



Los datos masivos (Big Data)



Imagen: Osvaldo Romero

El objetivo de esta nota es explicar qué son los datos masivos o *Big Data*, para qué se usan, quiénes son los principales actores en su manejo y desarrollo y cuál es la legislación actual en México y en otros países del mundo. El concepto ha atraído la atención de instituciones gubernamentales, académicas, la iniciativa privada y el público en general. En español se le conoce como **macrodatos, datos masivos o inteligencia de datos**, pero es común usar el anglicismo *Big Data*.

Introducción. ¿Por qué hablar de *Big Data*?

Aunque no existe una definición formal del término, su uso se refiere a grandes cantidades de datos o información digital que requiere equipos de cómputo de alto rendimiento y programas o técnicas de análisis especializadas para su procesamiento e interpretación.

Su gran atractivo es que permite obtener relaciones, patrones y resultados, que no son accesibles mediante otras metodologías. Por esta razón se ha sugerido, a manera de definición, que *Big Data* se refiere a todo aquello que se puede obtener de colecciones de datos en gran escala, pero que no podría conseguirse en escalas menores.¹

Dada la popularidad del término, es frecuente encontrarlo usado de manera incorrecta o imprecisa. Como regla general, si la información puede procesarse en unas hojas de cálculo (por ejemplo, en un formato "Excel"), no es *Big Data*.

RESUMEN

- Los datos masivos, también llamados **macrodatos** o *Big Data*, son grandes cantidades de información que por su volumen, variedad y velocidad de obtención requieren tecnologías y métodos especializados para su aprovechamiento.
- Al igual que cualquier otro conjunto de datos, pueden tener sesgos y errores.
- El conjunto de técnicas para procesar la información contenida en las grandes bases de datos se denomina **Big Data Analytics**, como por ejemplo la minería de datos, el aprendizaje computacional (*machine learning*) y el análisis de redes.
- Sus beneficios pueden llegar a todos los sectores de la sociedad; las empresas incrementarían sus márgenes de operación hasta en 60% y las dependencias del sector público generarían grandes ahorros.
- Es una actividad de alto valor agregado y la tecnología para implementarla ya es accesible, pero su uso responsable representa un reto legislativo que debe ser atendido.
- En México existen leyes federales que establecen el marco legal de uso de los datos personales.
- En el plano internacional, los marcos jurídicos de los Estados Unidos de América (EUA) y la Unión Europea son contrastantes entre sí.

Se ha estimado que el uso de los macrodatos en el sector salud de los EUA podría dejar beneficios por 300 mil millones de dólares anuales y ayudaría a reducir el gasto correspondiente en 8%. Igualmente, una empresa en la actualidad podría mejorar su margen de operaciones en 60%.² Sin embargo, su manejo inadecuado puede ocasionar pérdidas, problemas de seguridad e incluso, infracciones a la ley.

México se encuentra en un periodo de transformación digital,^{3,4} donde la eficiencia de los procesos de empresas, instituciones gubernamentales y académicas se han incrementado con las nuevas tecnologías digitales. En Latinoamérica se ha pronosticado un gran crecimiento en este terreno y México aparece como uno de los punteros de la región.⁵⁻⁷ En este contexto, el buen uso de los datos masivos puede representar una magnífica oportunidad de crecimiento.

Antecedentes. ¿De dónde surge?

El término *Big Data* saltó a la fama en 2008 con un artículo que publicó la revista electrónica *Wired*, donde se proponía que cantidades masivas de información vuelven obsoleto el método científico o el uso de modelos o teorías para explicar un fenómeno.⁸ Es decir, que bastaba una gran cantidad de datos para poder hacer predicciones sin tener que contar con un mayor conocimiento del tema. Sin embargo, existen ejemplos que muestran los grandes riesgos de usar estas predicciones sin tener un modelo o una explicación de fondo (Recuadro 1).

La necesidad de manejar eficientemente los macrodatos surge del crecimiento acelerado de la información que se genera y procesa en la actualidad, la cual se duplica casi cada dos años, es muy variada⁹ y se almacena en diversos formatos (audio, fotos, videos, texto, coordenadas). Por ejemplo las de origen médico que se usan para encontrar factores de riesgo de enfermedades, o la producida por los mismos dispositivos conectados a internet, que se comunican entre sí y generan aún más información. (Nota INCYTU No.13 *Biobancos y Registros Médicos Electrónicos*)

Recuadro 1. *Google Flu Trends*.

En 2009, la prestigiosa revista científica *Nature* publicó un artículo¹⁰ que llamó mucho la atención: un grupo de ingenieros de la empresa *Google*, mediante el análisis de simples patrones de búsqueda de usuarios en internet, pudieron predecir con gran precisión el progreso del número de casos de influenza en EUA. Esta herramienta se denominó *Google Flu Trends*. Las predicciones fueron validadas en 97% al ser cotejadas con la información oficial provista por el *Centro para el Control y la Prevención de las Enfermedades*. Una gran ventaja de la predicción de *Google* era que podía obtenerse hasta con 10 días de anticipación respecto a las cifras oficiales, además tenía un bajo costo operativo. Debido al éxito obtenido, el proyecto se amplió para también predecir el número de casos de dengue. Las predicciones se hicieron para varios países, incluido México.

Este caso se volvió un ejemplo paradigmático de *Big Data*, ya que fue diseñado a partir de una gran base de datos (de las búsquedas de los usuarios en *Google*) y, sin tener conocimiento del origen del problema (lo llevaron a cabo ingenieros en lugar de médicos), se lograron predicciones precisas.

Después las cosas cambiaron, en especial a partir de la epidemia de influenza del 2011 del virus A (H1N1), cuando las predicciones de *Google Flu Trends* fallaron notablemente, con errores de hasta 100%.¹¹ A partir de entonces el proyecto ha dejado de publicar sus resultados y se ha asociado con entidades académicas para investigar mejor el problema.¹²

Por estas razones, el caso también se volvió un ejemplo paradigmático de lo que puede salir mal al usar *Big Data* sin tener un conocimiento adecuado. Debido a esto algunos actores enfatizan la necesidad de tener personal con experiencia en programación así como en otras ramas del conocimiento y cuya formación esté certificada por instituciones de prestigio.¹³

¿Cómo se trabaja con Big Data? Estructura y procesamiento

Las tres propiedades principales de los datos, denominados las 3 v's, son:

1. **Volumen.** Se refiere a su gran volumen, que sobrepasa la capacidad de almacenamiento o procesamiento de un equipo de cómputo personal.

2. **Velocidad.** Se generan y usan a grandes velocidades, a veces incluso en tiempo real.
3. **Variiedad.** Tienen distintos orígenes y pueden ser de distinto tipo. En una base de datos pueden coexistir elementos almacenados en forma de texto, imágenes, video y audio.

Adicionalmente, entre muchas otras, se pueden incluir propiedades como los cambios en el tiempo (variabilidad), la confiabilidad de las fuentes (veracidad) y qué tan útiles son (valor). Estas propiedades pueden mezclarse dentro del mismo conjunto, pero antes de analizarlos es necesario depurarlos, estandarizarlos y conectarlos o catalogarlos de manera adecuada. A este proceso se le denomina **gestión de datos**.¹⁴

Para esto se hace uso de los **metadatos**, información complementaria que puede indicar su origen, calidad o su confiabilidad. Por ejemplo, si se tiene una colección de fotografías, éstas serían los datos y los metadatos pueden ser la fecha de creación, si ha sido editada o no, el autor, la cámara utilizada, etc. La información útil o sensible puede estar tanto en los datos como en los metadatos.

Para extraer sus cualidades se utilizan métodos avanzados de procesamiento que manejan el vasto volumen y su complejidad. A todo este conjunto de técnicas se le conoce como **Big Data Analytics** o simplemente **Analytics**. Algunos ejemplos de estas técnicas son:¹⁴

- **Minería de datos (*Data Mining*):** Se usa para encontrar patrones que no puedan localizarse con métodos tradicionales, ya sea por la complejidad o por el volumen del conjunto de datos.¹⁵ Este concepto es frecuentemente mal usado ya que su nombre pareciera indicar que consiste en extraer los datos, pero realmente consiste en obtener conocimiento de ellos, como identificar patrones de relación.
- **Aprendizaje computacional o automático (*Machine learning*):** Es una aplicación de la inteligencia artificial donde una máquina o computadora "aprende" a partir de un gran número de casos.¹⁶ Por ejemplo, a partir de analizar muchas partidas de ajedrez puede inferir las reglas del juego y cómo jugarlo. (Nota INCYTU No. 12 *Inteligencia Artificial*)
- **Análisis de redes:** Son estudios donde primero se establece un criterio para generar relaciones y representarlas en una red, luego se procede a analizar su estructura y principales componentes.

La complejidad y variedad de técnicas para el análisis ha provocado el surgimiento de una disciplina enfocada a ello, llamada **Ciencia de Datos**. (Nota INCYTU No. 12 *Inteligencia Artificial*)

¿Dónde se aplica el Big Data?

Los avances que permitieron el surgimiento del *Big Data* tienen su origen en desarrollos de la ciencia básica; desde el nacimiento de las redes de cómputo a finales de la década de los sesenta; posteriormente el descubrimiento que permitió el desarrollo de los discos duros modernos (Nobel de Física 2007),¹⁷ hasta la creación de la *World Wide Web* (WWW), que es la estructura que permitió el uso masivo de las redes de cómputo y el internet. La WWW se creó entre 1989 y 1990 por físicos de la Organización Europea para la Investigación Nuclear (CERN por sus siglas en francés), donde se desarrollan técnicas para el procesamiento de grandes cantidades de información

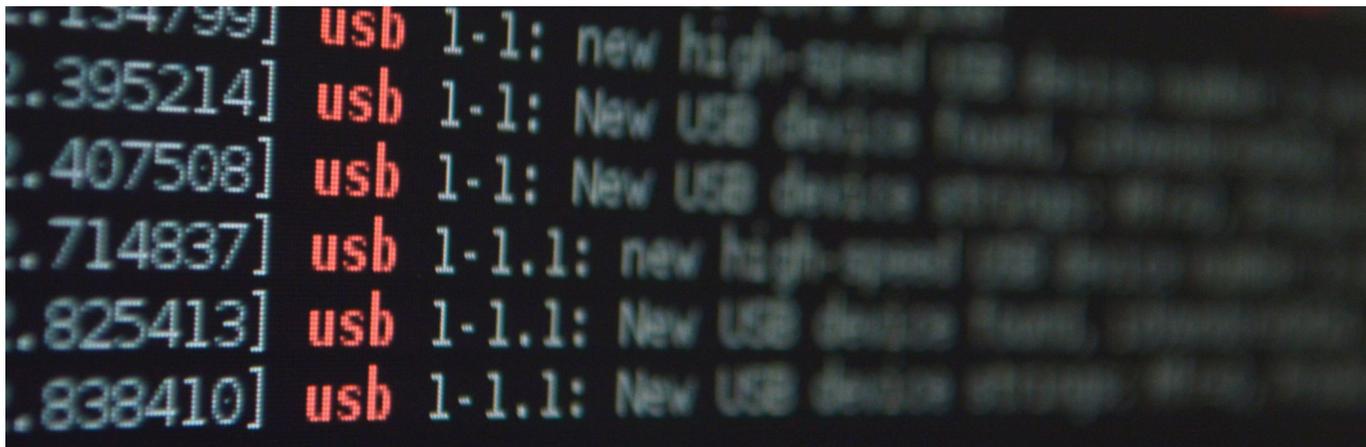


Imagen: Moisés Sánchez

y han creado la red (*Worldwide LHC computing grid*) que da acceso a miles de científicos de muchos países casi en tiempo real.^{18,19} Por otro lado, cuando está en operación el Gran Colisionador de Hadrones, el acelerador de partículas más potente del mundo, se genera 1PB (Petabytes) de información por segundo. Como referencia del significado de este volumen de información, la Biblioteca del Congreso de los EUA, la más grande del mundo, tiene una colección impresa que equivale aproximadamente a 0.01 PB²⁰ (Recuadro 2).

Recuadro 2. Unidades de información.²⁰

Todos los sistemas computacionales usan un sistema numérico binario, es decir, un sistema donde sólo se usan dos números, uno y cero. La unidad más pequeña de almacenamiento es aquella que sólo alberga un 1 o un 0, y se le denomina *bit*.

Un *byte* (B) es un arreglo de 8 bits, u ocho cifras y es la unidad más usada para medir el almacenaje de información digital. Para cada factor de mil se utiliza un prefijo diferente para ayudar en la notación. Por ejemplo, en lugar de hablar de mil bytes se usa el término *kilobyte* (kB)

Nombre (símbolo)	Equivalencia	Etimología (significado)
Kilobyte (kB)	1kB = 1000 B = 10 ³ B = 1,000 B	Χίλιοι (mil)
Megabyte (MB)	1MB = 1000 kB = 10 ⁶ B = 1,000,000 B	Μέγας (grande)
Gigabyte (GB)	1GB = 1000 MB = 10 ⁹ B = 1,000,000,000 B	Γίγας (gigante)
Terabyte (TB)	1TB = 1000 GB = 10 ¹² B = 1,000,000,000,000 B	Τέρας (monstruo)
Petabyte (PB)	1PB = 1000 TB = 10 ¹⁵ B = 1,000,000,000,000,000 B	-
Exabyte (EB)	1EB = 1000 PB = 10 ¹⁸ B = 1,000,000,000,000,000,000 B	-

Después de Tera, los prefijos ya no tienen una etimología precisa, sin embargo, siguen un patrón. La palabra “tera” es muy cercana a “tetra” (τετρα), que significa cuatro. De este modo, los siguientes sufijos vienen de los números griegos pero ligeramente modificados. Así, “peta” viene de “pente” (πέντε), cinco; “exa” viene de “hex” (ἕξ), seis; etc.

La Biblioteca del Congreso de los EUA cuenta con una colección impresa que equivale aproximadamente a 10TB. Al ser la biblioteca más grande del mundo, se usa como punto de referencia y a veces como unidad de medida.

En el sector comercio, el valor del *Big Data* es ampliamente reconocido. Por ejemplo, *Walmart* planea construir una “nube” o red de almacenamiento de datos para procesar aproximadamente 2.5 PB de información cada hora.²¹ Este centro de datos o *datacenter*, tendrá una capacidad inicial de aproximadamente 40 PB, y recopilará información de tipo mercantil, meteorológica y de redes sociales. Los eventos meteorológicos modifican el comportamiento de los consumidores y los patrones de oferta demanda, por lo que su estudio es de gran valía para las empresas.²²

De igual manera que el uso de estas tecnologías en el sector comercial ha dado luz a lo que se conoce como **Inteligencia de negocios**, en el gubernamental ahora apareció la **Inteligencia de valor público**.¹⁴ Existen numerosos ejemplos donde los macrodatos incrementan el valor público,¹⁴ que van desde aplicaciones en tránsito vehicular hasta administración de registros médicos electrónicos y biobancos (Nota INCyTU No.13 *Biobancos y Registros Médicos Electrónicos*). Gracias a su utilidad pública, se ha planteado que los macrodatos podrían ser un instrumento fundamental en ayudar a disminuir la pobreza extrema en el mundo.²³ Al uso de macrodatos en la toma de decisiones se le denomina **Data-Driven Decision Making** (toma de decisiones guiada por los datos).

¿Quién puede trabajar en Big Data? Demanda de recursos humanos y tecnológicos

Los altos requerimientos tecnológicos se pueden satisfacer mediante la oferta de computación en la “nube”, es decir, de un sistema rentado de cómputo remoto con precios accesibles. Una pequeña o mediana empresa puede alquilar todo el poder de cómputo que necesite por menos de 40 dólares por mes, de modo que no se necesita de grandes inversiones de capital. Como las computadoras alquiladas están en la red, pueden estar ubicadas físicamente en otro país, inclusive en uno distinto al que provee el servicio, por lo que se deben de revisar las leyes nacionales e internacionales al contratar estos servicios.

En relación a la formación de recursos humanos, hay una gran abundancia de materiales en internet con los que un usuario puede entrenarse en los métodos más usados. Sin embargo, algunos expertos advierten que aunque estas herramientas pueden ser muy benéficas, también pueden produ-

Legislación y Política Pública

Un estudio realizado por el *McKinsey Global Institute*² identificó seis rubros para la creación de políticas públicas adecuadas:

1. Formar capital humano de calidad.
2. Asegurar el desarrollo saludable de un mercado de datos.
3. Salvaguardar la privacidad y seguridad de los datos.
4. Establecer marcos claros de propiedad intelectual.
5. Superar las barreras tecnológicas.
6. Promover la infraestructura de las tecnologías de la información y comunicación.

México

En el ámbito legislativo, las leyes dictaminan el balance entre la protección de datos personales y la facilidad de acceso o uso de éstos, pero la necesidad de información se suele anteponer a la necesidad de privacidad.

En México existen varias leyes que rigen el uso de datos particulares como la Ley Federal de Protección de Datos Personales en Posesión de Particulares (LFPDPPP), que concierne principalmente al sector privado, y la Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados (LGPDPPSO), para el sector público. Dentro de la constitución se tiene los artículos 6, fracciones I y II, junto con el 16 y el 73.²⁸

A pesar de que estas leyes favorecen la protección de los datos personales, hace falta que los sectores de la sociedad se apeguen a ellas. Algunas organizaciones civiles y privadas han señalado que es común que las empresas no tengan las plataformas necesarias para garantizar la protección de los datos, pero que las quejas de la población al respecto tampoco son suficientes.

La implementación de protocolos para adecuarse a la ley es costoso y puede ser prohibitivo para organizaciones pequeñas y medianas, aunque algunos grupos civiles señalan que las empresas sortean con cierta facilidad las multas asociadas por no adherirse a la ley.²⁹ Dentro del sector pú-

blico, las críticas a la LGPDPPSO radican en que no siempre es lo suficientemente clara como para que los actores sepan cómo adherirse a la ley y no hay protocolos estandarizados que garanticen a las organizaciones gubernamentales cumplir con ésta.

A su vez, existen programas gubernamentales que promueven el buen uso de datos masivos, como el programa de Datos Abiertos de Presidencia o el laboratorio de datos de la Secretaría de Desarrollo Social (SEDESOL), donde se busca anonimizar la información para salvaguardar la privacidad de la ciudadanía.

Contexto Internacional

En el ámbito internacional, destacan dos actores que contrastan: EUA y la Unión Europea (UE).

En 1995, la UE creó la Directiva de Protección de Datos (Directiva 95/46/EC), la cual enfatiza el derecho de la población a la privacidad y protección de sus datos. Por ejemplo, el análisis de los datos está prohibido salvo que exista consentimiento explícito de los usuarios o que esté permitido por alguna cláusula particular. La Regulación General de la Protección de Datos (GDPR por sus siglas en inglés), implementada en mayo de 2018, actualiza esta directiva. La GDPR hace énfasis en la privacidad y homogeneiza la regulación de datos dentro de la UE para facilitar el mercado de datos internacional y establece un marco normativo claro para su flujo. Además, **la GDPR permite a los ciudadanos europeos cuestionar cualquier decisión tomada por un algoritmo o procedimiento computacional.**³⁰

En contraste, en los EUA no existe una ley federal que regule de manera general la protección de datos personales. Existen leyes estatales y algunas federales que establecen control sobre tipos particulares de datos personales. Por lo general, el análisis y uso de los datos personales no requiere el consentimiento explícito de los usuarios.

Referencias

1. Mayer-Schönberger, Viktor & Cukier K. Big Data. La revolución de los datos masivos. Primera edición. Latin Trade. Madrid. 2013.
2. McKinsey & Company. Big data: The next frontier for innovation, competition and productivity. McKinsey Global Institute. 2011.
3. El amanecer digital: actual reto para las empresas de México y Latinoamérica [Internet]. IDC Releases. [consultado 28/02/2018]. Disponible en: <http://mx.idclatin.com/releases/news.aspx?id=2135>
4. Big data será imprescindible en los próximos 20 años [Internet]. El Financiero. [consultado 28/02/2018]. Disponible en: <http://www.elfinanciero.com.mx/tech/big-data-el-petroleo-del-siglo-xxi.html>
5. En América Latina, habrá un incremento del 129% de gasto en Big Data y analítica para mejorar la experiencia del cliente hacia 2020: IDC [Internet]. IDC Releases. [consultado 28/02/2018]. Disponible en: <http://mx.idclatin.com/releases/news.aspx?id=2227>
6. Para 2018, se espera que la industria TIC en México crezca 3.8% [Internet]. IDC Releases. [consultado 28/02/2018]. Disponible en: <http://mx.idclatin.com/releases/news.aspx?id=2280>
7. Mercado de Big Data y Analytics se expande en México tres veces más que en Latinoamérica [Internet]. SAS Latin America. [consultado 28/02/2018]. Disponible en: <https://blogs.sas.com/content/sas-la/2015/02/26/mercado-de-big-data-y-analytics-se-expande-en-mexico-tres-veces-mas-que-en-latinoamerica/>
8. Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete [Internet]. Wired. [consultado 06/03/2018]. Disponible en: <https://www.wired.com/2008/06/pb-theory/>
9. Gantz J, Reinsel D. THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Idc. December 2012. p. 1-16.
10. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009. 457(7232). p. 1012-4.
11. Butler D. When Google got flu wrong. Vol. 494, Nature. 2013. p. 155-6.
12. Google Flu Trends [Internet]. [consultado 27/02/2018]. Disponible en: <https://www.google.org/flutrends/about/>
13. Conversación privada con expertos (ver agradecimientos).
14. Rodríguez P, Palomino N, Mondaca J. El uso de datos masivos y sus técnicas analíticas para el diseño e implementación de políticas públicas en Latinoamérica y el Caribe. 2017.
15. Conceptos de minería de datos | Microsoft Docs [Internet]. [consultado 09/03/2018]. Disponible en: <https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts>
16. ¿Qué es machine learning? [Guía completa para principiantes] [Internet]. [consultado 09/03/2018]. Disponible en: <https://blog.adext.com/es/machine-learning-guia-completa>
17. The 2007 Nobel Prize in Physics - Press Release [Internet]. [consultado 08/03/2018]. Disponible en: https://www.nobelprize.org/nobel_prizes/physics/laureates/2007/press.html
18. Computing | CERN [Internet]. [consultado 08/03/2018]. Disponible en: <https://home.cern/about/computing>
19. The Large Hadron Collider | CERN [Internet]. [consultado 08/03/2018]. Disponible en: <https://home.cern/topics/large-hadron-collider>
20. Lyman P, Varian HR. How Much Information? [Internet]. [consultado 08/03/2018]. Disponible en: <http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/>
21. Penta Analytics. Big Data en Walmart: Insights De Una Nube de 40+ Petabytes [Internet]. [consultado 09/03/2018]. Disponible en: <http://www.analytics.cl/big-data-walmart-insights-una-nube-40-petabytes/>
22. What Wal-Mart Knows About Customers' Habits [Internet]. The New York Times. 2004 [consultado 08/03/2018]. Disponible en: <http://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html>
23. Kim J. Using Big Data and the Internet of Things to Help End Poverty [Internet]. 2018 [consultado 12/03/2018]. Disponible en: <https://www.linkedin.com/pulse/using-big-data-internet-things-help-end-poverty-jim-kim/>
24. Conversación privada con Datank.
25. Tendencias actuales del mercado laboral | OLA [Internet]. [consultado 13/03/2018]. Disponible en: http://www.observatoriolaboral.gob.mx/static/estudios-publicaciones/Tendencias_actuales.html
26. Científicos de datos en México ganan hasta 7 veces más que ingenieros [Internet]. [consultado 13/03/2018]. Disponible en: <http://www.elfinanciero.com.mx/empresas/cientificos-de-datos-en-mexico-ganan-hasta-veces-mas-que-ingenieros.html>
27. Sweeney L. Simple Demographics Often Identify People Uniquely [Internet]. [consultado 27/02/2018]. Disponible en: <https://dataprivacylab.org/projects/identifiability/paper1.pdf>
28. Ornelas Nuñez L. El derecho a la protección de datos personales [Internet]. IFAI. 2011 [consultado 12/04/2018]. Disponible en: https://www.itei.org.mx/v3/micrositios/diplomado02/gdl/adjuntos/derecho_proteccion_datos_personales.pdf
29. Daniel Villegas. El INAI impone millones en multas, pero empresas multadas evaden pagarlas [Internet]. Son Tus Datos. [consultado 27/04/2018]. Disponible en: <https://sontusdatos.org/2016/12/08/empresas-evaden-pagar-multas/>
30. EU GDPR Information Portal [Internet]. [consultado 27/04/2018]. Disponible en: <https://www.eugdpr.org/>
31. <https://twitter.com/Nrg8000/status/957318498102865920> [consultado 09/03/2018].
32. Strava: cómo una aplicación de deportes dejó al descubierto secretos de bases militares de Estados Unidos - BBC Mundo [Internet]. [consultado 09/03/2018]. Disponible en: <http://www.bbc.com/mundo/noticias-42859883>
33. Strava Data Heat Maps Expose Military Base Locations Around the World [Internet]. [consultado 09/03/2018]. Disponible en: <https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/>