BIG DATA: ¿SOLUCIÓN O PROBLEMA?

Estrella Pulido Cañabate
Universidad Autónoma de Madrid

RESUMEN

Los que hayan leído o escuchado algo sobre el Big Data puede que se estén debatiendo entre la esperanza y la preocupación porque les surjan preguntas como ¿ayudará el Big Data a curar enfermedades? ¿O dará lugar a nuevas desigualdades médicas? ¿Contribuirá el Big Data a prevenir suicidios? ¿O se usará para rastrear los movimientos de los manifestantes en las calles?

En las siguientes líneas tratamos de aclarar lo que significa el Big Data, de describir un mundo que está lleno de luces, y también de sombras, y presentar una panorámica, tanto de lo que el Big Data puede aportar en la construcción de una sociedad mejor, como de los problemas que puede generar.

1. CARACTERÍSTICAS BÁSICAS DEL BIG DATA

Vivimos rodeados de teléfonos inteligentes, electrodomésticos inteligentes, sistemas GPS, sensores, medidores, cámaras de circuito de televisión… Todos estos dispositivos electrónicos tienen la capacidad de ejecutar numerosas aplicaciones, de comunicarse entre ellos y de generar cantidades inimaginables de datos. Además, gracias a los avances en las telecomunicaciones, las personas y las cosas estamos casi permanentemente interconectadas y generamos también una gran cantidad de información de forma consciente o inconsciente.

Lo cierto es que la información disponible ha aumentado en los últimos años de forma exponencial.

a) ¿De dónde proceden todos estos datos?

Esta inmensa cantidad de datos procede de fuentes muy diversas. Una parte son datos que se recogen sobre nuestras llamadas telefónicas, transacciones bancarias, pagos con tarjeta o búsquedas en Google. Incluso sobre nuestros movimientos a través de las señales GPS procedentes de nuestros teléfonos móviles. Otros datos los generamos nosotros de forma voluntaria cuando publicamos entradas en blogs, subimos imágenes o vídeos a YouTube, o enviamos mensajes a través de redes sociales como Facebook o Twitter.

La ciencia también genera una gran cantidad de datos en campos como la genómica, la física de partículas, la biología, las ciencias medioambientales, la astronomía o la meteorología.

Otra cantidad importante de datos procede de sensores que monitorizan los objetos, capturan datos sobre ellos y comunican esa información a través de la red. Es lo que se conoce como el *Internet de las Cosas*. Se estima que en 2020 habrá en el mundo 26.000 millones de dispositivos conectados (Gartner, 2014) que generarán el 40% de la totalidad de los datos creados (Capgemini, 2014).

Nota: Este texto es un extracto de la Lección inaugural del curso 2014-2015 que con el mismo título la autora pronunció en la Universidad Autónoma de Madrid el 10 de Septiembre de 2014.

En el mundo de ese año 2020, no es difícil ni inverosímil imaginar una vida como la de Andrés, que en el tren de vuelta a casa está utilizando su teléfono inteligente para consultar su blog favorito de cocina y acaba de decidirse por una receta para la cena de sus dos hijos. Envía una petición a su frigorífico inteligente que, a su vez, confecciona una lista de la compra y la remite al supermercado para asegurarse de que los ingredientes esenciales no estarán agotados cuando Andrés llegue. Como sus hijos no son muy aficionados a ir de compras en el coche cuando salen de la guardería, Andrés envía una petición a su sistema de entretenimiento doméstico para que seleccione un episodio de la serie favorita de televisión de sus hijos y lo descargue en el sistema multimedia del coche. Si fuéramos más allá con el ejemplo, Andrés podría dar una orden al horno para que se precalentara, comprobar si se ha cargado la batería de su coche eléctrico, y quizá también averiguar, con la ayuda de los chips cosidos en la ropa de sus hijos, si tiene que recogerles en el patio o en la sala de música de la guardería (Cramer, 2012).

Esta situación está más cerca de ocurrir de lo que pensamos, porque los circuitos integrados cada vez son más baratos y pueden añadirse sensores a casi todo. Las fábricas, las ciudades y el transporte están transformándose en fábricas inteligentes, ciudades inteligentes y transporte inteligente.

Las *fábricas inteligentes* (Van Rijmenam, 2013a) integran sensores y software en la maquinaria industrial, consiguiendo así optimizar sus procesos de producción, ahorrar energía y crear mejores productos.

Un ejemplo de *ciudad inteligente* o "smartcity" es Songdo (Lobo, 2014), una ciudad en Corea del Sur que empezó a construirse en el año 2000 y en la que viven en la actualidad unas 40.000 personas. La infraestructura de esta ciudad contiene sensores que monitorizan y regulan desde la temperatura hasta el consumo de energía y el tráfico. Casi cada dispositivo, edificio o carretera en Songdo está equipado con sensores inalámbricos o microchips, lo que permite, por ejemplo, que el número de farolas encendidas se ajuste automáticamente al número de personas que haya en ese momento en la calle. Además, todas las casas están equipadas con sensores controlados mediante domótica, y el tráfico se regula también mediante sensores instalados en los coches que envían datos de geo-localización y permiten monitorizar las áreas congestionadas de la ciudad.

Un ejemplo que permite ilustrar lo que puede significar el *transporte inteligente* es el de la compañía US Xpress (Van Rijmenam, 2014a) que ha instalado, en cada uno de sus camiones, casi mil sensores que generan datos sobre el consumo de carburante, la velocidad, el estado de los neumáticos o el uso de los frenos. Combinando los datos recibidos de estos sensores con información sobre las condiciones de las carreteras, el estado del tráfico, los datos meteorológicos o la localización de estaciones de servicio, es posible optimizar rutas de envío, reducir el consumo de combustible y la contaminación, aumentar la seguridad en las carreteras y llevar un control de inventario mucho más preciso. Y no sólo los camiones. Los trenes, los aviones, los barcos… Cualquier medio de transporte puede monitorizarse en tiempo real utilizando sensores y navegación por satélite.

En resumen, y simplificando mucho, podemos decir que las principales fuentes de los grandes datos son (1) los datos procedentes de transacciones como compras por Internet o presenciales, transacciones bancarias o búsquedas online; (2) los datos procedentes de cualquier tipo de máquina, ya sea maquinaria industrial, un equipo de secuenciación de genomas, el acelerador de partículas del CERN, un tomógrafo con el que se realizan TACs o el motor de un avión en vuelo; y (3) los datos compartidos por los usuarios a través de redes sociales (Arcplan, 2012).

b) Entonces, ¿por qué ha surgido el Big Data?

En primer lugar, porque el almacenamiento se ha abaratado enormemente. Hace dos décadas hacía falta una máquina del tamaño de un frigorífico y con un peso de 800 kilos para almacenar un

gigabyte de datos. Hoy en día, la mayoría de nosotros llevamos, al menos, 16 gigabytes de datos en nuestro teléfono inteligente.

El precio de los dispositivos de almacenamiento también ha bajado. En esas dos décadas de las que hablamos, el coste de almacenar un gigabyte ha pasado de más de 700 euros a cuatro o cinco céntimos.

Además, la velocidad de procesamiento ha aumentado enormemente con la aparición, a principios de los años 2000, de la computación paralela masiva. En vez de procesar tareas de una en una, ahora los ordenadores pueden procesar una gran cantidad de tareas en paralelo, es decir, todas a la vez. Así es como Google, Facebook o Amazon han sido capaces de construir sus servicios.

No sólo el hardware ha aumentado su velocidad. También ha sido decisiva la aparición de software inteligente que permite sacar partido de la capacidad de procesamiento paralelo, analizando grandes cantidades de datos en muy poco tiempo. Un detalle importante es que este software inteligente puede analizar, no sólo datos almacenados y estáticos, sino también datos volátiles que se analizan en tiempo real a la vez que se producen.

Además de todo esto, han aparecido nuevos tipos de datos que se generan de forma masiva, como correos electrónicos, mensajes de Facebook, tuits o vídeos de YouTube.

La principal diferencia entre estos nuevos tipos de datos y los que se han venido recogiendo hasta ahora es que estos últimos son datos estructurados, es decir, bien organizados, como los que pueden aparecer en una hoja de cálculo o en una base de datos, mientras que los primeros carecen de una estructura identificable. El análisis de estos datos no estructurados es mucho más difícil, pero a la vez también mucho más valioso.

c) ¿Y más o menos de cuántos datos estamos hablando?

Para poder estimar lo grande que es el Big Data, empezaremos por el megabyte, que equivale a 1 millón de bytes y nos permitiría almacenar un libro de unas 500 páginas que tuviera sólo texto (Grant, 2012).

En un terabyte, equivalente a un millón de megabytes, podríamos almacenar 2.767 copias de la Enciclopedia Británica, ó 16.667 horas de música, ó 1.333 horas de video.

Se necesitarían casi 5 exabytes, ó 5 millones de terabytes, para almacenar las secuencias del genoma humano de todas las personas del mundo.

Pues bien, la compañía americana IDC estima que en 2011 se crearon 1,8 zettabytes de información, es decir, 1.800 exabytes, una cantidad de datos con la que podrían llenarse 57.500 millones de iPads de 32 gigabytes. Con este número de iPads podría construirse una pared dos veces más alta que la muralla china.

En 2013 se generaron 4,4 zettabytes de datos, de los cuales el 49% eran datos no estructurados o semiestructurados. IDC pronostica que en 2020 generaremos 44 zettabytes, es decir, 10 veces más (IDC, 2014; TCS, 2013).

Toda esta cantidad de información se genera a una enorme velocidad, tan alta que el 90% de los datos que existían en el mundo en 2012 se había creado en los dos años anteriores. Cada minuto, por ejemplo, se envían 150 millones de mensajes de correo electrónico, se realizan más de 2 millones de búsquedas en Google y se visualizan casi tres millones de vídeos en YouTube (Excela, 2016). En media hora, el motor de un avión que vuela de Londres a Nueva York genera 10 terabytes de datos

(MacKinnon, 2013). En 2016 se publican al día 500 millones de tuits y se comparten casi 6000 millones de mensajes "me-gusta" en Facebook (Gwava, 2016).

2. CONTEXTO Y REPERCUSIONES DEL BIG DATA

La empresa consultora Gartner define el Big Data como "aquellos recursos de información caracterizados por su alto volumen, velocidad o variedad, que requieren formas de procesamiento innovadoras y eficientes para la mejora del conocimiento y la toma de decisiones" [1].

De acuerdo con esta definición, las características que definen el Big Data pueden resumirse en lo que se conoce como las tres Vs: *Volumen, Velocidad y Variedad* (Laney, 2011).

A estas tres Vs que forman parte de la definición intrínseca del Big Data, los expertos añaden una cuarta V que tiene que ver con la *Veracidad,* ya que es fundamental que los datos en los que se basan los análisis sean correctos y no contengan sesgos o ruido que puedan distorsionar las conclusiones que se extraigan de ellos (Normandeau, 2013).

La definición de Big Data propuesta por Gartner tiene una segunda parte, también importante, que tiene que ver con la capacidad para analizar los datos y extraer de ellos información relevante. Andreas Weigend, antiguo científico de Amazon y profesor en varias universidades americanas, afirma que los datos son el nuevo petróleo, no sólo en el sentido económico, sino también porque, al igual que el petróleo, es necesario refinarlos y depurarlos para que aporten valor (Brustein, 2014a).

Adoptando la definición propuesta por Gartner, en el resto de este artículo usaremos el término Big Data para referirnos de manera indisoluble a los grandes conjuntos de datos y a los resultados que puedan derivarse del análisis de los mismos.

El Big Data está causando una revolución en el mundo empresarial, modificando de manera sustancial los negocios existentes y creando otros completamente nuevos.

La Comisión Europea estima que el sector del Big Data crece a un ritmo del 40% anual a nivel mundial. Aunque en España el crecimiento es un poco más bajo y se sitúa en torno al 30%, supone siete veces más que el ratio de crecimiento conjunto de todas las Tecnologías de la Información y la Comunicación (ISDI, 2016).

Según un informe del World Economic Forum, entre los años 2015 y 2020 se crearán 2,1 millones de puestos de trabajo relacionados con las TIC y una gran parte de ellos tendrán que ver con el análisis de datos (WEF, 2016).

El problema es que no habrá suficientes especialistas para cubrir esta demanda. Se estima que en el 2018, sólo en Estados Unidos puede haber una carencia de entre 140.000 y 190.000 profesionales con conocimientos profundos sobre técnicas de análisis (McKinsey, 2011) y capaces de presentar, muchas veces de forma visual, los resultados de los análisis a los ejecutivos responsables de tomar las decisiones.

Teniendo en cuenta todos estos datos, está claro que la figura del científico de datos será indispensable para las empresas, por lo que es fundamental que las Universidades se planteen como una prioridad la formación de este tipo de perfil profesional.

3. APLICACIONES DEL BIG DATA

Los más optimistas opinan que en la que parece inminente era del Big Data lograremos una mejora sustancial de nuestra capacidad para realizar diagnósticos y pronósticos fiables en numerosas áreas de la vida social.

Esta nueva capacidad de análisis y predicción puede ser aprovechada por las empresas con fines puramente lucrativos utilizando distintas técnicas, como la publicidad personalizada o el análisis de sentimientos.

a) La publicidad personalizada

Localizar patrones y tendencias permite a las empresas de comercio online adecuar los productos y servicios a los clientes, anticipar la demanda o mejorar las ventas a través de incentivos, como descuentos, envíos gratuitos o facilidades de pago.

b) Análisis de sentimientos

Analizar las enormes cantidades de datos que se mueven en las redes sociales también puede ser de gran utilidad para las empresas, sobre todo teniendo en cuenta el número de usuarios cada vez mayor que interacciona en ellas. Facebook, por ejemplo, tiene 1550 millones de usuarios, Twitter 320 millones, y Linkedin 100 (The Social Media Hat, 2016).

Gracias a estos análisis, las empresas pueden entender lo que el público opina sobre distintos productos, servicios, anuncios publicitarios o series de televisión, y saber quiénes son los usuarios con más influencia en las redes. De esta forma pueden mejorar sus productos o incluso desarrollar otros nuevos, de acuerdo con el deseo de los consumidores. Es lo que se conoce como el análisis de sentimientos o minería de opiniones (Van Rijmenam, 2013b).

c) El Big Data y el gobierno

Como hemos visto, los datos recogidos sobre los clientes pueden ser utilizados por las empresas para mejorar la experiencia del consumidor y, en último término, para incrementar las ganancias. Los Gobiernos también pueden usar estas herramientas para el espionaje masivo (Naughton, 2014) o para predecir resultados electorales como el sitio web FiveThirtyEight[2] que en 2008 logró predecir correctamente el vencedor de 49 de los 50 Estados norteamericanos en las elecciones presidenciales que ganó Barack Obama.

d) El Big Data y el desarrollo

La revolución de los datos no se limita al mundo industrializado y está ocurriendo también en los países en vías de desarrollo. En países de África y del sur de Asia, el 50% de los adultos son propietarios de un teléfono móvil, y entre un 10 y un 20% más, aunque no son propietarios, tienen acceso al móvil de amigos, familiares o vecinos (Cartesian, 2014).

e) El Big Data y la salud

Un campo importante de aplicación del Big Data es el de la salud y la atención sanitaria. Hoy en día se almacenan todo tipo de datos sobre pacientes, enfermedades, tratamientos, medicación y resultados. Nuestro cuerpo se ha convertido en una fuente más de datos. Radiografías, mamografías, TACs, resonancias magnéticas, historiales médicos, … El análisis de todos estos datos juega un papel fundamental en lo que, según los expertos, será la medicina del futuro o medicina "4P" (Hood y Galas, 2008): una medicina personalizada, predictiva, preventiva y participativa.

4. EL BIG DATA EN LA EDUCACIÓN Y LA INVESTIGACIÓN

a) Educación

Dicen que lo que no se puede medir no se puede mejorar, y si hay un área que puede beneficiarse del análisis de los grandes conjuntos de datos es el de la educación. Al igual que las empresas y la Administración, la Universidad viene recogiendo datos sobre los estudiantes desde hace mucho tiempo, la mayoría de ellos relacionados con su rendimiento escolar. Se almacenan las calificaciones de los trabajos que entregan, sus resultados en los exámenes, el número de convocatorias que necesitan para superar una asignatura o cuánto tiempo tardan en finalizar sus estudios.

La novedad es que las universidades cada vez ofrecen más materiales educativos a través de plataformas de e-learning como Moodle, y el aprendizaje online a través de los MOOCs o Massive Open Online Courses cada vez está más extendido. Estos cursos online ofrecen numerosas ventajas, porque permiten a los estudiantes acceder a la enseñanza en el momento y lugar preferidos y les proporcionan un feedback inmediato y constante sobre su rendimiento, que influye muy positivamente en la motivación.

Plataformas de enseñanza online, como edX, Coursera o Udacity, hacen posible que 100.000 estudiantes repartidos por todo el planeta puedan asistir a una clase de un profesor de la Universidad de Harvard. Pero el gran impacto que pueden tener los MOOCs en la educación no tiene que ver sólo con esto, sino también con los conjuntos enormes de datos sobre el comportamiento de los estudiantes que recogen estas herramientas, cuyo análisis puede ayudar a mejorar el rendimiento académico de los estudiantes, a disminuir las tasas de abandono y a personalizar la educación adaptando los contenidos, las tareas y el feedback a las necesidades de cada estudiante (Guthrie, 2013; Nielson, 2013).

Analizar toda la variedad de datos que se recogen en el proceso de formación de los estudiantes es lo que se conoce como Analítica del Aprendizaje (Ferguson, 2014). Estos datos pueden tener que ver con el rendimiento académico, pero también con la interacción de los estudiantes con el campus universitario en general, como el número de accesos a la plataforma docente virtual, el uso de la biblioteca, la participación en actividades deportivas, la asistencia a reuniones con el tutor, o incluso el uso que hacen del aparcamiento.

Este análisis puede servir a los estudiantes individuales para reflexionar sobre sus resultados y modelos de comportamiento en relación con otros compañeros; a los profesores para identificar qué estudiantes tienen más riesgo de abandonar o fracasar y necesitan más apoyo y atención; a los encargados de la calidad docente para introducir mejoras en las asignaturas o desarrollar nuevos planes de estudio; y a los administradores para tomar decisiones sobre temas relacionados con la promoción de los estudios, la distribución de los recursos o el proceso de admisión.

Como todo, el uso del Big Data en el entorno universitario también puede ser controvertido (Renton, 2014). Una de las principales críticas que se plantea con respecto a la analítica del aprendizaje es que puede predecir los resultados académicos que obtendrá un estudiante, pero no sirve para identificar las causas de su éxito o fracaso.

Otro conjunto de críticas tiene que ver con el hecho de que el análisis de datos identifique perfiles de estudiantes que reúnen determinadas características. La primera de ellas es que el conocimiento de estos perfiles por parte de los profesores puede sesgar sus expectativas sobre los estudiantes. Es lo que se conoce como el efecto Pigmalión, o la teoría de la profecía auto cumplida, que puede definirse como una predicción que, una vez hecha, es en sí misma la causa de que se haga realidad (Díaz, 2014).

Por otro lado, si estos perfiles se utilizan pensando sólo en el beneficio económico, identificar el perfil de los estudiantes más propensos al abandono puede provocar que no se les admita, o que se les asignen menos recursos porque resultan menos rentables.

b) El Big Data y la investigación

La computación intensiva de datos juega un papel fundamental en el descubrimiento científico. Hallazgos como el Bosón de Higgs, aunque pendiente de la confirmación definitiva, o la secuenciación del genoma humano no habrían sido posibles sin el Big Data.

Los científicos están acostumbrados a los grandes conjuntos de datos que puede generar, por ejemplo, un secuenciador de ADN, un acelerador de partículas, un telescopio o las estaciones y satélites meteorológicos.

El Big Data permite entender mejor el flujo de tráfico en las autopistas, el uso de energía doméstica, las condiciones en el interior de los volcanes, los huracanes, las enfermedades o el universo (CRA, 2011).

También se puede utilizar para estudiar las variedades dialectales del español como han hecho dos investigadores del Instituto de Física Interdisciplinar y Sistemas Complejos a partir de 50 millones de tuits geo localizados escritos en español recogidos durante dos años (Gonçalves y Sánchez, 2014).

En economía, el Big Data puede utilizarse, por ejemplo, para estimar el índice de inflación. Es lo que hace el proyecto BPP (Billion Prices Project) del MIT a partir de los precios de más de 50.000 productos recogidos diariamente de cientos de tiendas online (Einav y Levin, 2013).

Otra área en la que el Big Data está teniendo un gran impacto es la ciencia social computacional, una disciplina a medio camino entre la psicología y la sociología de la que ya se hablaba en 2009 (Lazer et al., 2009) y que analiza los datos de nuestras interacciones en Internet para revelar patrones de comportamiento individual y grupal.

No resultaría novedoso decir que el análisis de grandes cantidades de datos se utiliza en disciplinas como la física de partículas, las ciencias medioambientales, la bioinformática o la microbiología. Pero quizá puedan resultar menos familiares disciplinas como la cliometría, que aplica la econometría a la historia; la estilometría, que es el estudio del estilo de escritura de un autor; o la culturómica, que describe las investigaciones cuantitativas en ciencias sociales y humanidades. En todas estas disciplinas el Big Data tiene mucho que decir.

Como vemos, el Big Data no sólo tiene aplicación en la industria o en la investigación científica, sino también en campos más relacionados con las humanidades, como la sociología, las ciencias políticas o la economía.

5. BIG DATA: GOBIERNO ABIERTO, PRIVACIDAD Y TRANSPARENCIA

a) Gobierno abierto

No podemos hablar de los grandes datos sin hablar también de los datos abiertos, que la red para el Conocimiento Abierto define como “cualquier contenido, información o datos que las personas pueden usar, reusar y redistribuir sin ninguna restricción legal, tecnológica o social”[3].

Todas las definiciones de datos abiertos incluyen, como características básicas, que deben estar disponibles de forma gratuita o a un coste mínimo en un formato que facilite su uso, y deben poder ser utilizados sin discriminación de personas, grupos o aplicaciones.

Podemos citar muchos ejemplos de datos que son o deberían ser abiertos. Por ejemplo, los datos sobre títulos y autores de obras culturales que se guardan en galerías de arte, bibliotecas, archivos y museos; los relacionados con la investigación científica, con los presupuestos del Estado o con los mercados financieros; los datos estadísticos como el censo o los indicadores socioeconómicos; datos meteorológicos, climatológicos o medioambientales relativos a la presencia y el nivel de contaminantes, o sobre la calidad de ríos y playas...

Muchos de estos datos, como los procedentes de la investigación científica, las bases de datos meteorológicas o el censo, pueden ser clasificados dentro del conjunto de los grandes datos y, al igual que éstos, su análisis puede beneficiar a las empresas, las organizaciones y el público en general, porque les permite crear nuevos negocios, detectar patrones y tendencias y tomar decisiones informadas (Gurin, 2014). Por el contrario, los datos que no están abiertos al público sólo benefician a aquellos que los controlan.

Por ello, el gobierno, que tiene la capacidad y los fondos para recoger grandes cantidades de datos, tiene también la responsabilidad de transformar en datos abiertos aquellos subconjuntos que puedan ser más beneficiosos para los ciudadanos y que les permitan participar en las decisiones que les afectan de forma continua y no sólo depositando su voto en una urna cada cuatro años.

Así queda reflejado en la Ley de Transparencia, Acceso a la Información Pública y Buen Gobierno aprobada en diciembre de 2013, cuyo preámbulo dice textualmente que "La transparencia, el acceso a la información pública y las normas de buen gobierno deben ser los ejes fundamentales de toda acción política. Sólo cuando la acción de los responsables públicos se somete a escrutinio, cuando los ciudadanos pueden conocer cómo se toman las decisiones que les afectan, cómo se manejan los fondos públicos o bajo qué criterios actúan nuestras instituciones podremos hablar del inicio de un proceso en el que los poderes públicos comienzan a responder a una sociedad que es crítica, exigente y que demanda participación" (BOE, 2013).

Este párrafo describe perfectamente el concepto de gobierno abierto, cuyos pilares fundamentales son la transparencia, la participación y la colaboración, (Preciado, 2012) y que Javier Llinares, uno de los autores más referenciados en este tema, define como "aquel que entabla una constante conversación con los ciudadanos con el fin de oír lo que ellos dicen y solicitan, que toma decisiones basadas en sus necesidades y preferencias, que facilita la colaboración de los ciudadanos y funcionarios en el desarrollo de los servicios que presta y que comunica todo lo que decide y hace de forma abierta y transparente" (Llinares, 2007).

b) La privacidad

Una primera cuestión ética que debe abordarse al tratar con el Big Data es la de quién es el dueño de los datos que se analizan. Los datos que se recogen sobre nuestras llamadas telefónicas, por ejemplo, ¿pertenecen a la persona que llama, a la compañía telefónica o a cualquier agencia de espionaje del gobierno que pueda acceder a ellos? Cuando nuestros coches sean monitorizados mediante sensores, todos los datos que generen, ¿pertenecerán a los conductores, a los propietarios, o a los fabricantes del vehículo? (TCS, 2013)

Tradicionalmente las organizaciones han venido utilizando diversas estrategias para asegurar la privacidad, entre ellas, anonimizar los datos de las identidades reales, creando así conjuntos de datos anónimos.

Normativas vigentes relacionadas con la privacidad, como la Ley de Protección de Datos de Carácter Personal o LOPD, recogen distintos principios relacionados con la privacidad. Uno de ellos es el principio de minimización, que establece que sólo deben guardarse aquellos datos personales que

sean necesarios para conseguir objetivos legítimos y especificados, y que éstos deben ser destruidos tan pronto como no sean relevantes para la consecución de dichos objetivos (Tene y Polonetsky, 2012).

Otro principio importante para proteger la privacidad, también recogido en la LOPD, es el de "información y consentimiento" según el cual debe explicarse a los individuos qué uso se dará a la información que se recoja sobre ellos y éstos, a su vez, deben poder decidir si quieren o no que se recojan sus datos (Velásquez, 2014).

Llevar a la práctica todos estos principios, en lo que al Big Data se refiere, no parece tarea fácil. Por un lado, existen técnicas para desanonimimizar o restablecer la identidad de los datos, como la que utilizaron a principios de 2013 varios expertos en seguridad informática para identificar los apellidos correspondientes a 50 genomas humanos, supuestamente anónimos (Gymrek et al., 2013).

Por otro, gran parte del valor de los datos suele estar en usos secundarios distintos de aquellos para los que se recogieron inicialmente, por lo que existe el riesgo de que los datos personales y privados se recojan y analicen con fines que los sujetos de los datos ni conocen ni aprueban. Esto hace que el mecanismo de información y consentimiento pueda no ser suficiente.

Además, los datos personales reúnen unas características que hacen especialmente difícil la tarea de establecer normativas (Rose y Kalapesi, 2012). En primer lugar, la naturaleza digital de los datos personales implica que pueden copiarse infinitamente y distribuirse globalmente, eliminando, por tanto, muchas de las fronteras comerciales que existen para los bienes materiales. Una segunda característica es que los datos no se agotan cuando se usan, se pueden reutilizar y aumentan su valor cuando se conectan con otros datos. Por último, y a diferencia de otros bienes materiales, los datos personales están íntimamente ligados a la historia e identidad del individuo.

A pesar de estas dificultades, es necesario establecer normas que (1) protejan los datos personales frente a las infracciones y abusos intencionados y no intencionados; (2) establezcan derechos, responsabilidades y permisos para que se equilibren los intereses de todos los participantes en el intercambio de datos; y (3) definan los requisitos que una empresa debe cumplir para que pueda ser considerada responsable con respecto a la protección, la seguridad y el uso de datos personales.

La privacidad no debe preocupar sólo a los gobiernos y a las empresas, sino también a los usuarios, que podemos protegerla utilizando, por ejemplo, la red TOR[4], donde los mensajes intercambiados entre los usuarios no revelan su identidad porque viajan desde el origen al destino a través de ordenadores seleccionados al azar.

Existen también las credenciales anónimas que permiten a un usuario demostrar que posee autorización para acceder a un sistema sin tener que revelar ninguna otra información sobre sí mismo, o protocolos como el de la recuperación privada de información (PIR)[5], que permite a un usuario obtener información de una base de datos sin que se sepa qué información concreta se ha consultado.

c) La transparencia

Además de proteger la privacidad de los datos, es necesario que los individuos tengan acceso a los datos que se recogen sobre ellos. Existen varias iniciativas en este sentido. Una de ellas es el proyecto *midata* del gobierno de Reino Unido (Cable, 2014) cuyos objetivos son, por un lado, conseguir que las empresas proporcionen a los consumidores acceso electrónico y seguro a los datos personales que han recogido sobre ellos y, por otro, animar a las empresas a desarrollar aplicaciones que ayuden a los consumidores a usar sus datos de forma efectiva.

Proporcionando acceso a estos datos, se consigue un mayor empoderamiento de los consumidores, porque pueden entender mejor su propio comportamiento, realizar elecciones más informadas de productos y servicios, y gestionar sus vidas de forma más eficiente. Tener acceso, por ejemplo, a los datos que una compañía de telefonía móvil tiene sobre nuestros usos y consumos puede ayudarnos a elegir una tarifa mejor.

Cabe señalar finalmente que en Estados Unidos han surgido iniciativas similares al proyecto midata, como el Blue Button[6], un símbolo que aparece en determinadas páginas web y permite a los pacientes acceder a sus historiales médicos. De forma equivalente, existen las iniciativas Green Button para datos de consumo energético, y Red Button para datos académicos.

Está claro que el Big Data puede transformar los negocios, el gobierno y la sociedad y puede ayudarnos a mejorar el mundo o, por el contrario, a promover la discriminación, la invisibilidad o el control de los ciudadanos por parte de los gobiernos.

Recordarán ustedes que al comienzo de este artículo planteábamos la disyuntiva de si las soluciones que aporta el Big Data son más numerosas que los problemas que crea. Cómo se resuelva esta disyuntiva dependerá, en gran parte, de que los Gobiernos, las empresas y los ciudadanos apostemos porque las transformaciones que el Big Data traerá consigo indudablemente conduzcan a un mundo más justo y más democrático.

Un mundo como el que describe la red para el Conocimiento Abierto, en el que "el conocimiento suponga poder para todos y no sólo para unos pocos, en el que los datos nos permitan elegir de forma informada cómo vivir y en el que la información y el conocimiento sean accesibles y transparentes para todos".

BIBLIOGRAFÍA

Arcplan (2012): *Big Data FAQs. A primer*. Business Intelligence Blog. 23/03/2012. www.arcplan.com/en/blog/2012/03./big-data-faqs-a-primer/

BOE (2013): *Ley 19/2013, de 9 de diciembre, de transparencia, acceso a la información pública y buen gobierno*. http://boe.es/buscar/pdf/2013/BOE-A-2013-12887-consolidado.pdf

Brannock, J. (2014): *I knew you before I met you: How social media has changed the way we communicate*. The Faculty Voice. 18/03/2014. http://imerrill.umd.edu/facultyvoice1/?p=3171

Brustein, J. (2014a): *Consultant Andreas Weigend on Big Data refineries*. Bloomberg Business Week. 06/03/2014. http://www.businessweek.com/articles/2014-03-06/consultant-andreas-weigend-on-big-data-refineries

Cable, V. (2014): *Providing better information and protection for consumers*. Department for Business, Innovation & Skills. Gov.uk. https://www.gov.uk/government/policies/providing-better-information-and-protection-for-consumers/supporting-pages/personal-data

Capgemini (2014): *The Internet of Things: Are organizations ready for a multi-trillion dollar prize?* Capgemini Consulting. http://www.capgemini-consulting.com/resource-file-access/resource/pdf/the-internet-of-things.pdf

Cartesian (2014): *Using mobile data for development*. Bill and Melinda Gates Foundation. http://www.cartesian.com/wp_content/upload/Using-Mobile-Data-for-Development.pdf

CRA (2011): *Advancing discovery in science and engineering. The role of basic computing research.* Computing Research Association. http://www.cra.org/ccc/files/docs/Natl_Priorities/web_data_spring.pdf

Cramer, I. (2012): *Big Data and the Internet of Things*. Business Technology, 9. https://www.bosch-si.com/media/bosch_software_innovations/documents/publication/english_1/2012/2012-07-BigData_IndustrialIT_byIMCramer_published_on_bosch-sicom.pdf

Davenport, T.H.; Dyché, J. (2013) : *Big Data in big companies*. International Institute for Analytics. http://www.sas.com/resources/asset/Big-Data-in-Big-Companies.pdf

Díaz, J. (2014): *Analíticas de aprendizaje: Consideraciones éticas*. Aprendizaje y conocimiento: el ecosistema digital y su impacto en la formación del siglo XXI. 03/02/2014. http://javierdisan.com/tag/big-data/

Einav, L.; Levin, J. (2013): *The data revolution and economic analysis*. Stanford University. http://web.stanford.edu/~jdlevin/Papers/BigData.pdf

Excela (2016). *What happens in an internet minute?* http://www.excelacom.com/resources/blog/2016-update-what-happens-in-one-internet-minute

Ferguson, R. (2014): *Learning analytics don't just measure students' progress – they can shape it*. The Guardian. 26/04/2014. http://www.theguardian.com/education/2014/mar/26/learning-analytics-student-progress

Gartner (2012): *Gartner says EMEA IT spending will grow 1.4 percent in 2013 after declining in 2012*. Nota de prensa. 05/11/2012. http://www.gartner.com/newsroom/id/2225616

Gartner (2014): *Gartner says the Internet of Things will transform the data center*. Gartner News Room. http://www.gartner.com/newsroom/id/2684616

Gonçalves, B.; Sánchez, D. (2014): *Crowdsourcing Dialect Characterization through Twitter*. (Institute for Cross-Disciplinary Physics and Complex Systems. http://ifisc.uib-csic.es/publications/publication-detail.php?indice=2550

Grant, E. (2012): *The promise of big data*. Harvard School of Public Health. HSPH News. http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/

Gurin, J. (2014): *Big data and open data: what's what and why does it matter?* The Guardian. 15/04/2014. http://www.theguardian.com/public-leaders-network/2014/apr/15/big-data-open-data-transform-government

Guthrie, D. (2013): *The coming Big Data education revolution*. US News. 15/08/2013. http://www.usnews.com/opinion/articles/2013/08/15/why-big-data-not-moocs-will-revolutionize-education

Gwava (2016): *How much data is created on the internet each day?* https://www.gwava.com/blog/internet-data-created-daily

Gymrek, M.; McGuire, A.L.; Golan, D.; Halperin, E.; Erlich, Y. (2013): *Identifying personal genomes by surname inference*. Science, 339 (6117), 321-324. http://www.sciencemag.org/content/339/6117/321

Hood, L.E.; Galas, D.J. (2008): *P4 Medicine: Personalized, predictive, preventive, participatory. A Change of view that changes everything*. Computing Research Association. http://www.cra.org/ccc/files/docs/init/P4_Medicine.pdf

IDC (2014): *The digital universe of opportunities: Rich data and the increasing value of the Internet of Things*. http://www.emc.com/leadership/digital-universe/2014iview/index.htm

ISDI (2016): *El Big Data registra un crecimiento anual del 30% en España* https://isdionline.com/es/blog/big-data-registra-crecimiento-anual-30-en-espana

Laney, D. (2001): 3D *data management: Controlling data volume, velocity and variety*. META Group. 06/02/2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lazer, D.; Pentland, A.; Adamic, L.; Aral, S.; Barabási, A.; Brewer, D.; Christakis, N.; Contractor, N.; Fowler, J.; Gutmann, M.; Jebara, T.; King, G.; Macy, M.; Roy, D.; Van Alstyne, M. (2009): *Computational social science*. Science, 323, 721-723. http://gking.harvard.edu/files/LazPenAda09.pdf

Lobo, R. (2014): *Could Songdo be the world's smartest city?* World Finance. 21/01/2014. http://www.worldfinance.com/inward-investment/could-songdo-be-the-worlds-smartest-city

Llinares, J. (2007): *Open government: La idea*. Javier Llinares, open governance, administración pública y otros temas. 29/12/2007. http://www.javierllinares.es/?p=476

MacKinnon, A. (2013): *What's so big about big data?* Online Educa Berlin News Portal. 25/09/2013. http://www.online-educa.com/OEB_Newsportal/whats-so-big-about-big-data/

McKinsey (2011): *Big Data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Naughton, J. (2014): *We're all being mined for data – but who are the real winners?* The Guardian. 08/06/2014. http://www.theguardian.com/technology/2014/jun/08/big-data-mined-real-winners-nsa-gchq-surveillance

Nielson, B. (2013): *MOOC analytics: What corporate training can learn from Big Data*. Your Training Edge. 24/06/2013. http://www.yourtrainingedge.com/mooc-analytics-what-corporate-training-can-learn-from-big-data/

Normandeau, K. (2013): *Beyond volume, variety and velocity is the issue of big data veracity*. insideBIGDATA. 12/09/2013. http://inside-bigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

Noyes, D. (2014): *The top 20 valuable Facebook statistics*. Zephoria Internet Solutions. 13/06/2014. http://zephoria.com/social-media/top-15-valuable-facebook-statistics/

Preciado, L. (2012): *Gobierno abierto y transparencia de la mano del "big data"*. IV Jornadas ASTICNET. ASTIC Boletic, 62, 44-45. http://www.astic.es/sites/default/files/articulosboletic/mono9_laura_preciado.pdf

Renton, S. (2014): *Snooping on students' digital footprints won't improve their experiences*. The Guardian. 26/03/2014. http://www.theguardian.com/education/2014/mar/26/students-digital-footprints-experience

Rose, J.; Kalapesi, C. (2012): *Rethinking personal data: Strengthening trust*. BCG Perspectives. 16/05/2012. https://www.bcgperspectives.com/content/articles/digital_economy_technology_software_rethinking_personal_data_strengthening_trust/

TCS (2013): *The emerging big returns on Big Data. A TCS 2013 global trend study*. http://www.tcs.com/SiteCollectionDocuments/Trends_Study/TCS-Big-Data-Global-Trend-Study-2013.pdf

Tene, O.; Polonetsky, J. (2012): *Privacy in the age of Big Data. A time for big decisions*. Stanford Law Review. 02/02/2012. http://www.stanfordlawreview.org/online/privacy-paradox/big-data

The Social Media Hat (2016): *Social media active users*. Infografía. http://www.thesocialmediahat.com/active-users

Van Rijmenam, M. (2013a): *The industrial Internet will bring a revolution to the manufacturing industry*. Bigdata-startups. 14/10/2013. http://www.bigdata-startups.com/industrial-internet-bring-revolution-manufacturing-industry/

Van Rijmenam, M. (2013b): *Why sentiment analytics should be a no-brainer for organisations*. Bigdata-startups. 01/10/2013. http://www.bigdata-startups.com/sentiment-analytics-no-brainer-organisations/

Van Rijmenam, M. (2014a): *Trucking company US Xpress drives efficiency with Big Data*. Bigdata-startups. http://www.bigdata-startups.com/BigData-startup/trucking-company-xpress-drives-efficiency-big-data/

Velásquez, H. (2014): *Big Data en el "Universo Compliance"*. Diariojuridico.com. 12/03/2014. http://www.diariojuridico.com/big-data-en-el-universo-compliance/

World Economic Forum. (2016): *The future of jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution*. http://www3.weforum.org/docs/WEF_Future_of_Jobs.pdf

[1] http://www.gartner.com/it-glossary/big-data/
[2] http://fivethirtyeight.com/
[3] https://okfn.org/opendata/
[4] https://www.torproject.org/
[5] http://crypto.stanford.edu/pir-library/
[6] http://bluebuttondata.org/