décadas de 1970 y 1980. Actualmente, las firmas de externalización *offshore* han reducido aún más el valor de la programación; lo que en tiempos era el parangón de la inteligencia técnica ahora no es más que un mecanismo de desarrollo para los pobres del mundo. Con esto no queremos decir que la formación en datos masivos carezca de importancia. Pero no es la fuente de valor más crucial, puesto que uno puede importarla del exterior.

Hoy por hoy, en las primeras fases de la era de los datos masivos, las ideas y las capacidades parecen atesorar todo el valor, pero a la larga lo más valioso serán los propios datos: podremos hacer más cosas con la información, y además los dueños de los datos sabrán apreciar mejor el valor potencial de ese activo que poseen. En consecuencia, probablemente se aferren a él con más fuerza que nunca, y vendan muy caro el acceso a terceros. Por seguir con la metáfora de la mina de oro: lo que más importará será el oro en sí.

Sin embargo, hay una dimensión importante, digna de destacarse, en el ascenso a largo plazo de los dueños de datos. En algunos casos, surgirán “intermediarios de datos” capaces de recopilar datos de múltiples fuentes, agregarlos y hacer cosas innovadoras con ellos, y los dueños de los datos se lo permitirán porque sólo así podrá extraerse todo el valor de los datos.

Un ejemplo nos lo ofrece Inrix[96], una firma de análisis de tráfico con base en las afueras de Seattle. Inrix recopila datos de geolocalización en tiempo real de cien millones de vehículos en Europa y Norteamérica: automóviles de BMW, Ford y Toyota, entre otros, así como flotillas comerciales como las de taxis y furgonetas de reparto. Obtiene asimismo datos de los móviles de los conductores (sus apps gratuitas para teléfonos inteligentes tienen su importancia aquí: los usuarios reciben información sobre el tráfico, y a cambio Inrix capta sus coordenadas). Luego combina esta información con datos sobre los patrones históricos del tráfico, el clima y otros factores locales, para predecir la fluidez de la circulación. El resultado se transmite desde su cadena de montaje de datos a los sistemas de navegación de los vehículos, y es utilizado por flotillas oficiales y comerciales.

Inrix supone la quintaesencia del intermediario de datos independiente. Recoge información de numerosas empresas automovilísticas rivales y genera así un producto más valioso del que ninguna de ellas podría haber desarrollado por su cuenta. Puede que cada fabricante de coches disponga de unos cuantos millones de puntos de datos gracias a sus vehículos en la carretera. Aunque podría usar esos datos para predecir flujos de tráfico, esas predicciones no resultarían ni muy exactas ni completas, porque la calidad mejora según aumenta la cantidad de datos. Además, las firmas automovilísticas pueden no disponer de la capacidad: sus competencias se hallan fundamentalmente en el trabajo del metal, no en la ponderación de distribuciones de Poisson. Así pues, todos se ven incentivados para dirigirse a una tercera parte que realice el trabajo. Por otro lado, aunque la predicción del tráfico sea importante para los conductores, apenas influye en que un particular adquiera o no un coche determinado. Así que a las empresas competidoras no les importa aunar sus fuerzas de este modo.

Por descontado, son muchas las empresas de diferentes sectores que han compartido informaciones en el pasado, particularmente las aseguradoras y los sectores en red como la banca, la energía y las telecomunicaciones, donde el intercambio de información resulta crucial para evitar problemas, y los reguladores lo requieren en ocasiones. Las empresas de estudios de mercado han agregado datos industriales durante décadas, igual que otras dedicadas a tareas especializadas como el control de la circulación de la prensa. Para algunas asociaciones de comercio, se trata de una tarea fundamental.

Ahora, la diferencia estriba en que los datos son una materia prima que se introduce en el mercado: un activo independiente de lo que había intentado medir previamente. Por ejemplo, la información de Inrix es más útil de lo que podría parecer a primera vista. Su análisis de tráfico se emplea para medir la salud de las economías locales, porque puede brindar pistas acerca del desempleo, las ventas en tiendas y las actividades de ocio. Cuando la recuperación económica de Estados Unidos empezó a perder fuerza en 2011, pese a los desmentidos políticos de que tal cosa estuviera produciéndose, los signos fueron aparentes en el análisis del tráfico: las horas punta se habían vuelto menos tumultuosas, lo que apuntaba a un mayor desempleo. Además, Inrix ha vendido sus datos a un fondo de inversiones que usa los patrones de tráfico alrededor de la tienda de una marca importante como aproximación a sus ventas, que el fondo usa para negociar las acciones de la compañía antes de hacer público el anuncio de sus ganancias trimestrales. La abundancia de vehículos en la zona se correlaciona con mejores ventas.

Están apareciendo otros intermediarios similares en la cadena de valor de datos masivos. Uno de los primeros fue la firma Hitwise, posteriormente adquirida por Experian, que alcanzó una serie de acuerdos con proveedores de servicios de internet para recopilar la información de sus flujos de clics a cambio de unos ingresos extra. Los datos se licenciaron por un pequeño importe fijo, en lugar de un porcentaje del valor que producían. Hitwise, en tanto que intermediario, se quedó con la mayor parte del valor. Otro ejemplo lo constituye Quantcast, que mide el tráfico online de páginas web para ayudarlos a saber más sobre el perfil demográfico y los patrones de uso de sus visitantes. Esta empresa regala un instrumento online para que las páginas web puedan seguirles la pista a sus visitantes; a cambio, Quantcast puede ver los datos, lo que le permite mejorar el enfoque de su publicidad.

Estos nuevos intermediarios han identificado nichos de mercado lucrativos sin suponer una amenaza para los modelos de negocio de los dueños de los datos que les brindan la información. Por el momento, la publicidad en internet es uno de esos nichos, ya que es ahí donde se halla la mayor parte de los datos, y donde existe la necesidad acuciante de extraerlos para enfocar con precisión los anuncios. Pero conforme el mundo se vaya datificando más y haya más sectores que se den cuenta de que su negocio principal es aprender de los datos, estos intermediarios independientes de la información aparecerán igualmente en otros lugares.

Algunos de los intermediarios pueden no ser empresas comerciales, sino entidades sin ánimo de lucro. Por ejemplo, el Health Care Cost Institute[97] [instituto del coste de la atención sanitaria] fue creado en 2012 por algunas de las principales aseguradoras sanitarias de Estados Unidos. Sus datos agregados ascendían a cinco mil millones de demandas (anonimizadas) que implicaban a treinta y tres millones de personas. Compartir los archivos permitió a las firmas detectar tendencias que podrían no haber sido capaces de advertir en sus conjuntos de datos individuales, más pequeños. Entre los primeros descubrimientos estaba que los costes médicos en Estados Unidos habían crecido tres veces más deprisa que la inflación entre 2009 y 2010, pero con diferencias pronunciadas si se analizaba de cerca: los precios de urgencias habían crecido un 11 por 100, mientras que los de las residencias de ancianos de hecho habían disminuido. Claramente, las aseguradoras médicas jamás habrían facilitado sus preciadas informaciones a nadie que no fuese un intermediario sin ánimo de lucro. Las motivaciones de una entidad sin ánimo de lucro resultan menos sospechosas, y la organización puede diseñarse pensando desde el principio en su transparencia y en la responsabilidad de rendir cuentas.

La diversidad de firmas de datos masivos muestra cómo está cambiando el valor de la información. En el caso deDecide.com[98], los datos de precios son suministrados por una serie de páginas web con las que comparten los ingresos.Decide.com percibe comisiones cuando la gente compra bienes a través de la página web, pero las compañías que suministraron los datos también se llevan una parte del negocio. Esto indica que ha madurado algo la forma de trabajar con los datos: ITA no recibió ninguna comisión por los datos que facilitó a Farecast, sólo un canon básico por la licencia. Ahora, los proveedores de datos están en condiciones de negociar acuerdos más atractivos. Por lo que se refiere a la próxima *start up* de Etzioni, es de presumir que tratará de suministrar los datos él mismo,

dado que el valor ha migrado de la capacidad a la idea, y está desplazándose ahora a los datos mismos.

Los modelos de negocio están viéndose trastocados con ese desplazamiento del valor hacia quienes controlan los datos. Aquel fabricante automovilístico europeo que cerró el acuerdo de propiedad intelectual con su proveedor contaba con un potente equipo interno de análisis de datos, pero tuvo necesidad de trabajar con un vendedor de tecnología externo a la empresa para descubrir qué le decían los datos. La firma de tecnología cobró su trabajo, pero el fabricante de coches se quedó con la mayor parte de los beneficios. Al detectar la oportunidad, sin embargo, la compañía tecnológica ha retocado su modelo de negocio para compartir parte del riesgo y de las recompensas con los clientes. Ha probado a trabajar por menos remuneración, a cambio de compartir parte de la riqueza que su análisis saca a la luz. (En cuanto a los fabricantes de componentes para la industria del automóvil, probablemente resulte seguro afirmar que en el futuro todos querrán incorporar sensores de medición a sus productos, o insistir en tener acceso a la información de rendimiento como parte estándar del contrato de venta, para poder mejorar continuamente sus componentes).

En cuanto a los intermediarios, sus vidas son complicadas, porque necesitan convencer a las compañías de que compartir resulta rentable. Por ejemplo, Inrix ha empezado a reunir otras informaciones además de las de geolocalización. En 2012, llevó a cabo una prueba para analizar cuándo y dónde se disparan los sistemas de frenado automático de los coches (ABS), por encargo de un fabricante de automóviles que había diseñado su sistema de telemetría para recoger la información en tiempo real. La idea consiste en que si salta con frecuencia el ABS[99] en un determinado trecho de la carretera quizá sea porque las condiciones en ese tramo son peligrosas, y los conductores deban contemplar rutas alternativas. Con estos datos, Inrix podía recomendar no sólo el camino más corto, sino también el menos peligroso.

Sin embargo, el fabricante de coches no tiene previsto compartir esta información con terceros. Al contrario, insiste para que Inrix instale el sistema exclusivamente en sus coches. Considera que el valor de proclamar la existencia de esa prestación supera con creces la ganancia derivada de agregar esos datos con los de otros fabricantes para mejorar la fiabilidad global del sistema. Inrix, por su parte, cree que con el tiempo todos los fabricantes de automóviles verán la utilidad de agregar todos sus datos. En tanto que intermediario de datos, Inrix tiene un incentivo poderoso para aferrarse a ese optimismo: su negocio se levanta por completo sobre el acceso a múltiples fuentes de datos.

Las compañías están experimentando también con distintas formas de organización en el negocio de los datos masivos. Inrix no dio por casualidad con su modelo de negocio, como suele pasarles a muchas *start ups*, sino que estableció su papel de intermediario intencionadamente. Microsoft, propietaria de las patentes esenciales de la tecnología, pensó que una firma pequeña e independiente —antes que una gran empresa conocida por sus tácticas agresivas— podría ser percibida como algo más neutral y atraer a otros rivales del sector para sacarle así el máximo partido a su propiedad intelectual. Del mismo modo, el MedStar Washington Hospital Center que usó el *software* Amalga de Microsoft para analizar los reingresos de pacientes sabía exactamente lo que hacía con sus datos: el sistema Amalga era en origen el *software* propio del servicio de urgencias del hospital, llamado Azyxxi, que le vendió en 2006 a Microsoft para que lo desarrollara mejor.

En 2010, UPS vendió una unidad interna de análisis de datos, llamada UPS Logistics Technologies, a la sociedad anónima privada Thoma Bravo. Operando ahora como Roadnet Technologies, la unidad tiene más libertad para efectuar análisis de ruta para más de una compañía. Roadnet recopila datos de muchos clientes y brinda un servicio de fijación de precios que emplean UPS y sus competidores por igual. Llamándose UPS Logistics, jamás habría logrado persuadir a las firmas rivales para que compartieran sus conjuntos de datos, explica el director general de Roadnet, Len Kennedy.[100] Ahora bien, cuando se hizo independiente, los competidores de UPS se sintieron más cómodos suministrándole sus datos, y, en último término, todos se beneficiaron de las mejoras en la exactitud que se consiguieron con la agregación.

La prueba de que son los datos mismos, antes que las capacidades o la mentalidad, los que acabarán convirtiéndose en lo más valorado puede hallarse en las numerosas adquisiciones que se han dado en el negocio de los datos masivos. Por ejemplo, en 2006 Microsoft premió la mentalidad *big data* de Etzioni al comprarle Farecast por alrededor de ciento diez millones de dólares. Y, dos años más tarde, Google pagó setecientos millones para adquirir al proveedor de datos de Farecast, ITA Software.

LA DESAPARICIÓN DEL EXPERTO

En la película *Moneyball*, que cuenta cómo los Athletics de Oakland se convirtieron en un equipo de béisbol ganador al aplicarle al juego analítica y nuevos tipos de indicadores, hay una escena deliciosa en la que unos viejos ojeadores canosos, sentados alrededor de una mesa, comentan a los jugadores. El espectador no puede evitar sentir vergüenza ajena, no sólo porque la escena expone la forma en que se adoptan decisiones sin datos en que apoyarlas, sino porque todos nos hemos hallado en situaciones donde la "certeza" se basaba en los sentimientos más que en la ciencia.

—Tiene físico de béisbol… buena cara —dice uno de los ojeadores.

—Tiene un *swing* precioso. Cuando consigue atizarle, impulsa la pelota, la hace salir despedida del bate —interviene un tipo frágil de pelo gris que lleva un audífono.

—Mucho impulso en el bate —abunda otro ojeador.

Un tercer hombre interrumpe bruscamente la conversación declarando:

—Tiene una novia fea.

—¿Y eso qué significa? —pregunta el ojeador que dirige la reunión.

—Una novia fea significa falta de seguridad en sí mismo —explica con total naturalidad el disidente.

—Vale —dice el jefe, satisfecho con la respuesta.

Después de una animada discusión, un ojeador que hasta entonces ha guardado silencio toma la palabra:

—Este tío tiene actitud, y eso es bueno. Quiero decir, que es de esa clase de tíos que entra en una habitación con la polla por delante.

Y otro añade:

—El examen visual lo aprueba. Tiene la pinta que hay que tener, está listo para representar el papel. Sólo necesita algo de tiempo de juego.

—Sigo diciendo —reitera el que lleva la contraria— que a su novia, como mucho, yo le pongo un seis.

La escena ilustra a la perfección las carencias del juicio humano. Lo que se presenta como un debate razonado no se basa en realidad en nada concreto. Las decisiones sobre unos contratos de jugadores valorados en millones de dólares se toman por instinto puro y visceral, sin apoyarse en datos objetivos. Sí, no es más que una película, pero la vida real no es demasiado diferente. La misma clase de razonamientos hueros se emplea en todo tipo de lugares: desde las salas de juntas de Manhattan hasta el Despacho Oval, pasando por cafeterías y cocinas de todo el mundo.

Moneyball, basada en el libro de Michael Lewis, refiere la historia real de Billy Beane, el director general de los Athletics de Oakland, que prescindió de las reglas tradicionales de valoración de los jugadores a favor de un método de fuerte carga matemática que contemplaba el juego con un conjunto de indicadores nuevos. Se acabaron las medidas estadísticas consagradas por el tiempo como el "promedio de bateo" y se introdujeron unas formas de pensar acerca del juego aparentemente extrañas, como el "porcentaje de colocación en la base". Este enfoque inspirado en los datos sacó a la luz una dimensión del deporte que siempre había estado presente, aunque oculta entre los cacahuetes y las palomitas. No importaba cómo llegara a la base el jugador, arrojándose al suelo o caminando tranquilamente, mientras lo consiguiese. Cuando los datos mostraron que los robos de bases eran ineficientes, se acabó uno de los elementos más excitantes, pero menos "productivos" del juego.

Envuelto en una controversia considerable, Beane consagró en la dirección del equipo el método conocido como *sabermétrica*, término acuñado por el periodista deportivo Bill James en referencia a la Society for American Baseball Research, que hasta entonces había sido coto privado de una subcultura alternativa. Beane estaba poniendo en cuestión el dogma del banquillo, igual que las ideas heliocéntricas de Galileo afrentaron en su momento a la autoridad de la iglesia católica. En última instancia, llevó al sufrido equipo a ganar en 2002 la American League West, con una racha de veinte victorias consecutivas. Desde entonces, los estadísticos sustituyeron a los ojeadores como sabios del deporte. Y muchos otros equipos se apresuraron a adoptar también la *sabermétrica*.

Siguiendo ese espíritu, el mayor impacto de los datos masivos consistirá en que las decisiones fundamentadas en datos probablemente están llamadas a incrementarse o a invalidar el juicio humano. En su libro *Super Crunchers*, Ian Ayers, economista y profesor de derecho en Yale, argumentaba que los análisis estadísticos fuerzan a las personas a pensar dos veces lo que les dice el instinto. A través de los datos masivos, esto se vuelve todavía más esencial. El experto en un área temática, el especialista sustantivo, perderá parte de su lustre en beneficio del estadístico y el analista de datos, quienes no se ven lastrados por las viejas formas de hacer las cosas y dejan que los datos hablen. Este nuevo marco se basará en unas correlaciones libres de juicios previos y prejuicios, de la misma forma que Maury no dio por buenas sin más las historias que los viejos capitanes le contaban acerca de un pasaje determinado mientras se tomaban una pinta en el *pub*, sino que confió en que los datos agregados revelaran verdades prácticas.

Estamos asistiendo a la desaparición en muchas áreas de la influencia de los expertos temáticos. En el ámbito de los medios de comunicación, el contenido que se crea y publicita en páginas web como The Huffington Post, Gawker y Forbes se ve determinado por los datos, no sólo por la opinión de los editores humanos. Los datos pueden revelar qué quiere leer la gente mejor que el instinto de los periodistas veteranos. La firma de educación online Coursera emplea datos acerca de qué secciones de una conferencia en vídeo vuelven a ver los alumnos para aprender qué pasaje puede no haber quedado claro, y retroalimenta la información a los profesores para que puedan explicar mejor. Como se ha comentado anteriormente, Jeff Bezos se deshizo de los autores de reseñas literarias de la propia empresa cuando los datos demostraron que las recomendaciones algorítmicas propiciaban más ventas.

Esto significa que las aptitudes necesarias para tener éxito en el trabajo están cambiando. Modifica lo que se supone que han de aportar los trabajadores a su empresa. La doctora McGregor, para cuidar de bebés prematuros en Ontario, no necesita ser la médica más lista del hospital, ni la principal autoridad del mundo en atención neonatal. De hecho, ni siquiera es médico: es doctora en informática. Pero se vale de datos que representan más de una década de años/paciente[101], que el ordenador procesa y ella convierte en recomendaciones de tratamiento.

Como hemos visto, es frecuente que los pioneros de los datos masivos provengan de campos ajenos al sector en el que acaban dejando su impronta. Se trata de especialistas en análisis de datos, inteligencia artificial, matemáticas o estadística, y aplican esas aptitudes a sectores específicos. Los ganadores de los concursos Kaggle, la plataforma virtual para proyectos *big data*, suelen ser novicios en el sector en el que triunfan, explica el director general de Kaggle, Anthony Goldbloom. Un físico británico desarrolló unos algoritmos casi imbatibles para predecir reclamaciones al seguro e identificar coches usados defectuosos. Un analista de seguros de Singapur encabezó una competición para predecir respuestas biológicas a compuestos químicos. Mientras tanto, en el grupo de traducción automática de Google, los ingenieros festejan sus traducciones de idiomas que no conoce nadie de la oficina. También a los estadísticos de la unidad de traducción automática de Microsoft les encanta repetir una vieja broma: que la calidad de las traducciones mejora cada vez que un filólogo abandona el equipo.

Seguramente, los expertos en campos determinados no desaparecerán por completo, pero su supremacía menguará. A partir de ahora, van a tener que compartir el podio con los *geeks* de los datos masivos, de la misma forma que la principesca causalidad tiene que convivir con la humilde correlación. Esto transforma nuestra manera de valorar el conocimiento, porque tendemos a pensar que las personas muy especializadas valen más que los generalistas: que la fortuna favorece a quien profundiza. Sin embargo, la especialización es como la exactitud: apropiada para un mundo escueto en datos, donde uno no tiene nunca información suficiente o correcta, y, por consiguiente, ha de apoyarse en la intuición y la experiencia para hallar el camino. En un mundo así, la experiencia desempeña un papel crucial, puesto que es la acumulación prolongada del conocimiento latente —conocimiento que no puede transmitirse con facilidad ni aprenderse en los libros, o quizá ni siquiera ser consciente— lo que le permite a uno tomar decisiones más adecuadas.

Pero cuando uno está atiborrado de datos hasta las orejas, lo que puede hacer en cambio es explotarlos para sacarles todo su rendimiento. Así, aquellos que analizan datos masivos pueden ver más allá de las supersticiones y el pensamiento convencional, no porque sean más listos, sino porque disponen de más datos. (Y al ser ajenos al ámbito concreto que analizan, son imparciales respecto a las trifulcas sectoriales que a veces sesgan la opinión del experto). Esto sugiere que lo que necesita un empleado para aportarle valor a la empresa cambia con el tiempo. Lo que uno necesita saber cambia, como cambian las personas a las que necesita conocer, y lo que necesita estudiar de cara a su vida profesional.

Las matemáticas y la estadística, tal vez con una pizca de programación y de ciencia de las redes, resultarán tan

fundamentales para el puesto de trabajo moderno como lo eran hace un siglo los conocimientos básicos de aritmética, y la capacidad de leer y escribir antes aún. En el pasado, para ser un biólogo excelente uno necesitaba conocer a montones de biólogos. Eso no ha cambiado del todo. Sin embargo, hoy en día también importa la amplitud de los datos masivos, no sólo el conocimiento exhaustivo de una especialidad. La solución de un problema de biología intrigante puede venir con la misma probabilidad por la asociación con un astrofísico que con un diseñador de visualización de datos.

El de los videojuegos es un sector en el que los lugartenientes de los datos masivos ya se han abierto paso a codazos hasta situarse a la altura de los generales especialistas, transformándolo de paso. Se trata de un gran negocio, que genera mayores rendimientos anuales que la taquilla cinematográfica mundial de Hollywood. En tiempos, las compañías diseñaban un juego, lo ponían a la venta y confiaban en que fuera un gran éxito. Basándose en las cifras de ventas, luego preparaban una secuela o se embarcaban en un nuevo proyecto. Las decisiones acerca[102] del ritmo del juego y los elementos del mismo como personajes, argumento, objetos y acontecimientos se basaban en la creatividad de los diseñadores, que se tomaban su trabajo con la misma seriedad que Miguel Ángel la Capilla Sixtina. Era arte, no ciencia; un mundo de corazonadas e instintos, muy parecido al de aquellos ojeadores de béisbol de *Moneyball*.

Pero esos tiempos han pasado. FarmVille, FrontierVille, FishVille y otros juegos de Zynga se hallan online y son interactivos. A primera vista, las partidas en red permiten a Zynga estudiar las estadísticas de uso y modificar los juegos según cómo se estén jugando en realidad. Así, si a los jugadores les cuesta pasar de un nivel a otro, o tienden a abandonar la partida en un momento determinado porque pierde ritmo, Zynga puede detectar esos problemas en los datos y buscarles remedio. Lo que resulta menos evidente es que la compañía puede ajustar los juegos a las características de los distintos jugadores. No existe una sola versión de FarmVille, sino cientos.

Los analistas de datos masivos de Zynga examinan si en las ventas de los bienes virtuales influye su color, o si las usan los amigos de los jugadores. Por ejemplo, después de que los datos mostrasen que los jugadores de FishVille compraban un pez traslúcido seis veces más caro que otras criaturas, Zynga puso a la venta más especies traslúcidas y aumentó considerablemente sus beneficios. En el juego Mafia Wars, los datos revelaron que los jugadores adquirían más armas con bordes dorados y compraban cachorros de tigre enteramente blancos.

Estas no son las cosas que podrían habérsele ocurrido a un diseñador de juegos trabajando en el estudio, pero los datos hablaron. “Somos una compañía de analítica que se hace pasar por una empresa de juegos. Todo lo llevan los números”, explicó Ken Rudin,[103] el entonces director de análisis de Zynga, antes de ser contratado por Facebook en un puesto similar. El aprovechamiento de los datos no garantiza en absoluto el éxito empresarial, pero sí muestra hasta dónde se puede llegar.

Empezar a tomar decisiones basadas en datos supone un cambio profundo. La mayoría de la gente basa sus decisiones en una combinación de hechos y reflexión, más una buena dosis de suposiciones. “Un aluvión de visiones subjetivas / sentimientos en el plexo solar”, en las inolvidables palabras del poeta W. H. Auden. Thomas Davenport, profesor de empresariales en el Babson College de Massachusetts y autor de numerosos libros de analítica, lo llama “las tripas de oro”. Los ejecutivos suelen estar seguros de lo que hacen por instinto visceral, así que lo siguen. Sin embargo, esto está empezando a cambiar, en la medida en que las decisiones directivas empiezan a adoptarse, o cuando menos a confirmarse, mediante modelos predictivos y análisis de datos masivos.

Por ejemplo,The-Numbers.com usa montones de datos y matemáticas para decirles a los productores independientes de Hollywood qué ingresos puede generar una película mucho antes de que se ruede su primera escena. La base de datos de la compañía procesa alrededor de treinta millones de registros que abarcan todas las películas comerciales producidas en Estados Unidos desde hace décadas. Estos incluyen el presupuesto de cada filme, su género, reparto, equipo técnico, premios recibidos, al igual que los rendimientos (de la taquilla nacional y extranjera, los derechos internacionales, las ventas y alquileres de vídeos, etc.) y demás. La base de datos también contiene un foco de conexiones humanas, del estilo de “este guionista ha trabajado con este director; este director ha rodado con este actor”, explica su fundador y presidente, Bruce Nash.

The-Numbers.com es capaz de hallar correlaciones intrincadas que predicen los ingresos de esos proyectos. Los productores llevan entonces la información a los estudios o a los inversores para conseguir financiación. La firma puede incluso jugar con las variables para aconsejar a sus clientes cómo incrementar sus ganancias (o minimizar el riesgo de pérdidas). En cierta ocasión, sus análisis concluyeron que un proyecto dado tendría muchas más probabilidades de ser un éxito si el papel protagonista lo interpretaba un actor de primera fila: específicamente, uno

nominado al Oscar y que cobraba en torno a los cinco millones de dólares. En otra, Nash informó al estudio IMAX de que un documental sobre navegación a vela sólo resultaría rentable si se recortaba su presupuesto de doce millones de dólares a ocho. “El productor se puso muy contento —dice Nash—; el director, bastante menos”.

Desde decidir hacer o no una película hasta qué jugador contratar para un equipo de béisbol, el cambio en la toma de decisiones corporativas se está empezando a notar en los balances. Erik Brynjolfsson, profesor de empresariales en la Sloan School of Management del MIT y algunos de sus colegas estudiaron los resultados de varias empresas que destacan en la toma de decisiones basadas en datos y los compararon con los de otras firmas. Descubrieron que los niveles de productividad eran hasta un 6 por 100 mayores en esas empresas que en las que no se apoyaban sobre todo en los datos para adoptar decisiones. Esto le concede una ventaja significativa a las firmas guiadas por los datos, aunque —al igual que las ventajas de mentalidad y aptitudes— puede resultar de corta duración a medida que sean cada vez más las empresas que vayan adoptando enfoques de datos masivos en su actividad.

CUESTIÓN DE UTILIDAD

A medida que los datos masivos se vayan convirtiendo en una fuente de ventaja competitiva para muchas compañías, la estructura de algunas industrias completas se verá reconfigurada. Sin embargo, no todos ganarán lo mismo: las más beneficiadas serán las firmas grandes y las pequeñas, en perjuicio de la masa central.

Los actores principales, como Amazon y Google, continuarán medrando. A diferencia de la situación que se dio en la era industrial, sin embargo, su ventaja competitiva no se hallará en la escala física. La vasta infraestructura técnica de centros de datos a sus órdenes es importante, sí, pero no constituye su cualidad más esencial. Cualquier empresa puede ajustar su potencia de computación y archivo a sus necesidades reales, si cuenta con una abundante capacidad de archivo y procesamiento digital que se pueda arrendar de forma económica y abastecer en cuestión de minutos. Convertido lo que era un coste fijo en uno variable, este cambio erosiona las ventajas de escala basadas en la infraestructura técnica que durante tanto tiempo han disfrutado las grandes compañías.

La escala aún tiene su importancia, pero ha variado. Lo que cuenta es la escala en materia de datos. Esto significa ser propietario de grandes conjuntos de datos y, al mismo tiempo, ser capaz de aprehender cada vez mayores cantidades con facilidad. Así pues, los grandes titulares de datos prosperarán a medida que recopilen y almacenen cada vez más cantidad de la materia prima de su negocio, que podrán reutilizar para crear valor adicional.

El reto para los ganadores de un mundo parco en datos y para los campeones fuera de línea —compañías como Walmart, Procter & Gamble, GE, Nestlé y Boeing— consiste en apreciar el poder de los datos masivos y recopilar y usar los datos de forma más estratégica. El fabricante de motores de avión Rolls-Royce[104] transformó de arriba abajo su negocio a lo largo de la pasada década al analizar los datos procedentes de sus productos, no sólo fabricándolos. Desde su centro de operaciones en Derby, en el Reino Unido, la compañía monitoriza de forma continua el rendimiento de más 3700 motores a reacción a lo largo y ancho del mundo para detectar los problemas antes de que se produzca una avería. Empleó los datos para ayudarse a transformar el negocio: ahora, Rolls-Royce vende los motores, pero también se ofrece a monitorizarlos, cobrando a los clientes en función del tiempo de uso (y los repara o reemplaza en caso de haber problemas). Esos servicios representan hoy en día alrededor del 70 por 100 de los ingresos anuales de la división de motores para aviación civil.

Las empresas emergentes, igual que los viejos incondicionales en nuevas áreas de negocio, están tomando posiciones para capturar grandes flujos de datos. La incursión de Apple en la telefonía móvil constituye un buen ejemplo. Antes del iPhone, los operadores de telefonía móvil amontonaban datos de los abonados, datos con un uso potencialmente valioso, pero a los que no lograban sacar partido. Apple, en cambio, estableció en sus contratos con los operadores que recibiría buena parte de la información más útil. Y así, con los datos de montones de operadores por todo el mundo, Apple consigue una visión panorámica del uso del móvil mucho más rica de lo que cualquier operador pueda alcanzar a ver por sí solo.

Los datos masivos ofrecen oportunidades excitantes también a los que se hallan en el otro extremo en cuestión de tamaño. Los pequeños actores astutos y ágiles pueden disfrutar de la “escala sin la masa”, como dijo ingeniosamente el profesor Brynjolfsson.[105] Es decir, pueden tener una gran presencia virtual aunque sus recursos físicos no sean considerables, y pueden difundir ampliamente sus innovaciones a coste reducido. Lo que resulta aún más importante, dado que algunos de los mejores servicios de datos masivos se basan primordialmente en ideas innovadoras, puede que no precisen de grandes inversiones iniciales. Las firmas pequeñas pueden licenciar la información en vez de guardársela, llevar a cabo sus análisis en plataformas de computación en la nube por un coste reducido, y pagar los cánones de las licencias con un porcentaje de los ingresos obtenidos.

Existe una buena oportunidad de que estas ventajas a ambos extremos del espectro no se limiten a los usuarios de los datos, sino que también beneficien a sus propietarios. Los grandes titulares de datos experimentan fuertes incentivos para acumular aún más, puesto que así mejoran los beneficios por un coste tan sólo marginal. En primer lugar, ya tienen instalada la infraestructura, en términos de almacenaje y procesamiento. En segundo lugar, hay un

valor especial en la combinación de conjuntos de datos. Y en tercer y último lugar, una tienda única en la que obtener los datos simplifica la vida de los usuarios de éstos.

Sin embargo, y eso es lo más curioso, puede que surja una nueva especie de propietarios de datos en el otro extremo: el de los individuos. Conforme se vaya volviendo más visible el valor de los datos, es posible que las personas deseen ampliar su poder en tanto que titulares de información que les pertenece: por ejemplo, sus preferencias de compra, sus hábitos como espectadores y quizá, también, sus datos de salud.

La propiedad de sus datos personales puede brindar a los consumidores individuales unas formas de poder que no se habían contemplado antes. Los ciudadanos pueden querer decidir por sí mismos a quién licenciar sus datos, y por cuánto. Por descontado, no todo el mundo va a querer vender sus bits al mejor postor; muchos se conformarán con verlos usados de nuevo gratis a cambio de un mejor servicio, como, por ejemplo, recomendaciones acertadas de libros de Amazon o mejor experiencia de usuario en Pinterest. Ahora, para un número significativo de consumidores duchos digitalmente, la idea de promocionar y vender su información personal puede llegar a resultar tan natural como llevar un blog, tuitear o editar una entrada de la Wikipedia.

Sin embargo, para que esto funcione, es necesario algo más que un cambio en las preferencias y la sofisticación del consumidor. Hoy en día, a la gente le resultaría demasiado complicado y costoso ceder sus datos personales, y a las empresas otro tanto negociar con cada individuo para obtenerlos. Lo más probable es que asistamos al nacimiento de nuevas firmas que recopilarán datos de muchos consumidores, ofreciendo una forma fácil de licenciarlos, y automatizarán las transacciones. Si los costes resultan lo suficientemente bajos, y si suficientes personas les otorgan su confianza, es concebible que se pueda establecer un mercado para los datos personales. Algunas empresas como la inglesa Mydex y grupos como ID^3, cofundado por Sandy Pentland, el gurú de la analítica de datos personales en el MIT, ya están trabajando para hacer realidad esta visión.

Sin embargo, hasta que estos intermediarios estén operativos y los usuarios de datos hayan empezado a recurrir a ellos, quienes deseen convertirse en propietarios de sus propios datos sólo tienen a su alcance unas opciones extremadamente limitadas. En el ínterin, si quieren conservar sus opciones para el momento en que ya estén habilitados la infraestructura y los intermediarios, quizá les interese revelar menos que más.

Para las empresas de tamaño medio, sin embargo, el enfoque de datos masivos resulta menos rentable. Existen ventajas de escala para las muy grandes, y ventajas de coste e innovación para las pequeñas, como argumenta Philip Evans[106] del Boston Consulting Group, un gran pensador sobre tecnología y negocios. En los sectores tradicionales, existen firmas de tamaño medio porque combinan un tamaño mínimo que les permite cosechar los beneficios de escala con cierta flexibilidad de la que carecen las grandes. Ahora bien, en el mundo de los datos masivos, no existe una escala mínima que deba alcanzar la empresa para poder financiar sus inversiones en infraestructura productiva. Los usuarios de datos masivos que deseen conservarse flexibles y triunfar se darán cuenta de que no les es necesario alcanzar un umbral de tamaño. Por el contrario, pueden seguir siendo pequeños y aun así florecer (o ser adquiridos por un gigante de los datos masivos).

Los datos masivos comprimen la parte media de la industria, empujando a las firmas a ser muy grandes o pequeñas, y a reaccionar rápido… o morir. Muchos sectores tradicionales terminarán redefiniéndose en términos de datos masivos, de los servicios financieros a los farmacéuticos, pasando por las manufacturas. Los datos masivos no eliminarán todas las firmas de tamaño medio en todos los sectores, pero, desde luego, presionarán sobre todo a las de los sectores más vulnerables a las convulsiones que genera el poder de los datos masivos.

Los datos masivos también están a punto de trastornar las ventajas competitivas de los países. En una época en que la producción industrial se ha perdido en buena medida a favor de los países en vías de desarrollo, y la innovación parece estar disponible para cualquiera, las naciones avanzadas conservan una ventaja en el sentido de que poseen la información y saben cómo usarla. La mala noticia es que esta ventaja no es sostenible. Igual que ocurrió con la informática y con internet, la temprana delantera de Occidente en el terreno de los datos masivos irá disminuyendo conforme otras partes del mundo vayan adoptando la tecnología. Sin embargo, la buena noticia para las actuales firmas poderosas de los países desarrollados es que los datos masivos probablemente exacerben las fortalezas y debilidades de las empresas. Así que, si una compañía domina el ámbito de los datos masivos, tiene una oportunidad no sólo de superar las prestaciones de sus pares, sino de ampliar su ventaja.

La carrera ha empezado. Igual que el algoritmo de búsqueda de Google necesita los desechos de los datos de los

usuarios para funcionar bien, e igual que aquel proveedor alemán de piezas de automóvil supo percibir la importancia de los datos para mejorar sus componentes, todas las empresas pueden obtener beneficios explotando los datos de formas inteligentes.

A pesar de los beneficios prometedores, no obstante, también hay razones para preocuparse. A medida que el sistema de datos masivos vaya haciendo predicciones cada vez más acertadas sobre el mundo y nuestro lugar en él, puede que no estemos preparados para su impacto sobre nuestra privacidad y nuestro sentido de la libertad. Nuestras percepciones e instituciones fueron creadas para un mundo de escasez, que no exceso, de información. En el siguiente capítulo exploraremos el lado oscuro de los datos masivos.

VIII

RIESGOS

Durante cerca de cuarenta años, hasta que cayó el muro de Berlín en 1989, el organismo de seguridad del estado de la República Democrática Alemana, conocido por Stasi, espió a millones de personas. Con un personal a tiempo completo de cerca de cien mil agentes, la Stasi[107] vigilaba desde vehículos y a pie de calle. Abría el correo y controlaba las cuentas corrientes, ponía micrófonos en los pisos y pinchaba las líneas telefónicas. Inducía a amantes y parejas, padres e hijos, a espiarse unos a otros, traicionando la confianza más básica que se pueden tener los seres humanos. Los archivos resultantes —por lo menos 39 millones de fichas y más de 100 km de documentos— registraron en detalle los aspectos más íntimos de la vida de la gente corriente. Alemania Oriental fue uno de los estados policiales más exhaustivos jamás vistos.

Veinte años después de la desaparición de Alemania Oriental, se están recogiendo y almacenando más datos sobre cada uno de nosotros que nunca antes. Estamos bajo vigilancia continua: cada vez que usamos nuestras tarjetas de crédito para pagar, nuestros teléfonos móviles para comunicarnos, o nuestro número de la seguridad social para identificarnos. En el año 2007, la prensa británica se deleitaba con la ironía de que hubiese más de treinta cámaras de vigilancia en un radio de doscientos metros alrededor del apartamento londinense en el que George Orwell escribió *1984*. Mucho antes del advenimiento de internet, ya había empresas especializadas como Equifax, Experian y Acxiom que recopilaban, tabulaban y ofrecían acceso a la información personal de cientos de millones de personas de todo el mundo. Internet ha hecho más fácil, más barato y más útil seguir un rastro. Y no son sólo organismos gubernamentales clandestinos con siglas de tres letras los que nos espían. Amazon monitoriza nuestras preferencias de compra, y Google nuestros hábitos de navegación, mientras que Twitter sabe qué pensamos. Facebook parece capturar asimismo toda esa información, junto con nuestras relaciones sociales. Los operadores de telefonía móvil no sólo saben con quién hablamos, sino también a quién tenemos cerca.

Con los datos masivos prometiendo revelaciones valiosas a quienes los analicen, todas las señales parecen indicar que la recopilación, almacenaje y reutilización de nuestros datos personales tienden a más. El tamaño y la escala de los conjuntos de datos seguirá acrecentándose a paso agigantado a medida que los costes de almacenamiento sigan cayendo en picado y las herramientas analíticas se tornen cada vez más poderosas. Si la era de internet era una amenaza para la privacidad, ¿no corre aún más peligro la mala utilización de los datos masivos? ¿Es ése el lado oscuro de los datos masivos?

Pues sí, y no es el único. También en este caso, lo esencial es saber que un cambio de escala lleva a un cambio de estado. Como ya veremos, esta transformación no sólo hace mucho más arduo proteger la privacidad, sino que también anuncia una amenaza enteramente nueva: la de los castigos basados en las propensiones. Es decir, la posibilidad de usar predicciones acerca de las personas basadas en datos masivos para juzgarlas y castigarlas antes incluso de que hayan actuado. Algo que niega toda idea de igualdad, justicia y libre albedrío.

Además de la privacidad y la propensión, hay un tercer peligro. Nos arriesgamos a ser víctimas de una dictadura de los datos, por la que fetichizaremos la información, el fruto de nuestros análisis, y acabaremos usándola mal. Manejados de forma responsable, los datos masivos son una herramienta útil para adoptar decisiones racionales. Empleados equivocadamente, pueden convertirse en un instrumento de poder, que algunos pueden convertir en una fuente de represión, bien simplemente frustrando a consumidores y empleados, o bien —y es peor— perjudicando a los ciudadanos.

La apuesta es de mayor cuantía de lo que se suele reconocer. Los riesgos de fracasar en la gestión de los datos masivos en lo referente a la privacidad y la predicción, o de ser engañados acerca del significado de los datos, van

mucho más allá de meras minucias como los anuncios personalizados online. La historia del siglo XX está tinta en sangre con situaciones en que los datos contribuyeron a fines ruines. En 1943, la oficina del censo de Estados Unidos dio a conocer las direcciones de las manzanas[108] (pero no los nombres y números de las calles, para mantener la ficción de que protegía la privacidad) en que residían estadounidenses de origen japonés, para facilitar su internamiento en campos. Los registros civiles de los Países Bajos, conocidos por su exhaustividad, fueron utilizados por los invasores nazis para sus redadas de judíos. Los números de cinco dígitos tatuados en el antebrazo de los prisioneros internados en campos de concentración nazis correspondían inicialmente a números de tarjetas perforadas Hollerith de IBM; el procesamiento de los datos facilitó el asesinato a escala industrial.

Pese a su potencia informativa, había muchas cosas que la Stasi no podía hacer. No podía saber, sin un gran esfuerzo, a dónde se desplazaban y cuándo sus compatriotas, ni con quién hablaban. Hoy en día, sin embargo, la mayor parte de esta información la recopilan los operadores de telefonía móvil. El estado germano oriental no podía predecir cuáles de sus ciudadanos se convertirían en disidentes, ni nosotros tampoco. Pero las fuerzas de policía están empezando a usar modelos algorítmicos para decidir dónde y cuándo patrullar, indicando el camino al que parece abocado el mundo. Estas tendencias hacen que los riesgos inherentes a los datos masivos sean tan grandes como los propios conjuntos de datos.

LA PARÁLISIS DE LA PRIVACIDAD

Resulta tentador extrapolar los peligros de privacidad que plantea el crecimiento de los datos digitales y hallar paralelismos con *1984*, la distopía de George Orwell sobre un mundo vigilado. Sin embargo, la situación es más compleja. Para empezar, no todos los conjuntos masivos de datos contienen información de carácter personal. No la tienen, por ejemplo, los datos de los sensores de las refinerías y la maquinaria de las plantas industriales, ni los datos sobre explosiones de tapas de registro ni sobre condiciones meteorológicas en el aeropuerto. BP y Con Edison no necesitan ni quieren información personal para extraer valor de la analítica que llevan a cabo. El análisis de datos masivos de este tipo no plantea prácticamente ningún riesgo para la privacidad.

Aun así, muchos de los datos que hoy en día se generan *sí* incorporan información de carácter personal. Es más, las empresas tienen un montón de incentivos para recopilar más, mantenerlos más tiempo, y reutilizarlos a menudo. Puede que los datos ni siquiera parezcan información personal de forma explícita pero, mediante su procesamiento en tanto que datos masivos, pueden remontarse fácilmente hasta el individuo al que se refieren, o deducirse detalles íntimos acerca de la vida de una persona dada.

Por ejemplo, las empresas de servicios públicos de Europa y Estados Unidos están desarrollando contadores eléctricos "inteligentes"[109] que recopilan información a lo largo de todo el día, a intervalos de hasta seis segundos incluso: mucho más que el goteo de información sobre el uso total de energía que recogían los contadores tradicionales. Lo relevante es que la forma de consumir energía de los electrodomésticos crea una "firma de carga" que es exclusiva de cada aparato. Así pues, un calentador de agua no se parece a un ordenador, que a su vez difiere de unos focos de crecimiento para el cultivo de marihuana. De este modo, el consumo de energía de una familia revela información privada, ya sea acerca de las actividades diarias de los residentes, su estado de salud o sus actividades ilegales.

La cuestión fundamental, sin embargo, no es tanto si los datos masivos aumentan los riesgos para la privacidad (que sí lo hacen), sino si cambian el carácter del riesgo. Si la amenaza es sencillamente mayor que antes, entonces puede que las leyes y reglamentos de protección del ámbito privado sigan sirviendo en la era de los datos masivos: lo único que se precisa es intensificar los esfuerzos actuales. Por otra parte, si el problema cambia, puede que necesitemos nuevas soluciones.

Y desgraciadamente, el problema se ha transformado. Con los datos masivos, el valor de la información ya no reside sólo en su propósito primero. Ahora se halla en los usos secundarios.

Este cambio socava el protagonismo que ahora tiene el individuo en las leyes de privacidad. Hoy en día, a éste se le comunica en el momento de reunir la información qué datos se recogen, y con qué fin; tiene entonces la oportunidad de manifestar su consentimiento, antes de la recopilación. Aunque este concepto de "notificación y consentimiento" no es la única forma legal de recoger y procesar datos de carácter personal, según Fred Cate, experto en privacidad en la universidad de Indiana, ha sido transformado en la piedra angular de los principios de privacidad en todo el mundo. (En la práctica, ha supuesto sobre todo unas alertas de privacidad de tamaño exagerado que rara vez lee alguien, y menos aún comprende; pero esa es otra historia).

Curiosamente, en la era de los datos masivos, la mayor parte de los usos secundarios más innovadores no han sido concebidos aún cuando se recogen los datos por primera vez. ¿Cómo podrían las empresas informar de un propósito que aún no existe? ¿Cómo podrían los individuos otorgar un consentimiento informado a algo que desconocen? Sin embargo, en ausencia de consentimiento, cualquier análisis de datos masivos con información personal podría requerir volver a ponerse en contacto con todas las personas afectadas y recabar su autorización para cada reutilización. ¿Cabe imaginar a Google intentando contactar a millones de usuarios para que le autoricen el uso de sus antiguas búsquedas y predecir con ellas la evolución de la gripe? Ninguna empresa podría asumir el coste, aun cuando la tarea fuera técnicamente factible.

La alternativa, solicitar a los usuarios en el momento de la recogida la aceptación de cualquier empleo futuro de

sus datos, tampoco resulta muy práctica. Un permiso al por mayor de esas características pervierte la noción misma del consentimiento informado. En el contexto de los datos masivos, el concepto probado y fiable de notificación y consentimiento[110] resulta muchas veces demasiado restrictivo para sacar a la luz el valor latente de la información, o, por el contrario, demasiado vacío para proteger la vida privada de los individuos.

Otras formas de proteger el ámbito privado fracasan asimismo. Cuando la información de todo el mundo está incluida en un conjunto de datos, hasta el optar por "salirse" puede dejar un rastro. Piénsese en el Street View de Google. Sus vehículos recopilaron imágenes de calles y casas en muchos países. En Alemania, Google tuvo que hacer frente a una amplia protesta ciudadana y de los medios de comunicación. La gente temía que las fotos de sus casas y jardines pudieran servir de ayuda a bandas de ladrones a la hora de elegir objetivos lucrativos. Bajo la presión normativa, Google aceptó que los propietarios de viviendas que lo desearan quedaran fuera, haciendo que sus fincas se vieran borrosas en las imágenes. Sin embargo, esa retirada resulta visible en Street View —se ven las casas tachadas—, y los ladrones en potencia podrían interpretarlo como una señal de que se trata de blancos particularmente atractivos.

La aproximación técnica a la protección de la intimidad —la anonimización— tampoco funciona bien en muchos casos. La anonimización consiste en despojar los conjuntos de datos de todos los rasgos identificadores personales, como pueden ser nombre, dirección, número de tarjeta de crédito, fecha de nacimiento, o número de la seguridad social. Los datos resultantes pueden ser analizados y compartidos sin poner en peligro la privacidad de nadie. Pero esto sólo funciona en un mundo escaso en información. Los datos masivos, al incrementar la cantidad y diversidad de la información, facilitan la reidentificación. Esto se ve en el siguiente caso.

En agosto de 2006, AOL[111] hizo públicas un montón de antiguas búsquedas en internet, con la idea bienintencionada de que los investigadores pudieran analizarlas para obtener percepciones interesantes. El conjunto de datos, integrado por veinte millones de búsquedas obra de 657 000 usuarios, efectuadas entre el 1 de marzo y el 31 de mayo de ese año, había sido cuidadosamente anonimizado. Se habían borrado las informaciones personales como nombre de usuario y dirección IP, y sustituido por identificaciones numéricas únicas. La idea era que los investigadores pudieran vincular las búsquedas de una misma persona, pero sin acceder a información personal.

Sin embargo, en cuestión de días, *The New York Times* casó búsquedas como "solteros 60", "té saludable" y "paisajistas en Lilburn (Georgia)" para lograr identificar al usuario número 4417749 como Thelma Arnold, una viuda de sesenta y dos años de Lilburn (Georgia). "Cielo santo, si es toda mi vida —le dijo ésta al periodista del *Times* que llamó a su puerta—. No tenía ni idea de que tuviera a alguien mirando por encima del hombro". El escándalo provocó el despido del director de tecnología de AOL y de otros dos empleados.

Sin embargo, escasamente dos meses después, en octubre de 2006, el servicio de alquiler de películas Netflix llevó a cabo algo similar al lanzar su premio Netflix. La empresa hizo públicos cien millones de registros de alquileres de cerca de medio millón de usuarios, y ofreció una recompensa de un millón de dólares para el equipo que consiguiera mejorar por lo menos en un 10 por 100 su sistema de recomendaciones de películas. Una vez más, los datos habían sido cuidadosamente expurgados de todos los identificadores personales. Pero aun así, uno de los usuarios fue identificado de nuevo: una madre de familia, y lesbiana que no había salido del armario, de la región conservadora del Medio Oeste, que por ello demandó posteriormente a Netflix bajo el seudónimo "Jane Doe".

Investigadores de la universidad de Texas en Austin compararon los datos de Netflix[112] con otras informaciones públicas. Descubrieron rápidamente que las calificaciones de un usuario anonimizado coincidían con las de un colaborador de la página web Internet Movie Database (IMDb). De forma más general, la investigación demostró que calificar sólo seis películas poco conocidas (de la lista de las quinientas primeras) permitía identificar al cliente de Netflix el 84 por 100 de las veces. Si además se sabía también la fecha en que había calificado las películas, podía identificárselo exactamente entre el cerca de medio millón de clientes del conjunto con un 99 por 100 de precisión.

En el caso de AOL, las identidades de los usuarios quedaron reveladas por el contenido de sus búsquedas. En el caso de Netflix, la identidad salió a la luz por la comparación de los datos con otras fuentes. En ambos ejemplos, las empresas no lograron advertir por qué los datos masivos contribuyen a desanonimizar. Por dos razones: se capturan más datos y se combinan más datos.

Paul Ohm, profesor de derecho de la universidad de Colorado en Boulder, y experto en los daños que puede provocar la desanonimización, explica que no hay ninguna solución fácil. Si se dispone de los datos suficientes, la anonimización perfecta resulta imposible, por mucho que se intente. Peor aún, recientes investigaciones han

mostrado que no sólo los datos convencionales, sino también la gráfica social —las interconexiones de las personas — son vulnerables a la desanonimización.

En la era de los datos masivos, las tres principales estrategias empleadas de antiguo para asegurar la privacidad —notificación y consentimiento individual, exclusión voluntaria y la anonimización— han perdido buena parte de su efectividad. Hoy en día, son ya muchos quienes sienten su intimidad violada: que esperen a que se generalicen más las prácticas basadas en los datos masivos.

En comparación con la Alemania Oriental de hace un cuarto de siglo, la vigilancia sólo se ha tornado más fácil, más económica y más poderosa. La capacidad de captar datos personales a menudo está profundamente arraigada en las herramientas que manejamos a diario, de las páginas web a las aplicaciones para el teléfono inteligente. Las grabadoras de datos que llevan la mayoría de los coches[113] para registrar todos los movimientos del vehículo, en los segundos antes de que se active el *airbag*, han llegado a servir de "testigos" de cargo contra el conductor tras un accidente.

Por supuesto, cuando las empresas recopilan datos para mejorar su balance, no tenemos por qué temer que su vigilancia tenga las mismas consecuencias que un teléfono pinchado por la Stasi. No acabaremos en la cárcel si Amazon descubre que nos gusta leer el *Libro rojo* de Mao. Google no nos condenará al exilio por buscar "Bing". Las empresas pueden ser poderosas, pero aún no disponen de los poderes coercitivos del estado.

Así que, aunque no se presentan en casa para llevársenos a rastras de madrugada, las firmas de todo tipo amasan montones de informaciones personales relacionadas con todos los aspectos de nuestras vidas, las comparten con otras sin nuestro conocimiento, y las usan de maneras que difícilmente hubiéramos imaginado.

El sector privado no es el único que está actuando con los datos masivos. Los gobiernos también lo hacen. Por ejemplo, se dice que la Agencia Nacional de Seguridad de Estados Unidos. (NSA en sus siglas en inglés) intercepta y almacena 1700 millones de correos electrónicos, llamadas telefónicas y otras comunicaciones cada día, según una investigación realizada por *The Washington Post* en 2010. William Binney, un antiguo agente de la NSA, estima que el gobierno ha recopilado unos "20 billones de transacciones" entre ciudadanos estadounidenses y otros: quién llama a quién, manda correos electrónicos a quién, transfiere dinero a quién, etcétera.

Para desentrañar el sentido de todos los datos, Estados Unidos está construyendo gigantescos centros de datos, como una instalación de mil doscientos millones de dólares de la NSA en Fort Williams, Utah. Y todos los ámbitos del gobierno demandan ahora más información que antes, no sólo los servicios secretos implicados en la lucha antiterrorista. Cuando la recogida de datos[114] se extiende a las transacciones financieras, los historiales de salud y las actualizaciones de estado en Facebook, la cantidad que se está cosechando es inconcebiblemente grande. El gobierno no puede procesar tanta información. Entonces, ¿por qué la recopila?

La respuesta nos da una pista de cómo ha cambiado la vigilancia en la era de los grandes datos. Antes, los investigadores "pinchaban" los cables telefónicos para averiguar lo más posible acerca de un sospechoso. Lo que importaba era profundizar y llegar a conocer a ese individuo. El enfoque moderno es diferente. En el espíritu de Google o Facebook, se piensa que las personas son la suma de sus relaciones sociales, interacciones online y conexiones con contenido. Para poder investigar exhaustivamente a un sujeto, hay que examinar la penumbra de datos lo más amplia posible que rodea a esa persona: no sólo a quién conoce, sino también a quiénes conocen esas personas a su vez, y de ahí en adelante. Esto era técnicamente muy difícil de hacer antes, pero hoy resulta más fácil que nunca. Y como el gobierno nunca sabe a quién va a querer escrutar, recopila, almacena y garantiza el acceso a la información, no necesariamente para monitorizar a todo el mundo todo el tiempo, sino para que cuando alguien caiga bajo sospecha, se hallen en condiciones de investigar de inmediato en vez de tener que empezar a reunir información desde cero.

El de Estados Unidos no es el único gobierno que amasa montones de datos sobre personas, ni quizá sea el más atroz en sus prácticas. Sin embargo, por turbadora que pueda resultar esa capacidad para conocer nuestra información personal, con los datos masivos surge un problema nuevo: el uso de predicciones para enjuiciarnos.

PROBABILIDAD Y CASTIGO

John Anderton es jefe de una unidad especial de la policía en la ciudad de Washington. Esta mañana, irrumpe en una casa de las afueras segundos antes de que Howard Marks, presa de un ataque de ira, le clave unas tijeras en el pecho a su mujer, a la que ha sorprendido en la cama con otro hombre. Para Anderton, sólo es un día más evitando que se cometan crímenes. "Por orden de la División de PreCrimen del Distrito de Columbia —recita—, lo declaro en arresto por el futuro asesinato de Sarah Marks que iba a tener lugar en el día de hoy…".

Otros agentes empiezan a maniatar a Marks, que grita: "¡No he *hecho* nada!".

En la película *Minority Report*[115] se describe una sociedad en la que las predicciones parecen tan exactas que la policía arresta a los individuos por delitos aún no cometidos. La gente es encarcelada no por lo que ha hecho, sino por lo que se prevé que vaya a hacer, aun cuando, de hecho, nunca lleguen a cometerlo. El largometraje atribuye esta aplicación presciente y preventiva de la ley a las visiones de tres adivinos, no al análisis de datos. Pero el inquietante futuro que describe podría hacerse realidad con el análisis irrestricto de datos masivos, en el que los juicios de culpabilidad se basarán en predicciones individualizadas de comportamientos futuros.

Ya empezamos a asistir a los albores de ello. Las juntas de libertad condicional en más de la mitad de los estados de Estados Unidos usan predicciones basadas en el análisis de datos como elemento de juicio a la hora de decidir si excarcelar a alguien o mantenerlo entre rejas. Cada vez más lugares de Estados Unidos —desde los distritos en Los Ángeles hasta ciudades como Richmond (Virginia)— recurren a la "policía predictiva",[116] usando el análisis de datos masivos para seleccionar qué calles, grupos e individuos han de ser sometidos a un escrutinio reforzado, simplemente porque un algoritmo los ha señalado como con mayores probabilidades de cometer un crimen.

En la ciudad de Memphis (Tennessee), un programa llamado Blue CRUSH (por las siglas en inglés de Reducción del Crimen Utilizando el Historial Estadístico) indica a los agentes de policía las áreas de interés relativamente precisas en términos de localización (unas pocas manzanas) y tiempo (unas cuantas horas de un determinado día de la semana). El sistema contribuye sin duda a que las fuerzas de la ley apliquen mejor sus escasos recursos. Desde su lanzamiento en 2006, los delitos graves contra la propiedad y los violentos han caído en una cuarta parte, según una medida (aunque, por descontado, esto nada dice acerca de la causalidad; nada permite indicar que la disminución sea debida al Blue CRUSH).

En Richmond (Virginia), la policía correlaciona los datos sobre crímenes con otros conjuntos de datos, como la información sobre cuándo pagan las nóminas a sus empleados las grandes compañías de la ciudad, o las fechas de conciertos o acontecimientos deportivos. Con ello se ha podido confirmar y, en ocasiones, refinar las sospechas de la policía acerca de las tendencias delictivas. Por ejemplo, la policía de Richmond sospechaba de antiguo que se producía un incremento de los delitos con violencia a continuación de las ferias de armamento; el análisis de datos masivos demostró que tenían razón, pero con una particularidad: el repunte se producía a las dos semanas del certamen, y no inmediatamente después.

Estos sistemas buscan prevenir los crímenes prediciéndolos, en última instancia, descendiendo hasta el nivel de los individuos que podrían cometerlos. Esto apunta al empleo de los datos masivos con una nueva finalidad: prevenir que se produzca el crimen.

Un proyecto de investigación desarrollado bajo los auspicios del Departamento de Seguridad Interior (DHS) de Estados Unidos, con el nombre de FAST ("rápido", por las siglas en inglés de Tecnología de Exploración de Futuros Atributos), trata de identificar a los potenciales terroristas monitorizando los indicadores vitales, el lenguaje corporal y otros patrones fisiológicos. La idea consiste en que la vigilancia del comportamiento de las personas puede servir para detectar sus intenciones dañinas. En pruebas, el sistema resultó ser exacto en un 70 por 100, según el DHS. (Lo que esto significa no está del todo claro: ¿se pidió a los sujetos de la investigación que fingieran ser terroristas para ver si se detectaba su "mala intención"?). Aunque estos sistemas están iniciando su desarrollo, la cuestión es que las fuerzas de seguridad se los toman muy en serio.

Impedir un delito parece un proyecto sugerente. ¿Acaso no resulta mucho mejor prevenir las infracciones antes de que se produzcan que castigarlas una vez cometidas? ¿El adelantarse a los crímenes no beneficiaría tanto a aquellos que podrían haber sido sus víctimas, como a toda la sociedad en su conjunto?

Pero esa es una senda peligrosa. Si mediante los datos masivos somos capaces de predecir quién podría cometer un crimen en el futuro, no nos conformaremos meramente con impedir que ese crimen suceda; probablemente deseemos también castigar al probable perpetrador. Esto es lógico. Si nos limitamos a intervenir para impedir que se produzca el hecho ilícito, el delincuente putativo podrá volver a intentarlo de nuevo con total impunidad. Por el contrario, usando los datos masivos para hacerlo responsable de sus actos (futuros) podemos disuadirlo a él, y de paso a otros.

Estos castigos basados en predicciones parecen una mejora respecto a otras prácticas que ya hemos acabado aceptando. Prevenir los comportamientos insalubres, peligrosos o de riesgo es una piedra angular de la sociedad moderna. Hemos hecho que sea más difícil fumar para impedir el cáncer de pulmón; exigimos llevar el cinturón de seguridad para evitar muertes en la carretera; no permitimos que los pasajeros suban armados a los aviones para evitar los secuestros. Todas estas medidas preventivas constriñen la libertad, pero muchos las consideran un precio razonable a cambio de prevenir daños mayores.

En numerosos contextos, el análisis de datos ya se está usando en nombre de la prevención. Se usa para agruparnos en cohortes de gente como nosotros, etiquetándonos de paso. Las tablas de los analistas de seguros[117] indican que los varones de más de cincuenta años son proclives al cáncer de próstata, por lo que ese grupo puede tener que pagar más por sus seguros médicos aun cuando nunca lleguen a padecer ese cáncer. En tanto que colectivo, los estudiantes de instituto con buenas notas tienen menos probabilidades de verse envueltos en accidentes de coche; por consiguiente, algunos de los menos aventajados tienen que pagar primas de seguro más elevadas. Los individuos que presentan ciertas características sufren más registros en el aeropuerto.

Esa es la idea subyacente al “perfilado” (*profiling*) en el mundo escaso en datos de hoy. Se encuentra una asociación común en los datos, se define un grupo de personas a las que es aplicable, y se las somete entonces a un escrutinio añadido. Es una regla generalizable que se aplica a todos los integrantes del grupo. *Profiling*, por supuesto, es una palabra con mucha carga, y el método presenta serios inconvenientes. En caso de emplearse mal, no sólo puede llevar a discriminar a ciertos grupos, sino también a la “culpabilidad por asociación”.

En cambio, las predicciones sobre personas a partir de datos masivos son otra cosa. Mientras que, hoy en día, las previsiones de comportamientos probables —en cosas como primas de seguros y calificaciones crediticias— normalmente se apoyan en un puñado de factores basados en un modelo mental de la cuestión de que se trate (es decir, problemas de salud anteriores o historial de pago de los créditos), con el análisis no causal del enfoque de datos masivos a menudo simplemente identificamos los factores predictivos más adecuados en un mar de información.

Lo importante es que, mediante el empleo de los datos masivos, esperamos ser capaces de identificar a personas específicas en vez de a grupos; eso nos libera de la desventaja, inherente al perfilado, de hacer de cada sospechoso un caso de culpabilidad por asociación. En un mundo de datos masivos, una persona de nombre árabe que haya pagado en efectivo un viaje sólo de ida en primera clase ya no tiene que verse sometido a una exploración exhaustiva en el aeropuerto si otros datos específicos de su persona hacen sumamente improbable que sea un terrorista. Con los datos masivos podemos zafarnos de la camisa de fuerza de las identidades grupales, sustituyéndolas por predicciones mucho más individualizadas.

La promesa de los datos masivos consiste en que hacemos lo que hemos estado haciendo todo el tiempo —perfilar—, pero mejor, de forma menos discriminatoria y más personalizada. Esto suena aceptable si el objetivo es sencillamente impedir acciones no deseadas, pero se vuelve muy peligroso si usamos predicciones basadas en datos masivos para decidir si una persona es culpable y debe ser castigada por un comportamiento que aún no se ha producido.

La mera idea de un castigo basado en las propensiones resulta nauseabunda. Acusar a una persona de un posible comportamiento futuro niega el fundamento mismo de la justicia: que uno ha de haber hecho algo antes de que se le puedan exigir cuentas. Al fin y al cabo, no es pensar en cosas malas lo que es ilegal, sino hacerlas. Es un principio fundamental de nuestra sociedad que la responsabilidad está ligada a la elección de cada individuo de cómo actuar. Si a uno lo obligan a abrir la caja fuerte de la empresa a punta de pistola, no hay elección posible, ni, por consiguiente, responsabilidad.

Si las predicciones basadas en datos masivos fueran perfectas, si los algoritmos pudieran prever nuestro futuro con infalible claridad, no tendríamos elección para obrar en el futuro. Nos comportaríamos exactamente a tenor de lo predicho. De ser posibles las predicciones perfectas, quedaría negada la voluntad humana, nuestra capacidad de vivir libremente nuestras vidas. Y, además, no sin ironía, al privarnos de elección nos librarían de toda responsabilidad.

Por supuesto, la predicción perfecta es imposible. Antes bien, el análisis de datos masivos lo que predecirá es que, para un individuo específico, hay cierta probabilidad de que tenga un comportamiento futuro determinado. Véase, por ejemplo, la investigación llevada a cabo por Richard Berk,[118] profesor de estadística y criminología de la universidad de Pensilvania. Berk afirma que puede predecir si una persona en libertad bajo fianza se verá envuelta en un homicidio (matando o siendo asesinada). Para ello utiliza numerosas variables propias del caso, entre ellas, el motivo del encarcelamiento y la fecha del primer delito, pero también datos demográficos, como la edad y el sexo. Berk sostiene que puede predecir un futuro asesino entre los presos en libertad condicional con una probabilidad de acierto mínima del 75 por 100. No está mal. Sin embargo, también significa que si los comités de libertad condicional se basan en el análisis de Berk, se equivocarán una de cada cuatro veces, y eso no es poco.

El problema central de basarse en estas predicciones no consiste tanto en que hacen correr un riesgo a la sociedad, cuanto en que con ellas esencialmente castigamos a la gente antes de que haga nada malo. Y al intervenir *antes* de que obren (por ejemplo, negándoles la libertad condicional si las predicciones indican que hay una elevada probabilidad de que cometan un asesinato), nunca sabremos si lo habrían cometido o no. No dejamos que se cumpla el destino, y aun así hacemos responsables a los individuos de lo que habrían podido hacer, según unas predicciones que nunca pueden ser desmentidas.

Esto niega el concepto mismo de la presunción de inocencia, el principio básico de nuestro sistema legal y de nuestro sentido de lo que es justo. Y si hacemos responsable a la gente de unos actos futuros pronosticados, que puede que nunca lleven a cabo, también negamos la capacidad humana de realizar elecciones morales.

Lo importante aquí *no* es sólo una cuestión de mantenimiento del orden público. El peligro es de mucha mayor amplitud que la justicia criminal; se extiende a todas las áreas de la sociedad, a todos los casos del juicio humano en que se emplean predicciones basadas en datos masivos para decidir si alguien es culpable de actos futuros o no. Estos incluyen desde la decisión adoptada por una empresa de despedir a un trabajador hasta la negativa de un cirujano a operar a un paciente, pasando por un cónyuge que presenta una demanda de divorcio.

Quizá con un sistema como ése la sociedad sería más segura o más eficiente, pero una parte esencial de lo que nos hace humanos —nuestra capacidad de elegir las acciones que llevamos a cabo y de rendir cuentas por ellas— quedaría destruida. Los datos masivos se habrían convertido en una herramienta para colectivizar la elección humana y abandonar el libre albedrío en nuestra sociedad.

Por supuesto, los datos masivos ofrecen numerosos beneficios. Lo que los convierte en un arma deshumanizadora no es una desventaja de los datos masivos en sí, sino de las formas en que usamos sus predicciones. El punto crucial radica en que hallar culpable a la gente de hechos vaticinados antes de que puedan cometerlos de verdad utiliza predicciones a partir de datos masivos basadas en correlaciones para adoptar decisiones causales sobre la responsabilidad individual.

Los datos masivos son útiles para comprender el riesgo presente y futuro, y para ajustar nuestras acciones en consonancia. Sus predicciones ayudan a pacientes y aseguradoras, prestamistas y consumidores. Pero no nos dicen nada acerca de la causalidad. En cambio, asignar “culpa” —culpabilidad individual— requiere que las personas a las que juzgamos hayan elegido actuar de determinada manera. Su decisión debe ser causa de la acción subsiguiente. Precisamente porque los datos masivos están basados en correlaciones, constituyen una herramienta del todo inadecuada para juzgar la causalidad y asignar, pues, la culpabilidad individual.

El problema radica en que los seres humanos estamos preparados para ver el mundo a través de las lentes de la causa y el efecto. Así pues, el recurso a los datos masivos se halla bajo la amenaza constante del mal uso por propósitos causales, de verse ligado a visiones optimistas como la de pensar que nuestro juicio y nuestra capacidad para asignar culpabilidades resultarían más efectivos si estuviéramos pertrechados de predicciones basadas en datos masivos.

Es la quintaesencia de la pendiente resbaladiza, la que lleva directamente a la sociedad que describía *Minority Report*, un mundo del que han sido eliminados la elección individual y el libre albedrío, en el que nuestra brújula moral individual ha sido sustituida por algoritmos predictivos, y las personas quedan expuestas al impacto libre de trabas del *fiat* colectivo. Así empleados, los datos masivos amenazan con aprisionarnos —quizá de forma literal—

en las probabilidades.

LA DICTADURA DE LOS DATOS

Los datos masivos erosionan la privacidad y amenazan a la libertad. Pero también exacerban un problema muy antiguo: el de confiar en los números cuando son harto más falibles de lo que pensamos. Nada ilustra mejor las consecuencias de un análisis de datos sesgado que la historia de Robert McNamara.[119]

McNamara era un hombre de números. Nombrado secretario de Defensa de Estados Unidos cuando se iniciaron las tensiones en Vietnam a principios de la década de 1960, insistió en recabar datos acerca de todo lo que pudo. Pensaba que sólo mediante la aplicación del rigor estadístico se podría comprender una situación compleja y adoptar las decisiones correctas. En su opinión, el mundo era una masa de información ingobernable que, una vez delineada, denotada, demarcada y cuantificada, podría ser domesticada por el hombre y por la voluntad humana. McNamara buscaba la Verdad, y esa Verdad podría encontrarse en los datos. Entre los números que llegaron a sus manos se hallaba el "recuento de cuerpos".

McNamara desarrolló su pasión por los números en la Harvard Business School, primero como estudiante y después como profesor asociado más joven de la institución, con sólo veinticuatro años. Aplicó este rigor cuantitativo durante la Segunda Guerra Mundial en tanto que integrante de un equipo de élite del Pentágono llamado Control Estadístico, que introdujo la toma de decisiones basadas en datos en una de las mayores burocracias del mundo. Anteriormente, los militares estaban ciegos. No sabían, por ejemplo, el tipo, cantidad y localización de los recambios de aeroplano disponibles. Los datos acudieron al rescate. Sólo el hacer más eficiente el suministro de armamento supuso un ahorro de 3600 millones de dólares en 1943. La guerra moderna consistía en la asignación eficiente de los recursos: el trabajo del equipo fue un éxito asombroso.

Acabada la guerra, el grupo decidió permanecer unido y ofrecer sus servicios a la empresa privada. La Ford Motor Company estaba en dificultades, y Henry Ford II, desesperado, les confió las riendas del negocio. Al igual que lo ignoraban todo sobre lo militar cuando ayudaron a ganar la guerra, tampoco tenían la menor idea sobre la fabricación de automóviles. Aun así, los llamados "cerebritos"[120] consiguieron enderezar la empresa.

McNamara ascendió rápidamente, elaborando un punto de datos para cada situación. Los gerentes de planta, agobiados, le facilitaban los datos que demandaba, fuesen estos correctos o no. Cuando llegaron órdenes desde las alturas de que había que agotar las existencias completas de un modelo de coche antes de poder iniciar la producción del siguiente, los jefes de línea, exasperados, sencillamente arrojaron las piezas sobrantes a un río cercano. Los superejecutivos de la sede central asintieron con aprobación cuando los capataces les enviaron estadísticas que confirmaban que la orden se había cumplido. La broma que circulaba en la fábrica afirmaba que uno podía caminar sobre las aguas… pisando piezas oxidadas de coches de 1950 y 1951.

McNamara era el epítome del directivo de mediados del siglo XX, el ejecutivo híperracional que confiaba en sus cifras antes que en la intuición, y que podía aplicar sus habilidades cuantitativas en cualquier sector que eligiera. En 1960 fue nombrado presidente de la Ford, un puesto que desempeñó sólo unas cuantas semanas, antes de que el presidente Kennedy lo nombrara secretario de Defensa.

A medida que el conflicto de Vietnam fue escalando y Estados Unidos enviaba más tropas, se hizo evidente que ésta era una guerra de voluntades, no de territorios. La estrategia estadounidense consistía en llevar al Viet Cong a palos hasta la mesa de negociaciones. La forma de medir el progreso, por lo tanto, pasaba por contar a los enemigos muertos, y este recuento se publicaba a diario en la prensa. Para los que apoyaban la guerra, era una prueba de progreso; para los críticos, una demostración de su inmoralidad. El recuento de cuerpos es el punto de datos que definió toda una época.

En 1977, dos años después de que el último helicóptero despegara del tejado de la embajada de Estados Unidos en Saigón, un general retirado del ejército de tierra, Douglas Kinnard, publicó un estudio que hizo época acerca de las opiniones de los generales. El informe, titulado *The War Managers* [Los gestores de la guerra], revelaba el lodazal de la cuantificación. Un escaso 2 por 100 de los generales estadounidenses consideraba el recuento de

cuerpos una forma válida de medir el progreso. Cerca de las dos terceras partes dijeron que las cifras a menudo estaban infladas. “Una falsificación: totalmente inútiles”, escribió un general en sus comentarios. “A menudo eran mentiras patentes”, escribió otro. “Muchas unidades las exageraban considerablemente, sobre todo por el increíble interés que habían demostrado personas como McNamara”, indicó un tercero.

Al igual que los operarios de la fábrica que tiraron piezas de motores al río, los oficiales de inferior rango le facilitaron ocasionalmente a sus superiores cifras apabullantes para conservar su mando o promover sus carreras, diciéndoles a los de más galones lo que querían oír. McNamara y su equipo confiaban en las cifras, las fetichizaban. Con su perfecto peinado hacia atrás y su corbata impecable, McNamara pensaba que sólo podía comprender lo que pasaba en el terreno estudiando una hoja de cálculo: todas esas filas y columnas, cálculos y gráficas perfectamente ordenados, que parecían llevarlo una desviación estándar más cerca de Dios.

El uso, abuso y mal uso de los datos por parte de los militares estadounidenses durante la guerra de Vietnam constituye una lección terrible acerca de los límites de la información en una época de datos escasos, lección que ha de ser recordada ahora que el mundo se precipita hacia la era de los datos masivos. La calidad de los datos subyacentes puede ser mediocre. Pueden estar sesgados, analizarse mal o emplearse de forma engañosa; y, de forma incluso más decisiva, pueden no captar bien lo que se proponen cuantificar.

Somos más sensibles de lo que pensamos a la “dictadura de los datos”; es decir, a permitir que los datos nos gobiernen de formas que pueden resultar tan dañinas como provechosas. La amenaza consiste en que nos dejemos atrapar irracionalmente por el resultado de nuestros análisis, aun cuando tengamos motivos razonables para sospechar que algo está mal. O en que nos acabemos obsesionando por la recopilación de hechos y cifras por el mero amor a los datos. O en que les atribuyamos un grado de veracidad que no merecen.

Según se van datificando más aspectos de la vida, la solución que políticos y hombres de negocios se ponen a buscar antes que nada es la de conseguir más datos. “En Dios confiamos:[121] que todos los demás traigan datos” es el mantra del directivo moderno que resuena en los cubículos de Silicon Valley, en las cadenas de montaje de las fábricas y por los pasillos de los organismos oficiales. La idea es válida, pero uno puede dejarse engañar por los datos.

¿Parece que la educación se está yendo al traste? Foméntense los tests para evaluar el rendimiento, y penalícese a los profesores o escuelas que no den la talla. Que los tests sirvan para medir las aptitudes de los alumnos, la calidad de la enseñanza o las necesidades de una fuerza laboral moderna, creativa y adaptable es una cuestión aún pendiente; pero es una cuestión que los datos no admiten.

¿Se desea prevenir el terrorismo? Créense capas de listas de vigilancia y de personas a las que se prohíbe volar para vigilar policialmente los cielos. Pero resulta dudoso que esos conjuntos de datos ofrezcan la protección que prometen. Según la famosa anécdota, el difunto Ted Kennedy, senador por Massachusetts, fue víctima de la lista de personas excluidas de los vuelos; fue detenido e interrogado sencillamente por llamarse igual que una persona incluida en la base de datos.

Quienes trabajan con datos tienen una expresión para designar algunos de estos problemas: “entra basura, sale basura”. En algunos casos, la razón es la calidad de la información subyacente, pero casi siempre se trata de un análisis erróneo. Con datos masivos, estos problemas pueden presentarse con más frecuencia o tener mayores consecuencias.

Google, como hemos mostrado en numerosos ejemplos, gestiona todo en función de los datos. Es obvio que a esa estrategia debe buena parte de su éxito, pero también le hace dar traspiés de vez en cuando. Durante mucho tiempo, sus cofundadores, Larry Page y Sergey Brin, insistieron en conocer, para todos los candidatos a puestos de trabajo en la firma, sus notas de final de secundaria, así como su nota media al licenciarse en la universidad. En su opinión, el primer resultado medía el potencial y el segundo los logros. Para su absoluto desconcierto, algunos directivos sobradamente preparados de cuarenta y tantos años que tomaban parte en procesos de selección quedaron descartados por esas puntuaciones. La compañía siguió analizándolas incluso mucho después de que sus estudios internos demostraran que no había correlación entre las notas y el rendimiento en el trabajo.

Google debería saber lo suficiente para no dejarse seducir por los engañosos atractivos de los datos. Esas medidas no dejan apenas margen para los cambios en la vida de las personas. Tienen menos en cuenta el conocimiento real de la vida que el de los ratones de biblioteca, y puede que no reflejen adecuadamente las

cualificaciones de la gente con formación en humanidades, donde quizá el conocimiento sea menos cuantificable que en las ciencias o la ingeniería. La obsesión de Google por este tipo de datos en el ámbito de los recursos humanos resulta especialmente llamativa, por cuanto los fundadores de la compañía son antiguos alumnos de escuelas Montessori, que ponen el énfasis en el aprendizaje, no en las notas. Por añadidura, incurre en los mismos errores que otras grandes empresas tecnológicas que se ufanaban de los currículos de la gente más que de sus capacidades reales. En tanto que doctorandos fracasados, ¿habrían podido tener Larry y Sergey la menor oportunidad de llegar a ocupar puestos directivos en los legendarios laboratorios Bell? De acuerdo con los estándares de Google, ni Bill Gates, ni Mark Zuckerberg, ni Steve Jobs habrían sido contratados nunca, al carecer de licenciaturas universitarias.

La confianza de la firma en los datos parece excesiva a veces. Cuando Marissa Mayer era directiva allí, ordenó al personal que hiciera una prueba con cuarenta y un matices distintos del color azul para ver cuáles eran más utilizados por la gente, para así decidir el color de una barra de herramientas de la página web. La deferencia de Google hacia los datos ha sido llevada al extremo. Incluso ha suscitado revueltas.

En 2009, el diseñador jefe de Google, Douglas Bowman, se despidió airado porque no podía soportar la constante cuantificación de todas las cosas. “Hace poco tuve una discusión acerca[122] de si un margen debía tener tres, cuatro o cinco píxeles, y se me pidió que demostrara mi punto de vista. No puedo trabajar en esa clase de entorno —escribió en un blog, en él anunciaba su dimisión—. Cuando una empresa está llena de ingenieros, recurre a la ingeniería para solucionar problemas. Se reduce cada decisión a un simple problema de lógica. Esos datos terminan por convertirse en una muleta para todas las decisiones, paralizando a la compañía”.

La brillantez no depende de los datos. Puede que Steve Jobs[123] mejorara el Mac portátil de forma continua a lo largo de los años basándose en los informes de campo, pero utilizó su intuición, no los datos, para lanzar el iPod, el iPhone y el iPad. Confió en su sexto sentido. “No es trabajo de los consumidores saber qué quieren”, dijo de forma memorable, al contarle a un periodista que Apple no llevó a cabo ningún estudio de mercado antes de lanzar el iPad.

En su libro *Seeing like a State*, el antropólogo James Scott, de la universidad de Yale, documenta las formas en que los gobiernos, llevados por su fetiche de la cuantificación y los datos, acaban por amargar la vida de los ciudadanos en vez de mejorarla. Utilizan mapas para reorganizar las comunidades en lugar de aprender de la gente de allí. Usan largas tablas de datos sobre las cosechas para decidir la colectivización de la agricultura sin saber nada de cultivos. Cogen todas las formas orgánicas e imperfectas en que ha interactuado la gente, y las retuercen para que se amolden a sus necesidades, en ocasiones únicamente para satisfacer su deseo de un orden cuantificable. El uso de los datos, en opinión de Scott, sirve a menudo para dar más poder a los poderosos.

He aquí la dictadura de los datos escrita con todas las letras. Y fue una desmesura similar la que llevó a Estados Unidos a la escalada del conflicto en Vietnam, en parte porque se basaron en el recuento de cuerpos y no utilizaron parámetros más significativos. “Es bastante cierto que no todas las situaciones humanas complejas posibles pueden reducirse a las curvas de un diagrama, ni a los puntos porcentuales de una gráfica, o las cifras de un balance —declaró McNamara en un discurso de 1967, cuando estaban aumentando las protestas en el país—. Pero se puede razonar sobre toda la realidad. Y no cuantificar pudiendo hacerlo equivale a contentarse con menos que el pleno alcance de la razón”. Con tal de que se usaran los datos correctos de forma correcta, en vez de respetarlos por su mera condición de datos.

Robert Strange McNamara pasó a dirigir el Banco Mundial en la década de 1970, y luego se volvió pacifista en los 80. Se convirtió en activista antinuclear y defensor del medio ambiente. Hacia el final de su vida, experimentó una conversión intelectual y escribió unas memorias, *In Retrospect*, en las que criticaba las ideas bélicas y sus propias decisiones en tanto que secretario de estado de Defensa. “Fue un error, un error terrible”, escribió. Pero se refería a la estrategia general de la guerra. Respecto a la cuestión de los datos, y del recuento de cuerpos, en particular, se mostró impenitente. Reconoció que las estadísticas podían “inducir a error o eran erróneas”; “pero las cosas que se pueden contar, deberíamos contarlas. La pérdida de vidas es una de ellas…”. McNamara falleció en 2009 a los noventa y tres años. Un hombre inteligente, pero no un sabio.

Los datos masivos pueden llevarnos a cometer el pecado de McNamara:[124] obsesionarnos tanto con ellos, con el poder y la promesa que ofrecen, que se nos olviden sus limitaciones. Para hacernos una idea del equivalente del recuento de cuerpos en términos de datos masivos, sólo necesitamos volver a contemplar Google Flu Trends.

Considérese una situación, no del todo inverosímil, en la que una variedad de gripe mortífera se propaga con ferocidad por el país. Los profesionales de la medicina agradecerían disponer de la capacidad de pronosticar en tiempo real, por medio de las búsquedas de información en Google, cuáles serán los mayores focos: así sabrían dónde acudir con ayuda.

Pero supóngase que, en un momento de crisis, los líderes políticos argumentan que simplemente saber dónde es probable que empeore la enfermedad para intentar frenarla no basta. Así que reclaman una cuarentena general, pero no para toda la población de esas regiones, pues resultaría innecesario y excesivo. Los datos masivos nos permiten mostrarnos más selectivos, por lo que la cuarentena se le aplicará sólo a los usuarios individuales de internet cuyas búsquedas hayan mostrado la correlación más alta con padecer la gripe. Aquí disponemos de los datos para saber a quién hay que buscar. Agentes federales, pertrechados con listas de direcciones IP e información de los GPS de los móviles, reúnen a los internautas y los ponen en cuarentena.

Pero, por muy razonable que pueda parecerles a algunos este supuesto, es sencillamente falso. Las correlaciones no implican causalidad. Esas personas pueden tener la gripe, o no. Habría que someterlas a pruebas. Serían reos de una predicción, pero aún más importante, serían víctimas de una visión de los datos que no aprecia lo que significa realmente la información. La clave del estudio de Google Flu Trends radica en que ciertos términos de búsqueda están *correlacionados* con el brote de gripe; pero la correlación puede darse, por ejemplo, si alguien con buena salud, al oír estornudar en la oficina, se conecta a internet para averiguar cómo protegerse, no porque esté enfermo.

EL LADO OSCURO DE LOS DATOS MASIVOS

Como hemos visto, los datos masivos permiten una mayor vigilancia de nuestras vidas y vuelven obsoletos, en buena medida, algunos de los medios legales de proteger la intimidad. También vuelven ineficaz el método técnico central para preservar el anonimato. Igualmente inquietante, las predicciones sobre individuos basadas en datos masivos pueden ser utilizadas en la práctica para castigar a la gente por sus propensiones, y no por sus acciones. Esto niega el libre albedrío y erosiona la dignidad humana.

Al mismo tiempo, existe un riesgo real de que los beneficios de los datos masivos nos induzcan a aplicar las técnicas en situaciones a las que no se ajustan del todo, o a confiar excesivamente en los análisis. Conforme las predicciones basadas en datos masivos vayan mejorando, el recurrir a ellas se irá tornando más atractivo, alimentando una obsesión por esos datos que tanto pueden aportarnos. Esa fue la maldición de McNamara, y la lección que ofrece su historia.

Debemos guardarnos contra la confianza excesiva en los datos, no vayamos a caer en el error de Ícaro, quien adoraba su capacidad técnica de volar pero no supo usarla y se precipitó en el mar. En el próximo capítulo, veremos algunas formas de controlar los datos masivos para evitar que sean ellos los que nos controlen a nosotros.

IX

CONTROL

Los cambios en la forma en que producimos la información e interactuamos con ella provocan, a su vez, cambios en las reglas que usamos para gobernarnos y en los valores que la sociedad necesita proteger. Veamos un ejemplo tomado de un diluvio de datos anterior, el que desencadenó la imprenta.

Antes de que Johannes Gutenberg inventase los caracteres móviles hacia 1450, la difusión de las ideas se limitaba en líneas generales a las conexiones entre personas. Los libros estaban confinados en su mayoría en las bibliotecas monásticas, vigilados de cerca por los monjes, que actuaban en nombre de la iglesia católica para proteger y preservar su dominio. Fuera del ámbito de la iglesia, los libros eran extremadamente raros. Unas pocas universidades habían reunido sólo docenas, o tal vez un par de cientos de libros. La de Cambridge[125] poseía escasamente ciento veintidós volúmenes a principios del siglo XV.

En cuestión de unas décadas a partir de la invención de Gutenberg, su imprenta se había extendido por toda Europa, haciendo posible la producción masiva de libros y panfletos. Cuando Martín Lutero tradujo la Biblia latina al alemán corriente, hubo de repente un motivo para aprender a leer: leyendo la Biblia por cuenta propia, se podría prescindir de los curas para aprender la palabra de Dios. La Biblia se convirtió en un éxito de ventas. Y, una vez que la gente supo leer, siguió haciéndolo. Algunos incluso decidieron escribir. En menos de lo que dura la vida de una persona, el flujo de información había pasado de goteo a torrente.

Este cambio radical abonó asimismo el terreno para nuevas reglas de gobierno de la explosión informativa que provocaron los caracteres móviles. Conforme el estado secular consolidaba su poder, estableció la censura y las licencias para contener y controlar la palabra impresa. Se introdujo la propiedad intelectual, brindando a los autores incentivos legales y económicos para crear. Más adelante, los intelectuales presionarían para lograr reglamentos que protegieran las palabras de la censura gubernativa; llegado el siglo XIX, la libertad de expresión se convirtió en un derecho garantizado por la constitución cada vez en más países. Pero estos derechos traían aparejadas unas responsabilidades. En la medida en que algunos periódicos vitriólicos pisoteaban la intimidad o ensuciaban reputaciones, surgieron reglas para proteger el ámbito privado de las personas y permitirles presentar demandas por difamación.

Sin embargo, estos cambios en la gobernanza reflejan asimismo una transformación más honda y fundamental de los valores subyacentes. A la sombra de Gutenberg, empezamos a comprender el poder de la palabra escrita; y, a la larga, también la importancia de la información que se disemina sin trabas por la sociedad. Con el paso de los siglos, optamos por más, y no menos, flujos de información y por protegernos de sus excesos no por medio de la censura principalmente, sino a través de normas que limitaban el mal uso.

Según se vaya adentrando el mundo en el ámbito de los datos masivos, la sociedad experimentará un desplazamiento tectónico similar. Los *big data* ya están transformando muchos aspectos de nuestra vida y forma de pensar, forzándonos a reconsiderar algunos principios básicos acerca de su crecimiento y su potencial dañino. Sin embargo, a diferencia de nuestros antepasados en tiempos de la revolución de la imprenta y después, no dispondremos de siglos para adaptarnos: puede que sólo se trate de unos pocos años.

No serán suficientes unos simples cambios de las reglas actuales para gobernar en la era *big data* y atemperar su lado oscuro. Más que un cambio de valores, la situación exige un cambio de paradigma. La protección de la privacidad requiere que los usuarios de datos masivos asuman mayor responsabilidad por sus actos. Al mismo tiempo, la sociedad tendrá que redefinir la misma noción de justicia para garantizar la libertad de actuación del ser humano (y, por consiguiente, la de ser considerado responsable de esas acciones). Por último, hará falta que surjan

nuevas instituciones y profesionales para interpretar los complejos algoritmos que subyacen a los hallazgos de los datos masivos, y para defender a aquellas personas que podrían verse perjudicadas por ellos.

DE LA PRIVACIDAD A LA RESPONSABILIDAD[126]

Durante décadas, un principio esencial de las leyes de protección de la vida privada alrededor del mundo pasaba por atribuirle el control a los individuos, dejándoles decidir si, cómo y por quién podría ser procesada su información personal. En la era de internet, este loable ideal se ha transformado a menudo en un sistema formal de "notificación y consentimiento". Al llegar los datos masivos, sin embargo, cuando la mayor parte del valor de estos reside en unos usos secundarios, que puede que ni siquiera se hubiesen concebido cuando se recogieron, un mecanismo así ya no sirve para asegurar la privacidad.

Para la era de los datos masivos, prevemos un marco muy diferente centrado menos en el consentimiento individual en el momento de la recogida de los datos, y más en hacer responsables a los usuarios de lo que hacen. Las firmas valorarán formalmente una reutilización determinada de los datos, basada en el impacto que tenga sobre los individuos cuya información personal están procesando. Las futuras leyes de protección de la privacidad no tienen por qué detallar de forma exhaustiva todos los casos, sino que definirán categorías amplias de uso, incluyendo las que son permisibles sin cortapisas, o con sólo algunas, limitadas y estandarizadas. Para las iniciativas de mayor riesgo, los legisladores establecerán reglas básicas para que los usuarios valoren los peligros de un uso previsto, y encuentren la forma de evitar o mitigar los daños potenciales. Esto estimulará la reutilización creativa de los datos, y al mismo tiempo asegurará que se adopten las medidas suficientes para no perjudicar a los individuos.

Valorar formal y correctamente el uso de datos masivos e implementar sus conclusiones con precisión ofrece beneficios tangibles a los usuarios de datos: en muchos casos, serán libres de buscar usos secundarios para los datos personales sin tener que volver a ponerse en contacto con los individuos para recabar su consentimiento explícito. Por otra parte, unas valoraciones chapuceras, o la mala implementación de las salvaguardias, expondrán a los usuarios de datos a responsabilidades legales, y a mandatos, multas y quizá incluso procesos penales. La responsabilidad del usuario de datos sólo será efectiva si tiene rigor.

Para ver cómo funcionaría esto en la práctica, tómese el ejemplo de la datificación de traseros del capítulo v. Imagínese que una compañía vende un servicio antirrobo para vehículos que emplea la postura del conductor al sentarse como identificador único. Posteriormente, vuelve a analizar la información para predecir los "estados de atención" del conductor, como, por ejemplo, si está soñoliento, o bebido, o furioso, para así enviar alertas a otros conductores cercanos y evitar accidentes. De acuerdo con las actuales reglas de privacidad, la firma podría pensar que necesitaba una nueva tanda de notificaciones y consentimientos, porque no había recibido autorización previa para cambiar de uso la información. Ahora bien, en un sistema de responsabilidad del usuario de datos, la compañía podría valorar los peligros del uso previsto y, en caso de parecerle mínimos, podría seguir adelante con sus planes… mejorando de paso la seguridad viaria.

El que el peso de la responsabilidad se desplace desde el público a los usuarios de los datos[127] tiene sentido por varias razones. Ellos saben mucho más que cualquiera y, desde luego, más que los consumidores o los reguladores, sobre qué uso pretenden darles. Al efectuar la valoración por sí mismos (o contratar a expertos para llevarla a cabo en su nombre), evitarán el problema de revelar estrategias de negocio confidenciales a gente externa. Y tal vez aún más importante, los usuarios de los datos se llevan la mayor parte de los beneficios de los usos secundarios, así que es justo que respondan de sus actos, y que tengan que asumir la carga de la valoración.

En un marco alternativo de protección de la privacidad como este, los usuarios de datos ya no estarán sujetos a la obligación legal de borrar la información personal una vez haya servido para su primer objeto, como exigen actualmente la mayoría de las leyes de privacidad. Este es un cambio importante, puesto que, como hemos visto, sólo explotando el valor latente de los datos pueden florecer los Maurys de nuestro tiempo, extrayendo el máximo valor posible de ellos en beneficio propio, y de la sociedad. A cambio, se les permitirá mantener la información personal durante más tiempo, aunque no indefinidamente. La sociedad necesita sopesar cuidadosamente los beneficios de la reutilización frente a los riesgos de desvelar demasiado.

Para alcanzar el equilibrio adecuado, los reguladores pueden escoger diferentes marcos temporales para la reutilización, dependiendo de los riesgos inherentes a los datos, y de los valores de diferentes sociedades. Algunas naciones pueden ser más precavidas que otras, igual que algunas clases de datos pueden ser más delicadas que otras. Este enfoque elimina también el espectro de la "memoria permanente": el riesgo de nunca poder huir uno de su pasado porque siempre se podrían recuperar los archivos digitales correspondientes. De otra manera, nuestros datos personales se cernirán sobre nosotros como una espada de Damocles, amenazando con atravesarnos a años vista con algún pormenor privado o alguna adquisición deplorable. Los límites temporales suponen también un incentivo para que los datos se usen a tiempo. Esto establece lo que, en nuestra opinión, es el mejor equilibrio para la era de los datos masivos: las empresas obtienen el derecho a usar datos personales durante más tiempo, pero, a cambio, se obligan a asumir la responsabilidad por el uso que les den, y también a eliminar los datos personales al cabo de un determinado periodo.

Además de un cambio regulatorio desde la "privacidad por consentimiento" a la "privacidad a través de la responsabilidad", prevemos que la innovación técnica ayudará a proteger la vida privada en determinados casos. Un enfoque innovador es el de la "privacidad diferencial":[128] hacer borrosos los datos de forma que una consulta en un gran conjunto de ellos no arroje resultados exactos, sino sólo aproximados. Esto hace difícil y costoso asociar unos puntos de datos particulares a unas personas específicas.

Empañar la información suena como si pudiera destruir perspectivas valiosas, pero no tiene por qué; o, por lo menos, la disyuntiva puede resultar favorable. Por ejemplo, expertos en política tecnológica han apuntado que Facebook se apoya en una forma de privacidad diferencial[129] cuando comunica información sobre sus usuarios a anunciantes potenciales: las cifras que indica son aproximadas, por lo que no pueden ayudar a revelar identidades individuales. Buscar mujeres asiáticas de Atlanta que estén interesadas en el yoga Ashtanga dará como resultado "unas 400", en vez de un número exacto, haciendo imposible el uso de la información para acotar estadísticamente a alguien en particular.

El cambio en el control desde el consentimiento individual a la responsabilidad del usuario de los datos es un cambio fundamental y esencial, necesario para una gobernanza efectiva del ámbito de los datos masivos. Pero no es el único.

PERSONAS CONTRA PREDICCIONES

Los tribunales hacen responsables de sus actos a las personas. Cuando el juez pronuncia su veredicto imparcial al acabar un juicio justo, se ha hecho justicia. Sin embargo, en la era de los datos masivos, tenemos que redefinir la noción de justicia para preservar la idea de la capacidad de decisión del ser humano: el libre albedrío por el que la gente elige sus actos. Se trata de la sencilla idea de que los individuos pueden y deben ser considerados responsables de su comportamiento, no así de sus propensiones.

Antes de los datos masivos, esta libertad fundamental resultaba obvia. Tanto, de hecho, que apenas necesitaba articularse. Al fin y al cabo, así es como funciona nuestro sistema legal: hacemos responsables de sus actos a las personas, valorando lo que han hecho. En cambio, con los datos masivos podemos predecir las acciones humanas cada vez con mayor exactitud, lo que podría incitarnos a juzgar a las personas no por lo que han hecho, sino por lo que hemos predicho que harían.

En la era de los datos masivos, tendremos que ampliar nuestra visión de la justicia y exigir que incluya salvaguardias para el albedrío humano, del mismo modo que, en la actualidad, velamos por la imparcialidad procesal. Sin esas salvaguardias, la idea misma de la justicia podría debilitarse por completo.

Al garantizar la capacidad de decisión del ser humano, nos aseguramos de que el gobierno juzga nuestro comportamiento basándose en acciones reales, no simplemente en análisis de datos masivos. Así pues, sólo debe hacernos responsables de nuestras acciones pasadas, no de predicciones estadísticas de unas acciones futuras. Y cuando el estado juzgue actos anteriores, no debería basarse exclusivamente en datos masivos. Por ejemplo, considérese el caso de nueve compañías sospechosas[130] de amañar los precios. Resulta del todo aceptable emplear análisis de datos masivos para identificar posibles colusiones, de forma que los reguladores puedan investigar y levantar el caso por medios tradicionales. Pero no se puede hallar culpables a estas empresas sólo porque los datos masivos sugieran que probablemente hayan cometido un delito.

El mismo principio debería aplicarse a las empresas privadas que adoptan decisiones importantes sobre las personas: contratarnos o despedirnos, ofrecernos un préstamo hipotecario o negarnos una tarjeta de crédito. Cuando estas decisiones se basen principalmente en predicciones a partir de datos masivos, recomendamos que se adopten determinadas salvaguardias. La primera es la transparencia: los datos y el algoritmo en que se fundamenta la predicción que afecta al individuo han de estar disponibles. La segunda es la certificación: una tercera parte experta ha de certificar que el algoritmo es correcto y válido para determinados usos sensibles. La tercera es la refutabilidad: hay que especificar formas concretas de que las personas puedan refutar una predicción sobre ellas. (Esto viene a ser análogo a la tradición, en el ámbito de la investigación científica, de revelar aquellos factores que pudieran debilitar los hallazgos de un estudio).

Lo más importante es que una garantía del albedrío humano nos protege de esa posible dictadura de los datos, la de atribuirles más sentido e importancia de la que merecen.

Resulta igualmente crucial que protejamos la responsabilidad individual. La sociedad sentirá la gran tentación de dejar de hacer responsables a los individuos, y dedicarse en cambio a gestionar riesgos, es decir, a basar las decisiones sobre las personas en valoraciones de posibilidades y probabilidades. Con tantos datos aparentemente objetivos a nuestra disposición, puede resultar atractiva la idea de desemocionalizar y desindividualizar la toma de decisiones, de basarnos en algoritmos en lugar de en las apreciaciones subjetivas de jueces y evaluadores, y formular las decisiones no en el lenguaje de la responsabilidad personal, sino en términos de riesgos más “objetivos” y cómo evitarlos.

Por ejemplo, los datos masivos suponen una fuerte tentación de predecir qué personas son más susceptibles de cometer delitos y someterlas a un tratamiento especial, escrutándolas una vez y otra en nombre de la reducción del riesgo. Los incluidos en esta categoría pueden pensar, y no sin razón, que se les está castigando sin haberse visto confrontados nunca a un comportamiento real, ni tenidos por responsables del mismo. Supóngase que un algoritmo

identifica a un adolescente determinado como muy probable autor de un delito grave en el transcurso de los siguientes tres años, y por ello se designa a un asistente social que lo visite una vez al mes, para tenerle echado un ojo y ayudarlo a no meterse en líos.

Si el adolescente y sus familiares, amigos, profesores o empleadores consideran un estigma esas visitas, entonces la intervención surte efecto de castigo, de condena por algo que no ha sucedido. Y la situación no mejora gran cosa si esas visitas se consideran no como un castigo, sino simplemente como un intento de evitar problemas futuros; como una forma de minimizar el riesgo (en este caso, el riesgo de un crimen que socavaría la seguridad pública). Si evolucionamos desde hacer responsables de sus actos a las personas a confiar en intervenciones basadas en datos para reducir su riesgo en la sociedad, estaremos devaluando el ideal de la responsabilidad individual. El estado predictivo es el estado niñera, como poco. Negarle a la gente la responsabilidad de sus actos destruye su libertad fundamental de elegir su comportamiento.

Si el estado basa muchas decisiones en predicciones y en el deseo de mitigar el riesgo, nuestras elecciones individuales —y, por consiguiente, nuestra libertad de acción individual— dejan de importar. Sin culpa, no puede haber inocencia. Ceder a un enfoque similar no mejoraría nuestra sociedad, sino al contrario.

Un pilar fundamental de la gobernanza de los datos masivos ha de ser la garantía de que seguiremos juzgando a las personas a tenor de su responsabilidad personal y de su comportamiento real, no procesando datos “objetivamente” para determinar si son probables malhechores. Sólo de esta manera los estaremos tratando como a seres humanos: como personas que tienen la libertad de decidir sus actos y el derecho a ser juzgadas por ellos.

ROMPER LA CAJA NEGRA

Los sistemas informáticos actualmente basan sus decisiones en reglas que se han marcado explícitamente en su programación. Así pues, cuando una decisión resulta incorrecta, como ocurre inevitablemente de vez en cuando, podemos volver atrás y averiguar por qué la adoptó el ordenador. Por ejemplo, podemos investigar preguntas como: "¿Por qué el sistema de piloto automático hizo ascender cinco grados al avión cuando un sensor exterior detectó un incremento repentino de la humedad?". El código informático de hoy puede abrirse e inspeccionarse, y los que saben interpretarlo pueden rastrear y comprender el fundamento de sus decisiones, por complejas que sean.

Con el análisis de datos masivos, sin embargo, esta trazabilidad se volverá mucho más difícil. La base de las predicciones de un algoritmo puede resultar a menudo demasiado intrincada para que lo entienda la mayoría de la gente.

Cuando los ordenadores estaban programados para obedecer series de instrucciones, como el remoto programa de traducción de ruso a inglés de IBM en 1954, el ser humano podía entender fácilmente por qué el *software* sustituía una palabra por otra. Ahora bien, Google Translate incorpora miles de millones de páginas de traducciones a la hora de juzgar si la palabra inglesa *light* debería traducirse como "luz" o como "ligero". Resulta imposible para el ser humano seguir las razones precisas que hay detrás de las elecciones de palabras del programa, porque están basadas en cantidades masivas de datos y en vastas computaciones estadísticas.

Los datos masivos operan a una escala que trasciende nuestro entendimiento corriente. Por ejemplo, la correlación que identificó Google entre un puñado de términos de búsqueda y la gripe fue el resultado de la prueba de 450 millones de modelos matemáticos. En cambio, Cynthia Rudin inicialmente diseñó 106 factores de predicción de si una tapa de registro podía explotar, y pudo explicarles a los directivos de Con Edison por qué su programa daba las prioridades de inspección que daba. La "explicabilidad", como se la conoce en los círculos de la inteligencia artificial, es importante para los mortales, que tendemos a querer saber por qué, y no sólo qué. Pero, ¿y si, en lugar de 106 factores de predicción, el sistema hubiese generado de forma automática nada menos que 601, la mayor parte de ellos con ponderaciones muy bajas, pero que tomados todos a la vez mejoraban la precisión del modelo? La base de cualquier predicción podría resultar asombrosamente compleja. ¿Qué habría podido contarles entonces a los directivos para convencerlos de que reasignasen su limitado presupuesto?

Mediante estos escenarios podemos advertir el riesgo de que las predicciones basadas en datos masivos, y en los algoritmos y conjuntos de datos que tienen detrás, se conviertan en cajas negras que no nos ofrecen ninguna rendición de cuentas, trazabilidad o confianza. Para impedirlo, los datos masivos requerirán monitorización y transparencia, lo que a su vez hará necesarios nuevos tipos de expertos e instituciones, capaces de ayudarnos a escrutar las predicciones basadas en datos masivos y a que quienes se sientan perjudicados por ellos puedan solicitar reparación.

Como sociedad, ya hemos visto surgir nuevas entidades similares cuando un campo determinado adquiría tal complejidad y especialización que se necesitaban expertos para gestionar las nuevas técnicas. Ciertas profesiones como el derecho, la medicina, la contabilidad y la ingeniería experimentaron la misma transformación hace más de un siglo. Más tarde, aparecieron especialistas en seguridad y privacidad informática para certificar que las compañías estaban aplicando las mejores prácticas determinadas por organismos como la Organización Internacional de la Normalización (que, a su vez, había sido creada para responder a la necesidad de nuevas directrices en este campo).

Los datos masivos precisarán de un nuevo modelo de profesionales que asuma este papel. Tal vez se los conozca por "algoritmistas". Podrían revestir dos formas —entidades independientes que monitorizaran a las firmas desde fuera, y empleados o departamentos propios para monitorizarlas desde el interior—, de la misma manera que las empresas tienen contables internos y auditores externos para revisar sus finanzas.

EL AUGE DEL ALGORITMISTA

Estos nuevos profesionales serían expertos en las áreas de ciencia de la computación, matemáticas y estadística; actuarían como revisores de análisis y predicciones de datos masivos. Los algoritmistas tendrían que hacer un voto de imparcialidad y confidencialidad, parecido al que ya hacen los contables y otros profesionales. Evaluarían la selección de fuentes de datos, la elección de herramientas analíticas y predictivas, incluyendo algoritmos y modelos, y la interpretación de los resultados. En caso de disputa, tendrían acceso a los algoritmos, a las aproximaciones estadísticas y al conjunto de datos que hubieran dado lugar a una decisión determinada.

De haber habido un algoritmista en la plantilla del Departamento de Seguridad Interior en 2004, podría haber evitado que este organismo generase una lista de excluidos de volar tan deficiente que incluía al senador Kennedy. También podrían haber desempeñado un papel en Japón, Francia, Alemania e Italia, donde algunas personas se han quejado de que la función "autocompletar" de Google, que elabora una lista de términos de búsqueda comunes asociados con el nombre que se teclea, los ha difamado. La lista se basa fundamentalmente en la frecuencia de búsquedas anteriores: los términos aparecen en el orden de su probabilidad matemática. Aun así, ¿quién no se enfadaría si la palabra "presidiario" o "prostituta" apareciese junto a su nombre cuando los posibles socios o pretendientes de uno se conectaran a la red para buscar información sobre su persona?

Imaginamos que los algoritmistas aportarán a problemas como estos un enfoque orientado al mercado que podría anticiparse a otras formas de regulación más invasivas. Satisfarían una necesidad similar a la que cubrieron los contables y auditores cuando aparecieron a principios del siglo XX para hacerse cargo del nuevo diluvio de información financiera. La ofensiva numérica resultaba difícil de entender para la mayoría de la gente; precisaba de especialistas organizados de forma ágil y autorregulada. El mercado respondió dando origen a un nuevo sector de firmas competitivas especializadas en la vigilancia financiera: una nueva especie de profesionales que reforzó la confianza de la sociedad en la economía. El mundo de los datos masivos podría, y debería, beneficiarse del refuerzo de confianza que aportarían los algoritmistas.

ALGORITMISTAS EXTERNOS

Imaginamos a los algoritmistas externos actuando como auditores imparciales, revisando la exactitud o validez de las predicciones basadas en datos masivos cada vez que el gobierno así lo requiera, bien por mandato judicial o en cumplimiento de las leyes. También podrán tener por clientes a las compañías de datos masivos, llevando a cabo auditorías para las firmas que soliciten apoyo especializado. Y podrán certificar la corrección de las aplicaciones basadas en datos masivos, como las técnicas antifraude o los sistemas de inversión bursátil. Por último, los algoritmistas externos estarán preparados para asesorar a los organismos oficiales sobre el buen uso de los datos masivos en el sector público.

Es de prever que, como en la medicina, el derecho y otras ocupaciones, esta nueva profesión se regule a sí misma dotándose de un código de conducta. La imparcialidad, confidencialidad, competencia y profesionalidad de los algoritmistas se verán reforzadas por severas reglas de responsabilidad; en caso de no cumplir con esos estándares, quedarán expuestos a demandas legales. También podrán ser citados como testigos expertos en juicios, o actuar como “peritos judiciales”, para asistir a los jueces en las cuestiones técnicas de los casos particularmente complejos.

Es más, las personas que piensen que se han visto perjudicadas por predicciones basadas en datos masivos —un paciente al que se le haya negado una intervención quirúrgica, un cliente al que no se le haya concedido la hipoteca — podrán acudir a los algoritmistas, como se hace ahora con los abogados, para que les expliquen o los ayuden a apelar contra esas decisiones.

ALGORITMISTAS INTERNOS

Los algoritmistas internos trabajan en el seno de una organización monitorizando lo que hacen con los datos masivos. Velan no sólo por los intereses de la empresa, sino también por los de las personas que se ven afectadas por sus análisis. Supervisan las operaciones de datos masivos y constituyen el primer punto de contacto para cualquiera que se sienta perjudicado por las predicciones basadas en datos masivos que haya efectuado su organización. También comprueban la integridad y exactitud de los análisis antes de permitir que se hagan públicos. Para llevar a cabo el primero de estos dos papeles, los algoritmistas deben disponer de cierto grado de libertad e imparcialidad en el seno de la organización para la que trabajan.

La noción de que una persona que trabaja para una compañía conserve la imparcialidad respecto a las operaciones de ésta puede antojarse contraintuitiva, pero es una situación ya bastante común. Las divisiones de vigilancia de las principales instituciones financieras son un ejemplo; otro tanto pasa con los consejos de administración de muchas firmas, que responden ante los accionistas, no ante la dirección. Y muchos medios de comunicación, entre ellos *The New York Times* y *The Washington Post*, emplean a defensores del lector cuya principal responsabilidad es velar por la confianza del público. Estos trabajadores se ocupan de las quejas de los lectores y a menudo critican públicamente a su empresa si creen que ha obrado mal.

Y existe una analogía aún más cercana con el algoritmista interno: el profesional encargado de asegurar que no se hace mal uso de la información personal en el marco corporativo. Por ejemplo, Alemania obliga a las compañías que superan cierto tamaño (generalmente, diez o más empleados dedicados a procesar información personal) a nombrar un responsable de la protección de datos.[131] Desde la década de 1970, estos representantes internos han desarrollado su ética profesional y su espíritu corporativo. Se reúnen periódicamente para compartir formación y las mejores prácticas, y disponen de sus propios medios de prensa y conferencias especializados. Es más, han logrado mantener una doble lealtad, a sus empleadores y a sus obligaciones en tanto que revisores imparciales, logrando actuar como defensores del pueblo de la protección de datos, al tiempo que incrustaban valores de privacidad informativa en las operaciones de sus empresas. En nuestra opinión, los algoritmistas internos pueden hacer lo mismo.

GOBERNAR A LOS SEÑORES DE LOS DATOS

Los datos son a la sociedad de la información lo que el combustible a la economía industrial: el recurso esencial que alimenta las innovaciones que usa la gente. Sin un suministro de datos rico y vibrante, y un mercado de servicios robusto, la creatividad y productividad potenciales pueden quedarse en nada.

En este capítulo hemos expuesto tres nuevas estrategias fundamentales para la gobernanza de los datos masivos, en lo referente a la privacidad, las tendencias de futuro y la auditoría de algoritmos. Confiamos en que, una vez afianzadas éstas, el lado oscuro de los datos masivos quede contenido. Sin embargo, conforme vaya desarrollándose la joven industria de los datos masivos, un nuevo desafío fundamental será el de salvaguardar unos mercados de datos masivos competitivos. Debemos impedir la aparición de "señores de los datos" del siglo XXI, el equivalente moderno de los barones rapaces que dominaron los ferrocarriles, la siderurgia y las redes telegráficas de Estados Unidos en el siglo XIX.

Para controlar a aquellos industriales, Estados Unidos estableció unas reglas antimonopolio extremadamente adaptables. Concebidas originalmente para los ferrocarriles a principios del siglo XIX, estas normas se aplicaron luego a firmas que eran guardabarreras del flujo de información del que dependen las empresas, desde National Cash Register en la década de 1910 a IBM en la de 1960 y, después, a Xerox en los 70, AT&T en los 80, Microsoft en los 90, y Google hoy. Las tecnologías en las que fueron pioneras estas firmas se convirtieron en componentes centrales de la "infraestructura de la información" de la economía, y fue precisa la fuerza de la ley para impedir un predominio malsano.

Para que se dé un bullicioso mercado de datos masivos en condiciones óptimas, necesitaremos medidas comparables a las que establecieron la competencia y la supervisión en aquellas primeras áreas tecnológicas. Deberíamos permitir las transacciones de datos, por ejemplo a través de licencias y de la interoperabilidad.[132] Esto plantea la cuestión de si podría beneficiarse la sociedad de un "derecho de exclusión" de datos cuidadosamente perfilado y equilibrado (similar al derecho de propiedad intelectual, por sorprendente que pueda parecer). Hay que reconocer que esto supondría todo un reto para los legisladores, un reto lleno de riesgos para todos los demás.

Obviamente, resulta imposible pronosticar cómo va a desarrollarse una tecnología; ni siquiera recurriendo a los datos masivos se puede predecir cómo van a evolucionar los datos masivos. Los reguladores tendrán que alcanzar un equilibrio entre la cautela y la osadía, y la historia de la legislación antimonopolios apunta a una forma posible de conseguirlo.

La regulación antimonopolios contuvo los abusos de poder. Curiosamente, sin embargo, sus principios se trasladaron magníficamente de un sector a otro, y a través de diferentes tipos de redes industriales. Es justo la clase de regulación poderosa —que no privilegia a un tipo de tecnología frente a otra— que resulta útil, puesto que protege la competencia, pero no pretende ir mucho más allá. Por consiguiente, la ley antimonopolios puede ayudar a los datos masivos a desarrollarse, exactamente igual que hizo con los ferrocarriles. Además, los gobiernos, que son de los principales acumuladores de datos del mundo, deberían hacer públicos los suyos. Resulta alentador que algunos ya estén haciendo ambas cosas; por lo menos, hasta cierto punto.

La lección que brinda la regulación antimonopolios es que, una vez que se han identificado unos principios generales, los legisladores pueden aplicarlos a garantizar que habrá un grado justo de protección. Del mismo modo, las tres estrategias que hemos propuesto —desplazar las protecciones de la privacidad del consentimiento individual a la responsabilidad de los usuarios de los datos; consagrar la voluntad humana frente a las predicciones; y crear una nueva casta de auditores de datos masivos a los que llamamos algoritmistas— pueden servir de fundamento para una gobernanza justa y efectiva de la información en la era de los datos masivos.

En muchos campos, desde la tecnología nuclear hasta la bioingeniería, primero construimos herramientas que nos

perjudican y sólo después nos dedicamos a diseñar los mecanismos de seguridad que nos protejan de esas nuevas herramientas. A este respecto, el ámbito de los datos masivos se sitúa junto a otras áreas de la sociedad que presentan retos sin soluciones absolutas, tan sólo preguntas en curso sobre cómo ordenamos nuestro mundo. Cada generación tiene que hacer frente de nuevo a esos problemas. Nuestra tarea consiste en valorar los peligros de esta poderosa tecnología, apoyar su desarrollo… y hacernos con sus recompensas.

La imprenta dio pie a una serie de cambios en la forma de gobernarse a sí misma la sociedad, y otro tanto hacen los datos masivos. Nos obligan a asumir retos antiguos de maneras nuevas, y nos enfrentan a problemas nuevos armándonos de principios consagrados por el tiempo. Para asegurarnos de que la gente esté protegida, al mismo tiempo que se promueve la tecnología, no debemos permitir que los datos masivos se desarrollen más allá del alcance de la capacidad humana de darle forma a la tecnología.

X

A PARTIR DE AHORA

A principios de la década de 2000, Mike Flowers[133] era un abogado que trabajaba en la oficina del fiscal del distrito de Manhattan llevando todo tipo de procesos, desde homicidios a delitos de Wall Street, cuando se cambió a un lujoso bufete corporativo. Después de un año tedioso en un despacho, decidió abandonar también ese trabajo y, buscando algo que tuviera más sentido, pensó en ayudar a reconstruir Irak. Un socio y amigo del bufete hizo unas cuantas llamadas a gente bien situada, y poco después, Flowers ya se hallaba de camino hacia la Zona Verde, el área de seguridad para las tropas estadounidenses en pleno centro de Bagdad, formando parte del equipo legal para el juicio contra Sadam Huseín.

La mayor parte de su trabajo resultó ser de tipo logístico, y no legal. Tenía que identificar áreas sospechosas de ocultar fosas comunes para saber dónde enviar a excavar a los investigadores. Tenía que introducir testigos en la Zona Verde evitando que saltaran por los aires en los numerosos ataques con artilugios explosivos improvisados, o IED (por sus siglas en inglés), que eran un siniestro incidente diario. Le llamó la atención que los militares trataban esas tareas como problemas de información. Y los datos acudieron en su ayuda. Los analistas de inteligencia militar combinaban los informes de campo con detalles sobre la localización, la hora y las bajas causadas por anteriores ataques con IED para predecir la ruta menos peligrosa cada día.

Unos cuantos años después, ya de vuelta en Nueva York, Flowers comprendió que esos métodos indicaban una forma más potente de combatir el crimen que todas las que había tenido a su disposición como fiscal. Y halló una auténtica alma gemela en el alcalde de la ciudad, Michael Bloomberg, quien había hecho fortuna con los datos, suministrando información financiera a la banca. En 2009, Flowers fue asignado a una unidad especial encargada de procesar datos que ayudaran a desenmascarar a los villanos del escándalo de las hipotecas *subprime*. La unidad tuvo tanto éxito que, un año más tarde, el alcalde Bloomberg le pidió que ampliara su campo de actuación. Flowers se convirtió en el primer "director analítico" de la ciudad. Su misión: constituir un equipo con los mejores científicos de datos que pudiera encontrar y explotar los montones de información virgen de la ciudad para aumentar la eficiencia en todos los terrenos posibles.

Flowers buscó a fondo hasta dar con las personas adecuadas. "No tenía ningún interés en estadísticos con mucha experiencia —explica—. Me preocupaba bastante que se mostraran reticentes a adoptar este enfoque novedoso de la resolución de problemas". Anteriormente, cuando había entrevistado a varios estadísticos para el proyecto sobre el fraude financiero, éstos habían mostrado cierta tendencia a expresar preocupaciones abstrusas acerca de los modelos matemáticos. "Yo ni siquiera pensaba en qué modelo iba a usar. Quería percepciones con las que poder actuar, eso era lo único que me importaba", dice. Al final, escogió un equipo de cinco personas a los que llama "los chicos". Todos menos uno eran recién licenciados en economía, salidos de la universidad hacía sólo uno o dos años, sin mucha experiencia sobre la vida en una gran ciudad, y todos tenían un acusado lado creativo.

Entre los primeros desafíos a los que se enfrentó el equipo estaban las "conversiones ilegales": la práctica de subdividir un alojamiento en muchas unidades más pequeñas para acabar acomodando hasta a diez veces más personas de lo proyectado. Esta clase de viviendas supone un gran riesgo de incendios, además de ser foco de delitos, drogas, enfermedades y de plagas de insectos. Por las paredes puede que culebree un lío de alargadores de cable; suele haber infiernillos eléctricos en equilibrio inestable sobre los cabeceros de las camas. La gente que vive hacinada de este modo corre un gran riesgo de perecer en incendios. En el año 2005, dos bomberos murieron al intentar rescatar a unos ciudadanos. La ciudad de Nueva York recibe aproximadamente unas 25 000 quejas anuales por conversiones ilegales, pero sólo dispone de doscientos inspectores para investigarlas. No parecía haber ningún sistema para distinguir los casos meramente molestos, de los que estaban a punto de estallar. Para Flowers y sus

chicos, sin embargo, este problema iba a poder resolverse con muchos datos.

Empezaron con una lista de todas las propiedades de la ciudad: las 900 000 que hay. A continuación, le agregaron conjuntos de datos procedentes de diecinueve organismos distintos que indicaban, por ejemplo, si el propietario del inmueble no pagaba los impuestos inmobiliarios, si había habido ejecuciones hipotecarias, y si alguna anomalía en el consumo o falta de pago de los servicios habían supuesto algún tipo de corte del suministro. También incorporaron información sobre la clase de edificio y su fecha de construcción, amén de visitas de ambulancias, tasas de delitos, quejas por roedores y demás. Luego, compararon toda esta información con cinco años de datos sobre incendios, clasificados por grado de gravedad, y buscaron correlaciones para intentar generar un sistema que permitiera predecir qué quejas deberían ser atendidas con la mayor urgencia.

Al principio, buena parte de los datos no estaban en un formato aprovechable. Por ejemplo, los responsables de los archivos de la ciudad no tenían una forma única y normalizada de describir la localización; cada organismo y cada departamento parecía usar su propio sistema. El departamento de inmuebles le asigna un número único a cada estructura, pero el de mantenimiento de la vivienda emplea un sistema de numeración diferente. Por su parte, el departamento de hacienda le atribuye a cada propiedad una referencia basada en el distrito municipal, la manzana y la parcela, mientras que la policía utiliza coordinadas cartesianas. Los bomberos se basan en un sistema de proximidad a las cabinas de teléfonos de emergencia, en relación con el emplazamiento de los cuarteles de bomberos, aun cuando esas cabinas ya no existen. Los chicos de Flowers abarcaron este desorden inventando un sistema de identificación de los edificios: usaron un área pequeña de la fachada de cada casa basada en coordenadas cartesianas, y luego importaron otros datos de geolocalización de las bases de datos de los demás organismos. Su método era inherentemente inexacto, pero la ingente cantidad de datos que eran capaces de usar compensaba con creces las imperfecciones.

El equipo, sin embargo, no se contentó con procesar datos. Salieron al terreno con los inspectores para ver cómo trabajaban. Tomaron copiosas notas y les hicieron todo tipo de preguntas a los profesionales. Cuando un canoso jefe de inspectores decía refunfuñando que el edificio que estaban a punto de visitar no sería un problema, los chicos quisieron saber por qué estaba tan seguro. El hombre no supo explicárselo del todo, pero los chicos determinaron gradualmente que su intuición se basaba en que el edificio tenía la fachada revocada hacía poco, lo que sugería que el propietario se preocupaba por el sitio.

Los chicos regresaron a sus cubículos y se preguntaron cómo podrían introducir “revocado reciente” en su modelo, a guisa de señal. Al fin y al cabo, los ladrillos no están datificados… aún. Pero, por supuesto, es preceptiva una autorización municipal para realizar cualquier trabajo en las fachadas. Añadir la información sobre los permisos mejoró las prestaciones predictivas del sistema al indicar que algunas propiedades bajo sospecha probablemente no constituyeran riesgos mayores.

La analítica también mostraba a veces que ciertas maneras de hacer las cosas, consagradas por el tiempo, no eran las mejores, igual que aquellos ojeadores de *Moneyball* habían tenido que aceptar las deficiencias de su intuición. Por ejemplo, el número de llamadas al 311, el teléfono de quejas urgentes de la ciudad, se consideraba indicativo de qué edificios estaban más necesitados de atención: más llamadas equivalían a un problema más grave. Pero ésta resultó ser una medida que inducía a error. Una sola rata detectada en el Upper East Side, el barrio *chic*, podía generar treinta llamadas en el espacio de una hora, pero hacía falta un batallón entero de roedores antes de que los residentes del Bronx se sintieran impelidos a marcar el 311. De igual modo, la mayoría de las quejas por una conversión ilegal podían estar relacionadas con el ruido, no con las situaciones de riesgo.

En junio de 2011, Flowers y sus chicos pusieron en marcha su sistema. Todas las quejas que caían en la categoría de conversión ilegal se procesaban semanalmente. Reunieron todas las que quedaron clasificadas en el 5 por 100 superior como las de mayor riesgo de incendio y se las pasaron a los inspectores para su inmediata investigación. Cuando llegaron los resultados, todo el mundo se quedó asombrado.

Antes del análisis de datos masivos, los inspectores investigaban las quejas que les parecían de peor agüero, pero sólo en el 13 por 100 de los casos hallaban condiciones lo bastante graves para requerir una orden de desalojo del inmueble. Ahora cursaban órdenes de desalojo en más del 70 por 100 de los edificios que inspeccionaban. Determinando así qué edificios requerían más urgentemente su atención, los datos masivos multiplicaron por cinco la eficiencia de la inspección. Y su trabajo se volvió además más satisfactorio: se concentraban en los problemas más graves. La eficacia redoblada de los inspectores tuvo asimismo beneficios indirectos. Los incendios en las conversiones ilegales suponen una probabilidad quince veces mayor de provocar heridos o muertos entre los

bomberos que intervienen, por lo que el departamento de bomberos se mostró encantado. Flowers y sus chicos parecían magos con una bola de cristal que les permitía ver el futuro, y predecir qué sitios presentaban mayor riesgo. Tomaron cantidades masivas de datos que llevaban años tirados, en su mayor parte sin usar desde su recogida, y los explotaron de forma novedosa para sacarles valor real. Usar un gran corpus de información les permitió advertir conexiones que no eran detectables en cantidades más pequeñas: esa es la esencia de los datos masivos.

La experiencia de los alquimistas analíticos de la ciudad de Nueva York pone de relieve muchos de los temas tratados en este libro. Usaron una cantidad monstruosa de datos, no sólo algunos; su lista de inmuebles de la ciudad representaba N = todo, nada menos. Los datos eran confusos, como la información sobre localización o los registros de las ambulancias, pero eso no los desanimó. De hecho, los beneficios de usar más datos compensaron con creces las desventajas de la información menos depurada. Fueron capaces de lograr su objetivo porque muchísimos rasgos de la ciudad habían sido datificados (si bien de forma incoherente), permitiéndoles procesar la información.

Las sospechas de los expertos tuvieron que ceder protagonismo ante el enfoque basado en datos. Al mismo tiempo, Flowers y sus chicos sometieron continuamente a prueba su sistema con los inspectores veteranos, aprovechando su experiencia para mejorar el funcionamiento. Pero, con todo, la razón principal del éxito del programa fue que prescindió de la causalidad en favor de la correlación.

“No me interesa la causalidad salvo en la medida en que lleva a la acción —explica Flowers—. La causalidad es para otras personas, y francamente, a mí se me antoja muy arriesgado empezar a hablar de causalidad. No creo que exista ni una sola causa entre el día en que alguien presenta una demanda de ejecución hipotecaria sobre una propiedad dada y el que esa finca tenga o no un riesgo histórico de incendio estructural. Me parece que sería obtuso pensarlo. Y de hecho, nadie saldría a decir eso. Pensarían: no, son los factores subyacentes. Pero ni siquiera quiero entrar en eso. Necesito un punto de datos específico al que pueda acceder, y saber su importancia. Si es significativo, actuaremos en consecuencia. Si no lo es, no haremos nada. Mire, tenemos auténticos problemas que resolver. Sinceramente, no puedo andar perdiendo el tiempo pensando en otras cosas, como la causalidad, ahora mismo”.

CUANDO HABLAN LOS DATOS

Los efectos de los datos masivos son considerables desde un punto de vista práctico, a medida que la tecnología se aplica a buscar soluciones a problemas cotidianos irritantes. Pero eso no es más que el principio. Los datos masivos están a punto de remodelar nuestro modo de vivir, trabajar y pensar. El cambio al que nos enfrentamos es, en ciertos sentidos, incluso mayor que el derivado de otras innovaciones que hicieron época, y que ampliaron acusadamente el alcance y la escala de la información en la sociedad. El suelo que pisamos se está moviendo. Las certezas anteriores se ven cuestionadas. Los datos masivos exigen una nueva discusión acerca de la naturaleza de la toma de decisiones, el destino, la justicia. Una visión del mundo que creíamos hecha de causas se enfrenta ahora a la primacía de las correlaciones. La posesión de conocimiento, que en tiempos significó comprender el pasado, está llegando a ser una capacidad de predecir el futuro.

Estos problemas son mucho más importantes que los que se nos presentaron cuando empezamos a explotar el comercio electrónico, a vivir con internet, a adentrarnos en la era de los ordenadores, o a adoptar el ábaco. La idea de que nuestra búsqueda para comprender las causas tal vez esté sobrevalorada —que en muchos casos puede resultar más ventajoso dejar de lado el *porqué* a favor del *qué*— sugiere que estos asuntos son fundamentales para nuestra sociedad y nuestra existencia. Los retos que plantean los datos masivos pueden no tener respuestas fijas tampoco. Más bien forman parte de un debate intemporal sobre el lugar del hombre en el universo y su búsqueda de sentido entre el tumulto de un mundo caótico e incomprensible.

En última instancia, los datos masivos señalan el momento en el que la "sociedad de la información" por fin cumple la promesa implícita en su nombre. Los datos son el eje de todo. Todos esos fragmentos digitales que hemos reunido pueden explotarse ahora de formas novedosas para servir a nuevos propósitos y liberar nuevas formas de valor. Pero esto requiere una forma de pensar nueva, y supondrá un desafío para nuestras instituciones e incluso para nuestro sentido de la identidad. La única certeza radica en que la cantidad de datos seguirá creciendo, igual que la capacidad de procesarlos todos. Pero mientras que la mayoría de la gente ha considerado los datos masivos como un asunto tecnológico, centrándose en el *hardware* o el *software*, nosotros creemos que hay que fijarse más bien en lo que ocurre cuando los datos hablan.

Podemos captar y analizar más información que nunca. La escasez de datos ya no define nuestros esfuerzos por interpretar el mundo. Podemos explotar cantidades enormes de datos y, en algunos casos, aproximarnos al todo. Pero eso nos fuerza a operar de maneras no tradicionales y, en particular, cambia nuestra idea acerca de qué constituye información útil.

En vez de obsesionarnos con la precisión, exactitud, limpieza y rigor de los datos, podemos permitirnos alguna flexibilidad. No deberíamos aceptar datos que sean directamente incorrectos o falsos, pero algo de confusión sí puede aceptarse a cambio de captar un conjunto de datos mucho más amplio. De hecho, en algunos casos, lo masivo y lo confuso pueden hasta representar una ventaja, dado que cuando intentamos usar únicamente una porción pequeña y exacta de los datos acabamos perdiendo la amplitud de detalle que encierra tanto conocimiento.

Las correlaciones pueden hallarse más deprisa y con menos coste que la causalidad, y de ahí que a menudo resulten preferibles. En ciertos casos, aún seguiremos necesitando estudios causales y experimentos controlados con datos cuidadosamente seleccionados, por ejemplo para probar los efectos secundarios de un fármaco o para diseñar una parte esencial de un aeroplano. Sin embargo, para muchos usos cotidianos, saber *qué* y no *por qué* es más que suficiente. Y las correlaciones de datos masivos pueden indicar el camino hacia otras áreas prometedoras en las que explorar las relaciones causales.

Estas rápidas correlaciones nos permiten ahorrar dinero en billetes de avión, predecir brotes de gripe y saber qué tapas de registro o pisos sobrepoblados se deben inspeccionar cuando los recursos son limitados. Pueden permitirles a empresas de seguros médicos ofrecer cobertura sin un reconocimiento previo y reducir el coste de recordarles a los enfermos que tienen que tomar su medicación. Se traducen idiomas y los coches se conducen solos basándose en

predicciones efectuadas a partir de correlaciones de datos masivos. Walmart puede averiguar qué sabor de Pop-Tarts[134] le conviene poner bien visible antes de un huracán. (Respuesta: fresa). Por supuesto, la causalidad está muy bien cuando se puede obtener. El problema es que, a menudo, resulta difícil de conseguir y, cuando pensamos que la hemos encontrado, muchas veces nos estamos engañando a nosotros mismos.

Las nuevas herramientas, desde los procesadores más rápidos y con más memoria al *software* y los algoritmos más inteligentes, son sólo una de las razones por las que podemos hacer todo esto. Aun cuando las herramientas son importantes, lo fundamental es que tenemos más datos, que se están datificando más aspectos del mundo. Por supuesto, la ambición humana de cuantificar el mundo antecede con mucho a la revolución informática, pero las herramientas digitales facilitan considerablemente ese proceso. No sólo permiten que los teléfonos móviles rastreen a quién llamamos y adónde vamos, sino que los datos que recopilan pueden emplearse para detectar si nos estamos poniendo enfermos. Pronto, los datos masivos serán capaces de decirnos si nos estamos enamorando.

Nuestra capacidad de hacer cosas nuevas, más cosas, mejor y más deprisa tiene el potencial de liberar un enorme valor, creando ganadores y perdedores nuevos. Gran parte del valor de los datos provendrá de sus usos secundarios, su valor de opción, y no simplemente de su uso primario, como hemos pensado hasta ahora. A consecuencia de ello, para la mayoría de los tipos de datos parece razonable recopilar cuantos más se pueda y mantenerlos durante todo el tiempo que sigan añadiendo valor, y permitir que otros los analicen si están mejor equipados para extraerles ese valor (siempre y cuando pueda uno compartir los beneficios que produzca el análisis).

Las compañías que puedan situarse en medio de los flujos de información y recopilar datos prosperarán. La explotación de los datos masivos requiere en efecto capacidades técnicas y un montón de imaginación: una mentalidad *big data*. Pero lo esencial del valor puede ir a parar a manos de quienes tienen los datos. En ocasiones, el activo importante no será sólo la información visible, sino los desechos de datos producidos por las interacciones de los individuos con la información, y que una empresa inteligente puede usar para mejorar un servicio que ya presta o para lanzar uno nuevo.

Al mismo tiempo, los datos masivos nos sitúan frente a enormes riesgos. Vuelven ineficaces los principales mecanismos técnicos y legales que existen actualmente para proteger la privacidad. Antes, sabíamos muy bien lo que constituía información personalmente identificable —nombres, números de afiliación a la seguridad social, registros fiscales, etcétera— y, por ende, resultaba relativamente sencilla de proteger. Hoy en día, hasta los datos más inocuos pueden revelar la identidad de una persona si un recopilador de datos ha reunido los suficientes. La anonimización o el ocultamiento a plena vista ya no funciona. Es más, hacer a un individuo blanco de una vigilancia implica hoy día una invasión mucho más extensa que nunca de la vida privada, puesto que las autoridades no sólo quieren obtener toda la información posible sobre una persona, sino también, en toda su amplitud, sobre sus relaciones, conexiones e interacciones.

Además de suponer una amenaza a la privacidad, estos usos de los datos masivos suscitan otro problema único y preocupante: el riesgo de que podamos juzgar a la gente no sólo por su comportamiento real, sino por las propensiones a las que apunten los datos. Según vayan volviéndose más precisas las predicciones basadas en datos masivos, la sociedad puede usarlas para castigar a personas por comportamientos predichos: actos que aún no se hayan cometido. Resulta axiomáticamente imposible demostrar la falsedad de esas predicciones; por consiguiente, las personas que se vean así acusadas nunca lograrán exculparse. Los castigos así fundados niegan el concepto de libre albedrío y la posibilidad, por pequeña que sea, de que una persona pueda elegir un camino diferente. Como la sociedad asigna una responsabilidad individual (e inflige los castigos), la voluntad humana ha de ser considerada inviolable. El futuro tiene que seguir siendo algo que podemos moldear según nuestras intenciones. Si no, los datos masivos habrán pervertido la esencia misma de la humanidad: el pensamiento racional y la libertad de elección.

No existen métodos infalibles que nos preparen plenamente para el mundo de los datos masivos; tendremos que establecer principios nuevos para nuestro autogobierno. Existen ciertas modificaciones importantes en nuestras prácticas que pueden ayudar a la sociedad mientras se va familiarizando con el carácter y las deficiencias de los datos masivos. Tenemos que proteger la privacidad desplazando la responsabilidad de los individuos hacia los usuarios de datos: es decir, que rindan cuentas por su uso. Si vamos a vivir rodeados de predicciones, resulta vital que nos aseguremos de que la voluntad humana sigue siendo sacrosanta, y preservemos no sólo la capacidad de las personas de efectuar elecciones morales, sino la responsabilidad individual por las acciones individuales. La sociedad debe diseñar salvaguardias para permitir el surgimiento de una nueva clase profesional de “algoritmistas” que evalúen la analítica de datos masivos; de forma que un mundo que se ha vuelto menos arbitrario a fuerza de

datos masivos no se convierta en una caja negra, en la que, simplemente, se sustituye una clase de incógnita por otra.

Los datos masivos se convertirán en una parte elemental de nuestra comprensión y forma de enfrentarnos a nuestros acuciantes problemas globales. Hacer frente al cambio climático requiere analizar datos de contaminación para comprender dónde resulta mejor concentrar nuestros esfuerzos y encontrar formas de mitigar los problemas. Los sensores colocados por doquier, incluidos los incrustados en los teléfonos inteligentes, proporcionan una plétora de datos que nos permitirán modelizar el calentamiento global con un nivel de detalle más preciso. Entretanto, mejorar y abaratar la atención sanitaria, especialmente para los pobres, consistirá en buena medida en automatizar tareas que ahora parecen necesitar del juicio humano pero que podrían ser asumidas por un ordenador, como, por ejemplo, examinar biopsias en busca de células cancerosas o detectar infecciones antes de que los síntomas se hagan plenamente visibles.

Los datos masivos ya se han usado en el campo del desarrollo económico y en la prevención de conflictos. Mediante el análisis de los movimientos de los usuarios de teléfonos móviles, han revelado que hay áreas de los suburbios[135] africanos que son comunidades vibrantes de actividad económica. Han descubierto áreas al borde de una confrontación étnica, e indicado cómo podrían evolucionar las crisis de refugiados. Y sus usos sólo seguirán multiplicándose conforme se vaya aplicando la tecnología a cada vez más aspectos de la vida.

Los datos masivos nos ayudan a hacer mejor lo que ya hacemos, y nos permiten hacer cosas del todo nuevas. Sin embargo, no son una varita mágica. No van a traer la paz mundial, ni a erradicar la pobreza, ni a crear al próximo Picasso. Los datos masivos no pueden hacer un bebé, pero sí salvar a los prematuros. Con el tiempo, llegaremos a esperar que se apliquen en todas las facetas de la vida (y quizá nos sintamos ligeramente alarmados cuando falten en alguna), de la misma forma que esperamos que un médico pida una radiografía para descubrir los problemas que no podría advertir con un examen físico.

Conforme los datos masivos se vayan normalizando, pueden incluso afectar a nuestra forma de pensar en el futuro. Hace alrededor de quinientos años, la humanidad experimentó una profunda transformación en su percepción del tiempo,[136] cuando Europa se volvió más secular, más científica y más ilustrada. Hasta entonces, el tiempo se experimentaba como algo cíclico, al igual que la vida. Cada día (y año) era muy parecido al anterior, e incluso el término de la vida se parecía a su inicio, al volverse los mayores otra vez como niños. Más tarde, el tiempo pasó a considerarse lineal: una secuencia de días que se iban desplegando, durante los cuales podía conformarse el mundo e influirse en la trayectoria de la vida. Si antes pasado, presente y futuro estaban fusionados, ahora la humanidad tenía un pasado que contemplar y un futuro que esperar, mientras le daba forma al presente.

Mientras que el presente se podía moldear, el futuro cambió: de ser algo perfectamente predecible a algo abierto, prístino, un vasto lienzo vacío que los individuos podían llenar en función de sus propios valores y esfuerzos. Uno de los rasgos definitorios de los tiempos modernos es el sentido que tenemos de nosotros mismos como dueños de nuestro destino; esta actitud nos diferencia de nuestros antepasados, para quienes la normalidad pasaba por alguna forma de determinismo. Sin embargo, las predicciones basadas en los datos masivos dejan el futuro menos abierto e intacto. En lugar de ser un lienzo en blanco, nuestro futuro parece estar ya esbozado con ligeros trazos que sólo pueden discernir aquellos que disponen de la tecnología para hacerlos visibles. Esto parece disminuir nuestra capacidad de dar forma a nuestro destino. La potencialidad se ve sacrificada en aras de la probabilidad.

Al mismo tiempo, los datos masivos significan que somos por siempre prisioneros de nuestras acciones pasadas, que pueden usarse en contra nuestra por unos sistemas que se creen capaces de predecir nuestro comportamiento futuro: nunca podemos escapar de lo que ha acaecido antes. “Lo pasado es el prólogo”, escribió Shakespeare. Los datos masivos consagran esto de forma algorítmica, para bien y para mal. ¿Podrán acaso las predicciones omnipotentes apagar nuestro entusiasmo por ver salir el sol, nuestro deseo de dejar una huella humana en el mundo?

De hecho, será más bien al contrario. Sabiendo cómo podrían desarrollarse las acciones en el futuro podremos adoptar medidas correctoras para prevenir los problemas o mejorar los resultados. Detectaremos a los estudiantes que flojean mucho antes del examen final. Detectaremos pequeños cánceres y los trataremos antes de que la enfermedad tenga ocasión de aflorar. Veremos la probabilidad de un embarazo adolescente indeseado o de una vida de delitos, e intervendremos para cambiar, en todo cuanto podamos, ese resultado predicho. Impediremos que los incendios consuman unos atestados pisos pobres de Nueva York, porque sabremos qué edificios hay que inspeccionar primero.

Nada está predeterminado, porque siempre podemos responder y reaccionar a la información que recibimos. Las predicciones basadas en datos masivos no están grabadas en piedra: son sólo resultados probables y eso significa, que si queremos cambiarlos, podemos hacerlo. Podemos identificar cómo recibir mejor al futuro y convertirnos en sus amos, igual que Maury encontró caminos naturales en el seno del vasto y abierto espacio de viento y olas. Y para lograr esto, no necesitaremos comprender la naturaleza del cosmos ni demostrar la existencia de los dioses: bastará con los datos masivos.

DATOS AÚN MÁS MASIVOS

A medida que los datos masivos transforman nuestras vidas —optimizando, mejorando, haciéndonos más eficientes, y capturando beneficios—, ¿qué papel les queda a la intuición, la fe, la incertidumbre y la originalidad?

Si los datos masivos nos enseñan algo, es que actuar mejor, introducir mejoras sin más —sin una comprensión más profunda— a menudo resulta suficiente. Hacerlo de forma continua es una virtud. Aunque no sepas por qué tus esfuerzos funcionan, estás generando mejores resultados de los que obtendrías sin hacer esfuerzos parecidos. Puede que Flowers y sus "chicos" de Nueva York no encarnen la ilustración de los sabios, pero salvan vidas.

El de los datos masivos no es un mundo helado de algoritmos y autómatas. Hay en él un papel esencial para las personas, con todas nuestras flaquezas, malentendidos y errores, porque estos rasgos de carácter van de la mano con la creatividad humana, el instinto y la genialidad. Los mismos procesos mentales confusos que a veces nos hacen quedar en ridículo o equivocarnos también dan lugar a éxitos y a que tropecemos con la grandeza. Esto sugiere que, igual que estamos aprendiendo a aceptar los datos confusos porque sirven a un propósito más elevado, deberíamos darle la bienvenida a la inexactitud que forma parte de la naturaleza humana. Al fin y al cabo, la confusión es una propiedad esencial tanto del mundo como de nuestras mentes; en ambos casos, aceptándola y aplicándola no obtendremos más que ventajas.

Si los datos moldean las decisiones, ¿qué propósito queda para las personas, o para la intuición y el ir en contra de los hechos? Si todo el mundo apela a los datos y aprovecha las herramientas de datos masivos, quizá lo que se convierta en el punto central de diferenciación sea la imposibilidad de predecir: el elemento humano del instinto, la asunción de riesgos, el accidente y el error.

De ser así, entonces surgirá una necesidad especial de hacerle sitio a lo humano: de reservar un espacio para la intuición, el sentido común y la buena suerte, para asegurar que no los eliminen los datos y las respuestas elaboradas por máquinas. Lo más grande de los seres humanos es precisamente lo que no revelan los algoritmos y los chips de silicio, aquello que no pueden revelar porque no puede ser capturado en forma de datos. No "lo que es", sino "lo que no es": el espacio vacío, las grietas de la acera, lo que aún no se ha dicho ni se ha pensado.

Esto tiene implicaciones importantes para la noción de progreso en la sociedad. Los datos masivos nos permiten experimentar más deprisa y explorar más pistas, ventajas que deberían producir más innovación. Pero la chispa de la invención se convierte en lo que los datos no dicen. Eso es algo que ninguna cantidad de datos podrá confirmar ni corroborar nunca, porque aún está por existir. Si Henry Ford hubiese interrogado a los algoritmos de datos masivos para saber qué querían sus clientes, habrían contestado "un caballo más rápido" (por reformular su famoso dicho). En un mundo de datos masivos, son nuestros rasgos más humanos los que necesitaremos potenciar —nuestra creatividad, intuición y ambición intelectual—, dado que nuestro ingenio es la fuente de nuestro progreso.

Los datos masivos son un recurso y una herramienta. Sirven para informar antes que para explicar; nos indican el camino para comprender, pero aun así pueden inducirnos a error, dependiendo de lo bien o mal que se manejen. Y, por deslumbrante que nos parezca el poder de los datos masivos, nunca debemos permitir que su brillo seductor nos ciegue a sus imperfecciones inherentes.

La totalidad de la información del mundo —el definitivo N = todo— nunca podrá ser recopilada, almacenada o procesada por nuestras tecnologías. Por ejemplo, el CERN,[137] el laboratorio de física de partículas en Suiza, recoge menos del 0,1 por 100 de la información que genera durante sus experimentos; el resto, aparentemente sin utilidad alguna, se deja que se disipe en el éter. Pero esto no es ni mucho menos una verdad nueva. La sociedad siempre se ha visto entorpecida por las limitaciones de las herramientas que usamos para medir y conocer la realidad, del compás y del sextante, pasando por el telescopio y el radar, hasta el GPS de hoy. Mañana nuestras herramientas pueden ser dos, o diez, o mil veces más potentes que hoy, haciendo que lo que ahora sabemos nos parezca

insignificante entonces. Nuestro actual mundo de datos masivos nos parecerá, dentro de poco, tan raro como se nos antojan ahora los 4 k de memoria RAM del ordenador que dirigía y controlaba el *Apolo 11*.

Lo que somos capaces de recopilar y procesar siempre será una fracción minúscula de la información que existe en el mundo. Sólo puede ser un simulacro de la realidad, como las sombras en la pared de la cueva de Platón. Como nunca podemos disponer de información perfecta, nuestras predicciones resultan inherentemente falibles. Tampoco significa que sean incorrectas, sólo que siempre están incompletas. Esto no niega las percepciones que ofrecen los datos masivos, pero los pone en su sitio: el de una herramienta que no ofrece respuestas definitivas, sólo algunas suficientes para ayudarnos por ahora, hasta que aparezcan métodos mejores y, por consiguiente, respuestas mejores. También sugiere que debemos usar esta herramienta con una generosa dosis de humildad… y de humanidad.

AGRADECIMIENTOS

Ambos hemos sido afortunados al haber podido trabajar y aprender con un gigante precursor en el campo de las redes de información y la innovación, Lewis M. Branscomb. Su inteligencia, elocuencia, energía, profesionalidad, ingenio e insaciable curiosidad siguen inspirándonos. A su simpática y sabia compañera, Connie Mullin, le presentamos nuestras disculpas por no haber hecho caso de su sugerencia de titular el libro "Superdata".

Momin Malik ha resultado un excelente ayudante de investigación por su excepcional inteligencia y laboriosidad. Como autores, tenemos el privilegio de que nos representen Lisa Adams y David Miller de Garamond Agency, que se han mostrado sencillamente estupendos en todos los aspectos. Eamon Dolan, nuestro editor, ha sido un fiel representante de esa rara especie de editores que poseen un sentido casi perfecto de cómo editar un texto y plantearle retos a nuestro pensamiento, de manera que el resultado es mucho mejor de lo que hubiéramos podido esperar nunca. Le estamos agradecidos a James Fransham de *The Economist* por su excelente comprobación de datos y sus agudas críticas al manuscrito.

Estamos especialmente agradecidos a todos aquellos expertos en datos masivos que dedicaron tiempo a explicarnos su trabajo, y notablemente a Oren Etzioni, Cynthia Rudin, Carolyn McGregor y Mike Flowers.

En cuanto a los agradecimientos personales de Viktor: quiero dar las gracias a Philip Evans, que siempre va dos pasos por delante de mí al razonar y sabe expresar sus ideas con precisión y elocuencia, por conversaciones que han abarcado más de una década. También deseo darle las gracias a mi antiguo colega David Lazer, quien desde hace años es un sólido teórico de los datos masivos, y a cuyo consejo he recurrido en numerosas ocasiones.

Deseo expresar mi gratitud a los participantes en el Oxford Digital Data Dialogue de 2011 (centrado en los datos masivos), y muy especialmente a su copresidente, Fred Cate, por mantener conmigo tantas discusiones fecundas.

El Oxford Internet Institute, donde trabajo, con tantos de mis colegas dedicados a la investigación sobre los datos masivos, ofrecía justo el ambiente apropiado para este libro. No se me podría haber ocurrido un sitio mejor para escribirlo. Reconozco asimismo con gratitud el apoyo del Keble College, a cuyo claustro pertenezco. Sin él, nunca hubiese logrado acceder a algunas de las importantes fuentes primarias empleadas en el libro.

La familia siempre es la más perjudicada cuando uno se dedica a escribir un libro. No es sólo por la cantidad de tiempo que he pasado delante de la pantalla del ordenador, en mi despacho, sino también por las muchas, muchísimas horas que, aún presente físicamente, he estado ensimismado en mis pensamientos, por lo que necesito pedirles perdón a mi mujer Birgit y al pequeño Viktor. Les prometo mejorar.

Por lo que se refiere a los agradecimientos personales de Kenn: quiero dar las gracias a muchos grandes científicos de los datos por su ayuda, y en particular a Jeff Hammerbacher, Amr Awadallah, DJ Patil, Michael Driscoll, Michael Freed, y a muchas personas de Google a lo largo de los años (incluyendo a Hal Varian, Jeremy Ginsberg, Peter Norvig y Udi Manber, entre otros, mientras que las charlas, demasiado breves, a mi pesar, con Eric Schmidt y Larry Page resultaron inapreciables).

Mis ideas se han enriquecido gracias a Tim O'Reilly, un sabio de la era de internet. También agradezco a Marc Benioff de Salesforce.com, quien ha sido un maestro para mí. El discernimiento de Matthew Hindman resultó inconmensurable, como siempre. James Guszcza de Deloitte fue de increíble ayuda, al igual que Geoff Hyatt, viejo amigo y emprendedor de datos en serie. Mi especial agradecimiento a Pete Warden, que es a la vez un filósofo y un experto en datos masivos.

Muchos amigos ofrecieron ideas y consejos, entre ellos John Turner, Angelika Wolf, Niko Waesche, Katia Verresen, Anna Petherick, Blaine Harden y Jessica Kowal. Otros, que inspiraron temas tratados en el libro, incluyen

a Blaise Aguera y Arcas, Eric Horvitz, David Auerbach, Gil Elbaz, Tyler Bell, Andrew Wyckoff y otros muchos en la OCDE; Stephen Brobst y el equipo de Teradata; Anthony Goldbloom y Jeremy Howard en Kaggle; Edd Dumbill, Roger Magoulas y el equipo de O'Reilly Media; y Edward Lazowska. James Cortada se merece un monumento. Gracias también a Ping Li de Accel Partners y a Roger Ehrenberg de IA Ventures.

En *The Economist*, mis colegas me ofrecieron ideas y un apoyo tremendo. Doy las gracias particularmente a mis editores Tom Standage, Daniel Franklin y John Micklethwait, así como a Barbara Beck, que editó el informe especial "Data, Data Everywhere", que constituyó la génesis de este libro. Mis colegas de Tokio, Henry Tricks y Dominic Zeigler, merecen mi admiración por buscar siempre lo novedoso y expresarlo bellamente. Oliver Morton aportó su acostumbrada sabiduría cuando más la necesitaba.

El Salzburg Global Seminar en Austria me ofreció la combinación perfecta de reposo idílico y estímulo intelectual que me ayudó a escribir y a pensar. Doy las gracias a los participantes y al organizador, Charlie Firestone, de la mesa redonda celebrada en el Aspen Institute, en julio de 2011, donde surgieron muchas ideas. Del mismo modo, deseo transmitir mi agradecimiento a Teri Elniski por su tremendo apoyo.

Frances Cairncross, rectora de Exeter College, Oxford, me brindó un lugar tranquilo en el que estar, y muchos ánimos. Reflexionar sobre unas cuestiones de tecnología y sociedad que surgen de sus planteamientos de hace década y media en *The Death of Distance*, obra que me inspiró mucho cuando era un joven periodista, supone toda una lección de humildad. Me resultaba satisfactorio cruzar todas las mañanas el patio de Exeter sabiendo que tal vez podría pasar a otros la antorcha que ella había llevado, aunque la llama ardía con mucha más brillantez en sus manos.

Mi más profundo agradecimiento es para mi familia por soportarme o, de manera más habitual, por soportar mi ausencia. Mis padres, hermana y demás parientes merecen que les dé las gracias, pero la mayor parte de mi gratitud la reservo para mi mujer Heather y para nuestros hijos Charlotte y Kaz, sin cuyo apoyo, ánimo e ideas no habría sido posible este libro.

Los dos estamos agradecidos a muchísimas personas que discutieron con nosotros la cuestión de los datos masivos, mucho antes incluso de que se popularizara el término. En este sentido, reservamos un agradecimiento especial a todos los que participaron, a lo largo de los años, en la Conferencia Rueschlikon sobre política de la información, que Viktor coorganizaba y en la que Kenn era ponente. Nos gustaría darles las gracias especialmente a Joseph Alhadeff, Bernard Benhamou, John Seely Brown, Herbert Burkert (quien nos descubrió al comodoro Maury), Peter Cullen, Ed Felten, Urs Gasser, Joi Ito, Jeff Jonas, Nicklas Lundblad, Douglas Merrill, Rick Murray, Cory Ondrejka, y Paul Schwartz.

VIKTOR MAYER-SCHÖNBERGER
KENNETH CUKIER
Oxford-Londres, agosto de 2012

BIBLIOGRAFÍA

ALTER, Alexandra, "Your E-Book Is Reading You", *The Wall Street Journal*, 29 de junio de 2012 (online.wsj.com).

ANDERSON, Benedict, *Imagined Communities*. Verso, 2006.

ANDERSON, Chris, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete", *Wired*, vol. 16, núm. 7, julio de 2008 (www.wired.com).

ASUR, Sitaram, y Bernardo A. HUBERMAN, "Predicting the Future with Social Media", *Proceedings of the 2010 ieee/wic/acm International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 492-499. (Hay versión online, disponible en: www.hpl.hp.com [PDF]).

AYRES, Ian, *Super Crunchers: Why Thinking-By-Numbers Is the New Way to be Smart*, Bantam Dell, 2007.

BABBIE, Earl, *Practice of Social Research*, Wadsworth, 2010, 12.ª ed.

BACKSTROM, Lars, Cynthia DWORK y Jon KLEINBERG, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography", *Communications of the acm*, diciembre de 2011, pp.133-141.

BAKOS, Yannis, y Erik BRYNJOLFSSON, "Bundling Information Goods: Pricing, Profits, and Efficiency", *Management Science*, 45, diciembre de 1999, pp. 1613-1630.

BANKO, Michele, y Eric BRILL, "Scaling to Very Very Large Corpora for Natural Language Disambiguation", *Microsoft Research*, 2001, p. 3 (acl.ldc.upenn.edu [PDF]).

BARBARO, Michael, y Tom ZELLER Jr., "A Face Is Exposed for AOL Searcher No. 4417749", *The New York Times*, 9 de agosto de 2006 (www.nytimes.com).

BARNES, Brook, "A Year of Disappointment at the Movie Box Office", *The New York Times*, 25 de diciembre de 2011 (www.nytimes.com).

BEATY, Janice, *Seeker of Seaways: A Life of Matthew Fontaine Maury, Pioneer Oceanographer*, Pantheon Books, 1966.

BERGER, Adam L., et al., "The Candide System for Machine Translation". *Proceedings of the 1994 ARPA Workshop on Human Language Technology*, 1994 (aclweb.org [PDF]).

BERK, Richard, "The Role of Race in Forecasts of Violent Crime", *Race and Social Problems*, núm. 1, 2009, pp. 231-242.

BLACK, Edwin, *IBM AND THE HOLOCAUST*, Crown, 2003.

BOYD, danah, y Kate CRAWFORD, "Six Provocations for Big Data", ponencia presentada en el simposio "A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society" del Oxford Internet Institute, 21 de septiembre de 2011 (ssrn.com).

BROWN, Brad, Michael CHUI, y James MANYIKA, "Are You Ready for the Era of 'Big Data'?", *McKinsey Quarterly*, octubre de 2011, p. 10.

BRYNJOLFSSON, Erik, Andrew MCAFEE, Michael SORELL, y Feng ZHU, "Scale Without Mass: Business Process Replication and Industry Dynamics". Harvard Business School Working Paper 07-016, septiembre de 2006

(www.hbs.edu [PDF]; también en: http:hbswk.hbs.edu).

BRYNJOLFSSON, Erik, Lorin HITT, y Heekyung KIM, "Strength in Numbers: How Does Data-Driven Decision-Making Affect Firm Performance?", *ICIS 2011 PROCEEDINGS*, documento 13 (aisel.aisnet.org; disponible asimismo en:papers.ssrn.com).

BYRNE, John, *The Whiz Kids*, Doubleday, 1993.

CATE, Fred H., "The Failure of Fair Information Practice Principles", en Jane K. Winn (ed.), *Consumer Protection in the Age of the "Information Economy"*, Ashgate, 2006, véase p. 341 y ss.

CHIN, A., y A. KLINEFELTER, "Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study", *North Carolina Law Review*, vol. 90, núm. 5, 2012, pp. 1417-1456.

CROSBY, Alfred, *The Measure of Reality: Quantification and Western Society, 1250-1600*. Cambridge University Press, 1997.

CUKIER, Kenneth, "Data, Data Everywhere", *The Economist*, 27 de febrero de 2010, pp. 1-14.

—, "Tracking Social Media: The Mood of the Market", *Economist.com*, 28 de junio de 2012 (www.economist.com).

DAVENPORT, Thomas H., Paul BARTH, y Randy BEAN, "How 'Big Data' Is Different", *MIT Sloan Management Review*, 30 de julio de 2012 (sloanreview.mit.edu).

DI QUINZIO, Melanie, y Anne MCCARTHY, "Rabies Risk Among Travellers", en: *Canadian Medical Association Journal*, vol. 178, núm. 5, 2008, p. 567.

DROGIN, Marc, *Anathema! Medieval Scribes and the History of Book Curses*, Allanheld and Schram, 1983.

DUGAS, A. F., et al., "Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics", *Clinical Infectious Diseases*, 8 de enero de 2012. DOI 10.1093/cid/cir883.

DUGGAN, Mark, y Steven D. LEVITT, "Winning Isn't Everything: Corruption in Sumo Wrestling", *American Economic Review*, núm. 92, 2002, pp. 1594-1605 (pricetheory.uchicago.edu [PDF]).

DUHIGG, Charles, *The Power of Habit: Why We Do What We Do in Life and Business*. Random House, 2012.

—, "How Companies Learn Your Secrets", *The New York Times*, 16 de febrero de 2012 (www.nytimes.com).

DWORK, Cynthia, "A Firm Foundation for Private Data Analysis", *Communications of the acm*, enero de 2011, pp. 86-95 (http://dl.acm.org/citation.cfm?id=1866739.1866758).

Economist, The, "Rolls-Royce: Britain's Lonely High-Flier", *The Economist*, 8 de enero de 2009 (www.economist.com).

—, "Building with Big Data: The Data Revolution Is Changing the Landscape of Business", *The Economist*, 26 de mayo de 2011 (www.economist.com).

—, "Official Statistics: Don't Lie to Me, Argentina", *The Economist*, 25 de febrero de 2012 (www.economist.com).

—, "Counting Every Moment", *The Economist*, 3 de marzo de 2012 (www.economist.com).

—, "Vehicle Data Recorders: Watching Your Driving", *The Economist*, 23 de junio de 2012 (www.economist.com).

EDWARDS, Douglas, *I'm Feeling Lucky: The Confessions of Google Employee Number 59*, Houghton Mifflin Harcourt, 2011.

EHRENBERG, Rachel, "Predicting the Next Deadly Manhole Explosion", en *Wired*, 7 de julio de 2010 (www.wired.com).

EISENSTEIN, Elizabeth L., *The Printing Revolution in Early Modern Europe*, Canto/Cambridge University Press,

1993.

ETZIONI, Oren, C. A. KNOBLOCK, R. TUCHINDA, y A. YATES, "To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price", SIGKD D'03, 24 a 27 de agosto de 2003 (http://knight.cis.temple.edu/~yates//papers/hamletkdd03.pdf).

FREI, Patrizia, et al., "Use of Mobile Phones and Risk of Brain Tumours: Update of Danish Cohort Study", *BMJ*, núm. 343, 2011 (www.bmj.com).

FURNAS, Alexander, "Homeland Security's 'Pre-Crime' Screening Will Never Work", *The Atlantic Online*, 17 de abril de 2012 (www.theatlantic.com).

GARTON ASH, Timothy, *The File*, Atlantic Books, 2008.

GERON, Tomio, "Twitter's Dick Costolo: Twitter Mobile Ad Revenue Beats Desktop on Some Days", *Forbes*, 6 de junio de 2012 (www.forbes.com).

GINSBURG, Jeremy, *et al.*, "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature* núm. 457, 19 de febrero de 2009, pp. 1012-1014 (www.nature.com).

GOLDER, Scott A., y Michael W. MACY, "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures", *Science*, vol. 333, núm. 6051, 30 de septiembre de 2011, pp. 1878-1881.

GOLLE, Philippe, "Revisiting the Uniqueness of Simple Demographics in the US Population", *Association for Computing Machinery Workshop on Privacy in Electronic Society*, 5 (2006), pp. 77-80.

GOO, Sara Kehaulani, "Sen. Kennedy Flagged by No-Fly List", *The Washington Post*, 20 de agosto de 2004, p. A01 (www.washingtonpost.com).

HAEBERLEN, A., et al., "Differential Privacy Under Fire", *SEC'11: Proceedings of the 20th USENIX Conference on Security*, p. 33 (www.cis.upenn.edu [PDF]).

HALBERSTAM, David, *The Reckoning*, William Morrow, 1986.

HALDANE, J. B. S., "On Being the Right Size", *Harper's Magazine*, marzo de 1926 (harpers.org).

HALEVY, Alon, Peter NORVIG, y Fernando PEREIRA. "The Unreasonable Effectiveness of Data", IEEE Intelligent Systems, marzo/abril de 2009, pp.8-12.

HARCOURT, Bernard E., *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, University of Chicago Press, 2006.

HARDY, Quentin, "Bizarre Insights from Big Data", *NYTimes.com*, 28 de marzo de 2012 (bits.blogs.nytimes.com).

HAYS, Constance L., "What Wal-Mart Knows About Customers' Habits", *The New York Times*, 14 de noviembre de 2004 (www.nytimes.com).

HEARN, Chester G., *Tracks in the Sea: Matthew Fontaine Maury and the Mapping of the Oceans*, International Marine/McGraw-Hill, 2002.

HELLAND, Pat, "If You Have Too Much Data then 'Good Enough' Is Good Enough', *Communications of the acm*", junio de 2011, p. 40 y ss.

HILBERT, Martin, y Priscilla LÓPEZ, "The World's Technological Capacity to Store, Communicate, and Compute Information", *Science*, vol. 332, núm. 6025, 1 de abril de 2011, pp. 60-65.

—, "How to Measure the World's Technological Capacity to Communicate, Store and Compute Information?", *International Journal of Communication*, vol. 6, 2012, pp. 956-979 (ijoc.org/ojs/index.php/ijoc/article/viewFile/1562/742).

HOLSON, Laura M., "Putting a Bolder Face on Google", *The New York Times*, 1 de marzo de 2009, p. BU1 (www.nytimes.coml).

HOPKINS, Brian, y Boris EVELSON, "Expand Your Digital Horizon with Big Data", Forrester, 30 de septiembre de 2011.

HOTZ, Robert Lee, "The Really Smart Phone", *The Wall Street Journal*, 22 de abril de 2011 (online.wsj.com).

HUTCHINS, John, "The First Public Demonstration of Machine Translation: The Georgetown-IBM System, 7th January 1954", noviembre de 2005 (hutchinsweb.me.uk [PDF]).

INGLEHART, R., y H. D. KLINGEMANN, *Genes, Culture and Happiness*, MIT Press, 2000.

ISAACSON, Walter, *Steve Jobs*, Simon & Schuster, 2011.

KAHNEMAN, Daniel, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

KAPLAN, Robert S., y David P. NORTON, *Strategy Maps: Converting Intangible Assets into Tangible Outcomes*. Harvard Business Review Press, 2004.

KARNITSCHNIG, Matthew, y Mylene MANGALINDAN, "AOL Fires Technology Chief After Web-Search Data Scandal", *The Wall Street Journal*, 21 de agosto de 2006.

KEEFE, Patrick Radden, "Can Network Theory Thwart Terrorists?", *The New York Times*, 12 de marzo de 2006 (www.nytimes.com).

KINNARD, Douglas, *The War Managers*, University Press of New England, 1977.

KIRWAN, Peter, "This Car Drives Itself", *Wired UK*, enero de 2012 (www.wired.co.uk).

KLIFF, Sarah, "A Database That Could Revolutionize Health Care", *The Washington Post*, 21 de mayo de 2012.

KRUSKAL, William, y Frederick MOSTELLER, "Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939", *International Statistical Review*, vol. 48, núm. 2, agosto de 1980, pp.169-195.

LANEY, Doug, "To Facebook You're Worth $80.95", *The Wall Street Journal*, 3 de mayo de 2012 (blogs.wsj.com).

LATOUR, Bruno, *The Pasteurization of France*, Harvard University Press, 1993.

LEVITT, Steven D., y Stephen J. DUBNER, *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. William Morrow, 2009.

LEVY, Steven. *In the Plex*, Simon and Schuster, 2011.

LEWIS, Charles Lee, *Matthew Fontaine Maury: The Pathfinder of the Seas*, U.S. Naval Institute, 1927.

LOHR, Steve, "Can Apple Find More Hits Without Its Tastemaker?", *The New York Times*, 19 de enero de 2011, p. B1 (www.nytimes.com).

LOWREY, Annie, "Economists' Programs Are Beating U.S. at Tracking Inflation", *The Washington Post*, 25 de diciembre de 2010 (www.washingtonpost.com).

MACRAKIS, Kristie, *Seduced by Secrets: Inside the Stasi's Spy-Tech World*, Cambridge University Press, 2008.

MANYIKA, James, et al., "Big Data: The Next Frontier for Innovation, Competition, and Productivity". McKinsey Global Institute, mayo de 2011 (www.mckinsey.com).

MARCUS, James, *Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut*, The New Press, 2004.

MARGOLIS, Joel M., "When Smart Grids Grow Smart Enough to Solve Crimes", *Neustar White Paper*, 18 de marzo de 2010 (energy.gov [PDF]).

MAURY, Matthew Fontaine, *The Physical Geography of the Sea*, Harper, 1855.

MAYER-SCHÖNBERGER, Viktor, "Beyond Privacy, Beyond Rights: Towards a 'Systems' Theory of Information Governance", *California Law Review*, vol. 98, pp. 1853-1885, 2010.

—, *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press, 2011, 2.ª ed.

MCGREGOR, Carolyn, Christina CATLEY, Andrew JAMES, y James PADBURY, "Next Generation Neonatal Health Informatics with Artemis", European Federation for Medical Informatics, *User Centred Networked Health Care*, editado por A. Moen *et al.* (IOS Press, 2011), p. 117 y ss.

MCNAMARA, Robert S., y Brian VANDEMARK, *In Retrospect: The Tragedy and Lessons of Vietnam*. Random House, 1995.

MEHTA, Abhishek, *Big Data: Powering the Next Industrial Revolution*, Tableau Software White Paper, 2011.

MICHEL, Jean-Baptiste, *et al.*, "Quantitative Analysis of Culture Using Millions of Digitized Books", *Science*, vol. 331, núm. 6014, 14 de enero de 2011, pp. 176-182 (www.sciencemag.org).

MILLER, Claire Cain, "U.S. Clears Google Acquisition of Travel Software", *The New York Times*, 8 de abril de 2011 (www.nytimes.com).

MILLS, Howard, "Analytics: Turning Data into Dollars", *Forward Focus*, diciembre de 2011 (www.deloitte.com [PDF]).

MINDELL, David A., *Digital Apollo: Human and Machine in Spaceflight*, MIT Press, 2008.

MINKEL, J. R., "The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II", *Scientific American*, 30 de marzo de 2007 (www.scientificamerican.com).

MURRAY, Alexander, *Reason and Society in the Middle Ages*, Oxford University Press, 1978.

NALIMOV, E. V., G. M. HAWORTH, y E. A. HEINZ, "Space-Efficient Indexing of Chess Endgame Tables", en: *icga Journal* vol. 23, núm. 3, septiembre de 2000, pp. 148-162.

NARAYANAN, Arvind, y Vitaly SHMATIKOV, "How to Break the Anonymity of the Netflix Prize Dataset", 18 de octubre de 2006, arXiv:cs/0610105 (arxiv.org).

—, "Robust De-Anonymization of Large Sparse Datasets", *Proceedings of the 2008 ieee Symposium on Security and Privacy*, p. 11 (www.cs.utexas.edu [PDF]).

NAZARETH, Rita, y Julia LEITE, "Stock Trading in U.S. Falls to Lowest Level Since 2008", *Bloomberg*, 13 de agosto de 2012 (www.bloomberg.com).

NEGROPONTE, Nicholas, *Being Digital*, Alfred Knopf, 1995.

NEYMAN, Jerzy, "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection", *Journal of the Royal Statistical Society*, vol. 97, núm. 4, 1934, pp. 558-625.

OHM, Paul, "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization", *UCLA Law Review*, vol. 57, 2010, p.1701.

ONNELA, J. P., *et al.*, "Structure and Tie Strengths in Mobile Communication Networks", *Proceedings of the National Academy of Sciences of the United States of America (pnas)* 104, mayo de 2007, pp. 7332-7336 (nd.edu [PDF]).

PALFREY, John, y Urs *GASSER, Interop: The Promise and Perils of Highly Interconnected Systems*, Basic Books, 2012.

PEARL, Judea, *Models, Reasoning and Inference*, Cambridge University Press, 2009.

POLLACK, Andrew, "DNA Sequencing Caught in the Data Deluge", *The New York Times*, 30 de noviembre de 2011 (www.nytimes.com).

President's Council of Advisors on Science and Technology [Consejo de asesores del Presidente sobre ciencia y tecnología], *Report to the President and Congress Designing a Digital Future: Federally Funded Research and*

Development in Networking and Information Technology, diciembre de 2010 (www.whitehouse.gov [PDF]).

PRIEST, Dana y William ARKIN, "A Hidden World, Growing Beyond Control", *The Washington Post*, 19 de julio de 2010 (projects.washingtonpost.com).

QUERY, Tim, "Grade Inflation and the Good-Student Discount", *Contingencies Magazine*, American Academy of Actuaries, mayo-junio de 2007 (www.contingencies.org [PDF]).

QUINN, Elias Leake, *Smart Metering and Privacy: Existing Law and Competing Policies; A Report for the Colorado Public Utility Commission*, Primavera de 2009 (http://www.w4ar.com/Danger_of_Smart_Meters_Colorado_Report.pdf).

RESHEF, David, *et al.*, "Detecting Novel Associations in Large Data Sets", *Science* núm. 334, 2011, pp. 1518-1524.

ROSENTHAL, Jonathan, "Banking Special Report", *The Economist*, 19 de mayo de 2012, pp.7-8.

ROSENZWEIG, Phil, "Robert S. McNamara and the Evolution of Modern Management", *Harvard Business Review*, diciembre de 2010, pp. 87-93 (hbr.org).

RUDIN, Cynthia, *et al.*, "21st-Century Data Miners Meet 19th-Century Electrical Cables", en: *Computer*, junio de 2011, pp. 103-105.

—, "Machine Learning for the New York City Power Grid", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, núm. 2, 2012, pp. 328-345 (hdl.handle.net).

RYS, Michael, "Scalable SQL", *Communications of the acm*, vol. 54, núm. 6, junio de 2011, pp. 48-53.

SALATHÉ, Marcel, y Shashank KHANDELWAL, "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control", *PloS Computational Biology*, vol. 7, núm. 10, octubre de 2011.

SAVAGE, Mike, y Roger BURROWS, "The Coming Crisis of Empirical Sociology", *Sociology*, vol. 41, núm. 5, 2007, pp. 885-899.

SCHLIE, Erik, Jörg RHEINBOLDT, y Niko WAESCHE, *Simply Seven: even Ways to Create a Sustainable Internet Business*, Palgrave Macmillan, 2011.

SCANLON, Jessie, "Luis von Ahn: The Pioneer of 'Human Computation'", en: *Businessweek*, 3 de noviembre de 2008 (www.businessweek.com).

SCISM, Leslie, y Mark MAREMONT, "Inside Deloitte's Life-Insurance Assessment Technology", *The Wall Street Journal*, 19 de noviembre de 2010 (online.wsj.com).

—, "Insurers Test Data Profiles to Identify Risky Clients", *The Wall Street Journal*, 19 de noviembre de 2010 (online.wsj.com).

SCOTT, James, *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*, Yale University Press, 1998.

SELTZER, William, y Margo ANDERSON, "The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses", *Social Research*, vol. 68, núm. 2, verano de 2001, pp. 481-513.

SILVER, Nate, *The Signal and the Noise: Why So Many Predictions Fail, But Some Don't*, Penguin, 2012.

SINGEL, Ryan, "Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims", *Wired*, 17 de diciembre de 2009 (www.wired.com/).

SMITH, Adam, *The Wealth of Nations* (1776), reimpresión en Bantam Classics, 2003. Hay una versión electrónica gratuita en: www2.hn.psu.edu [PDF].

SOLOVE, Daniel J., *The Digital Person: Technology and Privacy in the Information Age*, NYU Press, 2004.

SUROWIECKI, James, "A Billion Prices Now", *The New Yorker*, 30 de mayo de 2011 (www.newyorker.com).

TALEB, Nassim Nicholas, *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*, Random House, 2008.

—, *The Black Swan: The Impact of the Highly Improbable*, Random House, 2010, 2.ª ed.

THOMPSON, Clive, "For Certain Tasks, the Cortex Still Beats the CPU", *Wired* núm. 15-07, 25 de junio de 2007 (www.wired.com).

THURM, Scott, "Next Frontier in Credit Scores: Predicting Personal Behavior", *The Wall Street Journal*, 27 de octubre de 2011 (online.wsj.com).

TSOTSIS, Alexia, "Twitter Is at 250 Million Tweets per Day, iOS 5Integration Made Signups Increase 3x", *TechCrunch*, 17 de octubre de 2011 (techcrunch.com).

VALERY, Nick, "Tech.View: Cars and Software Bugs", *The Economist*, 16 de mayo de 2010 (www.economist.com).

VLAHOS, James, "The Department Of Pre-Crime", *Scientific American* núm. 306, enero de 2012, pp. 62-67.

VON BAEYER, Hans Christian. *Information: The New Language of Science*, Harvard University Press, 2004.

VON AHN, Luis, *et al.*, "RECAPTCHA: Human-Based Character Recognition via Web Security Measures", *Science*, vol. 321, núm. 5895, 12 de septiembre de 2008, pp. 1465-1468 (www.sciencemag.org).

WATTS, Duncan, *Everything Is Obvious Once You Know the Answer: How Common Sense Fails Us*, Atlantic, 2011.

WEINBERGER, David, *Everything Is Miscellaneous: The Power of the New Digital Disorder*, Times, 2007.

WEINBERGER, Sharon, "Intent to Deceive. Can the Science of Deception Detection Help to Catch Terrorists?", *Nature*, vol. 465, 27 de mayo de 2010, pp. 412-415 (www.nature.com).

—, "Terrorist 'Pre-crime' Detector Field Tested in United States", *Nature*, 27 de mayo de 2011 (www.nature.com/).

WHITEHOUSE, David, "UK Science Shows Cave Art Developed Early", *BBC NEWS ONLINE*, 3 de octubre de 2001 (news.bbc.co.uk).

WIGNER, Eugene, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences", *Communications on Pure and Applied Mathematics*, vol. 13, núm. 1 (1960), pp. 1-14.

WILKS, Yorick, *Machine Translation: Its Scope and Limits*, Springer, 2009.

WINGFIELD, Nick, "Virtual Products, Real Profits: Players Spend on Zynga's Games, but Quality Turns Some Off", *The Wall Street Journal*, 9 de septiembre de 2011 (online.wsj.com).

VIKTOR MAYER-SCHÖNBERGER (Zell am See, Austria - 1966) es profesor de Internet Governance and Regulation at the Oxford Internet Institute, University of Oxford que lleva a cabo la investigación sobre la economía de la red. Anteriormente estuvo en Harvard's John F. Kennedy School of Government. Es co-autor de *Big Data. La revolución de los datos masivos* (HMH, 2013) y autor de *Delete: The Virtue of Forgetting in the Digital Age* (Princeton, 2009), que obtuvo el premio Marshall McLuhan en 2010 como libro excepcional y el Don K. Price Award for Best Book in Science and Technology Politics (2010). Ha escrito más de un centenar de artículos y capítulos para libros.

Su madre era dueña de un cine de barrio. Después de salir de la escuela secundaria en su ciudad natal, estudió derecho en la Universidad de Salzburgo . Durante este tiempo, compitió con éxito en International Physics Olympiad para jóvenes programadores concurso.

En 1986, fundó Ikarus Software que desarrolla utilidades para virus informáticos. Posteriormente obtuvo una licenciatura en Derecho en la Universidad de Salzburgo y en la Escuela de Derecho de Harvard. En 1992 obtuvo un máster (Econ) de la Escuela de Economía de Londres , y en 2001 la venia docendi en leyes (entre otros) e información en la Universidad de Graz. También trabajó en el negocio de contabilidad de su padre.

En 1998 estuvo en la facultad Harvard Kennedy School, donde trabajó y enseñó durante diez años. Después, tres años en Lee Kuan Yew School of Public Policy, National University of Singapore. Mayer-Schönberger es actualmente presidente de Internet Governance and Regulation at the Oxford Internet Institute. Como experto en leyes de información y regulación, ha asesorado a empresas, gobiernos y organizaciones internacionales.

KENNETH CUKIER es editor senior de *The Economist* responsable de datos y productos digitales (web, aplicaciones y desarrollo de nuevos productos). Fue corresponsal en Tokio (2007-2012), y antes, corresponsal de tecnología del periódico en Londres, donde su trabajo se centró en la innovación, la propiedad intelectual y el gobierno de Internet. Es co-autor de *Big Data. La revolución de los datos masivos* junto a Viktor Mayer-Schönberger (2013), traducido a más de 20 idiomas. El libro ganó el premio The book won the National Library of China's Wenjin Book Award, y fue finalista para el FT Business Book of the Year. En 2014 publicaron un trabajo sobre el seguimiento y aprendizaje con grandes volúmenes de datos: Learning with Big Data: The Future of Education.

Anteriormente, Cukier fue el editor de tecnología de *The Wall Street Journal* (Asia) en Hong Kong y comentarista

habitual en la *CNBC* (Asia). También fue editor europeo de *Red Herring* así como en *The International Herald Tribune* en París. De 2002 a 2004 trabajó de investigador en la Harvard's Kennedy School of Government, sobre Internet y las relaciones internacionales.

Sus escritos han aparecido en *The New York Times*, *The Washington Post*, *Prospect*, *The Financial Times* y *Foreign Affairs*, entre otros . Ha sido comentarista habitual sobre asuntos de negocios y tecnología para la *CBS*, *CNN*, *NPR*, la *BBC* entre otros. Es miembro del World Economic Forum's global agenda council on data-driven development.

Cukier es miembro de International Bridges to Justice, una ONG con sede en Ginebra que promueve los derechos legales en los países en desarrollo. También es director de The Open String, que proporciona instrumentos de cuerda a los niños desfavorecidos. Además, es miembro de la junta de asesores de Daniel Pearl Foundation y miembro de Council on Foreign Relations.

Notas

[1] **Google Flu Trends**: Jeremy Ginsburg *et al.*, "Detecting Influenza Epidemics Using Search Engine Query Data", *Nature*, núm. 457, 2009, pp. 1012-1014 (www.nature.com). <<

[2] **Estudio de Google Flu Trends por investigadores de la Johns Hopkins**: A. F. Dugas *et al.*, "Google Flu Trends: Correlation with Emergency Department Influenza Rates and Crowding Metrics", CID Advanced Access, 8 de enero de 2012; DOI 10.1093/cid/cir883. <<

[3] **Adquisición de billetes de avión, Farecast**: la información proviene de Kenneth Cukier, “Data, Data Everywhere”, en *The Economist*, 27 de febrero de 2010, pp. 1-14, y de entrevistas con Etzioni celebradas entre 2010 y 2012. <<

[4] **Proyecto Hamlet de Etzioni**: Oren Etzioni, C. A. Knoblock, R. Tuchinda, y A. Yates, “To Buy or Not to Buy: Mining Airfare Data to Minimize Ticket Purchase Price”, SIGKDD’03, 24 a 27 de agosto de 2003 (knight.cis.temple.edu [PDF]). <<

[5] **Precio pagado por Microsoft por la compra de Farecast**: varios artículos en los medios de comunicación, en particular “Secret Farecast Buyer Is Microsoft”, en Seattlepi.com, 17 de abril de 2008 (blog.seattlepi.com). <<

[6] **Una forma de pensar sobre los datos masivos**: hay un debate intenso e improductivo acerca del origen del término *big data* y cómo definirlo a la perfección. Las dos palabras han aparecido juntas ocasionalmente desde hace décadas. Un artículo de investigación de 2001, escrito por Doug Laney de Gartner, definió las “tres Vs” de los datos masivos (volumen, velocidad, y variedad), lo que resultó útil en su día, aunque imperfecto. <<

[7] **Astronomía y secuenciación de ADN**: Cukier, “Data, Data Everywhere”. <<

[8] **A diario cambian de manos siete mil millones de acciones**: Rita Nazareth y Julia Leite, “Stock Trading in U.S. Falls to Lowest Level Since 2008”, *Bloomberg*, 13 de agosto de 2012 (www.bloomberg.com/).

Google 24 petabytes: Thomas H. Davenport, Paul Barth, y Randy Bean, “How ‘Big Data’ Is Different,” *Sloan Review*, 30 de julio de 2012, pp. 43-46 (sloanreview.mit.edu/).

Estadísticas de Facebook: folleto de oferta de suscripción pública de acciones de Facebook, “Form S-1 Registration Statement”, U.S. Securities And Exchange Commission, 1 de febrero de 2012 (sec.gov).

Estadísticas de YouTube: Larry Page, “Update from the CEO”, Google, abril de 2012 (investor.google.com).

Número de tuits: Tomio Geron, “Twitter’s Dick Costolo: Twitter Mobile Ad Revenue Beats Desktop on Some Days”, *Forbes*, 6 de junio de 2012 (www.forbes.com). <<

[9] **Información sobre la cantidad de datos**: Martin Hilbert y Priscilla López, “The World’s Technological Capacity to Store, Communicate, and Compute Information”, *Science*, 1 de abril de 2011, pp. 60-65; Martin Hilbert y Priscilla López, “How to Measure the World’s Technological Capacity to Communicate, Store and Compute Information?”, *International Journal of Communication* 2012, pp. 1042-1055 (ijoc.org/).

Estimación de la cantidad de información almacenada hasta 2013: entrevista de Hilbert por Cukier. <<

[10] **La imprenta y ocho millones de libros**; más que lo producido desde la fundación de Constantinopla: Elizabeth L. Eisenstein, *The Printing Revolution in Early Modern Europe*, Cambridge: Canto/Cambridge University Press, 1993, pp. 13-14.

Analogía de Peter Norvig: de las charlas de Norvig basadas en el artículo de A. Halevy, P. Norvig, y F. Pereira, "The Unreasonable Effectiveness of Data", *IEEE INTELLIGENT SYSTEMS*, marzo-abril de 2009, pp. 8-12. (Obsérvese que el título es un guiño al del famoso artículo de Eugene Wigner, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences", en el que considera por qué la física puede expresarse claramente con matemáticas básicas, pero las ciencias humanas se resisten a fórmulas de esa elegancia. Véase: E. Wigner, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences", *Communications on Pure and Applied Mathematics*, vol. 13, núm. 1, 1960, pp. 1-14. Entre las charlas de Norvig sobre el artículo se halla: "Peter Norvig: The Unreasonable Effectiveness of Data", conferencia en la universidad de Columbia Británica, vídeo en YouTube, 23 de septiembre de 2010 (www.youtube.com). <<

[11] **Picasso comenta sobre las imágenes de Lascaux**: David Whitehouse, “UK Science Shows Cave Art Developed Early”, *BBC NEWS ONLINE*, 3 de octubre de 2001 (news.bbc.co.uk).

Acerca de cómo afecta el tamaño físico a las leyes físicas operativas (aunque no es del todo correcto), la referencia citada más a menudo es J. B. S. Haldane, “On Being the Right Size”, *Harper’s Magazine*, marzo de 1926 (harpers.org/). <<

[12] **Sobre Jeff Jonas y cómo hablan los datos masivos**: conversación con Jeff Jonas, diciembre de 2010, París. <<

[13] **Historia del censo de Estados Unidos**: U.S. Census Bureau, “The Hollerith Machine”, texto online (www.census.gov). <<

[14] **Contribución de Neyman**: William Kruskal y Frederick Mosteller, “Representative Sampling, IV: The History of the Concept in Statistics, 1895-1939”, *International Statistical Review*, 48, 1980, pp. 169-195, pp. 187-188. El famoso artículo de Neyman es: Jerzy Neyman, “On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection”, *Journal of the Royal Statistical Society*, vol. 97, núm. 4, 1934, pp. 558-625.

Basta con una muestra de 1100 observaciones: Earl Babbie, *Practice of Social Research*, 12.ª ed., 2010, pp. 204-207. <<

[15] **Efecto del teléfono móvil**: “Estimating the Cellphone Effect”, 20 de septiembre de 2008 (www.fivethirtyeight.com); para saber más acerca de los sesgos en las muestras y otras revelaciones estadísticas, véase: Nate Silver, *The Signal and the Noise: Why So Many Predictions Fail, But Some Don’t*, Penguin, 2012. <<

[16] **Secuenciación de los genes de Steve Jobs**: Walter Isaacson, *Steve Jobs*, Simon and Schuster, 2011, pp. 550-551. <<

[17] **Google Flu Trends predicción a nivel de ciudad**: Dugas *et al.*, “Google Flu Trends”.

Etzioni acerca de los datos temporales: Entrevista por Cukier, octubre de 2011. <<

[18] **Cita de John Kunze**: Jonathan Rosenthal, “Special Report: International Banking”, *The Economist*, 19 de mayo de 2012, pp. 7-8.

Combates de sumo amañados: Mark Duggan y Steven D. Levitt, “Winning Isn’t Everything: Corruption in Sumo Wrestling”, *American Economic Review* 92, 2002, pp. 1594-1605 (pricetheory.uchicago.edu [PDF]). <<

[19] **Los once millones de rayos de luz de Lytro**: extraído de la web corporativa de Lytro (http://www.lytro.com).
<<

[20] **Sustitución del muestreo en las ciencias sociales**: Mike Savage y Roger Burrows, "The Coming Crisis of Empirical Sociology", *Sociology*, Vol. 41, núm. 5, 2007, pp. 885-899.

Análisis de datos ingentes de un operador de telefonía móvil: J.-P. Onnela *et al.*, "Structure and Tie Strengths in Mobile Communication Networks", *Proceedings of the National Academy of Sciences of the United States of America (pnas)*, 104, mayo de 2007, pp. 7332-7336 (nd.edu [PDF]). <<

[21] **Crosby**: Alfred W. Crosby, *The Measure of Reality: Quantification and Western Society 1250-1600*, Cambridge University Press, 1997. <<

[22] **Acerca de las citas de lord Kelvin y Bacon**: estos aforismos son ampliamente atribuidos a ambos hombres, aunque la formulación exacta en sus obras publicadas es ligeramente diferente. En el caso de Kelvin, forma parte de una cita más extensa sobre la medición, tomada de su conferencia “Electrical Units of Measurement”, 1883. En cuanto a Bacon, se suele considerar una traducción aproximada del latín, de *Meditationes Sacrae*, 1597. <<

[23] **Muchas formas de referirse a IBM**: DJ Patil, “Data Jujitsu: The Art of Turning Data into Product”, *O’Reilly Media*, julio de 2012 (oreillynet.com). <<

[24] **30 000 transacciones por segundo en la bolsa de Nueva York**: Colin Clark, "Improving Speed and Transparency of Market Data", de 9 de enero de 2011 en el blog NYSE EURONEXT, (exchanges.nyx.com).

La idea de que "2+2 = 3,9": Brian Hopkins y Boris Evelson, "Expand Your Digital Horizon with Big Data", *Forrester*, 30 de septiembre de 2011. <<

[25] **Mejoras en los algoritmos**: President's Council of Advisors on Science and Technology [Consejo de asesores del Presidente sobre ciencia y tecnología], *Report to the President and Congress, Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*, diciembre de 2010, p. 71 (www.whitehouse.gov [PDF]).

Tablas de fases finales de partidas de ajedrez: La tabla más amplia de finales de partida disponible públicamente, la "Nalimov tableset" (así llamada por uno de sus creadores), recoge todas las jugadas con seis o menos piezas. Su tamaño supera los siete terabytes, y comprimir la información que contiene es todo un reto. Véase E. V. Nalimov, G. McC. Haworth, y E. A. Heinz, "Space-efficient Indexing of Chess Endgame Tables", *ICGA JOURNAL*, vol. 23, núm. 3, 2000, pp. 148-162.

Microsoft y rendimientos de algoritmos: Michele Banko y Eric Brill, "Scaling to Very Very Large Corpora for Natural Language Disambiguation", *Microsoft Research*, 2001, p. 3 (acl.ldc.upenn.edu [PDF]). <<

[26] **Demo de IBM, palabras y cita**: IBM, “701 Translator”, comunicado de prensa, archivos de IBM, 8 de enero de 1954 (www-03.ibm.com). Véase asimismo John Hutchins, “The First Public Demonstration of Machine Translation: The Georgetown-IBM System, 7th January 1954”, noviembre de 2005 (www.hutchinsweb.me.uk [PDF]). <<

[27] **Programa Candide de IBM**: Adam L. Berger *et al.*, "The Candide System for Machine Translation", *Proceedings of the 1994 arpa Workshop on Human Language Technology*, 1994 (aclweb.org [PDF]).

Historia de la traducción automática: Yorick Wilks, *Machine Translation: Its Scope and Limits*, Springer, 2008, p. 107. <<

[28] **Los millones de textos de Candide frente a los miles de millones de Google**: entrevista a Och por Cukier, diciembre de 2009.

El corpus de 95 000 millones de frases de Google: Alex Franz y Thorsten Brants, “All Our N-gram are Belong to You”, post de 3 de agosto de 2006 en Google (googleresearch.blogspot.co.uk). <<

[29] **Corpus de Brown y el billón de palabras de Google**: Halevy, Norvig, y Pereira, “The Unreasonable Effectiveness of Data”.

Cita del artículo coescrito por Norvig: *ibid*. <<

[30] **BP, corrosión de tuberías y entorno hostil a las comunicaciones inalámbricas**: Jaclyn Clarabut, “Operations Making Sense of Corrosion”, *BP MAGAZINE*, núm. 2, 2011 (www.bp.com [PDF]). La dificultad de las lecturas inalámbricas de datos está tomada de: Cukier, “Data, Data, Everywhere”. El sistema, obviamente, no es infalible: un incendio en la refinería BP de Cherry Point en febrero de 2012 fue atribuido a una tubería corroída. <<

[31] **Proyecto de los Mil Millones de Precios**: Entrevista de Cukier a los cofundadores. Véase asimismo: James Surowiecki, “A Billion Prices Now”, *The New Yorker*, 30 de mayo de 2011; pueden hallarse datos y detalles en la página web del proyecto (http://bpp.mit.edu/); Annie Lowrey, “Economists’ Programs Are Beating U.S. at Tracking Inflation”, *The Washington Post*, 25 de diciembre de 2010 (washingtonpost.com). <<

[32] **Acerca de PriceStats como verificador de las estadísticas nacionales**: “Official Statistics: Don’t Lie to Me, Argentina”, *The Economist*, 25 de febrero de 2012 (www.economist.com). <<

[33] **Número de fotos en Flickr**: información tomada de la página web de Flickr (www.flick.com).

Sobre el reto que supone clasificar la información, véase: David Weinberger, *Everything Is Miscellaneous: The Power of the New Digital Disorder*, Times, 2007. <<

[34] **Pat Helland**: Pat Helland, “If You Have Too Much Data Then ‘Good Enough’ Is Good Enough”, *Communications of the ACM*, junio de 2011, pp. 40-41. Existe un vivo debate en el seno de la comunidad de las bases de datos acerca de los modelos y conceptos mejor adaptados a las necesidades de los datos masivos. Helland representa al bando que apuesta por una ruptura radical con las herramientas empleadas en el pasado. Michael Rys, de Microsoft, sostiene en “Scalable SQL”, *Communications of the acm*, junio de 2011, p. 48, que versiones muy adaptadas de las herramientas ya existentes funcionarán de maravilla. <<

[35] **VISA utiliza Hadoop**: Cukier, “Data, data everywhere”. <<

[36] **Sólo el 5 por 100 de las bases de datos están bien estructuradas**: Abhishek Mehta, “Big Data: Powering the Next Industrial Revolution”, Tableau Software White Paper, 2011 (www.tableausoftware.com). <<

[37] **Información sobre Linden, así como la "voz de Amazon"**: entrevista de Linden por Cukier, marzo de 2012.

***The Wall Street Journal* sobre los críticos de Amazon:** citado en James Marcus, *Amazonia: Five Years at the Epicenter of the Dot.Com Juggernaut* (New Press, 2004), p. 128. <<

[38] **Cita de Marcus**: Marcus, *Amazonia*, p. 199. <<

[39] **Las recomendaciones suponen la tercera parte de los ingresos de Amazon**: esta cifra nunca ha sido confirmada oficialmente por la compañía, pero ha aparecido en numerosos informes de analistas y artículos en la prensa, incluyendo "Building with Big Data: The Data Revolution Is Changing the Landscape of Business", *The Economist*, 26 de mayo de 2011 (www.economist.com). La cifra fue mencionada asimismo por dos antiguos directivos de Amazon en sendas entrevistas con Cukier.

Información de precios de Netflix: Xavier Amatriain y Justin Basilico, "Netflix Recommendations: Beyond the 5 stars (Part 1)", Netflix blog, 6 de abril de 2012. <<

[40] **Dejarse “engañar por la aleatoriedad”**: Nassim Nicholas Taleb, *Fooled by Randomness*, Random House, 2008; para más información, véase Nassim Nicholas Taleb, *The Black Swan: The Impact of the Highly Improbable*, 2.ª ed., Random House, 2010. <<

[41] **Walmart y las Pop-Tarts**: Constance L. Hays, “What Wal-Mart Knows About Customers’ Habits”, *The New York Times*, 14 de noviembre de 2004 (www.nytimes.com/). <<

[42] **Ejemplos de modelos predictivos de FICO, Experian y Equifax**: Scott Thurm, “Next Frontier in Credit Scores: Predicting Personal Behavior”, *The Wall Street Journal*, 27 de octubre de 2011 (online.wsj.com/).

Modelos predictivos de Aviva: Leslie Scism y Mark Maremont, “Insurers Test Data Profiles to Identify Risky Clients”, *The Wall Street Journal*, 19 de noviembre de 2010 (online.wsj.com/). Véase asimismo: Leslie Scism y Mark Maremont, “Inside Deloitte’s Life-Insurance Assessment Technology”, *The Wall Street Journal*, 19 de noviembre de 2010 (online.wsj.com). Véase también: Howard Mills, “Analytics: Turning Data into Dollars”, *Forward Focus*, diciembre de 2011 (www.deloitte.com [PDF]). <<

[43] **Ejemplo de Target y la adolescente embarazada**: Charles Duhigg, “How Companies Learn Your Secrets”, *The New York Times*, 16 de febrero de 2012 www.nytimes.com). El artículo está adaptado del libro de *Duhigg The Power of Habit: Why We Do What We Do in Life and Business*, Random House, 2012; Target ha declarado que hay inexactitudes en los informes periodísticos sobre sus actividades, pero no ha querido precisar cuáles son esas supuestas inexactitudes. Al ser preguntados al respecto para este libro, un portavoz de Target comentó: “La meta es usar los datos del cliente para intensificar su relación con Target. Nuestros clientes desean recibir un valor elevado, ofertas relevantes, y una experiencia superior. Como muchas otras compañías, hacemos uso de herramientas de investigación para ayudarnos a entender las tendencias y preferencias de compra de la clientela, para poder brindarles a nuestros clientes ofertas y promociones que sean relevantes para ellos. Nos tomamos muy en serio la responsabilidad de proteger la confianza de nuestros clientes hacia la empresa. Una de las formas que tenemos de hacerlo es aplicando una política de privacidad muy amplia, que compartimos abiertamente en Target.com, y ofreciendo formación periódica a los miembros de nuestro equipo acerca de cómo proteger la información de los clientes”. <<

[44] **Analítica predictiva de UPS**: entrevistas de Cukier a Jack Levis, marzo, abril, y julio de 2012. <<

[45] **Prematuros**: basado en entrevistas a McGregor in enero de 2010 y abril y julio de 2012. Véase asimismo: Carolyn McGregor, Christina Catley, Andrew James, y James Padbury, "Next Generation Neonatal Health Informatics with Artemis", en European Federation for Medical Informatics, *User Centred Networked Health Care*, editado por A. Moen *et al.*, IOS Press, 2011, p. 117. Algunos detalles están sacados de Cukier, "Data, Data, Everywhere". <<

[46] **Sobre la correlación entre la felicidad y el nivel de renta**: R. Inglehart y H.-D. Klingemann, *Genes, Culture and Happiness*, MIT Press, 2000.

Acerca del sarampión y los gastos en atención sanitaria, así como las nuevas herramientas no lineales para el análisis de correlación: David Reshef *et al.*, "Detecting Novel Associations in Large Data Sets", *Science* núm. 334, 2011, pp. 1518-1524. <<

[47] **Kahneman**: Daniel Kahneman, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011, pp. 74-75. <<

[48] **Pasteur**: a los lectores interesados en la considerable influencia de Pasteur en nuestra manera de percibir las cosas, les recomendamos la lectura de Bruno Latour *et al.*, *The Pasteurization of France*, Harvard University Press, 1993.

Riesgo de contraer la rabia: Melanie Di Quinzio y Anne McCarthy, “Rabies Risk Among Travellers”, *CMAJ*, vol.178, núm. 5, 2008, p. 567. <<

[49] **Raras veces puede demostrarse la causalidad**: Judea Pearl, científica de la computación galardonada con un Premio Turing, ha desarrollado una nueva manera de representar formalmente la dinámica causal; aunque dista de constituir una prueba formal, esto supone una aproximación pragmática al análisis de posibles conexiones causales. Véase: Judea Pearl, *Models, Reasoning and Inference*, 2009. <<

[50] **Ejemplo del coche naranja**: Quentin Hardy. “Bizarre Insights from Big Data”, nytimes.com, 28 de marzo de 2012 (http://bits.blogs.nytimes.com/2012/03/28/bizarre-insights-from-big-data/); y Kaggle: “Momchil Georgiev Shares His Chromatic Insight from Don’t Get Kicked”, *post* en el blog, 2 de febrero de 2012 (http://blog.kaggle.com/2012/02/02/momchil-georgiev-shares-his-chromatic-insight-from-dont-get-kicked/). <<

[51] **Peso de las tapas de registro eléctrico, número de explosiones y altitud alcanzada por el estallido**: Rachel Ehrenberg, "Predicting the Next Deadly Manhole Explosion", *WIRED*, 7 de julio de 2010 (http://www.wired.com/wiredscience/2010/07/manhole-explosions). <<

[52] **Con Edison y los estadísticos de la universidad de Columbia**. Este caso aparece descrito para un público no especialista en: Cynthia Rudin *et al.*, "21st-Century Data Miners Meet 19th-Century Electrical Cables", *Computer*, junio de 2011, pp. 103-105. Las descripciones técnicas del trabajo están disponibles en los artículos académicos de Rudin y sus colaboradores en sus páginas web, en particular, Cynthia Rudin *et al.*, "Machine Learning for the New York City Power Grid", en *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 34, núm. 2, 2012, pp. 328-345 (hdl.handle.ne).

Confusión en torno al término "caja de acometida" (service box): la lista está tomada de Rudin *et al.*, "21st-Century Data Miners Meet 19th-Century Electrical Cables".

Cita de Rudin: entrevista con Cukier, marzo de 2012. <<

[53] **Punto de vista de Anderson**: Chris Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, *Wired*, junio de 2008 (http://www.wired.com/science/discoveries/magazine/16-07/pb_theory/). <<

[54] **Marcha atrás de Anderson**: National Public Radio, “Search and Destroy”, 18 de julio de 2008 (www.onthemedia.org). <<

[55] **Sobre las elecciones que influyen en nuestro análisis**: danah boyd y Kate Crawford, “Six Provocations for Big Data”, ponencia presentada en el simposio del Oxford Internet Institute “A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society”, el 21 de septiembre de 2011 (http://ssrn.com/abstract=1926431). <<

[56] **La información sobre la vida de Maury** está reunida en numerosos trabajos escritos por él y sobre él. Entre ellos, Chester G. Hearn, *Tracks in the Sea: Matthew Fontaine Maury and the Mapping of the Oceans*, International Marine/McGraw-Hill, 2002; Janice Beaty, *Seeker of Seaways: A Life of Matthew Fontaine Maury, Pioneer Oceanographer*, Pantheon Books, 1966; Charles Lee Lewis, *Matthew Fontaine Maury: The Pathfinder of the Seas*, U.S. Naval Institute, 1927 (archive.org); y Matthew Fontaine Maury, *The Physical Geography of the Sea*, Harper, 1855. <<

[57] **Citas de Maury**: tomadas de Maury, *Physical Geography of the Sea*, “Introduction”, pp. XII, vi. <<

[58] **Datos de los asientos de los coches**: Nikkei, “Car Seat of Near Future IDs Driver’s Backside”, 14 de diciembre de 2011. <<

[59] **Cuantificar el mundo**: en buena medida, el pensamiento de los autores acerca de la historia de la datificación ha sido inspirado por Crosby, *The Measure of Reality*. <<

[60] **Los europeos nunca se vieron expuestos al ábaco**: *Ibid.*, p.112.

Calcular seis veces más deprisa con guarismos arábigos que con tableros de recuento: Alexander Murray, *Reason and Society in the Middle Ages*, Oxford University Press, 1978, p. 166. <<

[61] **Total de libros publicados en el mundo, y estudio de Harvard sobre el proyecto de escaneado de libros de Google**: Jean-Baptiste Michel *et al.*, "Quantitative Analysis of Culture Using Millions of Digitized Books", *Science*, núm. 331, 14 de enero de 2011, pp. 176-182 (http://www.sciencemag.org/content/331/6014/176.abstract). Para una vídeo conferencia sobre el artículo, véase Erez Lieberman Aiden y Jean-Baptiste Michel, "What We Learned from 5 Million Books", TEDx, Cambridge, Massachusetts, 2011 (www.ted.com). <<

[62] **Acerca de los módulos inalámbricos en coches y las primas de seguros**: Véase Cukier, “Data, Data Everywhere”. <<

[63] **Jack Levis de UPS**: Entrevista por Cukier, abril de 2012.

Datos acerca de los ahorros de UPS: Institute for Operations Research and the Management Sciences (INFORMS), "UPS Wins Gartner BI Excellence Award," 2011 (www.informs.org). <<

[64] **Investigaciones de Pentland**: Robert Lee Hotz, “The Really Smart Phone”, *The Wall Street Journal*, 22 de abril de 2011 (online.wsj.com).

Estudio de Eagle de los suburbios: Nathan Eagle, “Big Data, Global Development, and Complex Systems”, Santa Fe Institute, 5 de mayo de 2010 (www.youtube.com [Vídeo]). Asimismo, entrevista por Cukier, octubre de 2012. <<

[65] **Datos de Facebook**: tomados del folleto de oferta de suscripción pública de acciones de Facebook, 2012 (sec.gov).

Datos de Twitter: Alexia Tsotsis, "Twitter Is at 250 Million Tweets per Day, iOS 5 Integration Made Signups Increase 3x", *TechCrunch*, 17 de octubre de 2011 (techcrunch.com).

Uso de Twitter por fondos de inversión: Kenneth Cukier, "Tracking Social Media: The Mood of the Market", *Economist.com*, 28 de junio de 2012 (economist.com). <<

[66] **Twitter y la predicción de ingresos en taquilla en Hollywood**: Sitaram Asur y Bernardo A. Huberman, "Predicting the Future with Social Media", *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 492-499; versión online disponible en: www.hpl.hp.com [PDF].

Twitter y estados de ánimo globales: Scott A. Golder y Michael W. Macy, "Diurnal and Seasonal Mood Vary with Work, Sleep, and Day-length Across Diverse Cultures", *Science*, núm. 333, 30 de septiembre de 2011, pp. 1878-1881. <<

[67] **Twitter y la vacunación contra la gripe**: Marcel Salathé y Shashank Khandelwal, “Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control”, *PLoS Computational Biology*, octubre de 2011.

Patente de “suelo inteligente” de IBM: Lydia Mai Do, Travis M. Grigsby, Pamela Ann Nesbitt, y Lisa Anne Seacat, “Securing Premises Using Surface-Based Computing Technology”, patente estadounidense, núm. 8138882, otorgada el 5 de febrero de 2009. <<

[68] **Movimiento del “ser cuantificado”**: “Counting Every Moment”, *The Economist*, 3 de marzo de 2012.

Auriculares Apple para biomediciones: Jesse Lee Dorogusker, Anthony Fadell, Donald J. Novotney, y Nicholas R Kalayjian, “Integrated Sensors for Tracking Performance Metrics”, solicitud de patente estadounidense 20090287067. Solicitante: Apple. Fecha de solicitud: 23 de julio de 2009. Fecha de publicación: 19 de noviembre de 2009. <<

[69] **Derawi Biometrics**, “Your Walk Is Your PIN-Code”, comunicado de prensa, 21 de febrero de 2011 (biometrics.derawi.com).

Información sobre iTrem: véase la página web del proyecto iTrem en el Landmarc Research Center de Georgia Tech (eosl.gtri.gatech.edu), más intercambio de correos electrónicos.

Investigadores de Kyoto sobre acelerómetros triaxiales: iMedicalApps Team, “Gait Analysis Accuracy: Android App Comparable to Standard Accelerometer Methodology”, *mHealth*, 23 de marzo de 2012. <<

[70] **Los periódicos permitieron el auge del estado-nación**: Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism*, Verso, 2006.

Los físicos sugieren que la información es la base de todo: Hans Christian von Baeyer, *Information: The New Language of Science*, Harvard University Press, 2005. <<

[71] **Historia de Luis von Ahn**: basado en entrevistas de Von Ahn por Cukier en 2010 y 2011. Véase asimismo: Clive Thompson, "For Certain Tasks, the Cortex Still Beats the CPU", *Wired*, 25 de junio de 2007 (www.wired.com); Jessie Scanlon, "Luis von Ahn: The Pioneer of 'Human Computation'", *Businessweek*, 3 de noviembre de 2008 (www.businessweek.com/). Su descripción técnica de los reCaptchas se halla en: Luis von Ahn *et al.*, "ReCAPTCHA: Human-Based Character Recognition via Web Security Measures", *Science*, núm. 321, 12 de septiembre de 2008, pp. 1465-1468 (www.sciencemag.org). <<

[72] **Ejemplo del fabricante de alfileres, de Smith**: Adam Smith, *The Wealth of Nations*, reimpresión, Bantam Classics, 2003, libro I, capítulo 1 (**versión electrónica gratuita disponible en**: www2.hn.psu.edu [PDF]).

Almacenamiento: Viktor Mayer-Schönberger, *Delete: The Virtue of Forgetting in the Digital Age*, Princeton University Press, 2011, p. 63. <<

[73] **Sobre el consumo de energía de los coches eléctricos**: IBM, "IBM, Honda, and PG&E Enable Smarter Charging for Electric Vehicles", comunicado de prensa, 12 de abril de 2012 (www-03.ibm.com). Véase asimismo: Clay Luthy, "Guest Perspective: IBM Working with PG&E to Maximize the EV Potential", *PGE CURRENTS MAGAZINE*, 13 de abril de 2012 (www.pgecurrents.com). <<

[74] **Amazon y los datos de AOL**: entrevistas de Cukier a Andreas Weigend, 2010 y 2012. <<

[75] **Programa Nuance y Google**: Cukier, “Data, Data Everywhere”. <<

[76] **Empresa logística**: Brad Brown, Michael Chui, y James Manyika, “Are You Ready for the Era of ‘Big Data’?”, *McKinsey Quarterly*, octubre de 2011, p. 10. <<

[77] **Telefónica comercializa información de móviles**: “Telefonica Hopes ‘Big Data’ Arm Will Revive Fortunes”, *BBC ONLINE*, 9 de octubre de 2012. (www.bbc.co.uk).

Estudio de la Asociación Danesa contra el Cáncer: Patrizia Frei *et al.*, “Use of Mobile Phones and Risk of Brain Tumours: Update of Danish Cohort Study”, *BMJ*, núm. 343, 2011 (www.bmj.com), y entrevista por Cukier, octubre de 2012. <<

[78] **Google, registros de los GPS de Street View y coche autoconducido**: Peter Kirwan, “This Car Drives Itself”, *Wired uk*, enero de 2012 (www.wired.co.uk). <<

[79] **Corrector ortográfico de Google y citas**: entrevista celebrada por Cukier en el Googleplex, Mountain View, California, diciembre de 2009; algún material aparece también en Cukier, “Data, Data Everywhere”. <<

[80] **Intuición de Hammerbacher**: entrevista con Cukier, octubre de 2012. <<

[81] **Barnes & Noble, análisis de datos de su libro electrónico Nook**: Alexandra Alter, "Your E-Book Is Reading You", *The Wall Street Journal*, 29 de junio de 2012 (online.wsj.com).

Clase Coursera de Andrew Ng y datos: entrevista por Cukier, junio de 2012. <<

[82] **Política de gobierno abierto de Obama**: Barack Obama, “Presidential memorandum”, Casa Blanca, 21 de enero de 2009. <<

[83] **Acerca del valor de los datos de Facebook**: puede hallarse un excelente análisis de la discrepancia entre el valor de mercado y el valor contable de la oferta inicial al público de las acciones de Facebook en Doug Laney, "To Facebook You're Worth $80.95", *The Wall Street Journal*, 3 de mayo de 2012 (blogs.wsj.com). Para hacer una valoración de la empresa, Laney extrapoló a partir del crecimiento de Facebook para estimar los 2,1 billones de unidades de "contenido monetizable". En su artículo del *Wall Street Journal* valoró las unidades a tres centavos cada una, porque estaba empleando una estimación anterior del valor de mercado de Facebook, de 75 000 millones de dólares. Al final, el valor superaba los 100 000 millones de dólares, o cinco centavos la unidad, según nuestra propia extrapolación basada en su cálculo. <<

[84] **Desfase en el valor de los activos físicos e intangibles**: Steve M. Samek, “Prepared Testimony: Hearing on Adapting a 1930’s Financial Reporting Model to the 21st Century”, Comité del Senado de Estados Unidos sobre la Banca, la Vivienda y los Asuntos Urbanos, Subcomité sobre Valores, 19 de Julio de 2000. <<

[85] **Valor de los intangibles**: Robert S. Kaplan y David P. Norton, *Strategy Maps: Converting Intangible Assets into Tangible Outcomes*, Harvard Business Review Press, 2004, pp. 4-5. <<

[86] **Cita de Tim O’Reilly**: de una entrevista por Cukier, febrero de 2011. <<

[87] **Información sobre Decide.com**: cruce de correos electrónicos entre Cukier y Etzioni, mayo de 2012. <<

[88] **Informe de McKinsey**: James Manyika *et al.*, “Big Data: The Next Frontier for Innovation, Competition, and Productivity,” McKinsey Global Institute, mayo de 2011 (www.mckinsey.com), p. 10. <<

[89] **Cita de Hal Varian**: entrevista por Cukier, diciembre de 2009. <<

[90] **Cita de Carl de Marcken, de ITA**: intercambio de correos electrónicos con Cukier, mayo de 2012. <<

[91] **Acerca de MasterCard Advisors**: entrevista de Cukier a Gary Kearns, directivo de MasterCard Advisors, durante la conferencia “The Ideas Economy: Information” auspiciada por *The Economist*, Santa Clara, California, 8 de junio de 2011.

Accenture y la ciudad de St. Louis, Missouri: entrevistas de Cukier a empleados municipales, febrero de 2007. <<

[92] **Microsoft Amalga Unified Intelligence System**: “Microsoft Expands Presence in Healthcare IT Industry with Acquisition of Health Intelligence Software Azyxxi”, comunicado de prensa de Microsoft, 26 de julio de 2006 (www.microsoft.com/). El servicio Amalga forma parte ahora de la *joint venture* de Microsoft con General Electric llamada Caradigm. <<

[93] **Bradford Cross**: entrevistas por Cukier, marzo-octubre de 2012. <<

[94] **Amazon y el “filtrado colaborativo”**: folleto de oferta de suscripción pública de acciones, mayo de 1997 (www.sec.gov [TXT]). <<

[95] **Microprocesadores en coches**: Nick Valery, “Tech.View: Cars and Software Bugs”, *Economist.com*, 16 de mayo de 2010 (www.economist.com).

Maury llamaba a los barcos “observatorios flotantes”: Maury, *The Physical Geography of the Sea*. <<

[96] **Inrix**: entrevista de Cukier a varios directivos, mayo y septiembre de 2012. <<

[97] **Sobre el Health Care Cost Institute**: Sarah Kliff, “A Database That Could Revolutionize Health Care”, *The Washington Post*, 21 de mayo de 2012. <<

[98] **Acuerdo de uso de datos de Decide.com**: intercambio de correos electrónicos entre Cukier y Etzioni, mayo de 2012. <<

[99] **Inrix y ABS**: entrevista de Cukier a directivos de Inrix, mayo de 2012. <<

[100] **Historia de Roadnet y cita de Len Kennedy**: entrevista por Cukier, mayo de 2012.

Acuerdo de Google con ITA: Claire Cain Miller, "U.S. Clears Google Acquisition of Travel Software", *The New York Times*, 8 de abril de 2011 (www.nytimes.com).

Diálogos de la película *Moneyball*, dirigida por Bennett Miller, Columbia Pictures, 2011. <<

[101] **Los datos de McGregor representan más de una década de años/paciente**: entrevista con Cukier, mayo de 2012.

Cita de Goldbloom: de una entrevista por Cukier, marzo de 2012. <<

[102] **Acerca de los ingresos del cine frente a los de los videjuegos**: respecto a las películas, véase Brooks Barnes, "A Year of Disappointment at the Movie Box Office", *The New York Times*, 25 de diciembre de 2011 (www.nytimes.com). Sobre los videojuegos, véase "Factbox: A Look at the $65 billion Video Games Industry", Reuters, 6 de junio de 2011 (uk.reuters.com).

Analítica de datos de Zynga: Nick Wingfield, "Virtual Products, Real Profits: Players Spend on Zynga's Games, but Quality Turns Some Off", *The Wall Street Journal*, 9 de septiembre de 2011 (online.wsj.com). <<

[103] **Cita de Ken Rudin**: de una entrevista de Rudin por Niko Waesche, citada en Erik Schlie, Jörg Rheinboldt, y Niko Waesche, *Simply Seven: Seven Ways to Create a Sustainable Internet Business*, Palgrave Macmillan, 2011.

Cita de Thomas Davenport: entrevista a Davenport por Cukier, diciembre de 2009.

The-Numbers.com: entrevistas de Cukier a Bruce Nash, octubre de 2011 y julio de 2012.

Estudio de Brynjolfsson: Erik Brynjolfsson, Lorin Hitt, y Heekyung Kim, "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?", documento de trabajo, abril de 2011 (papers.ssrn.com). <<

[104] **Acerca de Rolls-Royce**: véase “Rolls-Royce: Britain’s Lonely High-Flier”, *The Economist*, 8 de enero de 2009 (www.economist.com). Cifras actualizadas con datos de la oficina de prensa, noviembre de 2012. <<

[105] **Erik Brynjolfsson, Andrew McAfee, Michael Sorell, y Feng Zhu, "Scale Without Mass**: Business Process Replication and Industry Dynamics," Harvard Business School, documento de trabajo, septiembre de 2006 (www.hbs.edu [PDF] y también hbswk.hbs.edu).

Acerca del movimiento hacia titulares de datos cada vez más grandes: véase también Yannis Bakos y Erik Brynjolfsson, "Bundling Information Goods: Pricing, Profits, and Efficiency", *Management Science*, núm. 45, diciembre de 1999, pp. 1613-1630. <<

[106] **Philip Evans**: entrevistas con los autores, 2011 y 2012. <<

[107] **Sobre la Stasi**: por desgracia, gran parte de la literatura está en alemán, pero una excepción bien documentada es Kristie Macrakis, *Seduced by Secrets: Inside the Stasi's Spy-Tech World*, Cambridge University Press, 2008; Thimoty Garton Ash comparte una historia muy personal en *The File*, Atlantic Books, 2008. Recomendamos asimismo la película ganadora de un Óscar *La vida de los otros*, dirigida por Florian Henckel von Donnersmarck, Buena Vista / Sony Pictures, 2006.

Cámaras de vigilancia cerca del apartamento de Orwell: "George Orwell, Big Brother Is Watching Your House", en *The Evening Standard*, 31 de marzo de 2007 (www.thisislondon.co.uk).

Sobre Equifax y Experian, véase: Daniel J. Solove, *The Digital Person: Technology and Privacy in the Information Age*, NYU Press, 2004, pp. 20-21. <<

[108] **Acerca de las direcciones de las manzanas de Washington en las que vivían japoneses, comunicadas a las autoridades estadounidenses**: J. R. Minkel, "The U.S. Census Bureau Gave Up Names of Japanese-Americans in WW II", *Scientific American*, 30 de marzo de 2007 (www.scientificamerican.com).

Sobre los datos usados por los nazis en los Países Bajos: William Seltzer y Margo Anderson, "The Dark Side of Numbers: The Role of Population Data Systems in Human Rights Abuses", *Social Research*, vol. 68, núm. 2, 2001, pp. 481-513.

Información sobre IBM y el Holocausto: Edwin Black, *IBM AND THE HOLOCAUST*, Crown, 2003. <<

[109] **Acerca de la cantidad de datos que recogen los contadores inteligentes, véase**: Elias Leake Quinn, *Smart Metering and Privacy: Existing Law and Competing Policies; A Report for the Colorado Public Utility Commission*, primavera de 2009 (www.w4ar.com [PDF]). Véase asimismo Joel M. Margolis, "When Smart Grids Grow Smart Enough to Solve Crimes," *Neustar*, 18 de marzo de 2010 (energy.gov [PDF]) <<

[110] **Fred Cate sobre la notificación y el consentimiento**: Fred H. Cate, "The Failure of Fair Information Practice Principles", en Jane K. Winn (ed.), *Consumer Protection in the Age of the "Information Economy"*, Ashgate, 2006, p. 341 y ss. <<

[111] **Sobre la publicación de datos de AOL**: Michael Barbaro y Tom Zeller Jr., “A Face Is Exposed for AOL Searcher No. 4417749”, *The New York Times*, 9 de agosto de 2006. Véase asimismo Matthew Karnitschnig y Mylene Mangalindan, “AOL Fires Technology Chief After Web-Search Data Scandal”, *The Wall Street Journal*, 21 de agosto de 2006. <<

[112] **Sobre la identificación de una usuaria de Netflix**: Ryan Singel, "Netflix Spilled Your Brokeback Mountain Secret, Lawsuit Claims", *Wired*, 17 de diciembre de 2009 (www.wired.com).

Acerca de la publicación de datos de Netflix: Arvind Narayanan y Vitaly Shmatikov, "Robust De-Anonymization of Large Sparse Datasets", en *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, p. 111 *y ss*. (www.cs.utexas.edu [PDF]); Arvind Narayanan y Vitaly Shmatikov, "How to Break the Anonymity of the Netflix Prize Dataset", 18 de octubre de 2006, arXiv:cs/0610105[cs.CR] (arxiv.org).

Identificar a la gente a partir de tres características: Philippe Golle, "Revisiting the Uniqueness of Simple Demographics in the US Population", *Association for Computing Machinery Workshop on Privacy in Electronic Society*, 5 (2006), p. 77.

Sobre la debilidad estructural de la anonimización: Paul Ohm, "*Broken Promises of Privacy: UCLA LAW REVIEW*", vol. 57, 2010, p.1701.

Acerca del anonimato de la gráfica social: Lars Backstrom, Cynthia Dwork, y Jon Kleinberg, "Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography", *Communications of the Association of Computing Machinery*, diciembre de 2011, p. 133. <<

[113] **Las “cajas negras” de los coches**: “Vehicle Data Recorders: Watching Your Driving”, *The Economist*, 23 de junio de 2012 (www.economist.com). <<

[114] **Recogida de datos por la NSA**: Dana Priest y William Arkin, “A Hidden World, Growing Beyond Control”, *The Washington Post*, 19 de julio de 2010 (projects.washingtonpost.com). Juan Gonzalez, “Whistleblower: The NSA Is Lying-U.S. Government Has Copies of Most of Your Emails”, *Democracy Now*, 20 de abril de 2012 (www.democracynow.org). William Binney, “Sworn Declaration in the Case of Jewel v. NSA”, archivado el 2 de julio de 2012 (publicintelligence.net/).

Cómo ha cambiado la vigilancia con los datos masivos: Patrick Radden Keefe, “Can Network Theory Thwart Terrorists?”, *The New York Times*, 12 de marzo de 2006 (www.nytimes.com). <<

[115] **Diálogos de la película *Minority Report***, dirigida por Steven Spielberg, DreamWorks/20th Century Fox, 2002. El diálogo que citamos ha sido acortado muy ligeramente. La película está basada en un relato corto de 1958 de Philip K. Dick, pero hay diferencias sustanciales entre las dos versiones. En concreto, la escena inicial con el marido cornudo no aparece en el libro, y el acertijo filosófico que plantea el PreCrimen se presenta de forma más tajante en el filme de Spielberg que en el relato. De ahí que hayamos preferido establecer nuestros paralelismos con la película. <<

[116] **Ejemplos de "policía predictiva"**: James Vlahos, "The Department Of Pre-Crime", *Scientific American*, núm. 306, enero de 2012, pp. 62-67.

Acerca de la Tecnología de Exploración de Atributos Futuros (FAST), véase: Sharon Weinberger, "Terrorist 'Pre-crime' Detector Field Tested in United States", *Nature*, 27 de mayo de 2011 (www.nature.com); Sharon Weinberger, "Intent to Deceive", *Nature*, 465, mayo de 2010, pp. 412-415. Sobre el problema de los falsos positivos, véase Alexander Furnas, "Homeland Security's 'Pre-Crime' Screening Will Never Work", *The Atlantic Online*, 17 de abril de 2012 (www.theatlantic.com). <<

[117] **Sobre las notas de los estudiantes y las primas de las pólizas de seguros**: Tim Query, "Grade Inflation and the Good-Student Discount", *Contingencies Magazine*, American Academy of Actuaries, mayo-junio de 2007 (www.contingencies.org/ [PDF]).

Acerca de los peligros del "perfilado": Bernard E. Harcourt, *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*, University of Chicago Press, 2006. <<

[118] **Acerca del trabajo de Richard Berk**: "The Role of Race in Forecasts of Violent Crime", *Race and Social Problems*, núm. 1, 2009, pp. 231-242, y entrevista por correo electrónico por Cukier, noviembre de 2012. <<

[119] **Sobre la pasión de McNamara por los datos**: Phil Rosenzweig, “Robert S. McNamara and the Evolution of Modern Management”, *Harvard Business Review*, diciembre de 2010 (hbr.org). <<

[120] **Sobre el éxito de los "cerebritos" durante la Segunda Guerra Mundial**: John Byrne, *The Whiz Kids*, Doubleday, 1993.

Sobre McNamara en la Ford: David Halberstam, *The Reckoning*, William Morrow, 1986, pp. 222-245.

Libro de Kinnard: Douglas Kinnard, *The War Managers*, University Press of New England, 1977, pp. 71-25. Esta sección se ha beneficiado de una entrevista por correo electrónico con el Dr. Kinnard, a través de su asistente, por la que los autores desean expresar su gratitud. <<

[121] **La cita “En Dios confiamos**…” se atribuye a menudo a W. Edwards Deming.

Acerca de Ted Kennedy y la lista de excluidos de los vuelos, véase: Sara Kehaulani Goo, “Sen. Kennedy Flagged by No-Fly List”, *The Washington Post*, 20 de agosto de 2004, p. A01 (www.washingtonpost.com). <<

[122] **Acerca de los métodos de contratación de Google, véase**: Douglas 210, *I'm Feeling Lucky: The Confessions of Google Employee Number 59*, Houghton Mifflin Harcourt, 2011, p. 9. Véase asimismo: Steven Levy, *In the Plex*, Simon and Schuster, 2011, pp. 140-141. Resulta irónico que los cofundadores de Google quisieran contratar a Steve Jobs como director general (a pesar de carecer de título universitario); Levy, p. 80.

Sobre la prueba con cuarenta y un matices de azul, véase: Laura M. Holson, "Putting a Bolder Face on Google", *The New York Times*, 1 de marzo de 2009 (www.nytimes.com).

Acerca de la dimisión del diseñador jefe de Google: la cita está extractada (sin elipsis en aras de la legibilidad) de Doug Bowman, "Goodbye, Google", *post* en el blog 20 de marzo de 2009 (stopdesign.com). <<

[123] **La cita de Steve Jobs** procede de Steve Lohr, “Can Apple Find More Hits Without Its Tastemaker?”, *The New York Times*, 18 de enero de 2011, p. B1 (www.nytimes.com/).

El libro de Scott: James Scott, *Seeing like a State: How Certain Schemes to Improve the Human Condition Have Failed*, Yale University Press, 1998.

La cita de McNamara de 1967 procede de un discurso en Millsaps College, Jackson, Mississippi, citado en el *Harvard Business Review* de diciembre de 2010. <<

[124] **Acerca de la disculpa de McNamara, véase**: Robert S. McNamara con Brian VanDeMark, *In Retrospect: The Tragedy and Lessons of Vietnam*, Random House, 1995, pp. 48, 270. <<

[125] **Sobre la colección de libros de la biblioteca de la universidad de Cambridge, véase**: Marc Drogin, *Anathema! Medieval Scribes and the History of Book Curses*, Allanheld and Schram, 1983, p. 37. <<

[126] **Acerca de la responsabilidad y la privacidad**: el Centre for Information Policy Leadership [Centro de Liderazgo de la Política de la Información] se ha implicado en un proyecto multianual sobre el interfaz de responsabilidad y privacidad; véase www.informationpolicycentre.com/. <<

[127] **En cuanto a las fechas de expiración de los datos, véase** Mayer-Schönberger, *Delete*. <<

[128] “**Privacidad diferencial**”: Cynthia Dwork, “A Firm Foundation for Private Data Analysis”, *Communications of the acm*, enero de 2011, pp. 86-95. <<

[129] **Facebook y la privacidad diferencial**: A. Chin y A. Klinefelter, “Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study”, *North Carolina Law Review*, vol. 90, núm. 5, 2012, p. 1417; A. Haeberlen *et al.*, “Differential Privacy Under Fire”, www.cis.upenn.edu [PDF]. <<

[130] **Empresas sospechosas de colusión**: ya hay trabajos en este campo; véase Pim Heijnen, Marco A. Haan, y Adriaan R. Soetevent, “Screening for Collusion: A Spatial Statistics Approach”, Discussion Paper TI 2012-058/1, Instituto Tinbergen, Países Bajos, 2012 (www.tinbergen.nl [PDF]). <<

[131] **Sobre los representantes corporativos alemanes de la protección de datos, véase**: Viktor Mayer-Schönberger, "Beyond Privacy, Beyond Rights: Towards a 'Systems' Theory of Information Governance", *California Law Review*, vol. 98, 2010, p.1853. <<

[132] **Acerca de la interoperabilidad, consúltese**: John Palfrey y Urs Gasser, *Interop: The Promise and Perils of Highly Interconnected Systems*, BasicBooks, 2012. <<

[133] **Mike Flowers y la analítica de la ciudad de Nueva York**: basado en una entrevista hecha por Cukier, julio de 2012. Puede hallarse una buena descripción en: Alex Howard, "Predictive Data Analytics Is Saving Lives and Taxpayer Dollars in New York City", *O'Reilly Media*, 26 de junio de 2012 (strata.oreilly.com). <<

[134] **Walmart y las "Pop-Tarts"**: Hays, "What Wal-Mart Knows About Customers' Habits". <<

[135] **Sobre el uso de los datos masivos en los suburbios y en la modelización de movimientos de refugiados, véase**: Nathan Eagle, “Big Data, Global Development, and Complex Systems”, www.youtube.com [Vídeo]. <<

[136] **La percepción del tiempo**: Benedict Anderson, *Imagined Communities*, Verso, 2006.

"**Lo pasado es el prólogo**": William Shakespeare, *La tempestad*, acto 2, escena I. <<

[137] **El experimento del CERN y el almacenamiento de datos**: cruce de correos electrónicos entre Cukier e investigadores del CERN, noviembre de 2012.

Sobre el sistema informático del Apolo 11: David A. Mindell, *Digital Apollo: Human and Machine in Spaceflight*, MIT Press, 2008. <<