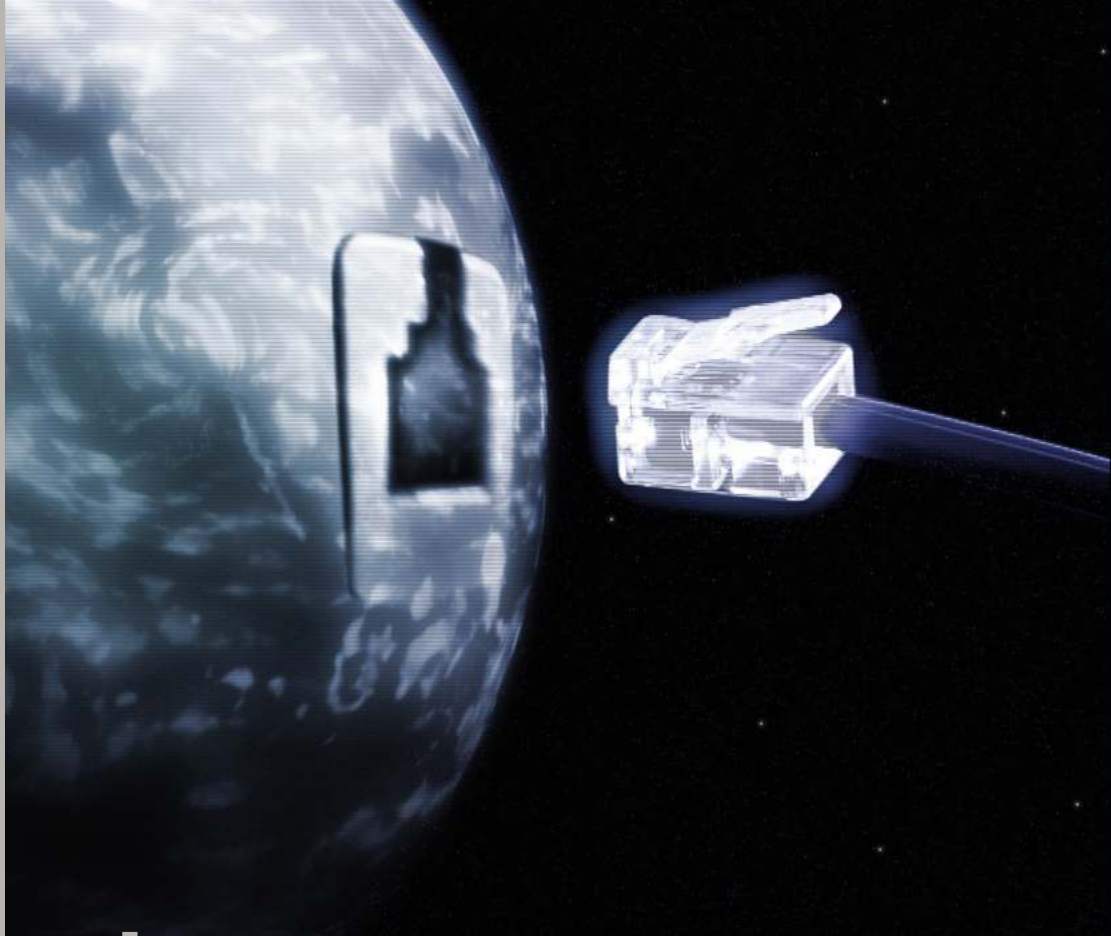

El imperio de los datos

Xavier Duran Escriba

El Big Data, la privacidad y la sociedad del futuro

CÀTEDRA DE DIVULGACIÓ DE LA CIÈNCIA
PUV
sin fronteras

El imperio de los datos

El Big Data, la privacidad y la sociedad del futuro

Xavier Duran

PREMIO EUROPEO DE DIVULGACIÓN CIENTÍFICA
ESTUDI GENERAL 2017

PUV PUBLICACIONS VNIVERSITAT DE VALÈNCIA CAM CÀTEDRA DE DIVULGACIÓ DE LA CIÈNCIA UCC+i

sin fronteras

Directora de la colección:
Carolina Moreno

Coordinación:
Soledad Rubio

Unitat de Cultura Científica
i de la Innovació de la Universitat de València
www.valencia.edu/cdciencia
cdciencia@uv.es

Publicacions de la Universitat de València, 2019
www.uv.es/publicacions
publicacions@uv.es

Producción editorial: Maite Simón

Interior
Diseño y maquetación: Inmaculada Mesa
Corrección: Letras y Píxeles S.L.

Cubierta
Diseño original: Enric Solbes
Grafismo: Celso Hernández de la Figuera

ISBN: 978-84-9134-362-2
Depósito Legal: V-269-2019
Impresión: La Imprenta Comunicación Gráfica, S. L.

La tecnología no es ni buena, ni mala, ni neutral.

A pesar de que la tecnología puede ser un elemento principal en muchos asuntos públicos, los factores no técnicos deben tener preferencia en las decisiones sobre política tecnológica.

(1.ª y 4.ª leyes de Kranzberg)
MELVIN KRANZBERG (1986)

El objetivo principal de toda ciencia es la libertad y la felicidad del hombre.

THOMAS JEFFERSON (1810)

Es un gran error elaborar teorías antes de tener datos. Inconscientemente, empiezas a distorsionar los hechos para adaptarlos a la teoría en vez de adaptar la teoría a los hechos.

(Sherlock Holmes)
ARTHUR CONAN DOYLE
(*Escándalo en Bohemia*)

Premios Literarios Ciutat d'Alzira 2017

Esta obra obtuvo el XXIII Premio Europeo de Divulgación Científica Estudi General, instituido por la Universitat de València y el Ayuntamiento de Alzira. Formaban parte del jurado Carmen Agustí, Pilar Campins, Andreu Escrivà, Lucía Hipólito y Fernando Sapiña.

ÍNDICE

Introducción
YO SOY YO Y MIS DATOS

Datos, datos, datos a montones... Vivimos en un mundo de datos, almacenados en formas variadas: textos, números, imágenes, gráficos... «Yo soy yo, mis circunstancias... y mis datos», diría hoy Ortega y Gasset.

Hay tantos y tantos que ya no hablamos de datos, sino de Big Data, grandes datos. El concepto ha hecho fortuna y pese a que a menudo se deja en inglés, también se adapta a cada idioma. En castellano de habla de megadatos o de datos masivos. Utilizaremos preferentemente Big Data, pero también usaremos las dos traducciones. La idea siempre es que hay muchos datos.

Big Data nos hace pensar en archivos digitales y en consultas por internet. Pensamos en Google y en las montañas de información por donde debe moverse este buscador para buscar lo que le pedimos. Y quizá pensamos en Facebook y en Instagram. Pero, como iremos explicando a lo largo del libro, parece que nada queda al margen del Big Data: ni mensajes privados por WhatsApp, ni llamadas telefónicas, ni compras con tarjeta, ni siquiera los paseos con el móvil encendido. Pero hay muchas más fuentes de datos: los que mandan los satélites, los que proporcionan sensores repartidos por las ciudades, por el campo o por los océanos, las imágenes de cámaras de seguridad, los datos que proporcionan aparatos médicos o los llamados *wearables* –una especie de captadores de datos portátiles, que pueden consistir en un brazalete o en una prenda, como una camiseta.

Ya escribió el filósofo inglés Francis Bacon, a finales del siglo XVI, que «Conocimientos es poder». Pero datos y conocimiento no son lo mismo. De hecho, incluso hay entre ellos un paso intermedio, que es la información. Confundimos datos con información y son cosas distintas. Un grupo de músicos tocando por su cuenta, por afinadamente que lo hagan y por virtuosos que sean, son datos. Todos ellos tocando en armonía a las órdenes de un director de orquesta es información.

Los datos son el combustible que permite resolver problemas –a veces, creados por los datos mismos–. Pero un combustible solo no sirve de nada. Los datos sirven para que funcione la maquinaria que busca las respuestas a los problemas. Por eso, los datos son imprescindibles, pero sin una estrategia para tratarlos y transformarlos no tendríamos nunca información. Y una vez reunida suficiente información, aún nos queda el trabajo de analizara y de reflexionar. De la manera como la utilicemos para producir conocimiento dependerá la calidad de este.

Aun así, no podemos negar que, hoy en día, los datos son poder. Hay quien los llama «el petróleo del siglo XXI». Volvemos a la metáfora del combustible, pero en este caso para alimentar máquinas de fabricar dinero –y de construir poder–. Quien tiene muchos datos tiene mucho poder, si sabe cómo utilizarlos... o si los vende a alguien a quien le interese hacerlo.

Aquí explicaremos de dónde surgen tantos datos, cómo circulan, cómo se guardan. Mostraremos cómo se procesan –algo que se puede hacer bien o muy mal–. Y describiremos los beneficios que aportan y los riesgos que representan. Muchos posibles beneficios y muchos posibles riesgos. Algunos ya son palpables –tanto las derivaciones positivas como los peligros– y otros están a punto de llegar, aunque parezcan fantasías de película de serie B.

En definitiva, proporcionaremos al lector muchos datos, transformados en información, con la esperanza de que generen conocimiento. No sabemos si nuestra aportación será valiosa, pero no tenemos ninguna duda de que intentarlo es necesario. Pueden existir datos sin información, pero difícilmente habrá información sin datos. Y aún menos, conocimiento. Para que el imperio de los datos no nos engulla, hay que estar medianamente preparados. Solamente si los ciudadanos tienen suficientes datos y los saben procesar podrán presionar para que la información y el conocimiento que se derivan de ellos sean beneficiosos para la sociedad.

Capítulo 1
VIAJE AL PAÍS DE LOS DATOS

> Había 5 exabytes de información creados desde el alba de la civilización hasta 2003, pero esta información ahora se genera cada dos días.
>
> ERIC SCHMIDT (2010)

> El mundo ya no está dominado por las armas, ni por la energía, ni por el dinero. Está dominado por unos y ceros, por pequeños bits de datos. Todo está en los electrones.
>
> COSMO, personaje de la película *The sneakers* (*Los fisgones*, 1992)

> A lo largo del siglo XX han tenido gran repercusión tres conceptos científicos profundamente desestabilizadores que lo han dividido en tres partes desiguales: el átomo, el bit y el gen. [...] Cada uno tiene su origen en una noción científica abstracta, pero crece hasta acabar invadiendo un gran número de disciplinas humanas y transformando la cultura, la sociedad, la política y el lenguaje.
>
> SIDDHARTHA MUKHERJEE

Fremont Rider levantó la vista para contemplar las estanterías llenas de libros, suspiró e inmediatamente pensó en un futuro más bien negro o, por lo menos, muy complejo. Rider era escritor y bibliotecario de la Universidad Wesleyana en Middleton (Connecticut, Estados Unidos) y en el año 1944 lanzó un grito de alarma respecto a la cantidad de libros que se publicaban anualmente. Calculó que las bibliotecas norteamericanas duplicaban su tamaño cada dieciséis años. Según Rider, a este ritmo, la biblioteca de la Universidad de Yale, una de

las principales del país, tendría, en el año 2040, «aproximadamente 200.000.000 de volúmenes, que ocuparían 9.656 kilómetros de estanterías». El problema no sería solo de espacio, sino también de gestión. Rider calculaba que esta cantidad de libros haría necesario un equipo de más de seis mil personas para catalogarlos.

Más de siete décadas después del aviso de Rider, el problema ya no son tanto los libros editados como el conjunto de la información. Internet ha provocado una explosión de datos. Solamente con los que procesa cada día Google se podrían editar volúmenes suficientes para que, apilados, llegasen a la mitad de camino entre la Tierra y la Luna. Quizá Rider ni tan solo tendría ánimos de calcular cuánto personal se necesitaría para catalogarlos –una sencilla regla de tres con los datos del bibliotecario americano revela que serían más de 118.000 personas.

Afortunadamente, estos datos no se encuentran en papel, sino que más del 90 % se hallan en soporte digital. Desgraciadamente, no tenemos que considerar solo las búsquedas en Google, sino todo lo que se genera en el universo digital en distintos formatos.

De vez en cuando, alguien realiza cálculos parecidos a los Rider, pero ya no se pueden limitar al papel. Además, suelen quedar obsoletos al cabo de poco tiempo. En 1997, Michael Lesk, un informático y experto en sistemas de información, se entretuvo en calcular cuánta información existía en el mundo. Empezó describiendo la Biblioteca del Congreso en Washington, con sus veinte millones de libros, trece millones de fotografías, cuatro millones de mapas, más de medio millón de películas y tres millones y medio de registros de sonido.

Pero Lesk no se podía limitar a una biblioteca, por grande que fuera, ni tan solo al material editado. Añadía que en un año se filmaban miles de películas, se realizaban miles de mi-

llones de fotografías, se emitían millones de horas de televisión y de radio, se editaban más de 400 millones de CD y más de 300 millones de casetes –muchos duplicados, sin duda, porque de algunos se hacían miles de copias–, había billones de minutos de conversaciones telefónicas... Realizando cálculos aproximados y basándose en otras fuentes, señalaba que quizá en el mundo había 12.000 petabytes (PB) de información. Esto significa 12.000 millones de gigas, por usar una unidad de medida que a mucha gente le resulta familiar.

Pese a estas cifras, concluía que en la Tierra habría suficiente capacidad de almacenamiento para todo lo que la gente escribiese, dijese, fotografiase o filmase en el futuro.

De todo ello se cumplen algo más de veinte años y la cantidad de información ha aumentado de forma exponencial. Y parece que sí, que la tecnología, al menos de momento, está solucionando el problema de guardarla e incluso de hacerla accesible. Pero ¿qué utilidad puede tener tanta información? ¿Y cómo podemos gestionarla?

El nacimiento de los datos

Los datos nacen de la necesidad. Hubo datos antes de que hubiese métodos para representarlos de forma comprensible para todo el mundo. Primero fueron los datos y, tiempo después, aparecieron los números. Hace miles de años, un pastor veía que de su corral salían muchas ovejas y que después de pasturar entraban muchas. Pero ¿cómo podía saber si volvían todas?

Para estar seguro de que no perdía ninguna oveja debía tomar una piedra o una ramita por cada una que salía del corral. Y cuando después de pasturar volvían a entrar, debía retirar una piedra o ramita del montón por cada una. Si no

quedaba ninguna, todas habían vuelto. Si quedaban piedras en el montón, alguna se había escapado. Y si seguían llegando ovejas y ya había acabado las piedras y las ramitas, o bien se había descontado, o bien había ganado algún ejemplar extra.

Más tarde llegarían los sistemas para simbolizar las cantidades. Las sociedades evolucionaban, se hacían más complejas. Había más producción agrícola y había más rebaños. Y se hacían intercambios comerciales. Así nacieron los números. No los números actuales, sino otros sistemas simbólicos para representar cantidades. Hace más de cinco mil años ya había fichas de arcilla con símbolos que correspondían a cantidades e incluso a cálculos.

Pero la información, los datos, no era simplemente numérica. Había textos, había representaciones simbólicas, había grabados. Ahorrémonos unos cuantos milenios y saltemos al siglo XV. Con la imprenta, la información editada con libros y documentos estalla y hay quien ve un alud difícil de gestionar. Los primeros escépticos sobre la capacidad humana para asimilar tantos libros no pudieron ver que cualquiera de sus previsiones se quedaba corta en pocas décadas.

Hagamos nuevamente un gran salto. A mediados del siglo XX, la cantidad de información era inmensa y a alguien se le ocurrió que tenía que haber alguna manera de cuantificarla. En 1948, el norteamericano John W. Tukey, matemático y pionero de la informática, creó el bit, como abreviatura de *BInary digiT*. Aparte de la contracción del concepto en tres letras, debía jugar con el significado de bit en inglés, 'pieza pequeña'. Ya tenemos la unidad de información digital.

Al cabo de pocos años, en 1956, el ingeniero electrónico Werner Buchholz –norteamericano nacido en Alemania, de donde se marchó huyendo del nazismo– creó el *byte*. En los

años cincuenta, Buchholz trabajaba en la IBM y formó parte del equipo que diseñó los primeros ordenadores, como el IBM 701. El bit era demasiado pequeño para medir la cantidad mínima de información, un solo carácter, y por eso surgió el byte. Al principio, no había una equivalencia estándar y un byte, según el sistema o el ordenador utilizados, podía variar. Ahora, un byte equivale a 8 bits y por eso a veces se le llama octeto.

Ya tenemos el byte, pero pese a la necesidad de definir la unidad que equivale a un solo carácter, una medida tan pequeña tiene poca utilidad cuando hablamos de grandes cantidades de información. Sería como medir distancias astronómicas en centímetros. Por ello, en seguida surgieron los múltiplos: kilobyte, megabyte... Pero mega, un millón, se queda corto en muchos casos y por eso aparecieron el giga (mil millones) y otros que progresivamente multiplican el anterior por mil: tera, peta, exa, zetta, yotta... Con este último llegamos al cuatrillón.

Explicábamos antes que Lesk había situado en 12.000 petabytes la cantidad de información que había en el mundo en 1997. Pero con estas cifras a mucha gente le pasa como con los presupuestos estatales o con los beneficios de las grandes empresas. Nos pueden hablar de 17.000 millones de euros, de 80.000 millones o de 250.000 millones. Comprendemos que es muchísimo, pero somos incapaces de hacernos una idea.

Por eso, algunas comparaciones serán útiles. Un byte es un solo carácter. Por tanto, una sola letra ocupa un byte. Si creamos un documento con una sola letra, «pesará» un byte. A partir de aquí, el primer paso no es difícil. Un kilobyte (KB) equivale a media página, unos mil caracteres. Y un megabyte podría ser una novela corta.

Hagamos un breve inciso. A menudo leemos que 1 KB son 1.024 bytes. Esto se debe al origen del byte y a que los informáticos trabajan en sistema binario y, por lo tanto, con potencias de dos. Como 2^{10} es 1.024, esta es la equivalencia que se utiliza a menudo en el ámbito de los ordenadores. Pero para el sistema internacional de medidas, 1 KB son mil bytes.

Pero la información no está solo en forma de texto o de cifras. Podemos tener gráficos, dibujos, fotografías... Incluso películas o sonidos. Cada añadido aumenta la cuantidad de información. Una fotografía con buena definición puede ocupar dos megabytes. Es decir, como dos novelas cortas.

Una hilera de diez metros de libros equivale a un gigabyte (GB). Y con seis millones de libros tendríamos un terabyte (TB). Si reuniéramos siete millones de horas de televisión de alta definición tendríamos un petabyte (PB). ¡Y Lesk decía que toda la información que había en el mundo ocupaba 12.000 petabytes! Hoy, en tan solo una hora ya se transmiten en todo el mundo 500 petabytes de información, equivalentes a 6.600 años de vídeo de alta definición o a diez veces todas las obras escritas por la humanidad desde los inicios de la historia.

Todas estas comparaciones son aproximadas. La cuantidad de bytes que tiene un texto también depende de las órdenes de estilo que incorpore –formato, estilo y tamaño de letra...–. Una fotografía puede tener mucha calidad o muy poca y lo mismo pasa con una película. Por otro lado, se hacen comparaciones con cosas muy difíciles de medir con exactitud. Así, se ha dicho que todas las palabras pronunciadas por toda la humanidad a lo largo de la historia ocuparían cinco exabytes (EB). La idea también ha sido rebatida y nuevos cálculos hablan de 42 zetabytes (ZB). Pero es muy probable que nos falten muchos elementos para poder valorarlo con precisión.

Unidades de información y sus equivalencias

(Cada una multiplica por mil la anterior)

1 byte	1 carácter
1 kilobyte (KB)	Media página mecanografiada
1 megabyte (MB)	Una novela corta
1 gigabyte (GB)	Una película de dos horas
1 terabyte (TB)	Seis millones de libros
1 petabyte (PB)	2.000 años seguidos de música
1 exabyte (EB)	100.000 veces todo el material impreso –libros, revistas, documentos– de la Biblioteca del Congreso de Washington
1 zetabyte (ZB)	152 millones de años de vídeo de alta definición
1 yotabyte (YB)	Toda la información que puede contener el centro de datos de la NSA (National Security Agency) de Estados Unidos en Utah, que tiene una superficie de 92.000 metros cuadrados

Toda la información en unos y ceros

Algo que sí se puede calcular con más certeza, aunque también tendrá imprecisiones, es la capacidad de almacenaje de la información. Así, en 1986 se podían guardar, con los dispositivos existentes en todo el mundo, 2,6 exabytes, y en 2007 ya podían ser 295 EB. Esto significa que en 1986 había el equivalente a menos de un CD por persona y en 2007 ya eran unos 61 CD por persona (Marinescu, 2013: 196) –no es una relación lineal porque la capacidad había aumentado, pero la

El imperio de los datos

sin fronteras

Vivimos en un mundo de datos. Los generamos y los recibimos en el móvil, el ordenador, el coche y en los utensilios más diversos, aunque no seamos conscientes de ello. Producimos datos cuando telefoneamos, cuando ponemos un «me gusta» en Facebook, cuando pagamos con tarjeta de crédito, cuando realizamos una búsqueda en internet, cuando nos hacen un reconocimiento médico o, simplemente, cuando nos movemos con el navegador del coche conectado. Hay billones y billones de datos y por eso hablamos de Big Data, megadatos o datos masivos.

Esta obra explica cómo se generan los datos, cómo se procesan, para qué sirven y, sobre todo, para lo que no deberían servir. Así, sin apostar por un mensaje catastrofista, el libro proporciona al lector información y consejos para concienciarlo sobre las grandes oportunidades que implica este imperio de los datos, tanto para la investigación como para la gestión y para otros ámbitos, pero también sobre los peligros y sobre la parte de responsabilidad que tenemos en el uso (y en el mal uso) de datos de todo tipo.

UCC+i
UNITAT DE CULTURA CIENTÍFICA I DE LA INNOVACIÓ
VNIVERSITAT DE VALÈNCIA

VNIVERSITAT DE VALÈNCIA
PUBLICACIONS
PUV

CÀTEDRA DE DIVULGACIÓ DE LA CIÈNCIA
VNIVERSITAT DE VALÈNCIA · Caixa Bank