El impacto del Big-data en la Sociedad de la Información. Significado y utilidad

Antonio Monleón-Getino
Universidad de Barcelona
amonleong@ub.edu

Recibido: 13 de septiembre de 2015
Aceptado: 25 de noviembre de 2015

Resumen
Inmersos en la revolución digital, generamos constantemente datos y la mayoría son almacenados. Es lo que se ha denominado los "datos grandes" o Big-data. Junto con el capital y la fuerza de trabajo, los datos se han convertido en un valor añadido para la economía que refleja un futuro con un paradigma revolucionario en el que la sociedad será dirigida por los datos. El futuro está en la investigación, tratamiento y aplicación de los datos que aportarán prosperidad a nuestra sociedad. En el presente artículo se recogen las normas y leyes que existen hoy en día en el Big-data y la regulación existente. Para analizar Big-data la solución pasa por el aprendizaje automático (Machine-Learning) que se ocupa de la construcción y el estudio de los algoritmos que pueden aprender a partir de datos. Existen muchas técnicas, como estadística descriptiva, clasificación o agrupamiento. En este artículo se recoge las tecnologías que se utilizan para almacenar y analizar los "Big-data".

Palabras clave: Big-data; comunicación; tecnologías de la información; sociedad digital; protección de datos; informática; estadística; aprendizaje automático.

Big-data, a digital ocean in the Information Society

Abstract
Immersed in the digital revolution, we generate data consistently and most are stored. This is what has been called the Big-data. Along with capital and labor force data have become an added value to the economy that reflects a future with a revolutionary paradigm in which society will be directed by the data. The future is in research, treatment and application data that will bring prosperity to our society. In this article the rules and laws that exist today in the Big-data and existing regulations are collected. To analyze the Big-data passes through the machine learning that deals with the construction and study of algorithms that can learn from data solution. There are many techniques such as descriptive statistics, sorting or grouping. This article describes the technologies that are used to store and analyze the "Big data" is collected.

Keywords: Big-data; digital society; data protection; information; communication; information technology; statistics; machine learning.

Referencia normalizada
Monleón-Getino, A. (2015). El impacto del Big-data en la Sociedad de la Información. Significado y utilidad. *Historia y Comunicación Social.* Vol 20, número 2, páginas 427-445.

ISSN: 1137-0734
http://dx.doi.org/10.5209/rev_HICS.2015.v20.n2.51392

4.1.2. Machine Learning (Aprendizaje automático, ML) y Big-data. 4.1.3. Big-data y estadísticas. 5. Conclusiones. 6. Agradecimientos. 7. Referencias bibliográficas.

1. Introducción

Nuestra vida cotidiana está siendo observada en todo momento desde el mismo momento de abandonar nuestra casa, sin querer estoy generando datos. Cuando conectas un dispositivo digital, como el GPS del coche estás dejando un rastro digital cargado de información. Al igual que cuando envías un email o manejas un teléfono inteligente (smartphone), usas una red social, usas una tarjeta de crédito o haces la compra semanal. Son sólo unos ejemplos de la cotidianidad de la información digital.

La popularización del término Big-data hace muy pocos años se ha debido a su utilización por las grandes compañías de las tecnologías de la información a mejorar las demandas electrónicas (ventas on-line) de sus clientes e intentar orientar las compras de forma que estas fueran más dirigidas al propio cliente, más amigables y más próximas. Para ello ha sido necesario recoger la información almacenada de los clientes. La primera pregunta que la sociedad se plantea es ¿con el debido tratamiento estadístico estos datos pueden ser usados para mejorar nuestra vida o por el contrario para ser un instrumento de control por parte de las grandes corporaciones o de los propios gobiernos? ¿Cuál es la tendencia futura?

Este artículo pretende recoger el significado de estas dudas generadas por Big-data y aportar sentido crítico desde el punto de vista de un estadístico, una persona que analiza información pero que de repente se ve desbordado por una gran cantidad de datos que no cesan de llegar.

A todos nos surgen preguntas respecto al uso y existencia del océano digital. ¿Es la proliferación de datos la prueba de que el mundo es cada vez más intrusivo? ¿Podemos estar seguros de que hay un peso y un valor económico detrás de toda esta información masiva? ¿Debemos dejar a las máquinas la tarea de filtrar información y seleccionar lo que es relevante? ¿Debemos legislar el uso de esta información? Son algunas preguntas que se van a tratar de resolver, o al menos tratar de dejar planteadas.

1.1.Datos inmanejables: ¿Qué es el Big-Data?

La sociedad crea datos y más datos y cada vez existen más dispositivos y más eficientes para almacenarlos. Los datos son vistos como una infraestructura o un capital en sí mismos para la organización ya sea pública o privada que disponga de ellos. Según Chui (2011) estas grandes cantidades de datos se están convirtiendo en factores de producción esenciales dentro de cada sector productivo.

Dos estudios realizados por Manyika y otros (2011) del McKinsey Global Institute y por Andrew McAfee y Erik Brynjolfsson (2012) de la Harvard Business Review

indican que el número de datos en actualmente inmanejable. Aquí van unos ejemplos citados por estos estudios:

- El 90 por ciento de los datos del mundo ha sido creado en los últimos dos años.
- Un disco duro que contiene toda la música del mundo sólo vale unos 500€
- En el año 2010 habían ya 5.000 millones de teléfonos móviles.
- Treinta mil millones de contenidos han sido compartidos en Facebook tan sólo en un mes.
- 235 Terabytes de información fueron almacenados por la Biblioteca del Congreso estadounidense en abril de 2011.
- 15 de los 17 sectores productivos de la economía norteamericana tienen más datos almacenados en cada compañía que la Biblioteca del Congreso de los Estados Unidos de América (la mayor del mundo).
- Durante 2012, cada día se generaron alrededor de 2,5 exabytes de información. Este número se dobla aproximadamente cada 40 meses.

Las empresas capturan miles de millones de bytes de información sobre sus clientes, proveedores y sus operaciones. Millones de sensores conectados en red están presentes en dispositivos tales como teléfonos móviles, sistemas de detección o redes sociales. Las personas, bien sea con teléfonos inteligentes (smartphones) o a través de redes sociales estimulan el crecimiento exponencial de la información.

El término Big-data es confuso, ya que si son grandes datos ¿a qué tamaño se refiere? Así según la bibliografía consultada no se refiere a un tamaño de información específica (IBM, 2014), ya que es usualmente utilizado cuando se habla en términos de petabytes (PB) y exabytes (EB) de datos. La información digital se mide en bytes[1] que es la unidad básica de información, a partir de ésta se construye la escala de medida digital de bytes:

- Kylobyte (KB) = 10^3 = 1,000 bytes
- Megabyte (MB) = 10^6 = 1,000,000 bytes
- Gigabyte (GB) = 10^9 = 1,000,000,000 bytes
- Terabyte (TB) = 10^{12} = 1,000,000,000,000 bytes
- Petabyte (PB) = 10^{15} = 1,000,000,000,000,000 bytes. BIG-DATA
- Exabyte (EB) = 10^{18} = 1,000,000,000,000,000,000 bytes
- *Zettabyte (ZB)* = 10^{21} bytes
- *Yottabyte (YB)* = 10^{24} bytes
- *Quintillón (QB)* = 10^{30} bytes

En términos comparativos humanos para hacernos una idea de qué supone esta información:

- 1 Byte -Una letra
- 10 Bytes -Una o dos palabras

- 100 Bytes -Una o dos frases
- 1 kB -Una historia muy corta
- 10 kB -Una página de enciclopedia
- 100 kB -Una fotografía de resolución mediana
- 1 MB -Una novela
- 10 MB -Dos copias de la obra completa de Shakespeare
- 100 MB -1 metro de libros archivados
- 1 GB -Un pen-drive lleno de páginas con texto
- 1 TB -50.000 árboles de papel
- 10 TB -La colección impresa de la biblioteca del congreso de EEUU BIG-DATA
- 1 QB – Datos que se generan en el mundo en 1 día

Según IBM (2014) cada día se generan más de 1 QB, que surgen de fuentes tan diferentes como los datos de clientes, proveedores, operaciones financieros en línea u obtenidos de dispositivos móviles, análisis de redes sociales, ubicación geográfica mediante GPS. En muchos países se gestionan gigantescas bases de datos que contienen datos de impuestos, censo de población, registros médicos, etc., (IBM, 2014).

En un estudio realizado por la empresa tecnológica Cisco, entre el 2011 y el 2016 los datos móviles crecerán anualmente un 78% y el número de dispositivos móviles que están conectados a Internet superará la población de la Tierra. Así según un cálculo realizado en 2016 habrán unos 19 mil millones de dispositivos conectados a la red, más de 2 por habitante del planeta; entonces el tráfico global de datos móviles alcanzará 130 EB anuales. Este volumen de tráfico previsto para 2016 equivale a 33 mil millones de DVDs anuales, simplemente inacabables (Cisco, 2014).

No sólo se registran datos entre personas en el océano digital, también las máquinas los registran. 30 millones de sensores interconectados envían instantáneamente datos en el sector del automóvil, eléctrico, comercio, logístico, industrial, científico, etc. Pensemos en los contadores digitales eléctricos que las compañías están instalando en nuestros hogares. Enviarán nuestros consumos eléctricos a las compañías a intervalos regulares. Las compañías podrán disponer de un perfil fidedigno de nuestra actividad diaria, millones de TB a almacenar y analizar. Es la denominada comunicación M2M (máquina a máquina o machine-to-machine) que genera también grandes cantidades de información. Esta crecerá cada año exponencialmente.

2. Cómo afecta el Big-data a la sociedad en general

No sólo son importantes los datos y el conocimiento que nos aportan los mismos (Monleón, 2010), sino que están cambiando la economía mundial. En nuestro entorno, la Unión Europea concentra gran parte de sus actividades de investigación

e innovación en el denominado Programa Marco que en esta edición se denominará Horizonte 2020 (H2020). En el período 2014-2020 y mediante la implantación de tres pilares, contribuye a abordar los principales retos sociales, promover el liderazgo industrial en Europa y reforzar la excelencia de su base científica. H2020 promueve la generación de una economía basada en el conocimiento, así uno de los objetivos que ha fijado dentro del H2020 es el de desarrollar tecnologías y sus aplicaciones para mejorar la competitividad europea, contando y promocionando inversiones en tecnologías clave para la industria, como Tecnologías de la Información y Comunicación (TIC) (Ministerio de Economía y Competitividad, 2014).

En julio de 2014, la Comisión presentó una nueva estrategia sobre Big Data, para apoyar y acelerar la transición hacia una economía basada en los datos en el espacio europeo. La economía basada en datos estimulará la investigación y la innovación en general, mientras que lleva a más oportunidades de negocio y un aumento de la disponibilidad de los conocimientos y el capital, en particular para las pequeñas y medianas empresas (PYME). Estas afirmaciones de la Unión Europea se recogen en su artículo "Towards a thriving data-driven economy" (Unión Europea, 2014a) donde citando otras fuentes, especialmente americanas indican que se espera que la tecnología y los servicios basados en Big-Data crezcan en todo el mundo a una tasa compuesta de crecimiento anual del 40% - cerca de siete veces la del mercado de las Tecnologías de la información (TIC) en general.

¿Pero, cómo puede esta recopilación tan masiva y su posterior análisis mejorar nuestra vida? Para responder a esta pregunta debemos vislumbrar que el quid de la cuestión es el propósito con qué se hace. Según indica Sánchez (2013) en el interesante artículo periodístico "Big-data: presente y futuro para las empresas", las grandes compañías tecnológicas disponen de centros de almacenamiento para guardar estas grandes fuentes de información y tras su análisis pueden estudiar el comportamiento de los clientes para realizar acciones comerciales más efectivas o focalizar la publicidad los intereses del consumidor.

La Comisión Europea también ha publicado otro interesante artículo "Making Big Data work for Europe" (Comisión Europea. 2014b) donde indica porqué es importante el Big-data. Los datos se han convertido en un activo clave para la economía y nuestras sociedades similares a las categorías clásicas de los recursos humanos y financieros, la necesidad de dar sentido a Big-data está dando lugar a innovaciones en la tecnología, el desarrollo de nuevas herramientas y nuevas habilidades. El buen uso de los datos puede traer oportunidades a sectores más tradicionales, como el transporte, la salud o de fabricación. La Comisión Europea (2014b) cita algunos ejemplos de cómo el análisis y tratamiento de datos, sobre todo de Big-data, cambiarán la sociedad:

- Transformaran las industrias de servicios de Europa mediante la generación de una amplia gama de productos y servicios de información innovadores;
- Aumentaran la productividad de todos los sectores de la economía;
- Mejorarán la investigación y acelerar la innovación;

- Lograrán reducciones de costos a través de servicios más personalizados
- Aumentarán la eficiencia en el sector público.

En la figura 1 se presentan algunos ejemplos reales de cómo Big-data afecta o nos afectará a nuestra vida cotidiana, en cualquier ámbito y lugar.

Figura 1: Algunos ejemplos de cómo el Big-data nos afecta y afectará en nuestra vida cotidiana.

Se ha tomado una fotografía de satélite de Google-maps del entorno del Hospital de Bellvitge en Hospitalet de Llobregat (Barcelona) (Basado en un ejemplo citado en "Making Big Data work for Europe", Comisión Europea. 2014b)

Uno de los campos más prometedores es el campo de la medicina, así el análisis de los Big-data está contribuyendo a reducir los elevados costes de la investigación clínica, proporcionando medidas reales del desempeño de nuestro sistema sanitario y ayudando a los médicos y pacientes a tomar mejores decisiones (Science Spain, 2014).

Pablo Serrano, director médico del Hospital de Fuenlabrada (Madrid), durante el 59º Congreso de la Sociedad Española de Farmacia Hospitalaria (SEFH) celebrado el pasado octubre de 2014 señaló nuevos retos en el uso de estos datos, así "en el ámbito de la farmacia hospitalaria, la tecnología Big-data ayudaría a comprender mejor la utilización de los medicamentos y los integrarían en el conjunto del hospital

para un conocimiento mayor de la morbilidad y el uso de recursos". (Science Spain, 2014).

Otro ejemplo sanitario comentado por Esteban (2014) en el artículo "Cinco ejemplos de cómo el 'Big-data' puede mejorar la sociedad" sería el de las pandemias, como el ébola que recientemente se ha convertido en un problema mundial. Así mediante el Big-data se puede descubrir el riesgo de una pandemia en tiempo real a través de las tendencias que se registran en un buscador de internet como Google u otros.

Sin embargo el campo de la biología y en especial de la genética es uno de los más prometedores. Así, el avance en los últimos años en el campo de la biología y el de la bioinformática ha creado la "Era ómica", una era donde se da una visión global de los procesos biológicos basada en el análisis de un gran volumen de datos, por lo que se necesita el apoyo de la bioinformática en la interpretación de los resultados obtenidos. El análisis y la interpretación de este Big-data permiten estudiar organismos que son ahora desconocidos así como sus funciones, todo a través de su rastro genético. También se ha denominado a este tipo de estudios ciencias ómicas: la genómica, la proteómica, la transcriptómica y la metabolómica. Todas estas especialidades han hecho avanzar a velocidades antes desconocidas la biomedicina y la biotecnología. Un ejemplo sería el estudio de asociación del genoma completo (GWAS (Genome-wide association study) o WGAS (Whole genome association study). Son análisis de la variación genética con un genoma humano completo y tiene como objetivo asociar el genoma con un rasgo observable (patología), por ejemplo ayudando a identificar si una persona tiene un determinado riesgo de sufrir una enfermedad. Estos estudios requieren genotipar a un gran número de personas, obtener muestras de su genoma y analizarlo. Hoy en día gracias a las técnicas de Big-data ya se están obteniendo resultados muy prometedores con aplicaciones biomédicas o biotecnológicas a corto o medio plazo.

¿Quizás en el futuro no habrá médicos sino robots que opinen sobre nuestras enfermedades y prescriban el mejor tratamiento en base a su experiencia? Los médicos hacen diagnósticos basados en su juicio y sus conocimientos. Pero con el tiempo, esto probablemente será considerado como un disparate. ¿Por qué no utilizar los Big-data? Se podría reunir la información de la práctica habitual y la experiencia de todos los médicos, y de cientos de millones de pacientes durante años, para identificar los mejores tratamientos para lograr los mejores resultados y detectar efectos secundarios de los medicamentos adversos ocultos. Después de todo, la suma de todo el conocimiento médico no está en la posesión de un único médico ni un único investigador. Pero si agregamos gran cantidad de información sanitaria junto con información genética del paciente y conocimiento científico, podemos aprender lo que funciona mejor. De momento este planteamiento choca contra las leyes de protección de datos actuales.

Otras soluciones que puede ofrecer Big-data a los retos de la sociedad se centran en el ámbito humano y la sostenibilidad (smart cities), un tema muy debatido hoy en día. Así, se están desarrollando sistemas de información inteligentes que a partir

de sensores electrónicos instalados a pie de calle permiten cambiar la duración de las luces de los semáforos en función de los datos que recojan en tiempo real. Otro ejemplo en el ámbito económico sería el desarrollado en el Massachusetts Institute of Technology (MIT) que recoge datos continuamente sobre precios de productos comercializados en Internet y los utiliza para estimar la tasa de inflación, el objetivo es identificar rápidamente picos de inflación, lo que es mucho más rápido que los métodos tradicionales (Esteban, 2014).

3. La privacidad y la ley

Por otro lado existe un problema muy importante para la sociedad: la privacidad y el control de la información. ¿Qué podría pasar si caen en manos ajenas todos nuestros datos personales?: consumo eléctrico diario, datos personales recopilados en internet, qué compramos, banca-online, que libros leemos, actividad diaria, datos genéticos de propensión a enfermedades, etc

Hoy en día existen sofisticadísimos algoritmos que pueden hacerse predicciones sobre lo que somos proclives a hacer, así con un método de análisis de tendencias a partir de Big-data de datos personales podríamos juzgar a las personas antes de que realmente hayan cometido la infracción. ¿Existe el libre albedrío con el Big-data? Así, se podría asignar una probabilidad del 95% que una determinada persona cometerá un determinado tipo de delito, o dejará de pagar el préstamo hipotecario, o no sobrevivirá a una determinada intervención quirúrgica, o vivirá tanto tiempo. Pero ¿pueden utilizar las empresas nuestros datos personales para hacer este tipo de análisis?

Actualmente, la privacidad es el gran problema y será un problema más grande en el futuro en la sociedad de la información. Hoy en día la legislación protege la privacidad de las personas a través de un método de notificación y consentimiento (la empresas informa a los usuarios qué datos se recopilan y cómo se utilizarán posteriormente, y el usuario da su consentimiento). Hay que tomar consciencia de que cuando firmamos un consentimiento para que un sistema informático grabe nuestros datos estamos ofreciendo información que puede ser utilizada posteriormente contra nosotros, es el rastro o huella digital.

No obstante, aún hay retos que afrontar, como la adopción de un marco regulatorio adecuado para el Big-data.

En la entrevista realizada a José Luís Antúnez director de "Big-data innovation" de la empresa Telefónica, se le preguntó por la protección de datos en este tipo de situaciones. Según indica Antúnez "Por definición, los datos individuales sólo se pueden usar para el propósito del servicio que se ha contratado: por ejemplo, si doy mis datos personales a una entidad financiera, ésta los deberá utilizar exclusivamente para mis operaciones bancarias; en cambio una operadora móvil los deberá usar para darme servicios de telecomunicaciones. El único propósito en que los datos se pueden utilizar individualizados y con toda su riqueza de detalle es en el contexto del servicio

primario (principal) que ofrece la empresa que nos presta un servicio" (Big-data: protección de datos personales. Utilidades en la sociedad, 2014).

Pero creo conveniente repasar qué dice la legislación actual tanto española como comunitaria. La Ley Orgánica de Protección de Datos (LOPD) es la que protege este uso, y la Agencia de Protección de Datos Española (APDE) es la que tiene competencias de que las empresas cumplan esta normativa: usar datos personales solamente para la prestación del servicio original. Esta normativa está recogida en la ley y está disponible en el portal de la APDE (http://www.agpd.es) donde existe un canal del ciudadano a través del que pueden preguntarse dudas o emitir quejas y reclamaciones. La APDE indica que la LOPD permite posicionarse en un ámbito seguro en el que no se está vulnerando ningún dato de carácter personal, trabajando con datos estadísticos, así para usos secundarios, se aplica una anonimización y agregación de los datos, que dan lugar a una información de carácter estadístico, como por ejemplo el censo, que ya se consideran como un cálculo con el que se puede trabajar abiertamente.

En la misma entrevista a Antúnez (2014), éste indica a propósito de la legislación Europea que tanto ha chocado con los gigantes informáticos como Google "En Europa también se están poniendo las medidas necesarias para proteger a la sociedad digital. En el caso de España, tenemos la segunda normativa de protección de datos personales más exigente después de la alemana, y eso da garantías". Actualmente, cuando se visita una página web aparece un indicador (banner) avisando de que van a recoger señales de nuestra actividad (cookies), así la actual legislación indica que sólo hay obligación de informar de qué datos se recogen pero no sobre el uso que se le va a dar a esos datos, como por ejemplo el análisis del comportamiento de los usuarios (" tracking"). Según parece la tendencia será a regular más y mejor el uso de los datos personales, ya que hoy en día no queda claro qué hacen las empresas con los datos del comportamiento online de los usuarios.

4. Tipos de Big-data

¿Pero, cómo funciona desde dentro el Big-Data? O mejor preguntado de otra forma ¿qué problema es el que se está tratando de resolver? (Soares, 2014).

Veamos ahora tipos de datos que vamos a encontrarnos en el Big-data según IBM (2014). Esta empresa ha clasificado en 5 los tipos de datos de Big-data, que aunque no es fácil de distinguir en algunos casos si pueden permitir tener una idea general. En la tabla 1 se presenta un ejemplo de estos 5 tipos de datos.

Tabla 1: Tipos de datos del Big-data según IBM
(Recopilado a partir de IBM, 2014)

Datos	Descripción	Ejemplos
Web and Social Media	Contenido web e información que es obtenida de las redes sociales	www, Facebook, Twitter, LinkedIn, blogs
Machine-to-Machine (M2M):	Tecnologías que permiten conectarse a otros dispositivos. M2M utiliza dispositivos como sensores o medidores que capturan algún acontecimiento en particular. Se transmiten a través de redes alámbricas, inalámbricas o híbridas.	Velocidad, temperatura, presión, variables meteorológicas, variables químicas
Big Transaction Data	Incluye datos procedentes de transacciones masivas de los centros de atención telefónica, de banca, finanzas, atención a clientes, etc	Incluye registros de facturación, en telecomunicaciones los llamados registros detallados de las llamadas (Call Detail Record o CDR), etc.
Biometrics	Información biométrica. En el área de seguridad e inteligencia, los datos biométricos son sumamente importantes para los gobiernos, seguridad privada, servicios de inteligencia, policía, etc	Huellas digitales, escaneo de la retina, reconocimiento facial, genética, etc
Human Generated	Datos digitales generados por las personas, en sentido genérico	Notas de voz, correos electrónicos, documentos electrónicos, resultados de estudios médicos, multas, etc

4.1. El análisis de la información

Es importante entender que las bases de datos convencionales son una parte importante y relevante para recoger los datos Big-data y su posterior análisis. Una plataforma de Big-data consiste en un conjunto de herramientas informáticas y estadísticas que permiten simplificar, administrar, coordinar y analizar grandes volúmenes de información.

Un ejemplo de plataforma Big-data es la plataforma de código abierto *Hadoop (Software for reliable, scalable, distributed computing) que puede encontrarse en* http://hadoop.apache.org/ . *Según* The Apache Software Foundation (2014) *Hadoop* está inspirado en el proyecto de Google File System (GFS) donde se utiliza un sistema de almacenamiento de ficheros distribuido denominado HDFS. Para gestionar todos estos ficheros de una manera eficaz se utiliza un proceso denominado *MapReduce*, este proceso consiste en dividir las tareas en dos partes (mapper – reducer) para manipular los datos distribuidos hacia los nodos de un clúster logrando así un gran

paralelismo en el procesamiento. Se puede encontrar más información en "Introducción a Hadoop y su ecosistema (Ticou, 2014).

No sólo existe Hadoop sino muchos otros sistemas, una buena revisión de los mismos puede encontrarse en el artículo ¿Qué es Big Data? IBM (2014).

4.1.1.Machine learning (Aprendizaje automático, ML) y Big-data

Pero surge la pregunta de ¿Cómo se procesa y analiza la información Big-data, almacenada y distribuida en Hadoop o en una base de datos o en cualquier otro tipo de almacenamiento? Es decir, ¿cómo analizar información de ese volumen de datos masivo sin perder información por el camino? La solución pasa por un viejo conocido: el aprendizaje automático (Machine Learning o ML).

Kovahi y Provost (1998) definen el ML como una disciplina proveniente de las ciencias y la ingeniería que se ocupa de la construcción y el estudio de los algoritmos que pueden aprender a partir de datos, en este caso de los Big-data. Los algoritmos ML intentan construir modelos basándose en los datos con objetivo de hacer predicciones o tomar decisiones, como si fueran verdaderos expertos, en lugar de seguir las instrucciones de manera explícita para lo que han sido programados (Bishop, 2006).

Las técnicas de ML que se revisan ampliamente en trabajos muy recientes como en "An Introduction to Statistical Learning: with Applications in R (James et al, 2013), proceden de campos como la Estadística, Reconocimiento de Patrones, Inteligencia Artificial y Minería de Datos. Estas técnicas transforman los datos en conocimiento y proporcionan sistemas de propósito general que se adaptan a las circunstancias, por ejemplo a los problemas computacionales de utilizar el Big-data. Entre las considerables aplicaciones pueden citarse: los sistemas de diagnóstico clínico, el reconocimiento de voz o de texto escrito a mano (OCR), navegación de robots autónomos (coches que circulan sin conductor, drones voladores sin pilotos o guía humana, etc), sistemas de recuperación de información, filtrado cooperativo (El algoritmo ML recomienda opciones en función de lo que el usuario ha comprado o ha realizado anteriormente), análisis genético de microarrays de ADN, etc.

Algunos sistemas ML son muy sofisticados e intentan eliminar cualquier necesidad de percepción o conocimiento experto de los procesos de análisis de datos, mientras otros tratan de establecer una colaboración entre la persona que entiende del tema concreto (experto) y la computadora, es decir una interacción. La intuición humana no puede ser reemplazada, ya que el creador del sistema ha de especificar la forma de representación de los datos, las formas de manipulación y caracterización de los mismos (Wikipedia, 2014).

Una muy buena revisión de los diferentes algoritmos de ML para el Big-Data puede encontrarse en "Too Big to Ignore: The Business Case for Big-data" (2013) o en "An Introduction to Statistical Learning" (2013) y también algunas técnicas para el análisis de datos masivos que nos ocupa puede encontrarse en "Programming with Big-data in R" (http://r-pbd.org/). Los diferentes algoritmos de ML se agrupan según

una taxonomía que se basa en los resultados deseados del algoritmo y que se presentan en la tabla 2, junto a algunos ejemplos.

Tabla 2: Algoritmos utilizados en el aprendizaje automático (ML) (Recopilado a partir de Wikipedia, 2014)

Tipo de aprendizaje	Descripción	Ejemplos
Aprendizaje supervisado	El algoritmo utilizado produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Este tipo de aprendizaje puede llegar a ser muy útil en problemas de investigación biológica, biología computacional y bioinformática.	Programa informático que clasifica el mail como " spam" o "no spam" Este es un problema de clasificación, donde el sistema de aprendizaje trata de etiquetar (clasificar) una serie de vectores utilizando una entre varias categorías (clases). La base de conocimiento del sistema está formada por ejemplos de etiquetados realizados anteriormente por el usuario.
Aprendizaje no supervisado	Todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías previas. El algoritmo tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas.	Un robot que minimiza la energía consumida en función de lo que indican los sensores que posee (temperatura, estado de la batería, etc)
Aprendizaje semisupervisado	Este tipo de algoritmos combinan los algoritmos anteriores para poder clasificar de manera adecuada. Se tiene en cuenta los datos marcados y los no marcados.	Algún dispositivo que permitiera una mezcla de los dos tipos anteriores.
Aprendizaje por refuerzo	El algoritmo aprende observando el mundo que le rodea. Su información de entrada es la retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error. Hay un supervisor que da información al agente sobre si lo está haciendo bien o mal, pero no exactamente lo que debe hacer.	Robot experto que aprende del mundo exterior en base a ensayo-error.

Transducción	Similar al aprendizaje supervisado, pero el algoritmo no construye de forma explícita una función, ya que los datos no tienen etiqueta, estan sin clasificar. Se pretende pronosticar las categorías de los futuros ejemplos basándose en los ejemplos de entrada, sus respectivas categorías y los ejemplos nuevos al sistema.	Análisis automático de texto, aplicaciones de la bioinformática.
Aprendizaje multi-tarea o multiinstancia	Éste algoritmo implica la resolución simultánea de distintas tareas; en particular, el aprendizaje de una tarea se ve mejorado y completado por el aprendizaje común con otras tareas relacionadas con la primera	Imputación de datos incompletos y clasificación de patrones mediante aprendizaje multitarea

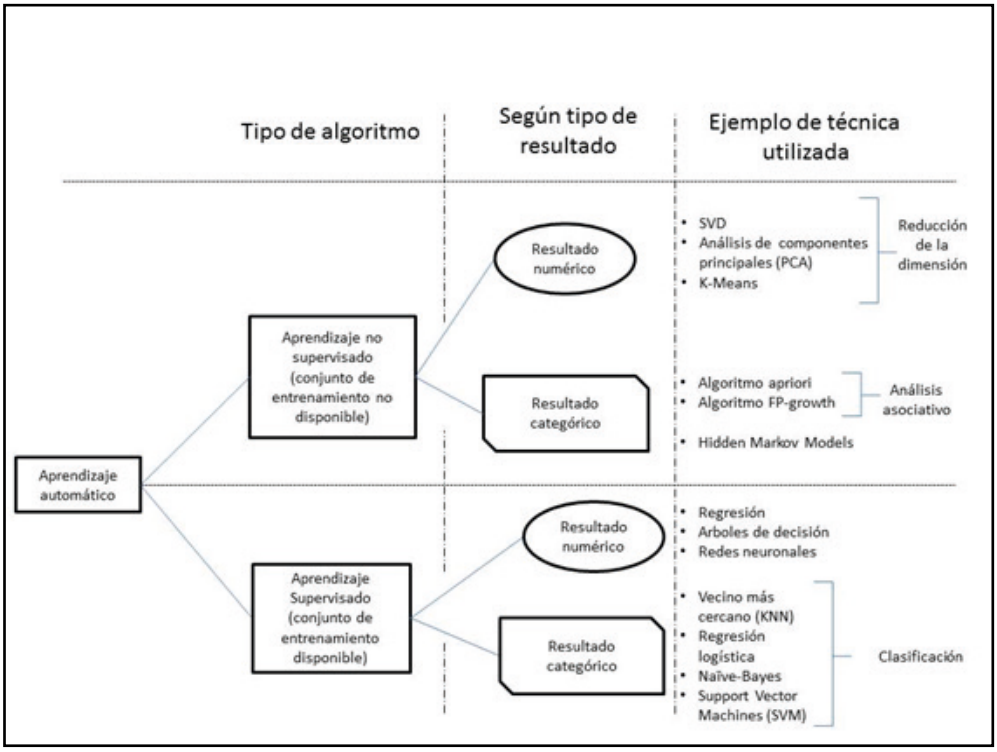
Actualmente para poder analizar datos tan masivos como los Big-data que en general se encuentran de manera distribuida en múltiples ordenadores, bases de datos, repositorios, etc mediante algoritmos basados en ML existen muchas interpretaciones, variantes y técnicas. Una de ellas es utilizar programas de código abierto y gratuitos como R package (http://www.r-project.org/) basado en el lenguaje R. Este es un lenguaje y un entorno para computación y gráficos estadísticos. R ofrece una amplia variedad de técnicas gráficas y estadísticas (análisis de series temporales, modelado lineal y no lineal, pruebas estadísticas clásicas, clasificación, agrupamiento, etc) y es altamente modulable y ampliable. Estos grandes grupos de técnicas y un esquema general pueden verse en las tablas 2 y 3.

4.1.2.Big data y estadística

Las principales técnicas estadísticas (Schmidberger, 2013) que se utilizan en el análisis del Big-Data y como parte del ML, pueden verse más en detalle en la figura 2 y pueden resumirse como:

- Clasificación: Consiste en asignar una clase a un determinado objeto o individuo. La salida del sistema es una "etiqueta". Así se podría clasificar un determinado producto comercial como "bueno" o "malo" según sus características.
- Regresión: Es una generalización del problema de la clasificación. La salida del sistema es un número o un vector de números reales. Se podría predecir el incremento de ventas de un determinado producto a partir de las consultas en web de un catálogo comercial.
- Clustering (agrupamiento): Técnicas para organizar objetos o individuos en grupos que tengan sentido. Agrupación de objetos o clases que puede ser jerárquica o no jerárquica.

Figura 2: Esquema de los principales algoritmos ML para el tratamiento de los Big-data clasificados según el tipo de resultado (numérico o categórico), así como ejemplos de técnicas utilizados dentro de los algoritmos.



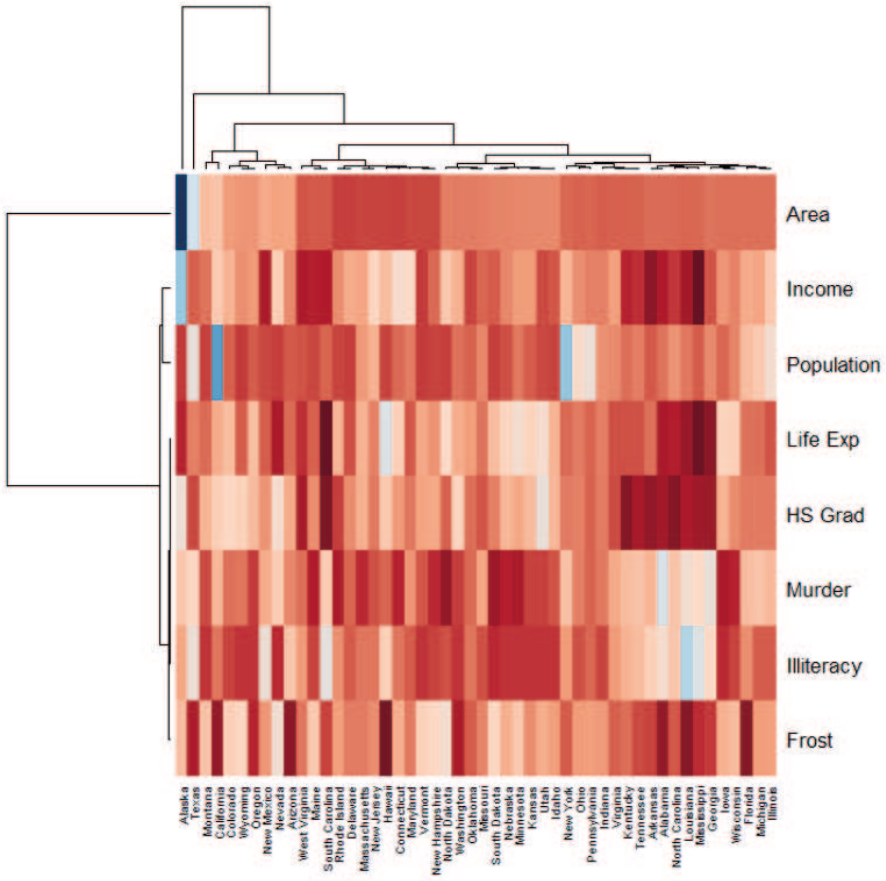
Existen numerosos packages de R que implementan los algoritmos ML comentados en las figuras 2 y tabla 2. Es ampliamente utilizado por la comunidad científica, académica y todo tipo de usuarios. Es gratuito y fácil de utilizar con un cierto tiempo de aprendizaje. Tiene gran variedad de librerías disponibles, y una muy buena revisión puede encontrarse en "Big-data in R" (2010) y en la recopilación "Machine Learning & Statistical Learning in R" (http://cran.r-project.org/web/views/MachineLearning.html).

Algunas librerías de R dedicadas al uso con los big-data son: RODBC Package, biglm, ff, bigmemory o snow. Podemos encontrar dentro de los ejemplos de la librería ff cómo realizar el tratamiento estadístico de los datos comerciales de la ASA 2009 (Airline on-time performance http://stat-computing.org/dataexpo/2009/). Estos datos consisten en los tiempos de llegada y partida de vuelos comerciales de aviones de Estados Unidos desde octubre de 1987 a abril de 2008. Son unos 120 millones de registros, de 29 variables que son intratables con un programa estadístico normal. En el artículo "Big-data in R" (2010) podemos encontrar cómo tratar y modelar este tipo

de información tan enorme para poder obtener resultados provechosos que optimicen las rutas de vuelo o expliquen las causas de los retrasos, etc.

Unos libros de referencia sobre machine learning y sobre técnicas de análisis estadístico con Big-data actuales son "Big-data Analytics with R and Hadoop" (Prajapati, 2013) y "Big-data Analyses with R" (Schmidberger, 2013). Contienen muchos ejemplos y tratan muchas técnicas para el análisis con este tipo de datos como los que se presentan en la figura 3, un mapa de colores que permite la interpretación y agrupación de datos y variables.

Figura 3: Heatmap biclustering:



El mapa de color es utilizado para la interpretación de análisis con Big-data. En este caso se trata de un ejemplo del análisis de datos poblacional de 50 estados de Estados Unidos, donde se han recogido datos sobre la población, ingresos, analfabetismo, esperanza de vida, asesinato, escuela secundaria, graduación, número de días con heladas y el área del estado Datos procedentes de Aedin Culhane (2012), analizados mediante el paquete estadístico R.

5. Conclusiones

La revolución digital ha hecho que dispositivos de cualquier tamaño y un sinfín de aplicaciones informáticas (ordenadores, teléfonos inteligentes, dispositivos digitales, sensores, micrófonos, cámaras, escáneres médicos, imágenes, redes sociales, etc) están presentes en nuestras vidas.

La sociedad genera constantemente datos y la mayoría son almacenados digitalmente; son información geográfica, estadística, datos meteorológicos, datos de la investigación, datos de transporte, datos de consumo de energía, datos de salud, redes sociales, banca on-line, etc. Es lo que se ha denominado los grandes datos o Big-data, aunque técnicamente se empieza a hablar de Big-data a partir de petabytes (PB) y exabytes (EB) de datos.

Cada día se registran 2,5 QB de datos, tanto que el 90% de los datos en todo el mundo se ha generado tan solo en los últimos dos años. Los datos generados a partir de estos elementos son ya el segmento más grande de toda la información disponible por la humanidad. Estos datos son almacenados en sistemas informáticos que pueden manejan estos grandes volúmenes de información, capaces de analizar y generar conocimiento.

El Big-data marca una nueva revolución como la de la llegada de la electricidad a nuestra sociedad, en cómo la sociedad genera y utiliza esta montaña de información instantánea y continua.

Junto con el capital y la fuerza de trabajo los datos se han convertido en un elemento fundamental para la economía que refleja una revolución en la que la sociedad será dirigida por los datos y la Comisión Europea quiere poner a Europa en la primera categoría mundial en materia de datos. En julio de 2014, la Comisión Europea presentó una nueva estrategia sobre Big Data, para apoyar y acelerar la transición hacia una economía basada en los datos en Europa. Se prevé la creación de miles de nuevos puestos de trabajo fundados en que la economía basada en los datos estimulará la investigación y la innovación aportando más oportunidades de negocio y el aumento de la disponibilidad del conocimientos y el capital, en particular para las pequeñas y medianas empresas (PYME), en toda Europa.

Con la capacidad de generar toda esta información valiosa procedente de los sistemas digitales, las empresas y los gobiernos están lidiando con el problema de analizar los datos para dos propósitos importantes: ser capaces de detectar y responder a los acontecimientos actuales de una manera oportuna, y para poder utilizar las predicciones del aprendizaje histórico con todos los datos recogidos anteriormente. Esta situación requiere del análisis tanto de datos continuos (datos actuales) como de datos precedentes (datos históricos), que son representados en diferentes y enormes volúmenes, variedades y velocidades.

Por otro lado existe un problema muy importante en el tratamiento y almacenamiento de los Big-data: la privacidad y el control de la información. La Ley Orgánica de Protección de Datos (LOPD) española es la que protege este uso, y la Agencia de

Protección de Datos Española (APDE) es la que tiene competencias de que las empresas cumplan esta normativa. No puede pararse la evolución tecnológica pero si debemos promover la regulación legal del tratamiento del big-data y su análisis en favor de los individuos, cuidando de la información que de ellos se recoge, cumpliendo la legislación vigente. Las leyes actuales aunque suficientes no responden a todos los retos que se están generando y esto ha abierto el debate en la sociedad que no ha hecho más que comenzar.

Para procesar y analizar la información Big-data, almacenada y distribuida en grandes sistemas distribuidos como Hadoop o en una base de datos o en cualquier otro tipo de almacenamiento la solución pasa por el aprendizaje automático (Machine Learning o ML). Las bases del ML provienen de las ciencias y la ingeniería y se ocupa de la construcción y el estudio de los algoritmos que pueden aprender a partir de datos, en este caso de los Big-data. Existen diversas técnicas disponibles, como modelado lineal y no lineal, estadística descriptiva, clasificación, agrupamiento, etc. Una es utilizar programas de código abierto y gratuitos como R package basado en el lenguaje R.

6. Agradecimientos

Mi agradecimiento al Dr. Esteban Vegas del Departamento de Estadística de la Universidad de Barcelona, especialista en Machine Learning y estadística por haber revisado el texto.

7. Referencias bibliográficas

BISHOP CM. 2006. *Pattern Recognition and Machine Learning*. Springer

FOSTER KEVIN, Nathan Senthil, RAJAN DEEPAK, Ballard Chuck, *IBM InfoSphere Streams: Assembling Continuous Insight in the Information Revolution,* IBM RedBooks, 2011

GARETH James; Daniela WITTEN; Trevor HASTIE; Robert TIBSHIRANI. 2013. *An Introduction to Statistical Learning*. Springer. p. vii.

JAMES G, WITTEN D, HASTIE T, TIBSHIRANI R. 2013. *An Introduction to Statistical Learning: with Applications in R.* Springer.

MANYIKA J, CHUI M, BROWN B, BUGHIN J, DOBBS R, ROXBURGH C, and HUNG-BYERS Angela. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. May 2011.

MONLEÓN-GETINO, A. 2010. El tratamiento numérico de la realidad. Reflexiones sobre la importancia actual de la estadística en la sociedad de la información. *Arbor*, Vol 186, No 743.

PENDLAND, Alez. 2014. Una sociedad dirigida por datos. Tecnologías de la información. Número especial: la era de los macrodatos. *Investigación y ciencia*. Enero 2014. Nº 448.

PRAJAPATI V. 2013. *Big Data Analytics with R and Hadoop*. — Packt Publishing. — 238 pp.

RON KOVAHI; Foster Provost. 1998. "Glossary of terms". *Machine Learning* 30: 271–274

SIMON P. 2013. *Too Big to Ignore: The Business Case for Big Data*. Wiley. p. 89.

YEAGER M, ORR N, HAYES RB, JACOBS KB, KRAFT P, WACHOLDER S, MINICHIELLO MJ, FEARNHEAD P, YU K, CHATTERJEE N, et al. 2007. *Genome-wide association study of prostate cancer identifies a second risk locus at 8q24*. Nature Genetics May; 39(5):645-9

ZIKOPOLOUS Paul, DEROOS Dirk, DEUTSCH Tom, LAPIS George. 2012. *Understanding Big Data: Analytics for Enterprise* Class Hadoop and Streaming Data, McGraw-Hill

7.1 Bibliografía (fuentes web)

AEDIN Culhane. 2012. *Introduction to Programming in R.* Course Website: http://bcb.dfci.harvard.edu/~aedin/courses/R/CDC/Intro2R.pdf. CENTERS FOR DISEASE CONTROL AND PREVENTION. 26-28TH JUNE 2012

Andrew MCAFEE, Erik BRYNJOLFSSON. 2012. *Big Data: The Management Revolution.* Harvard Business School. [Fuente 09/12/2013: https://hbr.org/2012/10/big-data-the-management-revolution/ar]

Aprendizaje automático (Machine learning). [Fuente: 2014 http://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico]

Big Data: presente y futuro para las empresas. Sánchez JM. ABC. Tecnología [Fuente: 09/12/2013 - 01.36h, http://www.abc.es/tecnologia/informatica-soluciones/20131207/abci-data-analisis-201312051430.html]

Big Data: protección de datos personales. Utilidades en la sociedad. [Fuente: 16/7/2014. http://thegreatiproject.com/big-data-respetando-la-proteccion-de-datos-personales-utilidades-en-la-sociedad/#]

Big-data in R, 2010. [Fuente: 09/12/2013: http://statistics.org.il/wp-content/uploads/2010/04/Big_Memory%20V0.pdf]

Cisco, 2014. *Internet será cuatro veces más grande en 2016,* Artículo Web [Fuente: 09/12/2013 http://www.cisco.com/web/ES/about/press/2012/2012-05-30-internet-sera-cuatro-veces-mas-grande-en-2016--informe-vini-de-cisco.html]

CLEGG Dai, Big Data: *The Data Velocity Discussion,* Artículo Web [Fuente 09/12/2013: http://thinking.netezza.com/blog/big-data-data-velocity-discussionn]

Comisión Europea. 2014a. *Towards a thriving data-driven economy.* [Fuente: http://ec.europa.eu/digital-agenda/en/towards-thriving-data-driven-economy.]

Comisión Europea. 2014b. *Making Big Data work for Europe.* [Fuente 09/12/2013: http://ec.europa.eu/digital-agenda/en/making-big-data-work-europe]

Flor de Esteban. *Cinco ejemplos de cómo el 'Big Data' puede mejorar la sociedad* [Fuente 09/12/2013: http://www.daemonquest.com/es/blog/cinco-ejemplos-de-como-el-big-data-puede-mejorar-la-sociedad/]

IBM. *¿Qué es Big Data?. [*Fuente 09/12/2013: http://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/]

KOBIELUS James, *Big Data Analytics Helps Researchers Drill Deeper into Multiple Sclerosis*, Artículo Web [Fuente 09/12/2013: http://thinking.netezza.com/blog/big-data-analytics-helps-researchers-drill-deeper-multiple-sclerosis]

MANYIKA J, Michael CHUI, Brad BROWN, Jacques BUGHIN, Richard DOBBS, Charles ROXBURGH, Angela HUNG BYERS McKinsey Global Institute. *Big data: The next frontier for innovation, competition, and productivity.* May 2011 [Fuente 09/12/2013: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation] |

Ministerio de Economía y Competitividad. Gobierno de España. 2014. *Horizonte2020*. [Fuente: http://eshorizonte2020.es/que-es-horizonte-2020]

Science Spain. *La tecnología 'Big data' podría agilizar y abaratar la gestión clínica* [Fuente 09/12/2013: http://www.dicyt.com/noticias/la-tecnologia-big-data-podria-agilizar-y-abaratar-la-gestion-clinica]

SOARES Sunil, *Not Your Type? Big Data Matchmaker On Five Data Types You Need To Explore Today.* [Fuente 09/12/2013: http://www.dataversity.net/not-your-type-big-data-matchmaker-on-five-data-types-you-need-to-explore-today/]

The Apache Software Foundation. 2014. *Hadoop* [Fuente 09/12/2013: *http://hadoop.apache.org/*]

Thiago FERRER MORINI. *'Big data' a la velocidad de la luz. La revolución económica que promete el análisis de grandes cantidades de datos se enfrenta al reto de que la infraestructura sea capaz de seguirle el ritmo* [Fuente 09/12/2013: http://economia.elpais.com/economia/2014/10/24/actualidad/1414168980_618761.html]

TICOUT, outsorcing center. 2014. *Introducción a Hadoop y su ecosistema* [Fuente 09/12/2013: http://www.ticout.com/blog/tag/big-data/page/2/]

Wikipedia. *Machine learning 2014*. [Fuente 09/12/2013: http://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico]

Notas

1 El byte es la unidad de información digital básica que es utilizada como un múltiplo del bit. Generalmente equivale a 8 bits. En español se le denomina octeto.