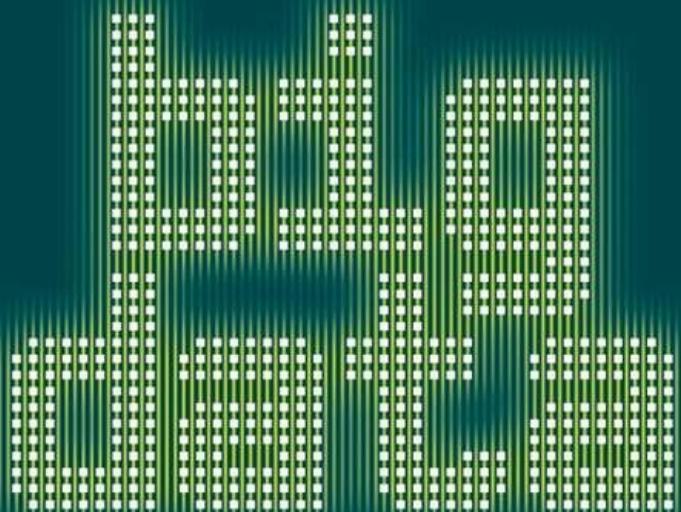


Mario Tascón
Arantza Coullaut



y el Internet de las cosas

*Qué hay detrás y
cómo nos va a cambiar*



CATARATA

Mario Tascón

es uno de los expertos en Internet más conocidos del panorama de medios español y ha dirigido, entre otras, las versiones digitales de El País o El Mundo. Es fundador de Prodigioso Volcán y coautor de los libros Twittergrafía (2011) y Ciberactivismo (2012).

Arantza Coullaut

ha sido periodista en El País y actualmente colabora en diferentes medios escritos. Cofundadora de CTXT y miembro de Prodigioso Volcán, es además directora y guionista de campañas audiovisuales.

Mario Tascón y Arantza Coullaut

Big Data y el Internet de las cosas

Qué hay detrás y cómo nos va a cambiar



diseño DE cubierta: Pablo Nanclares

© Mario Tascón y Arantza Coullaut, 2016

© Los libros de la Catarata, 2016

Fuencarral, 70

28004 Madrid

Tel. 91 532 05 04

Fax. 91 532 43 34

www.catarata.org

Big data y el Internet de las cosas.

Qué hay detrás y cómo nos va a cambiar

isbn: 978-84-9097-776-7

ISBN: 978-84-9097-074-4

DEPÓSITO LEGAL: M-2.884-2016

IBIC: UN/UNJ

este libro ha sido editado para ser distribuido. La intención de los editores es que sea utilizado lo más ampliamente posible, que sean adquiridos originales para permitir la edición de otros nuevos y que, de reproducir partes, se haga constar el título y la autoría.

PRESENTACIÓN

Big Data y el Internet de las cosas

Las empresas, administraciones e individuos tienen cada día más datos disponibles y mejores herramientas para analizarlos, pero la velocidad, el volumen y la fuerza con la que se están dando los cambios que agitan la tecnología configuran además lo que muchos denominan ya la “Cuarta Revolución Industrial”. Esas bases de datos gigantes a las que ya hoy podemos acceder y con las que podemos trabajar se denominan “Big Data”. Sacarles partido es uno de los retos de los próximos años, y parte de la clave para hacerlo está en la capacidad de estudiarlos de manera eficaz y en la capacidad de comunicación y de interoperar con la que podamos dotar a los objetos conectados a la Red. Todo va a estar enlazado; desde una impresora a un semáforo, pasando por nuestro propio coche y el garaje en el que aparcamos, los objetos empiezan a “hablar” entre ellos.

Internet está permitiendo que sobre el mundo físico se desarrolle, a una velocidad nunca vista, un nuevo ecosistema digital en el que los impulsos y la información circulan en forma de Big Data; los cables y las conexiones por ondas son el nuevo sistema nervioso, y ahora, a través del Internet de las cosas, estamos desplegando los sentidos, tanto los semejantes a los humanos de tacto, vista, oído, olfato, orientación y gusto, o los no humanos, como la ecolocalización o la electrorrecepción. Si añadimos a esto las partes móviles (robots y mecanismos que actúan) estamos configurando un nuevo entorno digital que, a su vez, modifica el mundo físico y que será el del futuro inmediato.

Capítulo 1

QUÉ ES BIG DATA

El mundo está construido sobre el poder
de los números.

Pitágoras

Eric Schmidt, antiguo director ejecutivo de Google, asegura que cada dos días generamos tantos datos como todos los que la humanidad entera había producido hasta el año 2003. No sabemos si ese cálculo es muy exacto, pero en cualquier caso sí sabemos que cada día se generan más de mil millones de comentarios en Facebook y que cada minuto se realizan cerca de cien mil transacciones con tarjeta bancaria. Y la mejor noticia es que una gran parte de todo esto se puede analizar y aprovechar.

Big Data es la utilización de grandes cantidades de información, que pueden provenir tanto de la actividad de una empresa como de la de los ciudadanos en su relación con la Administración pública, de su quehacer diario o de sus conversaciones en redes sociales, pero también de las estaciones meteorológicas, los sensores de tráfico desplegados por un ayuntamiento o los coches que circulan por las carreteras. Los datos en cantidades demasiado grandes o que se mueven muy rápido para las bases de datos convencionales son los llamados Big Data. Son aquellos que los sistemas tradicionales no pueden procesar ni almacenar y mucho menos analizar. Se definen a menudo por lo que se denomina las tres V: gran volumen de datos, velocidad de estos datos y variedad del origen de los mismos. Pero hay una cuarta V, cada vez más importante, que es la visualización.

La Fundación del Español Urgente emitió una recomendación sobre el uso del término en español. En ella dice que “la voz inglesa ‘Big Data’, que se emplea en el sector de las tecnologías de la información y de la comunicación, puede ser traducida por ‘macrodatos’”. “‘Big Data’ es una denominación con la que se alude a un conjunto de datos tal que por su volumen, variedad y por la velocidad a la que necesita ser procesado supera las capacidades de los sistemas informáticos habituales”.

El valor del Big Data radica en realidad en el análisis que sea capaz de realizar cada empresa, gobierno o individuo, de los datos disponibles.

Unos resultados que se pueden emplear, entre otros casos, para:

La realización de modelos predictivos de pautas de los consumidores, lectores, votantes o usuarios.

Modelos sobre comportamiento de vehículos en carreteras y ciudades.

El estudio de mecanismos de comportamientos sociales masivos.

Análisis de pautas de compra y consumo.

Ahorrar costes en un proceso industrial al poder detectar anomalías en el desarrollo del mismo que no son visibles con métodos convencionales.

El estudio previo mediante la realización de simulaciones a bajo coste para el lanzamiento de nuevos productos o servicios.

Mejorar los productos y servicios que se ofrecen logrando una mayor personalización, velocidad de fabricación y aumento de la calidad de los productos.

Personalizar de forma automática las recomendaciones de lectura o consumo.

Sugerir relaciones personales o de trabajo.

Mejorar la formación con un mejor conocimiento de las pautas de trabajo de una persona y, por tanto, con una formación personalizada que atienda específicamente a aquello más necesario.

Ofrecer información muy rica en tiempo real sobre situaciones que estén sucediendo para poder tomar decisiones relevantes de forma rápida y eficaz.

A pesar de todo ello, muy pocas empresas, organizaciones y gobiernos saben extraer el verdadero valor de los datos. Según el informe Índice del Valor de la Información, realizado por la consultora PwC, solo un 4% lo hacen. El estudio indica que el 43% de las empresas europeas y norteamericanas obtienen muy poco beneficio tangible de su información y un 23% no extrae absolutamente ninguno. El índice elaborado en ese trabajo, que llega a una puntuación media de 50,1 sobre 100 (un 46,6 en España —que ocupa el último lugar—, un 47,3 en Europa y un 52,9 en Norteamérica), confirma que la gran mayoría de las compañías —independientemente del tamaño, ubicación o sector— tienen un largo camino por delante antes de poder extraer el valor completo de la información que manejan o pueden obtener.

Pero, además, por encima de esas montañas de datos, se están desarrollando comunicaciones entre máquinas y medición de diferentes parámetros desde cada vez más dispositivos, y todo ello sin que exista intervención humana en el proceso.

El tamaño de los datos: los gúgolos

Los propios átomos del Big Data, los datos, no solo son muchos, variados y veloces, muchas veces son inmensos y otras de difícil asimilación por su desmesura. No es solo cuestión de volumen de datos, sino incluso de que cada dato único puede ser gigante. Uno de estos números, el más grande, es un gúgol.

El sonido de la palabra “gúgol” (googol, en inglés) se ha hecho tan familiar por su similitud con la pronunciación de la marca del buscador de Internet que es difícil tratar de convencer a alguien de que poco tiene que ver con la marca de la compañía de California. Aunque, en realidad, un poco sí tiene que ver. Larry Page, uno de los dos fundadores de Google, se equivocó cuando fue a registrar la marca el 17 de septiembre de 1997.

—Sergi, ¿cómo dijimos que se iba a llamar nuestro invento?

—Gúgol, Larry.

—Ok. El dominio está libre... ¡Registrado!

Se había confundido y escribió “Google”. Afortunadamente para Larry y Sergi, la nueva palabra estaba libre como marca y como dirección de Internet. “Googol” estaba registrado. Aquel baile de letras ha popularizado la variante y ha ocultado el significado del original.

Y así nuestro viaje vuelve al principio: ¿qué es un gúgol?

Pues bien, un gúgol es un número inmenso, gigante, casi inimaginable. Se podría representar por un uno seguido de cien ceros. Así, sin saber mucho de matemáticas, nos encontramos por puro sentido común delante de algo gigante, tan grande que nos resulta incomprensible. El primer intento didáctico que uno puede localizar navegando por la Red muestra una comparación: “El número de átomos de hidrógeno del universo no llegaría a conformar un gúgol”. Pero aquí empiezan las dificultades que queremos señalar. Se supone que el lector ha de tener en su cabeza la idea de cómo es de grande el universo o, al menos, saber que el hidrógeno es el elemento químico más abundante, porque, si no, ¿no se hace más compleja la explicación que el propio número?

En la era de los macrodatos o Big Data, de los guarismos incomprensibles por inconmensurables, todos los días leemos titulares que nos explican noticias como la siguiente: “Desde 2007 el rescate de la banca española nos ha costado 220.000 millones de euros”. Pero ¿cuánto son 220.000 millones de euros? ¿Y para los que todavía pensamos en pesetas? Una comparación puede ayudar: a cada español nos ha costado más de 3.000 euros, incluidos bebés, niños y ancianos. Multiplique usted por los miembros de su familia y sabrá (más o menos) lo que le ha menguado esto el patrimonio pecuniario de su hogar.

No solo los periodistas —también los escritores— intentan hacer que esos

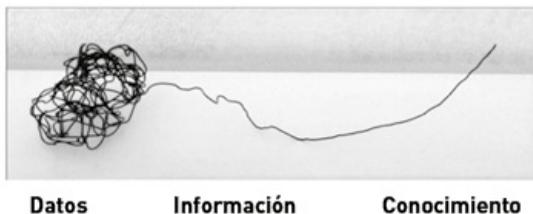
gúgolos sean más amables, más comprensibles: “Más de la mitad del planeta es lluvia en potencia. Cada segundo se evapora el equivalente a seis mil cuatrocientas piscinas olímpicas. Y todo volverá a caer”. Gracias a esa aclaración de Marina Bastos, en su texto para la revista Etiqueta Negra, se entiende mucho mejor qué cantidad de agua se evapora y tenemos una imagen más interesante para comprender. Poetas, matemáticos, periodistas o estadísticos se han de afanar cada día para que entendamos las nuevas cifras, tan grandes que no caben todos los átomos, ni los de hidrógeno ni del resto del universo en ellas, de una magnitud tal que nuestra cabeza necesita de explicadores de gúgolos para sobrevivir en este mar de datos. Veremos a nuevos carpinteros que con esa materia gigante construyan nuevas formas que nos traigan la comprensión de lo que ahora son meros números que se nos escapan.

Traducir, calcular, comparar y condensar son tareas necesarias para quienes trabajen con Big Data.

Datos, información y conocimiento

Un dato (Davenport y Prusak, 1999) es una representación simbólica (numérica, alfabética, algorítmica, etc.) de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades. Los datos son la mínima unidad, son elementos primarios de información que por sí solos son irrelevantes y no suelen decir nada sobre el porqué de las cosas. Un dato es un número de teléfono, un apellido, una hora, la coordenada de un lugar. Los datos pueden estar en la mente de una persona, en papel o en algún tipo de soporte digital. Generalmente, al hablar de Big Data nos referimos a los datos digitales que permiten el procesamiento de la información. En este sentido, la tercera acepción de “dato” para la Real Academia Española recoge su significado en la era del Big Data: “Información dispuesta de manera adecuada para su tratamiento por un ordenador”. Así, no sirve el dato puro si no está preparado adecuadamente para procesarse mediante software.

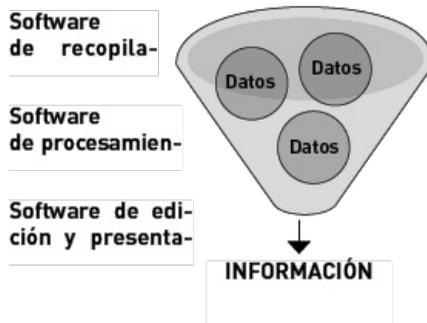
Camino al conocimiento



La información, por otro lado, se puede definir como un conjunto de datos procesados que tienen un significado (relevancia, propósito y contexto) y que, por lo tanto, son de utilidad, por ejemplo, en la toma de decisiones al ayudar a disminuir su incertidumbre.

El procesamiento de los datos

produce información



Para que los datos se conviertan en información hay que añadirles valor mediante alguna de las siguientes maneras:

Cálculo: los datos pueden haber sido procesados matemática o estadísticamente.

Categorización: se conocen las unidades de medida que ayudan a interpretar los datos.

Condensación: los datos se han podido resumir de forma más concisa (agregación).

Contextualización: se aporta el contexto y el propósito con el que se generaron.

Corrección: se eliminan los errores e inconsistencias de los datos.

Por tanto, la información es la comunicación de conocimientos o inteligencia, y es capaz de cambiar la forma en que el receptor percibe algo, impactando sobre sus juicios de valor y sus comportamientos. Davenport y Prusak, consultores y analistas norteamericanos, definen el conocimiento como una combinación de experiencia, valores, información y “saber hacer” que sirve como marco para la incorporación de nuevas experiencias e información, además de ser útil para la acción. Se origina y aplica en la mente de los individuos. En las organizaciones no solo se encuentra dentro de documentos o almacenes de datos, sino que también está implícito en los procesos, prácticas y normas.

Todo lo que nos rodea son datos. Como señala Wolfram Rozas, director académico de los programas de Big Data e Inteligencia de Negocio en EOI y Big Data Business Development Executive en IBM, “cada uno de los elementos en este nuevo mundo interconectado, sean seres vivos (hasta algunos ejemplares de los animales en peligro de extinción ya llevan chips) o

máquinas, emite datos en tiempo real y en múltiples formatos. Formatos cada vez más grandes y pesados. Hace unos años las transmisiones eran solo asequibles para gobiernos y empresas —empezaron enviando letras de texto (bytes como unidad de medida)—, pero ahora todos podemos enviar fotos de calidad (alrededor de 1-3 millones de letras o bytes) o pequeños vídeos (100-200 Mb)”. Twitter administra cerca de 12 terabytes de tuits diariamente y Facebook almacena alrededor de 100 petabytes de fotos y vídeos.

Los números marean: el informe Cisco Visual Networking señala que en 2019 habrá 3.900 millones de usuarios de Internet, aproximadamente el 51% de la población mundial prevista. Millones de usuarios manejando millones de datos. Estamos a las puertas de la era zettabyte: llegaremos a los 2 zettabytes en 2019. O, lo que es lo mismo, el tráfico IP global alcanzará los 168 exabytes mensuales, prácticamente multiplicándose por cinco desde los cerca de 31 exabytes mensuales registrados en 2011.

¿De dónde provienen tantos datos? Proceden de muchas fuentes: desde sensores utilizados para recoger información meteorológica, mensajes en medios sociales, fotografías y vídeos digitales hasta registros de transacciones de compra. Los datos se han vuelto enormes y Wolfran distingue tres categorías:

Datos estructurados, que son almacenables en filas y columnas, el tipo de datos más empleado por los sistemas de registro, que guardan las transacciones, los censos o las encuestas.

Datos semiestructurados que no se ajustan a un esquema fijo y explícito; documentos xml, los blogs o los sensores que emplean este tipo de datos.

Datos no estructurados, los más complejos, aquellos que se presentan en un formato que no puede ser fácilmente indexado en tablas relacionales para el análisis; son los datos de las imágenes, audio, vídeo o los de las redes sociales.

La consultora McKinsey sostiene que “Big Data se refiere a conjuntos de datos cuyo tamaño hace que estén fuera del dominio de las bases de datos relacionales para capturar, almacenar, gestionar y analizar”.

Clasificación de datos para IBM

Pero sigamos con otra clasificación. Según IBM, los tipos de datos utilizados más frecuentemente en trabajos de Big Data son:

1. Datos de Internet y redes sociales: Incluye contenidos web e información que es obtenida de lugares como Facebook, Twitter, LinkedIn, etc. En algunas ocasiones, también de blogs, wikis y otras plataformas.

2. Datos máquina a máquina (Machine to Machine, M2M): M2M son unas siglas referidas a las tecnologías que permiten conectarse entre dispositivos. M2M utiliza aparatos como sensores o medidores que capturan algún tipo magnitud en particular (velocidad, temperatura, presión, variables

meteorológicas, variables químicas como la salinidad, etc.) y transmiten esos datos a través de redes, sean estas alámbricas, inalámbricas o híbridas a otras aplicaciones que traducen estas variables en información significativa. El primer tipo de máquina específica conectado a Internet, por ejemplo, fue una máquina de Coca-Cola en la Universidad Carnegie Mellon a principios de 1980. Los profesores y estudiantes de programación podían conectarse a la máquina a través de Internet para comprobar la disponibilidad de latas frías antes de decidirse a hacer el viaje a la máquina expendedora.

3. Big Transaction Data: incluye registros de facturación o los registros detallados de las llamadas (CDR, Call Detail Record) que contienen información sobre su origen, destino y duración. Estos datos pueden ser usados para el control de la Red, contabilidad y facturación. “Como los teléfonos móviles se han convertido en omnipresentes, la extracción de los datos que generan puede realmente revolucionar el estudio del comportamiento humano”, afirma Ramón Cáceres, investigador principal de los laboratorios de AT&T en Nueva Jersey.

Estos datos transaccionales están disponibles en formatos tanto semiestructurados como no estructurados.

4. Biométricos: información procedente del cuerpo humano y su actividad en la que se incluyen huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc.

5. Generados por humanos: las personas generamos diversas cantidades de datos como la información que guarda un call center al establecer una llamada telefónica, notas de voz, correos electrónicos, documentos electrónicos, estudios médicos, etc. Un concepto relevante respecto a estos datos que generamos es el de huella digital, que es como se denomina al “rastros” que dejamos los usuarios al usar Internet y las redes sociales y que configura nuestra identidad en la Red. Las fotos y vídeos que subimos, cuando nos etiquetan en las redes sociales, los mensajes que publicamos, las búsquedas que realizamos, los “me gusta” que damos...

Las cuatro V

Como señalamos anteriormente, cuatro son las características que definen el Big Data: la velocidad, la variedad, el volumen de los datos y la visualización. Estas cuatro variables sirven para analizar y visualizar la información:

Velocidad: otra de las características importantes de los macrodatos es la velocidad a la que se mueven, se generan y se procesan. Vivimos en la época del tiempo real.

Variedad: en grandes volúmenes, se incluyen datos de cualquier tipo (estructurados y no estructurados) tales como textos, valores, localizaciones, audios o vídeos, entre otros. Al analizar estos datos juntos se encuentra información nueva.

Volumen: las administraciones y empresas poseen y hacen pública una cantidad cada vez mayor de datos de todo tipo. Los tamaños de esos archivos

son gigantescos.

Visualización: poder visualizar la información que extraemos de los datos de una forma comprensible es otra de las funciones necesarias del Big Data. Es una herramienta de difusión, pero también de análisis. La visualización ha ido adquiriendo cada vez más importancia.

Las cuatro V



CAPÍTULO 2

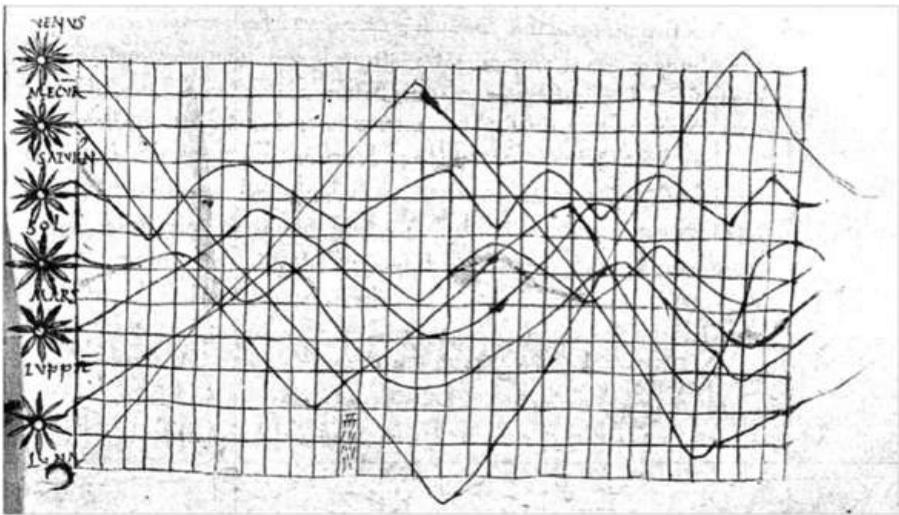
LA VISUALIZACIÓN DE DATOS

Marco Bressan, científico de datos del banco BBVA, cita en alguna de sus conferencias un antiguo caso en el que la visualización de datos ayudó a modificar el comportamiento de todo un país. Sucedió en Francia a comienzos de la década de los setenta, los accidentes automovilísticos y atropellos provocados por el aumento del parque de vehículos se habían disparado. El Gobierno lanzó una campaña publicitaria denominada “Mazamet villa muerta”, en la que mediante unas impactantes fotografías se veía a los habitantes de ese pequeño pueblo tumbados en el suelo sin moverse, como si hubieran sido atropellados y muertos todos en sus calles. Mazamet tenía en aquel momento 16.000 habitantes, el mismo número que los fallecidos por accidentes en Francia. Cada año desaparecían en el país galo pueblos enteros como Mazamet. Aquella forma tan visual de mostrar a la población francesa lo que estaba sucediendo, acompañada, por supuesto, de nuevas leyes y una campaña educativa consiguieron el punto de inflexión en una estadística que llevaba años aumentando. A partir de aquel momento el número de muertos anual en Francia disminuyó. Las imágenes ayudaron a entender la magnitud de la tragedia y sirvieron de forma muy eficiente y rápida para la educación vial de conductores y peatones.

La visualización de datos tiene una historia paralela a la de la infografía, donde se suele incluir. La importancia que ha empezado a tener hoy por la irrupción de nuevas herramientas y el gran volumen de datos ha hecho que se empiece a distinguir entre infografía, para referirse a los diagramas y esquemas, y visualización de datos, que comprende la representación con diversas tipologías de gráficos de datos, incluidos mapas sobre los que se georreferencian valores y, claro está, en muchas ocasiones con movimiento e interactividad, ya que el destino final lo permite, porque están hechos para ser consultados en dispositivos electrónicos.

BREVE HISTORIA DE LA VISUALIZACIÓN DE DATOS

La visualización de información es la línea de representación de la información que trata de manera más específica de cómo presentar cantidades y flujos numéricos. Históricamente ha estado muy asociada a la estadística, sobre todo a partir del siglo XVIII, y desde antes al mundo



Cuando en 1854 se produjo en Londres uno de los brotes de cólera más violentos de la historia de Inglaterra y unas 700 personas fallecieron en menos de una semana en un barrio (el Soho) de apenas medio kilómetro de diámetro, el doctor Snow, que tenía allí su consulta, empezó a dibujar en un mapa unas líneas negras que marcaban el lugar de las distintas muertes; igualmente, señaló en el mapa las fuentes de agua públicas, de tal forma que la visualización mostraba con claridad que la mayoría de los muertos estaban al lado de una de las bombas de agua. Fue una de las primeras representaciones de información georreferenciada y, desde luego, ha sido la que se ha llevado la fama como una de las primeras en las que se unía la incidencia de un fenómeno con la intensidad del mismo en el punto geográfico preciso. Este mapa fue decisivo para las tesis de Snow y sentó la base de los métodos geográficos de control de las epidemias.

Es mucho menos conocido el trabajo del español Rodrigo Onofre, que ya 50 años antes que Snow señalaba en un mapa de Málaga los lugares donde habían fallecido los enfermos de epidemias de fiebre amarilla que había azotado la ciudad en 1803. Sobre el mapa original hemos resaltado las cruces para que pueda percibirse mejor la geolocalización de los puntos.

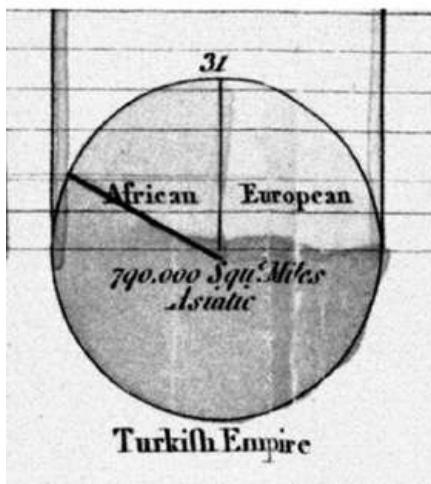


El mapa del doctor Snow. Los puntos representan muertes y las cruces las bombas de agua.



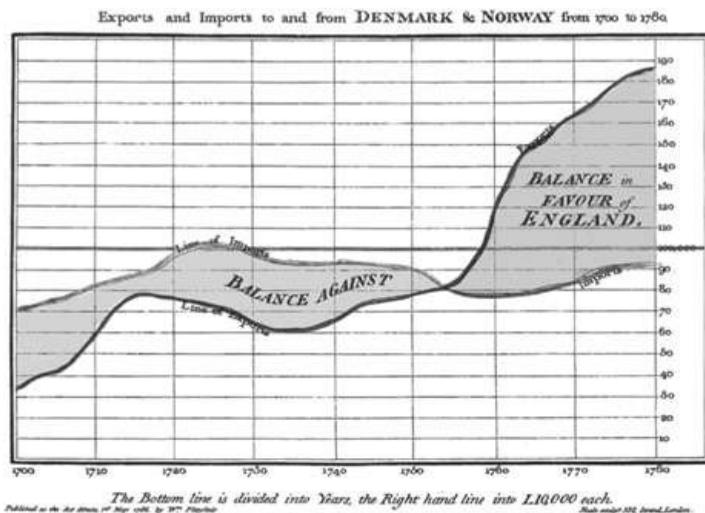
“Breve descripción de la fiebre amarilla padecida en Cádiz y pueblos comarcanos en 1800, en Medinasidonia en 1801, en Málaga en 1803 y en esta misma plaza y varias otras del Reyno en 1804”, por don Juan Manuel de Arejula.

Por esa misma época, el estadístico y delineante escocés William Playfair era el primero en introducir algunos de los gráficos estadísticos más comunes hoy día. En 1801 aparecía de su mano el que se considera el primer gráfico de tarta.



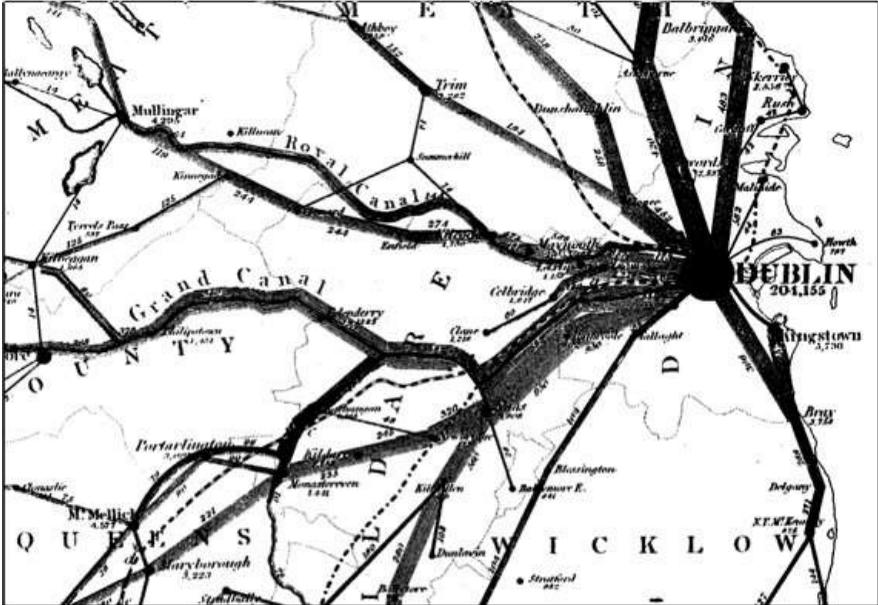
fuente: REPRODUCCIÓN DEL 'ATLAS COMERCIAL'.

Pero ya antes había utilizado gráficos de barras y de polígonos en su Atlas Comercial, publicado en 1786, del que podemos ver adjunta una muestra.



fuente: REPRODUCCIÓN DEL 'ATLAS COMERCIAL'.

Algo más tarde, en 1837, Henry Drury Harness imprimió unos mapas de flujo que hoy nos parecen modernos, pero que anticiparon en 200 años los actuales generados por ordenador. Drury representaba para la Comisión de los Ferrocarriles Irlandeses el tráfico de viajeros del transporte en Inglaterra, haciendo corresponder la anchura de las rutas con el número de viajeros que las transitaban.



The 1837 Henry Drury Maps, de Arthur H. Robinson.

Desde Drury saltamos a Minard, que fue el autor de una de las infografías más aplaudidas en la actualidad, ya que es la primera que representa en el mismo espacio una cronología (la marcha del ejército de Napoleón hacia Moscú), a la vez que el tamaño de la tropa (el ancho de las líneas como las de Drury), y todo ello sobre un mapa y, además, relacionado con la temperatura del famoso “general Invierno”. El mapa es hoy un icono de la visualización de datos.

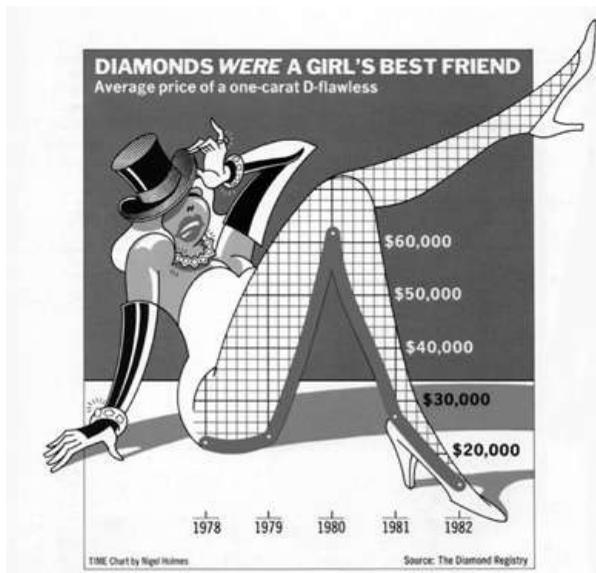
requerida.

La infografía vuelve a dar un salto en su popularización cuando, a finales de los años ochenta del pasado siglo, se empiezan a popularizar entre los profesionales herramientas informáticas que permiten desarrollar gráficos más precisos y trabajar de manera colaborativa. Son ejemplos de esa época las infografías de periódicos como el USA TODAY; sus Snapshots, gráficos estadísticos que aparecen en su primera página, se convierten gráficamente en una tendencia mundial.



UN EJEMPLO DE SNAPSHOT del periódico USA TODAY.

Los Snapshots intentan, mediante formas más atractivas, hacer más “digerible la información estadística.



UN GRÁFICO CLÁSICO DE NIGEL HOLMES PARA LA REVISTA TIME.

Otro ejemplo famoso de esa tendencia son los gráficos del infografista Nigel Holmes, por aquel entonces director de arte de la revista Time. Su escuela e influencia ha seguido hasta hoy, aunque siempre ha sido muy criticada porque muchos piensan que el componente de “golosina visual” es muy superior al de la información que transmite.

El Big Data y la representación visual de información

La llegada del Big Data ha traído al campo de la visualización nuevas herramientas, que son muy necesarias, puesto que las características del nuevo escenario (fundamentalmente las cuatro V a las que ya hemos aludido) implican problemáticas diferentes.

Así en los últimos años han aparecido visualizaciones como las siguientes:

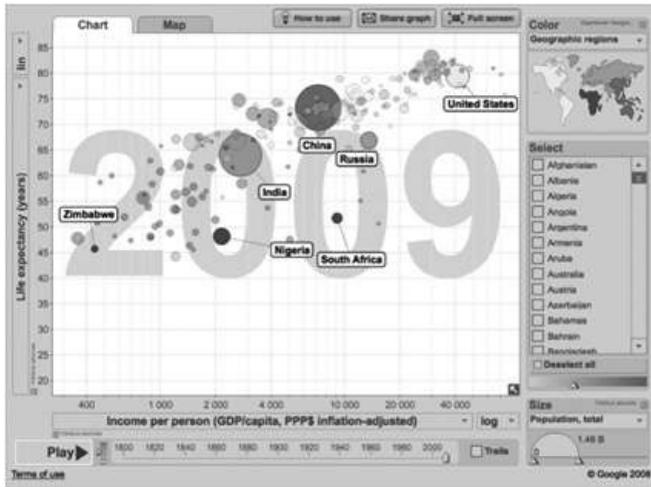
Nubes de palabras

Con diferentes formas se extraen los términos más frecuentes en un libro, discurso, texto... Las palabras más repetidas se resaltan por tamaño y color.

Fuente: Wikipedia.

Gráfico de burbujas

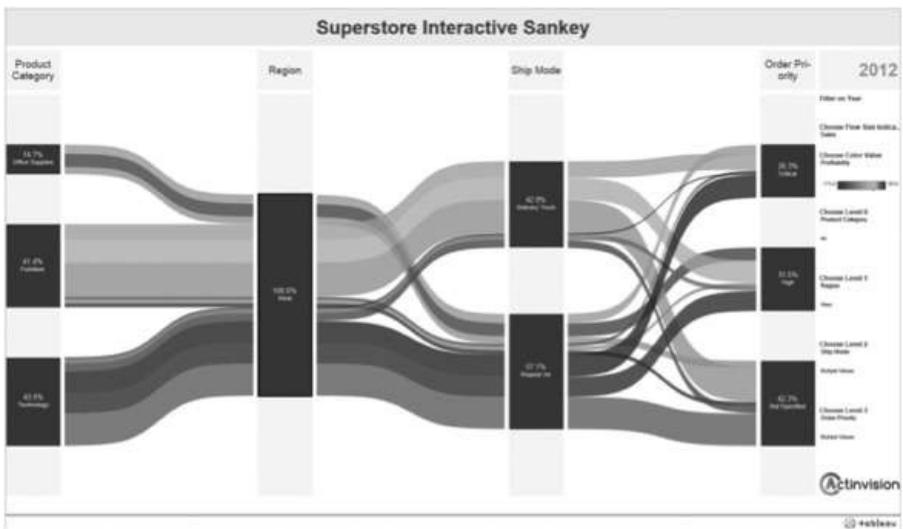
Las burbujas representan con sus diámetros una variable que se añade a dos que indica con su posición.



Fuente: Google Public Data Explorer.

Diagrama de Sankey

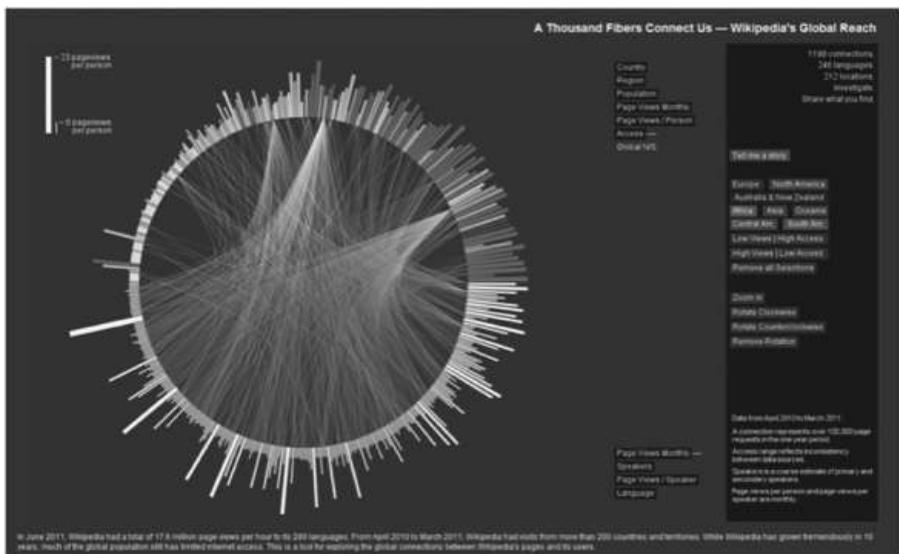
Este diagrama de flujos, que en realidad es muy antiguo ya que se inventó en el siglo XIX, ha vuelto a utilizarse gracias a que las nuevas herramientas han hecho que sea más fácil realizarlo que a mano. Los anchos de las flechas representan el volumen.



Fuente: Activision.

Rueda de acordes

Se utiliza para representar las relaciones en una matriz de datos.



Fuente: Tableau.

Árbol de nodos

Gráficos interactivos sobre resultados electorales publicados en las versiones digitales de El País a lo largo de 2015 y 2016.

Un buen muestrario de trabajos modernos de visualización de información puede encontrarse entre los que realizó el infografista Rafa Hörh, junto a desarrolladores de varias empresas de datos, para el diario español El País a lo largo de las sucesivas elecciones de 2015. En las páginas web y las versiones móviles se desplegó todo tipo de modalidades gráficas para poder explicar los resultados electorales tanto en tiempo real durante los conteos de los votos como, posteriormente, para facilitar la comprensión de los mismos a los lectores. Los millones de datos procedentes de los colegios electorales de cada comunidad, ciudad y provincia se convertían en visualizaciones sencillas de entender, en las que era mucho más fácil encontrar aquello que se buscaba, y comprender qué era lo que sucedía en el panorama político español, que hacerlo entre los datos en bruto que facilitaba el Ministerio del Interior.

EL TRABAJO CON BIG DATA

De forma esquemática, el trabajo con macrodatos sigue este diagrama sintético:

Recolección de datos → Almacenamiento → Análisis → Visualización

Vamos a detallarlo por partes:

1. Recolección de datos: una de las disciplinas que más ha variado en menos tiempo es la de la recogida de datos e información. Cada vez hay más datos públicos que se producen en cifras millonarias (datos personales en las redes sociales) y dispositivos desperdigados por todo el planeta que emiten, procesan y recogen números de las más diversas actividades (posicionamiento de vehículos e individuos en ciudades o mediciones de contaminación). No hace falta ir muy lejos para encontrar este tipo de aparatos, ya que uno de los principales, el teléfono, se encuentra en nuestro bolsillo.

Muchas empresas se están especializando en la recogida de esos datos por medio de diversos programas que, en muchos casos, añaden algo de inteligencia y consiguen que los números recolectados sean coherentes y homogéneos para poder realizar mejor los procesos posteriores. Las arañas de un buscador de noticias estarían en este campo.

2. Almacenamiento: ante el gran volumen de datos, se debe llevar a cabo un almacenamiento escalable. Muchas compañías tienen archivados sus datos, pero no saben cómo procesarlos. El llamado almacenamiento escalable tiene que ser mucho más transparente para poder permitir su ampliación, que es el objetivo de este tipo de arquitectura tecnológica.

EMC, por ejemplo, sugiere optar por infraestructuras orientadas directamente a Big Data a través de almacenamientos que incluyen tecnología cloud, en los que se puede disponer de un repositorio único de más de 10 pe-

tabytes en lugar de islas de datos, que suelen requerir una administración manual y además están soportadas en sistemas dispares. La llegada del modelo de almacenamiento y trabajo en “nube” está también ayudando a la consolidación del nuevo mundo del Big Data.

3. **Análisis:** es una de las áreas específicas para sacarle provecho a Big Data, porque solo con su análisis se puede revelar información que parecía oculta en el almacenamiento de los datos. La clave para obtener resultados es ser capaz de procesar y, además, hacerlo en un tiempo razonable, con lo que nos alejamos de los estudios de mercado estáticos.

Las soluciones tradicionales de análisis de datos suelen ser predefinidas y lentas, con lo que ante un incremento del volumen de los datos y una variedad en su origen proporcionan una información limitada porque solo pueden analizar terabytes de datos estructurados y agrupados.

Por ello, los expertos están recomendando aplicaciones diseñadas específicamente para Big Data, porque ofrecen una analítica más ágil y proactiva de este tipo de información que se relaciona entre ella.

Algunas de estas soluciones son, entre otras, Apache Hadoop, un framework de código abierto para el procesamiento, el almacenamiento y el análisis de grandes volúmenes de datos de diversas fuentes. Es una solución que usan Amazon, Facebook, Google, IBM, Intel Research!, entre otros.

Otro software de código abierto que se utiliza para el análisis es el software R, un lenguaje de programación de código abierto que facilita tanto el análisis de datos como el desarrollo de nuevo software de estadística.

En un análisis de la revista Computing se recomienda que, a la hora de formar un equipo de análisis en las organizaciones empresariales, este debe orientarse hacia esquemas más colaborativos, formados por los administradores de bases de datos del departamento TI junto con equipos dedicados a la ciencia de los datos, creándose así un nuevo perfil de profesional: los científicos de datos.

4. **Visualización:** los gráficos y los mapas interactivos son algunas de las herramientas utilizadas para visualizar estos análisis de datos. A través de estos mapas, el usuario incluso puede realizar su propio filtrado para organizar la información según le convenga, como, por ejemplo, en los mapas del crimen de Stamen.

En el caso de las gráficas, un ejemplo puede ser la tecnología desarrollada por Tableau Software. Aunque en su origen se creó solo para datos estructurados, ha sabido encajar en sus gráficas los datos no estructurados, como los mensajes de Twitter. En principio no precisa formación avanzada para manipular los datos.

Profesión: expertos en datos

En 2018, la consultora McKinsey calcula que serán necesarios entre 140.000

y 190.000 data scientist en Estados Unidos. En todo el planeta, Gartner Group señala que harán falta 4,4 millones de expertos en datos. ¿Qué se necesita para ser un analista Big Data? Aunque en los primeros años el perfil de un experto en datos era el de matemático, estadístico o informático, la cosas van, poco a poco, cambiando.

El artículo “El ‘Big Data’ arrasa en las escuelas de negocios” del periódico Cinco Días, publicado en 2015, refleja muy bien al auge de esta profesión: “Los expertos en data analytics escasean. Se trata de un perfil tan demandado como joven. Por eso las principales escuelas de negocios del país y la mayoría de las universidades han puesto en marcha, o están en proceso de hacerlo, estudios específicos relacionados con esta disciplina. Desde cursos de pocas semanas hasta másteres completos de 340 horas, los centros más punteros ya han hecho suya esta disciplina. El Big Data ha conquistado las escuelas de negocios”, destaca el artículo que recoge también las tres funciones básicas, según la compañía Acceso, que han de tener estos profesionales:

- Científico de datos. Es el encargado de traducir la información en conocimiento para que los analistas de negocio puedan tomar decisiones ágiles y con más inteligencia. Estos profesionales aplican técnicas avanzadas de análisis de la información (estadísticas predictivas y prescriptivas) y desarrollan pruebas de concepto rápidas ante nuevas preguntas de negocio, determinando la viabilidad y fiabilidad de las respuestas.
- Ingeniero de visualización. Este técnico diseña nuevas metáforas interactivas que den soporte al responsable de datos para encontrar eficientemente las respuestas más adecuadas en el nuevo proceso de análisis. Esto también permite al analista de negocio contextualizar y explicar la historia subyacente a los datos de la manera más eficaz desde el punto de vista del problema de negocio a resolver.
- Analista de negocio 2.0. Este profesional se sitúa entre los científicos de datos, los ingenieros de visualización y el consumidor final del análisis. Dotado con habilidades para procesar la información, tiene capacidad de análisis exploratorio y comprensión de los nuevos modelos de análisis visual. Debe ser además un buen narrador, ya que el storytelling es la clave para recorrer la parte final del camino de la información. El analista de negocio 2.0 es el encargado de traducir todo el trabajo anterior en propuestas concretas.

CAPÍTULO 3

Áreas e industrias modificadas

El sector de distribución es uno de los primeros que ha adoptado sistemas de Big Data, al igual que el financiero, pero también las aseguradoras y la Administración pública están poniendo en marcha proyectos orientados a sacar rentabilidad de los datos.

Asimismo, aunque las empresas de más de 500 empleados son las más avanzadas en conocimiento, las compañías de entre 100 y 500 empleados son las que lideran la adopción del modelo. De hecho, las consultoras estiman que un minorista, con la aplicación de sistemas Big Data, podría tener potencial para aumentar su margen de operación en más de un 60%. Aunque todo apunta a que los sectores de informática y productos electrónicos, los de información, finanzas, seguros y la administración son los que más beneficios pueden obtener en un menor plazo.

En nuestro país, uno de los principales obstáculos que puede presentar Big Data para las empresas españolas es la cuestión económica, junto con la falta de conocimiento del sector y la ausencia de la calidad de la materia prima. Estos son los tres factores que, según Curto Díaz, autor del libro *Introducción al Business Intelligence* (UOC, 2012), impiden a las empresas embarcarse en un proyecto de estas características.

Quizás algunos ejemplos prácticos nos ayuden a comprender mejor qué es y para qué sirve el Big Data, cómo está cambiando diferentes sectores y cómo podemos aprovechar mejor los datos en nuestra actividad económica, académica o profesional.

PERIODISMO DE DATOS

En los últimos años son múltiples las iniciativas que se desarrollan en medios de comunicación alrededor del periodismo de datos, que no es otra cosa que el uso de las técnicas de Big Data en el periodismo de investigación.



En la década de los setenta, el antecedente del periodismo de datos se denominaba periodismo de precisión, que el profesor José Luis Dader definía en 1995 como “la información periodística que se basa o desarrolla e interpreta datos empíricos recolectados, relacionados y verificados mediante métodos científicos de investigación socioestadística o informática”.

El periodismo de datos salió a la luz en 1967 al hilo del seguimiento de los conflictos raciales de Detroit, que se saldaron con 43 fallecidos. El equipo del Detroit Free Press fue galardonado el siguiente año con el prestigioso Premio Pulitzer por su cobertura de los hechos, pero sobre todo por su “precisa investigación sobre las causas subyacentes en la tragedia”.

Los datos que se manejaban en aquellos tiempos no tenían todavía el volumen ni la variabilidad de los que hoy existen, pero ya necesitaban del auxilio de herramientas básicas como bases de datos primitivas u hojas de cálculo.

Es en la primera década de este siglo cuando el fenómeno, paralelo al Big Data, se desarrolla en muchas partes del mundo. En 2014 se lanzaron en Estados Unidos varios sites basados en el periodismo de datos.

Definir qué es el periodismo de datos se convierte en relevante cuando uno de los medios que apuesta por este tipo de información, Vox.com, decide dedicar un artículo a describir sus características.

En el artículo, Melissa Bel destaca que, del mismo modo que hay un tipo de informaciones basadas en entrevistas o documentación, este periodismo parte de datos en bruto. Después, su visualización puede adquirir diferentes formatos: vídeos, gráficos, infografías, interactivos, texto...

Una de las voces más autorizadas en este campo es Simon Rogers, periodista e infografista experto en datos que empezó en el diario británico The Guardian para crear el “periodismo de precisión en el siglo XXI”.

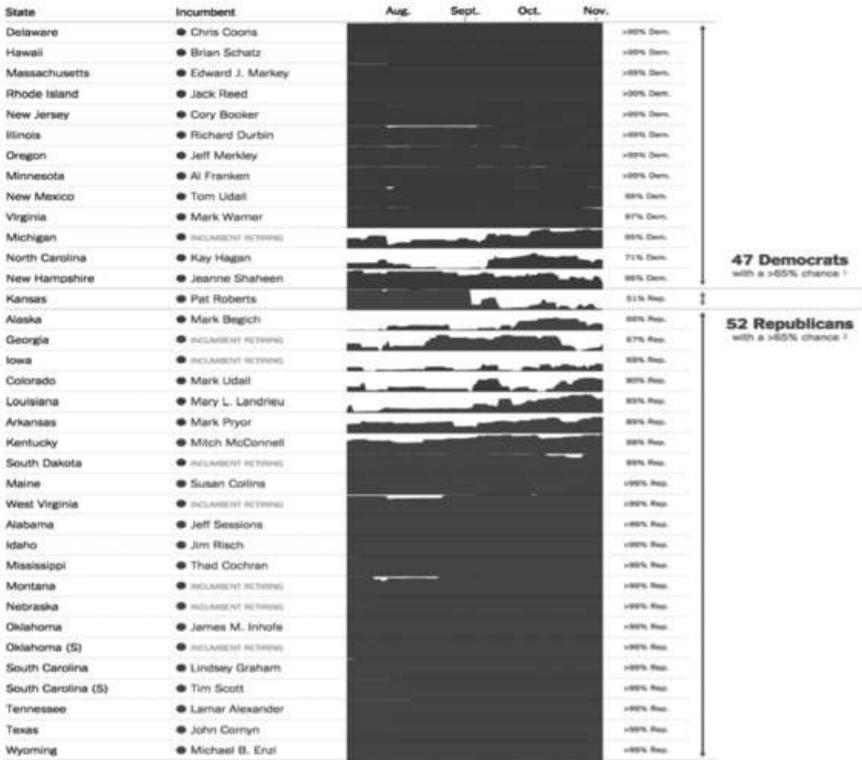
Rogers fue el encargado de formar a un equipo de periodistas especializados capaces de elaborar y poner a disposición del diario piezas periodísticas que permitan ampliar y enriquecer la interpretación habitual de las noticias. El

éxito de su trabajo le llevó a trabajar posteriormente como editor de datos en Twitter y actualmente editor de datos de Google.

The Upshot, lanzado en 2014 por The New York Times, es otro relevante proyecto de periodismo de datos. Fue dirigido por David Leonhardt, ganador de un Pulitzer. El objetivo del sitio web es explicar las principales noticias de actualidad a través de gráficos generados a partir de datos y estadísticas. Upshot se define por un diseño limpio y una gran coherencia temática. Los datos se presentan de forma clara y sencilla, con objeto de que la información pueda entenderse de manera rápida. Un buen ejemplo de este tipo de presentación es el que acompaña el artículo¹ “Who will win the Senate?”.

How Our Forecasts Have Changed

Our state-by-state forecasts are recalculated daily, adding new polls and responding to changes in the political climate. Below, how they have varied over the last four months. NOVEMBER 4, 2014



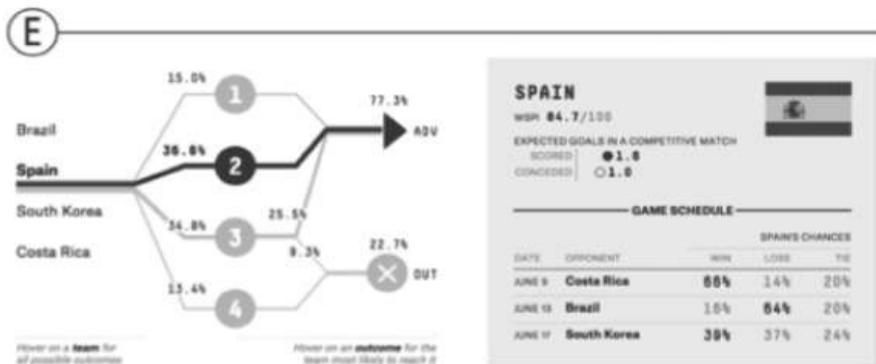
1. Democratic total includes two independents who caucus with the Democrats and 34 continuing seats.
 2. Republican total includes 30 continuing seats.

<p>TheUpshot Senate Forecasts</p>	<p>Overview A summary of our current forecasts and how we calculate them.</p>	<p>Other Forecasts We compile the latest ratings from other organizations into one scoreboard.</p>	<p>Changes See how our state-by-state forecasts have changed over time.</p>	<p>About A detailed explanation of everything that goes into making our forecasts.</p>
--	--	---	--	---

Evolución de predicciones de voto al Senado de Estados Unidos.

Fuente: NYT.

La web apareció después de que el periodista Nate Silver decidiera llevarse su proyecto “Five ThirtyEight” del The New York Times al grupo mediático norteamericano ESPN. “Five ThirtyEight”, que se define como una organización de periodismo de datos, ha configurado un espacio que combina, vinculándolos, el periodismo y los datos.



fuentes: Evolución de los resultados de España en el mundial de fútbol de 2014 según la web FiveThirtyEight.

Este medio tuvo su origen en el éxito de los pronósticos que su creador, Nate Silver, especialista en análisis de datos, alcanzó durante la segunda campaña presidencial de Obama, en la que predijo casi al milímetro el número de votos que le darían la victoria (acertó en 49 de 50 estados). Silver estaba entonces en The New York Times. El periodista es autor de un importante libro de divulgación, *La señal y el ruido* (2012), en el que plantea varias tesis basadas en que el ser humano “está obligado a planificar. A prever lo que podría ocurrir para estar preparado. Pero el mundo cada vez va más rápido, y la información de que disponemos se acumula a un ritmo cada vez mayor. Cualquier intento de organizar los datos que nos llegan y de utilizarlos para dilucidar qué podría ocurrir a continuación puede llevar al colapso y al aturdimiento”. En este libro intenta mostrarnos, en medio del universo de datos que nos rodean, cómo discriminar la información valiosa de la que no lo es.

Los periodistas que trabajan con datos suelen seguir un proceso como el que se describe en la siguiente figura. Partiendo de una idea se buscan los datos que, en muchos casos, hay que limpiar para que sean consistentes. A partir de esa homogeneización, se procesan y mejora nuestro material primigenio. Normalmente, estas son el tipo de tareas que se realizan en esas hojas de datos:

Convertir a unidades comunes. Los datos, al provenir de fuentes diversas, pueden estar unos en dólares y otros en euros, o unos en pulgadas y otros en centímetros.

Convertir a unidades comprensibles. Los periódicos suelen hacerlo, incluso en sus artículos, cuando al lado de una cantidad en dólares, por ejemplo, añaden su equivalente en euros si esta es la moneda corriente para sus lectores.

Redondear. En muchos casos, los decimales se redondean para hacer los

números más comprensibles y sencillos, sobre todo a la vista. El redondeo en el caso de grandes cantidades de números ha de realizarse con mucho cuidado y, si es posible, solo con fines de hacer las visualizaciones más simples, ya que una pequeña variante en un número, al estar hablando de estas cantidades, podría generar distorsiones significativas.

Simplificar combinando elementos. Combinar elementos puede crear números más comprensibles; por ejemplo, si queremos hablar de gastos en lotería por provincia puede ser más interesante el gasto por habitante, que sería la manera de poder comparar sin la distorsión que ofrece el volumen de habitantes de cada una de ellas, que nos llevaría a una apreciación errónea.

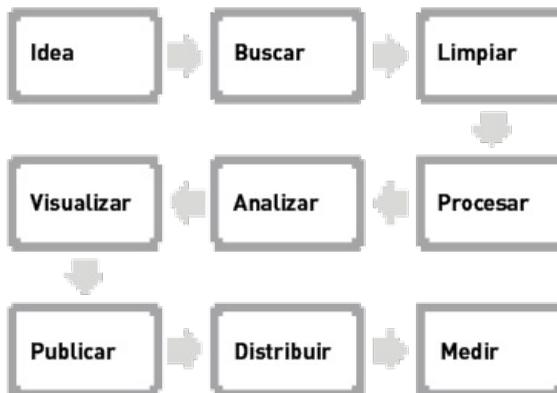
Reagrupar elementos.

Reordenar elementos. La simple ordenación de mayor a menor en una columna o gráfico ayuda a mejorar la comprensión.

Simplificar los datos.

Comprobar y cotejar con el original. Cuando se hacen cálculos, hay que estar siempre atento a no cometer errores y conviene comprobar que todo es coherente con la fuente original.

pasos del Periodismo de datos



Posteriormente, una vez consolidados y homogeneizados, se analizan y se visualizan en muchos casos por medio de infografías. Realizado todo ese trabajo y si en esta parte ha resultado algún tema o noticia interesante, se publica en el medio correspondiente. Hoy ha adquirido también mucha importancia la capacidad de medida de la repercusión que tiene la publicación de una noticia, esté o no basada en periodismo de datos, que introduce otra relevante aplicación del Big Data, como es el análisis de resultados.

Quien analiza datos y se encuentra ante una base de datos o una tabla con registros ha de estar atento a los siguientes signos de que algo puede ser

indicio de un problema... o una noticia:

Intervalos irregulares.

Datos incompletos.

Valores extremos.

Unidades diferentes.

Totales incorrectos.

Notas al pie, asteriscos...

DATOS Y PERIÓDICOS

Si el Big Data muestra que una de sus utilidades para los periodistas es la de ayudarlos a encontrar información, otra de ellas es la de contribuir a captar lectores y a satisfacer sus demandas. ¿Cómo se pueden adaptar los periódicos y las publicaciones al fenómeno de Big Data? Hasta ahora, muchos han optado por aplicar el mismo modelo de negocio para la impresión del periódico en papel y para sus ediciones en la web, incluso el modelo de la Red se adapta tal cual a las nuevas ediciones en tableta. ¿Funcionará este enfoque con el Big Data?

Uno de los retos del Big Data aplicado en los medios de comunicación es ofrecer a las personas información específicamente diseñada para ellos. Los cambios tanto en la recolección, análisis y visualización del gran volumen de datos tendrían que afectar no solo a la información que se genera para el lector, sino también al propio modelo de negocio de los medios de comunicación. Analizando el consumo de información y cómo este se produce, se pueden realizar modelos predictivos que mejoren los productos y la experiencia de los lectores.

Según el analista de medios Alan D. Mutter, para conseguir alcanzar esta meta los periódicos deben sumarse al rápido desarrollo de las aplicaciones en torno a Big Data, lo que favorecería tanto a sus lectores como al valor de sus anunciantes. Por ejemplo, Google ha logrado dominar la publicidad digital gracias a su tecnología en torno al análisis y la visualización de la gran cantidad de datos que maneja sobre gustos, preferencias y consumos de información en sus productos. Así, ha podido mejorar las conexiones entre los potenciales compradores y los anunciantes gracias al hábil manejo de los datos que generan los usuarios.

Está claro que una parte del uso del Big Data estará en un mayor y mejor conocimiento de lo que los lectores hacen, de ahí la importancia de profesionales y herramientas que lo analicen y sean propositivos a la hora de generar mejoras; pero, por otra parte, ¿no es el negocio y el trabajo de los propios datos el negocio y el trabajo de la información? ¿No deberían los periódicos tener una parte de su negocio enfocada a la “venta” de algunos tipos de datos o a su homologación?

Cada vez más aplicaciones basadas en Big Data permiten optimizar la publicación online con un análisis de titulares o fotografías en tiempo real y a medida que el usuario interactúa con la información. Outbrain es, por ejemplo, un conjunto de herramientas que ayudan a mejorar el criterio editorial y la toma de decisiones de los periodistas con el análisis predictivo y algorítmico, generando recomendaciones de contenido que luego el periodista puede, o no, seguir. Este tipo de software indica al periodista, por ejemplo, qué tipo de titular funciona mejor o qué foto atrae más visitas a la noticia. La decisión final es siempre del editor, pero cada día hay más sistemas que ofrecen información predictiva sobre el comportamiento de los lectores. Sin embargo, no es oro todo lo que reluce: muchos de estos sistemas necesitan y aprovechan los datos de los usuarios de esos medios a los que sirven, por lo que algunos editores prefieren no tener ingresos a corto plazo ni utilidades que les ayuden a mejorar para no compartir su bien máspreciado: los datos de sus lectores.

CIUDADES INTELIGENTES

“Ciudad inteligente” (Smart City) es un término del que estamos oyendo hablar mucho y del que todavía oiremos mucho más en los próximos años. Se estima que en 2020 estaremos gastando 400.000 millones al año en la construcción de ese tipo de ciudades, y si el proyecto “42 Greenway” es una ambiciosa planeación urbana para convertir una de las calles más emblemáticas de Nueva York en un entorno urbano sostenible, repleto de conexiones inalámbricas para los vecinos, células que gestionan los riegos adecuados de los parques, sistemas de transporte conectados a los móviles de sus vecinos o depósitos de reciclaje de residuos con innovadores sistemas de recolección y procesamiento automático, no es menos ambicioso el proyecto “París 2050” que pretende cambiar no ya una calle, sino toda la capital francesa. Una metrópoli más ecológica, con edificios que gestionan su propia energía y que limpian el aire de la urbe y que, para poder hacerlo, necesitan de toda la inteligencia digital disponible y de un nuevo “sistema nervioso”. El proyecto del arquitecto belga Callebaut ha sido tildado de utópico y quizás lo sea en la parte estructural y ecológica de los nuevos edificios que toman París (si los parisinos lo permiten), pero es bastante realista en la forma en que los datos pueden modificar las ciudades.

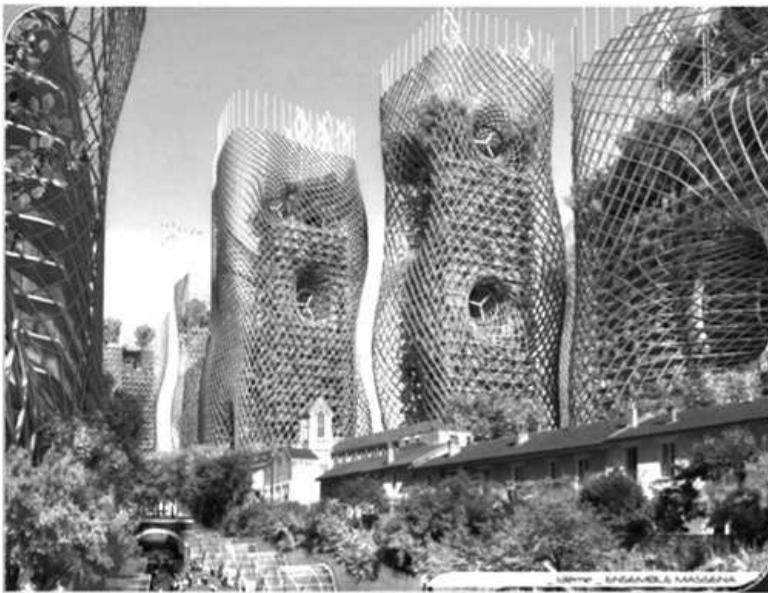


Imagen del proyecto “París 2050” del arquitecto Vicent Callebaut.

Una de las ideas básicas de las ciudades inteligentes es incorporar los avances en la tecnología y recopilación de datos que está haciendo el Internet de las cosas en las infraestructuras de los entornos en los que vivimos. Multinacionales como Cisco e IBM están trabajando con las universidades, los gobiernos y las áreas de planificación urbana en el desarrollo de nuevos sistemas basados en datos y aplicados al transporte, la gestión de residuos, la policía y el uso de energía etc., para que sean más eficientes y puedan mejorar la vida de los ciudadanos.

La ciudad de Boston, en Estados Unidos, también participa en un importante programa de Big Data con diferentes multinacionales, en el que se recoge una cantidad ingente de datos de muchas fuentes que podrían ser muy útiles para investigadores, desarrolladores, ingenieros de transporte, planificadores urbanos y, sobre todo, para los ciudadanos. Estos datos, sin embargo, a menudo están aislados y dispersos en varios departamentos y en diferentes formatos. Para lograr sus objetivos climáticos y de transporte, la ciudad de Boston necesita información oportuna, local y precisa sobre las condiciones del transporte en tiempo real. La alcaldía pretende, mediante el análisis de todos estos datos, la reducción de las emisiones de carbono asociadas a los desplazamientos de automóviles, que representan alrededor del 25% de las emisiones de carbono de la ciudad. También se pretenden analizar los patrones de viaje de los vehículos para reducir millas, lo que contribuiría a disminuir la congestión del tráfico y el dióxido de carbono (CO₂). Una buena gestión de esos análisis debería proporcionar datos a los residentes para tomar decisiones inteligentes acerca de las alternativas de transporte y eliminar atascos. Si estos sistemas están coordinados con los mecanismos como los de París para aligerar la contaminación mediante “edificios pulmón”

será todo más sencillo.

En Glasgow, Escocia, el Gobierno ha invertido 37 millones de euros en tecnología, que hará que la ciudad sea “más inteligente, más segura y más sostenible”.

Vamos a interactuar y obtener información de estos sistemas inteligentes usando en las ciudades nuestros teléfonos, relojes y otros dispositivos, y, fundamentalmente, las máquinas también van a hablar entre sí. Los camiones sabrán la ubicación de la basura que hay que recoger y los sensores en nuestros coches nos indicarán las plazas de aparcamiento disponibles.

Las aplicaciones previstas o desarrolladas en este tipo de ciudades incluyen la iluminación inteligente de las calles para ahorrar energía, y se apagarán cuando no haya nadie; el mapeo del uso de energía alrededor de la ciudad para entender mejor la demanda, y el mapeo de cómo la gente se desplaza para maximizar el uso de la bicicleta y los recorridos a pie.

Los sensores conectados a las luces de la calle y otro mobiliario urbano exterior medirán pisadas, los niveles de ruido y la contaminación del aire y estos datos serán utilizados para dar prioridad al funcionamiento de otros servicios. El Consejo de Estrategia Tecnológica del Gobierno, que coordina el proyecto, dice que más de 200 fuentes potenciales de los datos han sido identificadas y, aunque muchas de ellas ya están habilitadas para recoger los datos, la información se usa de manera aislada. Algo que cambiará. La extensa red de circuito cerrado de televisión en Glasgow será, por ejemplo, utilizada tanto para la vigilancia del tráfico y de la delincuencia como para el alumbrado público. Analizar datos de múltiples fuentes es siempre más valioso que hacerlo de las fuentes aisladas.

En Bristol, otro programa implica el desarrollo de una red inalámbrica dedicada específicamente al Internet de las cosas y comunicaciones entre dispositivos. Es un nuevo modelo de transmisiones que utiliza menos energía que el wifi y las redes móviles existentes, por lo que es más respetuoso con el medio ambiente y parece una mejor opción para los dispositivos que deben mantenerse en funcionamiento 24 horas los 7 días de la semana.

También los servicios públicos en las ciudades empiezan a utilizar de forma continua el Big Data: el Hospital Infantil de Seattle ha conseguido reducir los tiempos de espera de los pacientes usando una aplicación de la herramienta de visualización Tableau Software. Así, tras medir y evaluar los tiempos de espera, las causas y los factores que contribuían a su incremento, desde el centro hospitalario aseguran que han reducido estos tiempos y han podido atender a un mayor número de pacientes. “Estamos viendo que los analistas de datos, los administradores de empresas y los analistas financieros, así como los clínicos, médicos e investigadores, utilizan los resultados de la tecnología que ofrece Tableau para resolver diferentes problemas que antes no podíamos detectar por nuestra cuenta”, explica Ted Corbett, director de Gestión del Conocimiento de este centro hospitalario.

Ciudades inteligentes en España

Pero también en España hay ciudades que están experimentando con estos nuevos modelos. Una de ellas, Santander, preside la Red Española de Ciudades Inteligentes (RECI), que agrupa a más de 60 municipios que pretenden intercambiar sus experiencias en el ámbito de la innovación, impulsar el uso de las nuevas tecnologías para mejorar y hacer más eficientes los servicios públicos y promover la colaboración público-privada en este sector. Estos son algunos de los proyectos que está llevando a cabo el ayuntamiento cántabro:

Cloud City Center: el cerebro de la plataforma de análisis, desarrollo y actuación. No solo controlará todos los servicios públicos de la ciudad, sino que relacionará unos con otros para administrar y gestionar esa información de forma coordinada.

Plataforma: la columna vertebral y sistema nervioso del modelo. Esta plataforma estará constituida por los sensores y sistemas de comunicación. Hay dos grandes proyectos europeos en desarrollo:

- **SmartSantander:** son sistemas para la mejora del aparcamiento en superficie, sistemas de riego inteligente, realidad aumentada e información medioambiental.

- **Outsmart:** también se incorporará a la plataforma la información que se recabe del alumbrado público.

Por lo demás, en la ciudad se está entretejiendo una malla de sensores de tres tipos:

Sensores estáticos: se trata de los sensores que se encuentran instalados en un punto fijo de la ciudad, desde el que recogen los datos correspondientes que luego son procesados para las diferentes aplicaciones prácticas que va a tener esta información, entre ellos, sensores de detección de plazas de aparcamiento libres, sensores de humedad para el riego inteligente en zonas verdes o sensores medioambientales.

Sensores dinámicos: hay varios tipos de dispositivos de detección, según distintos parámetros, que están instalados en aparatos móviles.

Sensores ciudadanos: son los propios ciudadanos quienes, a través de su teléfono inteligente, recogen datos o incidencias que se están produciendo, en ese mismo momento, en el punto en el que se encuentren.

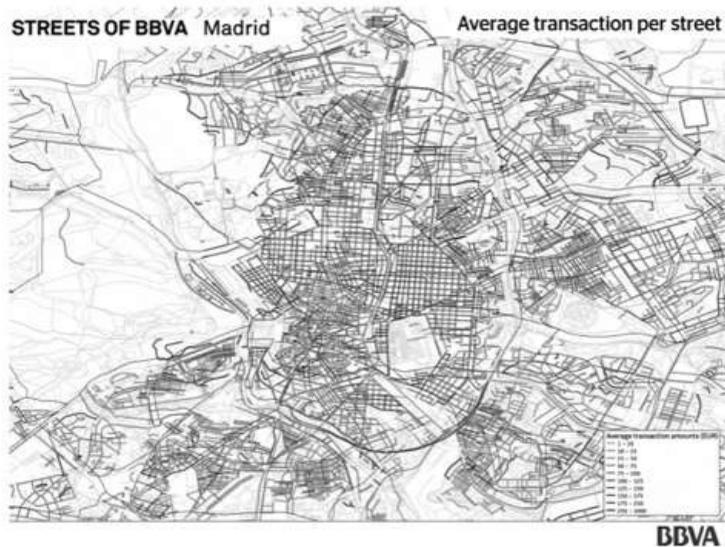
Uno de los proyectos que ha sido premiado por la RECI ha sido la app barcelonesa Shotl, que permite a los usuarios ponerse en contacto con otras personas que comparten su mismo destino para solicitar un medio de transporte. En pocos minutos, un minibús les recoge en la ruta establecida, lo que reduce el tráfico de la ciudad y las emisiones, además de ser más barato que el transporte privado. También ha destacado la plataforma Hackity app, que permite que los ciudadanos se pongan en contacto y propongan acciones para mejorar la ciudad.

Uno de los proyectos más ambiciosos de I+D+i que se ha puesto en marcha para impulsar la Smart City es Ciudad 2020, dotado con una financiación de alrededor de 16 millones de euros gracias al programa Innpronta del Centro para el Desarrollo Tecnológico Industrial (CDTI). El ciudadano recibe la información en tiempo real para llegar lo más rápido posible a su trabajo o ve cómo se regula la iluminación a su paso. Durante cuatro años se han desarrollado 33 activos tecnológicos experimentales en ciudades como Málaga, Zaragoza o la ya citada Santander. Coruña Smart City es otro de los proyectos que ambicionan una gestión integral y transversal gracias a la tecnología.

Las ciudades inteligentes son un verdadero desafío: en 2050, el 70% de la población mundial vivirá en pueblos y ciudades.

Ciudades inteligentes y cartografía

Y no solo los ayuntamientos tienen interés en los proyectos de ciudades inteligentes: Streets of BBVA fue un proyecto pionero de Smart Cities en el que se reflejan los primeros resultados de un trabajo en curso sobre la posible utilización de los datos transaccionales que genera BBVA, en combinación con otras fuentes de información. Estos mapas han sido realizados con datos anónimos, agregados y no personales, que tratan de describir la ciudad y su actividad.



Street of BBVA muestra una interesante combinación de datos para poder conocer las "zonas calientes" de las ciudades y así poder detectar los tipos de servicios que pueden ser susceptibles de ser utilizados por los ciudadanos, ya que contempla un estudio espacio-tiempo pormenorizado de las transacciones comerciales de Madrid y Barcelona en determinados momentos del año.

Elena Alfaro, CEO de BBVA Data & Analytics y experta en Smart Cities, en una entrevista en el Centro de Innovación BBVA cita otros casos conocidos, como el tráfico en Londres o Estocolmo. Son ciudades que “han reducido considerablemente el tráfico en el centro y mejorado la calidad del aire” gracias a las modificaciones surgidas de sus estudios sobre los movimientos del tráfico de vehículos.

El BBVA ha seguido con esa línea de trabajo². “En Barcelona en 2012, se hizo la medición del impacto económico del Mobile World Congress. Para ello, se extrajeron datos de las transacciones realizadas con tarjetas de crédito, tanto de la semana antes como de la semana en que tuvo lugar el evento.” Los resultados sirvieron para concluir los lugares, días y horarios donde más “movimiento” hubo, lo que puede servir a los comercios para reforzar sus acciones de marketing y venta de cara a eventos similares o a las propias ciudades para hacer lo propio con sus promociones turísticas.

Otro ejemplo de este análisis de flujos en ciudades realizado por BBVA fue el llevado a cabo junto con el Ayuntamiento de Madrid³ durante 2012, que analizaba su actividad comercial a partir de los datos de consumo de los turistas. Entre otros resultados, el estudio sirvió para cuantificar el impacto económico de la fiesta del Orgullo Gay en diversas zonas de la ciudad, mostrando que el gasto comercial aumentó un 24% respecto a la misma semana del mes anterior. Además, aportaba otros datos interesantes, como los que identificaban los turistas que más gastaban, en qué lo hacían, por dónde se movían, etc.

Los mapas son una de las herramientas de visualización, como es lógico, más utilizada para ver tendencias y patrones en ciudades. La firma estadounidense Stamen Design ha desarrollado el análisis de tendencias económicas, sociales y culturales en una ciudad a través de diferentes proyectos de visualización y desarrollo, por ejemplo, el mapa del crimen de San Francisco:

Crimespotting es una herramienta de procesamiento, análisis y visualización de los datos, frente a las aplicaciones de cartografía del crimen tradicionales que tienen una funcionalidad limitada y sistemas de representación anticuados. Con esta aplicación de Stamen se pueden seguir los casos de delincuencia en un barrio concreto y proporcionar servicios personalizados.

Cabsplotting es otra de las aplicaciones, pionera en su momento, diseñadas por Stamen que analiza los datos que proporcionan las señales de GPS de los taxis de la ciudad para revelar patrones ocultos de la vida urbana. También se ha aplicado en San Francisco, donde la información se recoge desde un servidor de la sede central de Yellow Cab, que recaba los datos GPS de cada taxi en tiempo real y los analiza. Hoy hay ya muchas ciudades que ofrecen datos similares, incluso compañías particulares como Uber o muchas de las que gestionan flotas de sus servicios públicos de forma similar, y ofrecen al público consultar la disponibilidad en tiempo real en sus teléfonos móviles.

Pero en análisis de datos con cartografía hay que citar sin duda a CartoDB, una empresa española que poco a poco ha ido situándose en la vanguardia

mundial de representaciones visuales a través de mapas. CartoDB ha conseguido una financiación de unos 30 millones de euros⁴ para arrancar el análisis y visualización de datos geolocalizados. Su fundador, Sergio Álvarez, sabe que los datos por sí solos no sirven de nada. Pero con un software que te permita ordenarlos, analizarlos y visualizarlos de forma sencilla sobre mapas se convierten en algo fundamental para las empresas, las administraciones y los ciudadanos.

CartoDB ha puesto en marcha numerosos mapas para ayudar a manejar de forma sencilla el mar de datos al que nos enfrentamos. Como en su momento Google y ahora CartoDB, parece que quien sabe ordenar el caos es candidato al éxito.

REVOLUCIÓN EN LA SALUD⁵

“Big Data ayuda a encontrar las preguntas que no sabes que quieres preguntar.” La frase del consultor tecnológico Eric D. Brown resume la ambición del sector sanitario con la recogida masiva de datos. Predecir lo que va a pasar es lo que buscan la mayoría de los médicos al aplicar la tecnología Big Data. Una tecnología que ha aterrizado en el sector sanitario y que, como señala el médico español Bernardo Valdivieso, director del Área de Planificación y del Área de Atención Domiciliaria y Telemedicina del Hospital Universitario y Politécnico La Fe de Valencia, va a suponer una verdadera revolución. “Va a ser el gran cambio. Somos el único sector productivo de servicios que nunca ha utilizado la información de manera adecuada para adaptar los servicios a las necesidades, para predecir expectativas o necesidades de usuarios y para evaluar las actuaciones.”

Del Big Data sanitario interesa el denominado Real World Data (RWD), que como señala el doctor Salvador Peiró en el Estudio Big Data Salud, “a diferencia de los ensayos clínicos y otros estudios realizados bajo condiciones controladas, refleja la atención real que reciben los pacientes reales en cada contexto concreto, y los resultados clínicos que realmente obtienen, que no son necesariamente iguales a los que se obtuvieron en los ensayos clínicos. El RWD recoge los beneficios y efectos adversos de las decisiones médicas en la práctica clínica habitual de millones de pacientes”.

Entre los RWD, se incluyen estudios observacionales de registro, datos de historias clínicas electrónicas, encuestas de salud y ensayos clínicos pragmáticos. ¿Y para qué sirven concretamente estos datos?:

Los RWD, por un lado, permiten identificar anticipadamente los pacientes crónicos en riesgo de descompensación (estratificación poblacional) para incluirlos en programas de específicos de atención (case management, disease management). También puede ayudar a la toma de decisiones clínicas en tiempo real (analizando casos similares y proponiendo alternativas de manejo) y reducir la variabilidad en la práctica médica. Adicionalmente, es posible trasladar información directamente a los pacientes, permitiéndoles un rol más activo en sus propios cuidados y, quizás, una mayor efectividad para modificar estilos de vida, controlar factores de riesgo y mejorar la adherencia a los tratamientos.

Peiró pone de ejemplo lo que se está realizando en Estados Unidos en investigación clínica, farmacológica y epidemiológica. El análisis de los RWD está suponiendo ya enormes beneficios para la población. Algunas aplicaciones, como el programa Mini-Sentinel (<http://www.mini-sentinel.org>) de la Agencia de Medicamentos estadounidense, han permitido detectar nuevas interacciones, efectos adversos de medicamentos y otros problemas de seguridad que han llevado a la retirada de fármacos o la modificación de sus indicaciones. Los RWD también permiten comparar diferentes tratamientos para una misma condición (comparative effectiveness) y establecer el papel de cada uno en el manejo de los diferentes pacientes, un aspecto con importantes implicaciones en efectividad, seguridad y costes. En gestión de la atención sanitaria, los RWD permiten el desarrollo de indicadores sofisticados para comparar la calidad de la atención que reciben los pacientes atendidos en diferentes centros o por diferentes médicos y desarrollar estrategias de mejora.

Como recalca Valdivieso:

Uno de los cambios que pretende el nuevo modelo es hacer una gestión proactiva del cuidado: no tendremos que esperar a que vengan los enfermos cuando se encuentren mal, sino nosotros, de alguna manera, gestionar la población para mantenerla sana y para tener estables las enfermedades y que no progresen. Usar la información que estamos acumulando para predecir qué va a pasar tiene que ayudarnos a tener esa proactividad.

Tanto en España como en Estados Unidos, el Big Data acaba de empezar. Un campo que comienza y que ofrece “más una oferta tecnológica que un proyecto real aplicado a la práctica clínica”, según el médico del hospital valenciano que actualmente está almacenando datos de 300.000 pacientes.

Otro proyecto ambicioso, fuera de España, es el IBM Watson Health. Ese el nombre de un superordenador capaz de procesar Big Data mediante un sistema cognitivo. Desde Boston y Nueva York, 2.000 profesionales entre consultores, médicos, personal sanitario, desarrolladores e investigadores trabajan en el desarrollo de las capacidades de IBM Watson Health. Los errores médicos son la tercera causa de muerte en Estados Unidos y Watson es capaz, entre otras cosas, de leer millones de artículos médicos. El gigante estadounidense se ha aliado con Apple, Johnson & Johnson y Medtronic, y con dos start-ups: Exploris, una spin-off de la Clínica Cleveland cuyos datos en 50 millones de pacientes se utilizan para detectar patrones de enfermedades, tratamientos y resultados sanitarios, y Phytel, de Dallas, fabricante de un software que gestiona la atención al paciente y reduce el ratio de readmisiones en hospitales para dar este paso de gigante en el mundo de los datos. Simplificando mucho el plan de IBM, se trata de que la tecnología Watson sirva como un servicio en cloud que permita a los hospitales, médicos, aseguradoras e incluso potenciales pacientes aprovechar su vasto almacenamiento de datos sanitarios. Aunque, como detallan en su página web, el proyecto no para de crecer: IBM tiene acuerdos con Boston Children's Hospital, Columbia University, ICON plc, Sage Bionetworks y Teva Pharmaceuticals, organizaciones que están utilizando la tecnología de IBM Watson para transformar aspectos clave del sector sanitario como son el

descubrimiento y desarrollo de nuevos fármacos, la medicina personalizada, la gestión de las enfermedades crónicas, la pediatría o la salud digital. Con Boston Children's Hospital, IBM integrará la capacidad de comprensión del lenguaje natural y de responder preguntas en la iniciativa OPENPediatrics de este hospital, a través de la cual comparten conocimiento médico con pediatras de todo el mundo. Junto a Columbia University Medical Center's, el departamento de Patología, Biología Celular y Sistemas de Biología va a colaborar para hacer pruebas con IBM Watson que ayuden a los oncólogos del Columbia Herbert Irving Comprehensive Cancer Center a diseñar tratamientos personalizados a partir del análisis de ADN. Por último, ICON es la primera organización de investigación clínica en utilizar Watson para analizar la compatibilidad de los pacientes con los ensayos clínicos. Se estima que el 80% de los ensayos clínicos fallan o se retrasan debido a que los pacientes que forman parte de ellos no son los más adecuados. En definitiva, IBM Watson Health es un sistema cognitivo que analiza los datos y que tiene la capacidad de encontrar conexiones en todo el conocimiento digital disponible. Watson aprende de los seres humanos a través de un lenguaje natural y por medio de algoritmos se lo pregunta todo.

Como indica el economista y gestor sanitario Miguel Ángel Mañez, las posibilidades en el ámbito de la salud del Big Data con muchas, entre las cuales destaca:

Predicción de hospitalizaciones por patologías basándose en factores ambientales, poblacionales, etc.

Identificación de pacientes de alto riesgo. De hecho, en muchos servicios de salud se han puesto en marcha programas de atención a la cronicidad que, basándose en los datos que constan en la historia clínica, han podido catalogar a cada paciente en un nivel de la conocida pirámide de Kaiser.

Toma de decisiones en la consulta.

Análisis del estado de salud de una población o territorio.

Seguimiento de tendencias.

Ensayos clínicos. En muchos casos, ya no será necesario reclutar a pacientes o revisar casi manualmente las historias clínicas, ya que los sistemas de Big Data permitirán encontrar a los pacientes para el ensayo y hacer el seguimiento de forma casi automática.

Efectividad de medicamentos y seguimiento de efectos adversos. ¿Cuántos pacientes han sufrido el mismo efecto adverso ante un medicamento? ¿Existe una mejora en los pacientes que utilizan el medicamento A frente a los que toman el medicamento B?

Evaluación de servicios sanitarios. Será mucho más sencillo obtener resultados de salud y comparar y evaluar iniciativas, proyectos, unidades, etc.

Vigilancia epidemiológica.

Una inmensidad de datos que, para el presidente de la Sociedad Española de Informática de la Salud, Luciano Sáez Ayerra, todavía hay que organizar. Un camino largo para el que se necesitan más recursos:

Llevamos planteando muchos años la necesidad de que se utilice la cuantiosa información clínica y de salud que se encuentra disponible en las organizaciones sanitarias con el fin de generar conocimiento y asegurar una toma de decisiones basada en la información. Esto es un reto a nivel de organización sanitaria, ya que esta función de explotar la información no está imbuida en nuestra cultura de una forma profesionalizada, no disponemos prácticamente en el sistema de expertos en análisis de datos. Lógicamente, existen algunas experiencias e iniciativas para anonimizar datos de salud y ofrecer esta información para la investigación. Por otra parte, los sistemas tecnológicos actuales no están preparados en general para abordar proyectos de Big Data, será necesario hacer importantes cambios en arquitecturas y comunicaciones. Pero antes de plantearse estos cambios, habrá que definir qué datos se van a utilizar, dónde se van a almacenar, quién lo va a coordinar y para qué se van a utilizar. Preguntas básicas para las que todavía no hay respuesta.

Luciano Sáez Ayerra pone mucho énfasis en la necesidad de contar con profesionales sanitarios expertos en TIC y también en que hay que hacer cambios en el modelo sanitario:

La existencia de un Big Data sanitario efectivamente puede cambiar nuestro escenario a medio plazo de una forma importante. Es el momento de definir una clara estrategia para el sistema sanitario en su conjunto, ya que el valor que aportaría el poder disponer de grandes volúmenes de información real y contrastada sería enorme en cuanto a generación de conocimiento para mejorar la seguridad de la atención sanitaria y su eficiencia. Desde mi punto de vista, estamos traspasando el momento adecuado para definir la estrategia y el modelo a aplicar en todo el sistema para que sea una realidad y que cumpla con su función de una forma ordenada, legal y técnicamente viable.

Industria farmacéutica

El sector farmacéutico también está poniendo los ojos en el Big Data. El responsable del área de Government Affairs en Roche Pharmaceuticals, Federico Plaza, califica de “brutal” la utilidad del Big Data en el plano epidemiológico:

El Big Data permite tener información epidemiológica de la vida real, conocer mejor el perfil de los pacientes y logra romper la falta de información que hay hoy en día entre los productos que salen al mercado, que vienen con toda la información de los estudios clínicos, y los productos que llevan ya tiempo en el mercado, en los que no hay ninguna información adicional y estructurada de cómo se utilizan.

También aborda el plano económico:

Va a dar lugar en un plazo que no sea muy largo a pagar los medicamentos en

función del valor que aportan, en función de los resultados en salud más que en función de los miligramos de sustancia, de principio activo. El Big Data es importantísimo para hacer sostenibles nuevos productos que tienen precios elevados, pues su desarrollo es costoso y van orientados a pocos pacientes. Si conseguimos conocer cuántos pacientes utilizan el producto en cada indicación y los resultados que tiene en salud, podemos conseguir que los sistemas sanitarios apliquen los recursos para financiar estos productos de forma más exigente.

Jesús Aguilar Santamaría, presidente del Consejo General de Colegios Farmacéuticos, define el Big Data Sanitario como

una minería de datos puesta al servicio de los pacientes y del sistema sanitario —con las máximas garantías de protección y confidencialidad de los datos— que supondría una auténtica revolución. Se avanzaría en el ámbito de la farmacoepidemiología, en la respuesta de los tratamientos farmacológicos según perfiles genéticos o estilos de vida; en la prevención y predicción de riesgos y reacciones adversas de los medicamentos; en la mejora de la adherencia a los tratamientos y, en consecuencia, habría un incremento exponencial de los resultados en salud.

Big Data Sanitario

Transformar los datos en conocimiento es el fin del Big Data Sanitario. Sin olvidar que detrás de esos datos hay pacientes y que las leyes tienen que garantizar la confidencialidad y la privacidad de las personas. Las técnicas de anonimización deben de asegurar que el tratamiento de los datos de salud no permita identificar a la persona a la que se refieren, salvo que exista consentimiento expreso. Luciano Sáez Ayerra se muestra tajante:

Lo más importante es tomar conciencia de que el tratamiento masivo de datos de salud, además de entrañar grandes beneficios para la investigación y la asistencia y gestión sanitaria, supone una serie de riesgos para la privacidad y la intimidad de los pacientes. Para lidiar con estos riesgos y poder identificarlos tempranamente y adoptar las medidas necesarias para eliminarlos o, al menos, mitigarlos hasta un nivel que resulte aceptable, es necesario e imprescindible que antes de abordar estos proyectos se lleve a cabo una Evaluación de Impacto en la Protección de Datos Personales (EIPD) —una de las herramientas más útiles del paradigma de la Protección de Datos desde el Diseño— que busca que los requerimientos y obligaciones en materia de protección de datos se tengan en cuenta desde las fases iniciales de cualquier proyecto que trate datos de carácter personal.

Josep María Argimon, el director de la Agencia de Calidad y Evaluación Sanitarias de Cataluña (AQuAS), uno de los impulsores del proyecto VISC+, que se basa en la cesión de la gestión de los expedientes de la sanidad pública a una empresa externa para que esta los venda a las sociedades (públicas o privadas) que los soliciten para fines científicos, tuvo que afrontar muchas suspicacias por la posibilidad de que esos pacientes pudiesen volver a ser identificados. Un proyecto que finalmente salió adelante en 2015:

Hay que tener en cuenta que no todo son ventajas en la reutilización de los

datos sanitarios; también conlleva algunos riesgos que hay que mitigar. Los principales riesgos son los que pueden afectar a la privacidad de las personas y al uso pernicioso que se podría hacer de estos datos. Por un lado, para evitar que se pueda afectar a la privacidad, existen técnicas como la desidentificación o la anonimización de los datos, que van mucho más allá de quitar el nombre y apellidos de las personas. Es equívoco hablar de una anonimización ilusoria: el riesgo de reidentificación es una cuestión de probabilidades y aunque el riesgo nunca es cero, sí se puede hacer extremadamente bajo. Por otra parte, igual que ocurre en los estudios clínicos, cualquier proyecto de investigación que implique la reutilización de datos sanitarios debe contar con el dictamen favorable de un comité ético y de investigación clínica. Son estos comités los que mejor pueden ponderar los beneficios sociales de la investigación y los riesgos individuales de la misma. Otro riesgo que podría afectar a la utilización de los datos sería el peligro de caer en un excesivo datocentrismo; es decir, la creencia de que en los datos se encuentran todas las soluciones y que se puede prescindir del uso de mecanismos más imperfectos para dar respuesta a los estudios, como los valores o la experiencia. Estos aspectos son especialmente importantes en una sociedad que a menudo piensa que todas las soluciones son computables y se encuentran en un servidor.

Aunque el Big Data sanitario está dando sus primeros pasos, dos cosas parecen claras: el potencial que tiene todavía es inimaginable y la preocupación por la privacidad y la confidencialidad va a protagonizar cualquier recogida masiva de datos.

LA PUBLICIDAD QUE TE SIGUE Y TE CONOCE

¿Nunca te has preguntado por qué aparece un anuncio que te invita a irte de vacaciones a Cádiz (por cierto, lo miraste ayer en una web de viajes) cuando navegas por una página del New York Times? Pues porque con tu rastro de navegación hay todo un mercado de datos donde se vende y compra la mercancía más valiosa: tú. Gracias al almacenamiento y procesado de lo que haces las empresas de publicidad saben más que nunca qué ofrecerte, y además en tiempo real, como nos gusta hoy: las cosas para ahora. No es raro que en tu móvil aparezcan menús cuando rondes uno de tus restaurantes favoritos —desconecta el GPS del teléfono si no quieres que cada vez que pases por esa calle te aparezca el plato del día— ni que te bombardeen con anuncios de clínicas de dientes cuando te has quejado de un dolor de muelas. Aunque creas que solo pensaste en ese dolor insoportable en tu cabeza, seguro que buscaste remedio en el buscador de Google o pediste ayuda en algún mensaje.

Somos datos y las empresas lo saben. Los datos se han convertido en la parte esencial de las estrategias comerciales y las compañías buscan transformarlos en negocios. Actualmente, hay 10.000 millones de aparatos conectados a Internet y en 2020 habrá unos 50.000 millones. Los análisis de datos nos permiten identificar patrones de comportamiento de los clientes. Si un cliente pinta una casa, hay muchas probabilidades de que quiera alquilar, las aseguradoras le podrán ofrecer seguros especiales u ofertas que le den más tranquilidad. Los datos ayudan a predecir su comportamiento; son vitales para que el departamento de marketing se anticipe a las necesidades de miles

de personas. Y por eso te llegan descuentos, durante días, de clínicas dentales. Aunque se te haya pasado, afortunadamente, el dichoso dolor de muelas.

La publicidad en los medios digitales y, en algunas ocasiones en los tradicionales como radio y televisión, ha cambiado considerablemente en pocos años debido a la irrupción del Big Data. Hoy la gestión de los anuncios publicitarios la hacen en gran parte programas y algoritmos basados en datos de diversa índole. Los sistemas de compra-venta programática de publicidad digital son métodos automatizados de comercialización de anuncios. La publicidad en medios digitales es actualmente un gran mercado de datos que funciona de forma similar a la bolsa, con compradores y vendedores. En la actualidad, mueven en torno al 15% de la actividad publicitaria en España, y la tendencia es que aumente y se consolide. Su gran ventaja, aparte de su rapidez, es que se basan fundamentalmente en datos de audiencia de los medios. Son datos extensos, fruto de la explosión del Big Data y las cookies (archivos que permiten conocer el comportamiento de los usuarios al navegar). Los datos que se manejan están relacionados con el comportamiento e historial digital del usuario: desde su navegador, sistema operativo, tipo de ordenador y tipo de móvil, aplicaciones descargadas y uso que hace de las redes sociales hasta algunas veces edad, género y geolocalización, por citar algunos indicadores.

Estos sistemas son, en cierto modo, la evolución natural de lo que comenzó hace una década con el behavioral targeting y las campañas dirigidas a segmentos muy definidos de usuarios. El behavioral targeting ha evolucionado y ha sido superado por su hermano mayor, el predictive behavioral targeting (PBT). Además del historial de navegación del usuario, tiene en cuenta otros parámetros y datos obtenidos de terceras partes, como actitudes de compra, localización, hora o perfiles sociodemográficos para predecir su comportamiento futuro. De este modo, el anunciante no pierde la ocasión de dirigirse directamente a un grupo de usuarios que quizás en el pasado no sabía si encajaban en su perfil, pero que por el cruce de datos se ajustan claramente a su audiencia potencial. Estos perfiles antes estaban solo en manos de los medios y portales, pero la diferencia sustancial es que ahora son agencias y marcas las que pueden acceder directamente al inventario de espacios disponibles de los medios de comunicación y configurar a su medida campañas dirigidas a los segmentos de audiencia que buscan, sin pasar por intermediarios.

Por otra parte, los sistemas programáticos permiten a los medios de comunicación online rentabilizar de una manera eficaz todo su inventario (no solo el remanente), monetizar su audiencia y satisfacer mejor las demandas del sector.

La publicidad programática es, por tanto, aquella que se compra y se vende de forma automatizada gracias al nuevo mercado que ofrecen los datos.

El término ha generado mucha confusión entre anunciantes, medios y agencias. Si hace unos años se compraban paquetes de publicidad por un volumen de impresiones a un precio pactado entre vendedor y comprador y a posteriori se analizaban los resultados, ahora la compra programática permite

decidir cuánto pagas por cada una de las impresiones de un usuario afín al target o segmento definido de la campaña: mucho más eficaz y con un retorno de la inversión medible.

La diferencia entre una compra programática de publicidad y la tradicional es que la relación entre compradores y vendedores ha implicado la aparición de nuevos actores. Antes, la publicidad tenía un esquema muy sencillo de vendedor de espacios, agencia intermediaria y cliente que necesitaba dar a conocer sus productos; hoy hay proveedores de datos de audiencia, consolidadores de esos datos, nuevos inventarios (tipos y número de espacios publicitarios a la venta) de los media de comunicación digital, mesas de compra automáticas y los data service provider (DSP). Es decir, el número de pasos y equipos que actúan en el proceso ha aumentado, pero también la manera de manejarlos se ha vuelto más efectiva y veloz. Y aunque consiguen mayores montos de ingresos para el editor, los diferentes agentes de esa cadena se quedan con 60 céntimos de cada euro que invierte un anunciante, llegando solo 40 para el editor dueño del espacio publicitario.

Entender qué papel juega cada uno de estos equipos es primordial para tener éxito en el manejo de esta publicidad; además, la programática es una tendencia imparable. Los sistemas de compra-venta programática mueven actualmente en torno al 15% de la actividad publicitaria en España, y la tendencia es que aumenten y se consoliden.

Su gran ventaja, a parte de su rapidez, es que se basa en datos de audiencia de los medios.

Los anunciantes y los medios de comunicación están integrando rápidamente las herramientas de venta y de compra de publicidad en modo programático.

En Estados Unidos, casi cuatro de cada cinco agencias y marcas están comprando su publicidad en programática, la cual superó el 50% del total de la publicidad digital en 2015.

Este tipo de compra incluye publicidad para los teléfonos móviles, que es donde se está produciendo una gran parte del incremento de la inversión publicitaria en programática. En teléfonos móviles supone un 46% y en vídeo el 36%. Serán las dos áreas que crezcan más rápido.

Pero es que, además, los ingresos por publicidad en Internet van camino de situarse muy cerca de los ingresos por publicidad en televisión.

Según el Entertainment and Media Outlook 2014-2018 de la consultora PwC, la diferencia de facturación publicitaria entre la televisión e Internet a nivel mundial se situará en torno a los 15.000 millones de euros en 2018, lo que significa que, de seguir la tendencia actual, en 2020 la publicidad en Internet podría superar a la televisiva.

Por lo que respecta a España, PwC prevé que los ingresos por publicidad en Internet crecerán a un ritmo anual del 4,7% en los próximos años, superando los 1.140 millones de euros en 2018.

El hecho de que el primer medio en el reparto de la tarta publicitaria provenga de la publicidad programática basada en macrodatos —en los países del primer mundo— es un indicador destacable de la importancia económica que adquiere el Big Data en muchos de los campos que analizamos.

Big Data y la predicción del comportamiento

Coca-Cola es una empresa pionera en la forma de analizar los millones de datos de sus consumidores. La multinacional tiene modelos que le permiten diagnosticar las causas y motivos de las ventas y, por tanto, planificar y poder hacer mejores predicciones sobre cuántos litros van a vender y dónde. Si Coca-Cola consiguiera incrementar un 0,1% sus ventas globales mediante estas técnicas, entonces ganaría 18 millones de dólares más. No parece mala cifra, ¿verdad?

Una de las compañías españolas que ya está utilizando tecnologías para el análisis de Big Data es el servicio de atención al cliente 1004 de Telefónica. La compañía está aplicando modelos predictivos con los que puede determinar el número de llamadas que recibirá para poder organizar el servicio, según explican los responsables del call center de Telefónica, quienes han detallado que el servicio del 1004 está realizando un plan de operaciones mensual y otro trimestral. Con el análisis de todos estos datos ha conseguido reducir los errores a menos de la mitad.

Estos análisis son claves porque en el call center existe una alta fluctuación en el volumen de demanda de clientes, y estar preparado para esas fluctuaciones es percibido positivamente por los usuarios. Esos ciclos, en parte, se pueden prever, ya que existe cierta estacionalidad. Por poner un ejemplo sencillo, conocer el dato histórico ya implica saber que si se tienen 2,4 millones de llamadas en agosto, se pasarán a recibir unas 3,6 en septiembre. Por supuesto, los cálculos del Big Data que maneja la operadora son mucho más sofisticados y permiten cubrir las necesidades con antelación suficiente de manera dinámica. A medida que se pueden procesar más datos y se mejoran los algoritmos de manejo, las capacidades aumentan y la satisfacción de los usuarios que llaman y no han de esperar una eternidad también.

Aunque también existen algunas nubes oscuras en el mundo de las predicciones del Big Data. En 2009 Google era protagonista en la revista Nature, maravillada ante la herramienta Google Flu Trends que permite predecir, de forma casi exacta, los casos de gripe en cada región gracias a las consultas del buscador. Google interpretaba millones de búsquedas relacionadas con la gripe —“síntomas gripe”, “dolores gripe”, “fiebre gripe”, por ejemplo— para señalar que habría colas en los centros hospitalarios en determinados puntos del país. Una información que permitía predecir el número de personas que acudiría a los centros sanitarios y ponía en alerta a médicos y sanitarios ante posibles epidemias. Cinco años después su herramienta fallaba —en 2013 predijo el doble de visitas al médico que las registradas— y el fiasco pasó a ser conocido como “la arrogancia del Big Data”. Google volvió a ser protagonista de otra revista prestigiosa científica, Science, pero con un titular muy crítico: “The parable of Google Flu: Traps in Big Data Analysis” (“La parábola de Google Flu: trampas en el análisis de

grandes datos”). El artículo tacha la herramienta de Google de “arrogante” y critica “su implícita suposición de que los grandes datos son sustituto de la recopilación de datos y del análisis tradicional, en lugar de un complemento para los mismos”. El gigante norteamericano no supo ver que muchos de los individuos que buscaban en el buscador no tenían gripe —tenían un simple resfriado o realizaban la búsqueda por la psicosis mediática— ni tampoco que los distintos (cientos, miles) de cambios que el propio Google introduce en el algoritmo de su buscador podían alterar los resultados.

El fundador de la red social de contactos profesionales LinkedIn, Reid Hoffman, ha declarado que la Web 3.0 procederá de los datos y de la analítica. De hecho, la compañía cuenta con un proyecto de Big Data en LinkedIn, como es InMaps, una visualización de datos que cartografía las conexiones en todo el universo profesional de un usuario. De este modo, LinkedIn aprovecha los datos generados por los mismos.

Las redes sociales

Las redes sociales nos permiten acceder a lo que las personas quieren publicar —aunque algunas veces no es tan claro que así lo quieran— y pretenden comunicar. Hay herramientas que en ese entorno virtual analizan la opinión expresada por las personas, incluso hay un software que pretende predecir nuestro comportamiento por lo que estamos diciendo en Facebook. ¿Nos vamos a separar? ¿Estamos a punto de cambiar de trabajo? Muchos de los signos de estos cambios se esconden en nuestros mensajes, aunque nosotros mismos no los sepamos ver, pero los robots que analizan los datos empiezan a hacerlo.

En enero de 2015, el Banco de Inglaterra anunciaba la creación de un equipo especial de analistas para estudiar el comportamiento de las personas en las redes sociales. En Estados Unidos, fondos de alto riesgo como Tashtego también analizan los perfiles de Twitter o Facebook de los inversores. Unos datos que guían a los fondos de alto riesgo para invertir en bolsa. La misión de estos equipos es predecir la economía. No es raro encontrarse con empresas que apuestan por la creación de grupos especiales para rastrear las huellas digitales. Aunque como señala Juan Merodio, experto en redes sociales, no solo hacen falta datos, también es preciso saber interpretarlos: “Hay que saber cruzarlos. Tenemos miles de datos almacenados, pero puede darse la paradoja de que no nos sirvan para nada. Por ello, en este caso me gusta hablar más de small data que de Big Data, es más lógico saber analizar pocos datos que almacenar miles sin ningún sentido”.

Las empresas van a dar cada vez más protagonismo a los equipos de analistas que sean capaces de interpretar los miles de datos que circulan por Internet y crear algoritmos para conocer las tendencias de los ciudadanos. También el futuro pasa por aplicaciones y webs predictivas. Una de las aplicaciones que rastrea nuestro comportamiento y es capaz de darnos información antes de que hayamos pensado en buscarla es Google Now. Así define el gigante estadounidense su aplicación:

Google te ofrece la información correcta incluso antes de que preguntes. Esas pequeñas píldoras informativas aparecen en forma de tarjetas cuando más lo

necesitas, organizando las cosas que necesitas saber y liberándote para centrarte en lo que es realmente importante para ti. Es útil durante el día, cuando te levantas con el tiempo y el tráfico. Y más adelante con noticias, deportes e información de bolsa en la que estás interesado. Cuando viajas, con recordatorios del vuelo y con cosas interesantes que visitar a tu alrededor.

Los supermercados estadounidenses Target, unos de los más innovadores en servicios para clientes, utilizan el posible poder predictivo de las redes sociales sumado a la gestión de los envíos y lectura de los correos promocionales: relaciona la lista de la compra con 25 productos y parece que puede saber detalles de sus compradores, como situación actual, embarazos, comienzos de vacaciones, etc., lo que le permite mandar publicidad específica de productos a esos grupos.

Amazon, el gran supermercado de la Red, con su aplicación abierta Machine Learning que permite a desarrolladores de cualquier lugar del mundo aplicar modelos de aprendizaje automático para máquinas, aspira a ayudar a las empresas a usar todos los datos que han recogido para mejorar la calidad de sus decisiones. Parece que el propio servicio de nueva creación de envíos en una hora, que acaba de arrancar el gran almacén virtual en la ciudad de Nueva York, no podría concebirse sin este software. Como explica la compañía:

Amazon Machine Learning puede ayudar a cualquier sitio web a proporcionar una experiencia al cliente más personalizada mediante el análisis predictivo para recomendar artículos u optimizar el flujo del sitio web a partir de las acciones anteriores del cliente. Puede procesar información de forma libre de sus clientes, incluidos mensajes de correo electrónico, comentarios o transcripciones de conversaciones telefónicas y recomendar acciones que aborden sus preocupaciones de la forma más adecuada. Puede utilizar Amazon Machine Learning para analizar el tráfico en los medios sociales, descubrir los clientes que tienen un problema relacionado con el soporte de un producto y ponerlos en contacto con el especialista de atención al cliente adecuado.

En la vida real, por ejemplo, esta aplicación permite prevenir que el cliente va a cambiar de compañía o predecir las nacionalidades de los huéspedes en los establecimientos hoteleros, lo que permite adelantarse a su llegada y ofrecerles servicios especiales que les gusten.

El estudio Predicting the Future With Social Media muestra la correlación entre la expectación que genera en Twitter con los ingresos en taquilla. Los modelos que se están desarrollando para la industria del cine o de la televisión sirven para predecir la taquilla de la primera semana en un 90% gracias al ruido que se ha generado antes del estreno. Otro ejemplo claro de las posibilidades del análisis de los datos que se dejan en las redes sociales es el de la cadena Wall Mart. Sus análisis permiten aumentar en un 24% las predicciones que se podrían hacer estudiando solo los métodos tradicionales. Gracias a ello, permite a la empresa calcular el volumen de compra y devoluciones y ahorrarse cantidades que rondan cifras de dos dígitos. Pero cuando, además, los datos ya no solo proceden de lo declarado en las redes

públicas, sino de medidas tomadas directamente de las personas, los análisis pueden ser extraordinariamente precisos. Así sucede cuando, mediante la colocación de pulseras y sensores, algunas compañías de videojuegos y de cine miden las reacciones de los espectadores y jugadores a lo que ven en pantalla. Parte del éxito de la película *The Revenant* del mexicano Alejandro González Iñárritu procede de que su departamento de marketing experimentó primero con las exhibiciones a puerta cerrada para prensa e invitados, y con los datos que indicaban en qué momentos la película emocionaba más, o ponía en tensión al público, determinaron la mejor forma de publicitarla. No sabemos qué parte del éxito ha tenido el Big Data, pero parece lógico atribuirle algunas entradas a la película.

El autor de la *Influencia oculta* de las redes sociales, Nicholas Christakis, destaca que “comprender las redes sociales y cómo se forman y funcionan nos puede ayudar a comprender desde crímenes o guerras a fenómenos económicos como el consumo o la caída de los mercados”. El estudio de IBM *Computational Discovery of Personality Traits from Social Media for Individualized Experience* demuestra cómo el análisis de los tuits y mensajes en Facebook puede definir la personalidad de alguien. Dependiendo del número de los mensajes que se publican en el muro de Facebook se puede conocer si una persona va a dejar de ser soltera y se va a casar. También sus características —hedonista, altruista, depresiva...—, una información que puede utilizar el departamento de recursos humanos de las empresas en los procesos de selección. Otro estudio, realizado por investigadores del *Psychometrics Centre* de la Universidad de Cambridge (Reino Unido) y del departamento de Ciencias de la Computación de la Universidad de Stanford estadounidense consiguieron, tras realizar un test de 100 preguntas a 90.000 usuarios de la red social, detectar los principales rasgos psicológicos de los encuestados gracias al estudio de los “Me gusta”. Las máquinas consiguieron, con solo 10 “Me gusta”, dibujar el perfil de la personalidad de los usuarios. Y acertaron mucho más que los amigos “reales” de estos usuarios. Los investigadores de Cambridge son los mismos que crearon un algoritmo que, tras analizar los likes, fotos y comentarios de los usuarios, permitía diferenciar entre usuarios blancos y de raza negra en el 95% de los casos; acertó en un 85% de las ocasiones determinar los usuarios demócratas y republicanos, y sobre la orientación sexual supieron quién era gay en el 88% de los casos. Con una base de datos de más de 58.000 personas y unos 55.000 “Me gusta” únicos, crearon una gigantesca matriz con cerca de 10 millones de resultados. En el 78% de los casos señaló quién era inteligente y en el 68% si alguien era emocionalmente estable. Como en tantos otros avances tecnológicos la frontera entre el abuso y la utilización positiva es de apenas un milímetro.

No solo las empresas saben que el estudio de las redes sociales es vital para sobrevivir en el futuro. El presidente estadounidense Barack Obama variaba su discurso en las pasadas elecciones en función de lo que ponían en las redes sociales los ciudadanos a los que se iba a dirigir en un mitin. Una estrategia que le permitió conectar más con esas personas al hablarles de lo que realmente les interesaba.

Pero el comportamiento de las personas en las redes presenta a su vez muchas dificultades para ser medido con precisión, dejando aparte el

importante hecho de que en muchos casos lo expresado no coincide con lo pensado (podemos decir que una cosa nos gusta solo por intentar ganar un concurso); el análisis del lenguaje humano, lleno de ironía, emoticonos, expresiones cambiantes, etc., hace muy difícil estimar con precisión las opiniones.

Otros sectores

La educación adaptativa

En Estados Unidos, los profesores tienen la posibilidad en los cursos online de ver lo que están consultando sus alumnos. Los datos de las visitas digitales a las lecciones se almacenan, lo que posibilita que el profesor se pueda dar cuenta de si sus alumnos asimilan los conocimientos. Uno de ellos se percató que cuando sus estudiantes llegaban a la lección ocho regresaban casi todos a la lección tres, que era una clase de repaso de matemáticas. El profesor supo, gracias a los datos que registraban el comportamiento de sus alumnos, que les debía preparar mejor si quería sacar el curso adelante. El Big Data también puede frenar las deserciones en la universidad como recoge un artículo de The Guardian. En Inglaterra, según el Consejo de Financiación de la Educación Superior de Inglaterra (HEFCE), más del 8% de los estudiantes abandonan sus estudios universitarios en el primer año. Esto le cuesta dinero a las universidades y los estudiantes pierden tiempo, dinero y autoestima. Algunas instituciones están utilizando los datos que poseen sobre sus alumnos para adaptarse a sus necesidades. En la Universidad de Nottingham Trent, el profesor Ed Foster ha estado dirigiendo un programa piloto que analiza factores que indican la participación de los alumnos en el mundo universitario: el uso de la biblioteca, los lectores de tarjetas de identificación en los edificios y el uso de los entornos de aprendizaje virtuales. Todo ello le sirve para identificar las necesidades de los alumnos o analizar su evolución. Y detectar los perfiles que están descontentos y desmotivados.

Varias tendencias educativas actuales como el “aprendizaje adaptativo” están basadas en la personalización de los itinerarios de estudios y estos solo pueden basarse en un manejo inteligente de los perfiles y datos de los alumnos.

Big Data y banca

Si Francisco González, presidente del BBVA, considera que en banca “el dinero son datos” y que “la clave del éxito de los bancos que sobrevivan en el nuevo entorno competitivo será maximizar el uso de los datos en beneficio de sus clientes. Los bancos del futuro serán compañías de software, orientadas a convertir esos datos en conocimiento útil para ofrecer a cada cliente las mejores soluciones y nuevos productos”, entonces no ha de extrañarnos que la banca en todo el mundo esté interesada en que se pueden calcular dos billones de bonos en un minuto para descubrir riesgos e identificar oportunidades, analizar 180 millones de préstamos en un día, estudiar 40 millones de correos electrónicos al mes para conocer las necesidades de los clientes... Todo ello es posible gracias al Big Data.

El experto de IBM Wolfram Rozas señala cuatro aspectos en los que la

recogida masiva de datos ayuda a la banca: optimización de ofertas, servicio a clientes y eficiencia, detección e investigación de fraude y riesgo de crédito y de contrapartida.

Los datos permiten a los bancos saber lo que dicen los otros clientes, una información que les permite modificar productos o crear nuevos para adelantarse a lo que va a pedir el mercado. Marco Bressan, científico de datos de BBVA Data & Analytics, destaca en una entrevista en la web del banco las posibilidades de la recogida masiva de datos: “Una gestión más eficiente del dato permite crear campañas comerciales mejor orientadas. Un mayor conocimiento del cliente nos permite también entender mejor su riesgo crediticio [...]”. Para Bressan, el uso de los datos ofrece beneficios inmediatos en áreas como la de riesgos o fraude. Y garantiza un conocimiento de gran alcance. Permite tener una visión más clara y objetiva de las operaciones, apoyar la toma de decisiones y anticipar su impacto. “Es habitual oír hablar sobre omnicanalidad, pero en la realidad no existe este concepto sin un verdadero conocimiento del cliente. Y no hay conocimiento del cliente sin intentar usar la traza que deja y que poseemos a nuestro alcance”. En los bancos se pueden encontrar millones de datos: desde los pagos hasta los movimientos de cuentas, pasando por contratos que dan información de millones de personas. En una sociedad bancarizada, el Big Data permite a las sucursales financieras conocer el tipo de público al que se dirigen, ayudar a la fidelización de clientes y la gestión de la omnicanalidad —experiencia de compra homogénea en cualquier canal— sean estos físicos (las oficinas) o digitales (las apps con las que manejamos nuestras cuentas).

A comienzos de 2016 algunos países del norte de Europa anunciaron que no se podría utilizar más dinero físico, que no se podría pagar nada con billetes o monedas. Este nuevo cambio en el propio corazón del sistema financiero capitalista no podría darse si no se hubieran extendido las terminales de pago para tarjetas, con diferentes tecnologías. Es un signo más de cómo avanza el Internet de las cosas y de cómo va a afectar a nuestros comportamientos habituales, a la decoración y formas de nuestros comercios e incluso a nuestros bolsillos y carteras.

El sector de la energía, obligado a utilizar Big Data

Un proverbio chino dice que cuando soplan vientos, unos construyen muros y otros molinos; por eso, cuando se trata de que la generación de energía para el consumo de industrias y ciudades sea objeto de mejoras gracias al Big Data, Vestas, compañía de sistemas de generación eólica que fabrica y usa molinos de aspas, adoptó una solución Big Data para poder procesar y analizar en tiempo real los grandes volúmenes de datos sobre las condiciones del viento y las cambiantes previsiones climáticas con la suficiente antelación para poder reaccionar. Sus sistemas procesan en tiempo real la actividad y condiciones de más de 45.000 aspas en 66 países, tiene datos de los últimos 15 años que hacen un total de 3 petabytes. Con este cambio en su estrategia, Vestas afirma que ahora prevé la colocación y puesta en marcha de una turbina en unos 15 minutos frente al mes que tardaba anteriormente.

La compañía de electricidad Iberdrola también está reenfocando toda la gestión de sus más de 11 millones de contadores que generan casi 300

millones de registros diarios, así como toda otra serie de aparatos colocados en las diferentes partes de la Red, con tecnologías de Big Data. De esta forma podrá planificar mejor el consumo eléctrico e incluso estimar con más exactitud las facturas porque está basada en datos reales y no solo en estadísticas.

Incluso ENDESA ha lanzado recientemente un concurso, “Endesa Datathon”, para recabar ideas que les permitan aprovechar mejor los 120 millones de registros diarios que consiguen en España.

Aunque quizás las prisas y el interés de las compañías eléctricas también vengan derivadas de la directiva europea que les obliga a teleadministrar todos los contadores residenciales antes de 2018.

La industria de la moda

A primera vista uno pensaría en que en la industria textil y de la moda el Big Data sirve para mejorar los procesos de distribución, tan importantes hoy, pero en realidad sirve para mucho más. El modelo de fast fashion que utilizan firmas como Zara o H&M ha de manejar con fluidez los datos de pedidos, las ventas o incluso las tendencias de colores o texturas. Zara, por ejemplo, lanza de forma continua unas 18.000 nuevas prendas cada año que van a las miles de tiendas que tiene en 88 países. Esta renovación continua ha sido uno de los cambios más importantes frente al tradicional modelo de lanzamiento de novedades una vez por temporada; dicen los expertos que es la clave del éxito de la compañía española. Para manejar este volumen de forma eficaz es necesario apoyarse en técnicas de Big Data. Las tiendas son ayudadas por algoritmos que predicen hasta las tallas que más se van a vender en cada comercio y que ayudan a los encargados a manejar las miles de unidades que reciben cada semana.

Pero no solo los grandes gigantes utilizan los datos; la empresaria y diseñadora estadounidense Michelle Lam tiene el mismo sueño con su compañía, True&Co: conseguir el sujetador perfecto. Para ello está analizando con tecnologías de Big Data millones de datos biométricos de mujeres para poder fabricar sostenes a medida, personalizados. Lam utiliza algoritmos para ajustar los tejidos a las formas del pecho femenino. “Cuanto más datos manejes, más bonita y cómoda será la prenda.” La diseñadora ha analizado la medida de más de 200.000 mujeres y ha determinado 6.000 patrones diferentes con los que fabrica su selección, que puede encontrarse online.

Por otro lado, y ya entrando en el mundo del Internet de las cosas, la moda parece ser uno de los sectores que más puede verse afectado. Los accesorios portátiles como relojes, gafas y bolsos están pasando a ser dispositivos que recogen y emiten información útil para quienes los llevan encima. Las gafas, como las que desarrolla Google, son una extensión no ya de la vista como hasta ahora, sino que proyectan y permiten que la persona que las lleva tenga acceso a muchos más datos sobreimpresionados a aquello que tiene delante.

Toda la gama de ropa tecnológica —wearable— también está ligada de forma íntima tanto a las aplicaciones relacionadas con la monitorización de

indicadores de salud en las personas como con la relación con el entorno en el que se llevan puestas. La moda tecnológica enganchada a Internet abre un amplio campo de posibilidades para las marcas y los individuos que se desarrollarán en los próximos años. En una estimación bastante conservadora, realizada por el Centro Tecnológico de Cataluña Eurecat, aseguran que la facturación por las prendas tecnológicas, entre las que se incluyen las conectadas, pasará de apenas 9 millones de euros en 2013 a más de 130 en 2018. Este mismo centro asegura que “es una revolución del textil, impulsada por los nuevos materiales, procesos de producción y canales, la proliferación de datos, dispositivos y sensores y la hiperconectividad”.

El fútbol

Ahora los entrenadores de fútbol tienen un nuevo aliado: Big Data. En equipos como el Real Madrid, el técnico tiene entre sus ayudantes a un especialista en análisis de información numérica. Las pizarras están siendo sustituidas por tabletas y allí se muestran las estadísticas en tiempo real del contrincante y propias, los análisis de resultados de otros partidos, estudios previos a jugadores antes incluso de que un ojeador del club los localice en Brasil o Argentina. Parte de las estrategias y cambios que hace un equipo como el Real Madrid tienen que ver hoy con una pantalla en la que el entrenador observa los resultados de múltiples algoritmos que procesan millones de datos en tiempo real.

Pero, además, el Real Madrid, al igual que los primeros equipos del mundo, ha incorporado nuevos sistemas de análisis para monitorizar el estado de su plantilla. Los jugadores llevan sensores durante los entrenamientos que recogen en tiempo real datos de sus movimientos, velocidad, esfuerzo realizado, fatiga acumulada, etc., con lo que el cuerpo técnico cuenta con una mayor precisión a la hora de analizar la capacidad de cada uno de los miembros del equipo.

Pero el fútbol todavía escapa del campo de las predicciones estadísticas de resultados a las que el béisbol parece estar más acostumbrado. El portal de Nate Silver, que tan buen resultado ha tenido en la anticipación de resultados electorales o de desenlaces en la liga de béisbol norteamericana, falló estrepitosamente en sus análisis sobre los posibles ganadores del Mundial de fútbol de 2014. Parece que todavía no hemos encontrado la forma de procesar los datos adecuados, o que los efectos del azar o factores todavía no contemplados son demasiado grandes.

CAPÍTULO 4

Cómo nos afecta en nuestra vida diaria

A comienzos de 2016, una información periodística disparaba todas las alarmas: si uno intentaba localizar webcams en “Shodan”, un buscador especializado en Internet de las Cosas, podía encontrarse desde imágenes en directo de cunas con bebés durmiendo a plantaciones privadas de marihuana. Un montón de sistemas de vigilancia de hogares abiertos al público. Un protocolo transmisión de datos poco seguro era el culpable, pero hay mucho más.

En el informe de McKinsey⁶ se pone el punto de mira en “las políticas relacionadas con la privacidad, seguridad, propiedad intelectual, e incluso con la responsabilidad”. Son aspectos, advierte dicho documento, que deberán ser abordados en breve para continuar con el desarrollo de los sistemas de Big Data.

El acceso a los datos críticos de las empresas es cada vez más una necesidad para poder integrar la información de múltiples fuentes de datos, a menudo de terceros, y analizarlos. “Hay que tener en cuenta los límites de esta normativa. Ver si el usuario ha habilitado el permiso para obtener esa información o no. Asimismo, hay que trabajar mucho las condiciones y términos de uso, ya que si no después nos encontraremos con un problema con el usuario”, explica Óscar López, presidente del Grupo Regulación de AUTELSI.

Gabriel Martín, directivo en España de Symantec, asegura que la seguridad “debe contemplarse desde el principio.”. En su opinión, “se necesita un sistema de acceso que permita determinar los niveles de acceso dependiendo de las edades”. Además, también se necesita un sistema que “deje una huella para que se pueda disponer de esos datos y que al mismo tiempo esté todo relacionado con el cumplimiento de normativas, tanto internas como legales”.

¿Cómo usar estos datos? Esta pregunta es una de las que más problemáticas está generando en torno a la puesta en práctica de las herramientas de Big Data. Así, en una entrevista para el Centro de Innovación BBVA, Elena Alfaro, gerente en el área de Business Discovery BBVA y experta en Smart Cities, aborda la cuestión sobre la privacidad de los datos y su uso. En este sentido, incide en que la información en su proyecto “se use para mejorar la ‘inteligencia de la ciudad’ y no tanto para determinar oportunidades de venta a personas concretas, a no ser que esas personas accedan a ello”.

“Los datos nos llegan ya como anónimos o enmascarados de acuerdo con la ley, de forma que no hay manera de identificar a ninguna persona o lo que

haya hecho. Nosotros creemos que hay muchas posibilidades en el uso y explotación de la información agregada, no personal. Digamos que nos interesa el qué, el dónde y el cuándo, pero no el quién”, explica Alfaro. Nos guste o no, no hay que olvidar que somos datos. Un estudio del MIT probó que los patrones de uso de las tarjetas permiten descubrir la identidad del 90% de una muestra de 1,1 millones de personas anónimas. Los investigadores aplicaron algoritmos a una base de datos de un gran banco y vieron cómo con un promedio de cuatro transacciones, el día y la tienda, les bastaba para identificar a las personas en el 90% de los casos. Una muestra del poder de los metadatos.

La protección de datos⁷

El director de la Agencia Vasca de Protección de Datos, Iñaki Pariente de la Prada, explica que en España la Ley General de Protección de Datos es del año 1999 (traspone una directiva europea del 95), engloba la protección de datos en general y no regula sectorialmente ningún campo de actuación. Una ley que se hizo cuando solo unos pocos estaban familiarizados con Internet y el término Big Data era prácticamente desconocido. El experto vasco remacha: “No hay una regulación específica ni para cloud computing o telefonía móvil. Lo que se hace es aplicar los principios generales de la protección de datos a cada sector, a cada ámbito nuevo que surge. Sería conveniente tener una ley más actual, pues la directiva tiene 20 años. Se hizo con unos parámetros en los que no había Internet y los actuales son que todo el mundo se pasea por la calle con varios dispositivos encendidos”. Pariente de la Prada añade: “Pese a no haber una ley específica de Big Data, se prevé la obligación a las administraciones y empresas de todo el mundo a que hagan tratamiento masivos de datos, la redacción de un texto denominado evaluación de impacto, que tiene que elaborarse antes de la recogida o tratamientos masivos de información, en donde se analice si se va a respetar o no la normativa de privacidad y cómo se va a hacer”.

El experto vasco incide en que hay que respetar la ley tanto en la recogida como en el tratamiento y en la cancelación de los datos personales:

Hay dos principios esenciales en la recogida de datos: información y consentimiento. Supuestamente, una persona cuando da sus datos a otra tiene que haber sido informada previamente de para qué se van a recoger —la finalidad—, cómo se van a utilizar, dónde se van a utilizar, en qué ficheros se van a guardar y la seguridad que van a tener. Teóricamente, una vez que he informado de todo esto, la persona consiente o no dar esa información. Y si recojo datos en una página web gracias a una casilla, el sistema tiene que permitir guardar esa casilla para acreditar que se dio el consentimiento. A efectos de prueba.

Cuando no se puede obtener ese consentimiento en la recogida, por ejemplo, de millones de datos para realizar investigaciones médicas o con otras finalidades, esos datos tienen que estar anonimizados para eliminar la vinculación de los datos a personas concretas y evitar su identificación. Señala el experto vasco:

Las empresas están almacenando actualmente miles de datos con los

denominados wearables. Nos ponemos pulseras, dispositivos y en 5 años tendrán muchísima información acumulada de cada uno. El problema es que no sabemos para qué están almacenando toda esa información, cómo la van a utilizar en el futuro. La finalidad no está nada clara. Muchas de ellas están en Estados Unidos, no están sujetas al derecho español. Eso implica que si tengo un problema no voy a poder acudir a que me tutelen o ayuden las instituciones estatales.

Para Pariente de la Prada no hay otro camino que la anonimización en la recogida masiva de datos. Aunque también alerta que esa anonimización puede ser reversible en algún momento: “Te dicen que técnicamente hay garantía de que no puede ser reanonimizado, pero realmente no sabes, aunque en ese momento sea cierto, si mañana la tecnología va a evolucionar. Pasa a ser un problema técnico”.

La recogida masiva de datos preocupa en Europa. Prueba de ello es que el grupo de expertos ha pasado de reunirse de forma esporádica cuando nació a hacerlo de forma asidua. La directiva de la Protección de Datos de 1995 creó un grupo de trabajo que reúne a todas las autoridades de protección de datos de los países miembros y ese grupo de expertos emite, ahora, cerca de 4 dictámenes al año. El de abril de 2014 fue sobre “técnicas de anonimización” y estableció los criterios técnicos para anonimizar los datos. “Es un dictamen muy interesante que marca las pautas que hay que seguir. De alguna forma es como si actualizase la ley. Trata cosas que la norma no contempló cuando nació y que hay que ir interpretando. La ley no cambia, pero con estos dictámenes se actualiza, son muy importantes”, explica el experto vasco.

Estas son algunas de las recomendaciones del dictamen europeo: “Existen limitaciones inherentes a algunas técnicas de anonimización. Los responsables del tratamiento deben ponderar seriamente estas limitaciones antes de escoger una técnica u otra para desarrollar un proceso de anonimización. Asimismo, deben atender a los fines previstos para la anonimización, como proteger la privacidad de las personas cuando se publica un conjunto de datos o permitir que se consulte algún tipo de información contenida en dicho conjunto”(Dictamen 05/2014 sobre técnicas de anonimización).

Y alerta sobre la dificultad de anonimizar los datos: “Las técnicas descritas en este documento no cumplen al cien por ciento los criterios de una anonimización efectiva, a saber: no es posible singularizar a una persona; no existe vinculabilidad entre los registros de una misma persona y no se puede inferir información sobre una persona. No obstante, dado que cada una de estas técnicas entraña, en mayor o menor medida, alguno de estos riesgos, es imprescindible diseñar cuidadosamente la aplicación de una determinada técnica a la situación concreta de que se trate, o bien la implementación de una combinación de estas técnicas, a fin de obtener resultados más sólidos” (Dictamen 05/2014 sobre técnicas de anonimización).

‘Selftracking’, la fiebre por los datos

¿Nos estamos convirtiendo en adictos a los datos? La medición constante de los datos vitales del cuerpo humano con wearables que permiten monitorizar

nuestra vida relacionándolas con las constantes vitales que en ese momento tengamos (la medida del ritmo cardíaco, la tensión arterial, la respiración, las calorías quemadas durante el ejercicio) aspiran a revolucionar el mundo de la salud. Relojes, gafas, ropa, pulseras, collares o zapatillas. Todo sirve para medir permanentemente el cuerpo. La tecnología se pega a la piel y acompaña a las personas desde que son bebés. Los dispositivos no solo sirven a los adultos. Los pañales inteligentes permiten a los padres escanear los fluidos del móvil para controlar la salud de sus hijos. También utilizan las pulseras en sus bebés, que, sin ser conscientes, son cobayas tecnológicas, early adopter. Anillos, tobilleras que calculan su ritmo cardíaco, que avisan al teléfono de sus padres si su cadencia respiratoria varía durante la noche. Hay wearables de todo tipo para las personas tecnoadictas que funcionan las 24 horas del día. La madre o el padre que monitoriza el sueño de niños sanos por la mañana seguramente utilizarán por la tarde un dispositivo que almacenará todas sus constantes vitales mientras hacen footing.

Todo lo que puede ser medido se mide. Un fenómeno que se conoce como selftracking (autorastreo o cuantificación). Los selftrackers utilizan apps y otros medios electrónicos para recopilar todos los datos de sus cuerpos, buscan averiguar el máximo sobre sí mismos. Uno de sus lemas es “Lo que puede ser medido puede ser mejorado” y por ello cuantifican todo. Gary Wolf y Kevin Kelly son los fundadores del movimiento Quantified Self, que nació en 2007 en Estados Unidos. Un modo de vida que gira alrededor de los datos. Los selftrackers necesitan saber exactamente lo que consumen, gastan, comen y duermen. Un selftracker sabe que ayer dio 5.997 pasos, los minutos que ha dormido, su índice de grasa corporal, la cantidad de azúcar en la sangre y los centilitros de vino que se ha bebido en los últimos 10 años. No solo mide su estado físico, también registra el entorno —la humedad o contaminación— y todo lo que le realiza cada día —uso de las redes sociales o las horas que pasa delante del televisor—. Creen en el autoconocimiento y la optimización de su estilo de vida gracias al seguimiento de las cifras que emite su cuerpo —mediante apps y dispositivos—, y llevan hasta el límite el aforismo “Conócete a ti mismo”. Algunos médicos alertan de esta práctica que califican de obsesiva y en algunos casos vana, pues se acumulan miles de datos sin ningún sentido.

El escándalo Snowden

Un informático de 30 años, Edward Snowden, nos despertó bruscamente del sueño acerca de las posibilidades y las bondades —pese a los recelos en el tema de la privacidad— que parecía brindarnos la recogida masiva de datos, y nos mostró la cara más turbia del Big Data. El caso Snowden, que estalló durante el verano de 2013 y que todavía está lejos de haber terminado, demostró que decenas de países estaban siendo espiados por Washington. El exanalista de la Agencia Nacional de Seguridad (NSA, por sus siglas en inglés) reveló la existencia de un programa global de espionaje de las comunicaciones telefónicas y del tráfico de Internet. Una información que fueron destapando con cuentagotas los periódicos The Guardian y Washington Post —ambos se llevaron el Premio Pulitzer en 2014 por la publicación de los documentos que prueban la interceptación sistemática de millones de comunicaciones de voz y datos— y que sacudió las relaciones de Estados Unidos con los países aliados.

La recopilación indiscriminada de datos por parte de la NSA afectó a decenas de políticos mundiales —la canciller alemana Angela Merkel o los tres últimos presidentes franceses, entre otros muchos— que fueron espiados por el país norteamericano y dejó al descubierto la connivencia de los gigantes tecnológicos estadounidenses —Google, Facebook, Microsoft— con la Agencia Nacional de Seguridad. Reino Unido también fue cómplice de este espionaje, según revelaron los documentos filtrados por Edward Snowden.

“Can you hear me now?” (“¿Pueden oírme ahora?”) es el primer tuit que curiosamente el exanalista escribió tras estrenar cuenta de Twitter a finales de septiembre de 2015. Antes, se escuchó y se habló mucho de Snowden tras la filtración de los documentos, tanto por su vida personal —David Miranda, su compañero sentimental, fue arrestado en Londres y se le aplicó la ley antiterrorista en un claro abuso de poder por parte de las autoridades inglesas— como por la profesional. En el aeropuerto Sheremétievo de la capital rusa, Snowden pasó 38 días en la zona de tránsito esperando la solución de su caso y terminó pidiendo asilo político a Rusia. Una decisión que agravó las relaciones entre Moscú y Washington. Todo esto ocurría en 2013 tras el escándalo mayúsculo. Desde ese año no han dejado de aparecer documentos que no solo han tenido consecuencias políticas, sino también económicas, pues miles de empresas desconfían de las compañías norteamericanas a la hora de trabajar con ellas. También han removido a los ciudadanos, que vieron incrédulos cómo en nombre de la seguridad se rastrearon sin pudor millones de comunicaciones, se gastaron más de 250 millones de dólares en influir en los diseños tecnológicos de los gigantes de Silicon Valley, se craquearon comunicaciones encriptadas o se almacenaron miles de llamadas telefónicas privadas. La Administración norteamericana realizó un espionaje masivo recolectando información a través de Google, Facebook, Apple y Skype gracias al programa Prisma NSA, que permite almacenar los correos electrónicos, búsquedas de Internet, archivos enviados y conversaciones online de cualquier ciudadano no estadounidense fuera de las fronteras de Estados Unidos.

El caso del exanalista de la CIA ya tiene una película, dirigida por Oliver Stone, que detalla la vida de uno de los hombres que para algunos es considerado un traidor —Estados Unidos justifica el espionaje masivo como arma indispensable para luchar contra el terrorismo— y para otros un héroe. Stone muestra todos los pasos del informático que, antes de llevarse de la oficina más de un millón de documentos secretos de la NSA y volar a Hong Kong con el propósito de contarle a un periodista el inmenso programa de vigilancia electrónica estadounidense, se preparó para ser marine en Irak tras rechazar la primera oferta que le hizo la NSA para trabajar como decodificador. Antes de su huida de Estados Unidos, Snowden trabajó en países como Suiza o Japón realizando labores de contraespionaje para la NSA y para otras agencias de la inteligencia de Estados Unidos. Aunque no ha vuelto a su país de origen, este informático que ahora tiene 32 años ha conseguido algún cambio en el mundo Big Data. Prueba de ello es que, a finales de 2015, el presidente de Estados Unidos Barack Obama rubricaba una ley para limitar los poderes de la Agencia de Seguridad Nacional. Algo que nadie había conseguido ni pensado tras los atentados del 11-S.

Datos para luchar contra la desigualdad

Erradicar la pobreza en el año 2030 es uno de los objetivos de los líderes mundiales. En septiembre de 2015, los representantes de las Naciones Unidas lanzaron en Nueva York los Objetivos de Desarrollo Sostenible (ODS), el proyecto más ambicioso en la historia del planeta en el empeño de construir una sociedad global y mejor. Para lograrlo señalaron la importancia de los datos. Los expertos advirtieron que si no consiguen una mejoría drástica de los datos sobre las personas más pobres del mundo, este ambicioso proyecto dejará de cumplir con su objetivo de “no dejar a nadie atrás”. El proyecto tiene 17 objetivos globales y 168 metas en educación, la seguridad alimentaria, el empleo o el medio ambiente. Pasa de gestionar millones — como sucedía con los Objetivos de Desarrollo del Milenio (ODM), que, tras 15 años de vigencia, ha dejado paso al nuevo programa— a billones, alrededor de 11,5 billones de dólares al año durante los próximos tres lustros. Para alcanzar el objetivo aspiran a mejorar esos datos utilizando datos. En la reunión señalaron que millones de personas se pierden dentro de la división de datos y recalcaron la importancia de la formulación de políticas de información con datos para ayudar a las personas más vulnerables del mundo. “Sin datos estamos volando a ciegas, y nosotros no podemos tomar las decisiones políticas que deben estar basadas en la evidencia.” Se estima que puede haber 350 millones más de personas que viven en la pobreza de las que se conoce actualmente. Tampoco se sabe el número de niñas que se casan antes de cumplir los 18 años ni cuántas mujeres hay en el porcentaje de los pobres del mundo. Tampoco el número de niños de la calle en todo el mundo ni el número de personas en el mundo que tienen hambre.

Un artículo de The Guardian recoge el desconcierto de muchos de los expertos que estuvieron en la jornada de la ONU —“La situación de los datos en muchos países es realmente preocupante. Si usted no sabe el número de personas que vive en su país ¿cómo se pueden hacer políticas serias?”—. El artículo incide en que la colaboración entre el sector privado y organizaciones no gubernamentales serán cruciales para tapan las lagunas de datos en los países en desarrollo. Los datos de las empresas privadas se pueden utilizar junto con los métodos de captura de datos tradicionales como las encuestas, que son costosos y difíciles de llevar a cabo. En Costa de Marfil, por ejemplo, la empresa de telecomunicaciones Orange ha lanzado una campaña de recogida de datos entre investigadores y universidades para encontrar soluciones a los problemas de desarrollo en el país. A medida que las relaciones público-privadas se hagan más fuertes, será de vital importancia la calidad de los datos y que estos cumplan un conjunto de normas internacionalmente reconocidas respetando la privacidad de los ciudadanos. Unos datos que, para la ONU, son necesarios para erradicar la pobreza.

Destrucción y creación de empleo

La automatización que permite el manejo de Big Data y el despliegue de los dispositivos del Internet de las cosas configuran un futuro sombrío para muchas profesiones actuales. Todo aquello que se puede automatizar se automatizará, y esta tautología es la causa de que cada día que pasa la mayoría del trabajo manual que venían haciendo miles de operarios pase a

“manos” de sensores, máquinas e incluso robots.

Hasta las cajas de supermercado están siendo sustituidas por cajas inteligentes que mediante las etiquetas de los productos calculan en un segundo el importe de la compra; el cliente no necesita apenas más que acercarse o dar una orden de voz a su teléfono móvil para realizar el pago de la operación. No hay ninguna persona intermedando. Los jardines y su riego se ajustan cada vez más con sensores e incluso la producción de leche empieza a estar cada vez más controlada por diferentes sensores. Estos son solo algunos de los muchos ejemplos de un cambio copernicano en el mundo laboral. Mientras aparecen nuevas profesiones, como los científicos de datos, se destruyen otras, lo que se está convirtiendo en uno de los mayores retos para gobiernos y sociedades de todo el mundo.

Capítulo 5

La hora del Internet de las cosas

Me da miedo entrar en la cocina
me da miedo lo que pueda haber
la tostadora se ha vuelto asesina
el lavaplatos no me puede ver
Se han rebelado todos a la vez
La Turmix, la plancha y la moulinex
Se han vuelto locos de repente, hay que ver
La aspiradora se niega a aspirar
dice que no, que no, ni hablar
y la nevera está leyendo a Marx
y me dice q la deje en paz
Se han rebelado todos a la vez
la Minipimer se ha unido también
Se han rebelado todos a la vez
la Minipimer se ha unido también
me han dicho que no me preocupe
que se irán, ¡dentro de un mes!
La rebelión de los electrodomésticos,
Alaska y Los Pegamoides

Un mundo conectado en el que también las máquinas y los objetos tienen un importante y nuevo papel al “hablar” entre ellos. Eso es el Internet de las

cosas (IoT por sus siglas en inglés, Internet of Things). Big Data e Internet de las cosas son dos conceptos íntimamente unidos. Mientras el primero es, de alguna forma y como señalan los expertos, el “petróleo” de la sociedad de la información, Internet de las cosas es un importante ecosistema donde se recoge, circula y se procesa ese “petróleo” formado por todo tipo de máquinas y objetos con sensores o disponibilidad para ejecutar acciones. Las personas cada vez portamos más sensores que incorporan y emiten información de manera continua. El móvil es uno de los principales, ya que solo con lo que él procesa se podría saber dónde estamos, el nivel de actividad que tenemos o, cada vez, más cosas de las que hacemos. Pero el Internet de las cosas va más allá del dispositivo que nos acompaña a todas horas, ya que la propia ropa está empezando a incorporar sensores o añadidos que realizan funciones concretas: hay hasta camisetas que se iluminan a la velocidad de los latidos del corazón y otras que muestran con su color los índices de contaminación del entorno recogidos por Internet desde un puesto municipal de análisis del aire cercano.

En 1999, Kevin Ashton, ingeniero del Instituto Tecnológico de Massachusetts (MIT), acuñaba el término Internet of Things —“Internet de las cosas”— para referirse a la relación entre la tecnología RFID (Radio Frequency Identification Device) e Internet. La tecnología de identificación por radiofrecuencia permite reconocer de forma automática cualquier objeto, animal o persona gracias a la información contenida en etiquetas electrónicas (tags) que portan. La etiqueta se compone de un pequeño chip de silicio unido a una antena de radiofrecuencia y, cuando entra en el área de cobertura de un lector, este envía una señal para que la etiqueta le transmita la información almacenada en su memoria, sin necesidad de contacto físico o visual, como sucedería, por ejemplo, para leer un código de barras. El sistema RFID permite comunicarse e identificarse a través de ondas de radiofrecuencia por un dispositivo transmisor/receptor.

La tecnología RFID es el germen de las tecnologías que, conectadas al protocolo de Internet IP, está cambiando la sociedad. Unas tecnologías que están presentes en todos los lugares en los que pasamos la mayor parte del tiempo, unas tecnologías que forman parte muchas veces de nuestro cuerpo, que están en los objetos que compramos y utilizamos, que miden lo que respiramos.

Somos datos —objetos, animales y personas— conectados. Ashton describió en 2009 el panorama tecnológico al que nos dirigimos actualmente:

Si tuviéramos ordenadores que supieran todo lo que hay que saber acerca de las cosas, mediante el uso de datos que ellos mismos pudieran recoger sin intervención humana, podríamos monitorizar, contar y localizar todo a nuestro alrededor, y, de esta forma, reducir increíblemente gastos, pérdidas y costes. Sabríamos cuándo reemplazar, reparar o recuperar lo que fuera y conocer si su funcionamiento está siendo correcto. El Internet de las cosas tiene el potencial para cambiar el mundo tal y como hizo la revolución digital hace unas décadas. Tal vez incluso hasta más.

Y definió en RFID Journal de esta manera el Internet de las cosas:

Es una infraestructura global interconectada, enlazando objetos físicos y virtuales a través de la explotación de la captura de datos y las capacidades de comunicación. Ofrecerá identificación específica de objetos y capacidades sensoriales y de conectividad como la base para el desarrollo de servicios cooperativos y aplicaciones independientes.

Buscamos analizar lo que la mente humana no puede. La conexión del universo físico y del universo virtual no deja de crecer y la conectividad a fuentes de información fijas o a elementos que pueden tener información de Internet transforma nuestras vidas. En los objetos inteligentes confluyen el mundo de la información (IT) con el mundo físico de las máquinas y de la producción (OT). Big Data e Internet de las cosas van de la mano y cambian la sociedad en donde los algoritmos son personalizados y el análisis masivo de datos revoluciona desde la forma de dirigir las empresas hasta nuestra vida diaria. El Internet de las cosas consigue que los objetos cotidianos sean inteligentes. Lo son las 24 horas del día gracias a la velocidad en la que se procesan los datos —Big Data— y la conexión desde cualquier lugar del planeta que permite a cualquier persona el acceso continuo a Internet.

En el siglo XIX las máquinas aprendieron a ejecutar órdenes; en el siglo XX aprendieron a pensar, y en el siglo XXI anticipan las cosas que van a suceder.

La consultora Mckinsey define nueve escenarios en los que el Internet de las cosas es protagonista y aporta un valor incalculable:

Humano: Los dispositivos que unimos a nuestro cuerpo o los chips que nos implantamos permiten monitorizar las constantes vitales y mejorar nuestra salud o incluso nuestra productividad. En las aplicaciones de la tecnología IoT en este contexto hay dos categorías principales: la salud y la productividad en el trabajo. El Internet de las cosas puede conducir a un cambio en la salud humana; por ejemplo, utilizando dispositivos conectados para monitorizar a los pacientes en su día a día —particularmente aquellos con enfermedades crónicas como la diabetes—. También mejora la adherencia a las terapias prescritas, evita hospitalizaciones (o complicaciones postoperatorias). En resumen, favorece la calidad de vida de cientos de millones de personas con el correspondiente impacto económico. Centrándonos en la productividad del trabajo: el ejemplo de las gafas de realidad virtual que muestran al trabajador lo que debe hacer para realizar su tarea de forma correcta es solo una muestra de lo que puede llegar a suponer el Internet de las cosas en el mundo laboral. En general, permite a las compañías rediseñar procesos y puestos de trabajo para conseguir una mayor eficiencia y efectividad.

Hogar: un amplio número de dispositivos logran que los hogares sean inteligentes y mejoran la seguridad. El Internet de las cosas inunda de dispositivos los hogares. Dispositivos que escanean lo que consumimos y nos dicen la cantidad de calorías que tiene nuestro plato favorito, frigoríficos que gestionan los alimentos y realizan los pedidos automáticamente o termostatos que regulan la temperatura de forma autónoma son algunos de los ejemplos que muestran cómo el Internet de las cosas facilita la vida diaria. Los sensores son capaces de identificar y monitorizar cualquier aparato conectado a Internet. A medida que estos dispositivos evolucionen, se estima

que el gran impacto económico en las casas vendrá de la mano de la automatización de las tareas del hogar. Se reducirá en 100 horas al año el tiempo que dedican las personas a las tareas del hogar. Otro de los grandes cambios llega gracias a la monitorización energética seguida de la seguridad en los hogares (con un alto impacto por los daños o incluso muertes que se pueden evitar).

Pequeñas y medianas empresas: el Internet de las cosas cambia los espacios en los que se fideliza al consumidor (tiendas, bancos o restaurantes) al ofrecer una experiencia más personalizada. Por ejemplo, permite guiar al consumidor en una tienda hacia un producto que ha visto online, o colocar los productos en una tienda basándose en los análisis del comportamiento de los clientes.

Oficinas: al igual que en los hogares, se monitoriza la energía y se garantiza la seguridad en las oficinas. Algunos de los beneficios claves en estos espacios de trabajo se relacionan con la seguridad mediante la utilización de cámaras de seguridad digitales a través de las cuales se podría monitorizar la actividad sin necesidad de vigilantes. Los espacios comerciales consumen el 20% de la energía en economías avanzadas como la estadounidense. Alrededor de la mitad de esta energía es consumida en oficinas y gracias a sistemas inteligentes de control se puede identificar, por ejemplo, si una habitación está vacía u ocupada, lo que permite activar mecanismos para el ahorro de dicha energía.

En fábricas y aquellos lugares con rutinas de trabajo repetitivas como hospitales o granjas, el Internet de las cosas permite la optimización del equipamiento y de las operaciones. Desde el punto de vista económico, el impacto es altísimo, pues incluye cualquier espacio empresarial con producciones en cadena. Abarca todo lo relacionado con mejoras en el mantenimiento de los equipos, la salud y la seguridad de los trabajadores. Juega un rol crítico en la fase posterior a la automatización, lo que se denomina la industria 4.0. Un término que se utiliza para describir la completa digitalización de los procesos de producción, que une el mundo digital y físico dentro de las fábricas. La industria 4.0 es la capacidad de monitorizar y controlar todas las herramientas de producción. Y utilizar esos datos para mejorar la productividad y la calidad de los productos.

Infraestructuras y construcción. Las grandes empresas petrolíferas, mineras o de construcción buscan ser más seguras y eficientes y han sido las primeras en adoptar las tecnologías del Internet de las cosas. Algunas explotaciones de crudo en el mar ya funcionan con 30.000 sensores para monitorizar su actividad. En la minería, por ejemplo, vehículos dirigidos ayudan en muchas de las operaciones, reduciendo costes y garantizando la seguridad. También se pueden utilizar sensores para controlar la maquinaria en uso, lo cual lleva a las compañías a cambiar la forma en la que realizan el mantenimiento de sus maquinarias. En vez de hacer un mantenimiento regular, periódico o reparar la maquinaria cuando se rompe, se pasa a un mantenimiento cuando existe una necesidad real que se conoce gracias a la predicción de los datos que analizan el funcionamiento de las máquinas.

Vehículos: el Internet de las cosas anticipa lo que va a ir mal en coches,

aviones o trenes. Los sensores aportan información básica sobre el entorno y los obstáculos al software de los vehículos. Evitan colisiones. Los asistentes de voz, la cartografía digital avanzada y los sensores y cámaras nos facilitan la conducción y los coches, en concreto, aspiran a ser autónomos, o semi, a corto plazo. Los expertos pronostican que gracias a la inteligencia artificial mejorarán el medio ambiente y la seguridad vial.

Ciudades: ecosistemas urbanos, espacios públicos e infraestructuras se benefician del Internet de las cosas. El control del tráfico, los metros inteligentes o la monitorización medioambiental son solo algunos de los ejemplos. Las ciudades son el motor del auge económico global —las 600 principales ciudades del mundo generarán el 65% del crecimiento económico en 2025— y la inteligencia busca su hueco. Utilizando la tecnología IoT, las ciudades se beneficiarán en cuatro áreas principalmente: transporte, seguridad y sanidad públicas, manejo de los recursos y prestación de servicios.

Outside: el Internet de las cosas mejora las rutas de barcos, aviones y otros vehículos gracias a sistemas de navegación avanzadas por medio de sensores.

Hoy en día casi todas las cosas pueden ser inteligentes en el sentido de poder comunicar información y poder interoperar a niveles básicos. El Internet de las cosas avanza a pasos agigantados para mejorar nuestras vidas. Para que la gente viva de forma más sana y segura. Los principales escollos a los que se enfrenta pasan por un colapso tecnológico en el transporte de datos o los requisitos de seguridad que garanticen la privacidad de las personas y un buen uso de los datos que no lleven a prácticas discriminatorias.

Para lograr el máximo potencial del Internet de las cosas es necesario que se produzcan avances tecnológicos, principalmente en tres áreas: por un lado, la mejora de ciertos elementos básicos (infraestructura), como son el abaratamiento de los costes, que los componentes hardware sean más “capaces” o mejoras en el software y en los datos analíticos; por otro lado, el desarrollo de determinadas normas (leyes y reglamentos favorables) y, por último, soluciones tecnológicas que permitan la interoperabilidad.

El informe *The Internet of Things: mapping the value beyond the hype* de McKinsey recalca como “esencial” conocer cuáles son los “dinamizadores” o “habilitadores” (técnicos y no técnicos) que van a permitir maximizar el impacto del IoT. Los dinamizadores técnicos (por ejemplo, el cifrado de alta seguridad) y los no técnicos (políticas o reglamentos) deben, al mismo tiempo, garantizar la privacidad, la confidencialidad y la seguridad de los datos del IoT.

Un hardware de bajo consumo y barato y una conectividad ubicua se antojan necesarios para conectar millones de dispositivos. El Internet de las cosas persigue que los ciudadanos se conviertan en sensores gracias a sus teléfonos móviles. Y aunque actualmente muchas aplicaciones son técnicamente funcionales, el alto coste de los componentes, como es el caso de los nodos de los sensores (con la comunicación y las fuentes de alimentación) hace dichas aplicaciones poco prácticas. Las personas pueden llevar actualmente decenas de dispositivos sin cables. Abaratar los costes de estos aparatos y eliminar las

baterías de los cientos de dispositivos con los que vamos a convivir es una de las obsesiones de la carrera en el Internet de las cosas. Prueba de ello es la competición de la NASA con varias empresas tecnológicas líderes estadounidenses para crear el sensor más pequeño del mundo. También la carrera se sigue en Europa: investigadores de la Universidad Tecnológica de Eindhoven han creado un sensor minúsculo —dos milímetros cuadrados— que no necesita batería: se alimenta del wifi gracias a la antena y el router que lleva incorporado.

Para que el Internet de las cosas tenga éxito, la conectividad juega también un papel esencial, especialmente en aquellas aplicaciones que requieren un análisis más complejo de datos provenientes de distintas fuentes. Una conectividad “ubicua” no está a disposición a día de hoy, especialmente en economías en vías desarrollo.

Como destaca la consultora, no debemos de olvidar el importante papel que juega el software, pues el auténtico valor de las aplicaciones del IoT viene de analizar datos de múltiples sensores y la correspondiente toma de decisiones basada en dichos datos. Y no hay que olvidar la interoperabilidad, de la que depende poder conseguir entre un 40% y un 60% del potencial total del IoT. Los fabricantes de equipos industriales ofrecen productos cada vez más “conectados”, con sensores y conectividad integrados; estos productos permiten a los fabricantes detectar anomalías e incluso llevar a cabo un mantenimiento predictivo. Algunos de estos fallos solo se pueden predecir cuando los datos provienen de múltiples sistemas, de ahí la importancia de la interoperabilidad.

La ausencia de un software común, formatos de datos estándares y protocolos de conectividad frena el avance del Internet de las cosas. También es necesario garantizar la privacidad, la protección de la propiedad intelectual y la cesión de la propiedad de los datos. Los políticos deben de jugar un papel clave, ayudando a abordar las preocupaciones existentes sobre la seguridad y la privacidad y fomentando el desarrollo de normas para promover la interoperabilidad. También tienen que garantizar la transparencia en la recogida y uso de datos para fomentar la confianza del consumidor. El diálogo con las empresas (proveedores de tecnología y usuarios), ciudadanos y expertos.

El coche conectado como nuevo paradigma

La llegada de los vehículos autónomos es uno de los grandes desafíos del Internet de las cosas, tanto desde el punto de vista técnico como del marco regulatorio. Estados Unidos acordó a comienzos de 2016 un fondo de 4.000 millones de dólares para impulsar el coche autónomo y que esté en las carreteras en menos de 10 años. La partida, que proviene del presupuesto federal de 2017, busca reducir los siniestros por fallos humanos —el 94% de las muertes en las carreteras están causadas por errores de los conductores, según las autoridades estadounidenses—, disminuir la contaminación y los atascos.

El Gobierno estadounidense quiere contar con la colaboración de las empresas tecnológicas y automovilísticas para diseñar un plan en el que se

van a estudiar tanto las carreteras como el marco regulatorio —que atañe tanto definir carreteras aptas para este tipo de coches, aparcamientos, normas de circulación y acuerdos con las compañías aseguradoras—. En un plazo de dos años, se prevé que cerca de 2.500 coches autónomos rueden por las calles norteamericanas. La seguridad es una de las obsesiones de los fabricantes e ingenieros. En 2015, los humanos evitaron 13 accidentes en tráfico real en los coches sin conductor de Google, según datos facilitados por la compañía. Apple, Google, Tesla, Ford, General Motors, Mercedes, Renault, Nissan, BMW son algunas de la treintena de compañías que están desarrollando la tecnología que revolucionará el mundo del transporte. Empresas que buscan coches 0,0: cero emisiones y cero accidentes. Para ello dotan a los vehículos de decenas de cámaras, radares y escáneres láseres que permiten identificar vehículos, objetos y personas para evitar cualquier accidente y garantiza una perfecta conducción autónoma. Algo que solo es posible gracias al Internet de las cosas.

GLOSARIO

Algoritmo: según la RAE es un conjunto ordenado y finito de operaciones que permite hallar la solución de un problema.

Almacenamiento escalable: es la capacidad de mejorar recursos progresivamente para ofrecer una mejor capacidad del servicio de almacenamiento. Requiere recursos adicionales en vez de realizar una modificación profunda o total del servicio o la tecnología de almacenamiento.

Anonimizar: “Expresar un dato relativo a entidades o personas eliminando la referencia a su identidad”. La necesidad de los cibernautas de ocultar información mientras navegan para evitar el uso de sus datos personales (como la dirección IP o la situación geográfica) con fines estadísticos, publicitarios o vandálicos ha causado la aparición del neologismo “anonimizar”, que se ha extendido después a otros ámbitos como el de la obligación de las empresas de trabajar con datos anonimizados para evitar conocer con detalle a las personas sobre las que se basan los análisis.

Base de datos: es un conjunto de datos pertenecientes a un mismo contexto y almacenados de forma sistemática para su posterior uso.

Científico de datos: profesionales con habilidades en matemáticas, estadísticas e ingeniería informática que son capaces de extraer el máximo valor de los datos de la organización, cerrando la brecha entre las necesidades del negocio y las TI.

Nube (Cloud computing): modelo de prestación de servicios de computación, información y aplicaciones a través de Internet donde la mayoría del software se ejecuta en la propia Red y también allí se almacenan los datos.

Datificación: el aumento y la conversión de nuestro mundo en datos medibles. La datificación del mundo es una realidad y es consecuencia directa del Big Data.

Dato: el dato es una representación simbólica (numérica, alfabética, algorítmica, etc.) de un atributo o variable cuantitativa o cualitativa. Los datos describen hechos empíricos, sucesos y entidades.

Datos abiertos (Open data): es una filosofía y práctica que persigue que determinados tipos de datos estén disponibles de forma libre para todo el mundo, sin restricciones de derechos de autor, de patentes o de otros mecanismos de control.

Datos estructurados: son aquellos datos que provienen de fuentes de información conocidas y con una forma y que, por tanto, son fáciles de medir y analizar a través de los sistemas tradicionales.

Datos no estructurados: son aquellos datos que provienen de la web, imágenes, vídeos, redes sociales... La variedad de su origen, además de la rapidez con la que se incrementa su volumen, son algunos de los factores que dificultan su análisis. Hay que realizar un trabajo previo de homogeneización y estandarización antes de que sean útiles.

Estadística: ciencia que utiliza conjuntos de datos numéricos para obtener, a partir de ellos, inferencias basadas en el cálculo de probabilidades. Estudio que reúne, clasifica y recuenta todos los hechos que tienen una determinada característica en común para poder llegar a conclusiones.

Gigabyte: un gigabyte es una unidad de almacenamiento de información cuyo símbolo es el GB, equivalente a 109.000 millones de bytes. Conocido coloquialmente como "Giga".

Hadoop: es un entorno de desarrollo y programación de código abierto para el procesamiento, el almacenamiento y el análisis de grandes volúmenes de datos de diversas fuentes.

Medidas de almacenamiento de datos:

1 bit es la unidad mínima de almacenamiento.

8 bits = 1 byte.

1024 bytes = 1 kilobyte.

1024 kilobytes = 1 megabyte.

1024 megabytes = 1 gigabyte.

1024 gigabytes = 1 terabyte.

1024 terabytes = 1 petabyte.

1024 petabyte = 1 exabyte.

1024 exabytes = 1 zettabyte.

1024 zettabyte = 1 yottaByte.

1024 yottabytes = 1 brontobyte.

1024 brontobytes = 1 geopbyte.

Metadatos: son aquellos datos que describen otros datos. Son como las fichas de una biblioteca que nos dan información sobre los libros que contiene.

Muestra: es un subconjunto de la población al que tenemos acceso y sobre el que podemos hacer las observaciones. Debería ser "representativo".

Muestra aleatoria: es una muestra representativa de la población elegida al azar. Ha de ser “confiable”.

Parámetro: es una cantidad numérica calculada para el total de la población (la media de salarios, por ejemplo).

Petabyte: es una unidad de almacenamiento de información que equivale a 1.000 terabytes (1.000.000.000.000.000 de bytes). Su símbolo es PB. Ahora también se está utilizando la unidad zetabytes (ZB), que equivale a 1021 bytes. Filmar toda la vida de una persona longeva (100 años) en alta definición a 50 fotogramas por segundo no llegaría a un petabyte.

Población: es el conjunto de elementos sobre el que estamos interesados en obtener conclusiones. También se denomina universo, que es el conjunto de elementos de referencia sobre el que se realizan las observaciones.

Publicidad programática: es aquella que se compra y se vende de forma automatizada. La publicidad programática ha permitido acceder al inventario publicitario de los medios digitales que no se vendía o que se vendía con unas tarifas muy bajas. Esto ha aumentado el inventario disponible y la posibilidad de acceder a gran velocidad y escalabilidad a las impresiones que se pueden comprar.

Quantified self: movimiento que incorpora las tecnologías para registrar todos los datos sobre aspectos de la vida diaria de una persona. En términos de inputs (alimentos que se consumen, la calidad del aire), estados (oxígeno en la sangre) y rendimiento (físico y mental).

R: programa de análisis estadístico y gráfico que es uno de los más utilizados en estadística y matemática. Es un proyecto de software libre.

Selftracking: “Autorastreo o autoanálisis de la actividad o datos de uno mismo. Seguimiento de los niveles de actividad, ciclos del sueño, glucosa, presión arterial, entre otros parámetros, gracias a aplicaciones y dispositivos tecnológicos.

Universo: población en estadística. Antes sobre el universo se seleccionaba una muestra representativa, ya que esta era abarcable. Hoy con las capacidades del Big Data algunas veces se estudia el universo completo.

Variable: es una característica observable que varía de unos individuos a otros de la población. La información que disponemos de cada individuo es resumida en variables. El dato sería un valor particular de una variable. Las variables pueden ser:

Cualitativas: no pueden asociarse de manera natural a un número (la religión, nacionalidad...).

- Nominales: no se pueden ordenar, por ejemplo, los colores.

- Ordinales: cuando pueden ordenarse, aunque no necesariamente en los

mismos intervalos, por ejemplo, "mucho", "poco", "nada".

Cuantitativas: los valores son numéricos.

- Discretas: los valores son enteros y están separados. El número de móviles que tiene un individuo sería una variable discreta.
- Continuas: entre dos valores hay infinitos valores posibles. La altura de los individuos sería una variable continua.

NOTAS

- 1 . Publicado en noviembre de 2014 (<http://www.nytimes.com/newsgraphics/2014/senate-model/comparisons.html>).
- 2 . Ebook BIG DATA del Centro de Innovación del BBVA.
- 3 . Dinámicas del Turismo en la Ciudad de Madrid. BBVA Centro de Innovación.
- 4 . “La española CartoDB seduce a los gurús de Silicon Valley”, El Confidencial (<http://goo.gl/6fMSJd>).
- 5 . “Informe Big Data y Salud”, Prodigioso Volcán y Planner Media, 2015.
- 6 . “Big data: The next frontier for innovation, competition, and productivity”.
- 7 . “Informe Big Data y Salud”, Prodigioso Volcán y Planner Media, 2015.

Índice

PRESENTACIÓN. BIG DATA Y EL INTERNET DE LAS COSAS

CAPÍTULO 1. QUÉ ES BIG DATA

CAPÍTULO 2. LA VISUALIZACIÓN DE DATOS

CAPÍTULO 3. ÁREAS E INDUSTRIAS MODIFICADAS

CAPÍTULO 4. CÓMO NOS AFECTA EN NUESTRA VIDA DIARIA

CAPÍTULO 5. LA HORA DEL INTERNET DE LAS COSAS

GLOSARIO

NOTAS

