ÍNDICE

I
AHORA

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

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

Unas cuantas semanas antes de que el virus H1N1 ocupase los titulares, dio la casualidad de que unos ingenieros de Google, el gigante de internet, publicaron un artículo notable en la revista científica *Nature*. Esta pieza causó sensación entre los funcionarios de sanidad y los científicos de la computación, pero, por lo demás, pasó en general

inadvertida. Los autores explicaban en ella cómo Google podía "predecir" la propagación de la gripe invernal en Estados Unidos, no solo en todo el ámbito nacional, sino hasta por regiones específicas, e incluso por estados. La compañía lo conseguía estudiando qué buscaba la gente en internet. Dado que Google recibe más de tres mil millones de consultas a diario y las archiva todas, tenía montones de datos con los que trabajar.

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

Aunque el personal de Google suponía que las búsquedas podrían centrarse en obtener información sobre la gripe –tecleando frases como "remedios para la tos y la fiebre"–, no era esa la cuestión: como no les constaba, diseñaron un sistema al que no le importaba. Lo único que hacía este sistema era buscar correlaciones entre la frecuencia de ciertas búsquedas de información y la propagación de la gripe a lo largo del tiempo y del espacio. Procesaron un total apabullante de cuatrocientos cincuenta millones de modelos matemáticos diferentes para poner a prueba los términos de búsqueda, comparando sus predicciones con los casos de gripe registrados por los CDC en 2007 y 2008. Así dieron con un filón: su software halló una combinación de cuarenta y cinco términos de búsqueda que, al usarse conjuntamente en un modelo matemático, presentaba una correlación fuerte entre su predicción y las cifras oficiales de la enfermedad a lo largo del país. Como los CDC, podían decir adónde se había propagado la gripe, pero, a diferencia de los CDC, podían hacerlo en tiempo casi real, no una o dos semanas después.

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

funcionarios de la sanidad pública consiguieron una herramienta de información incalculable.

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

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

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

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

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

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

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

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

Usando una muestra de doce mil registros de precios de vuelos, recabada a través de una web de viajes a lo largo de un periodo de cuarenta y un días, Etzioni creó un modelo predictivo que ofrecía a sus pasajeros simulados un ahorro estimable. El modelo no ofrecía

ninguna explicación del *porqué*, solo del *qué*. Es decir, no conocía ninguna de las variables que intervienen en la fijación de precios de las líneas aéreas, como el número de asientos sin vender, la estacionalidad, o si de alguna forma mágica la pernoctación durante la noche del sábado podría reducir el importe. Basaba su predicción en lo que sí sabía: probabilidades recopiladas de datos acerca de otros vuelos. "Comprar o no comprar, esa es la cuestión", se dijo Etzioni. Por consiguiente, denominó Hamlet a su proyecto.

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 por alrededor de ciento diez millones de dólares, y lo integró en el motor de búsqueda Bing. Para el año 2012, el sistema acertaba el 75 por 100 de las veces y le estaba ahorrando una media de cincuenta dólares por billete a los viajeros.

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

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

DEJAR HABLAR A LOS DATOS

Los frutos de la sociedad de la información están bien a la vista, con un teléfono móvil en cada bolsillo, un ordenador portátil en cada mochila, y grandes sistemas de tecnología de la información funcionando en las oficinas por todas partes. Menos llamativa resulta la información en sí misma. Medio siglo después de que los ordenadores se propagaran a la mayoría de la población, los datos han empezado a

acumularse hasta el punto de que está sucediendo algo nuevo y especial. No solo es que el mundo esté sumergido en más información que en ningún momento anterior, sino que esa información está creciendo más deprisa. El cambio de escala ha conducido a un cambio de estado. El cambio cuantitativo ha llevado a un cambio cualitativo. Fue en ciencias como la astronomía y la genética, que experimentaron por primera vez esa explosión en la década de 2000, donde se acuñó el término *big data*, "datos masivos". 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. Ese es el origen de las nuevas tecnologías de procesamiento, como MapReduce, de Google, y su equivalente de código abierto, Hadoop, que surgió de Yahoo. Con ellos se pueden manejar cantidades de datos mucho mayores que antes, y esos datos –esto es lo importante– no precisan ser dispuestos en filas ordenadas ni en las clásicas tabulaciones de una base de datos. Otras tecnologías de procesamiento de datos que prescinden de las jerarquías rígidas y de la homogeneidad de antaño se vislumbran asimismo en el horizonte. Al mismo tiempo, dado que las compañías de internet podían recopilar vastas cantidades de datos y tenían un intenso incentivo financiero por hallarles algún sentido, se convirtieron en las principales usuarias de las tecnologías de procesamiento más recientes, desplazando a compañías de fuera de la red que, en algunos casos, tenían ya décadas de experiencia acumulada.

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

Pero esto no es más que el principio. La era de los datos masivos pone en cuestión la forma en que vivimos e interactuamos con el

mundo. Y aun más, la sociedad tendrá que desprenderse de parte de su obsesión por la causalidad a cambio de meras correlaciones: ya no sabremos *por qué*, sino solo *qué*. Esto da al traste con las prácticas establecidas durante siglos y choca con nuestra comprensión más elemental acerca de cómo tomar decisiones y aprehender la realidad.

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

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

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

Las compañías de internet se han visto particularmente abrumadas. Google procesa más de 24 petabytes de datos al día, un volumen que representa miles de veces la totalidad del material impreso que guarda la Biblioteca del Congreso de Estados Unidos. A Facebook, una empresa que no existía hace una década, se suben más de diez millones de fotos nuevas cada hora. Sus usuarios hacen clic en el botón de "me gusta" o insertan un comentario casi tres mil millones de veces diarias, dejando un rastro digital que la compañía explota para descubrir sus preferencias. Entretanto, los ochocientos millones de usuarios mensuales del servicio YouTube de Google suben más de una hora de vídeo cada segundo. El número de mensajes de Twitter aumenta alrededor de un 200 por 100 al año, y en 2012 se habían superado los cuatrocientos millones de tuits diarios.

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

Son muchos quienes han intentado determinar la cifra exacta de la cantidad de información que nos rodea, y calcular a qué velocidad crece. Lo conseguido ha sido irregular, porque han medido cosas diferentes. Uno de los estudios más completos es obra de Martin Hilbert, de la Annenberg School de comunicación y periodismo de la universidad del Sur de California. Hilbert se ha esforzado por

cifrar todo cuanto ha sido producido, almacenado y comunicado, lo cual comprendería no solo libros, cuadros, correos electrónicos, fotografías, música y vídeo (analógico y digital), sino también videojuegos, llamadas telefónicas, hasta navegadores de vehículos y cartas enviadas por correo postal. También incluyó medios de emisión como la televisión y la radio, basándose en sus cifras de audiencia.

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

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

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

Las cosas se están acelerando de verdad. La cantidad de información almacenada crece cuatro veces más deprisa que la economía mundial, mientras que la capacidad de procesamiento de los ordenadores crece nueve veces más deprisa. No tiene nada de raro que la gente se queje de sobrecarga informativa. A todos nos abruman los cambios.

Tómese la perspectiva a largo plazo, comparando el actual diluvio de datos con una revolución de la información anterior, la de la imprenta de Gutenberg, inventada hacia 1439. En los cincuenta años que van de 1453 a 1503, se imprimieron unos ocho millones de libros, 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, comentó con mordacidad: "No hemos inventado nada".

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

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

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

En la información, como en la física, el tamaño sí importa. Por consiguiente, Google mostró que era capaz de determinar la prevalencia de la gripe casi igual de bien que los datos oficiales basados en las visitas de pacientes al médico. Google puede hacerlo peinando cientos de miles de millones de términos de búsqueda, y puede obtener una respuesta casi en tiempo real, mucho más rápido que las fuentes oficiales. Del mismo modo, el Farecast de Etzioni puede predecir la volatilidad del precio de un billete de avión, poniendo así un poder económico sustancial en manos de los consumidores. Pero ambos solo pueden hacerlo bien mediante el análisis de cientos de miles de millones de puntos de datos.

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

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

Esencialmente, los datos masivos consisten en hacer predicciones. Aunque se los engloba en la ciencia de la computación llamada inteligencia artificial y, más específicamente, en el área llamada aprendi-

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

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

MÁS, DE SOBRA, YA BASTA

Los datos masivos serán una fuente de innovación y de nuevo valor económico. Pero hay aún más en juego. El auge de los datos

masivos representa tres cambios en la forma de analizar la información que modifican nuestra manera de comprender y organizar la sociedad.

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

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

Esta forma de pensar resultaba de un entorno de "datos escasos": con tan pocas cosas que medir, teníamos que tratar de la forma más precisa posible lo que nos molestábamos en cuantificar. En cierto modo, esto es obvio: al llegar la noche, una tienda pequeña puede

contar el dinero que hay en la caja hasta el último céntimo, pero no haríamos lo mismo –de hecho, no podríamos– en el caso del producto interior bruto de un país. Conforme va aumentando la escala, también crece el número de errores.

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

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

Estos dos cambios conducen a un tercero, que explicamos en el capítulo IV: un alejamiento de la tradicional búsqueda de causalidad. Como seres humanos, hemos sido condicionados para buscar causas, aun cuando la búsqueda de la causalidad resulte a menudo difícil y pueda conducirnos por el camino equivocado. En un mundo de datos masivos, en cambio, no necesitamos concentrarnos en la causalidad; por el contrario, podemos descubrir pautas y correlaciones en los datos que nos ofrezcan perspectivas nuevas e inapreciables. Puede que las correlaciones no nos digan precisamente *por qué* está ocurriendo algo, pero nos alertan de que *algo* está pasando.

Y en numerosas situaciones, con eso basta. Si millones de registros médicos electrónicos revelan que los enfermos de cáncer que toman determinada combinación de aspirina y zumo de naranja ven remitir su enfermedad, la causa exacta de la mejoría puede resultar menos importante que el hecho de que sobreviven. Del mismo modo, si podemos ahorrarnos dinero sabiendo cuál es el mejor momento de comprar un billete de avión, aunque no comprendamos el método

subyacente a la locura de las tarifas aéreas, con eso basta. Los datos masivos tratan del *qué*, no del *porqué*. No siempre necesitamos conocer la causa de un fenómeno; preferentemente, podemos dejar que los datos hablen por sí mismos.

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

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

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

El gran avance hacia la gestión más eficiente de los datos llegó con el advenimiento de la digitalización: hacer que la información analógica fuese legible por los ordenadores, lo que también la vuelve más fácil y barata de almacenar y procesar. Este progreso mejoró drásticamente la eficiencia. La recopilación y el análisis de información, que en tiempos exigía años, podía ahora hacerse en días, o incluso menos. Pero cambió poco más. Los encargados de los datos muy a

menudo estaban versados en el paradigma analógico de asumir que los conjuntos de datos tenían propósitos específicos de los que dependía su valor. Nuestros mismos procesos perpetuaron este prejuicio. Por importante que resultase la digitalización para permitir el cambio a los datos masivos, la mera existencia de ordenadores no los hizo aparecer.

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

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

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

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

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

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

La sociedad cuenta con milenios de experiencia en lo que a comprender y supervisar el comportamiento humano se refiere, pero, ¿cómo se regula un algoritmo? En los albores de la computación, los legisladores advirtieron que la tecnología podía usarse para socavar la privacidad. Desde entonces, la sociedad ha erigido un conjunto de reglas para proteger la información personal. Sin embargo, en la era de los datos masivos, esas leyes constituyen una línea Maginot en

buena medida inútil. La gente comparte gustosamente información online: es una característica central de los servicios en red, no una vulnerabilidad que haya que evitar.

Entretanto, el peligro que se cierne sobre nosotros en tanto que individuos se desplaza del ámbito de lo privado al de la probabilidad: los algoritmos predecirán la probabilidad de que uno sufra un ataque al corazón (y tenga que pagar más por un seguro médico), deje de pagar la hipoteca (y se le niegue un crédito) o cometa un delito (y tal vez sea detenido antes de los hechos). Ello conduce a una consideración ética del papel del libre albedrío frente a la dictadura de los datos. ¿Debería imponerse la voluntad del individuo a los datos masivos, aun cuando las estadísticas argumenten lo contrario? Igual que la imprenta preparó el terreno para las leyes que garantizaban la libertad de expresión –que no existían antes, al haber tan poca expresión escrita que proteger–, la era de los datos masivos precisará de nuevas reglas para salvaguardar la inviolabilidad del individuo.

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

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

Los beneficios para la sociedad resultarán muy numerosos, conforme los datos masivos se conviertan en parte de la solución de ciertos

problemas globales acuciantes, como hacer frente al cambio climático, 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.