XIX Edición del Premio Protección de Datos Personales de Investigación
de la Agencia Española de Protección de Datos

ACCÉSIT 2015

Big data, privacidad y protección de datos

Elena Gil

AGENCIA ESPAÑOLA DE PROTECCIÓN DE DATOS

BOLETÍN OFICIAL DEL ESTADO
MINISTERIO DE LA PRESIDENCIA

BIG DATA, PRIVACIDAD Y PROTECCIÓN DE DATOS

BIG DATA, PRIVACIDAD Y PROTECCIÓN DE DATOS

Elena Gil González

Protección de Datos Personales
Accésit en el Premio de Investigación de 2015

Agencia Española de Protección de Datos

Agencia Estatal Boletín Oficial del Estado
Madrid, 2016

NIPO: 007-16-097-4
ISBN: 978-84-340-2309-3
Depósito Legal: M-16320-2016

IMPRENTA NACIONAL DE LA AGENCIA ESTATAL
BOLETÍN OFICIAL DEL ESTADO
Avda. de Manoteras, 54. Madrid 28050

Para que la sociedad de la información funcione,
debe haber un equilibrio entre la privacidad
y el flujo de datos.

AGRADECIMIENTOS

Quiero mostrar mi agradecimiento a todos aquellos que, de una u otra forma, han contribuido a la preparación de este libro. En concreto, debo hacer una mención especial a las siguientes personas.

En primer lugar, a mi familia, por ser los primeros a quienes recurro en los buenos momentos, y ser los primeros en llegar en los no tan buenos. Sois el motor de mi vida.

A Javier Puyol Montero, por tus explicaciones y por toda la documentación que me has facilitado. Y sobre todo, por confiar en mí desde el principio y empujarme para seguir superándome.

A Javier Aparicio Salom, por ser una fuente inagotable de buenos consejos. El más importante de todos, haberme aconsejado estudiar el *big data* y su problemática jurídica, que en última instancia me ha traído aquí. Gracias también por tus aportaciones durante el proceso de elaboración de este trabajo, que han sido un factor clave para poder elaborar una opinión.

Y a Luis Uribe Beyer, por todo tu tiempo, y por abrirme las puertas de Telefónica y a su equipo de inteligencia de negocio y *big data*.

Índice

PRÓLOGO

La Agencia Española de Protección de Datos convoca anualmente el Premio de Protección de Datos Personales de Investigación con el objetivo de fomentar y promocionar los trabajos más destacados en esta materia. En su edición XIX el Jurado ha otorgado el accésit al trabajo *Big data, privacidad y protección de datos*, de Elena Gil González, una obra que analiza el impacto en el derecho a la privacidad de este nuevo fenómeno que implica el tratamiento y almacenamiento masivo de información.

Uno de los retos actuales de la Agencia, tal y como se recoge en su Plan Estratégico 2015-2019, es que la innovación y la protección de datos discurran de forma paralela. La protección de datos es un factor crítico para conseguir un correcto afianzamiento de la sociedad de la información, determinante para proporcionar productos y servicios respetuosos y de calidad, y, especialmente, para generar confianza en los usuarios de los mismos. La Agencia quiere contribuir a que el sector empresarial consiga un elevado cumplimiento de las obligaciones que la normativa de protección de datos les impone, fomentando una cultura de la protección de datos que suponga una clara mejora de la competitividad, compatible con el desarrollo económico. Para ello, es necesario apostar por políticas proactivas de cumplimiento que permitan detectar el impacto que los nuevos desarrollos tecnológicos pueden tener en la privacidad de los ciudadanos, buscando mitigar los riesgos sin que, en ningún caso, haya que renunciar a las funcionalidades y beneficios que estos proporcionan.

En este sentido, el *Big Data* es un reflejo del paradigma actual de la sociedad de la información y del impacto de la tecnología en la esfera de la vida privada. El despliegue de tecnologías como el *big data*, el internet de las cosas, el uso de *wearables* o las *smartcities,* entre otras, requiere de un análisis y valoración técnica y jurídica para promover buenas prácticas que garanticen su adecuación a la normativa de protección de datos y, en consecuencia, el respeto por los derechos de los ciudadanos.

El trabajo de investigación que el lector encontrará a continuación aborda el *big data* desde una perspectiva optimista, buscando el acer-

camiento entre los beneficios sociales y económicos que puede aportar sin soslayar la garantía del derecho a la protección de datos de las personas.

La obra contribuye a difundir algunos aspectos esenciales de esta tecnología, aportando propuestas para minimizar los riesgos de intrusión en la privacidad. El deber de transparencia y consentimiento, las técnicas de anonimización, la privacidad por defecto y desde el diseño que se abordan en el trabajo son elementos esenciales para un desarrollo respetuoso del *big data*.

Mar España Martí
Directora de la Agencia Española de Protección de Datos

INTRODUCCIÓN

Las tecnologías irrumpen en todas las esferas de nuestra vida. Tal es el nivel de cambio que algunas de estas tecnologías conllevan, que se han llegado a describir como disruptivas. Así por ejemplo, el fenómeno del internet de las cosas, el *cloud computing* o el *big data* son tecnologías disruptivas, que están revolucionando la forma en la que funciona nuestro mundo.

Y con estas tecnologías, los datos se convierten en el activo más preciado. Hoy en día se crean más datos que nunca antes en la historia, y recogerlos, almacenarlos y tratarlos es posible de forma más sencilla que nunca antes. Desde las redes sociales, las compras con tarjetas, las llamadas telefónicas y tantos otros gestos cotidianos generan datos, cuyo estudio es una fuente de valor incalculable.

Las empresas ofrecen servicios de acceso gratuito a cambio de poder acceder a nuestros datos, y poder utilizarlos para una innumerable cantidad de fines, muchos de los cuales ni siquiera se conocen en el momento de analizar los datos.

En concreto, el *big data* es el conjunto de tecnologías que permiten tratar cantidades masivas de datos provenientes de fuentes dispares, con el objetivo de poder otorgarles una utilidad que proporcione valor. Éste puede ser descubrir patrones de comportamiento de los clientes de una organización para crear publicidad dirigida mucho más efectiva, predecir tendencias económicas o descubrir relaciones antes desconocidas entre variables que puedan abrir las puertas a la innovación.

Sin embargo, estas nuevas oportunidades del *big data* también van acompañadas de riesgos. Quizás uno de los más relevantes sea el riesgo que este análisis masivo de datos posa sobre la privacidad de las personas.

En este trabajo examinaremos con detalle este riesgo, y lo analizaremos desde un puto de vista legal. La tecnología evoluciona a un ritmo tan veloz que en ocasiones las normas no son capaces de dar solución a las nuevas problemáticas que se plantean.

El trabajo está dividido en seis capítulos. Así, en el Capítulo I nos acercaremos al concepto de *big data*, sus características y las formida-

bles oportunidades que puede aportar. También repasaremos de forma somera algunos de los riesgos que crea, en concreto, (i) el riesgo de caer en una confianza ciega en los algoritmos que analizan los datos, de modo que nadie revise las conclusiones lanzadas por las máquinas, dando lugar a una toma de decisiones automatizada; y (ii) el riesgo de no comprobar si las relaciones que parecen encontrarse entre las variables son verdaderas o responden al mero azar.

El resto del trabajo estará centrado en el riesgo que el *big data* entraña sobre la privacidad y la protección de datos. De este modo, en el Capítulo II introduciremos el concepto de dato de carácter personal y el marco jurídico de la protección de datos, así como el impacto que el *big data* tiene en esta normativa.

Los siguientes dos capítulos están dedicados a los instrumentos principales de la normativa de protección de datos. El Capítulo III tratará el consentimiento como instrumento que legitima que los datos de carácter personal puedan ser recabados y tratados de acuerdo a la legalidad. El Capítulo IV presenta la anonimización, el mecanismo por el que los datos se hacen anónimos, y que provoca que ya no sean datos personales, de modo que pueden ser tratados sin estar sujetos a las previsiones de la normativa de protección de datos.

Posteriormente, tras haber analizado la normativa actual de protección de datos y la problemática que el *big data* le causa, analizaremos en el Capítulo V si la nueva normativa, todavía en trámites de discusión, es capaz de dar solución a los problemas que hemos identificado.

Por último, el Capítulo VI cierra el trabajo con las conclusiones y propuestas para superar los límites de que adolece el sistema actual.

CAPÍTULO I. LA REVOLUCIÓN DEL *BIG DATA*

1.1 EL NUEVO PARADIGMA

Si hace tan solo unos años se hablaba de la revolución que ha provocado internet, hoy en día nos encontramos ante un nuevo fenómeno, una tendencia tecnológica menos visible pero con igual poder de transformación, el *big data*. Si bien es cierto que son realidades diferentes, también lo es que internet facilita mucho la recogida y la transferencia de datos sobre la que se basa el *big data*.

Big data es un término que alude al enorme crecimiento en el acceso y uso de información automatizada. Se refiere a las gigantescas cantidades de información digital controlada por compañías, autoridades y otras organizaciones, y que están sujetas a un análisis extenso basado en el uso de algoritmos[1]. No es una tecnología en sí misma, sino más bien un planteamiento de trabajo para la obtención de valor y de beneficios como consecuencia del tratamiento de los grandes volúmenes de datos que se están generando día a día[2].

La idea principal es que a partir del tratamiento de cantidades masivas de información, algo hasta ahora imposible, podamos comprender cosas antes desconocidas cuando solo analizábamos cantidades pequeñas de información, y permite descubrir o inferir hechos y tendencias ocultos en las bases de datos. Esta explosión de datos es relativamente reciente. En el año 2000, solamente un cuarto de toda la información mundial estaba almacenada en formato digital; el resto se almacenaba en medios analógicos como el

---

1 GRUPO DE TRABAJO DEL ARTÍCULO 29, «Opinion 03/2013 on Purpose Limitation» (2013).

2 Javier PUYOL. «*Big data* y Administraciones Públicas» [Conferencia]. Seminario Internacional *Big data* para la Información Oficial y la Toma de Decisiones (2014).

papel. Sin embargo, en la actualidad más del 98% de toda nuestra información es digital[3].

En conclusión, el concepto de *big* data se aplica a toda la información que no puede ser procesada o analizada utilizando herramientas o procesos tradicionales. El desafío consiste en capturar, almacenar, buscar, compartir y agregar valor a los datos poco utilizados o inaccesibles hasta la fecha. No es relevante el volumen de datos o su naturaleza. Lo que importa es su valor potencial, que sólo las nuevas tecnologías especializadas en *big data* pueden explotar. En última instancia, el objetivo de esta tecnología es aportar y descubrir un conocimiento oculto a partir de grandes volúmenes de datos[4].

Este fenómeno parte del hecho de que actualmente hay más información a nuestro alrededor de lo que ha habido nunca en la historia, y está siendo utilizada para nuevos usos. A nivel ilustrativo, podemos señalar que desde el inicio de la historia hasta 2003 los humanos habíamos creado 5 exabytes (es decir, 5 mil millones de gigabytes) de información. En 2011 ya creábamos esa misma cantidad de información cada dos días[5].

Además, el *big data* permite transformar en información muchos aspectos de la vida que antes no se podían cuantificar o estudiar, como los datos no estructurados (por ejemplo, datos no-texto como fotografías, imágenes y ficheros de audio). Este fenómeno ha sido bautizado como *dataficación (*o «*datafication*»[6] en inglés) por la comunidad científica. Así, nuestra localización ha sido dataficada, primero con la invención de la longitud y la latitud, y en la actualidad con los sistemas de GPS controlados por satélite. Del mismo modo, nuestras palabras ahora son datos analizados por ordenadores mediante minería de datos. E incluso nuestras amistades y gustos son transformados en datos, a través de los gráficos de relaciones de redes sociales o los *«likes»* de facebook.

---

3 Kenneth NEIL CUKIER y Viktor MAYER-SCHÖENBERGER. «The Rise of Big data. How It's Changing the Way We Think About the World». *Foreign Affairs* Vol. 92, n.º 3 (2013).

4 Javier PUYOL. «Una aproximación a big data». *Revista de Derecho de la Universidad Nacional de Educación a Distancia* (UNED), n.º 14 (2014).

5 Cálculos realizados por Dave Turek, responsable de desarrollo de superordenadores de IBM.

6 Kenneth NEIL CUKIER y Viktor MAYER-SCHÖENBERGER. «The Rise of Big data. How It's Changing the Way We Think About the World». *Foreign Affairs* Vol. 92, n.º 3 (2013).

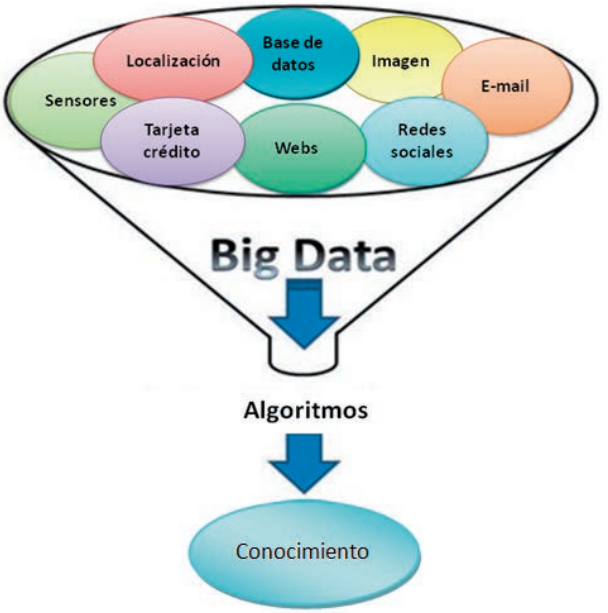


Gráfico: Funcionamiento del *big data*

Hay quienes hablan de los años 80 como del inicio del *big data*, cuando los procesadores y la memoria computacional hicieron posible analizar más información. Sin embargo, el fenómeno del *big data* no deja de ser el último paso de la humanidad en un camino ancestral: el deseo de comprender y cuantificar el mundo.

Pongamos, a modo de ejemplo, cómo el deseo de comprender el mundo dio origen a la ciencia de la astronomía. La curiosidad humana con respecto al día y la noche, el Sol, la Luna y las estrellas llevó a los hombres primitivos al estudio de sus movimientos, y a realizar cálculos para llegar a la conclusión de que los cuerpos celestes parecen moverse de forma regular. La utilidad de estas observaciones era poder orientarse y cuantificar los días y las estaciones. Pero esta datificación del movimiento de los planetas no se detuvo ahí. Más adelante comenzó a desarrollarse la ciencia de la astronomía, cuando cerca del año 300 a. C Artisaco de Samos realizó amplios análisis que permitieron calcular y cuantificar la distancia que separa a la Tierra de la Luna y el Sol, y creó un modelo heliocéntrico del Sistema Solar. Actualmente, la astronomía también hace uso del

big data gracias a la astro-estadística a la inteligencia artificial. El satélite Gaia toma instantáneas de la Vida Láctea desde 2013, y creará hasta un petabyte de información a lo largo de los cinco años que dura la misión. El reto de los astrónomos será organizar toda esta información sobre el universo, y su objetivo, extraer conocimiento de esos datos para clasificar millones de estrellas, descubrir nuevos cuerpos celestes e incluso comprender la evolución de nuestra galaxia.

Hoy en día, los nuevos datos se ponen al servicio de usos antes no conocidos, que ha sido posible desarrollar gracias al crecimiento de la capacidad de memoria de los ordenadores, los poderosos procesadores, el increíble abaratamiento de recopilar y almacenar toda esta cantidad de información, y el desarrollo de análisis matemáticos que provienen de la estadística tradicional. Cuando transformamos la realidad en datos, podemos transformar la información en nuevas formas de valor. Un ejemplo de estos nuevos servicios son los motores de recomendaciones automatizados, que no necesitan a un analista que revise los datos y realice estas recomendaciones. Así, Amazon utiliza datos que extrae de sus clientes para hacernos recomendaciones basadas en compras previas de otros clientes. Y la red profesional Linkedin también nos sugiere personas que podríamos conocer, seleccionando unos pocos focos de datos de entre una cantidad masiva de información sobre sus más de 300 millones de miembros[7].

Las conocidas como las tres «uves» de las bases de datos, variedad, volumen y velocidad, eran incompatibles años atrás, creando una tensión que obligaba a elegir entre ellas. Es decir, podíamos analizar un gran volumen de datos y a alta velocidad, pero era necesario que fueran datos sencillos, como datos estructurados en tablas; esto es, había que sacrificar la variedad de los datos. Del mismo modo, se podían analizar grandes volúmenes de datos muy variados, pero no a gran velocidad; era necesario dejar que los sistemas trabajaran durante horas, o incluso días.

---

7 Lutz FINGER. «Recommendation Engines: The Reason Why We Love Big data». *Forbes Tech* (2 de septiembre de 2014).

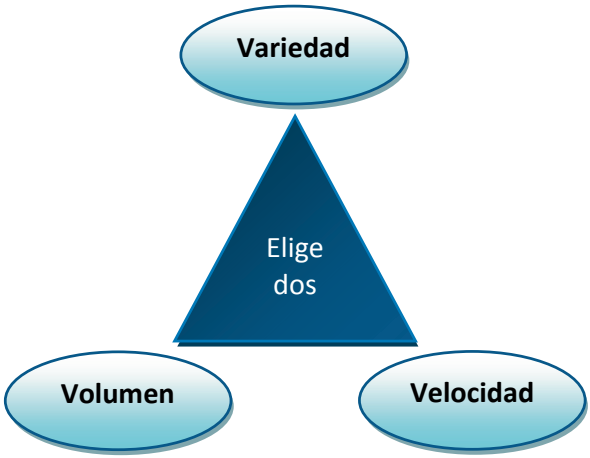


Gráfico: Las tres «uves» de las bases de datos

Sin embargo, con la aparición del *big data* estos tres atributos ya no son excluyentes sino complementarios:

- **Volumen:** el *big data* implica recoger, almacenar y tratar grandes cantidades de datos y metadatos[8], en lugar de estudiar una muestra, como hace la estadística tradicional. El Boston Consulting Group[9] ha estimado que se produce un crecimiento de 2,5 exabytes de información al día (sabiendo que un exabyte son 1.000.000.000.000.000.000 bytes). De hecho, son cantidades tan grandes que, para los no expertos, dejan de tener sentido y añadir más o menos ceros a la cifra ni siquiera nos permite ver la diferencia.

  Este volumen de datos es tan grande que ya no puede ser analizado mediante herramientas y procesos tradicionales tales como MS Excel o SQL. Ha sido necesario comenzar a utilizar nuevos sistemas, como NoSQL o el software Apache Hadoop,

---

8 Los metadatos son datos que describen a otros datos principales, a los que van asociados. Por ejemplo, una fotografía en formato digital tomada con un dispositivo móvil puede contener otra información en forma de metadatos como la localización o la fecha en que la imagen fue tomada.

9 THE BOSTON CONSULTING GROUP. «To get started with big data. BCG Perspectives» (2013), citado por Information Commissioner's Office (ICO, autoridad de protección de datos de Reino Unido), «Big Data and Data Protection» (2014).

que permite trabajar millones de *bytes* de información y organizarlos en miles de nodos.

- **Velocidad:** la velocidad a la que se crean y procesan los datos está en continuo aumento, y con frecuencia para las organizaciones es importante poder analizarlos de forma muy rápida, incluso en tiempo real, algo que en ocasiones es imposible con los sistemas tradicionales. El *big data* permite transferir datos de forma barata y eficiente, y así se pueden analizar tanto los datos dinámicos que se van creando, como los datos estáticos o históricos que ya han sido almacenados de forma previa.

  Por ejemplo, el análisis de datos en tiempo real puede ayudar a seguir la trayectoria de huracanes y su intensidad. Esto podría llegar a permitir realizar predicciones sobre dónde pueden producir daños con horas o incluso días de antelación.

- **Variedad:** los datos recogidos provienen tanto de fuentes estructuradas como no estructuradas: transacciones bancarias, imágenes de satélite, redes sociales, contenidos de páginas web, dispositivos móviles de geolocalización y miles de aplicaciones, las conexiones del internet de las cosas, los servicios web 2.0, e incluso el cuerpo humano (por ejemplo, cuando se utilizan sistemas de identificación biométricos). En la actualidad, solo el 20% de nuestros datos provienen de fuentes estructuradas, mientras que el 80% restante son datos no estructurados[10].

  Extraer información de datos tan diversos supone un gran reto. Las tecnologías que se han desarrollado para el *big data* permiten, entre otras soluciones, combinar datos a pesar de que no se encuentren almacenados en ficheros con la misma estructura. Así por ejemplo, una cadena de tiendas puede analizar de forma conjunta los datos de ventas con los datos de temperaturas para realizar un modelo predictivo en tiempo real para cada uno de sus locales comerciales.

  Algunos expertos consideran que, de todas las «uves», la variedad es la característica más relevante del *big data*[11]. Y ello

---

10 Cas PURDY. «Finding benefits in big data» [Infografía]. *Trustwave Blog* (2013).

11 INFORMATION COMMISSIONER'S OFFICE (ICO, autoridad de protección de datos en Reino Unido). «Big data and Data Protection» (2014).

porque, si por ejemplo una empresa desea extraer información de su propia base de datos de clientes, a pesar de que ésta sea muy grande, puede que no necesite utilizar nuevas herramientas de análisis o enfrentarse a nuevos problemas de privacidad. Sin embargo, cuando la empresa desea combinar esos datos con otras fuentes externas, entonces estará llevando a cabo actividades completamente diferentes que ya sí se podrían llamar *big data*.

Estas «tres uves» pueden ser, además, ampliadas con otras tres más: veracidad, visualización y valor de los datos.

- **Veracidad:** la veracidad hace referencia al nivel de fiabilidad o de calidad de los datos. Conseguir datos de alta calidad se ha convertido en todo un reto, principalmente importante cuando se trata de datos no estructurados. Sin embargo, tal y como IBM asegura[12], algunos datos son inciertos por naturaleza, como los sentimientos, el futuro, los sensores GPS que rebotan entre los rascacielos de una ciudad, o los datos creados en entornos humanos como las redes sociales; y ninguna limpieza de datos puede corregirlos. Así, manejar la incertidumbre es una cuestión esencial al tratar con tecnologías *big data*.

  Y tal y como IBM propone, para gestionar la incertidumbre los analistas han de crear un contexto en torno a los datos. Un modo posible de hacerlo es combinar diversas fuentes de datos para dar lugar a una información más fiable. Por ejemplo, cuando los comentarios en redes sociales se combinan con datos de geolocalización para analizar las reacciones de la gente y el impacto de un evento como un concierto multitudinario, un mitin electoral o partido de fútbol.

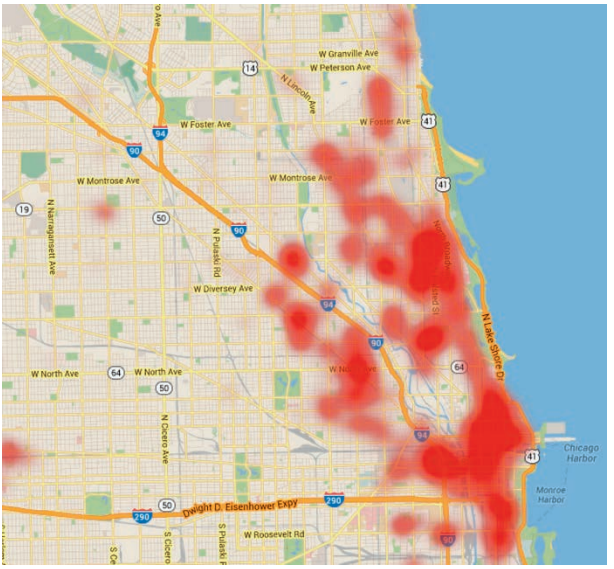
- **Visualización:** poder visualizar los datos es básico para comprenderlos y tomar decisiones en consonancia.

  Por ejemplo, la utilización de técnicas *big data* permite combinar datos y obtener una predicción de cuándo se producirán determinados tipos de crímenes y dónde (después de un partido

---

12 IBM INSTITUTE FOR BUSINESS VALUE, en colaboración con la Escuela de Negocios Säid de la Universidad de Oxford. «Analytics: el uso del *big data* en el mundo real». *IBM Global Business Services* (2012).

de fútbol multitudinario, etc.). Sin embargo, una lista de páginas interminables con coordenadas que muestren dónde se predicen los crímenes no resulta manejable. Los instrumentos de visualización mediante, por ejemplo, mapas en los que la intensidad del color muestre la probabilidad de que se produzca cada tipo de crimen puede ser crucial para llegar a comprender realmente los datos.



- **Valor:** la finalidad última de los procesos de *big data* es crear valor, ya sea entendido como oportunidades económicas o como innovación. Sin él, los esfuerzos dejan de tener sentido.

Aparte de estas características, el *big data* también se puede caracterizar en función de sus diferencias con respecto a las herramientas de procesamiento tradicional, tales como la utilización intensiva de algoritmos o la utilización de los datos para nuevos fines[13]:

- **La utilización intensiva de algoritmos:** antes del *big data*, para analizar un fichero de datos era necesario decidir en primer lugar qué se quería estudiar y qué se esperaba encontrar,

13 Ídem.

esto es, establecer una hipótesis, para así poder lanzar una búsqueda e identificar los datos relevantes asociados a esos parámetros. En ocasiones, cuando el análisis era más complejo, se podían ejecutar numerosos algoritmos sobre los datos para así encontrar correlaciones. Entonces, ¿qué aporta el *big data* de nuevo?

Encontrar una correlación significa hallar un fenómeno que desconocíamos, y esto implica hallar nueva información. Esta nueva información puede volverse a introducir en los algoritmos para que realicen búsquedas más exactas, de modo que los resultados anteriores perfeccionan el funcionamiento del algoritmo y el sistema «aprende». Esto es lo que se conoce como «*machine learning*» (que podríamos traducir como aprendizaje computacional). Pues bien, el uso de algoritmos de este modo es una novedad del *big data*.

> **Ejemplo: UCB e IBM colaboran para personalizar la atención de pacientes con epilepsia[14].**
>
> La empresa bio-farmacéutica UCB y la empresa tecnológica IBM están desarrollando de forma conjunta un proyecto basado en *machine learning* para mejorar los tratamientos que se aplican a las personas que padecen epilepsia. El proyecto trata de desarrollar un sistema capaz de analizar millones de datos clínicos de pacientes, así como literatura médica, para ayudar en el momento de la toma de decisiones.
>
> El sistema de análisis de datos combina técnicas de aprendizaje computacional y procesamiento de lenguaje natural, con el objetivo de que se puedan introducir los datos de un paciente concreto y obtener, por ejemplo, una predicción que determine la probabilidad de que un tratamiento concreto tenga éxito.

- **La utilización de los datos para nuevos fines:** la analítica *big data* a menudo reutiliza datos que fueron obtenidos para una primera finalidad, y les otorga una finalidad nueva. Esto es consecuencia de todo lo anterior. Si analizar cada vez más da-

---

14 Nota de prensa publicada por IBM Research (2013).

tos nos permite llegar a conocer más información, las organizaciones o gobiernos pueden identificar problemas antes desconocidos que pueden ser comprendidos o atajados con esta información.

Por ejemplo, imaginemos una planta de generación de energía eólica que instala sensores inteligentes en los molinos para monitorizar su funcionamiento. El objetivo de la compañía es poder gestionar de forma más eficiente cuántos técnicos serán necesarios en cada área para revisar las posibles averías. Sin embargo, de los datos sobre las averías de los molinos se pueden extraer beneficios para otros fines. Por ejemplo, la empresa puede observar qué partes se averían con más frecuencia, y estos datos a su vez se pueden combinar con los datos de los proveedores para considerar comprarle a otro fabricante esas piezas que fallan. O también se puede llegar a saber qué fallos ocurren más a menudo cuando el clima es caluroso y seco respecto de cuando el clima es frío y húmero; y con esto la empresa puede gestionar sus inventarios de forma más precisa.

Todos estos cambios están suponiendo una modificación de paradigma en la forma en que se analiza la información. Este cambio de paradigma se concreta en tres grandes tendencias que han sido ilustradas por Kenneth Neil Cukier y Viktor Mayer-Schöenberger[15].

En primer lugar, el cambio del «algo» al «todo». Tradicionalmente, la forma en que se trataban los datos era mediante muestreos representativos de la realidad. Confiábamos en que cantidades de información pequeñas, que se pudieran manejar de forma fácil, explicaran realidades complejas. En cuestiones generales, las muestras y la estadística funcionan bien, pero cuando queremos obtener conclusiones de subgrupos concretos de la muestra, la estadística deja de ser fiable. Ello es porque las muestras aleatorias son suficientes para describir la realidades globales, pero no para detectar comportamientos particulares de subgrupos.

---

15 Kenneth NEIL CUKIER y Viktor MAYER-SCHÖENBERGER. «The Rise of Big data. How It's Changing the Way We Think About the World». *Foreign Affairs* Vol. 92, n.º 3 (2013).

Por ejemplo, la estadística es capaz de dar respuesta a la pregunta de qué candidato electoral es el preferido por las mujeres solteras menores de 30 años. La estadística obliga a saber *a priori* qué queremos analizar y elegir una muestra en concordancia. Pero si una vez hecho el sondeo queremos volver a analizar un subgrupo de esta población, por ejemplo, el candidato preferido de mujeres solteras menores de 30 años, con educación universitaria, nacionalidad española y padres extranjeros, las conclusiones no serán válidas. Esto es porque en la muestra elegida probablemente no haya un grupo suficientemente grande con estas últimas características como para poder extraer conclusiones. Sin embargo, si recogemos información de forma masiva, este problema desaparece. No necesitamos saber de antemano para qué queremos la información, simplemente recogemos toda la información posible y de este modo podremos analizar el comportamiento del grupo principal así como de los diversos subgrupos que queramos crear *a posteriori*.

En segundo lugar, el cambio de lo «limpio» a lo «caótico». Si pretendemos recoger tal cantidad de datos, hemos de desistir en el intento de que toda esta información esté estructurada y limpia, y debemos aceptar algo de desorden. Los beneficios de analizar grandes cantidades son mayores que los inconvenientes de permitir estas pequeñas inexactitudes (siempre que los datos no sean completamente incorrectos).

Viktor Mayer Schöenberger pone como ejemplo los traductores automáticos. Sus inicios se remontan a los años 90, cuando el Gobierno canadiense tuvo necesidad de un medio eficiente para traducir sus escritos al inglés y al francés. En ese momento, IBM ideó un sistema estadístico de traducción, de forma que infería qué palabra del otro idioma era la mejor alternativa para la traducción. En la actualidad, Google ha tomado el relevo. Ahora, su traductor cuenta con más de 90 idiomas, que abarcan desde los más usados hasta otros como el cingalés o el kazajo. El funcionamiento del motor de traducción de Google no se basa en unas pocas traducciones perfectas, sino en enormes cantidades de datos de fuentes muy variadas: webs corporativas de empresas, documentos de la Unión Europea en todas sus versiones traducidas, etc. Los resultados de sus traducciones no son perfectos, pero son sin duda útiles en una gran cantidad de idiomas.

Detectar idioma	Chichewa	gallego	japonés	nepalí	tamil
afrikáans	chino	georgiano	javanés	noruego	tayiko
albanés	Cingalés	griego	jemer	persa	telugu
alemán	coreano	gujarati	kazajo	polaco	turco
árabe	criollo haitiano	hausa	lao	portugués	ucraniano
armenio	croata	hebreo	latín	punjabí	urdu
azerí	danés	hindi	letón	rumano	uzbeco
bengalí	eslovaco	hmong	lituano	ruso	vietnamita
bielorruso	esloveno	holandés	macedonio	serbio	yidis
Birmano	español	húngaro	malayalam	sesoto	yoruba
bosnio	esperanto	igbo	malayo	somalí	zulú
búlgaro	estonio	indonesio	malgache	suajili	
canarés	euskera	inglés	maltés	sueco	
catalán	finlandés	irlandés	maorí	sundanés	
cebuano	francés	islandés	maratí	tagalo	
checo	galés	italiano	mongol	tailandés	

Gráfico: Idiomas disponibles en el traductor de Google

En tercer lugar, se está produciendo un cambio de la «causalidad» a la «correlación». Ya no importa tanto descubrir la causalidad entre dos hechos, sino su correlación. Así, en lugar de tratar de comprender exactamente por qué una máquina se estropea o por qué los efectos secundarios de un medicamento desaparecen, el *big data* permite a los investigadores recoger y analizar cantidades masivas de datos sobre estos eventos y todo lo asociado a éstos, para encontrar relaciones entre variables que permitan descubrir patrones ocultos y predecir cuándo estos eventos pueden volver a ocurrir. En todo caso, hay que tener en cuenta que esta aproximación a la realidad también conlleva riesgos, que serán analizados más adelante en este trabajo.

1.2 BENEFICIOS

En este contexto, es claro que las oportunidades que genera el *big data* son enormes, y estas oportunidades, son ya hoy en día, en muchos casos, un beneficio tangible.

El universo digital es un área empresarial absolutamente en alza, que tendrá un enorme valor en el futuro. Algunos de los beneficios más relevantes del *big data* son poder ofrecer una visión cada vez más precisa de las fluctuaciones y rendimientos de todo tipo de recursos, permitir realizar adaptaciones experimentales a cualquier escala de un proceso y conocer su impacto en tiempo casi real, ayudar a conocer mejor la demanda y así realizar una segmentación mucho más

ajustada de la oferta para cada bien o servicio, o acelerar la innovación y la prestación de servicios cada vez más innovadores y más eficientes[16].

Los datos para obtener estos conocimientos provendrán tanto de las personas como de los objetos, y con mayor énfasis a medida que el denominado internet de las cosas se generalice. Sin embargo, las previsiones estiman que tan solo el 0,5% de la información será efectivamente procesada[17].

Por ejemplo, el Instituto Global McKinsey estima que la explotación de conjuntos de datos masivos alberga un potencial anual de hasta 240.000 millones de euros para la sociedad estadounidense, y una valor de hasta 200.000 millones de euros solamente para la administración de la Unión Europea, ayudando en cuestiones como la recaudación de impuestos, la eficiencia energética o a creación de *smart cities* que conjuguen el internet de las cosas con el tratamiento de grandes cantidades de datos[18].

Las grandes empresas supieron ver el valor potencial de las técnicas del *big data* y la minería de datos hace años, y así Axciom, Google, IBM o Facebook llevan años invirtiendo en descubrir nuevos usos de los datos, cómo tratarlos y cómo transformarlos en valor.

Siguiendo a los grandes pioneros, en la mayor parte de los sectores, tanto compañías maduras como nuevos entrantes están poniendo en marcha estrategias para innovar y capturar valor. Por ejemplo, en el sector sanitario, algunas empresas pioneras están analizando los resultados que determinados medicamentos ampliamente prescritos tienen sobre la salud, y están descubriendo beneficios y riesgos que no fueron descubiertos durante los ensayos clínicos. Otras empresas están recolectando datos provenientes de sensores integrados en productos tales como juguetes para niños o bienes industriales, para determinar cómo se están utilizando estos productos en la práctica. Con este nuevo conocimiento, las empresas son capaces de generar nuevos

---

16 Javier PUYOL. «*Big data* y Administraciones Públicas» [Conferencia]. Seminario Internacional *Big data* para la Información Oficial y la Toma de Decisiones (2014).

17 EMC DIGITAL UNIVERSE STUDY, «The Digital Universe and big data». (2014).

18 Ídem. No ha sido posible contrastar el informe de referencia.

servicios y diseñar productos futuros. De este modo, el análisis de datos se convierte en una importante ventaja competitiva para las empresas.

Un ejemplo concreto en el sector de la venta minorista es la cadena de supermercados Walmart, que recoge datos sobre las compras de sus clientes que después analiza para comprender sus hábitos de consumo. Con los millones de bytes de información que posee, la empresa decidió intentar realizar predicciones de ventas en determinadas circunstancias, como en situaciones de alarma de huracán. El análisis de los datos descubrió patrones tan sorprendentes como que el producto estrella que los consumidores compran antes de estos huracanes es cerveza, o que ante una amenaza de huracán, se disparan las ventas de los dulces de fresa «Pop Tarts» hasta siete veces por encima de las ventas ordinarias. Con este conocimiento, la cadena se abastece antes de un huracán, y esta información no es solo poder, sino también dinero.

Y no son únicamente las empresas las que utilizan los datos masivos para obtener beneficios.

Para las Administraciones Públicas el uso de *big data* puede implicar la toma de decisiones más rápida y eficaz, poder realizar análisis predictivos o una mejora continua de los sistemas de trabajo, además de mejorar la eficiencia en cuestiones tan sensibles como la protección ciudadana o la asistencia sanitaria.

El Ayuntamiento de la ciudad de Nueva York ha utilizado la analítica de datos masivos para fines tan dispares como prevenir atascos, pues la regulación del tráfico en las grandes ciudades en una causa de problemas conexos como la dificultad de atender a víctimas de incendios o la ineficiencia de los servicios en la ciudad. En cualquier caso, ningún sistema ha logrado aún dar con una solución estable a los problemas de tráfico.

En España numerosas empresas también se han sumado a los beneficios del *big data*.

Telefónica dedica importantes esfuerzos en desarrollar líneas de investigación utilizando técnicas de *big data*. Por ejemplo, el pasado mes de mayo de 2014 Telefónica I+D publicaba un informe elaborado de forma conjunta con RocaSalvatella sobre turismo en las ciudades de Madrid y Barcelona, realizado a partir de grandes cantidades de datos de dos empresas diferentes, Telefónica Móviles España y BBVA. En

concreto, el informe ha cruzado los datos de los terminales extranjeros que utilizaron la infraestructura de Telefónica con los datos de los pagos electrónicos por tarjetas extranjeras que usaban las infraestructuras de BBVA. Para publicar las conclusiones del informe garantizando la privacidad de los usuarios, los datos fueron previamente anonimizados, agregados y extrapolados mediante técnicas estadísticas.

De igual modo, la empresa O2, filial de Telefónica en Reino Unido, también declara cruzar grandes cantidades de datos provenientes de fuentes como historiales de pagos, redes sociales, a qué empresas llaman los clientes, preferencias de los consumidores y segmentación, etc. En una sesión interna liderada por Dave Watkins, Director de Analítica Estratégica e Inteligencia Empresarial de O2 en Reino Unido[19], se da cuenta de las grandes oportunidades que Telefónica ve en el *big data*. El análisis permitiría, afirman, añadir valor a los clientes mediante el estudio de cómo los productos de Telefónica se utilizan, observar la localización de los usuarios, sus hábitos televisivos, etc. Con ello podrían inferir conocimientos de gran valor para la estrategia de la empresa, como patrones de movilidad o círculos sociales de los clientes. También se pueden generar matrices de movimientos de la población o elaborar mapas de densidad de población, estimados por la concentración de llamadas. Esto podría, a su vez, ayudar a extraer datos valiosos para otros agentes sociales como agencias gubernamentales. O2 calcula que el aumento de beneficios que podría obtener de utilizar el *big data* ascendería a 434 millones de libras.

Sin embargo, como Telefónica manifiesta, el talón de Aquiles de las empresas que operan con datos es la privacidad, y el riesgo reputacional al que se enfrentan es muy elevado. Grandes empresas como Google, Facebook, AOL o Microsoft se encuentran entre las peor percibidas por los usuarios en términos de privacidad. Y es que, como veremos más adelante, la defensa de la privacidad y la protección de datos es uno de los retos más importantes a los que se enfrenta el *big data* en la actualidad.

En resumen, el *big data* ayudará a crear nuevas oportunidades de negocio e incluso nuevos mercados y nuevas categorías de empresas. Es de prever que muchas de estas nuevas empresas se sitúen en medio de

---

[19] La información que se expone a continuación proviene de una reunión con personal de Telefónica en su sede de Madrid (marzo de 2015).

los flujos de datos, para capturar y analizar la información sobre productos y servicios, proveedores y clientes, o preferencias de consumo.

Pero es más, algunas de las más importantes oportunidades de creación de valor de los datos personales todavía son desconocidas. Mucha de la información capturada reside actualmente en silos separados por diferentes normas legislativas y contratos, y la falta de un sistema efectivo para la transferencia de datos impide crear valor de forma segura. Sin embargo, es necesario mover los datos para producir valor. «Los datos por sí mismos situados en un servidor son como el dinero debajo del colchón. Es seguro, pero estancado e infrautilizado»[20].

Además de las enormes oportunidades que presenta el *big data*, no hay que olvidar que también presenta ciertas limitaciones, que veremos a continuación.

1.3 **RIESGOS**

En efecto, el *big data* debe hacer frente a determinados retos o limitaciones. En concreto, algunos de los retos más importantes (dejando de lado las dificultades técnicas de almacenamiento o investigación computacional) son: (i) el riesgo de caer en conclusiones erróneas que nadie revisa; (ii) el riesgo que para las personas pueda tener tomar decisiones automatizadas sin un sesgo humano; y (iii) el riesgo para la privacidad de las personas. En este epígrafe analizaremos los dos primeros riesgos de modo somero, para después centrar el resto de los capítulos de la investigación en los problemas que el *big data* plantea sobre la privacidad y la protección de datos.

(i) RIESGO DE CAER EN CONCLUSIONES ERRÓNEAS QUE NADIE REVISA: ERROR POR AZAR Y ERROR POR CONFUSIÓN

Una de las ideas fundamentales del *big data* es que el análisis masivo de datos pasados puede localizar patrones que permitan realizar

---

20 FORO ECONÓMICO MUNDIAL y THE BOSTON CONSULTING GROUP. «Rethinking Personal Data: Strengthening Trust» (2012). Proyecto «Rethinking Personal Data».

predicciones futuras. Pero después de analizar los datos es importante encontrar la verdadera relación entre las variables para poder crear un modelo predictivo. Es decir, es imprescindible poder diferenciar la *causalidad* de la *casualidad*. En ocasiones, tendemos a confundir ambos conceptos, pero aunque pueda parecer un trabalenguas, diferenciarlos tiene importantísimas consecuencias prácticas. Y de hecho, la causalidad es el ámbito de la estadística más malinterpretado y utilizado de forma incorrecta por los no especialistas. De modo que antes de poder establecer esta diferencia, comenzaremos por repasar el concepto de correlación estadística de forma muy simple.

En las ciencias estadísticas, la correlación es el grado de relación entre dos variables. Es decir, se dice que dos variables están correladas cuando el aumento o disminución de una provoca un cambio claro en la otra. De este modo, si el aumento de un valor va acompañado del aumento de otro valor, habrá una correlación positiva. Si el aumento de un valor hace que se observe una disminución en otro valor, tendremos una correlación negativa. Y si, a pesar de que un valor cambie, no observamos cambio alguno en otro valor, habrá una correlación cero. Por ejemplo, hay una correlación positiva entre el número de horas que una persona estudia y la calificación que obtiene en un examen. También se ha encontrado una relación entre el nivel de PIB de un país y el tamaño medio del pene de sus habitantes.

Pues bien, cuando dos variables presentan correlación, es posible que también presenten una relación de *causalidad*. Esto implica que un evento es consecuencia directa del otro, o lo que es lo mismo, que existe una relación causa-efecto de forma tal que la ocurrencia del primer suceso (que denominamos causa), provoca el segundo (que denominamos efecto). En el ejemplo mencionado, hay una relación de causalidad entre el número de horas de estudio y el resultado de un examen. Sin embargo, una correlación entre dos variables no siempre implica que haya causalidad.

En efecto, en ocasiones dos variables presentan una correlación, aunque ésta ocurre por mera *casualidad* o azar. A este tipo de relaciones se las denomina espurias o falsas. Por ejemplo, es mero azar que el nivel de PIB de un país esté correlacionado con el tamaño del pene de los hombres de dicho país.

Así, los datos estadísticos pueden mostrar una correlación, y esto resulta un punto de partida idóneo; pero tras ello está en nuestras

manos añadir un enfoque subjetivo y estudiar si ciertamente existe un patrón entre ambas variables que explique una verdadera conexión, o si por el contrario se trata de una mera coincidencia.

Pongamos otro ejemplo: hace años se observó que existía una correlación entre los casos de cáncer y el consumo de tabaco. En un primer momento no se sabía si el tabaco causaba realmente un aumento en la probabilidad de sufrir cáncer, de modo que hubo que iniciar una investigación médica para determinar que, ciertamente, había una relación causa-efecto entre ambas variables.

En concreto, podemos encontrar dos tipos de errores en la interpretación de relaciones espurias, el error por azar y el error por confusión. Pero si esto siempre ha sido un problema que era necesario tener en cuenta, ¿por qué es especialmente importante al hablar de *big data*?

En primer lugar, analicemos el error por azar. El estadístico Stanley Young viene alertando desde hace tiempo de lo que ha denominado «la tragedia de los grandes conjuntos de datos»: cuantas más variables se estudian en un gran conjunto de datos, más son las correlaciones que pueden evidenciar una significación estadística espuria o falsa. Así, cuantos más datos tengamos, más probabilidades habrá de encontrar relaciones ilusorias sin ninguna significación real, aunque ambas presenten una fuerte relación estadística. Del mismo modo, mediante el método de Montecarlo para generar variables aleatorias, se ve que las correlaciones espurias crecen de forma exponencial con respecto al número de variables. Esto conlleva que, de interpretarse de forma errónea, el analista puede terminar siendo engañado por los datos.

Comprobación: cuantas más variables se estudian, más correlaciones espurias aparecen.

Ricardo Galli, doctor en informática y activista del software libre, llevó a cabo el experimento siguiente[21]. Supongamos que tenemos los siguientes datos de evolución de cinco variables económicas de los últimos años.

---

21 Ricardo GALLI. «Sé cuidadoso con el *Big Data*». Blog De Software, libre, internet, legales (29 de mayo de 2013).

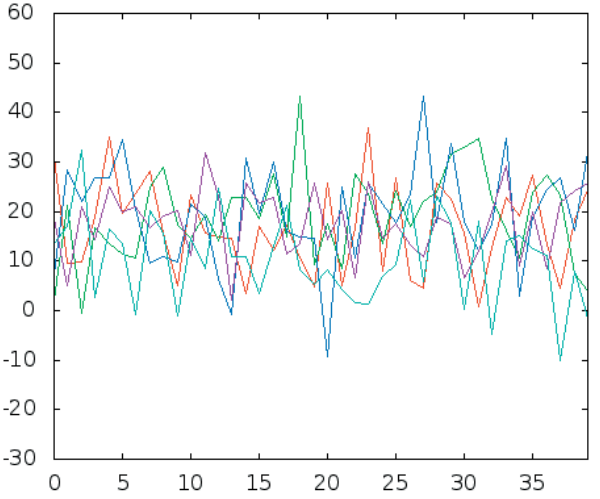


Gráfico: Variables 1 a 5

Esta pequeña cantidad de datos no nos muestra ninguna correlación entre las variables. Pero ahora supongamos que, en lugar de cinco variables, podemos analizar mil variables (algo similar a lo que ocurriría con el *big data*). Nuestro gráfico sería algo similar a esto:

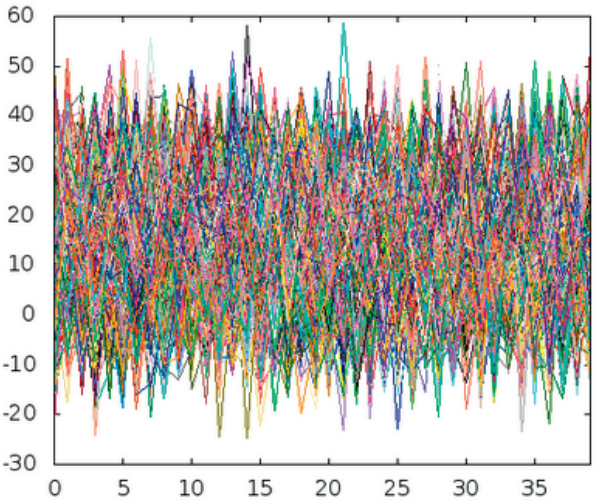


Gráfico: Variables 1 a 1000

Analizando tantos datos, encontramos variables correlacionadas positivamente (es decir, crecen y decrecen a la par), como por ejemplo, las presentadas en el siguiente gráfico. Es decir, cuando hemos ampliado el número de variables del estudio, ha aumentado el número de correlaciones que podemos observar entre dichas variables.

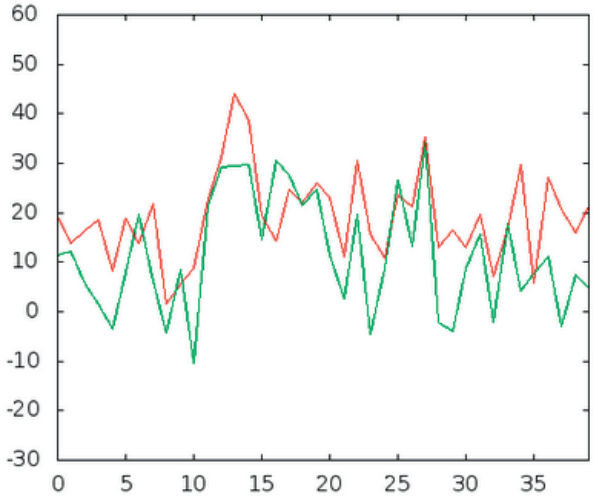


Gráfico: Dos variables muestran correlación positiva

El análisis de datos masivos ha permitido detectar correlaciones que desconocíamos, y que a su vez nos pueden permitir descubrir otra información que desconocemos. El problema es que esta relación se debe al azar.

En efecto, estos datos fueron creados de forma artificial por Galli. En su experimento, explica que generó estas variables económicas con números pseudo-aleatorios e independientes, de modo que las conclusiones que se obtienen también son aleatorias.

Imaginemos las consecuencias que una malinterpretación de variables económicas puede tener en la toma de decisiones de un país, o en las ciencias econométricas[22].

---

[22] La econometría es la rama de la economía que se vale de diversos recursos estadísticos y matemáticos para realizar análisis, interpretaciones y predicciones acerca de sistemas económicos.

Como ya hemos mencionado, lo que sucede en estas series es lo que en estadística se conoce como error por azar, la relación entre ambas variables es pura coincidencia.

Así las cosas, en un momento en el que almacenamos y analizamos cantidades masivas de datos, estamos expuestos a encontrar más relaciones espurias que nunca, que si no son cuestionadas, nos harán caer en conclusiones erróneas. Debido a la confianza ciega que parece haber en la actualidad por los datos, dichas conclusiones erróneas pueden determinar decisiones sobre las personas, acciones estratégicas empresariales o políticas gubernamentales que están basadas en el puro azar.

Encontrar las *causas* de un determinado hecho es lo que ciertamente nos permite conocer cómo funciona la realidad y predecir cómo funcionará en el futuro. Sin embargo, el modo de encontrar la causa real de una relación es mediante experimentos controlados. El problema es que, en la mayoría de las ocasiones, éstos son demasiado caros, largos o incluso imposibles de llevar a cabo técnicamente. Es por ello que éste podrá ser un área de investigación importante en los próximos años[23].

En segundo lugar, analicemos el llamado error por confusión con otro ejemplo.

> **Ejemplo: Relación entre la población de cigüeñas y la tasa de natalidad.**
>
> En 1952 el matemático polaco J. Neyman observó que en numerosas regiones rurales había una relación directa positiva entre el número de cigüeñas que habitaban en los pueblos y el número de nacimientos en dichos pueblos[24]. Es decir, la evolución de las series de población de cigüeñas y habitantes mostraba que aquellas regiones que tenían una mayor población de cigüeñas, también presentaban una mayor tasa natalidad.

---

23 Joris M. MOOIJ, Jonas PETERS, Dominik JANZING, Jakob ZSCHEISCHLER, Bernhard SCHÖLKOPF. «Distinguishing cause from effect using observational data: methods and benchmarks», versión 2. *Journal of Machine Learning Research, Cornell University* (2015).

24 J. NEYMAN. «Lectures and Conferences on Mathematical Statistics and Probability». Departamento de Agricultura de Estados Unidos (1952). Solo ha sido posible acceder a referencias del estudio.

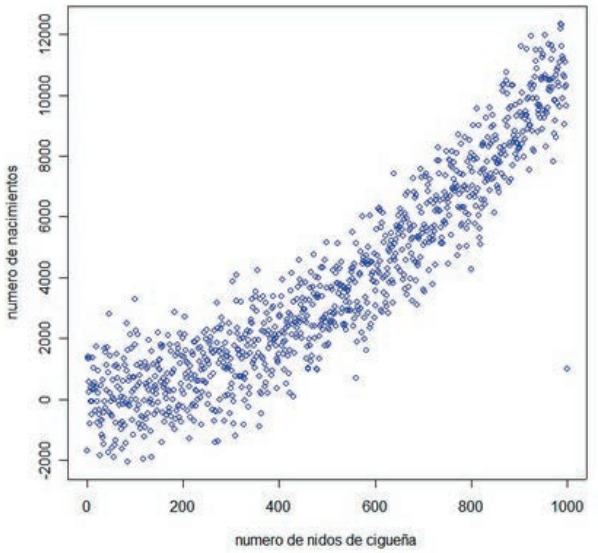


Gráfico: Tasa de natalidad y población de cigüeñas

¿Será ésta la demostración de que a los niños les trae la cigüeña? Quizás esto podría verse como la prueba de ello, pero sabemos que no es así.

En efecto, una vez más, ambas variables no tienen una relación de causa-efecto. La razón de esta relación es que tanto el número de cigüeñas como el número de niños dependen de una tercera variable, la calidad de las cosechas. En los años de bonanza, con más sol, lluvia y alimentos, las cigüeñas criaban más, al igual que los habitantes de dichas regiones.

Lo que sucede en esta serie es lo que se conoce como error por confusión. Es decir, existen dos eventos que son independientes, pero que parecen estar relacionados; en este caso la población de cigüeñas y la tasa de natalidad. Esta aparente relación se debe a que existe un tercer factor desconocido que está afectando a ambas series al mismo tiempo; en nuestro ejemplo, las buenas cosechas. Así se crea un triángulo en el que un lado es la relación que observamos, pero falsa, y los otros dos lados son relaciones ocultas, pero reales.

Gráfico: Error por confusión

Consciente de esta realidad, Tyler Vigen, un estudiante de Harvard, decidió aportar un poco de humor creando el sitio web «*Spurious Correlations*»[25], en el que pone a disposición pública innumerables relaciones espurias entre series de lo más dispar. Con esta creación, Vigen trata de demostrar que si introducimos variables de forma ciega en nuestros cálculos, existe una posibilidad aleatoria de que algunas variables parezcan relacionadas, aunque en realidad no lo estén.

Así, podemos observar que hay una fuerte correlación, de más del 87%, entre la edad de las mujeres que han ostentado el título de Miss América y los homicidios producidos con objetos calientes; o una relación casi perfecta, del 99.79%, entre el gasto de Estados Unidos en ciencia y tecnología, y el número de suicidios por ahorcamiento y métodos similares.

---

25 Spurious Correlations [Sitio web]. Disponible en: http://www.tylervigen.com/spurious-correlations.

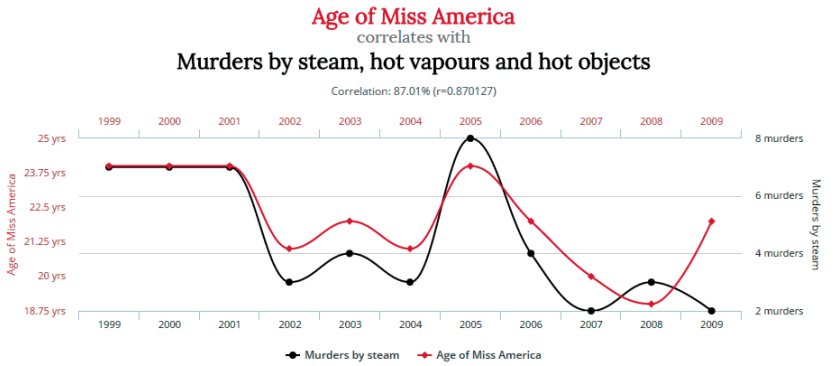


Gráfico: correlación entre la edad de Miss América y los homicidios producidos por vapor y objetos calientes

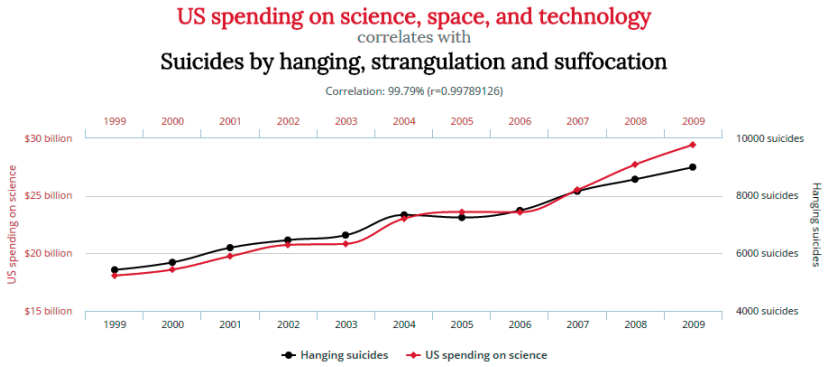


Gráfico: Correlación entre el gasto de Estados Unidos en ciencia, espacio y tecnología, y el número de suicidios por ahorcamiento y métodos similares

En conclusión, la correlación no implica causalidad, sino que puede haber un factor de puro azar o una tercera variable que esté influyendo en la relación, creando una falsa apariencia de causa-efecto. En el contexto del *big data*, este hecho se agudiza, porque el número de correlaciones que podemos encontrar aumenta significativamente.

Las conclusiones que debemos sacar de esto son: (i) es necesario ser críticos con los resultados numéricos que vemos; (ii) siempre que sea posible, debemos buscar la causa o el mecanismo por el que se produce una relación entre sucesos; y (iii) debemos exigir rigor científico en los resultados que arrojan los datos.

Aun así, y dentro de sus limitaciones, la estadística sigue siendo muy útil para obtener información y orientarnos en la dirección correcta. Algunas hipótesis han resultado acertadas. Por ejemplo, la relación entre el tabaco y el aumento del riesgo de cáncer. En este caso, la correlación pudo establecerse mucho antes de que se descubriesen los mecanismos y causas exactas; y éstas a su vez pudieron conocerse porque los investigadores no se conformaron con descubrir la relación, y los estudios continuaron en los años siguientes. En este sentido, «las estadísticas son como una farola para un borracho: deben servir más para apoyarnos que para iluminarnos», y la búsqueda de las causas reales existentes en las relaciones que surgen de las montañas de datos que se analizan todos los días será un reto en los años próximos.

(ii) RIESGO DE LA TOMA DE DECISIONES AUTOMATIZADAS

La evolución de la ciencia de datos nos hace cuestionarnos cuándo hace falta la intervención de una persona que supervise las conclusiones obtenidas de forma automatizada antes de que se transformen en decisiones; decisiones que pueden abarcar espectros tan amplios como la publicidad, la concesión de un préstamo o un diagnóstico médico.

Muchas de las operaciones que se realizan en internet están basadas en la toma de decisiones automatizadas sin intervención humana, salvo, evidentemente, la intervención previa para fijar los parámetros para la adopción de dicha decisión automatizada[26]. Es decir, hay intervención humana cuando se crean los algoritmos que analizarán los datos para tomar una decisión, pero en muchas ocasiones no vuelve a haber un control humano para comprobar dicha decisión.

La investigación básica y la intuición están siendo usurpadas por fórmulas algorítmicas[27]. Steve Lohr afirma que, de hecho, la toma de

---

26 Ana Victoria SÁNCHEZ URRUTIA, Héctor Claudio SILVEIRA GORSKI, Mónica NAVARRO MICHEL, Stefano RODOTÁ. «Tecnología, intimidad y sociedad democrática». *Icaria* (2003).

27 Richard ANDERSON. «Cómo los matemáticos dominan los mercados». *BBC Economía* (1 octubre 2011).

decisiones automatizadas está ideada para eliminar a los humanos de la ecuación, «pero el impulso de querer que una persona supervise los resultados que vomita el ordenador es muy humano»[28]. Como nuestra lógica es diferente a la de las máquinas, necesitamos sentir que las correlaciones se traducen a causas. Siempre hemos funcionado así: motivar nuestras conclusiones es una de las cuestiones a las que más tiempo dedicamos en nuestros análisis. Este esquema se desmorona si el *big data* nos dice lo que tenemos que hacer sin más justificación.

Confiar ciegamente en los algoritmos lleva a que en muchas ocasiones las empresas tomen decisiones sobre nosotros sin que podamos saber por qué las han tomado. Éste es uno de los caballos de batalla del *big data*.

Muchos consideran que el sector del marketing es el lugar idóneo para probar y corregir las nuevas herramientas matemáticas y tecnológicas, pues tiene pocos riesgos y muchos beneficios. En marketing, un error implica simplemente que un consumidor vea un anuncio erróneo, y un acierto puede suponer un aumento de las ventas.

Sin embargo, la preocupación aumenta cuando estas nuevas técnicas comienzan a utilizarse en otros sectores, como el bancario, el asegurador, y sobre todo, el sanitario. En éstos, surgen serias dudas sobre cuándo es realmente necesaria la intervención humana para supervisar los resultados alcanzados por los algoritmos. Una de las tendencias es hacer que el humano siga formando parte del proceso de toma de decisiones; pero otros opinan que esto sería contraproducente.

Así por ejemplo, en el sector bancario numerosas empresas están acudiendo al análisis de los macrodatos, siguiendo el principio básico de la banca de «conozca a su cliente». El análisis de datos permite conocer a los prestatarios mejor que nunca y predecir si devolverán sus préstamos de forma más certera que si únicamente se estudia su historial crediticio. Este sistema depende de algoritmos que analizan datos de forma compleja y automatizada (y hasta sus defensores tienen dudas sobre el proceso)[29].

Consciente de esto, IBM ha creado un súper-ordenador, llamado Watson que está siendo probado en el sector de la asistencia sanitaria.

---

28 Steve LOHR. «If Algorithms Know All, How Much Should Humans Help?» *The New York Times News Services* (6 abril 2015).

29 Ídem.

Es capaz de leer miles de documentos por segundo, una velocidad imposible de igualar por los humanos, buscando correlaciones y otras ideas importantes. El programa Watson Paths, concretamente, permite que los médicos vean las pruebas y el sendero de deducciones que ha seguido el ordenador para lanzar sus conclusiones (por ejemplo, cómo ha concluido el diagnóstico médico que realiza). Este experimento pionero intenta dar una solución a esta falta de supervisión humana de los resultados de los algoritmos, a la falta de una traducción máquina-humano, que sin duda seguirá avanzando a medida que la ciencia del *big data* progrese. El ordenador Watson no es la única iniciativa en este sentido. Algunas empresas aseguran que sus empleados revisan las recomendaciones que hacen sus ordenadores, «aunque es raro que rechacen lo que dictan los algoritmos» [30].

Hasta la fecha, la realidad de las organizaciones que se basan en la ciencia de datos para tomar sus decisiones es que muy pocas veces se revisan. Además, también existen partidarios de no concederle al humano un poder de veto sobre las decisiones tomadas de forma analítica a través de algoritmos. Afirman que esto introduciría un sesgo humano en un sistema en el que una de sus virtudes es, precisamente, que promete decisiones basadas en datos y no en la intuición o la arbitrariedad. Así los resultados proporcionados serán mejores.

Ante esta dicotomía, una posible solución ya apuntada por los analistas es programar o retocar los algoritmos de forma que otorguen una mayor protección a las personas, y que así el riesgo de tomar una decisión equivocada sobre una persona concreta disminuya. De este modo, «el objetivo no es necesariamente que un humano supervise el resultado *a posteriori*, sino mejorar la calidad de la clasificación de los individuos *a priori*» [31].

Por ejemplo, continuando con nuestra mención al sector bancario, esto podría traducirse en que los algoritmos que utilicen las empresas se ajusten para que la probabilidad de que un individuo sea asociado a un perfil de dudoso pago sea menor.

La nueva Propuesta de Reglamento de Protección de Datos (norma que modificará la actual regulación sobre protección de datos en

---

30 Ídem.
31 Ídem.

el ámbito europeo) también ha querido tener en cuenta esta problemática e introduce una disposición por la que prohíbe tomar decisiones trascendentales para una persona sobre la única base de un análisis automático de datos. Este punto será analizado más adelante en el Capítulo V.

Tras haber revisado los riesgos que el *big data* impone en relación a la toma de decisiones erróneas basadas en relaciones espurias, y el riesgo de la toma de decisiones automatizadas sin supervisión humana, queda por analizar el riesgo que el *big data* entraña sobre la privacidad y la protección de datos de los individuos. El resto del trabajo se centrará en este problema, tratado desde un punto de vista jurídico.

CAPÍTULO II. ¿QUÉ SE ENTIENDE POR DATOS DE CARÁCTER PERSONAL?

2.1 DEFINICIÓN

Se entiende por dato de carácter personal «cualquier información concerniente a personas físicas identificadas o identificables». Una persona es identificable cuando su identidad pueda determinarse, directa o indirectamente, mediante cualquier información referida a su identidad física, fisiológica, psíquica, económica, cultural o social, salvo que dicha identificación requiera actividades o plazos desproporcionados[32].

Los datos de carácter personal no se limitan únicamente a nombres y apellidos, sino que son una lista amplia y abierta, que va creciendo, y que incluye datos como nuestra voz, número de la Seguridad Social, nuestra dirección o datos económicos. Pero también son datos de carácter personal nuestros «*likes*» en Facebook, nuestro ADN o nuestra forma de caminar. Ni siquiera nosotros mismos somos conscientes de las formas en las que nuestro propio día a día nos hace identificables.

Así por ejemplo, aunque no nos hayamos registrado en un sitio web, éste puede utilizar técnicas analíticas para rastrear las huellas digitales que nuestras actividades han ido dejando hasta terminar identificándonos.

Mención señalada merecen los llamados datos especialmente protegidos, puesto que son aquellos datos que, de divulgarse de manera indebida, podrían afectar a la esfera más íntima del ser humano, tales

---

[32] Definición dada por la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (artículo 3), y su Reglamento de desarrollo (artículo 5), partiendo de la definición que aporta la Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (artículo 2).

como ideología, afiliación sindical, religión, creencias, origen racial o étnico, salud y orientación sexual. Estos datos requieren un nivel de protección mayor y la Ley les reserva un tratamiento especial.

Recuperemos la definición de dato de carácter personal:

> «cualquier información concerniente a personas físicas identificadas o identificables».

De esta definición pueden extraerse varias de sus características principales[33].

(i) «CUALQUIER INFORMACIÓN»

Tanto las normas comunitarias de protección de datos, como las nacionales abogan por un concepto amplio de protección de datos.

Desde el punto de vista del *contenido*, ya hemos adelantado que la lista de lo que constituye un dato de carácter personal es abierta y va creciendo a medida que la sociedad y la tecnología evolucionan.

Desde el punto de vista de la *naturaleza* de la información, el concepto de dato personal incluye tanto información objetiva (por ejemplo, la presencia de determinada sustancia en la sangre), como las evaluaciones subjetivas. De hecho, la información subjetiva constituye una parte considerable de los datos personales tratados en sectores como la banca («Fulano es un prestatario fiable»), el asegurador («no se espera que Fulano muera pronto») o el laboral («Fulano es un buen trabajador y merece un ascenso»).

Para que esas informaciones se consideren datos personales no es necesario que sean verídicas o estén probadas. De hecho, las normas de protección de datos prevén la posibilidad de que la información sea incorrecta y confieren al interesado el derecho de acceder a esa información y de refutarla a través de los medios apropiados.

Por último, desde el punto de vista del *formato* o soporte en el que está contenida la información, los datos pueden aparecer de forma alfabética, fotográfica, sonora o cualquier otro.

---

[33] GRUPO DE TRABAJO DEL ARTÍCULO 29. «Opinion 4/2007 on the Concept of Personal Data» (2007).

Es especialmente relevante el hecho de que no es necesario que la información esté recogida en una base de datos o en un fichero estructurado. También la información desestructurada contenida en un texto libre en un documento electrónico puede calificarse como datos personales.

Ejemplo: Banca telefónica

En las operaciones de banca telefónica, en las que la voz del cliente que da instrucciones al banco se graba en una cinta, las instrucciones grabadas deben ser consideradas como datos personales.

(ii) «PERSONA IDENTIFICADA O IDENTIFICABLE»

Se considera que una persona es *identificada* cuando la información disponible indica directamente a quién pertenece, sin necesidad de realizar una averiguación posterior. Por su parte, una persona es *identificable* cuando, aunque no haya sido identificada todavía, sea posible hacerlo.

En los casos en que, a primera vista, la información no permite singularizar a una persona determinada, ésta aún puede ser identificable, porque esa información puede ser combinada con otros datos, tanto si el responsable de su tratamiento tiene conocimiento de ellos como si no, que permitan distinguir a esa persona de otras.

En otras palabras, la posibilidad de identificar a una persona ya no equivale necesariamente a la capacidad de poder llegar a conocer su nombre y apellidos[34].

Así, por ejemplo, los ficheros informatizados de datos personales suelen asignar un identificador único a las personas registradas para evitar toda confusión entre dos personas incluidas en el fichero. También en internet, las herramientas de control de tráfico permiten identificar con facilidad el comportamiento de una máquina, por ejemplo a través de las denominadas *cookies*, y así, identificar al usuario que se encuentra detrás. Así pues, se unen las diferentes piezas que componen la personalidad del individuo con el fin de atribuirle determinadas de-

---

[34] GRUPO DE TRABAJO DEL ARTÍCULO 29. «Opinion 4/2007 on the Concept of Personal Data» (2007).

cisiones y poder así incluirle en una categoría, sobre la base de criterios socioeconómicos, psicológicos, o de otro tipo, como «mujer en el rango de edad 25-30 años residente en Madrid aficionada al deporte».

Por último, nuestra normativa introduce la idea de que no se considerará identificable a una persona si la identificación requiere plazos o actividades desproporcionados[35]. Esto significa que la mera e hipotética posibilidad de singularizar a un individuo no es suficiente. Se deben tener en cuenta todos los factores en lid, tales como el coste de realizar la identificación, la finalidad del tratamiento de esos datos, el beneficio que espera obtener el responsable del tratamiento, etc. Asimismo es de especial relevancia el grado de avance tecnológico en el momento del tratamiento del dato, y en el futuro. En todo caso, ahondaremos más en el concepto de «persona identificable» en el Capítulo IV, al tratar sobre la anonimización de datos.

2.2 MARCO JURÍDICO DE LA PROTECCIÓN DE DATOS

Los medios de comunicación actuales, especialmente las redes sociales, facilitan el intercambio de información y el acceso por parte de terceros a imágenes y datos sobre nuestros gustos, preferencias, hábitos, nuestras relaciones y, en general, aspectos de nuestra vida privada que deben ser garantizados y tutelados en virtud del derecho a la protección de datos[36].

Este derecho a la protección de datos ha sido definido por nuestros tribunales, estableciendo que «consiste en un poder de disposición y de control sobre los datos personales que faculta a la persona para decidir cuáles de estos datos proporcionar a un tercero, sea el

---

35 Considerando 26 de la Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos; y artículo 5 del Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de protección de datos de carácter personal.

36 Sor ARTEAGA JUÁREZ. «Las redes sociales digitales en la gestión y las políticas públicas. Capítulo 3: Implicaciones Legales en el uso de las redes sociales por las administraciones públicas». *Escola d'Administració Pública de Catalunya* (2013).

Estado o un particular, o cuáles puede este tercero recabar, permitiendo también al individuo saber quién posee esos datos personales y para qué, pudiendo oponerse a esa posesión o su uso. Su carácter de derecho fundamental le otorga determinadas características, como la de ser irrenunciable y el hecho de prevalecer sobre otros derechos no fundamentales» (STC 292/2000, de 30 de noviembre de 2000).

En el ámbito europeo, la protección de datos está configurada nada menos que como un derecho fundamental, recogido en el artículo 16 del Tratado de Funcionamiento de la Unión Europea y el artículo 8 de la Carta de los Derechos Fundamentales de la Unión Europea.

Ya en 1980 se publicaban las Directrices de Privacidad de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), y en 1981 el Consejo de Europa adoptó el Convenio n.º 108 para la Protección de las Personas con Respecto al Tratamiento Automatizado de Datos de Carácter Personal, que es el único instrumento internacional vinculante sobre protección de datos[37].

Estas normas dieron lugar a la principal norma actual en la materia, la Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (en adelante la «Directiva de Protección de Datos» o simplemente la «Directiva»). Las disposiciones de la Directiva pretenden regular el equilibro que debe existir entre la protección de la vida privada de las personas físicas y la libre circulación de datos personales dentro de la Unión Europea.

En el ordenamiento jurídico español, la protección de datos personales está desarrollada en la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal (en adelante, la «LOPD»), y su Reglamento de desarrollo, aprobado mediante Real Decreto 1720/2007, de 21 de diciembre (en adelante, el «Reglamento de Protección de Datos»). Asimismo, existen otras normas sectoriales (Ley 34/2002, de 11 de julio, de Servicios de la Sociedad de la Información y de Comercio Electrónico; Ley 9/2014, de 9 de mayo, General de Telecomunicaciones, etc.).

---

[37] Sitio web de la Agencia Española de Protección de Datos. Disponible en: http://www.agpd.es/portalwebAGPD/index-ides-idphp.php.

El objeto primero de esta protección, tal y como recoge la LOPD es «garantizar y proteger, en lo que concierne al tratamiento de los datos personales, las libertades públicas y los derechos fundamentales de las personas físicas, y especialmente de su intimidad personal y familiar». Para ello se basa en la diferenciación entre lo que se considera dato de carácter personal y los datos no personales, en el sentido ya analizado en el apartado anterior.

La legislación española en materia de protección de datos es aplicable en los siguientes casos[38]:

i. Cuando el tratamiento de los datos se realiza en territorio español en el marco de las actividades propias de un establecimiento del que sea titular el responsable del tratamiento de los datos.

ii. Cuando el responsable del tratamiento de los datos no está establecido dentro del territorio español, pero le es aplicable la legislación española conforme a las normas de Derecho Internacional Público.

iii. Cuando el responsable del tratamiento de los datos no está establecido en ningún país de la Unión Europea, pero en el tratamiento de los datos utiliza medios situados en territorio español, salvo que tales medios se utilicen únicamente con fines de tránsito.

Por el contrario, existen otros datos que se regulan por sus normas específicas, tales como los ficheros destinados a fines exclusivamente estadísticos, amparados por las normas sobre la función estadística pública; o los datos procedentes de imágenes y sonidos obtenidos por las Fuerzas y Cuerpos de Seguridad mediante videocámaras.

Por último, la normativa sobre protección de datos no será de aplicación a informaciones tales como los ficheros mantenidos por personas físicas en el ejercicio de actividades exclusivamente personales o domésticas, ni aquellos datos hechos anónimos de tal manera que ya no sea posible identificar al interesado[39]. Como ya se ha adelantado, en el Capítulo IV desarrollaremos el tratamiento de los datos anónimos.

---

38 Artículo 2 de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.

39 Considerando 26 de la Directiva 95/46/EC del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas

A estas normas hay que añadir los documentos publicados por el denominado Grupo de Trabajo del Artículo 29 (en adelante "GT 29"). Se trata de un órgano consultivo creado en virtud del artículo 29 de la Directiva de Protección de Datos, que integra a todas las autoridades de protección de datos de los Estados miembros. Así, la Agencia Española de Protección de datos, la autoridad española en la materia, forma parte del GT 29 desde su inicio. Los documentos que publica (dictámenes, documentos de trabajo, informes, etc.) no son jurídicamente vinculantes, pero tienen un importantísimo valor doctrinal y son frecuentemente citados por legisladores y tribunales nacionales y europeos[40]. En este sentido, el GT 29 se ha pronunciado en materias tan amplias como el consentimiento, la anonimización, el internet de las cosas o el *big data*, todo ello en relación a la aplicación de las normas sobre protección de datos.

2.3 IMPACTO DEL *BIG DATA* EN LA NORMATIVA DE PROTECCIÓN DE DATOS

El *big data* puede representar un reto para diferentes cuerpos normativos, tales como la protección de datos, la prohibición de la discriminación, la responsabilidad civil, el derecho de la competencia, los derechos de propiedad intelectual, etc. Este estudio está centrado en los problemas sobre la privacidad y la protección de datos.

Como ya hemos señalado, la normativa de protección de datos se aplica cuando la información de las personas físicas hace que éstas sean identificadas o identificables. *A sensu contrario*, cuando los datos no hacen identificable a una persona, no se aplica esta regulación. Es decir, cuando los datos se hacen anónimos a través de técnicas de anonimización, se convierten en datos no personales, y la privacidad de los individuos queda protegida, de modo que no es necesario aplicar ninguna norma sobre protección de datos. Junto con la anonimización, nuestra norma trata, además, de lo que denomina el proceso de disociación, que permite crear datos pseudónimos, una categoría

---

físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

40 Sitio web de la Agencia Española de Protección de Datos. Disponible en: http://www.agpd.es/portalwebAGPD/index-ides-idphp.php.

de datos que, sin ser anónimos, tienen más garantías para la privacidad que los datos puramente personales.

El big data desafía las normas de protección de datos al facilitar la re-identificación de los sujetos, ya no solo a partir de los datos pseudónimos, sino también a partir de datos que considerábamos anónimos. Es decir, las técnicas de anonimización ya no siempre son suficientes con la llegada del *big data*. Esto supone volver al debate de base de qué datos son personales y cuáles no personales. Todos estos conceptos serán tratados en detalle más adelante.

En conclusión, el *big data* amenaza la normativa de protección de datos, debido a diversos motivos:

i. *La normativa no se encuentra adaptada al nuevo entorno tecnológico*. La publicación de la Directiva de Protección de Datos, que es la norma de la que parten las demás normas de protección de datos de la Unión Europea, data de 1995; esto es, antes incluso de la generalización de Internet, y de fenómenos como la ubicuidad de los dispositivos móviles y de geolocalización o las redes sociales, por no mencionar las tecnologías disruptivas como el *big data* o el *cloud computing*.

ii. *El principio de «minimización de datos» no se cumple en la práctica*. Este principio implica que los datos recopilados no deben ser excesivos, sino que debe recopilarse solo la cantidad mínima necesaria para el fin para el que se recogen. Pues bien, en muy pocas ocasiones las autoridades de protección de datos obligan de forma efectiva a las empresas a rediseñar sus procesos para minimizar los datos recabados.

    Es más, el principio de minimización de datos se contrapone contra la misma lógica del *big data*. Los nuevos modelos analíticos se basan precisamente en el estudio de cantidades masivas de datos sin los cuales no podría extraerse el conocimiento que nos permite el *big data*.

iii. *La normativa confía demasiado en el consentimiento informado* del individuo para recopilar y tratar sus datos de carácter personal[41]. Esto supone un problema, dada la experiencia

---

41 En este sentido se pronuncian algunos de los referentes más importantes en la materia. Por todos, ver los trabajos de Ira S. RUBINSTEIN. «Big data: The

de que la gran mayoría de los individuos no lee las políticas de privacidad antes de prestar su consentimiento; y aquellos que lo hacen no las comprenden. Así, otorgar el consentimiento es, con carácter general, un ejercicio vacío.

iv. *La anonimización ha demostrado tener limitaciones*. Si bien se presentaba como la mejor solución para tratar los datos protegiendo la privacidad de los sujetos, en los últimos años se han dado numerosos casos de reidentificación de bases de datos que habían sido anonimizadas.

Cada vez se hace más sencillo reidentificar a los sujetos, ya no solo a través del análisis de distintas fuentes que contienen datos personales parciales de una persona, sino a través de datos no personales. Esto supone un debilitamiento de la anonimización como medida para asegurar la privacidad durante el tratamiento de datos.

v. *El big data aumenta el riesgo relacionado con la toma de decisiones de forma automática*. Esto hace que decisiones trascendentales para nuestra vida, tales como calcular nuestro riesgo crediticio, queden sujetas a algoritmos ejecutados de forma automática. El problema surge cuando los datos que son analizados por medio de los algoritmos no son precisos o veraces, pero los individuos no tienen incentivos para corregirlos porque no son conscientes de que están siendo utilizados para tomar decisiones que les afectan.

RETOS DE LAS NORMAS DE PROTECCIÓN DE DATOS

- No adaptación al entorno tecnológico
- No implementación efectiva del principo de minimización de datos
- Excesiva confianza en el consentimiento informado
- La anonimización ha demostrado tener limitaciones
- Riesgo de toma de decisiones de forma automática

End Of Privacy Or A New Beginning?». *International Privacy Law*, Vol. 3, n.º 2 (2013); Fred H. CATE. «The failure of Fair Information Practice Principles». *Consumer Protection In The Age Of Information Economy* (2006).

2.4 HACIA DÓNDE CAMINA EL FUTURO DE LA PROTECCIÓN DE DATOS

Consciente de esta realidad, la Unión Europea se encuentra inmersa en la renovación de la normativa de protección de datos, y así ha publicado una Propuesta de Reglamento Europeo sobre Protección de Datos (2012) (en adelante la «Propuesta de Reglamento de Protección de Datos» o simplemente la «Propuesta»).

En términos generales, la Propuesta mantiene los objetivos y los principios actuales sobre protección de datos, aunque introduce algunos nuevos cambios sobre cómo aplicarlos. Entre los cambios más notables se encuentran nuevos derechos de privacidad, concretamente el derecho al olvido y la portabilidad de datos. Asimismo, propone endurecer los deberes de transparencia e incluir el concepto de privacidad por defecto y privacidad desde el diseño. Y especialmente relevante es que la Propuesta continúa confiando en el consentimiento informado como primera herramienta para proteger los datos y la privacidad de los ciudadanos europeos; de hecho, se refuerza la importancia del consentimiento[42].

El hecho de que se haya escogido la figura del Reglamento para actualizar la normativa europea de protección de datos, que hasta ahora está regulada en forma de Directiva, no es casualidad. Tanto los Reglamentos como las Directivas son normas vinculantes, pero presentan diferencias. Una Directiva es una norma que vincula a todo Estado miembro en lo relativo al resultado que hay que alcanzar, pero que permite a las instancias nacionales libertad para elegir los medios y las formas de alcanzar dicho resultado. En la práctica esto implica que la Directiva necesita ser traspuesta a la normativa nacional, de forma tal que hasta que la norma nacional no se promulgue, los postulados de la Directiva no son de aplicación. Asimismo, cada Estado miembro tiene un margen de discrecionalidad a la hora de trasponer la norma, lo que ha dado lugar a una aplicación diferente de los postulados de la Directiva en cada uno de los 28 Estados miembros.

Por contra, un Reglamento tiene aplicabilidad directa en todos los Estados miembros desde el momento de su publicación. Esto en la

---

42 Juan Fernando LÓPEZ AGUILAR. «Por fin una ley europea de protección de datos (I)». *El Huffington Post Unión Europea* (24 de octubre de 2013).

«práctica significa que cualquier ciudadano de cualquier Estado miembro puede acudir a los tribunales nacionales para reclamar el cumplimiento del Reglamento europeo, a pesar de que no haya norma nacional. Además, de esta forma se pretende homogeneizar las distintas legislaciones nacionales de los Estados miembros.

Así las cosas, en España esto significa que la entrada en vigor del Reglamento, previsiblemente en 2018, conllevaría desplazar a la norma actual, la LOPD, en las materias en las que difieran.

Pese a todo, la aprobación de este nuevo marco normativo se ha ido retrasando por diversas razones: la divergencia de posturas entre los Estados, un número record de enmiendas, o el *lobby* ejercido por las multinacionales y por Estados Unidos. Curiosamente, el caso Snowden, la NSA, el programa PRISM y los últimos escándalos de espionaje están sirviendo de impulso a la necesidad de una normativa europea sólida en el ámbito de la protección de datos[43].

No obstante lo dicho, ¿es suficiente este futuro Reglamento para hacer frente a los retos que trae el *big data*? Intentaremos dar respuesta a esta pregunta a lo largo de la exposición.

2.5 FASES EN LAS QUE SE DESARROLLA EL *BIG DATA*

Antes de comenzar a tratar aspectos legales concretos, veo necesario hacer una precisión sobre cómo se aplican las tecnologías *big data* en la práctica. Esta idea parece encontrarse ya en la argumentación de algunos autores (véanse los razonamientos de Ira Rubinstein) y también han sido acogidos por otros juristas españoles como Javier Aparicio Salom[44].

Se trata de la idea de que el *big data* se desarrolla y se aplica en dos fases diferenciadas, y cada una de estas fases tiene consecuencias diferentes sobre la privacidad y la protección de datos. En muchas ocasiones ambas fases y sus riesgos se confunden, dando lugar a conclusiones erróneas sobre las posibles medidas para proteger a los individuos.

---

43 Ignacio BRUNA. «El futuro Reglamento General de Protección de Datos de la Unión Europea (o eso esperamos)». *Blogs KPMG, Ciberseguridad* (17 de noviembre de 2013).

44 Discusiones mantenidas en relación a una versión preliminar de este trabajo.

(i) PRIMERA FASE DEL *BIG DATA*

La primera fase comprende la recolección de información sobre los individuos y la aplicación de algoritmos y medios automatizados para observar correlaciones. De este modo se pueden extraer conclusiones sobre cómo afecta una circunstancia específica en el comportamiento del individuo. Por ejemplo, cómo afecta un determinado tratamiento médico a los síntomas que muestra el individuo en las sucesivas revisiones; o cómo afecta la edad, el nivel de formación y el lugar de residencia de una persona sobre la tasa de impago crediticio en una sucursal bancaria.

En esta fase es importante que los datos de cada persona puedan asociarse como pertenecientes a la misma persona. Pero esto no implica que podamos saber quién es esa persona. Por ejemplo, se puede utilizar un código (un número aleatorio) que nos permita relacionar todas las variables que pertenecen a una misma persona. Así podremos identificar qué variables coinciden o difieren entre los grupos de personas para poder obtener nuestras conclusiones. Por ejemplo, podremos analizar si hábitos de vida como la meditación o las actividades al aire libre influyen en el desarrollo de una enfermedad y en qué medida lo hacen, o si una enfermedad afecta con mayor probabilidad a individuos de raza caucásica o que posean una genética determinada.

Aquí cobran mucha importancia las técnicas de pseudonimización (que la LOPD llama disociación). Y ello porque el organismo que lleva a cabo una investigación puede recibir estos datos pseudónimos en los que no sabe a qué persona se refiere cada conjunto de datos, pero sí sabe que se refieren a la misma persona, puesto que está asociada a un código. Es decir, únicamente es necesario tener la certeza de que los datos han sido asociados de forma correcta, o lo que es lo mismo, que se garantice la calidad de los datos.

Es razonable pensar que, por ejemplo, un equipo de investigación, no tenga incentivos para invertir esfuerzos en identificar a estas personas, aunque en todo caso, habrán de preverse salvaguardas que protejan los datos. También serán necesarias medidas de seguridad efectivas que aseguren que dichos datos no serán accesibles por un tercero no autorizado (que sí podría tener incentivos para identificar a las personas).

En este sentido, las técnicas que persiguen proteger la privacidad de las personas disminuyendo la calidad de los datos pueden ser muy nocivas para la investigación. Este argumento será desarrollado con mucho más detalle en el Capítulo IV, dedicado a la anonimización de datos.

De hecho, ya durante la redacción de la actual LOPD se discutió la posibilidad de crear reglas de uso de estos códigos que se generasen a través de algoritmos de forma que impidieran un retorno, es decir, que impidieran que el código pudiera volverse a asociar a la identidad de la persona, todo ello en relación con información clínica para facilitar la investigación médica. De este modo, una vez que una entidad (que podría ser, por ejemplo, un equipo de investigación) obtiene datos sobre un conjunto de personas, los nuevos datos que se generen sobre esa misma persona en momentos posteriores podrán ser asociados de nuevo a esta misma persona a través del código. No obstante, esta propuesta no fue incluida en la redacción final de la norma[45].

El objetivo de esta fase es encontrar correlaciones entre los grupos de datos, manteniendo la confianza de que estos datos no están contaminados.

En este sentido, Javier Aparicio afirma[46] que en esta primera fase estaríamos ante lo que el Derecho clásico romano bautizó como el principio *ius usus inocui*, esto es, el derecho a aprovechar la cosa ajena usándola por razón de utilidad, sin que el dueño sufra perjuicio[47]. Tradicionalmente estos aprovechamientos se asientan en la tolerancia del dueño.

De forma análoga, el aprovechamiento de los datos y del conocimiento que de ellos se desprende podría llevarse a cabo bajo este principio jurídico mientras la finalidad del uso sea inocua para el individuo (como sería la elaboración de estudios o investigaciones).

En su caso, la falta de tolerancia puede ser manifestada por la persona a través de los denominados sistemas *opt-out*, que son aquellos en los que la persona manifiesta su voluntad de no ser parte del pro-

---

45 Ídem.

46 Ídem.

47 Definición de *ius usus inocui* utilizada en la Sentencia de la Audiencia Provincial de Murcia de 20 de febrero de 1960.

ceso, de cancelar los datos (en contraposición a los sistemas *opt-in* en los que el individuo debe manifestar su consentimiento para que los datos sean tratados, por ejemplo, marcando una casilla). El concepto de sistemas *opt-in* y *opt-out* será revisado en el Capítulo III, dedicado al consentimiento; basta ahora con enunciar esta breve definición.

En conclusión, esta primera fase implica una recolección de datos y su procesamiento automatizado para crear un modelo que nos muestre las correlaciones existentes entre las variables.

(ii) SEGUNDA FASE DEL *BIG DATA*

La segunda fase consiste en aplicar el modelo a una persona determinada. Así, los datos de una persona serán recogidos y procesados para que, de acuerdo con el modelo que hemos creado, podamos obtener conclusiones sobre ella. En este momento es necesario obtener el consentimiento informado de la persona.

Por ejemplo, supongamos que un nuevo paciente acude a un hospital. Sus datos forman un contexto, que son las circunstancias del paciente (su edad, sexo, patologías detectadas, alergias, etc). Estas variables son analizadas a través de nuestro modelo, de modo que podremos obtener conclusiones como qué medicación es la más adecuada para este paciente concreto.

Esta fase comporta un riesgo ético mucho mayor, en la medida en que ahora se toman decisiones que afectan a los individuos. El uso que se haga de las conclusiones puede ser beneficioso, pero también puede utilizarse de modo discriminatorio. Estos conceptos también van a ser desarrollados en los próximos Capítulos.

En concreto, en las próximas páginas nos aproximaremos al concepto de consentimiento y los problemas prácticos que se observan en la actualidad con la aparición de nuevas tecnologías y las técnicas de *big data*, que han modificado las reglas del juego hasta ahora conocidas.

De modo esquemático, el funcionamiento de la normativa de protección de datos es el siguiente. Los datos de carácter personal quedan regulados por la normativa de protección de datos que ha sido expuesta. El consentimiento es el instrumento más importante que plantea esta normativa para legitimar la recogida y el tratamiento de

datos personales. Por el contrario, los datos anónimos no se pueden asociar con una persona, de modo que no se consideran datos de carácter personal y su tratamiento queda libre de las disposiciones de la normativa de protección de datos. Los dos capítulos siguientes tratan estos dos instrumentos.

CAPÍTULO III. EL CONSENTIMIENTO

El artículo 7 de la Directiva (o el artículo 6 LOPD) establece los fundamentos jurídicos que permiten llevar a cabo el tratamiento de los datos de carácter personal. Así, el tratamiento de datos se puede realizar siempre que se cumpla alguna de las siguientes condiciones, enumeradas de forma somera:

a) contar con el consentimiento inequívoco del interesado;

b) que el tratamiento sea necesario para ejecutar un contrato en el que el interesado sea parte;

c) que el tratamiento sea necesario para cumplir una obligación legal;

d) que el tratamiento sea necesario para proteger un interés vital del interesado;

e) que el tratamiento sea necesario para el cumplimiento de una misión de interés público o inherente al ejercicio del poder público

f) que el tratamiento sea necesario para la satisfacción del interés legítimo perseguido por el responsable del tratamiento o por el tercero a quien se comuniquen los datos, siempre que no prevalezca el interés o los derechos y libertades fundamentales del interesado

En este capítulo desarrollaremos el concepto de consentimiento, en tanto instrumento fundamental de la protección de datos.

El consentimiento del interesado ha sido siempre un concepto clave de la protección de datos, como un medio que permite respetar la autonomía de los individuos sobre la toma de sus decisiones. Así, cuando la anonimización no se puede alcanzar en la práctica, o cuando no resulta útil, el consentimiento es la solución más utilizada.

El consentimiento informado es el corolario natural de la idea de que privacidad implica control sobre nuestra información. Sin embargo, no siempre está claro cuándo es necesario y cuáles son las condiciones que deben cumplirse para que este consentimiento sea válido.

El problema se ha agravado a medida que la actividad de tratamiento de datos personales ha ido adquiriendo una importancia creciente en la sociedad actual, tanto en entornos en línea como fuera de línea. Esto se da especialmente en el contexto de las técnicas de *big data*, en el que, por si esto no fuera suficiente, en el tratamiento de los datos participan a menudo diferentes países.

El esquema que sigue este capítulo es sencillo. En primer lugar estableceremos la definición del consentimiento en virtud de las normas de protección de datos, y llevaremos a cabo un análisis de los requisitos que debe tener para ser válido. Posteriormente veremos la influencia que tiene el *big data* sobre la forma de prestar el consentimiento, y los problemas que surgen en la práctica. Esto es de especial trascendencia porque, como veremos, los nuevos retos tecnológicos hacen que en la práctica las personas no presten un consentimiento válido para que las organizaciones recaben y traten sus datos, de modo que el instrumento principal de la protección de datos queda en entredicho.

3.1 MARCO JURÍDICO Y CARACTERÍSTICAS DEL CONSENTIMIENTO

La función del consentimiento ha sido reconocida expresamente en la Carta de los Derechos Fundamentales de la Unión Europea en relación con la protección de datos personales, cuando establece que los datos personales pueden ser tratados «sobre la base del consentimiento de la persona afectada o en virtud de otro fundamento legítimo previsto por la ley» (artículo 8.2). Por lo tanto, el consentimiento se reconoce como un aspecto esencial del derecho fundamental a la protección de datos de carácter personal.

En el ámbito de la UE, la Directiva de Protección de Datos define el consentimiento como:

> Artículo 2: «h) «consentimiento del interesado»: toda manifestación de voluntad, *libre*, *específica* e *informada*, mediante la que el interesado consienta el tratamiento de datos personales que le conciernan». (Cursiva añadida).

Por su parte, el artículo 7(a) añade el requisito de que el consentimiento debe ser inequívoco:

> Artículo 7: «Los Estados miembros dispondrán que el tratamiento de datos personales sólo pueda efectuarse si: (a) el interesado ha dado su *consentimiento de forma inequívoca* (...)». (Cursiva añadida).

Y el artículo 8 añade la palabra «explícito» cuando el tratamiento se refiere a categorías especiales de datos:

> Artículo 8: «Tratamiento de categorías especiales de datos
>
> 1. Los Estados miembros prohibirán el tratamiento de datos personales que revelen el origen racial o étnico, las opiniones políticas, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, así como el tratamiento de los datos relativos a la salud o a la sexualidad.
>
> 2. Lo dispuesto en el apartado 1 no se aplicará cuando: a) el interesado haya dado su *consentimiento explícito* a dicho tratamiento (...)». (Cursiva añadida).

España ha traspuesto la Directiva de forma tal que nuestra Ley Orgánica de Protección de Datos se refiere al consentimiento en los mismos términos. Es decir, de igual modo, exige que el consentimiento sea libre, inequívoco, específico e informado[48]. La única diferencia radica en que la LOPD es más estricta que la Directiva en lo que se refiere a las categorías especiales de datos: mantiene el requisito de que el consentimiento haya de recabarse de forma expresa, a lo que añade que también deba ser por escrito[49]. Las versiones preliminares de la Directiva también exigían que el consentimiento fuera por escrito, pero esto se eliminó en la versión final, de modo que abre a la puerta a que en otros países este consentimiento expreso pueda obtenerse de forma escrita o verbal.

---

48 Artículos 3 h) y 6 de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.

49 Artículo 7 de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.

Estas definiciones contienen diferentes elementos clave, que han sido analizados por el GT 29 en su Dictamen sobre la definición de consentimiento. Trataremos estas características a continuación con ejemplos prácticos sobre la base del mencionado Dictamen[50].

(i) «TODA MANIFESTACIÓN DE VOLUNTAD MEDIANTE LA QUE»

Los términos «manifestación» y «mediante la que» apuntan a que es necesaria una acción (frente a una situación en la que el consentimiento podría deducirse de la falta de acción).

Ejemplo: Paneles publicitarios de *bluetooth* o wifi

Un instrumento publicitario en expansión consiste en paneles que envían mensajes en los que se pide que se establezca una conexión *bluetooth* o wifi para enviar anuncios a las personas que pasan por las inmediaciones. Los mensajes se envían a las personas que tienen activados los mecanismos *bluetooth* o wifi en el dispositivo móvil. La mera activación de las funciones *bluetooth* o wifi no constituye un consentimiento válido (ya que pueden activarse para otros fines).

Por contra, cuando alguien está informado de este servicio y se aproxima a pocos centímetros del panel con el móvil, se produce, por lo general, una manifestación de voluntad: así se comprueba cuáles son las personas realmente interesadas en recibir los anuncios. El GT 29 considera que únicamente estas personas han manifestado su consentimiento, y sólo ellas deben recibir los mensajes por teléfono.

En 2012, la empresa inglesa Renew incorporó en diversas papeleras en Londres un dispositivo que permitía leer la dirección MAC de los *smartphones* de los transeúntes que caminaban por la ciudad con la función wifi activada. A través del software de Renew, eran capaces de identificar a esos mismos usuarios los días sucesivos, y así llegar a conocer sus rutas ha-

---

50 GRUPO DE TRABAJO DEL ARTÍCULO 29. «Opinion 15/2011 on the Definition of Consent» (2011).

bituales, la velocidad al caminar, etc., abriendo la posibilidad de mostrarles publicidad relevante basándose en estos datos. El problema de este sistema es que las antenas de wifi son capaces de identificar el número MAC asociado a un teléfono móvil, que es un dispositivo muy estrechamente ligado a cada dueño (no es normal compartirlo entre varios usuarios), de modo que la información que de ellos se obtiene está directamente referida al usuario (no obstante, no está claro si el número MAC es un dato de carácter personal en sentido legal). La posibilidad de que las papeleras pudieran identificar y enviar publicidad a los transeúntes sin haber prestado consentimiento para esta recogida y tratamiento de datos generó enormes protestas. En el primer mes de su implementación, una docena de estas papeleras inteligentes ya había identificado a más de un millón de dispositivos, con picos diarios de hasta 10.000 personas en horas punta. Debido a las críticas que despertó el sistema, Renew anunciaba su retirada poco después.

El uso del número MAC es un caso típico de dato límite, pues no queda claro si se trata de un dato personal en el sentido de la ley, y en su caso, si el usuario presta el consentimiento para su tratamiento.

Es dudoso si la falta de actuación -o el comportamiento pasivo- podría interpretarse como consentimiento en circunstancias muy concretas en las que se dé un contexto totalmente inequívoco. El concepto de «manifestación» es amplio, de modo que abre un gran margen de interpretación. El GT 29 señala que, a su parecer, debe haber una acción, aunque recordemos que sus impresiones no son jurídicamente vinculantes.

(ii) «MANIFESTACIÓN DE VOLUNTAD LIBRE»

El consentimiento únicamente puede ser válido si el interesado puede elegir una opción real y no hay ningún riesgo de engaño, intimidación o consecuencias negativas significativas en caso de que no consienta.

En ocasiones, el consentimiento no se presta de forma libre. Con frecuencia, esto se debe a que entre la persona afectada y quien recoge y

trata los datos hay una relación de subordinación (como una relación laboral de empleador-trabajador). En otras ocasiones la falta de libertad puede deberse a algún tipo de coacción social, financiera o psicológica.

(iii) «MANIFESTACIÓN DE VOLUNTAD ESPECÍFICA»

Para ser válido, el consentimiento debe ser específico. Así, el GT 29 considera que el consentimiento indiscriminado sin especificar la finalidad exacta del tratamiento no debería ser admisible.

Para ser específico, el consentimiento debe ser comprensible; es decir, referirse de manera clara y precisa al alcance y las consecuencias del tratamiento de datos. No puede referirse a un conjunto indefinido de actividades de tratamiento. Esto implica saber *a priori* cuáles son los datos recogidos y los motivos del tratamiento.

Ejemplo: redes sociales

El acceso a los servicios de redes sociales suele estar sujeto a la autorización de diferentes tipos de tratamiento de datos personales.

Al usuario se le puede pedir su consentimiento para recibir publicidad comportamental antes de poder inscribirse en los servicios de una red social, sin más especificaciones ni opciones alternativas. Considerando la importancia que han adquirido las redes sociales, ciertas categorías de usuarios (como los adolescentes) aceptarán recibir esta publicidad para evitar el riesgo de ser excluidos de las interacciones sociales.

En este contexto, el GT 29 opina que el usuario debería estar en condiciones de dar su consentimiento libre y específico para recibir la publicidad personalizada, independientemente de su acceso al servicio de la red social.

Sin embargo, en la práctica, es frecuente que al usuario se le impida utilizar una aplicación si no da su consentimiento para la transmisión de sus datos al promotor de la aplicación para diversos motivos, incluida esta publicidad comportamental y la reventa de datos a terceros. Dado que la aplicación puede funcionar sin necesidad de transmitir ningún dato al promotor de la aplicación, el GT 29 propugna un consentimiento del usuario diferenciado para los diferentes fines.

> Podrían utilizarse diferentes mecanismos, como las ventanas desplegables, para que el usuario tuviera la posibilidad de seleccionar la finalidad para la que otorga su consentimiento (transmisión al promotor, servicios de valor añadido, publicidad personalizada, transmisión a terceros, etc.).
>
> En cambio, existen otras voces que consideran que enviar publicidad comportamental es una potestad del responsable del tratamiento, que debería encontrarse detallada en los términos y condiciones de uso de la plataforma. Cada individuo tendría, pues, la opción de acceder al servicio y recibir publicidad, o de no acceder al servicio, salvo que se tratase de un servicio esencial.

El carácter específico del consentimiento también significa que si los fines para los que los datos son tratados por el responsable cambian en algún momento, el usuario deberá ser informado y estar en condiciones de dar su consentimiento para el nuevo tratamiento de datos. La información que se facilite debería mencionar las consecuencias del rechazo de los cambios propuestos.

Cuando hablamos de *big data* es especialmente relevante el hecho de que el consentimiento debe aplicarse a un contexto determinado, así como el hecho de que si el fin para el que se utilizarán los datos cambia, podría ser necesario volver a recabar el consentimiento.

Y ello porque, precisamente, el valor del *big data* reside en que la nueva información que se crea permite dar nuevos usos a los datos. Es precisamente en estos usos secundarios donde reside el potencial del *big data*. Esta forma de concebir el consentimiento obligaría a que cada vez que se descubra un nuevo uso para los datos, el responsable debería volver a pedir el consentimiento a cada uno de los individuos cuyos datos estén siendo tratados por segunda vez. Esto, en muchas ocasiones, podrá ser técnicamente inviable, por no decir que las empresas no podrían asumir los costes.

(iv) «MANIFESTACIÓN DE VOLUNTAD INFORMADA»

Para ser válido, el consentimiento debe estar informado. Esto implica que toda la información necesaria debe suministrarse en el momento en que se solicita el consentimiento, de forma clara y compren-

sible, y debe abarcar todas las cuestiones pertinentes. En principio, debe abarcar las informaciones enumeradas en el artículo 10 de la Directiva, pero también depende del momento y las circunstancias en que se solicite el consentimiento.

El consentimiento como manifestación de voluntad informada es especialmente importante en el contexto de las transmisiones de datos personales a terceros países, en la medida en que exige que el interesado esté informado acerca del riesgo de que sus datos se transfieran un país que carece de la protección adecuada.

En concreto, en España, la información que debe prestarse para obtener el consentimiento del interesado debe cumplir con los estrictos requisitos del artículo 5 de la LOPD.

Entre éstos, se exige que se dé información «de modo expreso, preciso e inequívoco» sobre la misma existencia del fichero y su finalidad (lo cual vuelve a resaltar las limitaciones que esto crea en el ámbito del *big data*, donde en muchas ocasiones no se conoce la finalidad del tratamiento de antemano); sobre el carácter obligatorio o facultativo de dar la información que se pide; de la posibilidad de ejercitar los denominados derechos ARCO (acceso, rectificación cancelación y oposición); y de la identidad del responsable del tratamiento de los datos[51].

(v) CONSENTIMIENTO INEQUÍVOCO

Como adelantábamos, de acuerdo con la Directiva, el consentimiento debe ser también inequívoco. Es decir, el procedimiento por el que se presta el consentimiento no debe dejar ningún lugar a dudas sobre la intención del interesado de dar su consentimiento. Si existe alguna duda sobre la intención del sujeto se producirá una situación equívoca.

Este requisito obliga a los responsables del tratamiento a crear procedimientos rigurosos para que las personas den su consentimiento. Se trata de, o bien buscar un consentimiento expreso, o bien basarse en procedimientos que permitan que las personas manifiesten un claro consentimiento deducible.

---

[51] Artículo 5 de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.

El GT 29 y el Supervisor Europeo de Protección de Datos (SEPD), han declarado en sus aportaciones a los debates sobre el nuevo marco de protección de datos que:

> «no siempre es fácil determinar qué constituye un consentimiento verdadero e inequívoco. Determinados responsables del tratamiento de datos explotan esta incertidumbre recurriendo a métodos que excluyen toda posibilidad de dar un consentimiento verdadero e inequívoco» [52].

Ejemplo: juego en línea

Imaginemos la situación en la que el proveedor de un juego en línea exige a los jugadores que faciliten su edad, nombre y dirección antes de participar en el juego (para realizar una distribución de jugadores por edades y direcciones). El sitio web contiene un anuncio, accesible a través de un enlace (aunque el acceso al anuncio no es para participar en el juego), que indica que al utilizar el sitio web, y por tanto al facilitar información, los jugadores consienten en que sus datos sean tratados para que el proveedor de juegos en línea y otros terceros les envíen información comercial.

En opinión del GT 29, el acceso y la participación en el juego no equivalen a dar un consentimiento inequívoco para el ulterior tratamiento de la información personal con fines distintos de la participación en el juego. Este tipo de comportamiento no constituye una manifestación inequívoca del deseo de la persona de que sus datos se utilicen para fines comerciales.

Ejemplo: parámetros de privacidad por defecto

Los parámetros por defecto de una red social, a los que los usuarios no acceden necesariamente al utilizarla, permiten, por ejemplo, que la totalidad de la categoría «amigos de amigos» pueda ver toda la información personal del usuario.

---

[52] Dictamen del Supervisor Europeo de Protección de datos, de 14 de enero de 2011, sobre la Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones — «Un enfoque global de la protección de los datos personales en la Unión Europea».

Los usuarios que no desean que su información sea vista por los amigos de amigos tienen que pulsar un botón. El responsable del fichero considera que si se abstienen de actuar o no pulsan el botón han consentido en que se puedan ver sus datos. Sin embargo, es muy cuestionable que no pulsar el botón signifique que por lo general las personas consienten en que su información pueda ser vista por todos los amigos de amigos.

Debido a la incertidumbre en cuanto a si la inacción significa consentimiento, el GT 29 considera que el hecho de no pulsar no puede considerarse consentimiento inequívoco.

(vi) CONSENTIMIENTO EXPRESO PARA CATEGORÍAS ESPECIALES DE DATOS

Los datos de naturaleza especialmente sensible requieren que el consentimiento del interesado sea expreso o explícito. Además, la norma española añade que debe ser prestado por escrito.

Ejemplo: historiales médicos electrónicos

Los datos médicos de los pacientes son un tipo de dato especialmente sensible, y la creación de historiales médicos electrónicos conlleva un tratamiento de estos datos. Así, es necesario que el consentimiento que presten los pacientes sea de forma explícita.

Esto implica que no son válidos aquellos mecanismos que se basan en entender que si el paciente no manifiesta explícitamente lo contrario es que ha prestado su consentimiento.

En el caso de los contextos en línea, el consentimiento expreso puede prestarse mediante el uso de la firma electrónica, pulsando un botón o una casilla en un sitio web, o enviando un correo electrónico de confirmación.

3.2 CONSENTIMIENTO *VS.* *BIG DATA*: RETOS ACTUALES

Como ya se ha mencionado, las nuevas tecnologías tales como los dispositivos móviles, los servicios de localización, el internet de las cosas y la existencia de sensores ubicuos, han puesto en entredicho

los medios para recabar el consentimiento de los usuarios para el tratamiento de sus datos personales.

La solución ha sido vista en las políticas de privacidad *online*, ofrecidas a los usuarios como términos unilaterales y (cuasi) contractuales, que se han convertido en la piedra angular de la protección de la privacidad *online*, a pesar de la aplastante evidencia de que la mayoría de las personas ni siquiera lee los términos o no los comprende.

Ante esta situación, los operadores jurídicos demandan mejoras, en especial en lo relativo a:

- la forma en la que las políticas de privacidad son redactadas, de modo que haya una notificación efectiva; y
- desarrollar mecanismos que permitan otorgar un consentimiento informado, con una especial importancia sobre los sistemas denominados *opt-in* y *opt-out*.

Por su parte, los retos surgidos del *big data* hacen que el consentimiento, por sí mismo, no sea suficiente. A continuación analizaremos los retos que sufren los modelos existentes de notificación y consentimiento.

(i) ¿ES EL LENGUAJE SENCILLO LA SOLUCIÓN?

Una política de privacidad ideal ofrecería a los usuarios verdadera libertad de elección, sobre la base de una comprensión suficiente de lo que implica dicha elección.

Algunos actores abogan por la utilización de un lenguaje sencillo, políticas fáciles de comprender y casillas o ventanillas fáciles de identificar en las que los usuarios pueden indicar su consentimiento.

Sin embargo, en el entorno actual de complejos flujos de datos y actores con intereses diferentes se ha desencadenado lo que Solon Baroccas y Helen Nissebaum han denominado «la paradoja de la transparencia»[53], en el sentido de que la simplicidad y la claridad conllevan, de forma inevitable, una pérdida de precisión.

---

[53] Solon BAROCCAS y Helen NISSEBAUM. «Privacy, big data and the public good»*;* Chapter 2*:* Big data's End Run Around Anonimity and Consent», *Cambrigde University Press* (2014).

Baroccas y Nissebaum exponen que la evidencia en este sentido es tajante: los pocos usuarios que leen las políticas de privacidad no las comprenden. De este modo, una redacción sencilla de estas políticas de privacidad podría facilitar su comprensión. No obstante, incluso cuando los usuarios comprenden las políticas de privacidad, los textos escritos en este lenguaje sencillo no permiten tener información suficiente para elaborar un consentimiento informado. Por contra, el detalle que sería necesario para que la política de privacidad diera información suficiente sería abrumador[54].

Así, por ejemplo, en el negocio de la publicidad personalizada, muy desarrollado en la era del *big data* actual, para que los usuarios puedan tomar una decisión informada sobre privacidad, deberían ser notificados sobre el tipo de información que se recoge, con quién se compartirá, bajo qué límites, y para qué fines. Un lenguaje sencillo no puede proveer de toda la información necesaria para que los usuarios tomen una decisión suficientemente informada.

En este sentido, se ha estimado que si todos los usuarios de internet estadounidenses leyeran las políticas de privacidad cada vez que visitan una nueva página web, el país perdería alrededor de 781.000 millones de dólares anuales por el coste de oportunidad del tiempo dedicado a leer estas políticas de privacidad[55].

(ii) DEBER DE INFORMACIÓN SOBRE DATOS PRIMARIOS Y SECUNDARIOS

Los mismos problemas surgen, no solo con respecto a la publicidad personalizada, sino de forma generalizada. Así, por ejemplo, consideremos algunos momentos en los que se generan y almacenan datos de forma cotidiana: abrirse un perfil en una red social, comprar a través de internet, descargarse una aplicación móvil o viajar. Todas estas actividades crean datos brutos cuyo tratamiento posterior justifica que el individuo otorgue su consentimiento.

---

54 Ídem.

55 Aleecia M. McDONALD y Lorrie FAITH CRANOR. «The Cost of Reading Privacy Policies». *Journal of Law and Policy of the Information Society,* Vol. 4, n.º 3 (2008).

Además, la cadena de emisores y receptores de datos es potencialmente infinita, e incluye actores e instituciones cuyo rol y responsabilidades no están delimitados o comprendidos. Así, la cesión de datos puede llegar a ser relativamente oscura.

Lo dicho hasta ahora nos hace preguntarnos ¿cómo ha de ser redactada la información para que los usuarios puedan otorgar su consentimiento informado? El funcionamiento del *big data* hace que esta tarea sea tremendamente difícil, por cuanto los datos se mueven de un lugar a otro, y de un receptor a otro de modo impredecible, ya que el valor de los datos no se conoce en el momento en que son recogidos. Así, el consentimiento se parece cada vez más a un cheque en blanco[56].

En esta situación, la pregunta que surge es si la obligación del responsable del tratamiento de informar sobre la recogida de los datos se circunscribe a la información que explícitamente recoge, o si debe adoptarse un criterio más amplio y entender que este deber de información también alcanza a aquella información que la institución pudiera obtener tras el tratamiento.

Numerosos autores opinan que los deberes de información y la necesidad de recabar el consentimiento debe referirse, no solo al hecho de que se recaben datos primarios, sino también a la información que se puede extraer de un análisis sofisticado de éstos, incluyendo la información que pueda extraerse de la agregación de datos que recaba la empresa con datos provenientes de otras fuentes y ficheros. No obstante, esta aproximación tiene muchas dificultades prácticas, en tanto que, por su propia naturaleza, el valor del *big data* reside precisamente en lo inesperado de los resultados que revela. Así, ¿cómo explica el responsable del tratamiento que resulta imposible saber con antelación qué información revelará el tratamiento de los datos recabados? Son muchos los autores que consideran que el consentimiento prestado bajo estas circunstancias no es el consentimiento informado que la ley exige[57].

---

[56] Solon BAROCCAS y Helen NISSEBAUM. «Privacy, big data and the public good; Chapter 2: Big data's End Run Around Anonimity and Consent», *Cambrigde University Press* (2014).

[57] Por todos, Fred H. CATE y Viktor MAYER-SCHÖENBERGER. «Notice and consent in a world of Big data». *International Data Privacy Law,* Vol 3, n.º 2 (2013); Omer TENE y Jules POLONESTSKY. «Big data for all: Privacy

(iii) DEDUCIR DATOS DE LA MAYORÍA A PARTIR DE DATOS DE LA MINORÍA

El dilema sobre el consentimiento para recoger y tratar datos propios también se agrava por lo que Baroccas y Nissebaum han denominado la «tiranía de la minoría». Su teoría parte de la premisa de que la información que voluntariamente comparten unos pocos individuos puede revelar la misma cantidad de información sobre aquellos que deciden no otorgar su consentimiento, en tanto que las instituciones con quienes esa minoría han consentido pueden inferir los mismos resultados para aquella mayoría que no ha consentido[58].

Esto supone que, en realidad, cada individuo no tiene una capacidad real de tomar una decisión que proteja sus intereses (en este caso, proteger su privacidad y la información que sus datos puedan revelar). Veamos algunos ejemplos prácticos.

Por ejemplo, una solicitud de amistad en una red social muestra una conexión entre ambas personas, que permite deducir algún tipo de vínculo común; ya sea compartir intereses, afinidades o algún momento de la historia personal. Esto permite crear inferencias en determinados comportamientos.

Bajo esta premisa, los científicos informáticos se han puesto manos a la obra para intentar responder a la pregunta de si las técnicas *big data* de análisis de redes sociales y minería de datos podrían ser utilizadas para inferir atributos de un usuario basándose en la información revelada por otro usuario. Los resultados de varios de los experimentos que se han podido llevar a cabo son reveladores.

Ejemplo: Inferir información de una persona a partir de la información facilitada por sus amigos en redes sociales

Los usuarios de las redes sociales crean perfiles que normalmente incluyen datos como localización geográfica, intereses y

---

and user control in the age on analytics». *Northwestern Journal of Technology and Intellectual Property*, Vol. 11, n.º 5 (2013); Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?» *International Privacy Law,* Vol. 3, n.º 2 (2013).

58 Solon BAROCCAS y Helen NISSEBAUM. «Privacy, big data and the public good; Chapter 2: Big data's End Run Around Anonimity and Consent», *Cambrigde University Press* (2014).

la universidad a la que acuden. Esta información es utilizada por las redes sociales para poder agrupar usuarios, compartir contenido y sugerir conexiones con otros usuarios. Pero no todos los usuarios revelan esta información.

En su experimento, Alan Misolve[59] quería responder a la siguiente pregunta: dados determinados atributos por algunos de los usuarios de una red social, ¿podemos inferir esos mismos atributos sobre otros usuarios, utilizando para ello los gráficos de relaciones sociales?

Para intentar responder a esta pregunta, el estudio recogió datos muy detallados que obtuvo de dos redes sociales. Observaron que los usuarios con características comunes tienen más probabilidades de ser amigos en las redes sociales, y en ocasiones crean densas comunidades. El estudio demostró que es posible inferir atributos gracias a las comunidades y relaciones de amistad en las redes sociales. De este modo, el estudio analizó la información que algunos usuarios de redes sociales publicaban sobre la titulación universitaria que cursaban, su año de graduación y su dormitorio en las residencias universitarias. El experimento fue capaz de deducir estos mismos atributos, con un alto grado de exactitud, de aquellos otros estudiantes que no habían revelado estos datos en redes sociales.

Con ello, el estudio concluyó que algunos atributos pueden ser inferidos con un alto grado de precisión a partir de los datos de tan solo el 20% de los usuarios.

Ejemplo: Inferir la orientación sexual de una persona a partir de sus amistades en redes sociales

Dos estudiantes del Massachusetts Institute of Technology (MIT) crearon en 2007 un programa de software llamado

---

[59] Alan MISOLVE *et al.* «You are who you know: inferring users profiles in online social networks». *Web Search and Data Mining (WSDM)*. ACM, Nueva York (2010); citado por Solon BAROCCAS y Helen NISSEBAUM. «Privacy, big data and the public good; Chapter 2: Big data's End Run Around Anonimity and Consent», *Cambrigde University Press* (2014). Solo ha sido posible acceder a un extracto del estudio original.

> Gaydar, que permite concluir la orientación sexual de los usuarios de redes sociales, como proyecto para su asignatura de Ética, Derecho e Internet[60].
>
> Analizaron los lazos de amistad en Facebook de 1544 hombres que se declaraban heterosexuales, 21 hombres que declaraban ser bisexuales y 33 homosexuales, e investigaron las correlaciones entre la orientación sexual del usuario y la de sus amigos. Los hombres homosexuales tenían una proporción mucho mayor de amigos homosexuales en la red social, de modo que con estos datos crearon un sistema para que el programa pudiera inferir la orientación sexual de otros usuarios en función de sus amigos.
>
> El estudio no pudo probar con rigor científico sus conclusiones, pues estuvo restringido por las limitaciones de ser un proyecto de fin de curso. Sin embargo, sí pudo demostrar que el programa podía deducir con un margen de error pequeño qué usuarios eran hombres homosexuales.
>
> Por contra, las predicciones sobre hombres bisexuales o mujeres homosexuales no fueron tan certeras.

Gaydar es solo uno de los muchos proyectos que pretender realizar minería de datos sobre la información de las redes sociales y las relaciones de amistad de los usuarios para obtener información potencialmente muy valiosa, pero personal.

El riesgo para la privacidad de las personas se acrecienta cuando, a partir de los datos confirmados de un número suficientemente grande de contactos de redes sociales, que revelan sus propios datos, se pueden inferir datos no revelados por otros usuarios. Debido a esto, determinadas personas deciden no estar activos en redes sociales como Facebook. Sin embargo, esta solución puede no ser suficiente. Veamos el siguiente ejemplo.

---

60 Carter JERNIGAN y Behran F. T. MISTREE. «Gaydar: facebook friendships expose sexual orientation». *First Monday*, Vol. 14, n.º 10 (2009); citado por *ABC News Tech*. Solo ha sido posible acceder a referencias del estudio.

Ejemplo: inferir información de individuos que no forman parte de una red social a partir de información obtenida en redes sociales

El experimento llevado a cabo por un equipo de la Universidad de Heidelberg (Alemania) quería analizar si es posible que los datos revelados por determinadas personas en redes sociales desprendan datos sobre otras personas que no formen parte de la red social[61].

Gracias al estudio de los gráficos de relaciones sociales en redes y de la dirección de correo electrónico de los miembros de las redes, el grupo creó un algoritmo capaz de inferir que dos personas ajenas a la red social (e incluso sin necesidad de conocerse entre ellas) compartían ciertas características, sobre la base de los datos obtenidos de un amigo común, presente en redes sociales.

En este sentido, la investigadora sobre comunicación y redes sociales danah boyd[62] afirma que «tu registro permanente ya no solo se conforma de lo que uno haga. Todo lo que otros hacen que nos concierna, nos implique o nos pueda influir pasará a formar parte de nuestro registro permanente»[63].

De este modo, incluso aunque un usuario realice esfuerzos para no revelar información personal (por ejemplo, cambiando los parámetros por defecto del ordenador, rechazando publicar información sobre su ideología política, religión u orientación sexual, o no publicando fotos), la información de sus contactos en las redes sociales, o incluso la misma lista de contactos en las redes sociales, puede permitir a otros deducir información sobre nosotros.

Pero todavía más asombroso es el hecho de que pueden hacerse inferencias similares sobre una población entera incluso cuando úni-

---

61 Emöke-Ágnes HORVÁT, Michael HANSELMANN, Fred A. HAMPRECHT, y Katharina A. ZWEIG. «One plus one makes three (for Social Networks)». *PLOS One Journal* (2012).

62 Danah Michele Mattas cambió legalmente su nombre en el año 2000 por danah boyd, escrito con minúsculas.

63 danah boyd. «Networked privacy». *Personal Democracy Forum*, Nueva York (2011).

camente una pequeña proporción de personas, con las que ni siquiera tienen conexiones o relaciones de amistad, revela sus datos.

Ejemplo: Predecir embarazos

La cadena de grandes almacenes Target llevó a cabo un estudio a través del cual podía predecir la tasa de embarazos de sus clientas. En este caso, no se realizaron deducciones sobre la base de los amigos en redes sociales. Target analizó los datos de sus ficheros sobre mujeres que habían celebrado un *baby shower*[64] para identificar a las mujeres que habían revelado el hecho de que estaban embarazadas, y estudió su cesta de la compra. Puesto que estos hábitos eran diferentes de los de otros clientes, Target pudo averiguar qué clientas podrían estar embarazadas sobre la base del cambio en sus hábitos de consumo y la información revelada por aquellas otras mujeres con las que no tenían ningún tipo de vínculo.

Así las cosas, la pregunta que surge es: ¿Cuál es la proporción mínima de personas que debe revelar sus datos sobre un atributo en concreto para que sea posible identificar qué otros miembros de la población total poseen ese mismo atributo?

Esta pregunta vuelve a traernos a los conceptos estadísticos: en tanto la muestra sea representativa y los atributos analizados sean estadísticamente relevantes será posible realizar inferencias con un menor margen de error a partir de una muestra menor.

En este sentido, el estudio de Alan Misolve y sus compañeros[65], al que hicimos referencia unas páginas más atrás, reveló que es posible desprender determinados atributos de toda la población con el hecho de que únicamente un 20% de esta población revele dichos atributos. Cabe poner de manifiesto que este umbral se obtuvo para este experimento concreto, que trataba de averiguar atributos relativamente simples (titulación estudiada, año de graduación y dormitorio), y úni-

---

64 Un «*baby shower*» es una forma de celebrar el futuro o reciente nacimiento de un bebé presentando regalos a los padres en una fiesta, muy común en países como Estados Unidos.

65 Alan MISOLVE *et al.* «You are who you know: inferring users profiles in online social networks». *Web Search and Data Mining (WSDM)*. ACM, Nueva York (2010).

camente analizaba los mismos atributos que luego se querían inferir, sin valorar otra gran cantidad de información que se pueden extraer de las redes sociales.

En cualquier caso, parece razonable concluir que el valor añadido del consentimiento prestado por un individuo concreto disminuye a medida que otros usuarios prestan su consentimiento y la muestra y la base de datos que se crea van alcanzando representatividad estadística.

Es de este modo que una minoría puede determinar los atributos que se infieran de la población total analizada, y desincentiva a los responsables del tratamiento de los ficheros de invertir en procesos que faciliten obtener el consentimiento del resto de usuarios una vez que el umbral mínimo de representatividad se ha alcanzado. En consecuencia, no prestar el consentimiento puede no cambiar la forma en que los responsables del tratamiento de datos categoricen o traten a un individuo.

(iv) PÉRDIDA DE BENEFICIO SOCIAL E INNOVACIÓN

Como ya quedó expuesto al inicio de este capítulo, el Dictamen 15/2001 del Grupo de Trabajo del Artículo 29, sobre la Definición del Consentimiento resalta que el consentimiento «parece implicar una necesidad de acción». En la práctica, esto implica priorizar sistemas *opt-in* sobre los sistemas *opt-out*.

Los sistemas *opt-in* y *opt*-out son dos formas de manifestar el consentimiento. El sistema *opt-in* se basa en que el usuario debe manifestar un consentimiento expreso y positivo y rellenando la casilla creada al efecto. Por su parte, en el sistema *opt-out* el individuo debe manifestar su oposición, bien rellenando la casilla correspondiente o manifestándolo a la organización por el medio oportuno (por ejemplo, en muchas ocasiones la forma de manifestar la oposición es accediendo a un enlace web para darse de baja).

En resumen, como ya sabemos, el consentimiento es el instrumento principal de la normativa de protección de datos, cuyo fundamento es otorgar al individuo un poder de control sobre sus datos. Los sistemas *opt-in* y *opt*-out permiten un diferente grado de control de

los individuos sobre sus datos. Y esta diferencia, ¿es relevante en la práctica?

Para analizarlo, hagamos primero un receso para señalar dos líneas doctrinales existentes en la actualidad. Por un lado, algunos autores[66] sostienen que todos los datos deberían ser considerados como personales, y así, estar sometidos a los requisitos de la normativa sobre protección de datos. Esto implicaría pedir el consentimiento para el tratamiento de cualquier dato. Sin embargo, una definición tan amplia de los datos de carácter personal sería fácticamente inmanejable.

Frente a esta corriente, otros autores[67] sostienen que debe primar una aproximación más pragmática sobre el consentimiento del individuo, y que el derecho a la privacidad y la protección de datos deben ponerse en equilibrio con otros valores sociales tales como la salud pública o la seguridad nacional. Así, en relación con los sistemas *opt-in* y *opt*-out mencionados, consideran que el sistema *opt-in* podría provocar una importante pérdida de beneficios sociales colectivos si los individuos deciden no rellenar la casilla para prestar su consentimiento.

En un intento de proponer una solución a las diferencias entre estas dos corrientes, Tene y Polonetsky proponen un marco basado en una matriz de riesgo: cuando los beneficios del tratamiento de los datos superen a los riesgos sobre la privacidad de los individuos, debe asumirse que el responsable del fichero tiene legitimidad para el tratamiento de los datos, incluso cuando los individuos no hayan prestado su consentimiento. Por ejemplo, el análisis de webs cuyo fin sea comprender y mejorar el uso de dicha web crea un alto valor, asegurando que los productos y servicios puedan ser mejorados para lograr un mejor servicio. Los riesgos a la privacidad son mínimos, pues si el sistema se implementa de modo correcto, únicamente trata datos estadísticos que no permiten identificar a una persona en concreto. En estas situaciones, exigir que los usuarios presten un consentimiento expreso (*opt-in*) al análisis amenazaría seriamente el uso del análisis. De hecho, una de las claves del *big data* es poder diferenciar a los in-

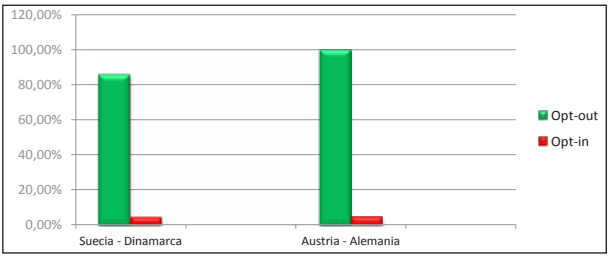
---

66 Paul OHM. «Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization». *UCLA Law Review*, Vol. 57 (2010).

67 Omer TENE y Jules POLONETSKY. «Privacy in the age of big data: a time for big decisions». *Standford Law Review Online*, Vol. 64, n.º 63 (2012).

dividuos para poder mantener la información de cada sujeto diferenciada de la de otros sujetos, pero sin que sea necesario identificarlos.

Pensemos también por ejemplo en los sistemas de donación de órganos. Aquellos países en los que se sigue un sistema *opt-in* la tasa de donaciones de órganos es mucho menor que aquellos países culturalmente similares pero con un sistema *opt-out*[68]. Así, la tasa de donaciones es muy superior en Suecia, que sigue un modelo *opt-out* que en Dinamarca, que sigue un modelo *opt-in*, a pesar de que son países culturalmente muy similares y se comportan análogamente en muchas otras esferas. Lo mismo ocurre entre Austria y Alemania.



Tene y Polonetsky no defienden que nunca haya que recabar un consentimiento expreso del individuo. En numerosas ocasiones será necesario el consentimiento (ya sea mediante un sistema *opt-in* u *opt-out*), tales como en lo relativo a servicios de marketing comportamental, tratamiento de datos por parte de terceros o servicios basados en geolocalización[69]. Sin embargo, una creciente atención por el consentimiento expreso y el principio de minimización de datos, sin tomar en consideración el valor o los usos de dichos datos, podría ralentizar la innovación y los avances sociales.

Tras haber expuesto las características del consentimiento tal y como está definido en la Directiva y la LOPD, y según la interpretación que aporta el GT 29, hemos analizado cómo influye el *big data* sobre este consentimiento. Las páginas anteriores han puesto de manifiesto que el consentimiento, tal y como está previsto en la actualidad, no soluciona los problemas prácticos que antes venía a solucionar. Las políticas de privacidad no son comprensibles para los

68 Ídem.
69 Ídem.

individuos, y la información necesaria para cumplir con el requisito legal del consentimiento informado no se puede detallar en el momento en que hay que recabar este consentimiento, precisamente porque los fines para los que los datos podrán ser utilizados no se conocen *a priori*.

En las próximas páginas pasamos a analizar el otro gran instrumento que permite obtener el valor de los datos preservando la privacidad de los individuos: las técnicas por las que los datos son hechos anónimos.

CAPÍTULO IV. ANONIMIZACIÓN Y PSEUDONIMIZACIÓN DE DATOS

En este punto ya sabemos que existe una dicotomía entre los datos de carácter personal (que están sujetos a las normas de protección de datos) y los datos anonimizados. Una vez que un set de datos ha sido anonimizado y los individuos no son identificables, la normativa de protección de datos no se aplica.

Tradicionalmente, la anonimización consistía en un proceso de dos fases principales. En primer lugar despojar a los conjuntos de datos de todos los rasgos identificadores personales (PII por sus siglas en inglés-*personal identificable information),* como pueden ser nombre, dirección, fecha de nacimiento o número de seguridad social. En segundo lugar, se modificaban o eliminaban otras categorías de datos que podían actuar como identificadores en dicho contexto concreto (por ejemplo, un banco eliminaría los números de tarjeta de crédito, y una universidad eliminaría los números de identificación de sus estudiantes).

De este modo, el resultado aunaba lo mejor de ambos lados: los datos continuaban siendo útiles, y podían ser analizados, compartidos o puestos a disposición del público al tiempo que los individuos no podían ser identificados, y por lo tanto se protegía su privacidad. La anonimización aseguraba la privacidad.

Sin embargo, con los nuevos avances, esta situación cambia. El *big data*, al incrementar la cantidad y diversidad de la información, facilita la reidentificación de individuos, incluso después de haber sido anonimizados[70].

En efecto, la práctica[71] constata que la creación de un set de datos verdaderamente anónimo no resulta tarea fácil. Por ejemplo, un set

---

70 Kenneth NEIL CUKIER y Viktor MAYER-SCHÖENBERGER. «Big data: A Revolution That Will Transform How We Live, Work And Think». *Houghton Mifflin Harcourt* (2013).

71 En este sentido, ver las demostraciones llevadas a cabo por Lantaya Sweeny, Arvind Narayanan y Vitaly Shmatikov, algunas de las cuales son expuestas más adelante en este trabajo.

considerado anónimo puede ser combinado con otros de forma tal que uno o más individuos puedan ser reidentificados. Es por ello que para las autoridades de protección de datos la anonimización es un proceso crítico.

En este sentido, la Federal Trade Commission de Estados Unidos ha declarado que:

> «Hay evidencias suficientes que demuestran que los avances tecnológicos y la posibilidad de combinar diferentes datos puede conllevar la identificación de un consumidor, ordenador o dispositivo, incluso si estos datos por sí mismos no constituyen datos de identificación personal. Es más, no solo es posible reidentificar datos que no son identificadores personales a través de medios diversos, sino que las empresas tienen fuertes incentivos para hacerlo» [72].

En este capítulo analizaremos el marco jurídico de los datos anónimos y los problemas que el desarrollo de la tecnología trae para proteger el anonimato de los datos.

4.1 MARCO JURÍDICO DE LA ANONIMIZACIÓN

En realidad, ni la normativa europea ni nuestra legislación española regulan los datos anónimos ni el proceso de anonimizar la información. Quizás la referencia más directa sea el considerando 26 de la Directiva de Protección de Datos. De acuerdo con la Directiva, la anonimización implica el tratamiento de datos de carácter personal de modo tal que no sea posible volver a identificarlos:

> «Los principios de la protección deberán aplicarse a cualquier información relativa a una persona identificada o identificable. Para determinar si una persona es identificable, hay que considerar el conjunto de los medios que puedan ser *razonablemente* utilizados por el responsable del tratamiento o por cualquier otra persona, para identificar a dicha persona.

---

[72] FEDERAL TRADE COMMISSION (FTC). «Protecting Consumer Privacy in an Era of Rapid Change. Recommendations for Businesses and Policymakers» (2012).

> Los principios de la protección no se aplicarán a aquellos datos hechos anónimos de manera tal que ya *no sea posible identificar al interesado* (...)»[73]. (Cursiva añadida).

La Directiva sobre Privacidad y Comunicaciones Electrónicas[74] se ha referido a los conceptos de «anonimización» y «datos anónimos» en términos similares.

Así, del texto de la Directiva de Protección de Datos pueden extraerse los elementos clave de la anonimización:

- «Hay que considerar el conjunto de medios que puedan ser razonablemente utilizados»: esto implica tener en cuenta el estado de la técnica en cada momento, con especial referencia al coste y al *know how* necesario para revertir la anonimización.

  Se podría criticar que la redacción de la norma utilice términos tan abstractos. No obstante, hemos de tener en cuenta que el desarrollo de la tecnología a lo largo del tiempo podría elevar el riesgo de reidentificación de los datos, de modo que para que la redacción legal pueda adaptarse al contexto de cada momento, debe mantenerse técnicamente neutral.
- «La anonimización debe realizarse de manera tal que ya no sea posible identificar al interesado»: ¿debe esto interpretarse en el sentido de que la Directiva establece el umbral de que la anonimización sea irreversible? Vamos a dar respuesta a esta pregunta en las páginas siguientes.

Nuestro análisis de la anonimización va a estar basado, en gran medida, en la Opinión que el Grupo de Trabajo del Artículo 29 ha

---

73 Considerando 26 de la Directiva 95/46/EC del Parlamento Europeo y del Consejo, de 24 de octubre de 1995 relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

74 Directiva 2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas (Directiva sobre la privacidad y las comunicaciones electrónicas); también conocida como «e-Privacy Directive».

publicado acerca de la efectividad y las limitaciones de las técnicas de anonimización actuales[75]. En este Capítulo nos referiremos a dicho informe como «la Opinión».

Pese a que los informes del GT 29 no tienen fuerza vinculante, resulta un buen punto de partida para saber el punto de vista de las autoridades de protección de datos de los países miembros de la Unión Europea. En concreto, este documento llama la atención y clarifica numerosas cuestiones relevantes. Sin embargo, en determinados momentos, propone cuestiones que parecen no ser consistentes con el nivel de la técnica actual, que también analizaremos en los próximos epígrafes.

4.2 ¿A PARTIR DE QUÉ UMBRAL CONSIDERAMOS QUE LOS DATOS SON ANÓNIMOS?

La Directiva ha sido traspuesta en las legislaciones nacionales de cada Estado miembro e interpretada por sus tribunales de formas muy diversas. Así, lo que se entiende por dato anónimo puede variar entre:

- Anonimización absoluta: implica la posibilidad cero de reidentificar, directa o indirectamente, a ninguno de los sujetos. En términos fácticos, este nivel de anonimización es imposible de conseguir en muchas ocasiones, principalmente cuando tratamos con ficheros muy ricos en datos.
- Anonimización funcional: implica un riesgo insignificante de reidentificación.

Sobre este extremo, la Opinión del GT 29 sobre técnicas de anonimización parece caer en algunas contradicciones que es importante analizar. Por un lado, la Opinión reconoce que existe un riesgo de reidentificación residual incluso después de aplicar las técnicas de anonimización. Pero, por otro lado, la Opinión también señala que la Directiva ordena que la anonimización sea «irreversible». Parece que

---

[75] GRUPO DE TRABAJO DEL ARTÍCULO 29. «Opinion 05/2014 on Anonymisation techniques» (2014).

ambos conceptos son contrarios. La importancia que en la práctica tiene asumir una u otra interpretación justifica el análisis.

A este respecto, el trabajo de Khaled El Emam y Cecilia Álvarez es de gran ayuda[76]. Si el umbral para considerar que un dato es anónimo es el «riesgo cero», la anonimización no sería posible en la práctica, y por lo tanto, no sería posible tratar o ceder datos sin consentimiento (o alguno de los otros fundamentos legales). Esto podría hacer cesar muchos flujos de información, con la pérdida de beneficio social que ello supondría, de forma más acuciante aún en el entorno *big data*. También podría suponer que las organizaciones no tuvieran incentivos para dificultar la identificación de datos y buscaran alternativas para seguir tratando los datos, entre las que cabe imaginar ampliar los fines primarios de los datos, lo que en última instancia podría ser más lesivo para la privacidad de los individuos.

Así, la interpretación más razonable es la de que el «riesgo cero» no debe ser el umbral a seguir. En su lugar, ha de acudirse al juicio de razonabilidad que establece el considerando 26 de la Directiva, arriba citado. Ésta es la aproximación que ha realizado la norma española cuando hace referencia a «plazos o actividades desproporcionados»[77].

La nueva Propuesta de Reglamento también ha adoptado este criterio, lo que, de ser finalmente aprobado, ayudará a una interpretación uniforme en todos los Estados miembros. Y, como recuerdan K. El Emam y C. Álvarez, ésta también es la definición que han desarrollado jurisdicciones no europeas, como Canadá[78] y Estados Unidos[79].

---

76 Khaled EL EMAM y Cecilia ÁLVAREZ. «A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques». *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.º 1 (2015).

77 Artículo 5.1 (o) del Reglamento de desarrollo de la LOPD.

78 Personal Health Information Protection Act, Ontario, Canadá (2004).

79 Health Insurance Portability and Accountability Act Privacy Rule (HIPAA), Estados Unidos (1996).

En tanto la anonimización absoluta no existe, el Profesor Mark Elliot alude al concepto de la «utilidad del riesgo de reidentificación»[80]. Esto es, para que los datos recogidos conserven su utilidad, se opera bajo un concepto de insignificancia *(«negligibility»)*. De acuerdo con esto, cuando el riesgo de reidentificación es tan pequeño que, a pesar de no ser estadísticamente cero, es considerado funcionalmente cero, se considera que la anonimización llevada a cabo es suficiente. De otro modo, operar con los datos de los que disponemos sería fácticamente imposible.

En un sentido similar, el Profesor Josep Domingo Ferrer, aclara cómo se lleva a cabo la anonimización en el 95% de los casos: mediante la aproximación que él mismo denomina *«utility first»*. Consiste en anonimizar los datos manteniendo su utilidad, y tras ello, analizar el nivel de riesgo de reidentificación. En caso de que el riesgo fuera muy elevando, habría que volver a anonimizar los datos con una mayor distorsión hasta que el análisis de riesgos diera un nivel aceptablemente bajo.

Todas estas aproximaciones de K. El Emam, C. Álvarez, Mark Elliot y de Josep Domingo Ferrer son similares en la medida en que tratan de preservar la utilidad de los datos.

Frente a esta aproximación, está la que el Profesor Ferrer denomina *«privacy first»,* y que ha sido acogida principalmente en el mundo académico, pero no utilizada con frecuencia en la práctica. Este método consiste en establecer *a priori* el nivel de privacidad deseado, y utilizar modelos de privacidad que garanticen este nivel de protección, sin tomar en cuenta el nivel de utilidad de los datos.

Así las cosas, la cuestión central es qué debe considerarse un nivel aceptable de riesgo de reidentificación, en tanto que los datos anónimos quedan fuera de la normativa de protección de datos.

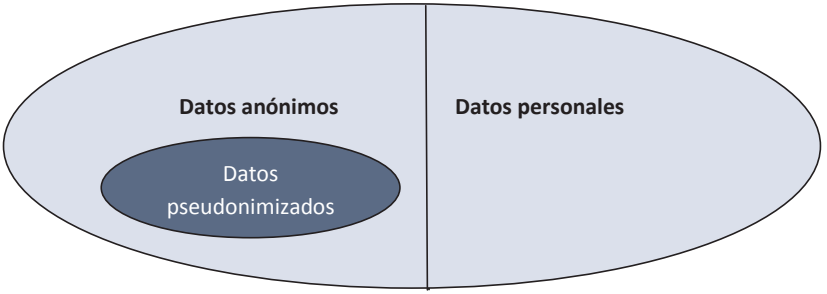
En el contexto de *big data*, el análisis de grandes cantidades de datos puede llegar a permitir identificar personas a partir de datos que nadie habría considerado de identificación personal o que antes eran anónimos, así como inferir gustos y comportamientos de los individuos a pesar de no conocer su identidad.

---

80 Mark ELLIOT. «To be or not to be (anonymous)? Anonymity in the age of big and open data». [Conferencia] *Computers, Privacy & Data Protection on the Move (CPDP)* (2015).

4.3 LA PSEUDONIMIZACIÓN NO ES ANONIMIZACIÓN

Tradicionalmente se incluía la pseudonimización como una técnica más de anonimización.



La pseudonimización consiste en remplazar un atributo de un set de datos (normalmente un atributo único que funciona de identificador directo, como el nombre y los apellidos) por otro atributo (como por ejemplo, el DNI, el número de Seguridad Social, o un código aleatorio que no pueda ser descifrado, de modo que no pueda conocerse a quién se refiere)[81].

Los métodos más extendidos de pseudonimización son la encriptación y la *tokenización*.

Encriptación y *tokenización*. Definición

La encriptación con una clave secreta permite que el dueño de la clave reidentifique a los sujetos desencriptando la clave (por ejemplo, volviendo a asociar cada número de la Seguridad Social con el nombre de la persona).

Por su parte, la *tokenización* se aplica principalmente en el sector financiero para el procesamiento de tarjetas de crédito. Normalmente, la creación del identificador (*token*) consiste en sustituir los números de DNI por valores de escasa utilidad para un posible *hacker*, pero que garantizan la misma operati-

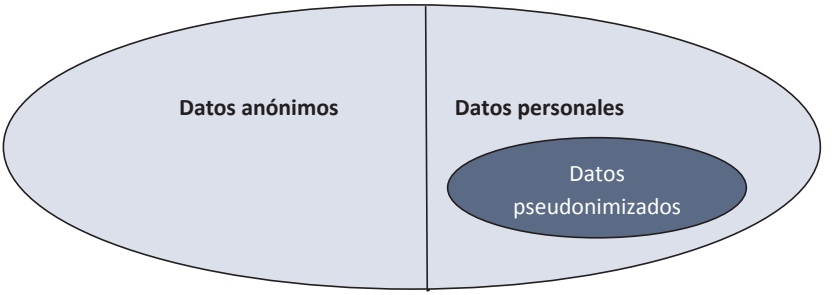
---

81 GRUPO DE TRABAJO DEL ARTÍCULO 29. «Opinion 05/2014 on Anonymisation techniques» (2014).

vidad, a través de un sistema de encriptado unidireccional que genera un número aleatorio.

Así, a pesar de que tradicionalmente los datos pseudonimizados eran considerados datos anónimos, en la actualidad la pseudonimización ya no se considera un método de anonimización, pues la persona es todavía identificable, aunque sea de forma indirecta.

En la medida en que reduce la asociación («*linkability*») entre la información y el sujeto de la que proviene, la pseudonimización es una medida de seguridad útil, aunque en todo caso continúa permitiendo la identificación de los sujetos. Así, actualmente se considera que los datos pseudonimizados son todavía datos de carácter personal y están sujetos a la normativa sobre protección de datos de carácter personal[82].



En este sentido, la historia del número de la Seguridad Social es ilustrativa. Asociando a cada persona con un número único, éste servía de identificador de la persona ante las administraciones, aunque fuera de éstas se trataba de un dato sin significado, anónimo. De hecho, el número era incluso un identificador más único que un nombre, pues los más comunes pueden aparecer más de una vez en una base de datos. A medida que su uso fue extendiéndose, las grandes compañías primero, y pequeñas empresas después, lo adoptaron también como forma de identificar a las personas, y en la actualidad

---

82 Vincet TOUBIANA. «To be or not to be (anonymous)? Anonymity in the age of big and open data». [Conferencia] *Computers, Privacy & Data Protection on the Move (CPDP)* (2015); Khaled EL EMAM y Cecilia ÁLVAREZ. «A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques». *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.º 1 (2015).

constituye un identificador único y de uso común. Por ello, actualmente, el número de Seguridad Social es un dato sensible que permite una identificación directa de la persona, y por tanto es un dato de carácter personal.

Así, cualquier dato aleatorio que pueda llegar a utilizarse como un identificador único será un pseudónimo y no un «identificador anónimo», cuyo valor anónimo decrece a medida que su uso se va extendiendo.

Ello no implica que el uso de pseudónimos carezca de valor. En efecto, limita la capacidad de inferir datos como género, raza, religión u origen nacional. Además, aquellos pseudónimos que no se comparten entre distintas bases de datos, no permiten dicha identificación directa. En principio, únicamente la institución que asigna este pseudónimo podrá reconocer a la persona a la que corresponde. Y cuando éste se elimina o se sustituye, ni siquiera la institución que lo asignó a una persona concreta podrá reconocerla por ese mismo pseudónimo[83]. Esto es lo que ocurre, por ejemplo, cuando una *cookie* que tenemos instalada en el ordenador y que permite identificarnos, se elimina del sistema cuando caduca o cuando la borramos.

Así las cosas, actualmente la pseudonimización no puede ser considerada una técnica de anonimización. No obstante, la equivocación de considerar que sí se trata de un método de anonimización es uno de los riesgos más importantes de las técnicas de pseudonimización.

En este sentido es especialmente revelador el trabajo de Narayanan y Shmatikov, que han mostrado el efecto de la reidentificación en el contexto de las redes sociales[84].

Caso: gráficos de amistad en redes sociales

Narayanan y Shmatikov han llevado a cabo un estudio mediante el que han demostrado que determinada información sensible de los usuarios de redes sociales puede ser extraída de

---

83 Solon BAROCCAS y Helen NISSEBAUM. «Privacy, big data and the public good. Chapter 2: Big data's End Run Around Anonimity And Consent». *Cambrigde University Press* (2014).

84 Arvind NARAYANAN y Vitaly SHMATIKOV. «De-anonymizing social networks». *The University of Texas at Austin, Simposio IEEE on Security and Privacy* (2009).

los gráficos de amistad en redes sociales («*social networks graphs*»), a pesar de que los datos hayan sido pseudonimizados mediante la utilización de apodos en lugar de los nombres reales de los sujetos. Esto es posible puesto que las relaciones de amistad entre los individuos de una red social son únicas y pueden servir de identificadores.

El experimento consistió en crear un algoritmo que permitía reidentificar a los usuarios de la red social Twitter a través de los gráficos de amistad de la propia red Twitter y de la red social Flickr, que se utilizó como fuente de información adicional.

Así, el primer gráfico estaba formado por las relaciones de seguidores de la red Twitter. El segundo gráfico estaba formado por la red de contactos de Flickr. Ambas redes requieren la utilización obligatoria de un nombre de usuario, y además permiten completar los campos adicionales de nombre y localización. Los grafos de relaciones utilizan únicamente los nombres de usuario (por ejemplo «xk_562»), de modo que son «anónimos» (pseudónimos en realidad).

El experimento también utilizó los gráficos que muestran las relaciones de amistad de la red de blogs LiveJournal, a la que los investigadores tuvieron acceso.

Para crear el algoritmo, Narayanan y Shmatikov hicieron pruebas con los datos conocidos de la red LiveJournal, y así pudieron constatar cuánta información auxiliar es necesaria para identificar un nodo en aquella red social que constituya nuestro objetivo.

Posteriormente, el algoritmo lleva a cabo lo que se denomina un «ataque de reidentificación» en dos fases. En primer lugar, el atacante identifica los nodos de los gráficos de amistad que se encuentran presentes en ambas redes sociales, es decir, (i) la red objeto de ataque, Twitter, y (ii) la red cuyo gráfico sirve de información auxiliar, Flickr. Posteriormente, ambos gráficos se superponen. La información que se obtiene al superponer ambos gráficos va creando nuevos datos que alimentan al algoritmo, que a su vez identifica nueva información para ampliar los gráficos de amistad. El resultado es una superposición de los grafos de la red auxiliar y la red objetivo de forma tal que permite reidentificar los sujetos de dichas cuentas.

> El resultado del estudio es que hasta un tercio de los individuos (con cuentas verificadas) que son miembros de ambas redes sociales, Twitter y Flickr, pueden ser reidentificados con tan solo un 12% de error. Los datos que se tomaron para llevar a cabo el experimento son más restringidos que los datos que un atacante real pudiera obtener, con lo que la tasa de aciertos en el experimento es menor que la que el atacante real alcanzaría.
>
> Asimismo, el número de individuos presentes en ambas redes sociales en el momento de realizar el experimento no era muy elevado (menos de un 15%), con lo que la superposición de gráficos inicial no era muy amplia. Sin embargo, otras redes sociales más extensas tienen un ratio de superposición mayor (por ejemplo, el propio estudio señala que cerca de un 64% de los usuarios de la red Facebook en ese momento eran también usuarios de la red social MySpace). Todo ello implica que este algoritmo pueda alcanzar un ratio de reidentificación mucho más elevado en otras redes sociales.
>
> Tal y como exponen en sus conclusiones «la principal lección de este trabajo es que la anonimización —que en realidad es pseudonimización— no es suficiente para alcanzar la privacidad cuando se trata de redes sociales. Hemos desarrollado un algoritmo genérico de reidentificación y hemos mostrado que se puede des-anonimizar de forma exitosa a varios miles de usuarios en el gráfico de red anónimo de la popular red de microblogs Twitter, utilizando una red social completamente diferente (Flickr) como fuente de información auxiliar» [85].

Un ejemplo real de los problemas que puede conllevar considerar que la pseudonimización es suficiente para lograr la anonimización es el conocido incidente sufrido por la empresa America On Line (AOL) en 2006 [86].

> **Caso AOL:**
>
> En 2006, el proveedor de servicios de Internet, AOL hizo públicos los datos de 20 millones de búsquedas en internet que

---

85 Ídem.

86 Paul OHM. «Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization». *UCLA Law Review*, Vol. 57 (2010).

> los usuarios habían realizado en su motor, correspondientes a 657.000 usuarios, con la única idea de favorecer la investigación libre. Los datos habían sido «anonimizados» (en realidad pseudonimizados), eliminando el nombre de usuario y la dirección IP y sustituyendo los datos por identificaciones numéricas únicas que permitían a los investigadores correlacionar diferentes búsquedas con usuarios individuales. El objetivo era que los investigadores pudieran vincular las búsquedas llevadas a cabo por la misma persona, sin acceder a la información personal.
>
> En cuestión de días, se relacionaron ciertas búsquedas, tales como las del usuario asignado al código 17556639: «cómo matar a tu mujer», «foto de accidente de coche», «fotos de personas muertas», aunque en este caso no se logró identificar al sujeto.
>
> Unos días después, sin embargo, el periódico New York Times sí logró identificar a Thelma Arnold, una mujer viuda de 62 años de la ciudad de Lilburn, a través de sus búsquedas.
>
> Como el GT 29 ha declarado, el historial de búsquedas en internet, unido a otros atributos como la dirección IP del cliente, u otra información de cliente, posee un alto poder de identificación[87].

Si el incidente hubiera tenido lugar en territorio europeo, el razonamiento jurídico sería el siguiente: en tanto los datos estaban meramente pseudonimizados o disociados, seguían estando sometidos a la Directiva de Protección de Datos. Por ello, el tratamiento de los datos debía ser compatible con la finalidad para la que se recogieron, que por supuesto, no incluía la publicación o el intento de reidentificación. El fallo de seguridad estuvo provocado por considerar que pseudonimizar era igual que anonimizar.

En ambos casos, los adversarios necesitaron una fuente de datos externa para volver a unir los datos con su identificador. De estos casos se extrae que, en lo que atañe a la privacidad, existe una preocupación particular con los nombres y apellidos como identificadores personales, pero en realidad, cualquier patrón suficientemente

---

87 GRUPO DE TRABAJO DEL ARTÍCULO 29. «Opinion 05/2014 on Anonymisation techniques» (2014).

único puede servir para reconocer a una misma persona en otras bases de datos[88].

4.4 CRÍTICA AL CRITERIO DE ANONIMIZACIÓN PROPUESTO POR EL GRUPO DE TRABAJO DEL ARTÍCULO 29

La Opinión del GT 29 comenta diferentes técnicas de anonimización, con niveles de robustez variables. En tanto ninguna de estas técnicas elimina el riesgo de reidentificación de forma total, normalmente será necesario combinar diferentes técnicas.

En el mencionado dictamen, el Grupo de Trabajo propone dos métodos para determinar si un fichero es anónimo:

1) Realizar un análisis sobre el riesgo de reidentificación de los datos.

2) Verificar que el fichero no tenga ninguna de las siguientes propiedades:

i. Capacidad de singularizar a un individuo («*singling out*»), que se define como la capacidad de poder aislar algunos o todos los datos de un mismo individuo dentro del fichero.

ii. Capacidad de asociar al menos, dos datos pertenecientes a un mismo sujeto o grupo de sujetos, ya sea en el mismo fichero o en dos ficheros diferentes («*linkability*»). El GT 29 continúa explicando que la diferencia entre este atributo y el anterior es que si un adversario es capaz de determinar que un dato se refiere al mismo grupo de individuos, pero no puede singularizar a los individuos de dicho grupo, se habrá producido una asociación, pero no una singularización.

iii. Capacidad de inferir nuevos datos sobre individuos («*inference*»): esto es, poder deducir, con una probabilidad

---

88 Solon BAROCCAS y Helen NISSEBAUM. «Privacy, big data and the public good. Chapter 2: Big data's End Run Around Anonimity And Consent». *Cambrigde University Press* (2014).

significativamente alta, un nuevo dato a partir de los datos supuestamente «anonimizados».

Estos métodos no se derivan de la Directiva ni de la Propuesta de Reglamento, y además comportarían reducir la utilidad de los datos de forma extrema, de modo que parece importante detenerse en este extremo.

Parece razonable el criterio del GT 29 de exigir que un fichero anónimo no permita singularizar a un individuo concreto a partir de los datos de dicho fichero. Ciertamente, esta propiedad implicaría que un individuo es identificable a través de los datos contenidos en el fichero, y que por tanto, dichos datos constituirían datos de carácter personal, y en ningún caso serían datos anónimos.

Por contra, a mi entender, es necesario matizar el razonamiento del GT 29 en lo que se refiere a no permitir que un fichero posea las propiedades de asociación e inferencia.

(i) LA ASOCIACIÓN DE DATOS PERTENECIENTES AL MISMO INDIVIDUO

La Opinión del GT 29 considera que la asociación de datos («*linkability*») es algo negativo, entendiendo por esto la capacidad de asociar al menos, dos datos pertenecientes a un mismo sujeto o grupo de sujetos, ya sea en el mismo fichero o en dos ficheros diferentes.

En primer lugar, cabe mencionar que la información que permite identificar a un grupo de sujetos, pero no a un individuo particular dentro de dicho grupo no es información personal. Esto se desprende de la propia definición de datos de carácter personal, y así lo han asumido otras instituciones de protección de datos, como el organismo inglés Information Commissioner's Office (ICO)[89].

Khaled El Emam y Cecilia Álvarez resaltan que poder vincular diferentes datos a un mismo individuo *dentro de un mismo fichero* es fundamental para poder crear sets de datos longitudinales. Así, en

---

[89] INFORMATION COMMISSIONER'S OFFICE (ICO, autoridad de protección de datos en Reino Unido). «Anonymisation: managing data protection risk code of practice» (2012).

este punto se oponen al criterio del GT 29 y destacan que existen métodos eficaces de anonimizar datos longitudinales, de modo que no es razonable impedir su creación[90].

Datos longitudinales. Definición

En términos muy sucintos, los estudios longitudinales son aquellos en los que se realizan mediciones repetidas o de seguimiento sobre los individuos[91]. Esto permite estudiar a un grupo de individuos de manera repetida a lo largo de los años.

Los datos longitudinales son muy utilizados en las investigaciones demográficas y médicas. Algunos de sus objetivos son, por ejemplo, describir la evolución de un paciente, ya sea antes o después de iniciar un tratamiento, o llegar a realizar predicciones de determinadas enfermedades[92].

En efecto, ser capaces de asociar, por ejemplo, todos los síntomas que un individuo muestra en sus revisiones médicas a lo largo de los años es la base para crear datos longitudinales. Para ello, es necesario poder asociar a ese mismo individuo dentro del fichero, y todos sus datos asociados.

La Opinión del GT 29 también menciona como un problema del método de pseudonimización que éste no elimine la propiedad de asociación de los datos de los individuos. Pero en realidad, tal y como señalan K. El Emam y C. Álvarez, precisamente una de las principales virtudes de los datos pseudónimos es permitir vincular o asociar datos que pertenecen al mismo sujeto sin necesidad de acudir a los identificadores personales. Con todo, como ya hemos mencionado previamente, la pseudonimización no es un método de anonimización.

---

90 Khaled EL EMAM y Cecilia ÁLVAREZ. «A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques». *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.º 1 (2015).

91 Miguel DELGADO RODRÍGUEZ y Javier LLORCA DÍAZ. «Estudios Longitudinales: concepto y características». *Revista Española de Salud Pública*, Vol. 78, n.º 2 (2004).

92 DOMINGO-SALVANY, BRUGAL PUIG y BARRIO ANTA. «Tratado SET de Trastornos Adictivos». *Sociedad Española de Toxicomanías. Editorial Médica Panamericana* (2006).

El Information Commissioner's Office (ICO) también resalta la importancia de mantener la propiedad de asociar datos para mantener su utilidad, en concreto, para llevar a cabo estudios longitudinales:

> «Aunque los datos pseudonimizados no identifiquen a un individuo en las manos de aquellos que no tienen la «llave», la posibilidad de asociar *múltiples bases de datos* al mismo individuo puede ser un precursor de la identificación. En cualquier caso, esto no significa que una anonimización efectiva realizada a través de pseudonimización sea imposible. El Comisionado de Información [ICO] reconoce que *determinados tipos de investigaciones, por ejemplo, los estudios longitudinales, únicamente pueden desarrollarse cuando diferentes datos pueden ser asociados de forma fiable al mismo individuo* (...)[93]». (Cursiva añadida).

Es decir, el ICO reconoce el riesgo que supone para la privacidad asociar datos que se encuentren en bases de datos diferentes. Sin embargo, cabe entender que, cuando se encuentran en la misma base de datos, mantener la capacidad de asociación de los datos permite obtener nuevos conocimientos que podrán ser utilizados de forma beneficiosa.

(ii) LA INFERENCIA

De forma similar a lo expuesto sobre la propiedad de asociación, el GT 29 también considera que la capacidad para realizar inferencias sobre los datos es algo negativo. La Opinión define la inferencia como la posibilidad de deducir, con un nivel significativo de probabilidad, el valor de un atributo a partir de los valores de un conjunto de otros atributos.

Restringir la capacidad de realizar inferencias tiene enormes consecuencias para el *big data*. Detengámonos una vez más es la estadística para comprender el concepto inferencia y su importancia.

---

[93] INFORMATION COMMISSIONER'S OFFICE (ICO). «Anonymisation: managing data protection risk code of practice» (2012).

Los métodos de inferencia estadística se pueden dividir en dos: los métodos de estimación de parámetros y los de contraste de hipótesis. Pues bien, cuando estimamos un parámetro, se comete un error de estimación, que es la diferencia entre nuestra estimación y el valor real del parámetro. Con nuestra estimación y este error se construye un intervalo de confianza, que es la probabilidad de que nuestro modelo estimado contenga el verdadero valor del parámetro.

Recordemos las características con las que definíamos el *big data*: volumen, variedad, velocidad. La enorme variedad de fuentes de datos, y la velocidad a la que se crean, provocan que los datos que sirven de base al *big data* sean, en muchas ocasiones, caóticos, incompletos, con ruido, o no representativos. Esto supone un reto para la estadística, que se puede suplir en parte porque, gracias al mayor volumen de datos que permite procesar el *big data*, las estimaciones que podemos realizar son más certeras. Esto es así porque un modelo es simplemente una representación de la realidad. Hay aspectos del modelo que son iguales a la realidad, pero otros aspectos no se encuentran en el modelo. El *big data* permite que, al disponer de un mayor volumen de datos, el modelo esté mejor representado. En todo caso, todos los modelos tienen un cierto grado de inexactitud o margen de error.

Por su parte, las tecnologías *big data* hacen uso de la minería de datos para buscar correlaciones en los datos almacenados. El profesor Aluja utiliza la definición de minería de datos ya adelantada por Hans (1998): «el proceso de análisis secundario de grandes bases de datos con el objetivo de encontrar relaciones insospechadas que son de interés o aportan valor al titular de la base de datos[94]».

La minería de datos es, en realidad, una evolución de la estadística de análisis de datos, que se sirve de la integración de algoritmos y la automatización del proceso. Así, una de sus aplicaciones más importantes es buscar asociaciones entre sucesos. Es decir, dar respuesta a la pregunta de ¿podemos inferir que determinados sucesos ocurren simultáneamente más de lo que sería esperable si fuesen independientes?[95]

---

94 Tomás ALUJA. «La minería de datos, entre la estadística y la Inteligencia Artificial». *Qüestiió, Universidad Politécnica de Catalunya,* Vol. 25, n.º 3 (2001).

95 Ídem.

Este proceso tiene aplicaciones en una infinidad de campos, desde el marketing (por ejemplo, sabiendo que un consumidor ha comprado el producto X, ¿podrá interesarle el producto Y?) hasta las investigaciones del proyecto Genoma (por ejemplo ¿qué secuencias de genes motivan la aparición de enfermedades?).

Lo que para algunos constituye la mayor amenaza para la privacidad es, irónicamente, lo que despierta mayor interés del *big data*: su capacidad de detectar correlaciones ocultas entre datos y así inferir conclusiones no evidentes a simple vista. En efecto, uno de los usos principales del *big data* es poder hacer un uso secundario de los datos que permita realizar inferencias para obtener nuevos conocimientos.

Sin embargo, la inferencia puede dar lugar a dos tipos de datos nuevos. Esta es la base del argumento de K. El Eman y C. Álvarez, cuando afirman que la Opinión del GT 29 trata dos tipos de inferencias de forma conjunta, a pesar de que normalmente no van juntas en la práctica: la revelación de la identidad de un sujeto, y la revelación de un atributo. En concreto, el GT 29 afirma que:

> «Debe quedar patente que "identificación" no solo significa la posibilidad de recuperar el nombre o la dirección de una persona, sino que también incluye la identificación potencial por singularización, asociación e inferencia» [96].

Por un lado, la revelación de la identidad de un sujeto se da cuando un adversario es capaz de asignar una identidad correcta a un dato. Es a esto a lo que nos referimos cuando hablamos de identificación. De este modo, una base de datos anónima es aquella en que la probabilidad de inferir la identidad de un sujeto es muy baja. Por otro lado nos encontramos con las revelaciones de atributos. Esto es, cuando un adversario es capaz de aprender algo nuevo de los sujetos gracias al análisis de los datos. Cuando estamos frente a múltiples variables, éste es un proceso complejo que se lleva a cabo mediante minería de datos y aprendizaje computacional («*machine learning*»). Un ejemplo podría ser construir un modelo a partir de variables como la

---

[96] GRUPO DE TRABAJO DEL ARTÍCULO 29. «Opinion 05/2014 on Anonymisation techniques» (2014).

edad, sexo, diagnósticos previos del paciente y síntomas, para predecir la probabilidad de padecer un tipo determinado de cáncer[97].

En este sentido, Brian Dalessandro ya afirmó que:

> «Se pueden predecir una gran cantidad de cosas sobre las acciones de una persona sin necesidad de conocer nada personal sobre ellas»[98].

Ésta es una afirmación de enorme trascendencia en la época de los datos masivos. Las conclusiones obtenidas a través del análisis de datos con técnicas de *big data* pueden mostrar nuevos hechos sobre los individuos a pesar de no tener conocimiento alguno sobre su identidad.

En lugar de cruzar entradas de datos asociados al mismo nombre u otra información personalmente identificable, la minería de datos arroja conclusiones que permiten que las compañías simplemente infieran estas características. Esto abre las puertas a que las cualidades que pueden ser inferidas vayan mucho más allá de la información que resida en las bases de datos.

Esto también explica las declaraciones que hace unos años realizó un ingeniero de Google:

> «No queremos el nombre. El nombre es ruido. Hay suficiente información en las enormes bases de datos de Google sobre búsquedas en internet, localización y comportamiento online que se puede saber mucho sobre una persona de forma indirecta»[99].

Una vez que hemos sentado las bases de cómo funciona el proceso de inferencia, surge la siguiente pregunta: ¿es el *big data* compatible

---

97 Khaled EL EMAM y Cecilia ÁLVAREZ. «A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques». *International Data Privacy Law Journal*, Oxford University Press, (2015).

98 Brian DALESSANDRO. «The science of privacy» (2013); citado por Solon BAROCCAS y Helen NISSEBAUM, «Privacy, big data and the public good; Chapter 2: Big data's End Run Around Anonimity And Consent», publicado en *Cambrigde University Press* (2014). La cita del trabajo original no ha podido ser contrastada.

99 Cita recogida por Quentin HARDY. «Rethinking privacy in an era of big data». *New York Times* (4 de junio de 2012).

con la privacidad? Y en su caso, ¿qué implicaciones legales puede tener la mayor facilidad de asociar patrones de conducta y otros datos con un mismo individuo a pesar de no conocer su identidad?

Para responder a esta pregunta me retrotraeré a la definición legal antes expuesta:

> «los principios de la protección deberán aplicarse a cualquier información relativa a una *persona* identificada o identificable (...)» [100]. (Cursiva añadida).

La norma parece establecer el requisito de que los datos identificables sobre «personas» se sometan a la normativa de protección de datos. Sin embargo, nada dice sobre la posibilidad de hacer inferencias sobre atributos (tendencias de consumo, gustos, etc). De este modo, si las técnicas *big data* o de minería de datos se utilizan simplemente para construir un modelo que permita inferir nuevas características o patrones de conducta sin ligarlas a la identidad de una persona concreta, no existe un tratamiento de datos personales.

Por otra parte, y como ya fue adelantado, el GT 29 proponía un segundo medio para determinar si un fichero es anónimo: realizar un análisis sobre el riesgo de reidentificar los datos. En este sentido, basta señalar que inferir datos no atenta contra la anonimización, precisamente porque a través del análisis del riesgo se detectará la posibilidad real de identificar a un sujeto o grupo de sujetos dentro del set de datos. De este modo, siguiendo esta aproximación también podremos utilizar el *big data* sin que las inferencias de nuevos datos resulten en datos personales que deban someterse a la normativa.

En resumen, la capacidad de deducir nuevas características o patrones de conducta no debe ser vista como un agravio a la privacidad de las personas, y los modelos de anonimización no deben tener como objetivo impedir estas inferencias.

Lo explicado hasta ahora se ha correspondido con lo que definimos como la primera fase del *big data*; esto es, un proceso que ha consistido en tomar nuestros datos, insertarlos en los algoritmos y

---

[100] Considerando 26 de la Directiva 95/46/EC del Parlamento Europeo y del Consejo, de 24 de octubre de 1995 relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

construir un modelo que nos permite realizar inferencias. A partir de este momento entramos en la segunda fase del *big data*, en la que el nuevo conocimiento se puede utilizar para tomar decisiones. Es en este paso cuando puede surgir un mayor riesgo para la privacidad de las personas.

Las decisiones se pueden tomar sobre grupos de individuos, sin conocer la identidad de cada persona (por ejemplo, cuando el Gobierno pone en marcha una política de vacunación contra un nuevo virus), o sobre personas determinadas (por ejemplo, cuando las personas en situación de riesgo de infección reciben información personalizada sobre su estilo de vida y medidas de precaución para evitar el contagio).

Como puntualizan K. El Eman y C. Álvarez, el modelo puede utilizarse para tomar decisiones beneficiosas como las que acabamos de describir, pero también decisiones que impliquen consecuencias negativas como discriminación. Por ejemplo, si se ha estimado que una persona tiene un alto riesgo de padecer una enfermedad rara y se le calcula una esperanza de vida de 40 años, esta información puede utilizarse para que a la persona le sea negada una beca de estudios o un tratamiento médico costoso[101].

En conclusión, como numerosos autores defienden, el problema no es la capacidad de realizar inferencias, ni el modelo que se crea con el nuevo conocimiento. El problema es el uso que se hace de ese modelo. Lo que se considere un uso adecuado en cada momento dependerá de las normas y los patrones socialmente aceptados, y será por lo tanto una cuestión subjetiva que habrá de ser analizada caso por caso.

La solución propuesta por K. El Eman y C. Álvarez es, por tanto, diferente a la propuesta por el GT 29. La Opinión del GT 29 recoge medidas basadas en algoritmos cuyo objetivo último es provocar una distorsión en los datos que no permita ni la singularización, ni la asociación ni la inferencia. Sin embargo, a la luz de todo lo expuesto, queda claro que esta solución reduciría la utilidad de los datos a la mínima expresión, impidiendo obtener todos los beneficios que los

---

101 Ejemplo de Javier Aparicio Salom. Discusiones mantenidas sobre una versión preliminar de este trabajo.

nuevos conocimientos nos aportan. La solución que proponen estos autores es la aplicación de un modelo de «ética de la privacidad» [102].

En la práctica, esto se lleva a cabo mediante la creación de un consejo ético en el seno de la organización responsable de los datos, que trabaje de forma independiente. Este consejo estaría formado por un representante de las personas sobre las que se tomarán las decisiones, un experto en privacidad y en ética, un representante de la empresa y un representante de la marca pública. Entre los criterios a tener en cuenta, se proponen algunos de los que el GT 29 ya ha propuesto en otros informes y trabajos:

- la relación entre el fin para el que se tomaron los datos y el fin del modelo actual;
- el contexto en el que se obtuvieron los datos y las expectativas de los sujetos;
- la naturaleza de los datos y su posible impacto;
- las medidas de salvaguarda que se apliquen, como la imposición de condiciones de uso de los datos cuando éstos se cedan a un tercero.

A mi entender, especial importancia merece este último criterio. En todo caso, será la combinación de estas medidas lo que pueda garantizar de forma mucho más precisa que el uso que se haga de los datos sea legal y legítimo, de modo que se puedan obtener los beneficios sociales derivados de este uso, al tiempo que se protege la privacidad de las personas.

Recurriendo una vez más a los términos estadísticos, esta situación es lo que la teoría de juegos definiría como una situación *win-win* (aquella en la que todos ganan).

Una vez analizados los criterios de anonimización que maneja el GT 29, de haberlos sometido a crítica y de haber propuesto una solución alternativa, pasemos ahora a desarrollar las técnicas de anonimización que se exponen en la Opinión del GT 29.

---

[102] Khaled EL EMAM y Cecilia ÁLVAREZ. «A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques». *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.º 1 (2015).

(iii) TÉCNICAS DE ANONIMIZACIÓN

La Opinión del GT 29 sobre la anonimización afirma que, en términos generales, existen dos aproximaciones a las técnicas de anonimización: la randomización y la generalización.

La randomización incluye a la familia de técnicas que alteran la veracidad de los datos con el objetivo de eliminar la fuerte asociación que existe entre los datos y el individuo al que se refieren. De esta forma, cuando los datos son suficientemente inciertos, no podrán ser asociados a una persona concreta.

Las técnicas más utilizadas de randomización son la adición de ruido y la permutación. La adición de ruido consiste en modificar los datos de forma que sean menos precisos, pero manteniendo la distribución general de los datos. Así, por ejemplo, si nuestra base de datos recoge las alturas de los individuos de un estudio de forma exacta, la adición de ruido podría consistir en modificar los datos para que tuvieran una precisión de ±5 centímetros.

Por su parte, la permutación consiste en intercambiar los atributos de los individuos, de forma que quedarían ligados artificialmente a otros sujetos. Esta técnica también permite mantener la distribución de valores, aunque la correlación entre individuos se modificará.

La segunda familia de técnicas de anonimización es la generalización. Consiste en generalizar o diluir los atributos de los sujetos, modificando su escala (por ejemplo, haciendo referencia a un país en vez de una ciudad; o a datos mensuales en vez de semanales).

Veamos un ejemplo de cómo ha funcionado la anonimización tradicionalmente, antes de la llegada de los datos masivos.

(iv) EJEMPLO

Imaginemos una base de datos de un hospital en la que se almacena información sobre las visitas y quejas de los pacientes[103]. En concreto, los datos relativos a la salud de las personas pueden resultar

---

[103] Adaptación del ejemplo utilizado por Paul OHM. «Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization». *UCLA Law Review*, Vol. 57 (2010).

especialmente sensibles, al tiempo que la utilidad de las conclusiones que se puedan extraer sobre esos datos también son de especial relevancia, pues pueden ayudar a elaborar un diagnóstico temprano o acertar con el mejor tratamiento para el caso concreto.

Tabla 1. Datos originales

Nombre	Raza	Fecha nacimiento	Sexo	Código postal	Patología
Ana	Negra	20/09/1965	F	02141	Problema respiratorio
Daniel	Negra	14/02/1965	M	02141	Dolor de pecho
Catalina	Negra	23/10/1965	F	02138	Problema de vista
María	Negra	24/08/1965	F	02138	Tos
Elena	Negra	07/11/1965	F	02138	Dolor articulaciones
Rosa	Negra	01/12/1964	F	02138	Dolor de pecho
Carmen	Blanca	23/10/1964	F	02138	Problema respiratorio
Hilaria	Blanca	15/03/1965	F	02139	Hipertensión
Felipe	Blanca	13/08/1964	M	02139	Dolor articulaciones
Jaime	Blanca	05/05/1964	M	02139	Fiebre
Sergio	Blanca	13/02/1967	M	02138	Vómitos
Adrián	Blanca	21/03/1967	M	02138	Dolor de espalda

Para tratar los datos protegiendo la privacidad de los pacientes, se llevan a cabo los siguientes pasos para anonimizar la base de datos:

1. *Determinar qué información es de identificación personal*: se trata de analizar qué campos pueden permitir una identificación personal de los pacientes. En esta definición entran los identificadores personales básicos como el nombre, pero también la combinación de ellos que puedan asociar una entrada con un paciente concreto. En ocasiones podremos acudir a otras fuentes como estadísticas, políticas internas de la organización o normativa para saber qué campos pueden ser identificadores personales.

En nuestro ejemplo, supongamos que el responsable del tratamiento del fichero concluye que los potenciales identificadores son: nombre, fecha de nacimiento, sexo y código postal.

2. *Supresión*: seguidamente, el responsable puede eliminar los campos mencionados, de forma que dé lugar a la siguiente tabla:

Tabla 2. Suprimir identificadores

Raza	Patología
Negra	Problema respiratorio
Negra	Dolor de pecho
Negra	Problema de vista
Negra	Tos
Negra	Dolor articulaciones
Negra	Dolor de pecho
Blanca	Problema respiratorio
Blanca	Hipertensión
Blanca	Dolor articulaciones
Blanca	Fiebre
Blanca	Vómitos
Blanca	Dolor de espalda

Tras esta operación, los pacientes no tienen de qué preocuparse en lo relativo a su privacidad. Sin embargo, los datos han perdido prácticamente toda su utilidad. Se podría analizar la influencia de cada enfermedad en función de la raza de cada persona, pero no pueden extraerse muchas más conclusiones.

3. *Randomización o generalización*: para mantener un mejor equilibrio entre el derecho a la privacidad y la utilidad de los datos, el responsable del tratamiento puede utilizar las técnicas de randomización o generalización a las que hacíamos mención en lugar de eliminar los datos. Por ejemplo, se puede suprimir el nombre, generalizar la fecha nacimiento únicamente por el año, y generalizar los códigos

postales dejando las tres primeras cifras. La tabla resultante sería la siguiente:

Tabla 3. Randomizar o generalizar identificadores

Raza	Fecha nacimiento	Sexo	Código postal	Patología
Negra	1965	F	021*	Problema respiratorio
Negra	1965	M	021*	Dolor de pecho
Negra	1965	F	021*	Problema de vista
Negra	1965	F	021*	Tos
Negra	1965	F	021*	Dolor articulaciones
Negra	1964	F	021*	Dolor de pecho
Blanca	1964	F	021*	Problema respiratorio
Blanca	1965	F	021*	Hipertensión
Blanca	1964	M	021*	Dolor articulaciones
Blanca	1964	M	021*	Fiebre
Blanca	1967	M	021*	Vómitos
Blanca	1967	M	021*	Dolor de espalda

Con esta tabla los resultados son más complicados de reidentificar que en la tabla original (tabla 1), y la utilidad de los datos para los investigadores es mucho mayor que suprimiendo todos los campos de información de identificación personal (tabla 2).

4. *Agregación*: en numerosas ocasiones, los investigadores o analistas únicamente necesitan estadísticas de datos, sin ser necesario trabajar con los datos originales (los que están sin anonimizar). De esta forma, si los investigadores solamente necesitan conocer cuántas mujeres tuvieron problemas de respiración, el responsable del tratamiento de datos podría proveer los datos agregados sobre las entradas:

Tabla 4. Estadística agregada

Mujeres con problemas respiratorios	2

Con estos datos, serían necesarios muchos otros datos adicionales para deducir que Carmen es una de las dos personas a las que se refiere la estadística. Su privacidad está protegida.

4.5 RIESGO DE REIDENTIFICACIÓN

Recordemos que la Opinión del GT 29 proponía dos criterios para determinar si una base de datos es anónima o no: (i) verificar que el fichero no tenga las propiedades de singularización, asociación e inferencia, y (ii) realizar un análisis sobre el riesgo de reidentificación de los datos.

Hasta ahora hemos repasado el primer criterio, para pasar ahora a analizar el segundo criterio propuesto.

Como adelantábamos en la introducción de este capítulo, la llegada de tecnologías como el *big data* y la minería de datos ha traído serias dudas sobre el poder de la anonimización.

De hecho, la promesa de alcanzar la anonimización absoluta es imposible de cumplir especialmente por dos circunstancias. Primera, porque aunque los datos no contengan información que pueda ser considerada de identificación personal (PII), en ocasiones dichos datos continúan siendo capaces de diferenciar a una persona de forma única, de modo tal que se puedan asociar esos datos a una persona concreta. Así por ejemplo, cuando los datos contienen información extremadamente rica (como por ejemplo datos de geolocalización), se puede identificar a una persona. Y segunda, porque cada vez son más frecuentes y sencillos los denominados ataques de reidentificación.

En palabras de Ohm, «los datos pueden ser útiles o perfectamente anónimos, pero nunca ambos»[104].

Esto no quiere decir que ninguna técnica de anonimización sirva para proteger la privacidad de las personas, pues algunas técnicas son verdaderamente difíciles de revertir. Sin embargo, la tecnología ha avanzado y los investigadores (entre otros) han demostrado de mane-

---

104 Paul OHM. «Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization». *UCLA Law Review*, Vol. 57 (2010).

ra más que suficiente que la anonimización ya no puede ser considerada la panacea de la protección de datos y la privacidad.

Veamos dos de los casos más paradigmáticos de cómo bases de datos que habían sido anonimizadas sucumbieron ante una ataque de reidentificación, poniendo en entredicho las técnicas utilizadas.

Caso GIC: identificación por el trío código postal, fecha de nacimiento y sexo

A mediados de los años 90, en Massachusetts, el denominado Group Insurance Commission (GIC) decidió hacer públicos datos relativos a las visitas al hospital de los funcionarios públicos, con el objetivo de que los investigadores pudieran analizarlos y sacar conclusiones. El GIC «anonimizó» previamente los datos, mediante la eliminación de identificadores personales explícitos, tales como el nombre, la dirección y el número de la Seguridad Social. Pese a ello, cerca de cien atributos de cada paciente y hospital permanecieron en los datos hechos públicos; entre éstos, el código postal, la fecha de nacimiento y el sexo de los individuos. En ese momento, el Gobernador de Massachussets, William Weld, aseguró públicamente que la privacidad de los individuos estaba asegurada.

Lantaya Sweeny, directora del Laboratorio de Privacidad de la Universidad de Harvard, realizó un estudio cuyo objetivo era poner de manifiesto las limitaciones de las normas de privacidad[105] y las medidas de seguridad para poder implementar mejoras mediante el uso de algoritmos más fuertes y complejos.

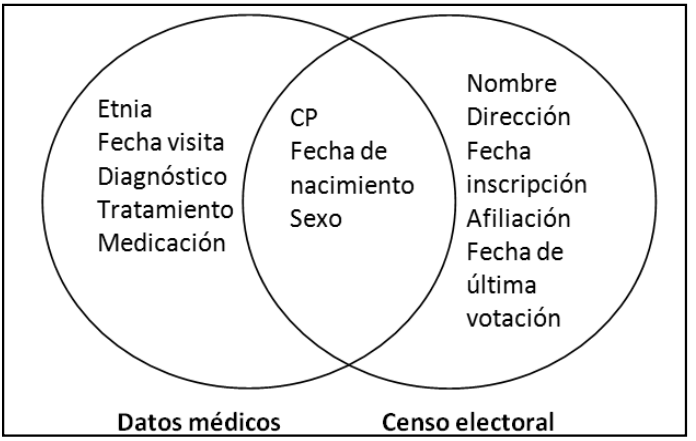
Sweeny pidió una copia de los datos publicados y comenzó a intentar reidentificar los datos del Gobernador. Sabía que éste residía en la ciudad de Cambridge (Massachusetts), una ciudad de 54.000 residentes y siete códigos postales. Además, pagando 20 dólares compró el último censo electoral de la ciudad de Cambridge, que contenía, entre otros datos, el

---

105 En concreto, el estudio de Sweeny se refiere a la norma de privacidad estadounidense HIPAA, que establece medidas para proteger la privacidad de datos médicos y de salud; no obstante, las conclusiones son extrapolables a otras jurisdicciones.

nombre, dirección, código postal, fecha de nacimiento y sexo de cada votante.

Combinando ambas bases de datos, los datos de GIC y el censo electoral, Sweeny consiguió identificar al Gobernador Weld sin dificultad: únicamente seis personas en Cambridge habían nacido el mismo día que el Gobernador, solo tres de estas personas eran hombres, y solamente él vivía en su código postal. Sweeny envió el historial médico del Gobernador, que incluía diagnósticos y prescripciones, a su oficina.



Pero Sweeny no se quedó ahí. Extendió su análisis hasta concluir que el 87.1% de los ciudadanos residentes en Estados Unidos son identificables mediante este trío de atributos: código postal, fecha de nacimiento y sexo.

Las conclusiones del experimento de Sweeny sobre este trío de identificadores han vuelto a ser analizadas y actualizadas. En un estudio llevado a cabo en la Universidad de Standford, se comprobó que en 2006 la proporción de personas residentes en Estados Unidos que podían ser reidentificadas utilizando únicamente la triada código postal, fecha de nacimiento y sexo había descendido al 63.3%. No obstante, el nuevo estudio volvió a demostrar que los ataques de reidentificación siguen siendo fáciles de realizar. Además, es necesario tener en cuenta que la disponibilidad de información que actualmente es pública es mucho mayor que en el momento en que se realizó el

estudio, y las técnicas de reidentificación son más exactas, de modo que es fácil imaginar que un atacante pueda tener acceso a más datos que el código postal, la fecha de nacimiento y el sexo de las personas cuyos datos se encuentran en la base de datos.

Otro de los casos más conocidos que pusieron de manifiesto las limitaciones de la anonimización (en este caso llevada a cabo mediante randomización) es el acaecido en relación a Premio Netflix[106].

Caso Netflix:

Netflix.Inc es la mayor empresa del mundo proveedora de una tarifa plana mensual multimedia para ver películas y series de televisión. En octubre de 2006, la compañía lanzó el denominado Premio Netflix. La empresa hizo públicos cien millones de registros de películas de 500.000 usuarios, y ofreció una recompensa a aquel que consiguiera mejorar su servicio de recomendación de películas (que se basa en las películas que otros usuarios con gustos similares puntuaron de forma muy alta). Los datos habían sido anonimizados, de forma que se eliminaron todos los identificadores personales excepto las calificaciones de las películas y la fecha de la calificación; además, se añadió ruido, de forma que las calificaciones de los usuarios fueron ligeramente incrementadas o reducidas.

Como fuente de datos externa se utilizó la base de datos pública de Internet Movie Database (IMB), una base de datos *online* que almacena información relacionada con películas.

La cuestión de la que partía el experimento era: ¿cuánto tiene que saber un adversario sobre un suscriptor de Netflix para poder identificar sus datos en la base de datos, y así, conocer su historial completo de películas? Es decir, en términos analíticos el estudio se basó en calcular el tamaño de los datos auxiliares que eran necesarios para reidentificar a los sujetos supuestamente anonimizados.

---

106 Arvind NARAYANAN y Vitaly SHMATIKOV. «Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)». *The University of Texas at Austin* (2008).

En este sentido, cabría preguntarse si realmente un suscriptor de Netflix considera privado su historial de películas vistas. Incluso aunque la respuesta fuera negativa (lo cual no se puede asumir), eso sería así solo en la medida en que no comprendamos las consecuencias reales de reidentificar estos datos.

Tal y como demostró el experimento, la correlación encontrada entre la base de datos anonimizada del Premio Netflix y la base de datos pública de IMB permite conocer información sensible y no pública sobre una persona, tal como preferencias políticas u orientación sexual.

En efecto, se logró identificar a uno de los usuarios, una madre de familia lesbiana que mantenía su orientación sexual en secreto, residente en una región muy conservadora de Estados Unidos, y que demandó a la empresa bajo el pseudónimo de Jane Doe.

Tras el escándalo, investigadores de la universidad de Texas compararon los datos de Netflix con otros datos públicos sobre calificación de películas. El estudio demostró que un usuario que hubiera calificado tan solo seis películas poco conocidas (aunque de la lista de las quinientas primeras películas), podría ser identificado en el 84% de los casos. Y esta proporción aumentaba al 99% de los casos si, además, se sabía en qué fecha se habían calificado las películas. Así, se demostró que la calificación de las películas creaba una huella personal.

Antes de este momento, nadie habría definido las calificaciones a películas como datos de identificación personal.

En ambos casos fue necesario combinar dos bases de datos que contenían datos parciales sobre las personas, y que muestra el principio de que, a pesar de que una base de datos parezca anónima, cuando se compara con una segunda base de datos, se encuentra información única sobre los sujetos, y la reidentificación de éstos se hace posible.

A esta información única sobre cada persona se le ha denominado «huella», es decir, combinaciones de valores de datos que una persona no comparte con nadie más incluido en la base de datos y que, por tanto, permiten identificarla.

(i) HUELLA DIGITAL

Esta huella ha generado innumerables debates y polémicas sobre lo que es información de identificación personal. Sin embargo, los ejemplos anteriores ponen de manifiesto una problemática que trasciende este hecho: se han encontrado huellas personales en datos que hasta entonces no habían sido considerados de identificación personal.

Es decir, a pesar de que en una base de datos no aparezcan nombres, se aprecian patrones. Y con estos patrones, una persona con suficientes conocimientos analíticos puede obtener nombres.

Así, cuando se trata de *big data*, 2+2 no son 4. Las sinergias que se forman al unir bases de datos diferentes, provocan que el resultado sea mayor que la mera suma de las partes.

El proyecto de la Electronic Frontier Foundation bautizado como *«How unique —and trackable— is your browser?»* («¿Cómo de único es tu navegador?»), analiza los datos del navegador que crean la huella digital de dispositivo para inferir cómo de única es, y qué capacidad tiene de singularizarnos ante otros usuarios.

Utilizando su herramienta denominada Panopticlik[107], se puede acceder a la información que nuestro navegador entrega voluntariamente cuando visitamos páginas webs. Haciendo una prueba con mi ordenador personal, los resultados son los siguientes[108]:

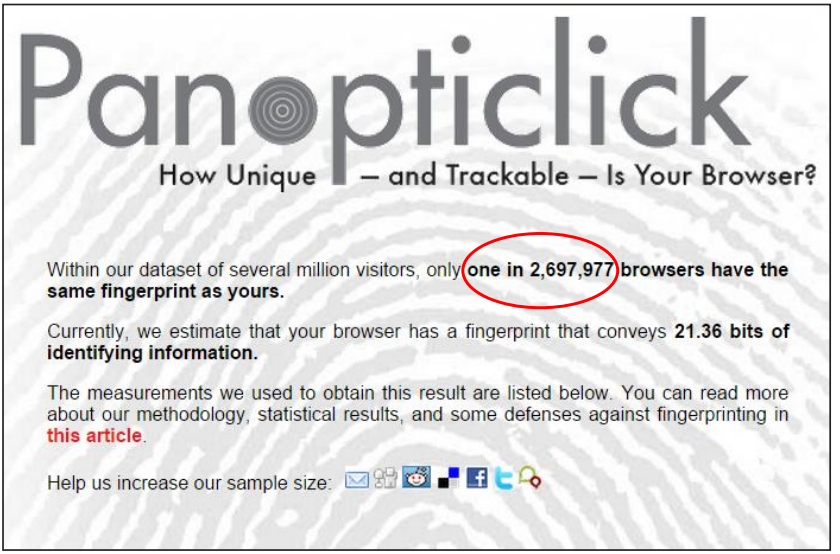
Antes de haber borrado las *cookies* de mi ordenador, el navegador es identificable entre 5,3 millones de usuarios.

Tras haber borrado mis *cookies*, mi navegador es único entre prácticamente 2,7 millones de usuarios. Es decir, parece que eliminar las *cookies* favorece que nuestro ordenador sea menos único, y por tanto, más anónimo. Sin embargo, es muy destacable que, aun sin *cookies* instaladas (al menos las *cookies* visibles), mi equipo siga siendo perfectamente singularizable entre tantos usuarios.

---

107 Panopticlick. Proyecto «How unique —and trackable— is your browser?» Disponible en: https://panopticlick.eff.org/

108 El experimento fue realizado a mero título orientativo, con fecha 25 de mayo de 2015.



(ii) TEST DE REIDENTIFICACIÓN ¿QUIÉN ES EL ADVERSARIO?

La jerga común ha aceptado denominar «adversario» a la persona o entidad que intenta llevar a cabo un proceso o ataque de reidentificación. De acuerdo con la Directiva:

> «(...) para determinar si una persona es identificable, hay que considerar el conjunto de los medios que puedan ser razonablemente utilizados por el *responsable del tratamiento* o por *cualquier otra persona*, para identificar a dicha persona»[109]. (Cursiva añadida).

La Opinión del GT 29 sobre técnicas de anonimización utiliza estos mismos términos. De ello se desprende que el adversario que puede realizar un ataque de reidentificación puede ser, bien el responsable del tratamiento, o bien cualquier otra persona.

---

[109] Considerando 26 de la Directiva 95/46/EC del Parlamento Europeo y del Consejo, de 24 de octubre de 1995 relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.

Sin embargo, la forma en la que se interprete quién puede ser esa otra persona tiene importantes consecuencias. Siguiendo de nuevo a K. El Eman y C. Álvarez[110], tomaremos el siguiente ejemplo para poner de manifiesto las posibles incongruencias que surgen al utilizar el concepto de «cualquier otra persona».

Ejemplo: importancia de contextualizar quién es el adversario

Imaginemos un hospital que cede datos previamente anonimizados a una compañía farmacéutica, que a su vez cuenta con altas medidas de seguridad, auditadas, y que sigue las mejores prácticas de conducta en lo que respecta al tratamiento de los datos. En este entorno, la probabilidad de que se llegue a realizar una reidentificación de los individuos, ya sea de forma deliberada o accidental, es muy baja.

De forma paralela, sabemos que existe la Profesora Slocum (personaje ficticio), conocida por su habilidad para reidentificar bases de datos médicas, y por publicar los resultados. En este contexto, la Profesora Slocum es un adversario que encaja en la definición de «cualquier otra persona» y que puede llevar a cabo un ataque de reidentificación.

Una interpretación estricta de la Directiva llevaría a concluir que cuando el hospital cede los datos anonimizados a la empresa farmacéutica se debe tener en cuenta el alto riesgo que representa la existencia de adversarios como la Profesora Slocum. Esto llevaría a utilizar técnicas de anonimización que distorsionen los datos de manera suficiente para que, en caso de que la Profesora Slocum tenga acceso a este fichero, sea incapaz de realizar la reidentificación. Esto, como ya hemos visto, implicaría que la utilidad de los datos quedara muy reducida.

Sin embargo, en la práctica, los datos se ceden únicamente a una entidad concreta como es la empresa farmacéutica. Las altas medidas de seguridad que protegen a estos datos y sus bue-

---

110 Khaled EL EMAM y Cecilia ÁLVAREZ. «A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques». *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.º 1 (2015).

> nas prácticas de la empresa hacen que la probabilidad real de que la Profesora Slocum tenga acceso a estos datos sea extremadamente baja.
>
> Si el hospital tiene la obligación de anonimizar los datos suponiendo en todo momento que la Profesora Slocum tendrá acceso a ellos y sin tener en cuenta el contexto real de la cesión, la utilidad de los datos quedaría muy mermada. Esto a su vez restaría valor a los esfuerzos que la farmacéutica hace por establecer fuertes medidas de seguridad y salvaguardas eficaces.

En efecto, tomar en cuenta el contexto es un criterio que ha sido abrazado por el propio GT 29. Éste es también el criterio propuesto por K. El Eman y C. Álvarez, que a mi entender soluciona los problemas prácticos al tiempo que mantiene la protección de la privacidad. Consiste en considerar que «cualquier otra persona» se refiere a cualquier tercero que se encuentre en el mismo contexto que aquél que va a recibir los datos, y que además sea un «adversario motivado» *(«motivated intruder»* [111]).

El concepto de adversario motivado ha sido desarrollado por la autoridad inglesa ICO, y ha sido ya mencionado en algunos de los trabajos del GT 29. El código de anonimización del ICO define a este adversario motivado como «aquella persona que comienza sin ningún conocimiento previo pero que pretende identificar al sujeto de cuyos datos personales se han obtenido los datos anonimizados» [112].

Se asume que este adversario motivado es una persona razonablemente competente, que tiene acceso a fuentes de datos tales como internet, bibliotecas y todos los documentos públicos, y que utilizará técnicas de investigación. Sin embargo, no cuenta con habilidades técnicas especiales como *hacker*, ni acceso a equipo especializado ni recurrirá a tretas ilegales como el robo de datos. En opinión del ICO, éste es un buen baremo, puesto que sitúa el listón por encima de una persona relativamente inexperta del público general, pero por debajo

---

111 Ídem.

112 INFORMATION COMMISSIONER'S OFFICE. «Anonymisation: managing data protection risk code of practice» (2012).

de una persona especialista, con grandes medios analíticos o con conocimientos previos sobre la persona.

En mi opinión, no obstante, quizás debiera considerarse que este adversario motivado sí tiene conocimientos técnicos y medios analíticos altos. Sí estoy de acuerdo, sin embargo, en considerar que el adversario no recurrirá a medios ilegales de reidentificación.

Así, una persona sin conocimientos técnicos especializados probablemente quede desincentivada a intentar lograr la reidentificación de sujetos si existen salvaguardas para ello. Sin embargo, muchos de los casos en los que el público general tenga acceso a bases de datos anónimas será cuando éstas se hagan accesibles al público sin restricciones. Y en estos casos, sí es necesario aplicar medidas de seguridad más elevadas porque no solo ese público inexperto tendrá acceso, sino también aquellos adversarios más expertos.

Por el contrario, una persona con conocimientos técnicos o medios elevados no quedará desincentivada para llevar a cabo un ataque de reidentificación si sabe que tiene capacidad para realizarlo. Además, ciertas categorías de datos son susceptibles de ser más atractivas que otras, ya sea por su valor económico o por su capacidad de sacar a la luz datos privados que puedan ser utilizados para ridiculizar a una persona, poner sus ideas en entredicho, etc. (por ejemplo, las búsquedas que una persona realiza en su navegador web). Así, en mi opinión, sería necesario considerar que el adversario sí tiene conocimientos o medios relativamente elevados. Sería necesario un análisis más profundo para determinar cómo definir entonces el concepto y el nivel de habilidades de *hackeo* y conocimientos que se le deben suponer a nuestro adversario motivado.

En efecto, nos encontramos en una sociedad que camina cada vez más hacia la transparencia —cuya máxima expresión son las prácticas de datos abiertos («*open data*»)—, y hacia la creación y el tratamiento de datos —gracias al internet de las cosas y la nube («*cloud computing*»)—. En este escenario, establecer un umbral de seguridad alto es una exigencia básica.

En cualquier caso, el código de prácticas de anonimización desarrollado por el ICO contiene recomendaciones de gran utilidad. Ante la falta de una normativa o un código de anonimización desarrollado por las autoridades españolas o comunitarias, algunas empresas

españolas (como Telefónica[113]) se han servido de este código elaborado por el ICO como ayuda para desarrollar sus protocolos internos de anonimización de datos.

En todo caso, ha de tenerse presente que cuando una organización crea datos personales mediante un proceso de reidentificación sin el conocimiento o el consentimiento del individuo, está obteniendo datos personales de forma ilegal, y en cuanto tal, es susceptible de ser multado.

4.6 OTRAS TÉCNICAS DE ANONIMIZACIÓN

Ante los retos de reidentificación, se han desarrollado técnicas complementarias para limitar aún más las posibilidades de deducir la identidad de los sujetos de un set de datos, tales como la privacidad diferencial o la k-anonimización.

Las técnicas englobadas en lo que se denomina privacidad diferencial son un área de investigación importante. Parten de la base de que el riesgo sobre la privacidad de una persona no debería aumentar por el hecho de que sus datos se encuentren en una base de datos. Es decir, que no debería ser posible obtener información acerca de un individuo de una base de datos que no pueda conocerse de otro modo, sin acceso a la base de datos. Así, para el individuo, el hecho de estar presente en las bases no supondría un riesgo añadido contra su privacidad. Éste es sin duda un área de desarrollo que ya está siendo investigada, y donde destacan los trabajos de Cynthia Dwork[114].

Anthony Tockar, de la Universidad de Northwestern, señala que la solución radica en añadir ruido en bases de datos relacionadas, de modo que se enmascare la identidad de cada uno de los individuos, manteniendo la utilidad de los datos y una alta precisión en la información.

Otra de las técnicas que está siendo desarrollada es la denominada k-anonimización. De forma intuitiva, la k-anonimización consiste en generalizar los atributos de diversos sujetos de forma que un número

---

113 Información obtenida en una reunión mantenida con el responsable de análisis de datos de Telefónica (septiembre de 2015).

114 Cynthia DWORK. «Differential Privacy». *Microsoft Research*.

«k» de sujetos comparten el mismo valor. Así, cuanto mayor es el valor «k», mayor es la garantía de privacidad. Esto impide que un sujeto sea individualizado, al menos, dentro de ese grupo[115].

En este capítulo hemos puesto de manifiesto los límites inherentes a la anonimización de datos, que era visto como mecanismo infalible para proteger la privacidad de las personas. Si, como hemos visto, tanto el consentimiento como la anonimización muestran graves limitaciones en la protección de los datos ¿cuál debe ser la solución?

---

115 Lantaya SWEENY. «K-Anonimity, a model for protecting privacy». *International Journal on Uncertainty*, Fuzziness and Knowledge-Based Systems, Vol. 10, n.º 5 (2002).

CAPÍTULO V. *BIG DATA VS.* PROPUESTA DE REGLAMENTO DE PROTECCIÓN DE DATOS

Hasta este momento, la aproximación a la cuestión de los problemas que el *big data* presenta sobre la normativa de protección de datos y la privacidad de las personas se ha basado en las normas actualmente en uso; esto es, la Directiva de Protección de Datos y las normas españolas.

Sin embargo, como ya adelantamos en el Capítulo II al introducir el marco normativo de la protección de datos, la Unión Europea, consciente de los límites de las normas actuales, se encuentra en un proceso de renovación de la Directiva de Protección de Datos. La figura que se ha elegido para la nueva normativa ha sido el Reglamento.

Este Reglamento está siendo discutido actualmente y se espera que durante el presente año 2015 sea aprobado, de modo que no se conoce la versión final del texto. No obstante, la Propuesta de Reglamento que se discute contiene ya las líneas generales de lo que será esta nueva normativa.

En este capítulo vamos a analizar algunas de las medidas que introduce esta Propuesta, examinando si vienen a solucionar los problemas que el *big data* está creando para la privacidad y la protección de datos de las personas.

La Propuesta de Reglamento introduce nuevas medidas para incrementar el poder de control de los sujetos sobre los datos, mediante la exigencia de una mayor transparencia, mediante el fortalecimiento del papel del consentimiento informado para tratar datos personales, declarando nuevos derechos individuales y modificando las provisiones sobre la creación de perfiles de los sujetos. Analicémoslas.

5.1 DEBER DE TRANSPARENCIA Y EL CONSENTIMIENTO

La nueva Propuesta de Reglamento otorga más importancia al principio de transparencia, requiriendo que los datos sean tratados de

forma transparente (artículo 5 de la Propuesta). Además, en el artículo 11 añade que el responsable del tratamiento aplicará políticas transparentes y «fácilmente accesibles» sobre el tratamiento de datos personales.

En cuanto al consentimiento, la Directiva actual obliga a que sea una manifestación de la voluntad libre, específica e informada. La Propuesta amplía su definición y le otorga mayor importancia (artículos 4.8 y 7.1 de la Propuesta). Así, se añade el requisito de que sea explícito. Es decir, el silencio o la falta de acción ya no se podrán considerar como consentimiento válido. Esto ha sido ya señalado en los informes del GT 29, pero recordemos que éstos no tienen fuerza vinculante, de modo que actualmente los agentes son libres de exigir un consentimiento explícito o no. Además, será el responsable del tratamiento quien deba probar que ha obtenido el consentimiento de los sujetos para el procesamiento de los datos.

A pesar de la utilidad de estas reformas, parece difícil que la nueva redacción vaya a solucionar los problemas que plantean las deficiencias del consentimiento informado, en la medida en que ya han sido expuestos en capítulos anteriores; ni que logre que los individuos comprendan sus derechos y actúen sobre ellos.

Tal vez el modelo de consentimiento informado ya no deba ser la piedra angular del tratamiento de datos. En estos términos se han expresado numerosos autores, entre los que destacaré la siguiente cita de Ira Rubinstein:

> «Mi argumento es simple aunque radical: el consentimiento informado está roto, sin posibilidad de que una norma lo repare, y el único modo de fortalecerlo es cambiando los mercados relevantes de información» [116].

En efecto, a pesar de que la nueva normativa provea de mecanismos de transparencia para que las personas puedan ejercer un consentimiento informado, y de que se fortalezca este consentimiento, quizás la medida no sea suficiente.

---

116 Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law,* Vol. 3, n.º 2 (2013).

Bien es cierto que, jurídicamente, el consentimiento es solo una de las condiciones alternativas que permiten el tratamiento de datos personales, en igualdad de importancia a otras tales como la necesidad de cumplir con una obligación legal o que el tratamiento sea necesario para satisfacer el interés legítimo perseguido por el responsable del tratamiento o por el tercero a quien se comuniquen los datos, siempre que no prevalezca el interés o los derechos y libertades fundamentales del interesado.

Sin embargo, en la práctica, el consentimiento no es simplemente una base legal más que legitima el tratamiento de datos, sino que se trata del instrumento principal.

Cualquier sistema en el que la base principal para el tratamiento de datos sea el consentimiento emplaza la responsabilidad en el individuo. A pesar de que las mayores garantías de transparencia que introduce la nueva normativa pueden ayudar a prestar un consentimiento mejor informado, en mi opinión, la carga de la responsabilidad no debe recaer en los individuos. Éstos deben mantener su poder de decisión, y para ello las medidas introducidas por la normativa son adecuadas. Pero, a mi entender, no son suficientes en la medida en que no equilibran la responsabilidad que pesa sobre los individuos. En el último capítulo, dedicado a las conclusiones, volveremos sobre este aspecto.

5.2 CREACIÓN DE PERFILES Y TRATAMIENTO AUTOMATIZADO DE DATOS

La elaboración de perfiles y la toma de decisiones individuales automatizadas ha sido una cuestión ampliamente discutida durante la elaboración de la Propuesta de Reglamento. Esto trae como causa la preocupación de que las nuevas técnicas analíticas sean la base de decisiones trascendentales para nuestra vida, que se toman utilizando algoritmos sin intervención humana.

(i) DEFINIENDO LA PROBLEMÁTICA

El problema que conlleva la toma de decisiones automatizadas basadas únicamente en las conclusiones que lanzan los algoritmos ya fue analizado en el Capítulo I, apartado 1.3 al que nos remitimos.

En lo relativo a la creación de perfiles, la Propuesta la define como toda forma de tratamiento automatizado destinado a evaluar o generar datos sobre aspectos propios de una persona física o a analizar o predecir su rendimiento profesional, su situación económica, su localización, su estado de salud, sus preferencias, su fiabilidad, su conducta o su personalidad (artículo 4 de la Propuesta).

Sin embargo, quizás resulte más útil la definición que da la Agencia Europea de los Derechos Fundamentales.

> «La elaboración de perfiles consiste en categorizar a individuos en función de sus características (tales como género, edad, hábitos y comportamientos). Los individuos son frecuentemente divididos en perfiles por las compañías aseguradoras para calcular su riesgo y sus precios (así por ejemplo, un fumador tendrá un riesgo más alto de tener problemas de salud que un no fumador), así como por empresas de marketing para determinar qué productos ofrecer a cada persona»[117].

De esta forma, la elaboración de perfiles permite que los individuos sean categorizados sobre la base de algunas características observables para así poder inferir otras que no son observables.

La categorización puede ser una herramienta muy útil, pero también entraña riesgos de cometer errores al conectar con una persona determinadas características con ciertos comportamientos. Por su parte, también hay riesgo de que los perfiles basados en características como raza, etnia o religión creen estereotipos poco precisos que causen discriminación.

Normalmente, la creación de perfiles se lleva a cabo en varios pasos. En primer lugar, los datos recogidos se anonimizan. En segundo lugar, se utilizan las técnicas de la minería de datos para conectar los datos y buscar correlaciones que logren crear nuevas categorías de información. Por último, se interpretan los resultados para obtener conclusiones y suposiciones sobre el comportamiento de las personas. Esto se lleva a cabo mediante inferencia, en el sentido en que ya ha sido expuesto al tratar sobre la anonimización. En este punto todavía

---

117 AGENCIA EUROPEA DE LOS DERECHOS FUNDAMENTALES. «Understanding and preventing discriminatory ethnic profiling» (2012).

nos encontramos en la primera fase de desarrollo del *big data*, tal y como ha sido definida anteriormente. Es decir, se ha creado un modelo a partir del cual se pueden tomar decisiones sobre las personas. Una vez más, cabe destacar que el problema no es la creación del modelo en sí mismo, sino el uso positivo o negativo que se pueda hacer de él en la segunda fase, una vez que se aplica a individuos concretos para categorizarlos.

(ii) MODIFICACIONES INTRODUCIDAS POR LA PROPUESTA DE REGLAMENTO

La redacción propuesta en el nuevo artículo 20 vendría a sustituir al artículo 15 de la actual Directiva; sin embargo, parece que no soluciona los problemas que éste planteaba. A los efectos de facilitar el análisis de estos dos artículos, se adjunta una tabla comparativa de ambos textos en la página siguiente.

Directiva 95/46/CE de Protección de Datos	Propuesta de Reglamento General de Protección de Datos
Artículo 15 Decisiones individuales automatizadas	Artículo 20 Medidas basadas en la elaboración de perfiles
1. Los Estados miembros reconocerán a las personas el *derecho a no verse sometidas a una decisión con efectos jurídicos sobre ellas o que les afecte de manera significativa, que se base únicamente en un tratamiento automatizado de datos destinado a evaluar determinados aspectos de su personalidad,* como su rendimiento laboral, crédito, fiabilidad, conducta, etc.	1. Toda persona física tendrá *derecho a no ser objeto de una medida que produzca efectos jurídicos que le conciernan o le afecten de manera significativa y que se base únicamente en un tratamiento automatizado destinado a evaluar determinados aspectos personales propios* de dicha persona física o a analizar o predecir en particular su rendimiento profesional, su situación económica, su localización, su estado de salud, sus preferencias personales, su fiabilidad o su comportamiento.
2. Los Estados miembros permitirán, sin perjuicio de lo dispuesto en los demás artículos de la presente Directiva, que una persona pueda verse sometida a una de las decisiones contempladas en el apartado 1 cuando dicha decisión:	2. A reserva de las demás disposiciones del presente Reglamento, una persona solo podrá ser objeto de una medida del tipo contemplado en el apartado 1 si el tratamiento:
a) se haya adoptado en el marco de la celebración o ejecución de un *contrato,* siempre que la petición de celebración o ejecución del contrato presentada por el interesado se haya satisfecho o que existan medidas apropiadas, como la posibilidad de defender su punto de vista, para la salvaguardia de su interés legítimo;	a) se lleva a cabo en el marco de la celebración o la ejecución de un *contrato,* cuando la solicitud de celebración o ejecución del contrato presentada por el interesado haya sido satisfecha ***o*** se hayan invocado medidas adecuadas para salvaguardar los intereses legítimos del interesado, como el derecho a obtener una intervención humana; o
b) esté *autorizada por una ley* que establezca medidas que garanticen el interés legítimo del interesado.	b) está expresamente autorizado por el *Derecho* de la Unión o de un Estado miembro que establezca igualmente medidas adecuadas para salvaguardar los intereses legítimos del interesado; ***o***

Directiva 95/46/CE de Protección de Datos	Propuesta de Reglamento General de Protección de Datos
Artículo 15 Decisiones individuales automatizadas	Artículo 20 Medidas basadas en la elaboración de perfiles
	c) se basa en el *consentimiento* del interesado, a reserva de las condiciones establecidas en el artículo 7 y de garantías adecuadas. 3. El tratamiento automatizado de datos personales destinado a evaluar determinados aspectos personales propios de una persona física *no se basará únicamente en las categorías especiales de datos personales contempladas en el artículo 9*. 4. En los casos contemplados en el apartado 2, *la información que debe facilitar el responsable del tratamiento* en virtud del artículo 14 incluirá información sobre la existencia del tratamiento por una medida del tipo contemplado en el apartado 1 y los efectos previstos de dicho tratamiento en el interesado. 5. La Comisión estará facultada para adoptar actos delegados, de conformidad con lo dispuesto en el artículo 86, a fin de especificar los criterios y condiciones aplicables a las medidas adecuadas para salvaguardar los intereses legítimos del interesado contempladas en el apartado 2.

Así, el actual artículo 15 de la Directiva reconoce el derecho de las personas a no verse sometidas a una decisión con efectos jurídicos sobre ellas o que les afecte de manera significativa, que se base únicamente en un tratamiento automatizado de datos destinado a evaluar determinados aspectos de su personalidad, como su rendimiento laboral, crédito, fiabilidad, conducta, etc. E introduce dos excepciones en las que sí se puedan tomar estas decisiones (que equivalen a crear perfiles individuales de los sujetos): cuando así se haya acordado por contrato o así lo prevea una ley.

Es decir, la Directiva no prohíbe la creación de perfiles, sino que establece límites para que los perfiles no sean creados de forma automatizada sin intervención humana. Ciertamente este artículo tiene limitaciones que necesitan revisión. En primer lugar, la prohibición se aplica exclusivamente cuando se cumplen todos los requisitos que establece la norma. Sin embargo, la redacción acaece de muchos términos ambiguos (tales como «que les afecte de manera *significativa*», «que se base únicamente en un tratamiento automatizado»). Dichas ambigüedades crean problemas en su aplicación práctica. ¿Qué papel debe tener un humano para que se considere que la decisión no es únicamente automatizada?[118]

Además, este derecho solamente puede ejercitarse cuando las personas son conscientes de que las decisiones que les afectan se toman de manera automatizada. Pero como ya hemos mencionado, los procesos analíticos por los que se analizan los datos masivos (fase 1 del *big data*) muchas veces ocurren sin conocimiento de los sujetos, de forma poco transparente. El problema vendría, pues, del hecho de que los individuos en muchas ocasiones no son conscientes de que se están tomando decisiones que les afectan (fase 2 del *big data*) con base en esos procesos automatizados. Por ello, este derecho podría quedar en la práctica en papel mojado. Para ello sería necesario que los individuos prestaran su consentimiento a que se tomen decisiones sobre ellos basadas en la creación de perfiles.

Así, la Propuesta de Reglamento vendría a sustituir este artículo 15 por el nuevo artículo 20, que introduce diversas modificaciones, con el ánimo de otorgar a los individuos más poder sobre sus datos. En

---

118 Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law,* Vol. 3, n.º 2 (2013).

este sentido, define el tratamiento automatizado de forma más amplia (apartado 1) e introduce una nueva excepción a esta prohibición dada por el consentimiento del sujeto (apartado 2). Éste es otro ejemplo más de cómo el consentimiento sigue teniendo un papel crucial en la nueva Propuesta.

Mención especial merece el apartado 3 del nuevo artículo, que prohíbe aquél tratamiento automatizado que se base únicamente en categorías especiales de datos personales. Recordemos que se trata de aquellos datos que, de divulgarse de manera indebida, podrían afectar a la esfera más íntima del ser humano, tales como ideología, afiliación sindical, religión, creencias, origen racial o étnico, salud y orientación sexual.

Algunas enmiendas presentadas a este nuevo artículo incluyen, además, la mención expresa de que queda prohibida la elaboración de perfiles que tenga como efecto la discriminación. A mi juicio, esto es especialmente relevante, pues precisamente la discriminación es uno de los mayores riesgos de la creación de perfiles. Los perfiles generales creados con datos agregados pueden discriminar a aquellos que no actúan de acuerdo con el perfil general.

Por último, el apartado 4 del artículo propuesto establece la obligación del responsable del tratamiento de proporcionar información sobre la creación de perfiles y los efectos previstos, una novedad respecto a la Directiva actual. Esto podría ser la gran solución al hecho de que los individuos no sepan cuándo se está utilizando la minería de datos para tomar decisiones que les afecten, y sin duda supondría una enorme mejora en el control de los sujetos sobre sus datos y el uso que se hace de ellos. En todo caso, esta redacción presenta tres problemas principales que han sido identificados por Ira Rubinstein[119]. En primer lugar, la letra del artículo no dice expresamente que el responsable del tratamiento deba informar de oficio, sino que podría llegarse a interpretar como un deber de informar solo después de que un ciudadano pida dicha información. Pero aunque ello no fuera así, es muy complicado prever los efectos, precisamente por la propia naturaleza de las técnicas de *big data* y de minería de datos. E incluso aunque los efectos pudieran ser previstos, existiría un tercer proble-

---

119 Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law,* Vol. 3, n.º 2 (2013).

ma: la información que suministran los responsables es muy difícilmente comprensible por un ciudadano medio.

Hemos realizado un análisis de las normas actuales de protección de datos, así como de la Propuesta de Reglamento que vendrá a renovar esta normativa. De nuestro análisis se concluye que ninguna de estas normas termina por dar solución a los retos que nos brindan las tecnologías disruptivas actuales, en concreto, el *big data*.

CAPÍTULO VI. PROPUESTA DE SOLUCIONES

Es posible que privacidad y protección de datos no sean compatibles con el *big data*, y que haya llegado el momento de reconfigurar los riesgos que esto puede suponer para las personas y los beneficios que el tratamiento de datos puede traernos.

Ciertamente, en el marco legal europeo, la protección de datos y la privacidad se derivan de un derecho fundamental cuya protección no puede obviarse, tal y como ha quedado expuesto en este mismo trabajo.

No obstante, las fuertes trabas legales que Europa impone para la recolección y el tratamiento de datos de carácter personal están suponiendo una merma en nuestra fuerza competitiva frente a otras áreas del planeta, tales como Asia. El desarrollo y la innovación que las técnicas de *big data*, la minería de datos o el *cloud computing* implican, serán fuente de beneficios sociales y económicos que no debemos desaprovechar.

Después de todo lo analizado, la cuestión que surge no es si el *big data* incrementa el riesgo sobre la privacidad (lo cual indudablemente hace), sino si la naturaleza de este riesgo cambia. Y ello porque si el riesgo es simplemente mayor, las normas y leyes sobre privacidad todavía cumplen su función en la era del *big data*; simplemente tendremos que redoblar nuestros esfuerzos. Sin embargo, si la naturaleza del problema cambia, necesitaremos nuevas soluciones[120].

Y efectivamente, el problema se ha transformado por las características del *big data*. El valor de la información ya no reside en su uso primario, sino que ahora está en los usos secundarios. Y esto afecta al centro de la normativa de protección de datos y el papel que tiene el individuo en ella.

---

120 Kenneth Neil CUKIER y Viktor MAYER-SCHÖENBERGER. «Big data: A Revolution That Will Transform How We Live, Work And Think». *Houghton Mifflin Harcourt* (2013).

En primer lugar, ya han sido expuestas en este trabajo las limitaciones de que adolecen las técnicas de anonimización. El advenimiento de nuevas capacidades técnicas, junto con la gran cantidad de datos disponibles que nos deja el *big data* hace que la reidentificación de sujetos sea cada día más sencilla.

En segundo lugar, el consentimiento también padece grandes limitaciones. El modelo actual se basa en que la organización que recoge los datos de los individuos informa a éste sobre el tipo de información que recabará, y los fines para los que la utilizará. Así, el individuo presta su consentimiento, que se ha convertido en la piedra angular de los modelos de protección de datos. No obstante, ya han quedado expuestas en este trabajo las limitaciones de este modelo de consentimiento.

Algunas de las propuestas más relevantes sobre soluciones a estos problemas se describen a continuación.

6.1 REUNIONES MICROSOFT

Con la llegada y generalización de las tecnologías que hacen uso del *big data*, la información guarda un mayor valor en los usos secundarios que se puedan hacer de ésta. Y muchos de estos usos secundarios no se conocen ni se han llegado a imaginar en el momento en el que se recogen los datos. Así, las organizaciones no tienen la capacidad de especificar *a priori* los fines para los que utilizarán los datos (lo cual constituye un requisito indispensable para cumplir el principio de limitación de los fines y el requisito del consentimiento informado que exige la normativa de protección de datos). Pero si las empresas no pueden especificar cuáles serán los nuevos fines para los que se tratarán los datos, las empresas deberían volver atrás para pedir de nuevo el consentimiento a cada individuo para dichos usos secundarios. Esto entrañaría unos costes y unos recursos imposibles de asumir para las empresas, con lo que muchos de estos usos secundarios, y el consiguiente valor que supondrían, quedarían en saco roto.

Además, hoy en día, gracias a la computación en la nube (*cloud computing*), los datos se transfieren y cambian de manos de forma veloz y poco transparente para el individuo, incluso fuera de las fronteras europeas. Esto provoca una pérdida de control de los datos por

parte del individuo, a quien le resulta muy complejo tomar decisiones para prestar su consentimiento. De hecho, como ya hemos expuesto previamente, la mayoría de los usuarios no lee las políticas de privacidad, y los pocos que las leen no las comprenden. A mayor abundamiento, los datos que forman la base del *big data* se recogen y se procesan tan a menudo que nos invaden las peticiones de consentimiento.

La otra cara de la misma moneda es que los mecanismos actuales de privacidad y protección de datos ralentizan la utilización de los datos y el nuevo conocimiento, retrasando la innovación.

Conscientes de estas circunstancias, Microsoft patrocinó en 2012 una iniciativa a nivel mundial con el objetivo de reunir a los mayores expertos de privacidad de cada área (representantes de los reguladores, de la industria, defensores del interés público y expertos académicos) para incentivar una discusión sobre el papel del consentimiento en el mundo actual, así como para proponer alternativas.

De este modo, Microsoft organizó diferentes encuentros regionales en Washington, Bruselas, Singapur, Sydney y Sao Paulo, a la que siguió una última discusión-encuentro global. Los Profesores Fred H. Cate y Viktor Mayer Schönberger, moderadores de los diálogos en Washington y Bruselas respectivamente, plasman las conclusiones de las discusiones en las que el valor del consentimiento fue uno de los elementos clave[121].

La percepción general es que las normas de protección de datos tienen como objetivo asegurar que sea el individuo quien tenga el control sobre la información. Sin embargo, debido a los factores ya comentados, los sistemas basados en la notificación y el consentimiento no son sostenibles. En su lugar, los debates regionales propusieron cambiar el foco de la responsabilidad y control del individuo al «usuario» de los datos (esto es, las organizaciones que realizan el tratamiento de los datos), y hacia sistemas de rendición de cuentas por una verdadera custodia responsable de los datos, en contra de un mero cumplimiento normativo.

---

121 Fred H. CATE y Viktor MAYER-SCHÖENBERGER. «Notice and consent in a world of *Big data*». *International Data Privacy Law,* Vol 3, n.º 2 (2013).

En otras palabras, el foco de atención ya no debe ser el momento de la recolección de los datos (y con ello los sistemas para prestar un verdadero consentimiento informado), sino en momento de la utilización de los datos.

En un nuevo sistema, el requisito del consentimiento debería reservarse para usos relevantes, de forma que los individuos presten una atención mayor cuando el consentimiento les es requerido, y de este modo sea un mecanismo más efectivo. Así por ejemplo, reservando el consentimiento a situaciones en las que sea necesario que el individuo renuncie a derechos o tratamientos que de otro modo serían esperables, la forma del consentimiento tomaría un valor real. Esta misma premisa también es apoyada por el Foro Económico Mundial[122] y por autores como Solon Baroccas y Helen Nissenbaum[123].

No obstante sus limitaciones, los autores consideran que el consentimiento seguirá siendo un instrumento importante en el futuro de la protección de datos.

Del mismo modo, la custodia responsable de los datos exige fuertes medidas de seguridad para proteger los datos personales contra ataques, pérdidas o fugas de información. Si bien, debido a la constante evolución de la tecnología, los estándares de seguridad deben ser flexibles, y adaptados a un entorno dinámico, en contraposición con una aproximación estática que defina un set de datos o unas medidas de seguridad como suficientes en un momento determinado. Esto refuerza de nuevo la idea de que ya no es el usuario el centro de responsabilidad de los datos, sino las empresas que quieren hacer uso de ellos.

Del mismo modo, son necesarios mecanismos de legalidad internacional. La falta de consenso a nivel global sobre los principios que deben regir un nuevo modelo impiden una labor de desarrollo normativo armonizada que causa perjuicios en un momento en el que el comercio electrónico, el *cloud computing* o el *big data* permiten una transferencia internacional de datos sin precedentes.

---

122 FORO ECONÓMICO MUNDIAL y THE BOSTON CONSULTING GROUP. «Rethinking Personal Data: Strengthening Trust» (2012). Proyecto «Rethinking Personal Data».

123 Solon BAROCCAS y Helen NISSEBAUM. «Privacy, big data and the public good. Chapter 2: Big data's End Run Around Anonimity And Consent». *Cambrigde University Press* (2014).

Pero si el consentimiento ya no debe ser la herramienta principal de las normas de protección de datos, las organizaciones deben ser más transparentes y proporcionar una información más clara a los usuarios. Los datos deben tener un tratamiento diferente en función del uso que se les vaya a dar, o del tipo de dato de que se trate. No obstante, una vez más, todavía falta consenso sobre qué usos de los datos deben ser permitidos y cuáles no, o bajo qué circunstancias. Las particularidades culturales de cada región, o el mero hecho de considerar la privacidad como un derecho fundamental en algunas jurisdicciones pero no en otras, posan un reto para alcanzar medidas globales conjuntas.

6.2 PRIVACIDAD POR DEFECTO Y PRIVACIDAD DESDE EL DISEÑO

La expansión de avances como el *big data* conlleva riesgos para la privacidad de los individuos. Así, si los sistemas de análisis de datos son utilizados para un fin que trasciende la legalidad, la privacidad puede resultar quebrada. Es por ello que las organizaciones que quieren tomar la delantera de los avances analíticos también deberían dar un paso atrás y reconsiderar el diseño que realizan de sus invenciones tecnológicas, para tomar en cuenta la seguridad de los datos y la privacidad ya desde el diseño de la arquitectura de la tecnología, el diseño de sistemas y los procedimientos operativos. Los principios de privacidad deben ser insertados en el código de funcionamiento del dispositivo.

Los riesgos que los avances suponen para la privacidad han hecho cambiar el foco. Ya no debe servir únicamente con cumplir con los principios de protección de datos, ahora también debemos tener en cuenta la privacidad desde el mismo momento en que la tecnología se está diseñando. Es decir, la privacidad debe dejar de ser un mero concepto legal para ser una prioridad de negocio. Tomando en cuenta la privacidad ya desde el primer momento, los diseñadores pueden desarrollar herramientas que aseguren un mayor grado de protección.

Los sistemas de privacidad desde el diseño implican que las tecnologías son construidas teniendo en cuenta la necesidad de la protección de la privacidad. Por su parte, los sistemas de privacidad por defecto conllevan que la tecnología está configurada para que las opciones que por defecto vienen establecidas sean las más protectoras

de la privacidad; y el individuo puede posteriormente cambiar la configuración para permitir otras utilidades que requieran un nivel de privacidad menor.

La mejor aproximación será incluir la configuración más segura por defecto en un sistema diseñado bajo los principios de privacidad desde el diseño. De este modo, los avances persiguen mantener todo el potencial de crear valor de las tecnologías, poniendo la protección del individuo como un factor más a tener en cuenta, junto con el resto de parámetros (como capacidad operativa, viabilidad económica, etc.). Así se crea una dinámica *win-win* (es decir, en la que ganan ambas partes) para las organizaciones y los individuos.

Las organizaciones también pueden encontrar un incentivo para convertir sus esfuerzos de proteger la privacidad en una ventaja competitiva respecto a sus competidores en el mercado.

Un ejemplo de tecnología que abraza este concepto es el sistema creado por Jeff Jonas, director científico de IBM Entity Analitics, para conseguir un sistema capaz de poner toda la información disponible en una base de datos en un contexto que dé un sentido al conjunto (el término concreto utilizado por Jeff Jonas para su tecnología es «*sensemaking system*»)[124]. Pero ¿en qué consiste esta tecnología?

Las tecnologías *sensemaking* son un tipo de tecnologías emergentes diseñadas con el objetivo de permitir que las empresas obtengan una mejor comprensión de su entorno. Para ello es necesario tomar en consideración datos que las empresas ya poseen (por ejemplo, datos de los sistemas financieros de la empresa, de recursos humanos, del sistema operativo, etc.), y otros que no controlan (por ejemplo, datos generados fuera de la empresa y no estructurados como aquellos generados en redes sociales). Estos sistemas permiten analizar cientos de millones de observaciones de fuentes muy diversas, y serán utilizados por las empresas para poder tomar mejores decisiones de forma más rápida[125].

La lógica de estas tecnologías es que una organización necesita sumar todas sus observaciones, datos y experiencia para obtener un conocimiento avanzado. En la actualidad, muchas empresas no se be-

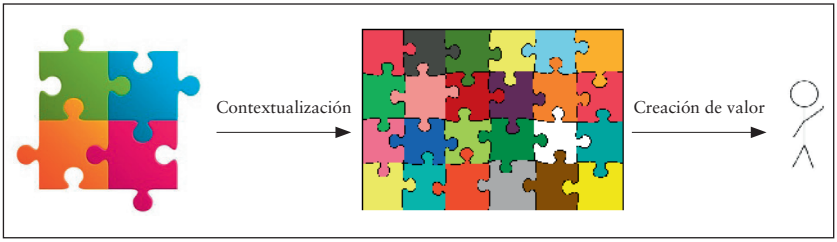
---

124 Jeff JONAS y Ann CAVOUKIAN. «Privacy by Design in the Age of Big data». *Privacy by Design (PbD)* (2012).

125 Ídem.

nefician del conocimiento que han generado en el pasado, o no llegan a generar conocimiento con los datos que recopilan. Por ejemplo, un estudio de una gran cadena minorista ha mostrado que de cada 1.000 empleados que contrataba la empresa, dos de ellos habían sido previamente despedidos por robar en la misma tienda en la que volvían a ser contratados. Las empresas generan tanta cantidad de nuevos datos que no son capaces de acudir a los datos ya creados y crear un conjunto con sentido.

De este modo, las tecnologías «*sensemaking*» pretenden integrar los nuevos datos con los ya disponibles, y unir las distintas piezas del puzzle para crear un contexto con sentido que permita a las empresas tomar mejores decisiones. De este modo se crea un contexto acumulativo que permite comprender lo que sucede en cada momento.



El proceso de creación del sistema de *sensemaking* desarrollado por Jeff Jonas incorporó los principios de privacidad desde el diseño.

Así por ejemplo, entre otras medidas de privacidad, Jeff Jonas introdujo métodos que favorecían los falsos negativos. En determinadas circunstancias, es menos dañino perder algunas cosas (es decir, obtener un falso negativo), que sacar conclusiones que no son verdaderas (es decir, obtener un falso positivo). Esta tecnología favorece los falsos negativos de modo que las decisiones erróneas que puedan tomarse sobre un individuo que tengan trascendencia en su vida se reducen.

6.3 NUEVO MODELO DE NEGOCIO

Las normas de privacidad actuales están construidas sobre la premisa de que compartir datos puede ser perjudicial para el individuo. Por ello se hacía necesario gestionar mecanismos como la notificación y el consentimiento para la recolección y tratamiento de los da-

tos. Sin embargo, existen muchos nuevos servicios actuales (tales como Twitter, Facebook, Foursqueare o Yelp) que ponen de manifiesto que los individuos sí quieren compartir datos personales[126].

Además, para crear valor, los datos necesitan moverse, y para moverse, se necesita confianza entre todos los agentes que participan en la cadena de valor[127] (individuos, empresas, organizaciones, gobiernos, reguladores, etc.). Y más aún, al contrario que con la mayoría de los bienes físicos, el valor de los datos crece con su uso: conectar dos datos crea un nuevo dato que a su vez puede servir para nuevas aplicaciones y para seguir creando valor.

La falta de confianza en los modelos actuales de consentimiento y anonimización han propiciado una corriente basada en los derechos de acceso y cancelación de los datos de los individuos de las bases de datos, así como de los sistemas *opt-in*. Por ejemplo, la nueva Propuesta de Reglamento de Protección de Datos introduce el denominado «derecho al olvido» que faculta a los ciudadanos europeos a pedir el borrado de sus datos bajo determinados requisitos, e introduce la obligación de que el consentimiento sea explícito.

Si bien estos derechos no deben eliminarse, deben regirse con flexibilidad. Algunos han visto en este sistema el riesgo de que se produzca lo que se ha dado en denominar «la tragedia de los datos comunes». El condominio de los datos, o estos datos comunes, son aquellos datos que han sido agregados, anonimizados y hechos públicos para favorecer el conocimiento y la innovación. La tragedia viene del riesgo de que los individuos vean un incentivo para borrar sus datos si aprecian un riesgo de que los sistemas de anonimización fallen y puedan ser reidentificados, sin importar cuán pequeño sea este riesgo.

Una vez más, (aparte de la evidente solución técnica), una solución complementaria puede verse en que los agentes compartan derechos y obligaciones. Efectivamente, el hecho de que un investigador, una em-

---

126 FORO ECONÓMICO MUNDIAL. «Unlocking The Economic Value Of Personal Data. Balancing Growth And Protection» (2012). Proyecto «Rethinking Personal Data».

127 FORO ECONÓMICO MUNDIAL y THE BOSTON CONSULTING GROUP. «Rethinking Personal Data: Strengthening Trust» (2012). Proyecto «Rethinking Personal Data».

presa, o una agencia gubernamental tengan la capacidad técnica de reidentificar a los sujetos, no quiere decir que deba hacerlo, ni que tenga derecho a ello. Un ejemplo visual de esto es el caso de que por el mero hecho de que los compuestos de los fertilizantes puedan ser utilizados para fabricar una bomba, no debemos parar de utilizar fertilizantes; en su lugar, habremos de prohibir la fabricación y uso de las bombas[128].

Para generar la confianza que el entorno necesita, las normas y las leyes no son suficientes por sí solas. Una de las soluciones que se ha propuesto es un modelo de co-regulación en la que las responsabilidades y los derechos sean compartidos entre todos los agentes, y que se base en la creación de los llamados «puertos seguros» en el uso de los datos. Y por supuesto, todo sistema debe ir acompañado de las mayores medidas de ciberseguridad disponibles en cada momento.

(i) EL PROBLEMA DE LA PROPIEDAD DE LOS DATOS

Uno de los principales dilemas de base que hay que superar es el de la propiedad de los datos. Muchos individuos opinan que en la medida en que la información se refiere a ellos, los datos les pertenecen. Por su parte, las empresas sostienen que los datos que se encargan de recoger, tratar y crear son de su propiedad, pues es el beneficio que les pueden reportar lo que les anima a invertir enormes cantidades de dinero, tiempo y otros recursos.

En su trabajo, el Foro Económico Mundial propone que, a pesar de que los datos se refieran a un individuo, se crean a partir de la interacción de diversas partes, de forma que todos ellos deben tener derechos y responsabilidades sobre esta información. Los derechos deben ser comunes, no exclusivos. Para explicar este concepto, expone un ejemplo ilustrativo por analogía. Los artistas musicales son generalmente los propietarios de su música, y tienen los derechos de autor sobre ésta. Sin embargo, también otros agentes como publicistas, agentes discográficos o distribuidores ostentan derechos sobre el uso de la música. De esta forma, una multitud de partes comparten el valor que se crea en las diferentes fases de la cadena de valor a lo largo del tiempo[129].

---

128 Ídem.
129 Ídem.

Pero eso no es todo. El concepto mismo de propiedad de los datos queda en entredicho debido a la capacidad actual de realizar copias prácticamente ilimitadas de los datos que las empresas recaban.

(ii) EL EMPODERAMIENTO DE LOS INDIVIDUOS

Así las cosas, las últimas corrientes actuales caminan hacia los modelos de empoderamiento de los individuos.

Los modelos futuros de creación de permisos deben venir de un diálogo entre agentes técnicos, regulatorios, empresariales y civiles. Uno de los modelos propuestos, y que está siendo probado en países como Reino Unido, es la creación de permisos para el tratamiento de los datos diferentes en función del contexto en el que se quieran utilizar (así por ejemplo, en el ámbito sanitario, el financiero, etc.).

Un nuevo modelo de negocio está siendo planteado basado en este empoderamiento de los individuos. Los individuos se convierten en un agente económico más, gestionan su propia información, para sus propios fines, y comparten una parte de esta información con las empresas para comunicar qué quieren, cómo y cuándo, y para obtener beneficios conjuntos[130].

Este modelo se basa en lo que se ha bautizado de forma genérica como «servicios de datos personales» («*personal data services*» o «PDS» por sus siglas en inglés). El Profesor Searls, director del Centro Berkman de Harvard describe estas herramientas bajo el nombre de «gestión de relaciones con proveedores» («*vendor relations management*» o simplemente «VRM»). Por su parte, el Foro Económico mundial utiliza la terminología de «archivos de datos personales» («*personal data lockers*»).

Independientemente del nombre, esto implica que las empresas ya no son las que monitorizan los comportamientos de los individuos de forma poco transparente, sino que el propio individuo utiliza estos

---

130 Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law,* Vol. 3, n.º 2 (2013), en referencia al proyecto de Ctrl-Shift. «The New Personal Data Landscape Report» (2011); y Doc Searls. «The Intention Economy: When Customers Take Charge». *Harvard Business Review* (2012).

sistemas de gestión VRM para comunicarse con el resto de agentes y decidir sobre sus datos.

Ira S. Rubinstein, Profesora de Derecho de la Información de la Universidad de Nueva York, sistematiza las características que tiene este nuevo modelo de negocio, sobre la base de las propuestas del Foro Económico Mundial, y de los trabajos de Searls (en el proyecto que dirige, «Project VRM»), de Ctrl-Shift y otros[131].

1. El individuo es el centro del sistema de recolección, gestión y uso de los datos.
2. El individuo decide qué información revelar, de forma selectiva.
3. Control sobre los fines para los que se usan los datos, tanto primarios como secundarios, así como sobre su duración, a través de contratos.
4. El individuo tiene mecanismos para comunicar lo que demanda de forma abierta y flexible, sin estar ligado a ninguna organización concreta.
5. Toma gran relevancia la gestión de los medios para permitir la autenticación de la identidad de los sujetos que accedan al sistema.
6. Medidas de seguridad al más alto nivel.
7. Portabilidad de datos, de modo que los individuos puedan obtener todos sus datos y moverlos de un proveedor de servicios VRM a otro.
8. Medidas para hacer a las empresas proveedoras de estos servicios responsables de la seguridad de los datos y protegerlos de acuerdo con los distintos niveles de permisos que el individuo ha decidido otorgar.

Para ser técnicamente viables, sería necesario que estos sistemas cumplieran dos requisitos. Por un lado, altas medidas de seguridad; y por otro lado, la capacidad de que los datos puedan estar asociados a

---

131 Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law,* Vol. 3, n.º 2 (2013).

«etiquetas» con metadatos que describan el nivel de privacidad de cada uno de ellos.

En tanto estos sistemas contendrían la información más privada de los individuos (desde partidas de nacimiento, pasaporte, contenidos de redes sociales e incluso contraseñas), serían objeto de todo tipo de ciberataques. Es por ello que las medidas de seguridad que deben ser desplegadas deben ser del máximo nivel.

Estas medidas de seguridad deberían incluir[132] (i) datos personales encriptados, tanto en el lugar de almacenamiento como durante las transferencias de datos; (ii) las claves para des-encriptar la información deben estar guardadas en silos externos; (iii) los metadatos también deben estar encriptados; (iv) las medidas de autenticación de los individuos que accedan a los datos deben ser del más alto nivel.

Tal y como señala Ira S. Rubinstein, este sistema está alineado con los principios de protección de datos europeos, al tiempo que ofrece una solución a los problemas que el *big data* impone sobre la gestión de los datos y la privacidad de los individuos.

El Foro Económico Mundial propone dos fases para desarrollar el nuevo modelo de negocio basado en esta creación de permisos diferenciados en función del contexto[133].

El primer paso para ayudar a los individuos a poder ejercitar un mayor control sobre sus datos es el derecho a obtener una copia de todos los datos que una organización tiene sobre el individuo. De esta forma, los individuos pueden conocer qué datos recogen las empresas sobre ellos. El proyecto Ctrl-Shift llevado a cabo en Inglaterra es un ejemplo de esta propuesta. El proyecto tiene como objetivo proporcionar a quien lo solicite una copia de sus datos obtenidos desde múltiples fuentes (desde bancos, compañías telefónicas, energéticas o el propio Gobierno). Así, los individuos pueden unir toda esta información en un único sistema, bajo su control, y utilizarla de modos que ofrezcan nuevos beneficios o permitan comparar su actividad con otros.

---

132 Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law,* Vol. 3, n.º 2 (2013).

133 FORO ECONÓMICO MUNDIAL y THE BOSTON CONSULTING GROUP. «Rethinking Personal Data: Strengthening Trust» (2012). Proyecto «Rethinking Personal Data».

Estas copias ayudarán a identificar los derechos de cada uno de los agentes, como primer paso para crear un verdadero sistema de derechos y responsabilidades conjuntos.

El segundo paso en este sistema que se propone es la gestión de esta información. Numerosos son los términos que se han acuñado para este método, desde «archivos de datos personales» («*personal data lockers*»), hasta «gestión de relaciones con proveedores» («*vendor relations management*» o simplemente «VRM»). Se trata de que cada individuo pueda almacenar y agregar datos provenientes de fuentes diversas, y conceder permisos a las organizaciones para poder acceder a dichos datos de una forma controlada. El objetivo último de estos servicios es conceder al individuo un mayor control sobre cómo se utilizan sus datos.

Esto implica un cambio de paradigma en la gestión de los datos con respecto a los modelos actuales. Ya no serían las organizaciones las que recolectan, almacenan y analizan los datos para dirigirse al consumidor de forma dirigida. Serían los individuos los que gestionarían sus datos, y a cambio podrían proporcionar sus datos de manera más directa y obtener parte de los beneficios, como por ejemplo una mejor tarifa en su compañía telefónica en función de su historial de llamadas. Es decir, el individuo deja de tener un papel pasivo, a obtener un papel proactivo en el control de sus datos, resultando empoderado.

Diversas *start ups* están poniendo en marcha este tipo de servicios, que constituyen un nuevo modelo de negocio (por ejemplo, Connect. Me o Personal. Com). Sin embargo, es en los sectores de salud y financiero donde este modelo puede tener más beneficios.

En todo caso, tal y como el Foro Económico Mundial ha resaltado, este modelo se encuentra todavía en su más tierna infancia.

Ahora bien, la pregunta que surge después de analizar esta propuesta de nuevo modelo de negocio es ¿qué incentivos tienen las empresas para apoyar este nuevo sistema que otorga el control al individuo?

En primer lugar, los servicios de datos personales cubren una necesidad. Ayudan a los individuos a gestionar de forma más fácil su vida, y les otorga las herramientas para apreciar el valor de sus datos (en tanto son fuente de nuevos datos). Además, también crea una nueva oportunidad de negocio para nuevos entrantes que no pueden benefi-

ciarse del *big data* alojado en los silos de las grandes empresas como Google o IBM, Axicom o Facebook.

En segundo lugar, la calidad de los datos es mayor, pues los individuos revelan la información de forma voluntaria, y así se mantiene actualizada, detallada, veraz y precisa. Esto supone una mejora respecto a la aproximación predictiva actual.

En tercer lugar, las organizaciones que confíen en las nuevas empresas prestatarias de servicios de protección de datos incurrirían en un menor coste de cumplimiento regulatorio, contra los costes en los que incurren al almacenar los datos en sus propios silos.

En conclusión, los modelos de negocio de empoderamiento de los individuos, y en particular los servicios de gestión de datos personales, son técnicamente viables, y potencialmente atractivos para el mercado. Sin embargo, se trata todavía de proyectos en fase de investigación, cuyos límites deben ser estudiados.

Para ello será necesario financiar proyectos de investigación en este campo, así como nueva investigación sobre si el uso de metadatos puede requerir un nuevo desarrollo legislativo[134].

6.4 CONSIDERACIONES FINALES

Las propuestas analizadas muestran aproximaciones al problema de la privacidad y la protección de datos desde puntos de vista diferentes. Los nuevos modelos de negocio basados en el empoderamiento del individuo pivotan sobre el hecho de otorgar un mayor control a las personas sobre sus datos. Por su parte, las propuestas basadas en poner el nuevo foco de atención sobre el uso de los datos y no sobre la recolección, pretenden otorgar un papel más relevante (y mayores responsabilidades) a las empresas y organizaciones que vayan a hacer uso de los datos.

Tal vez en el corto plazo sean estas últimas propuestas las que puedan adquirir más relevancia, a través de códigos de conducta a los que las organizaciones se adhieran. Sin embargo, las propuestas de los nuevos modelos de negocio no deben ser perdidas de vista, pues pueden ser soluciones innovadoras que vean la luz en el futuro de forma generalizada.

---

134 Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law,* Vol. 3, n.º 2 (2013).

BIBLIOGRAFÍA

LEGISLACIÓN Y OPINIONES DE ORGANISMOS DE PROTECCIÓN DE DATOS

1. Tratado de Funcionamiento de la Unión Europea.
2. Carta de los Derechos Fundamentales de la Unión Europea.
3. Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos.
4. Directiva 2002/58/CE del Parlamento Europeo y del Consejo, de 12 de julio de 2002, relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas.
5. Propuesta de Reglamento Europeo sobre Protección de Datos.
6. Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.
7. Real Decreto 1720/2007, de 21 de diciembre, por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.
8. Ley 34/2002, de 11 de julio, de Servicios de la Sociedad de la Información y de Comercio Electrónico.
9. Ley 9/2014, de 9 de mayo, General de Telecomunicaciones.
10. Sentencia del Tribunal Constitucional 292/2000, de 30 de noviembre de 2000.
11. Informes del Grupo de Trabajo del Artículo 29:
    a) Opinion 8/2014 on the Recent Developments on the Internet of Things.
    b) Statement of the WP29 on the impact of the development of big data on the protection of individuals with regard to the processing of their personal data in the EU adopted on 16 September 2014.
    c) Opinion 05/2014 on Anonymisation techniques.
    d) Opinion 03/2013 on Purpose Limitation.
    e) Opinion 15/2011 on the Definition of Consent.
    f) Opinion 13/2011 on Geolocation Services on Smart Mobile Devices.
    g) Opinion 4/2007 on the Concept of Personal Data.
12. Dictamen del Supervisor Europeo de Protección de datos, de 14 de enero de 2011, sobre la Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social Europeo y al Comité de las Regiones — «Un enfoque global de la protección de los datos personales en la Unión Europea».

13. Agencia Europea de los Derechos Fundamentales. «Understanding and preventing discriminatory ethnic profiling» (2012).
14. Sitio web de la Agencia Española de Protección de Datos.
15. Information Commissioner´s Office. «Big data and Data Protection» (2014).
16. Information Commissioner´s Office. «Anonymisation: managing data protection risk code of practice» (2012).
17. Federal Trade Commission (FTC). «Protecting Consumer Privacy in an Era of Rapid Change. Recommendations for Businesses and Policymakers» (2012).
18. Personal Health Information Protection Act, Ontario, Canadá (2004).
19. Health Insurance Portability and Accountability Act Privacy Rule (HIPAA), Estados Unidos (1996).

LIBROS, MANUALES Y ARTÍCULOS

20. Joris M. MOOIJ, Jonas PETERS, Dominik JANZING, Jakob ZSCHEISCHLER, Bernhard SCHÖLKOPF. «Distinguishing cause from effect using observational data: methods and benchmarks», versión 2. *Journal of Machine Learning Research, Cornell University* (2015).
21. Khaled EL EMAM y Cecilia ÁLVAREZ. «A critical appraisal of the Article 29 Working Party Opinion 05/2014 on data anonymization techniques». *International Data Privacy Law Journal*, Oxford University Press, Vol. 5, n.º 1 (2015).
22. Bert-Jaap KOOPS. «The trouble with European data protection law». *International Data Privacy Law Journal*, Oxford University Press, Vol. 14, n.º 14 (2014).
23. EMC «Digital Universe Study». *The Digital Universe and big data*. (2014).
24. Solon BAROCCAS y Helen NISSEBAUM, «Privacy, big data and the public good; Chapter 2: Big data's End Run Around Anonimity and Consent», *Cambrigde University Press* (2014).
25. Javier PUYOL. «Una aproximación a *big data*». *Revista de Derecho de la Universidad Nacional de Educación a Distancia* (UNED), n.º 114 (2014).
26. Kenneth Neil CUKIER y Viktor MAYER-SCHÖENBERGER. «The Rise of Big data. How It's Changing the Way We Think About the World». *Foreign Affairs* Vol. 92, n.º 13 (2013).
27. Kenneth Neil CUKIER y Viktor MAYER-SCHÖENBERGER. «Big data: A Revolution That Will Transform How We Live, Work And Think». *Houghton Mifflin Harcourt* (2013).
28. Fred H. CATE y Viktor MAYER-SCHÖENBERGER. «Notice and consent in a world of Big data». *International Data Privacy Law,* Vol 3, n.º 12 (2013).

29. Ira S. RUBINSTEIN. «Big data: The End Of Privacy Or A New Beginning?». *International Privacy Law*, Vol. 3, n.º 12 (2013).
30. Cas PURDY. «Finding benefits in big data». *Trustwave Blog* (2013).
31. Omer TENE y Jules POLONESTSKY. «Big data for all: Privacy and user control in the age on analytics». *Northwestern Journal of Technology and Intellectual Property*, Vol. 11, n.º 15 (2013).
32. Sor ARTEAGA JUÁREZ. «Las redes sociales digitales en la gestión y las políticas públicas. Capítulo 3: Implicaciones Legales en el uso de las redes sociales por las administraciones públicas». *Escola d'Administració Pública de Catalunya* (2013).
33. THE BOSTON CONSULTING GROUP. «To get started with big data. BCG perspectives» (2013).
34. Doc SEARLS. «The Intention Economy: When Customers Take Charge». *Harvard Business Review* (2012).
35. FORO ECONÓMICO MUNDIAL y THE BOSTON CONSULTING GROUP. «Rethinking Personal Data: Strengthening Trust» (2012). Proyecto «Rethinking Personal Data».
36. FORO ECONÓMICO MUNDIAL. «Unlocking The Economic Value Of Personal Data. Balancing Growth And Protection» (2012). Proyecto «Rethinking Personal Data».
37. Omer TENE y Jules POLONETSKY. «Privacy in the age of big data: a time for big decisions». *Standford Law Review Online*, Vol. 64, n.º 163 (2012).
38. Emöke-Ágnes HORVÁT, Michael HANSELMANN, Fred A. HAMPRECHT, y Katharina A. ZWEIG. «One plus one makes three (for Social Networks)». *PLOS One Journal* (2012).
39. Jeff JONAS y Ann CAVOUKIAN. «Privacy by Design in the Age of Big data». *Privacy by Design (PbD)* (2012).
40. IBM INSTITUTE FOR BUSINESS VALUE, en colaboración con la Escuela de Negocios Säid de la Universidad de Oxford. «Analytics: el uso del *big data* en el mundo real». *IBM Global Business Services* (2012).
41. CTRL-SHIFT. «The New Personal Data Landscape Report» (2011).
42. danah boyd, «Networked privacy». *Personal Democracy Forum*, Nueva York (2011).
43. Alan MISOLVE *et al.* «You are who you know: inferring users profiles in online social networks». *Web Search and Data Mining (WSDM)* Conference. ACM, Nueva York (2010).
44. Paul OHM. «Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization». *UCLA Law Review*, Vol. 57 (2010).
45. Arvind NARAYANAN y Vitaly SHMATIKOV. «De-anonymizing social networks». *The University of Texas at Austin, Simposio IEEE on Security and Privacy* (2009).

46. Carter JERNIGAN y Behran F. T. MISTREE. «Gaydar: facebook friendships expose sexual orientation». *First Monday,* Vol. 14, n.º 110 (2009); citado por *ABC News Tech*. Solo ha sido posible acceder a referencias del estudio.

47. Aleecia M. MCDONALD y Lorrie FAITH CRANOR. «The Cost of Reading Privacy Policies». *Journal of Law and Policy of the Information Society,* Vol. 4, n.º 13 (2008).

48. Arvind NARAYANAN y Vitaly SHMATIKOV. «Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)». *The University of Texas at Austin* (2008).

49. Fred H. CATE. «The failure of Fair Information Practice Principles». *Consumer Protection In The Age Of Information Economy* (2006).

50. DOMINGO-SALVANY, BRUGAL PUIG y BARRIO ANTA. «Tratado SET de Trastornos Adictivos». *Sociedad Española de Toxicomanías*. Editorial Médica Panamericana (2006).

51. Miguel DELGADO RODRÍGUEZ y Javier LLORCA DÍAZ. «Estudios Longitudinales: concepto y características». *Revista Española de Salud Pública,* Vol. 78, n.º 12 (2004).

52. Ana Victoria SÁNCHEZ URRUTIA, Héctor Claudio SILVEIRA GORSKI, Mónica NAVARRO MICHEL, Stefano RODOTÁ. «Tecnología, intimidad y sociedad democrática». *Icaria* (2003).

53. Lantaya SWEENY. «K-Anonimity, a model for protecting privacy». *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, n.º 15 (2002).

54. Tomás ALUJA. «La minería de datos, entre la estadística y la Inteligencia Artificial». *Qüestiió, Universidad Politécnica de Catalunya,* Vol. 25, n.º 13 (2001).

55. Cynthia DWORK. «Differential Privacy». *Microsoft Research*.

56. Steve LOHR. «If Algorithms Know All, How Much Should Humans Help?» *The New York Times News Services* (6 de abril de 2015).

57. Stephen F. DE ANGELIS. «Big data Analytics: Determining Causation rather than Correlation». *Enterra Insights Blog* (4 de febrero de 2015).

58. Lutz FINGER. «Recommendation Engines: The Reason Why We Love Big data». *Forbes Tech* (2 de septiembre de 2014).

59. Ignacio BRUNA. «El futuro Reglamento General de Protección de Datos de la Unión Europea (o eso esperamos)». *Blogs KPMG, Ciberseguridad* (17 de noviembre de 2013).

60. Juan Fernando LÓPEZ AGUILAR. «Por fin una ley europea de protección de datos (I)». *El Huffington Post Unión Europea* (24 de octubre de 2013).

61. Ricardo GALLI. «Sé cuidadoso con el *Big Data*». *Blog De Software, libre, internet, legales* (29 mayo 2013).

62. Harry WALLOP. «A like on Facebook tells the world all it needs to know about you». The Telegraph Technology (13 marzo 2013).
63. «La confusión de las cigüeñas». *Naukas: ciencia, escepticismo y humor* (3 diciembre 2012).
64. Quentin HARDY. «Rethinking privacy in an era of big data». *New York Times* (4 junio 2012).
65. Richard ANDERSON. «Cómo los matemáticos dominan los mercados». *BBC Economía* (1 octubre 2011).
66. CENTRO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS (CSIC). «Usos y abusos de la Estadística». (11 marzo 2010).
67. Ambrosio LICEAGA. «Causas y correlaciones». *Ciencia de Bolsillo.com* (30 junio 2008).

RECURSOS AUDIOVISUALES Y SITIOS WEB

68. Javier PUYOL. «*Big data* y Administraciones Públicas» [Conferencia]. Seminario Internacional *Big data* para la Información Oficial y la Toma de Decisiones (2014).
69. «User's control over their data: is prior consent the best way to monitor?» [Conferencia] Computers, Privacy & Data Protection on the Move (CPDP) (2015).
70. «To be or not to be (anonymous)? Anonymyity in the age of big and open data». [Conferencia] Computers, Privacy & Data Protection on the Move (CPDP) (2015).
71. «EU data protection reform: have we found the right balance between fundamental rights and economic interests?» [Conferencia] Computers, Privacy & Data Protection on the Move (CPDP) (2015).
72. «Ojo con tus datos» [Documental]. Documentos TV, Radio Televisión Española (2013).
73. Spurious Correlations [Sitio web]. Disponible en: http://www.tylervigen.com/spurious-correlations.
74. Panopticlick. Proyecto *How unique —and trackable— is your browser?* Disponible en: https://panopticlick.eff.org/.

La obra de Elena Gil González, que el lector tiene en sus manos, analiza el impacto en el derecho a la privacidad del nuevo fenómeno conocido como *Big data*, que implica el tratamiento y almacenamiento masivo de información.

Uno de los retos actuales de la sociedad de la información es que el sector empresarial alcance un elevado cumplimiento de las obligaciones que la normativa de protección de datos le impone, fomentando una cultura de preservación de datos de carácter personal, que suponga una clara mejora de la competitividad compatible con el desarrollo económico. En este sentido, el *Big data* es un reflejo del impacto de la tecnología en la esfera de la vida privada. El despliegue de tecnologías como el *Big data*, el internet de las cosas, el uso de wearables o las smartcities, entre otras, requiere de un análisis y valoración técnica y jurídica para promover buenas prácticas que garanticen su adecuación a la normativa en la materia y, en consecuencia, el respeto por los derechos de los ciudadanos.

La obra contribuye a difundir algunos aspectos esenciales de esta tecnología, aportando propuestas para minimizar los riesgos de intrusión en la privacidad. El deber de transparencia y consentimiento, las técnicas de anonimización, la privacidad por defecto y desde el diseño que se abordan en el trabajo son elementos esenciales para un desarrollo respetuoso del *Big data*.

ISBN 978-84-340-2309-3