BIG DATA

Hacia la cuarta revolución industrial

Antonio Monleón (coord.)
Esteban Vegas
Ferran Reverter

UBe

Índice

Prólogo

Los dispositivos digitales están presentes en nuestra vida: generamos constantemente datos y la mayoría son almacenados, ya se trate de información geográfica, estadística, datos meteorológicos, resultados de las investigaciones, datos de transporte, datos de consumo de energía o datos de salud. Es lo que se ha denominado los macrodatos o Big Data; un término que hasta ahora solo utilizaban los académicos, pero que ha sido incorporado rápidamente por la sociedad en general. Son grandes océanos de datos digitales que se pueden capturar, comunicar, almacenar y analizar, aunque varios estudios indican que actualmente son inmanejables. Junto con el capital y la fuerza de trabajo, los datos se han convertido en un valor añadido para la economía, que espera que se produzcan grandes cambios en una sociedad dirigida por los datos. En este sentido, la Comisión Europea está apoyando la transición hacia una economía basada en los datos, en la que se prevé la creación de miles de nuevos puestos de trabajo. Dicha economía estará basada en la investigación y la innovación, generará más oportunidades de negocio y una mayor disponibilidad del conocimiento y el capital, en particular para las pequeñas y medianas empresas (pymes).

El futuro está en la investigación, en el tratamiento y aplicación de los datos, que aportarán prosperidad a nuestra sociedad, pero existen reparos respecto a cómo estos amenazan nuestra privacidad. En el presente artículo se recogen las normas y leyes que regulan hoy en día los Big Data.

La capacidad predictiva de los Big Data es ingente, permite analizar grandes cantidades de datos relacionados y es una herramienta

de gran utilidad en sectores como la salud pública o la economía. En el presente trabajo, la mayor parte de la investigación en torno a estos «datos grandes» se ha centrado en su volumen, pero a lo largo del artículo se desarrollan varios ejemplos donde se demuestra que no solo es cuestión de cantidad sino de calidad. Para procesar y analizar la información Big Data, almacenada y distribuida en grandes sistemas como Hadoop, una base de datos o cualquier otro tipo de almacenamiento, es necesario el aprendizaje automático (*machine learning*, ML). El ML, cuyas bases proceden de las ciencias y la ingeniería, se ocupa de la construcción y el estudio de algoritmos que pueden aprender a partir de datos. Existen muchas técnicas disponibles, como modelado lineal y no lineal, estadística descriptiva, pruebas estadísticas clásicas, clasificación o agrupamiento. En este libro se recogen las tecnologías que se utilizan para almacenar y analizar los Big Data.

Nuestro más sincero agradecimiento a Biel Stela, que cursó el Postgraduate on Data Science and Big-data (Universidad de Barcelona), por aportar ideas y material sobre el tema del *deep learning*, i a Sergi Gallardo Masip, por la ilustración de la cubierta *Red neuronal daliniana*.

Introducción

Hace ya unos años, en el trabajo «El tratamiento numérico de la realidad. Reflexiones sobre la importancia actual de la estadística en la sociedad de la información» (Monleón, 2010) se revisó el papel de la estadística en el tratamiento de datos y su importancia en la sociedad actual. Este papel ha ido cobrando mayor relevancia debido a la gran cantidad de datos que cada día se generan: el 90% de los datos existentes ha sido producido en los últimos dos años.

Nuestra vida cotidiana está siendo observada constantemente: desde el mismo momento en que abandonamos nuestra casa, sin querer estamos generando datos. Cuando conectas un dispositivo digital, como el GPS del coche, estás dejando un rastro digital cargado de información. Lo mismo sucede cuando envías un correo electrónico o manejas un teléfono inteligente (*smartphone*), cuando usas una red social o una tarjeta de crédito para pagar la compra. Son solo unos ejemplos de la cotidianidad de la información digital.

El término «Big Data» se ha popularizado estos últimos años debido a que ha sido utilizado con frecuencia por las grandes compañías de las tecnologías de la información, que se sirven de esta gran cantidad de información para mejorar las demandas electrónicas (ventas en línea) de sus clientes e intentar orientar las compras de forma que estén más dirigidas al propio cliente, sean más amigables y más próximas. Para ello ha sido necesario recoger la información almacenada de los clientes. La primera pregunta que la sociedad se plantea es: con el debido tratamiento estadístico, ¿estos datos pueden ser usados para mejorar nuestra vida o, por el contrario, se

convierten en un instrumento de control por parte de las grandes corporaciones o de los propios gobiernos? ¿Cuál es la tendencia futura?

Este libro pretende recoger las dudas generadas por los Big Data y aportar sentido crítico desde el punto de vista de un estadístico: una persona que analiza información pero que de repente se ve desbordada por una gran cantidad de datos que no cesan de llegar.

A todos nos surgen preguntas respecto al uso y existencia del océano digital. ¿Es la proliferación de datos la prueba de que el mundo es cada vez más intrusivo? ¿Podemos estar seguros de que hay un peso y un valor económico detrás de toda esta información masiva? ¿Debemos dejar a las máquinas la tarea de filtrar la información y seleccionar lo que es relevante? ¿Debemos legislar el uso de esta información? Estas son algunas de las preguntas que vamos a tratar de resolver, o, al menos, tratar de dejar planteadas.

Datos inmanejables: ¿qué son los Big Data?

La sociedad crea datos y más datos, y cada vez existen más dispositivos y más eficientes para almacenarlos. Los datos son vistos en sí mismos como un bien o un capital para la organización, pública o privada, que disponga de ellos. Según Manyika (2011), estas grandes cantidades de datos se están convirtiendo en factores de producción esenciales dentro de cada sector económico.

Dos estudios al respecto, uno realizado por Manyika y colaboradores (2011), del McKinsey Global Institute, y otro por Andrew McAfee y Erik Brynjolfsson (2012), de la Harvard Business School, indican que el número de datos resulta actualmente inmanejable. Aquí van unos ejemplos citados por estos estudios:

- El 90% de los datos del mundo han sido creados en los últimos dos años.

- Un disco duro que contiene toda la música del mundo solo vale unos 500 €.
- En el año 2010 había ya 5.000 millones de teléfonos móviles.
- 30.000 millones de contenidos han sido compartidos en Facebook en tan solo un mes.
- 235 terabytes de información fueron almacenados por la Biblioteca del Congreso estadounidense en abril de 2011.
- 15 de los 17 sectores productivos de la economía norteamericana cuentan con más datos almacenados que la Biblioteca del Congreso de los Estados Unidos de América (la mayor del mundo).
- Durante 2012, cada día se generaron alrededor de 2,5 exabytes de información. Este número se dobla aproximadamente cada cuarenta meses.

Las empresas capturan miles de millones de bytes de información sobre sus clientes, proveedores y las operaciones que realizan. Millones de sensores conectados en red están presentes en dispositivos tales como teléfonos móviles, sistemas de detección o redes sociales. Las personas, bien sea con teléfonos inteligentes (*smartphones*) o a través de las redes sociales estimulan el crecimiento exponencial de la información.

El término «Big Data» es confuso: son grandes datos, pero ¿a qué tamaño se refiere? Según la bibliografía consultada, no se hace referencia a un tamaño de información específico (IBM, 2014), pero habitualmente se utiliza el término cuando se habla de petabytes (PB) y exabytes (EB) de datos. La información digital se mide en bytes,[1] que es la unidad básica de información, y a partir de esta se construye la escala de medida digital de bytes:

1 El byte es la unidad de información digital básica, y es un múltiplo del bit. Generalmente equivale a 8 bits. En español se le denomina octeto.

- Kylobyte (kB) = 10^3 = 1.000 bytes
- Megabyte (MB) = 10^6 = 1.000.000 bytes
- Gigabyte (GB) = 10^9 = 1.000.000.000 bytes
- Terabyte (TB) = 10^{12} = 1.000.000.000.000 bytes
- Petabyte (PB) = 10^{15} = 1.000.000.000.000.000 bytes → BIG+ DATA
- Exabyte (EB) = 10^{18} = 1.000.000.000.000.000.000 bytes
- Zettabyte (ZB) = 10^{21} bytes
- Yottabyte (YB) = 10^{24} bytes
- Quintillón (QB)= 10^{30} bytes

Para hacernos una idea de la cantidad de información que se puede almacenar, podríamos establecer la siguiente comparación:

- 1 byte: una letra
- 10 bytes: una o dos palabras
- 100 bytes: una o dos frases
- 1 kylobyte: una historia muy corta
- 10 kylobytes: una página de enciclopedia
- 100 kylobytes: una fotografía de resolución media
- 1 megabyte: una novela
- 10 megabytes: dos copias de la obra completa de Shakespeare
- 100 megabytes: un metro de libros archivados
- 1 gigabyte: un lápiz de memoria lleno de páginas con texto
- 1 terabyte: 50.000 árboles de papel
- 10 terabytes: la colección impresa de la Biblioteca del Congreso de los Estados Unidos.
- ... BIG DATA
- 1 quintillón: datos que se generan en el mundo en un día

Según IBM (2014), cada día se generan más de 1 quintillón de datos, que proceden de fuentes tan diferentes como clientes, proveedores, operaciones financieras en línea, u obtenidos de dispositivos

móviles, análisis de redes sociales, GPS. En muchos países se gestionan gigantescas bases de datos que contienen información de impuestos, censo de población, registros médicos, etc. (IBM, 2014).

Un estudio realizado por la empresa tecnológica Cisco calcula que entre 2011 y 2016 los datos móviles crecerán anualmente un 78% y el total de dispositivos móviles que están conectados a internet superará el número de habitantes de la Tierra. Así, se estima que en 2016 habrá unos 19.000 millones de dispositivos conectados a la red, más de 2 por habitante del planeta, por lo que el tráfico global de datos móviles alcanzará los 130 exabytes anuales. Este volumen de tráfico previsto para 2017 equivale a 33.000 millones de DVD anuales, una cifra del todo inabarcable (Cisco, 2014).

Y, en el océano digital, no solo las personas suministran datos con su actividad directa, también las máquinas los registran. Unos 30 millones de sensores interconectados envían instantáneamente datos en el sector del automóvil, eléctrico, comercio, logístico, industrial, científico, etc. Pensemos en los contadores digitales eléctricos que las compañías suministradoras están instalando en nuestros hogares. Enviarán nuestros consumos eléctricos a dichas compañías a intervalos regulares, de manera que estas podrán disponer de un perfil fidedigno de nuestra actividad diaria, millones de terabytes a almacenar y analizar. Es la denominada comunicación M2M (máquina a máquina o *machine-to-machine*), que genera también una gran cantidad de información que crecerá cada año de manera exponencial.

Finalmente, la última definición establecida de los Big Data, que es la recogida en la Recomendación UIT-T Y.3600 o de «Grandes volúmenes de datos – requisitos y capacidades basados en la computación en la nube», de 12/2015, dice:

> Big Data es un paradigma para hacer posible la recopilación, el almacenamiento, la gestión, el análisis y la visualización, potencialmente en condiciones de tiempo real, de grandes conjuntos de datos con características heterogéneas.

¿Afectan los Big Data a la sociedad en general?

Es tan importante el manejo de estos datos masivos que está produciendo cambios en la economía mundial. En nuestro entorno, la Unión Europea concentra gran parte de sus actividades de investigación e innovación en el llamado Programa Marco, que en esta edición se denominará Horizonte2020 (H2020). En el período 2014-2020, y mediante la implantación de tres pilares, se pretende abordar los principales retos sociales, promover el liderazgo industrial en Europa y reforzar la excelencia de su base científica. H2020 promueve la generación de una economía basada en el conocimiento; en consecuencia, uno de los objetivos que se ha fijado este programa marco es el de desarrollar tecnologías y sus aplicaciones para mejorar la competitividad europea; para ello cuenta con inversiones en tecnologías clave para la industria, como las tecnologías de la información y de la comunicación (TIC) (Ministerio de Economía y Competitividad, 2014).

En julio de 2014, la Comisión presentó una nueva estrategia sobre los Big Data, para apoyar y acelerar la transición hacia una economía basada en los datos en el espacio europeo. La economía basada en datos estimulará la investigación y la innovación en general, y dará lugar a más oportunidades de negocio y a un aumento de la disponibilidad de los conocimientos y el capital, en particular para las pequeñas y medianas empresas (pymes). Estas afirmaciones de la Unión Europea se recogen en su artículo «Towards a thriving data-driven economy» (Unión Europea, 2014a), donde, citando a otras fuentes, en especial americanas, indican que se espera que la tecnología y los servicios basados en Big Data crezcan en todo el mundo a una tasa compuesta de crecimiento anual del 40% —cerca de siete veces la del mercado de las TIC en general.

Pero ¿cómo puede esta recopilación tan masiva y su posterior análisis mejorar nuestra vida? Para responder a esta pregunta debemos tener en cuenta que el quid de la cuestión es el propósito con

que se hace. Según indica Sánchez (2013) en su interesante artículo periodístico «Big-data: presente y futuro para las empresas», las grandes compañías tecnológicas disponen de centros de almacenamiento para guardar estas grandes cantidades de información, y, en un análisis posterior, pueden estudiar el comportamiento de los clientes para luego realizar acciones comerciales más efectivas o focalizar la publicidad en los intereses del consumidor.

La Comisión Europea también ha publicado otro interesante artículo, «Making Big Data work for Europe» (Comisión Europea, 2014b), donde indica por qué son importantes los Big Data. Los datos se han convertido en un activo clave para la economía y nuestras sociedades, similar a las categorías clásicas de los recursos humanos y financieros. La necesidad de dar sentido a los Big Data está originando innovaciones en la tecnología, el desarrollo de nuevas herramientas y nuevas habilidades. Un buen uso de los datos puede crear oportunidades en sectores más tradicionales como el transporte, la salud o la fabricación. La Comisión Europea (2014b) cita algunos ejemplos de cómo el análisis y el tratamiento de datos, sobre todo de los Big Data, cambiarán la sociedad:

- Transformarán las industrias de servicios de Europa mediante la generación de una amplia gama de productos y servicios de información innovadores.
- Aumentarán la productividad de todos los sectores de la economía.
- Mejorarán la investigación y acelerarán la innovación.
- Lograrán reducciones de costos a través de servicios más personalizados.
- Aumentarán la eficiencia en el sector público.

En la figura 1 se presentan algunos ejemplos reales de cómo los Big Data afectan o afectarán a nuestra vida cotidiana, en cualquier ámbito y lugar.

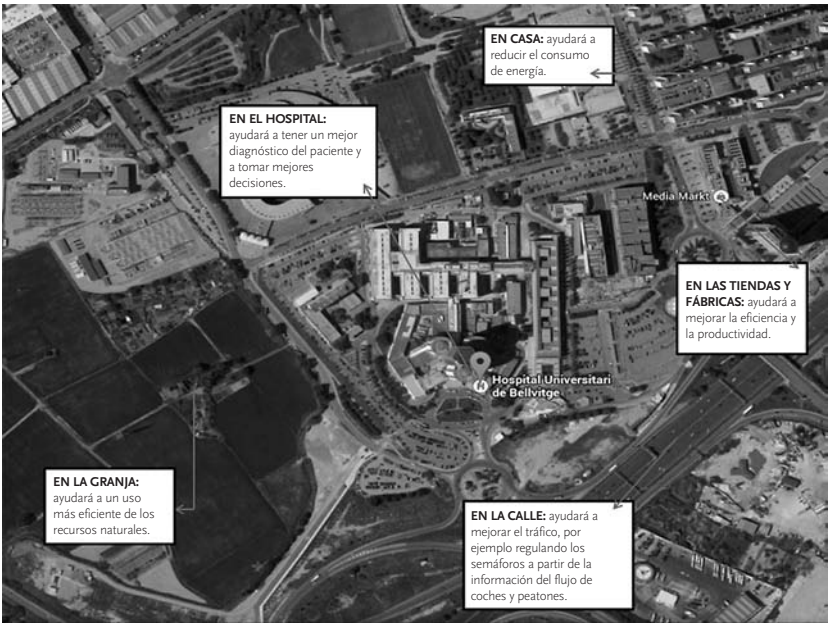

Figura 1. Algunos ejemplos de cómo los Big Data afectan y afectarán a nuestra vida cotidiana. (Basado en «Making Big Data work for Europe», Comisión Europea. 2014b).

Fuente: Fotografía de satélite de Google Maps; entorno del Hospital de Bellvitge en Hospitalet de Llobregat (Barcelona).

Uno de los campos más prometedores en este ámbito es la medicina; así, el análisis de los Big Data está contribuyendo a reducir los elevados costes de la investigación clínica, proporcionando medidas reales del desempeño de nuestro sistema sanitario y ayudando a los médicos y pacientes a tomar mejores decisiones (Science Spain, 2014).

Pablo Serrano, director médico del Hospital de Fuenlabrada (Madrid), durante el 59.º Congreso de la Sociedad Española de Farmacia Hospitalaria (SEFH), celebrado en octubre de 2014, señaló los nuevos retos en el uso de estos datos, así, «en el ámbito de la farmacia hospitalaria, la tecnología Big-data ayudaría a comprender mejor la utilización de los medicamentos y los integrarían en el conjunto del

hospital para un conocimiento mayor de la morbilidad y el uso de recursos» (Science Spain, 2014).

Otro ejemplo sanitario comentado por Esteban (2014) en el artículo «Cinco ejemplos de cómo el 'Big-data' puede mejorar la sociedad» sería el de las pandemias, como el ébola, que recientemente se ha convertido en un problema mundial. Así, gracias a los Big Data se puede determinar el riesgo de una pandemia en tiempo real, a través de las tendencias que se registran en un buscador de internet como Google u otros.

Sin embargo, el campo de la biología, y en especial la genética, es uno de los más prometedores. Así, los avances de los últimos años en el campo de la biología y la bioinformática han creado la «era ómica», una era donde se da una visión global de los procesos biológicos basada en el análisis de un gran volumen de datos, por lo que se necesita el apoyo de la bioinformática para la interpretación de los resultados obtenidos. El análisis y la interpretación de estos Big Data permiten estudiar organismos hasta ahora desconocidos, así como sus funciones, todo a través de su rastro genético. También se ha denominado a este tipo de estudios ciencias ómicas: la genómica, la proteómica, la transcriptómica y la metabolómica. Todas estas especialidades han hecho avanzar a una gran velocidad la biomedicina y la biotecnología. Un ejemplo sería el estudio de asociación del genoma completo (GWAS, Genome-wide association study, o WGAS, Whole genome association study). Son análisis de la variación genética con un genoma humano completo y tienen como objetivo asociar el genoma con un rasgo observable (patología), por ejemplo, ayudando a identificar si una persona tiene un determinado riesgo de sufrir una enfermedad. Estos estudios requieren genotipar a un gran número de personas, obtener muestras de su genoma y analizarlo. Hoy en día, gracias a las técnicas de Big Data ya se están obteniendo resultados muy prometedores, con aplicaciones biomédicas o biotecnológicas a corto o medio plazo. En la figura 2 se presenta un ejemplo de GWAS de Yeager (2007) para cáncer de próstata.

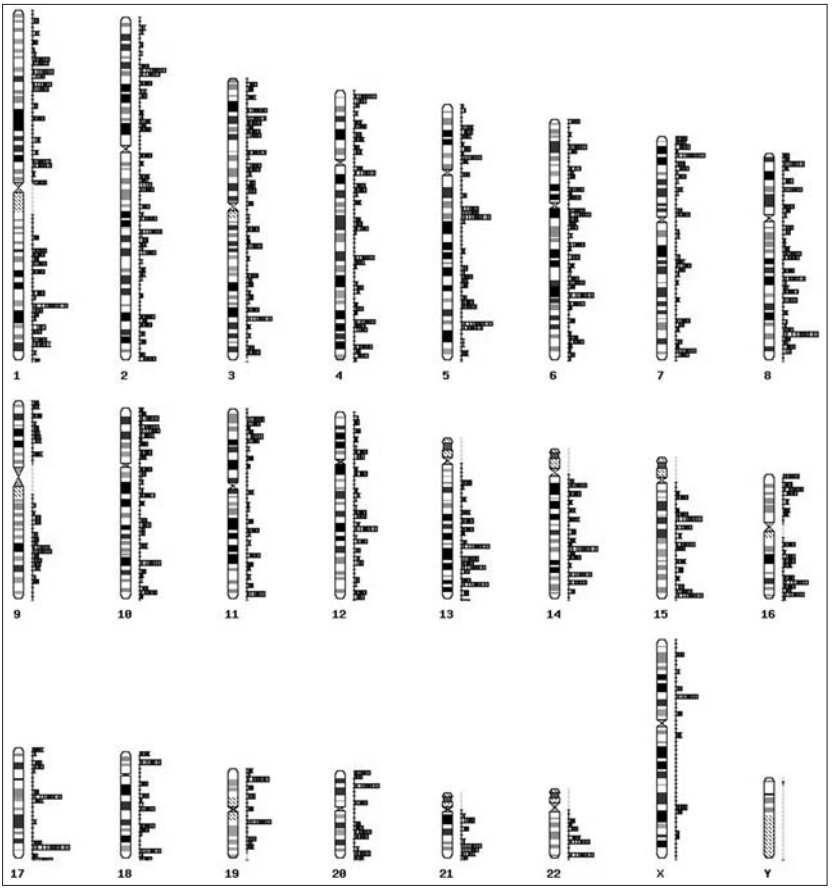

Figura 2. Análisis de asociación de genoma completo (Genome-wide association study, GWAS) de marcadores de riesgo para cáncer de próstata. En la imagen pueden verse los 23 cromosomas y diferentes tipos de marcadores genéticos (Yeager et al., 2007).
Fuente: http://www.gwascentral.org/study/HGVST1.

¿Es posible que en un futuro en lugar de médicos sean los robots quienes diagnostiquen nuestras enfermedades y prescriban el mejor tratamiento basándose en su experiencia? Los médicos hacen diagnósticos basados en su juicio y sus conocimientos. Sin embargo, con

el tiempo esto probablemente se considerará un disparate. ¿Por qué no utilizar los Big Data? Se podría reunir la información de la práctica habitual y la experiencia de todos los médicos, y de cientos de millones de pacientes durante años, para identificar los mejores tratamientos con el fin de lograr los mejores resultados y detectar ocultos efectos secundarios adversos de los medicamentos. Después de todo, un médico solo o un investigador no pueden poseer la suma de todo el conocimiento médico. Pero si agregamos gran cantidad de información sanitaria junto con información genética del paciente y conocimiento científico, podemos determinar qué es lo que funciona mejor. De momento este planteamiento choca contra las leyes de protección de datos actuales.

Otras soluciones que pueden ofrecer los Big Data a los retos de la sociedad actual se centran en el ámbito humano y en la sostenibilidad (*smart cities*), un tema muy debatido hoy en día. Así, se están desarrollando sistemas de información inteligentes que, a partir de sensores electrónicos instalados a pie de calle, permiten cambiar la duración de las luces de los semáforos en función de los datos recogidos en tiempo real sobre el tráfico. Otro ejemplo en el ámbito económico sería el método desarrollado en el Massachusetts Institute of Technology (MIT), que recoge continuamente datos sobre precios de productos comercializados en internet y los utiliza para estimar la tasa de inflación. El objetivo es identificar los picos de inflación de manera más rápida que cualquier otro método tradicional (Esteban, 2014).

¿Es la privacidad el problema de los Big Data? ¿Qué dice la ley?

Por otro lado existe un problema muy importante para la sociedad: la privacidad y el control de la información. ¿Qué podría pasar si caen en manos ajenas todos nuestros datos personales?: consumo