Diciembre 2015

frA
REVISTA DE CIENCIAS Y HUMANIDADES
FUNDACIÓN RAMÓN ARECES

BIG DATA

Impulsando el conocimiento

Núm. 14

REVISTA DE CIENCIAS Y HUMANIDADES DE LA FUNDACIÓN RAMÓN ARECES

frA

FUNDACIÓN RAMÓN ARECES

Compartimos el Conocimiento

Visítanos en fundacionareces.es *o en* fundacionareces.tv

y síguenos en flickr slideshare YouTube

DICIEMBRE '15

Edita
Fundación Ramón Areces

Director
Raimundo Pérez-Hernández y Torra

Consejo Asesor
Federico Mayor Zaragoza, Jaime Terceiro Lomba, Julio R. Villanueva, Juan Velarde Fuertes, Avelino Corma Canós, Alfonso Novales Cinca, Juan González-Palomino Jiménez

Director
Manuel Azcona

Servicio de Publicaciones
Consuelo Moreno Hervás

Diseño y maquetación
Omnívoros. Brand Design & Business Communication

Administración y redacción
Calle Vitruvio, 5. 28006 Madrid. Teléfono: 91 515 89 80. Fax: 91 564 52 43

***Coordinador del Especial* Big Data**
Julio Cerezo Gilarranz

Web
www.fundacionareces.es

Web TV
www.fundacionareces.tv

Blog Ágora
www.agorafundacionareces.es

Ilustraciones
Roberto Díez (Portada) y Carlos Pan

Fotomecánica
Gamacolor S.G.I.

Impresión
V.A. Impresores

Depósito Legal: M-51664-2009

Síguenos en
flickr slideshare
YouTube

ÍNDICE

- *Conferencia disponible en fundacionareces.tv*

EL FENÓMENO *BIG DATA* EN LA FUNDACIÓN RAMÓN ARECES

Raimundo Pérez-Hernández y Torra
Director de la Fundación Ramón Areces

La Fundación Ramón Areces, siempre atenta a los desarrollos de vanguardia, ha hecho de la Ciencia de los Datos (*Data Science* o *Big Data*) uno de los campos científicos prioritarios en su labor de mecenazgo, promoción y difusión del conocimiento en las Ciencias Sociales. Dentro de esta línea de actuaciones, que comenzaron a principios de 2013 con la primera jornada dedicada al impacto de la Nube y el *Big Data* en la Ciencia, se encuadran los dos seminarios programados para 2016 en Madrid y Barcelona.

La creciente utilización de bases de datos cada vez más grandes y heterogéneas hace del estudio de las técnicas aplicadas al *Big Data* una de las disciplinas más innovadoras y atractivas de los desarrollos científicos recientes así como de su aplicación empresarial.

Durante los últimos siglos la Ciencia, que fue eminentemente empírica con anterioridad, comenzó a adentrarse en la modelización y la formulación matemática en búsqueda de la generalización. En las últimas décadas los datos han vuelto a tomar la iniciativa en forma de Ciencia computacional relacionada con la simulación de procesos complejos y la utilización de datos masivos para la predicción de acontecimientos difíciles de prever. Los datos masivos que recogemos automáticamente por sensores digitales están transformando nuestra sociedad permitiendo mejores decisiones individuales y colectivas.

Este número monográfico de nuestra revista incluye artículos y presentaciones relacionadas con las Jornadas que la Fundación Ramón Areces ha dedicado a analizar las posibilidades y limitaciones del *Cloud Computing* y del *Big Data*, su impacto y contribución tanto en los procesos de investigación científica como en la gestión económica y empresarial así como en el estudio del cambio climático.

El conjunto de los textos aquí recogidos, del que son autores relevantes expertos en la materia, aportarán, sin duda, luz y mayor comprensión sobre el *BIG DATA* y todas sus aplicaciones. A todos ellos deseo expresarles mi mayor agradecimiento.

BIG DATA, LA NUEVA CIENCIA DEL SIGLO XXI

Por Julio Cerezo Gilarranz
Coordinador de las Jornadas sobre Nube y Big Data
celebradas en la Fundación Ramón Areces

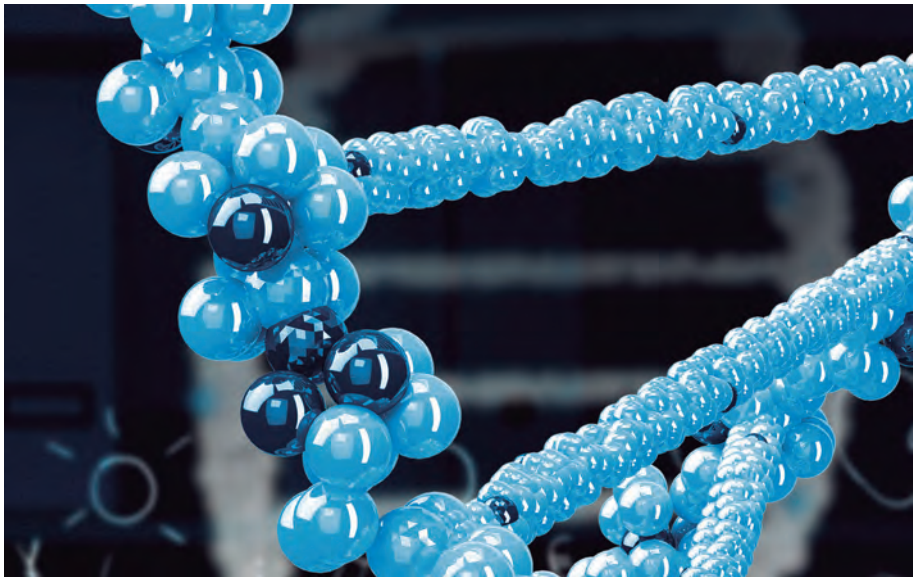
La Computación *Cloud* (la Nube) y el *Big Data* son dos de los principales campos de innovación actuales asociados a las tecnologías de la Información y Comunicación (TIC). Su irrupción ha implicado la aparición de nuevos desarrollos tecnológicos que están transformando profundamente nuestro entorno económico, empresarial, social y, por supuesto, también el científico. A pesar de su corta edad –-la primera mención al *Big Data* en un documento científico no llega a los 20 años (1)–, su conocimiento se ha popularizado en muy poco tiempo y, dado su carácter transversal, la Nube y el *Big Data* se han hecho omnipresentes en muy diversos campos de la actividad humana. De las redes sociales a las ciudades inteligentes, de los nuevos servicios de ocio en la Nube al periodismo de datos, de la creación de contenidos a la monitorización del cambio climático.

Y, por supuesto, también han impactado notablemente en el mundo de la Ciencia, no solo porque hayan sido científicos –ingenieros, matemáticos, físicos–, trabajando mayoritariamente en universidades y centros de investigación, quienes han definido y caracterizado estos nuevos sistemas y quienes están haciendo posible la evolución de las nuevas herramientas tecnológicas. La Ciencia también ha encontrado en las funcionalidades y posibilidades que ofrecen la Nube y el *Big Data* unos nuevos aliados para mejorar la efectividad de sus propios sistemas y medios de investigación. Y, además, han permitido ampliar las fronteras mismas de la Ciencia, situando a su alcance retos o fenómenos inabarcables o inaccesibles hasta ahora.

LA CIENCIA DE LOS DATOS

Una contribución de tal envergadura hasta el punto de transformar también la propia esencia de la investigación científica.

La Ciencia en sus inicios fue empírica, vinculada a la experiencia, y se centraba en describir los fenómenos naturales. Hace unos 400 años la Ciencia se abrió a la aproximación teórica: la formulación de teorías. Se generalizó el uso de modelos y fórmulas. Hace



unas décadas apareció la computación –Ciencia computacional– que permitió abordar la simulación de fenómenos complejos. Actualmente, la aparición de los datos masivos ha llevado a algunos autores a hablar de "la muerte del método científico" (2). Aunque esta afirmación resulte exagerada, lo cierto es que "la Ciencia actual está enfocada hacia la exploración del *Big Data*, que representa la unificación de la teoría, la experimentación y la simulación", como señalaba hace más de 10 años el matemático y científico en computación ya desparecido Jim Gray, quien habló por primera vez de la Ciencia de los Datos como el "cuarto paradigma" científico.

BIG DATA Y *CLOUD*, AL SERVICIO DE LA CIENCIA

En múltiples disciplinas –de la Astrofísica a la Medicina; de la Economía a la Biología– y en innumerables proyectos, los equipos de investigadores pueden acceder ahora a instrumentos y herramientas que hasta hace muy poco tiempo –por su tamaño, coste o accesibilidad física– no estaban a su alcance, del mismo modo que tampoco lo estaban, en muchos casos, los objetivos mismos de sus proyectos de estudio.

El *Big Data* y la computación en la Nube no solamente ayudan a mejorar y optimizar los resultados de los trabajos académicos sino que hoy día hacen posible el objeto mismo de muchas nuevas investigaciones.

Analizar esta nueva realidad presente en el campo de la Ciencia fue el objetivo de la primera jornada organizada por la Fundación Ramón Areces en la primavera de 2013; la primera vez en España que expertos de diferentes disciplinas científicas: astrónomos, físicos, médicos, biólogos, compartían sus experiencias y reflexiones en torno a la contribución e influencia de la Nube y el *Big Data* en sus propios campos de investigación

científica. Y también para hablar de los problemas y dificultades a resolver, no hay que olvidar que se trata de sistemas experimentales, en desarrollo, que tienen que hacer frente a enormes desafíos tecnológicos. De hecho, alguna definición del fenómeno *Big Data* que hace referencia precisamente a esta característica de desafío, lo caracteriza como "conjuntos de datos tan grandes que desafían el uso de herramientas de análisis de datos y elaboración tradicionales".

El volumen de los datos generados –junto a la velocidad, la variedad de su naturaleza y la veracidad– es una de las señas de identidad del fenómeno *Big Data*, señas que también identifican los principales retos y problemas que ha de afrontar en su desarrollo. Desde la llegada de Internet, los científicos de computación han trabajado para aumentar considerablemente el poder de procesamiento de las máquinas. "Durante las tres últimas décadas, cada diez años la velocidad de procesamiento de los ordenadores se ha multiplicado por mil. En 30 años, se ha multiplicado por mil millones. Esto ha supuesto que el supercomputador más rápido del mundo hace 12 años quepa hoy día en un chip" recuerda Mateo Valero, director del Barcelona Supercomputing Center (BSC), uno de los centros de computación más prestigiosos del mundo y pionero en las investigaciones de *Big Data*.

Junto a otros factores, a lo que se enfrenta hoy la Ciencia es a las consecuencias de ese aumento exponencial de los órdenes de magnitud.

Para entender de lo que estamos hablando, un ejemplo: el acelerador de partículas LHC (*Large Hadron Collider*) genera 1 PetaByte (1 millón de GigaByte) de datos por segundo. Esta gigantesca cantidad de datos producidos en un solo segundo es similar al volumen de información que ocupan 10.000 millones de fotografías o 13 años de televi-

sión de alta definición. Además, el Colisionador de Hadrones, que forma parte del Centro de Datos del CERN (Organización Europea para la Investigación Nuclear), comparte la información con 170 centros colaboradores de 36 países en todo el mundo que están conectados con el CERN. Esta red de centros pone en funcionamiento centenares de miles de ordenadores que proporcionan los recursos necesarios para almacenar, distribuir y procesar toda la información generada. El poder combinado de esta red en un solo día es el equivalente al trabajo continuado de un ordenador durante más de 600 años.

El acceso a las infraestructuras y servicios de computación y la gestión de datos se han convertido, por tanto, en elementos fundamentales para la investigación científica, especialmente para aquellas disciplinas donde estas facultades son más necesarias y relevantes, como por ejemplo la Astronomía, la Genética, la Ciencia del clima o la Biología molecular. Y son necesarias iniciativas institucionales y colectivas que respondan a esta nueva necesidad y hagan accesible estos nuevos medios a la comunidad científica. En España, la denominada "*e-investigación*" se articula a través de la Red Española de e-Ciencia, creada en 2007 y que ha llegado a movilizar 101 grupos de investigación de 76 instituciones diferentes y más de 1.000 investigadores españoles suscritos (3).

DE LA CIENCIA A LA SOCIEDAD

Pero la Ciencia de los Datos representa también una nueva realidad para la sociedad en su conjunto, en distintos ámbitos y disciplinas. Y un área donde el impacto está siendo especialmente significativo es el mundo económico y empresarial.

Como hemos visto, las tecnologías *Big Data* no solo ayudan a recopilar grandes cantidades de datos, sino que además permiten su almacenamiento, organización y recuperación para aprovechar todo su valor. Y con el objetivo puesto en que su uso permita optimizar la toma de decisiones.

El *Big Data* es al mismo tiempo un reto y una oportunidad tanto para las empresas como para las Administraciones públicas; las primeras, para mejorar su competitividad y adaptarse al nuevo escenario de la economía global y digital, en el que nuevos agentes están revolucionando las diferentes industrias con nuevos modelos de negocio y propuestas de valor; las Administraciones, para mejorar la calidad de los servicios públicos y ahorrar costes. Según el informe "Open Data in Europe", realizado por la Fundación DemosEuropa, el *Big Data* generará 4,4 millones de empleos en todo el mundo en los próximos cinco años. El comercio, la industria, la salud, la información, las comunicaciones, la banca, los seguros y la Administración pública son los sectores donde el aumento de la inversión será más relevante.

En las empresas, el *Big Data* está generando la aparición de un gran número de aplicaciones en diferentes ámbitos de la gestión, como la minería de datos de redes sociales

para explotación en el área del *Marketing*, la inteligencia y procesos de negocio, el comercio electrónico o la detección del fraude.

Internet, las redes sociales y la movilidad son los factores que explican la eclosión de los datos masivos en la sociedad de inicios del siglo XXI, con 5.000 millones de dispositivos de todo tipo conectados a Internet. Una cifra que en 2020 habrá alcanzado los 25.000 millones, debido al impacto de la Internet de las cosas y las ciudades inteligentes (4).

"El 90 por ciento de toda la información disponible actualmente se ha creado en los últimos dos años y el 80 por ciento es información no estructurada, procedente de vídeos, imágenes digitales, correos electrónicos, comentarios en las redes sociales y otros textos", señalaba un informe de IBM de 2013, que pone de relevancia otra de las características fundamentales del *Big Data*: la variedad de la naturaleza de los datos.

No resulta exagerado afirmar, por tanto, que el *Big Data* está cambiando nuestro mundo.

NOTAS

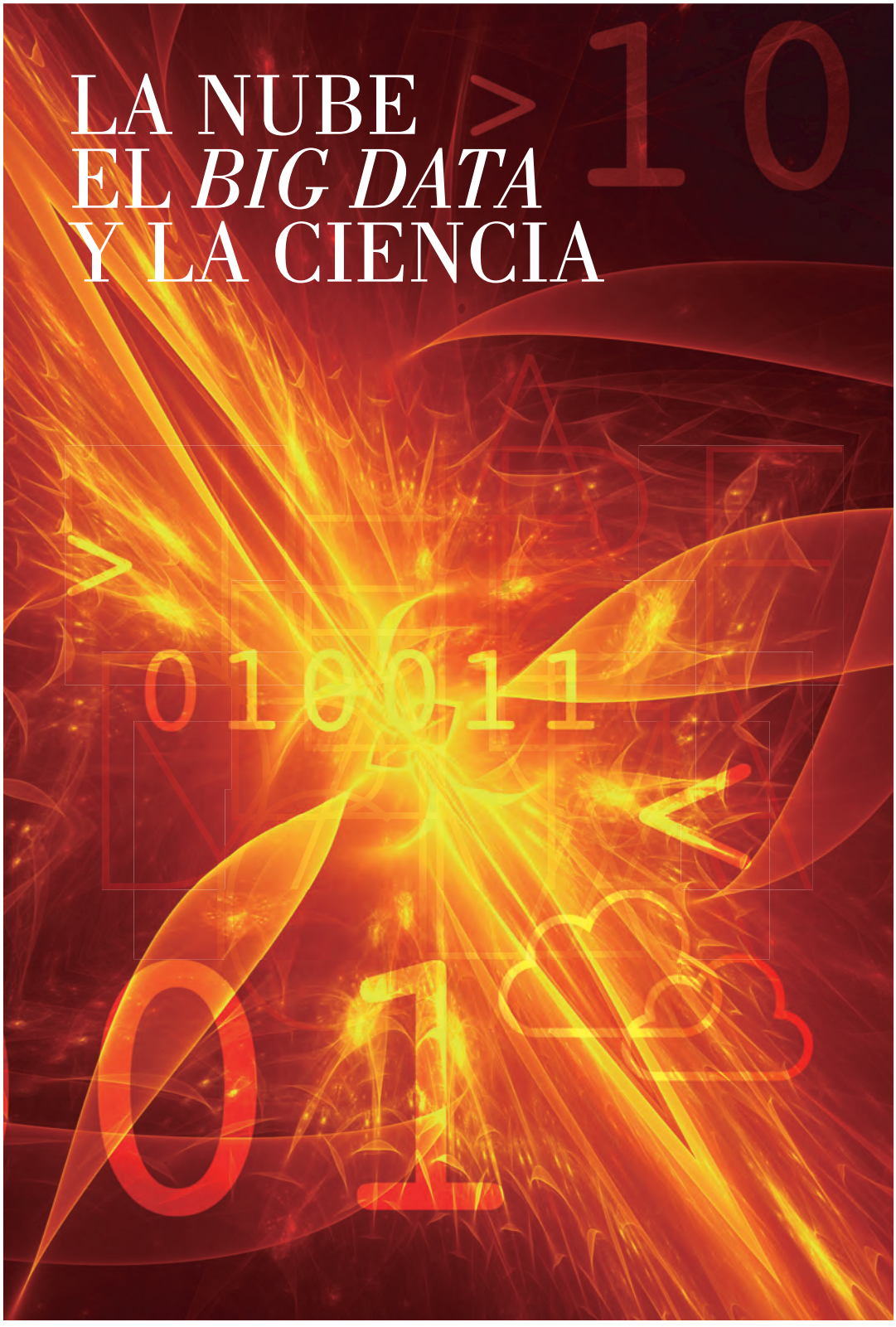
(1) Michael Cox y David Ellsworth, científicos del Centro de Investigación Ames de la NASA, publican un artículo en el que, por primera vez, se hace referencia al problema del *Big Data*: "*Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of Big Data.*"

(2) Chris Anderson. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete". Revista Wired. Julio 2008.

(3) Ignacio Blanquer. "Un balance de la e-Investigación en España". Conferencia dentro de la Jornada "El impacto de la Nube y el *Big Data* en la Ciencia". Fundación Ramón Areces. Marzo 2013.

(4) Gartner Symposium/ITxpo 2014 Barcelona.

LA NUBE EL *BIG DATA* Y LA CIENCIA



LA NUBE,

EL *BIG DATA*

Y LA CIENCIA

INTRODUCCIÓN GENERAL

La Fundación Ramón Areces dedicó en marzo de 2013 una jornada a analizar el impacto de la Nube y el Big Data *y sus beneficios para el mundo de la Ciencia y de la investigación científica. Por primera vez en España, una jornada reunió a científicos españoles y europeos para explicar y analizar los fundamentos de estas dos disciplinas y cómo pueden contribuir a la innovación y al impulso de la investigación científica. Para la Ciencia, los servicios de computación en la "Nube" y el* Big Data *–fenómeno asociado a la gestión de gigantescos volúmenes de datos, cuyo tratamiento no puede realizarse con las herramientas y analíticas convencionales– representan una oportunidad de impulso a la investigación, principalmente a través del acceso a plataformas de computación y de análisis de datos hasta ahora vedadas a pequeños grupos o proyectos de investigación.*

La jornada se estructuró en dos sesiones. La de la mañana, dirigida a explicar en profundidad la naturaleza y características principales de ambos fenómenos, mientras que la sesión de la tarde se orientaba a presentar algunos de los proyectos de investigación más relevantes en diversas disciplinas –Medicina, Astronomía, Física o Biología– en las que la Nube y el *Big Data* juegan un papel relevante para la consecución de los objetivos científicos definidos.

En la primera intervención, titulada *Cloud Computing* y *Big Data*, *la próxima frontera de la innovación*, el profesor **Jordi Torres** (UPC Barcelona Tech. Barcelona Supercomputing Center) presentó e introdujo el fenómeno de la Computación en la Nube. Para el profesor, del mismo modo que hace siglos se produjo un gran avance de la Ciencia cuando la teoría matemática permitió formalizar la experimentación, la aparición de los computadores representó otro paso fundamental para el avance de la ciencia, gracias a lo cual hoy en día disponemos de potentes supercomputadores que por medio de simulaciones nos permiten crear escenarios caros, peligrosos o incluso imposibles de reproducir en la vida real.

La supercomputación ha representado un destacado avance para la ciencia y el pro-

El profesor Blanquer asegura que España se encuentra excelentemente posicionada en la e-Ciencia a nivel internacional tanto en infraestructuras como en aplicaciones

greso. Y aunque hasta ahora, debido a los costes de crear y mantener las grandes infraestructuras de este tipo, la potencia de la supercomputación no ha estado al alcance de todo el mundo, reduciéndose a un conjunto limitado de grupos de investigación, la llegada de lo que se conoce como *Cloud Computing* ya está permitiendo que muchos otros ámbitos de la Ciencia que hasta ahora no podían beneficiarse de esta tecnología puedan hacerlo. Pero el hecho de que los datos disponibles para poder realizar los cálculos han adquirido dimensiones de gran magnitud –lo que se conoce por *Big Data*–, los sistemas de computación actuales presentan nuevos retos que la propia Ciencia informática ha empezado a abordar.

La presentación repasó las características y funcionalidades de estas nuevas herramientas que son los supercomputadores, como el "MareNostrum" del Barcelona Supercomputing Center, con una capacidad de 48.000 *cores* –es decir que cuenta con la misma capacidad de computación que 48.000 ordenadores personales–, así como las dificultades de gestión de estas potentísimas instalaciones.

Un centro de computación es una instalación destinada a la producción de información, que alberga miles de máquinas en un recinto de enormes dimensiones y de las que existen varias docenas en todo el mundo. Uno de los ejemplos es el *Data center* de Facebook (que ocupa una superficie de 28.000 m^2, similar a 4 campos de fútbol) y que consume 40 megawattios de energía al año.

La idea en la que se basa el *Cloud Computing* es la de eliminar los recursos propios y contar con una conexión que fa-

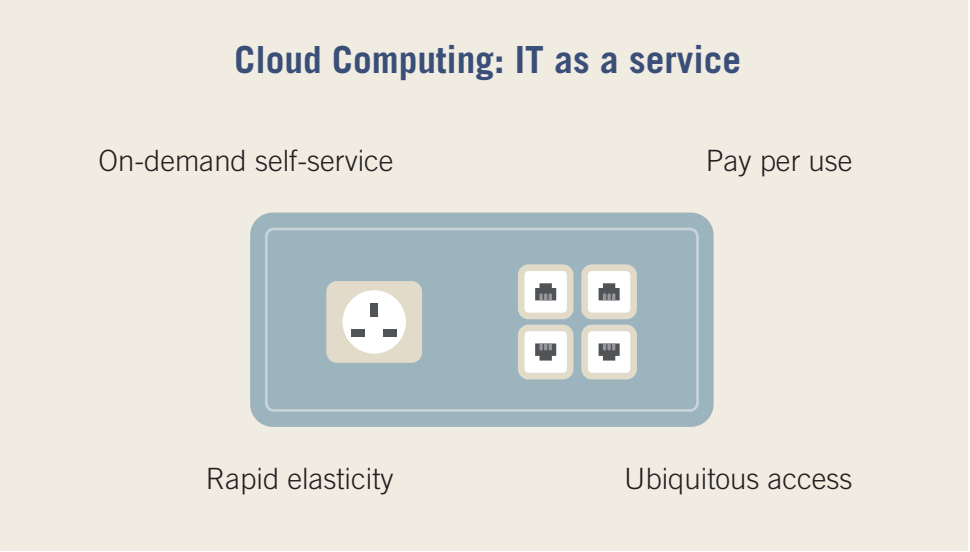


Fig. 1 / Source: http://www.telegrapgh.co.uk/technology/reviews/9241719/Power-Ethernet-Sockets-review.html.

cilite el acceso remoto y virtual a recursos externos de cálculo para el mismo fin. No es un concepto nuevo, ya ocurrió con la electricidad cuando hace más de un siglo las industrias fueron abandonando su producción y se engancharon a la red. Dejaron esa actividad a un agente especializado que, por economía de escala, podía prestar el servicio a un precio más barato. Y es la misma filosofía que rige para la Computación en la Nube. Estos grandes centros de computación, por economía de escala y por la complejidad intrínseca de su gestión, generan el mismo producto: "mi computación y mi almacenamiento de datos" más barato.

La tecnología –la informática– pasa a ser un servicio que se paga por uso, como la electricidad. (Figura 1) "Si yo dimensiono mi centro en casa, si no lo he hecho bien, puedo estar gastando innecesariamente por unos recursos que normalmente no utilizo o puedo haberme quedado corto en el diseño y no puedo ofrecer el servicio porque no tengo suficiente capacidad", explicaba el profesor Torres.

De las diferentes modalidades que existen en la Nube, la infraestructura como servicio es la que más se ajusta a la realidad de los centros de computación. La gran baza para los equipos de investigación es el precio del servicio: 10 céntimos de euro por hora de cálculo. Es una oportunidad que tienen ante sí los grupos de investigación y las empresas en general.

Por otra parte, el volumen de generación de datos ha crecido enormemente –el CERN produce 1 petabyte de información cada segundo (1 petabyte son 1 millón de gigas)– y el *Big Data* se ha convertido en un gran reto. No solo por el volumen y porque los datos exceden los sistemas de almacenamiento de que disponemos ahora, que hacen imposible almacenar todos los datos producidos, sino también por la velocidad en la que estos se generan. Realidades actuales como el Internet de las cosas o las *Smart Cities* (ciudades inteligentes) llevan en paralelo un proceso de sensorización masiva de dispositivos y el envío constante de información que han modificado también el concepto de las bases de datos porque el modelo tradicional de bases de datos estruc-

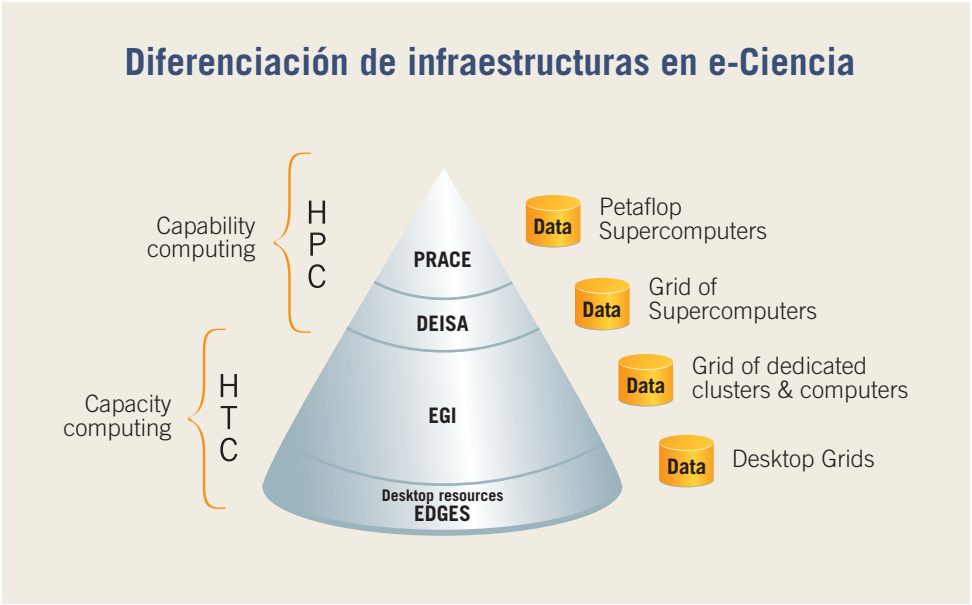


Fig. 2 / Las e-infraestructuras. Conceptos asociados.

turadas no da respuesta a las necesidades actuales del *Big Data*.

Pero el reto más importante al que responder es la forma de analizar los datos. Se puede disponer de muchos datos pero de escasa utilidad; la información es relevante pero lo importante es el conocimiento "accionable", algo que permita tomar una acción. Y no es un hecho trivial, porque los sistemas actuales no son válidos para millones de datos, del mismo modo que extraer valor de esos datos tampoco lo es. Hoy se aplican algoritmos de minería de datos para extraer la información, pero la magnitud del problema es enorme. Queda aún mucho recorrido y la actual coyuntura de crisis y recortes implican más dificultades para la solución de algunos de estos retos.

Un balance de la e-investigación en España

La segunda presentación de la jornada corrió a cargo de **Ignacio Blanquer**, Departamento de Sistemas Informáticos. Universidad Politécnica de Valencia, quien realizó *Un balance de la e-investigación en España*. Tomando la definición acuñada en 1999 por John Taylor, la e-Ciencia o e-investigación se traduce como la actividad científica mejorada (*enhanced-science*) mediante el uso de infraestructuras que integran recursos avanzados sobre Internet.

Blanquer aportó diversas visiones complementarias del concepto de e-Ciencia, deteniéndose en la que hace referencia a la evolución de la Ciencia desde su nacimiento, en el milenio pasado, como disciplina empírica dedicada a la descripción de fenómenos naturales; su etapa posterior, de aproximación teórica, con el uso de modelos, fórmulas y generalizaciones; la aparición en las últimas décadas de la rama computacional, destinada a la simulación de fenómenos complejos y para acabar, en la actualidad, centrada en la exploración de los datos, con la unificación de teoría, experimentación y simulación, gracias a la captura masiva de datos mediante instrumentos o generada mediante simulación y procesada por computador. La e-Ciencia se convierte así en una nueva visión de la Ciencia, fundamentada en una

colaboración global en áreas de la Ciencia y las infraestructuras que la dan soporte.

Bajo este concepto nacen programas de e-Ciencia como el NeSC en el Reino Unido, Open Science Grid en EE.UU., NAREGI en Japón y en España la Iniciativa Nacional de Grid (ES-NGI), la Red Española de Supercomputación (RES), los centros autonómicos de supercomputación y la Red Española de e-Ciencia.

La base para la e-Ciencia es la e-infraestructura (Figura 2). De acuerdo con la definición del grupo de trabajo de e-infraestructuras de la Comisión Europea, estas son el entorno de investigación en el que los investigadores tienen acceso compartido a una serie de recursos únicos o distribuidos, que incluyen datos, computación, almacenamiento, instrumentos… Si bien la definición de e-infrastructuras engloba un conjunto mayor de recursos (como instrumentación avanzada o bases de datos) accesibles de forma ubicua, se asocia principalmente este término a las infraestructuras informáticas que integran recursos de cómputo y almacenamiento de datos accesibles desde Internet.

Los principales conceptos asociados a las e-infraestructuras los integran el *middleware*, entendido como el conjunto de aplicaciones y servicios que permiten utilizar de forma coordinada y eficiente las e-infraestructuras (gestiona aspectos tales como el acceso, los permisos, el estado de los recursos, la distribución de la carga entre los diferentes sistemas, la indexación de los datos); los *sciencegateways* (pasarelas científicas), que facilitan el acceso y uso de las e-infraestructuras automatizando procesos y ofreciendo interfaces amigables. Y por encima de estos dos conceptos se encuentran los usuarios. La forma en que los investigadores se organizan en la e-Ciencia es también importante porque favorece la colaboración y al mismo tiempo permite administrar de forma efectiva los recursos. Cuando se trata de cientos de miles de *cores* distribuidos en cientos de instituciones, la forma de establecer rápidamente una política de acceso no es trivial. De aquí nace el concepto de *organización virtual*, entendida como una asociación encargada de gestionar el acceso a los recursos a todo el conjunto de usuarios de diferentes organizaciones reales que colaboran en ella. Así, en vez de gestionar el acceso de miles de usuarios personales se ordena la participación de unas pocas organizaciones.

La Comisión Europea en sus últimas directrices y estudios ha definido una serie de principios para la mejora de la investigación en Europa. Europa quiere ser el líder mundial de la e-Ciencia y para ello apuesta por el desarrollo de las infraestructuras. Y no solamente en el ámbito de la investigación, sino también para la innovación y el empleo.

Entre las principales iniciativas en marcha destaca la Iniciativa Europea de Grid (EGI). Se trata de un proyecto en el que participan 332 organizaciones de 58 países, orientada principalmente a la resolución de grandes problemas en las áreas de Física de altas energías, Biocomputación, Geofísica y Astrofísica. Cuenta con 20.000 usuarios que han desarrollado más de 1,7 millones

Según Martín Llorente, la computación cloud *mejorará la competitividad y productividad, reduciendo y eliminando barreras de entrada en determinados campos de investigación y generando nuevas líneas de investigación científica*

de trabajos, gracias a la disponibilidad de unos 320.000 *cores* con una capacidad de 152 PBytes. España participa a través de la Iniciativa nacional de Grid.

El proyecto PRACE (Partnership for Advanced Computing in Europe) es una iniciativa dirigida a fortalecer el uso de infraestructuras de supercomputación para permitir alcanzar un impacto importante en la investigación básica y aplicada. Participan 25 países miembros que aportan conjuntamente una potencia sostenida de más de 11,5 Pflops y casi 1 millón de *cores*.

En el ámbito de las redes, GÉANT es la red de investigación y educación paneuropea que conecta las redes nacionales europeas de investigación y educación (NRENs). Une más de 40 millones de investigadores y estudiantes en Europa y permite el acceso de banda ancha a diferentes recursos singulares para Física de altas energías, Radio Astronomía, Biomedicina, Cambio climático, Observación de la Tierra, Arte, etc.

Respecto del *Cloud Computing*, el profesor Blanquer destacó dos ejemplos de "nubes científicas" europeas: Helix Nebula y Venus C. El primero de ellos, enfocado a la "gran Ciencia" y en fase de definición, se estructura como un consorcio entre grandes actores científicos y grandes empresas para avanzar en la provisión sostenible de recursos de computación en la Nube, con tres casos de uso: Física de altas energías, Genómica y Observación de la Tierra. Venus Ces, por su parte, una experiencia piloto en el desarrollo de un *stock* de componentes para aplicaciones científicas en infraestructuras *cloud* públicas o privadas.

De esta forma, Europa defiende un ecosistema de múltiples soluciones de infraestructuras, formadas por *grids*, supercomputadores, nubes y computación voluntaria. No hay una única solución sino un conjunto de soluciones que buscan la interoperabilidad. Y en cuanto al futuro, en el horizonte de 2020, el plan apuesta por los datos y por la colaboración con la industria, con especial énfasis en servicios innovadores, "*digital curation*", acceso abierto, interoperabilidad y un mayor enfoque centrado en el usuario.

El profesor Blanquer destacó que España se encuentra excelentemente posicionada en la e-Ciencia a nivel internacional tanto en infraestructuras como en aplicaciones y señaló que el acceso a las infraestructuras internacionales es una oportunidad para el desarrollo y la innovación, especialmente en la situación actual de crisis económica. Repasó las diferentes iniciativas y proyectos vinculados con las infraestructuras *grid*, la Red Española de Supercomputación (RES) y los centros autonómicos, la participación española en proyectos de e-infraestructuras y la Red Española de e-Ciencia.

La iniciativa ES-NGI, en estrecha colaboración con la iniciativa Portuguesa (INGRID), engloba 17.690 *cores* de 28 centros y más de 100 usuarios, habiendo proporcionado en el año 2012, 175.000 millones de horas de CPU normalizadas a la comunidad científica española. Igualmente, la Red Española de Supercomputación (RES), liderada por el Barcelona Supercomputing Center, ha proporcionado en 2012 aproximadamente 90 millones de horas de cálculo a más de 200 grupos científicos, al igual que los centros autonómicos de supercomputación, entre los que destaca el gallego CESGA. La RES ha renovado recientemente sus recursos en la mayor parte de sus centros, realizando también una recolocación de los recursos del MareNostrum II entre varios de sus centros.

Tanto ES-NGI como la RES tienen una importantísima proyección internacional, con una destacada participación en la iniciativa de Grid Europea (participando tanto en el PMB como en el Council) y en PRACE (siendo uno de los cuatro 'hostingmembers').

La visión en las e-infraestructuras se completa con la participación española en proyectos e iniciativas destacadas como el *Large Hadron Collider* (LHC) *Computing Grid* o MAGIC. La participación de entidades españolas en proyectos europeos del séptimo programa marco en el ámbito de infraestructuras de investigación es destacada, con presencia en 19 proyectos y un presupuesto total superior a los 250 millones de euros.

En este contexto, la Red Española de e-Ciencia se crea en 2007 con el objetivo de dinamizar el diálogo entre los diferentes grupos que participan en este escenario. La Red Española de e-Ciencia llega a movilizar 101 grupos de 76 instituciones y más de 1.000 investigadores españoles suscritos que se organizaron en 2 áreas temáticas: infraestructuras (ES-NGI, RES y Red IRIS) y usuarios científicos. La Red Española de e-Ciencia logró durante sus cuatro años de existencia realizar seis reuniones plenarias, identificar 60 aplicaciones y dinamizar 11 proyectos piloto.

Finalmente, en los últimos años merece especial atención el esfuerzo que se ha dirigido hacia el uso de infraestructuras científicas en la nube, en la que el proyecto VENUS-C ha desarrollado un conjunto de utilidades de plataforma que han permitido la adaptación y despliegue de 27 aplicaciones científicas, 5 de ellas españolas. Estas aplicaciones han demostrado la idoneidad de este tipo de infraestructuras, más adecuadas para grupos pequeños y pymes innovadoras, en el desarrollo de la investigación.

Los beneficios y los riesgos del *Cloud Computing*

La tercera conferencia de la jornada, ¿*Qué ofrece la Nube a la investigación científica*?, fue impartida por **Ignacio Martín Llorente** (Open Nebula Project, C12G Labs. DSA Research Group, Universidad Complutense de Madrid), quien se centró en analizar el papel clave que la computación *cloud* está llamada a jugar en los procesos actuales de investigación científica, mejorando la competitividad y productividad, reduciendo y eliminando barreras de entrada en determinados campos de investigación y generando nuevas líneas de investigación.

El objetivo de la presentación fue el de describir las posibilidades y limitaciones del *Cloud Computing*, así como el impacto potencial de su adopción como plataforma de investigación. Es necesario redefinir los conceptos, los beneficios y los riesgos que aporta la Nube y tener claro que no va a resolver todos los problemas.

Hay un mensaje disruptivo y transformador del *Cloud*. Cuando aparece una tecnología nueva, en una primera fase todo el

El profesor Salvachúa afirma que las recetas que se trasladan del "bussines inteligence" al Big Data *no funcionan porque colapsan por problemas computacionales o de los algoritmos. Las soluciones pasan por diversos enfoques de sistemas distribuidos*

mundo trata de averiguar cómo esa tecnología se adapta a nuestros procesos. Pero la fase más importante es la que da comienzo cuando somos capaces de modificar los procesos para sacar el máximo provecho de las nuevas tecnologías.

El *Cloud* es un modelo de provisión de recursos (aplicaciones, plataformas e infraestructuras) como servicio, bajo demanda, y de forma elástica y dinámica. Y dentro de los diferentes servicios, el de las infraestructuras es el más disruptivo. Además, existe el de plataforma, orientado al desarrollador, y el *software* como servicio. Todos somos usuarios del *software* como servicio. Twitter o Gmail son algunos ejemplos. Es la capa que está más en contacto con el usuario, por debajo de ella está la capa de la plataforma y, debajo de todo, la capa de las infraestructuras. Y la innovación se desarrolla ahí aunque no lo veamos.

El concepto no es nuevo. En los años 60 algunos científicos ya concibieron la computación como un servicio de acceso bajo demanda, que pasaba por convertir la IT en una "commodity". El modelo de provisión del servicio ha evolucionado desde entonces. El modelo inicial fue el de "mainframe", un sistema muy centralizado, caro, difícil de gestionar, al que se accedía por terminales sencillos y con barreras de entrada muy altas. De ahí se evolucionó al modelo cliente/servidor, basado en el uso de ordenadores personales y servidores para computación y almacenamiento distribuidos; un modelo optimizado para obtener la máxima agilidad debido a su bajo coste. Y el tercer estadio de evolución, que representa la Nube, con grandes centros de datos con componentes con capacidad para escalar y donde el coste se determina en función del uso. Como modelo de uso, con la Nube solo nos preocupamos de los resultados y no de su implementación; como modelo de acceso, la aplicación puede usarse desde cualquier dispositivo y lugar; como modelo de infraestructura, la capacidad es elástica y como modelo de costes, solo se paga por el uso realizado, eliminando costes fijos.

Al tratarse de una tecnología disruptiva, el *Cloud Computing* aporta numerosos beneficios: 1) Ahorro de costes, al pagar solo por el uso del *software* y de la infraestructura. 2) Flexibilidad y tiempo de despliegue. Capacidad elástica e instantánea y rápido despliegue del servicio. 3) Comodidad, por la externalización de la configuración y gestión de la infraestructura. 4) Calidad y reproducibilidad, calidad de los resultados de la investigación y reproducibilidad. 5) Eficiencia y productividad. Inversión de tiempo en la investigación y no en la infraestructura. Simplicidad. 6) Aplicaciones actualizadas, colaborativas y accesibles desde clientes ligeros. 7) Colaboración a partir de la compartición de datos y aplicaciones y 8) Acceso asequible a recursos a quienes no tienen sistemas locales, como pymes y países en desarrollo. (Figura 3).

De igual forma, la adopción de un modelo *Cloud* entraña también algunos riesgos, siendo el principal la falta de control por el desconocimiento de la gestión interna del proveedor. Junto a este se encuentra la dependencia del proveedor (*lock-in*), muy importante hoy ya que, a pesar de los esfuerzos de estandarización, realmente sigue siendo difícil migrar de uno a otro proveedor. La disponibilidad o los cortes de servicio, las variaciones de rendimiento o los cuellos de botella en la transmisión de datos representan otro conjunto de riesgos significativos. Por último, los modelos de licencias y la se-

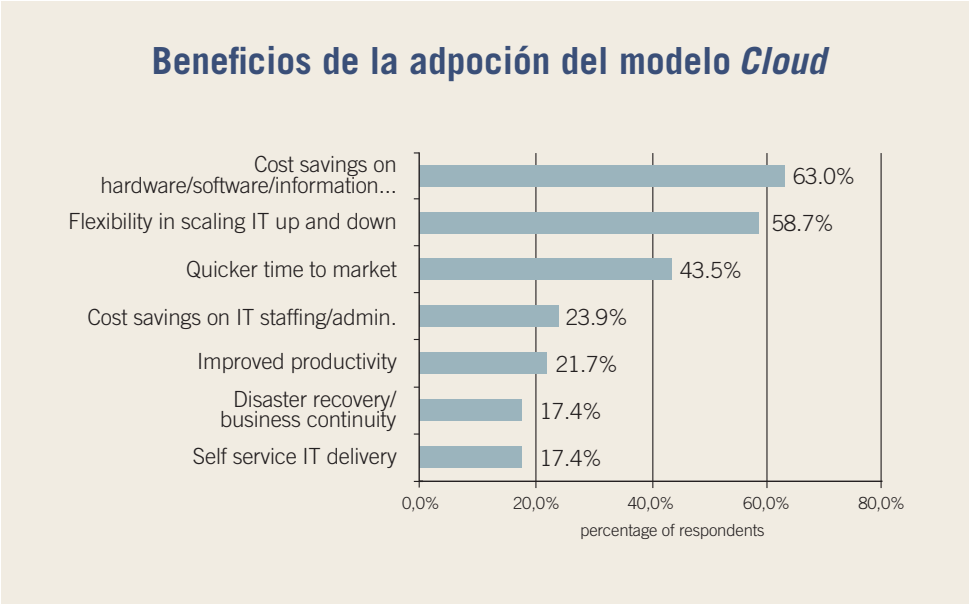


Fig. 3 / ¿Qué ofrece el *Cloud* a la investigación científica? Fuente: Cloud End User Survey, The 451 Group, 2011.

guridad y privacidad de los datos terminan de dibujar el cuadro de riesgos que representa la adopción del modelo *Cloud*.

Existen 4 razones para migrar al modelo *Cloud*: optimizar costes, reforzar la calidad, entrar en líneas de investigación ya existentes, eliminando barreras de entrada, o crear nuevas líneas de investigación. Y tres son las fases en el proceso de adopción de la Nube como soporte a la innovación:

1. Priorización de servicios. Migración gradual de aplicaciones y priorización en función del beneficio y de la afinidad.

2. Orientación a servicios. Cambiar la mentalidad y no pensar en las infraestructuras. El usuario no quiere hablar de *teras*, *cores*, de datos. Quiere hablar de necesidades y resultados. Es muy importante y difícil definir las aplicaciones en términos de calidad de servicio, y que el proveedor de infraestructuras lo traduzca en número de recursos físicos necesarios.

3. Selección del proveedor. Para ello hay que evaluar factores como las medidas de seguridad y protección de datos, la ubicación del centro, las garantías en la continuidad del servicio, compensaciones, servicios alternativos, etc.

Por último, la presentación abordó los diferentes tipos de *Cloud* en función de la propiedad de las infraestructuras: privados, cuando la infraestructura es propiedad de una organización y disponible solo para esa organización; públicos, cuando la infraestructura está disponible para otras organizaciones a través de Internet o redes virtuales; e híbridos, cuando la infraestructura es una composición de dos o más *clouds*.

La última ponencia de la mañana, titulada *Un universo de datos. El fenómeno Big Data y la Ciencia*, corrió a cargo de **Joaquín Salvachúa** (Departamento de Ingeniería de Sistemas Telemáticos [DIT]. Universidad Politécnica de Madrid). Para explicar lo que representa el *Big Data*, el profesor Salvachúa

utilizó como metáfora el movimiento browniano (movimiento aleatorio de las partículas en un medio fluido). Hasta ahora si teníamos partículas de polen flotando en un fluido solo podíamos seguir el movimiento de estas partículas por el fluido. Ahora podemos tener el movimiento de todas las moléculas de agua que están moviendo las partículas de polen. Es el significado de *Big Data*. De repente se ha abierto la puerta no solo de tener cierta cantidad de datos, sino todos los datos, lo que representa un cambio completo del discurso, que conlleva nuevos problemas y dificultades porque nos encontramos en el límite de la tecnología.

Un ejemplo de esta realidad la encontramos hoy en el movimiento browniano social, donde ya podemos disponer de toda la información de lo que hace una persona, de sus movimientos (*Smart City, Smart Car*), e incluso lo que piensa, siente o desea (Facebook, Twitter, etc.). Para el investigador en Ciencias Sociales se abre un mundo fascinante donde todo se convierte en una gigantesca fuente de datos. Estos datos, que pueden ser analizados casi en tiempo real, son de todo tipo, relevancia y veracidad.Y todos ellos pueden ser almacenados, procesados y guardados. Esta posibilidad ha llevado a algunos a asegurar la muerte del método científico. La revista Wired publicó hace 4 años un número especial sobre la muerte de la Ciencia.

Uno de los grandes problemas con los que se enfrenta la investigación científica es generar conocimiento nuevo. Los grandes descubrimientos llegaron cuando se consiguieron fórmulas analíticas que nos aportaban conocimiento extra que nos permitía predecir comportamientos nuevos. Ahora, con los datos, es posible que la investigación se oriente hacia materias o temas donde dispongamos de datos, generando agujeros en otras áreas de la Ciencia donde no se dé esta circunstancia.

Por la heterogeneidad de aplicaciones o soluciones no hay una única definición de *Big Data*, ya que la naturaleza de los datos es diferente según los casos. Hay varias características que acotan esta heterogeneidad: volumen, variedad en la naturaleza, velocidad en la generación de los datos y veracidad.

En cuanto al volumen, hay que tener en cuenta, en primer lugar, que escalar soluciones aparentemente sencillas puede llevar al colapso. Para ilustrar esta realidad, el profesor Salvachúa utilizó el ejemplo de una hipotética hormiga de cuatro metros de altura, que no sería viable simplemente aumentando la escala porque implicaría que su exoesqueleto tuviera un grosor imposible. Por esta razón, para animales de ese tamaño las respuestas son diferentes y se requiere el esqueleto de un elefante o de un dinosaurio de una naturaleza y estructura completamente diferentes a la de la hormiga. De análoga forma, las recetas que se trasladan del "bussines inteligence" al *Big Data* no funcionan porque colapsan por problemas computacionales o de los algoritmos. De esta forma, para resolver los problemas de almacenamiento y procesado de grandes cantidades de datos, las soluciones pasan por diversos enfoques de sistemas distribuidos.

Así, repasó las diferentes alternativas, recordando los primeros sistemas implementados por Google y que ahora han sido imitados por Hadoop (HDFS), basados en el diseño de una gigantesca base de datos no estructurada; los sistemas P2P, autorregulados y autoconfigurados, basados en DHT (*Distributed Hash Tables*) y sobre bases de datos NoSQL; o el sistema MapReduce, un sistema de procesado distribuido. La idea clave es la flexibilidad, dado que la computación puede ser llevada a cabo por un número variable de ordenadores.

Sobre la variedad de los datos, destacó el hecho de tratar con datos no estructurados, ni agregados implica la necesidad de "cocinarlos" previamente. La variedad también se da en la multiplicidad de fuentes no disjuntas e implica la búsqueda de soluciones "artesanales" adecuadas a cada proyecto.

La velocidad representa un problema a

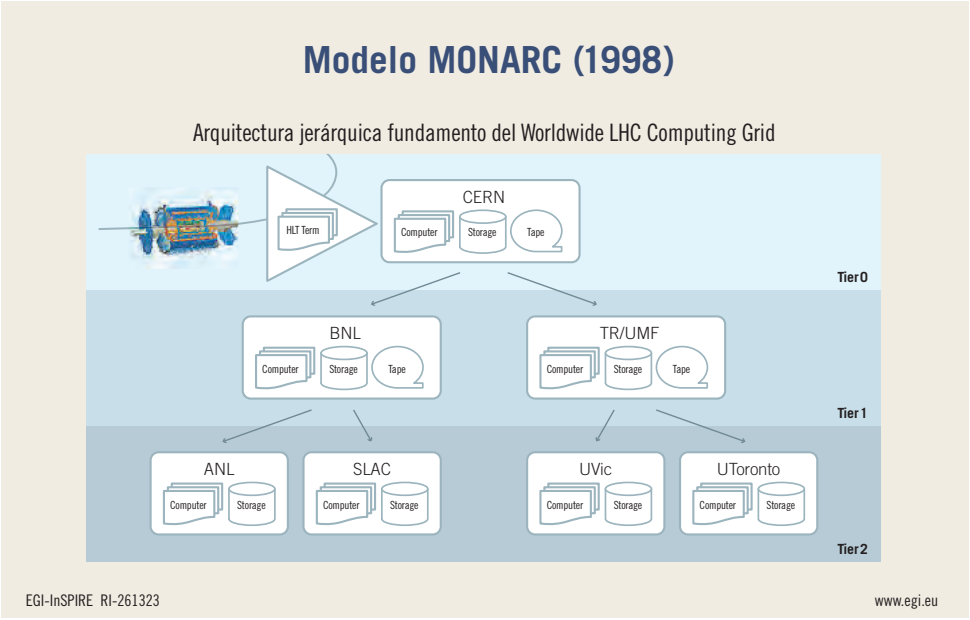


Fig. 4 /

la hora de transportar datos de un sensor a un almacenamiento o entre distintos almacenamiento. Existe la necesidad de procesarlos rápidamente, para lo cual se utilizan esquemas similares a los que se ofrecen en multimedia con el uso de GPUs y el procesado en *streamming*.

La visualización es un componente vital de todo análisis. Representa la parte artística del análisis y es un nicho de gran futuro, que requiere de profesionales con perfiles diferentes a los que desarrollan la investigación. También destacó que, tras la irrupción de fenómenos como el Internet de las cosas y el Social Media, que generan un enorme volumen de datos aportados por una abundancia de sensores, al igual que ocurre con las *Smart Cities*, el escenario se ha complicado. Ello implica problemas de almacenamiento que se han de resolver con el procesado distribuido.

Como un problema práctico en el ámbito de la privacidad de los datos, el profesor Salvachúa señaló el de sintetizar atributos no deseados a los que debemos aplicar la ley de protección de datos o la dificultad de la anonimización de los datos.

Como ejemplos de esta nueva realidad basada en la fuerza de los datos, citó el caso de "House of cards", la primera serie de televisión en la que los guionistas tienen información sobre cómo vemos la serie (duración, horario, paradas), de tal forma que pueden escribir los guiones de los siguientes capítulos valorando el comportamiento observado de los usuarios, cerrando así el bucle. Actuamos sobre un sistema en función de los datos que nos aporta el propio sistema.

Para terminar, hizo un repaso de los lenguajes disponibles –R, Julia y NumPy– y mencionó los problemas en el procesado de grafos y la dificultad de acceder a *datasets* interesantes, resaltando que para que la ciencia avance es necesario contar con enfoques abiertos y colaborativos.

La segunda sesión de la jornada se centraba en la descripción de diversas experiencias de "nubes científicas" y su aplicación en

Según Guillermo Antiñolo, el proyecto Medical Genoma Project *pretende estudiar y validar la forma en la que se procesan los datos, resolver los problemas de almacenamiento, decidir cómo se devuelve esta información en una forma clínicamente útil, ver cómo se integra con los actuales sistemas de información clínicos y hacer un análisis de validez analítica y de coste-eficiencia de NGS* vs *el análisis genético convencional*

diferentes disciplinas científicas, como la Física de partículas, la Medicina, la Biología o la Astrofísica, con ejemplos prácticos y reales de proyectos o centros de investigación que han encontrado en la Nube y el *Big Data* respuesta a las necesidades de computación y tratamiento de datos que requieren sus investigaciones.

Fernando Barreiro, responsable del proyecto Grid y recursos *Cloud* en el CERN, trató de *La iniciativa Helix Nebula y el impacto del Cloud Computing en los experimentos del LHC*. Empezó explicando que el CERN, donde está ubicado el LCH, es el laboratorio de Física más grande del mundo, en el que participan 20 estados y más de 10.000 usuarios en todo el mundo.

El objetivo último del CERN es entender el origen del universo y el *Big Bang*, ocurrido hace miles de millones de años y que comenzó cuando toda la materia estaba concentrada en un solo punto. Para comprender cómo el universo ha evolucionado desde ese primer momento hasta lo que es hoy, en el CERN se construyó el Gran Colisionador de Hadrones (*Large Hadron Collider*, LHC), que es el aparato científico más grande del mundo. Es un túnel de 27 km de circunferencia, a 100 metros bajo tierra, entre Suiza y Francia, en el que hay cuatro puntos donde colisionan las partículas, haces de protones que vienen de direcciones contrarias. El principal desafío del análisis de datos es su volumen y la necesidad de compartir los datos a través de la colaboración del LHC, ya que hay unos 10.000 físicos en todo el mundo que quieren analizar los datos que genera el LCH y cada experimento es muy voluminoso en términos de información. A día de hoy existen 140 petabytes almacenados.

Para el almacenamiento y procesamiento de datos, los modelos computacionales de los experimentos del LHC se diseñaron en torno al concepto de "grid computing" y, desde el inicio de la toma de datos, este modelo ha demostrado ser muy exitoso. El modelo computacional del LHC es el *Worldwide LHC Computing Grid* (WLCG), (Figura 4) con una carga de computación que representa entre 80.000 y 100.000 trabajos simultáneos. El funcionamiento es el siguiente: una colección de "granjas" elige las colisiones que recoge el detector ATLAS. Estos eventos significativos pasan al CERN y se almacenan, los datos se distribuyen a diversos centros de datos, con una disponibilidad cercana al 100%, en discos y cintas. Además, hay otros centros de computación más pequeños que desarrollan su labor en ámbitos más locales.

Modelo de integración básico entre *grid* y *cloud. The grid of clouds*

Los nuevos paradigmas de la informática, como son la virtualización y la computación en la Nube (*cloud computing*), ofrecen características atractivas para mejorar las operaciones y la elasticidad de la computación científica distribuida. Si bien no es posible sustituir el *grid* por la Nube, hay maneras de integrar recursos de la Nube en la infraestructura *grid* existente. Un proyecto de colaboración con la industria europea que ha re-

sultado exitoso es *Helix Nebula –the Science Cloud* o Nube Científica– que consiste en un esfuerzo de colaboración de varias organizaciones europeas, entre ellas el CERN, ESA y EMBL, para establecer alianzas público-privadas y la construcción de una infraestructura *cloud* europea capaz de soportar las misiones de estas organizaciones.

Entre las conclusiones de la experiencia en el CERN en torno a la computación y los datos, la computación *grid* y la Nube están vistas como tecnologías complementarias que van a convivir en diferentes niveles de abstracción. En cuanto a la simulación y procesado de datos, el modelo para ejecutar los trabajos en nubes externas es útil, pudiendo mejorar la automatización y monitorización, pero las necesidades actuales están cubiertas.

En lo referido a las cuestiones pendientes, Barreiro destacó la poca experiencia en el almacenamiento de datos en la Nube, la falta de adopción de estándares, tanto en las interfaces como en los servicios ofrecidos por los diferentes proveedores, así como la identificación de modelos de negocio para la colaboración con proveedores europeos.

La segunda presentación de la sesión dedicada a experiencias científicas en torno a la Nube y el *Big Data* correspondió a **Marco Aldinucci** (Computer Science Department. Universidad de Turín) y la ponencia titulada *Transformando el Big Data en conocimiento: gotas de sistemas biológicos en la Nube*.

La presentación se centró básicamente en explicar los usos que se hacen del *Cloud* en el dominio de los sistemas biológicos y en la biología en general, así como las facilidades que aporta y las limitaciones que tiene en diferentes órdenes. La Biología requiere de la producción de aplicaciones que sean eficientes en la Nube y que sean útiles para extraer datos, pero usando las técnicas utilizadas en Biología. En este sentido, la modelización es un referente. Y uno de los ejemplos más característicos es el de la modelación estocástica.

El modelado estocástico –sistemas regidos por la aleatoriedad– de los sistemas biológicos, unido a los modelos de simulación Monte Carlo, es una técnica cada vez más popular en Bioinformática. Para ser efectiva, las simulaciones estocásticas deben ser soportadas por herramientas poderosas de análisis estadístico. El flujo de procesos de análisis-simulación puede resultar costoso computacionalmente al reducir la interactividad necesaria en el ajuste del modelo.

Para hacer frente a estos desafíos, se aboga por el diseño de *software* de alto nivel para la construcción de simuladores paralelos eficientes y portátiles para la Nube. En particular, el grupo ha desarrollado el simulador de sistemas biológicos *Calculus of Wrapped Components* (CWC), que se diseña según el enfoque basado en el patrón de *FastFlow*, un *software* desarrollado por el grupo. Gracias al marco de *FastFlow*, el simulador CWC está diseñado como un flujo de trabajo de alto nivel que puede simular modelos, combinar los resultados de la simulación y analizarlos estadísticamente en un único flujo de procesos en paralelo en la Nube. Para mejorar la interactividad, las fases se implementan sucesivamente de tal manera que comienzan a generarse resultados del análisis inmediatamente después de arrancar la simulación así como realizar distintos análisis simultáneamente.

Guillermo Antiñolo, director científico del *Medical Genoma Project* presentó el *Medical Genoma Project*, con la biomedicina como protagonista. El proyecto *Medical Genome Project* (MGP) es un proyecto singular donde los principales objetivos son el descubrimiento de nuevos genes responsables de enfermedades de base genética y la caracterización de la variabilidad genética de individuos sanos fenotipados, mediante la secuenciación del genoma humano, usando las nuevas tecnologías de NGS. (NGS de sus siglas inglesas *next generation sequencing*).

El proyecto de secuenciación del geno-

Carlos Allende explica que el análisis de las observaciones ha sufrido una revolución gracias a los progresos en computación. Los simples modelos analíticos son reemplazados por sofisticadas simulaciones numéricas. Las estrellas, que solían ser bolas con simetría esférica, pasan a ser objetos cuadri-dimensionales con planetas en órbita, y las galaxias se transforman de estructuras axisimétricas aisladas en amasijos irregulares de gas, estrellas y materia oscura que interaccionan y evolucionan a la vez que el universo se expande

ma humano, que arranca en la década de los años 80, está basado en el concepto de "pensar genéticamente para actuar en la Medicina localmente". El descubrimiento de la reacción en cadena de la polimerasa (PCR), una técnica para amplificar fragmentos de ADN, hizo posible alcanzar los objetivos del proyecto Genoma Humano y modificar la forma de aproximación a las secuencias de ADN. Estos avances permitieron que en 2005 se publicara la primera secuenciación del genoma humano. Un cambio de paradigma y un cambio de actitud, se pasaba de analizar un gen concreto a analizar un exoma, una proporción de 1 a 3.000.

Este cambio de paradigma representa también un desafío en cuanto al volumen de información y datos a analizar, donde la tecnología NGS no es más que el inicio; el almacenamiento y la gestión del análisis de la información es el verdadero problema. El cambio del volumen de datos a gestionar es muy importante, multiplicándose por varios órdenes de magnitud (x 1.000). El análisis y procesamiento de datos tienen el riesgo de convertirse en un "cuello de botella" conforme vaya incrementándose el volumen de datos disponible en los procesos de secuenciación. En este nuevo escenario, las soluciones tradicionales de computación y bases de datos no son suficientes y se han de implementar nuevas respuestas de la mano del *Big Data* y de la computación *Cloud*.

En el corto periodo de tiempo desde 2005, NGS ha modificado la investigación genómica y ha permitido a los investigadores llevar a cabo experimentos a nivel del genoma completo que anteriormente no eran viables o asequibles. De esta manera NGS se ha empezado a aplicar ya con un gran éxito en el descubrimiento de genes de enfermedades mendelianas y en cáncer, y es la herramienta ideal para hacer realidad las promesas de la Medicina personalizada. Las tecnologías que constituyen este nuevo paradigma continúan evolucionando de forma muy rápida, de forma que las mejoras previsibles en la robustez tecnológica y el aumento de la eficiencia de los procesos allanarán el camino de la traslación del conocimiento generado al diagnóstico clínico. Empezamos a tener herramientas para que esta tecnología NGS nos ayude a determinar la existencia de enfermedades.

El proyecto *Medical Genoma Project* pretende estudiar y validar la forma en la que se procesan los datos, resolver los problemas de almacenamiento, decidir cómo se devuelve esta información en una forma clínicamente útil, ver como se integra con los actuales sistemas de información clínicos, y hacer un análisis de validez analítica y de coste-eficiencia de NGS *vs* el análisis genético convencional.

De esta forma, para llevar a cabo el proyecto MGP se utiliza un entorno de computación de alta capacidad junto con una infraestructura de almacenamiento distribuido para poder procesar el gran volumen de datos que se generan (figura 5). Además del procesamiento de datos, la interpretación

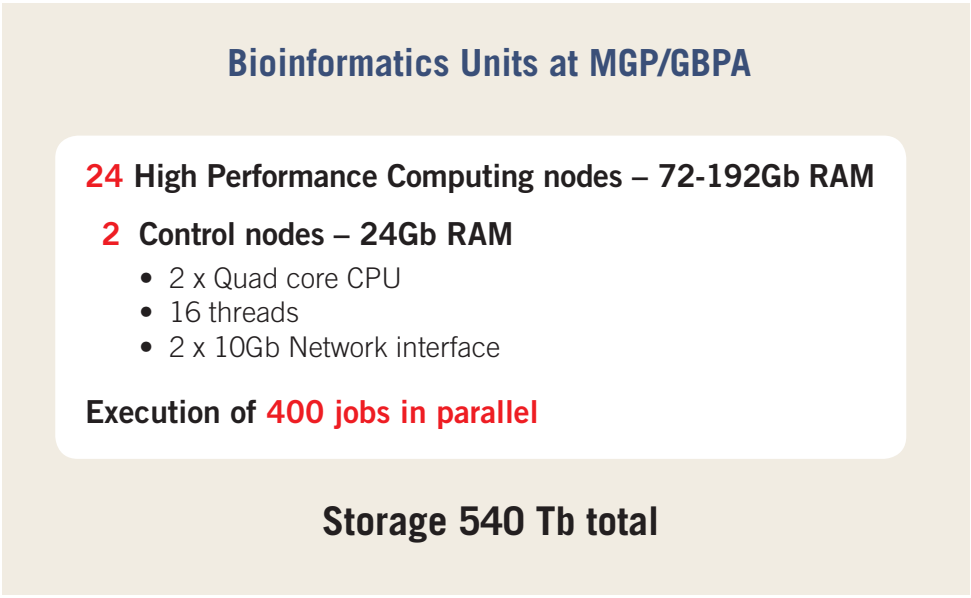


Fig. 5 /

de los resultados de secuenciación requiere de grandes bases de datos que alberguen una completa caracterización de las variaciones nucleotídicas presente en los genomas. Todo ello pone en un primer plano la relevancia de las infraestructuras de computación y almacenamiento a la hora de poder manejar, procesar y transformar en información útil el gran volumen de datos producido por las nuevas tecnologías de nueva secuenciación. El objetivo es reducir la enorme distancia que aún existe entre los resultados de las investigaciones y su aplicación cotidiana a los enfermos y uno de los caminos es la construcción de una base de datos para la identificación de terapias y medidas preventivas que permita obtener información para el correcto diagnóstico de enfermedades, como en el caso de la distrofia hereditaria de retina, donde los resultados del proyecto han permitido encontrar, después de seis meses de trabajo, las mutaciones responsables causantes de la enfermedad en seis de las siete familias de genes estudiados.

La última presentación de la jornada, *Las oportunidades del Big Data en la Astronomía moderna*, fue impartida por **Carlos Allende** (Instituto de Astrofísica de Canarias). Los métodos tradicionales de análisis en la Astronomía observacional han cambiado en la última década. Se ha pasado de proyectos relativamente modestos, realizados por una sola persona o un pequeño grupo de investigación y con unas pocas noches de observación seguidas por una reducción de datos y un análisis artesanales, a proyectos mucho más ambiciosos, que utilizan instrumentos altamente optimizados en operación continua y durante años. En Astronomía solo se puede observar y tratar de entender lo que ocurre. La mayor parte de la información que disponemos proviene de la luz, de la energía fotoeléctrica, que se observa en diferentes puntos del universo. El 99,9 de los estudios en Astronomía responden a esta realidad, a través de dos tipos de observaciones: imágenes y espectros.

El astrónomo va directamente a la montaña; apunta el telescopio (detector) y obtiene datos (fotografías). Después viene la fase

de extracción de la información, analizando la densidad en las placas fotográficas para ver la cantidad de luz en cada punto en función del ángulo. A continuación vienen la fase de mapeo o calibración, la reducción y el análisis de datos.

Hasta hace unos años, todo el proceso era bastante artesanal, cada proyecto de observación contaba con su propio *software* adaptado a sus necesidades particulares; este modelo implicaba varios problemas: lentitud en los procesos, repetición de esfuerzos y falta de homogeneidad en la operación de los instrumentos, la calibración, etc. El modelo alternativo, que ya está en marcha en algunos proyectos, consiste en la utilización de instrumentación y *sotfware* no genérico, que pueda ser útil para diferentes proyectos con el objetivo de mejorar el rendimiento y garantizar la homogeneidad de resultados en diferentes estudios.

El análisis de las observaciones ha sufrido una revolución gracias a los progresos en computación. Los simples modelos analíticos son reemplazados por sofisticadas simulaciones numéricas. Las estrellas, que solían ser bolas con simetría esférica, pasan a ser objetos cuadri-dimensionales con planetas en órbita, y las galaxias se transforman de estructuras axisimétricas aisladas en amasijos irregulares de gas, estrellas y materia oscura que interaccionan y evolucionan a la vez que el universo se expande.

El recorrido por los más ambiciosos programas proyectados o en marcha de la Astronomía observacional incluye los siguientes:

Sloan Digital Sky Survey. Lleva más de 10 años funcionando. Se basa en un único telescopio pequeño (2,5 metros de diámetro), diseñado para tener un campo de visión de enorme calidad y conseguir imágenes del cielo. La innovación de la cámara radica en que los dispositivos de carga acoplada (CCD) rotan al igual que el cielo, permitiendo la observación simultánea de cientos de objetos celestes. Con esta cámara se publicó la imagen más grande del mundo, de 26 gigapixeles.

Misión Espacial Gaia. Gaia es una de las principales misiones de la ESA. Después de una década de trabajos, se espera su lanzamiento en octubre de 2013. Es un instrumento, situado en un satélite, para escanear el cielo repetidamente y estudiar las posiciones, las velocidades, los colores de las estrellas de nuestra galaxia. Gaia va a dar información tridimensional para 1.000 millones de estrellas en la Vía Láctea. Uno de los problemas que tiene la misión es conseguir transmitir la información desde el satélite a la Tierra. El ritmo de transmisión es de unos 4 órdenes de magnitud menor que el que genera el propio instrumento, de varios gigabits por segundo. Para resolverlo, se hace una reducción de datos a bordo muy significativa en base a cálculos de supercomputación.

Proyecto APOGEE. Es un proyecto desde tierra y complementario de Gaia dirigido a obtener más información de los espectros –abundancias químicas–. Opera con luz infrarroja y puede llegar a distancias de 20.000 años luz. También resalta su precisión con una resolución espectral mucho mayor. El Instituto de Astrofísica de Canarias está muy involucrado en este proyecto, que tiene como objetivo construir un mapa químico de las galaxias en 3D.

Telescopio Big Boss. Es muy parecido a Sloan, pero mejorado y más grande (4 metros de diámetro). Dispone de un robot que posiciona 5.000 fibras. El único modelo basado en la Nube es el telescopio **Hetdex**, orientado a elaborar un censo completo de lo que hay en el Universo. Dispone de 30.000 fibras ópticas.

Hay un gran esfuerzo en desarrollo para poder tratar con los volúmenes ingentes de información que aportan los nuevos instrumentos de observación e interpretar las observaciones con la Física tradicional, de la mano de la computación del siglo XXI, los algoritmos avanzados, y las redes de alta velocidad.

Cloud Computing y *Big Data*, la próxima frontera de la innovación

Por Jordi Torres

UPC Barcelona Tech. Barcelona Supercomputing Center

Mateo Valero suele exponer una presentación en la que cuenta brevemente cómo comenzó la Ciencia. Empezó en el momento en que la Matemática, la teoría, permitió describir la experiencia. Éste fue un paso fundamental, pero ¿cuál es el siguiente paso fundamental? El siguiente paso es, o ha sido hasta ahora, la simulación. La simulación hecha por la supercomputación nos permite simular y crear escenarios que sin la supercomputación serían imposibles. Escenarios caros, peligrosos e imposibles. Primero fue la teoría, luego fue la simulación que nos ha permitido llegar hasta aquí y que se basa en muchas fórmulas, mucha matemática, y muchos cálculos. ¿Dónde se realizan estos cálculos? En este caso se hacen en Barcelona pero hay una red en España, la Red Española de Supercomputación, en la que los científicos españoles de diferentes áreas

La simulación hecha por la supercomputación permite crear escenarios que sin su intervención serían imposibles. Escenarios caros, peligrosos e imposibles.
Fuente: Prof. Mateo Valero, BSC-CNS 2010.

de investigación, no ingenieros informáticos sino precisamente de otras disciplinas, tienen una herramienta para desarrollar sus experimentos, escenarios caros, peligrosos o imposibles de crear.

Un supercomputador es una máquina de unas dimensiones y de unas características no normales para la mayoría de nosotros. En este caso, el MareNostrum, la máquina que tenemos en Barcelona y el nodo principal de esta red española de siete nodos, tiene 48.000 *cores*. Estas características implican unas dificultades importantes de gestión. Por ejemplo, existe un problema importante de infraestructura para disipar el calor, especialmente en latitudes como las nuestras, porque en Finlandia no tienen tantos problemas de refrigeración, y por tanto no tienen que asumir esos costes. Este supercomputador puede ser utilizado por muchos grupos españoles de investigación. Para ello existe un comité de expertos en diferentes materias que recibe propuestas de proyectos y que ordena y asigna los proyectos. Pero ¿qué pasa con el resto de grupos que no tienen acceso a un supercomputador? Por ejemplo, para ciertas empresas no es fácil entrar en esta red española de supercomputación, tienen que hacerlo a través de grupos de investigación, pero la investigación, por suerte, también se realiza en empresas. Por suerte también, a día de hoy el resto del mundo tiene el *cloud*. Amazon anunciaba hace un año que iba a contar con un supercomputador similar a los que tenemos en la red, que en su momento alcanzó el número 46 de una lista de 500 supercomputadores en el mundo.

La importancia del *Cloud Computing* para la Ciencia

El *Cloud Computing*, o computación en la Nube, es importante para la Ciencia porque, como servicio, ofrece lo que hasta ahora solo podían ofrecer ciertos centros muy especializados con unos costes muy elevados. Crear un centro de supercomputación es muy caro y hasta ahora los recursos tenían financiación pública, pero ya se

sabe cómo está la situación ahora mismo. Con lo cual, el *Cloud Computing* es algo que ya está aquí. WIRED, una revista técnica, publicaba hace un año un artículo titulado: "*Amazon builds world's fastest non existent supercomputer*" (Amazon construye el supercomputador, no existente, más rápido del mundo). Ahora todos tenemos acceso y capacidad para usar un supercomputador. *Cloud Computing* es, en el fondo, un gran número de máquinas en algún lugar del mundo, porque al final la computación y el almacenado sí existen, que se ubican en *data centers* (centros de datos) de los que, a día de hoy, hay decenas en el mundo. El de Amazon ocupa una superficie de 28.000 metros cuadrados, es decir, como cuatro campos de fútbol. El de Microsoft, por ejemplo, ocupa un 40% más, aunque su capacidad aumentó un 60%. La tecnología va avanzando.

Éstas son grandes factorías de información, grandes centrales de producción de información similares a las grandes centrales de producción eléctrica cuya existencia damos por descontado. Y sin embargo, algo similar pasó hace un siglo cuando las empresas dejaron de generar su propia electricidad y se conectaron a la red porque era más barato y les permitía centrarse en su negocio, dejando la producción de electricidad, que ya no era un elemento competitivo, a un profesional que, por economía de escala entre otras cosas, producía el mismo servicio más barato. Ahora está ocurriendo lo mismo en el ámbito de la computación y de los datos. Estos grandes centros de datos, por economía de escala y otros factores, generan el mismo producto, mi computación y mi almacenado, más barato. Así de simple. Y, además, se puede ubicar en Helsinki, donde el sistema de refrigeración es un 44% más eficiente que en Madrid, por ejemplo.

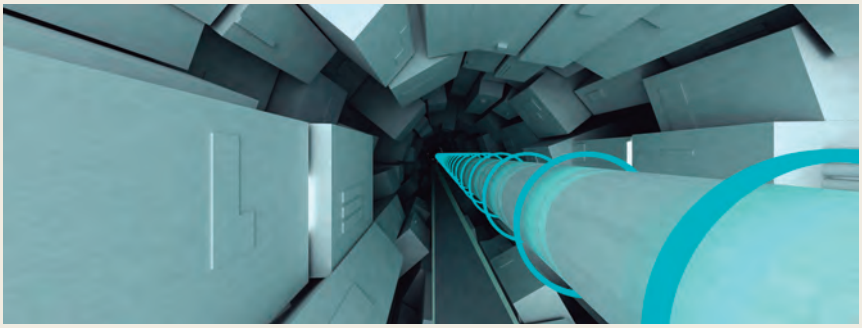
La idea es sencilla. La informática se convierte en un servicio. Un servicio que se paga por uso como la electricidad que pagamos en nuestras casas, donde si gastamos más, pagamos más, y viceversa, y donde puedo

utilizar puntas de energía si las necesito. Se elimina la posibilidad de un gasto innecesario, unas máquinas infrautilizadas o de un servicio insuficiente por no contar con suficientes máquinas. La idea es delegar la infraestructura y las necesidades de un supercomputador en un tercero. Por supuesto, no toda la Ciencia necesita supercomputación. La supercomputación es una parte de la computación que tiene unas características especiales, en lo que a tipo de *hardware* y de almacenado se refiere, que puede realizar un trabajo por partes y en paralelo. Es decir, si una empresa necesita 120 horas de computación para realizar una tarea, puede dividir la tarea por partes y utilizar un supercomputador para hacer todo el trabajo en una hora, porque es como si estuviese utilizando 120 máquinas. Y si lo hace en la Nube no necesita montar 120 máquinas con el coste que eso supone, además de los costes relacionados con el espacio, la refrigeración, la administración, los empleados, etc.

Para mí, éste es el *cloud* de verdad, el auténtico, el que supone la infraestructura como un servicio, una gran base para muchos grupos de investigación que necesitan hacer una simulación en un momento dado y pueden contratar este servicio. Una hora en Amazon cuesta 10 céntimos de euro. Es una oportunidad muy grande que tenemos todos los grupos de investigación y las empresas en general. Es algo muy simple, grandes centros de cálculo que son simples máquinas como las que vemos en casa, agrupadas, gestionadas por alguien y con acceso a través de Internet. Un centro de datos que no está en el piso -2 sino que está al otro lado del mundo y accesible por cables de fibra óptica en los que la latencia, el tiempo de acceso, puede ser pequeñísima si la red es del ancho de banda que requerimos.

En relación a los datos, el supercomputador produce un petabyte de información cada segundo. El *Big Data*, en sí mismo, es un reto muy grande que tenemos todos. Podemos definir el *Big Data* como aquellos problemas en los que los datos exceden los sistemas de almacenamiento que tenemos ahora. De hecho, una de las cuestiones es: ¿tenemos que almacenarlo todo o no? Y ¿cómo procesamos tanta información? La idea es que todo aquello que los sistemas convencionales hasta ahora existentes no soportan se denomina *Big Data*. Si el problema es asumible y tratable ahora, no es *Big Data*. Porque ¿qué es lo que hace que no podamos almacenar, gestionar estas cantidades de datos? No es solo una cuestión de volumen, sino también de la velocidad con la que se generan los datos que salen de sensores. Estamos ya en el mundo de la *Internet of things* (Internet de las cosas). Todo está o empieza a estar sensorizado y es una información de mucho valor que se debe usar, sobre todo en el mundo de la salud, que es uno de los temas más importantes de investigación. Es un *streaming* que va generando información constantemente. ¿Qué se hace con tanta información a esa velocidad? De repente llega el mundo sensorizado de las *Smart Cities*, otra de las áreas de investigación más relevantes, y aparece la posibilidad de añadir a su volumen de información factores como la contaminación, los recorridos de los au-

...”The LHC produces 1PetaByte of data every second, big data and lack of computing resources were becoming the European Organization for Nuclear Research’s biggest IT challenges...”



El “enorme” volumen de los datos es una de las variables que definen el fenómeno *Big Data*. El acelerador de partículas LHC produce 1 PetaByte (1 millón de GigaByte) de datos por segundo.

tobuses, los semáforos, para dar prioridad a los autobuses y que el tráfico sea más fluido. Estos datos no pueden ser procesados por modelos tradicionales de bases de datos estructuradas como hemos hecho hasta ahora.

Y finalmente, el último paso, y quizás el más importante, es cómo cambiamos la manera de analizar estos datos. Aplicamos algoritmos de minería de datos, de aprendizaje, etc., para extraer valor y conocimiento de los datos y muchos sistemas utilizan estos algoritmos para predecir escenarios a partir de los cuales nosotros podamos tomar decisiones. No obstante, estos algoritmos funcionan muy bien para miles de registros, miles de datos, pero no para millones de datos en tiempo real. La mayoría de los datos de la *Internet of things* no pueden ser almacenados, aunque los utilicemos en un momento dado. Bastante trabajo hay ya con los nuevos datos como para dedicar tiempo a los antiguos, con lo que el análisis se vuelve fundamental. En resumen, el mundo científico tiene cuatro retos fundamentales para poder aportar toda esta nueva tecnología que se llama *Big Data* al resto de grupos de investigación: almacenar, gestionar, procesar y analizar los datos. Todavía hay mucho por hacer, a pesar de las expectativas optimistas de mucha gente.

Los retos

Por ejemplo, ¿el almacenamiento de datos es viable económicamente? Claro que sí. Podemos conectarnos a Amazon y contratar dos terabytes por 82€, y esta capacidad de almacenamiento puede ser suficiente para muchas empresas que pueden almacenar el movimiento de una parte importante de su día. Es un gasto asumible, aunque hay que tener en cuenta que actualmente podemos leer discos a una velocidad de 100 Mb/s, por lo que necesitaríamos 5 horas para poder leer dos terabytes. Sin embargo, esto es un problema porque muchas empresas necesitan tomar decisiones empresariales con rapidez. ¿Qué hace Google? En mi opinión, nos ha hecho un flaco favor

La Nube es una oportunidad para acercar la supercomputación a todos aquellos grupos de investigación que hasta ahora no podían contar con ella

porque estamos acostumbrados a ir tecleando y que nos dé sugerencias de búsqueda relacionadas con búsquedas anteriores, ya que Google ha dotado a su buscador de una función de aprendizaje. Pero Google cuenta con 20.000 discos que, en paralelo, leen dos terabytes en un segundo. Lo mismo que se hace en computación, se hace también en almacenamiento. Aunque el primer reto implica cambiar el modelo de procesado. Existen iniciativas como Reduce, Storm o S4, pero el problema no ha sido resuelto todavía. Sobre todo en lo referente al tiempo real, como las decisiones que deben tomarse en una *Smart City* (ciudad inteligente), en donde hay situaciones que requieren encontrar una solución en menos de un segundo.

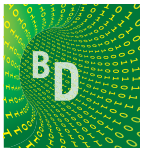
El almacenamiento es otro de los retos. Hasta ahora se utilizaba la RAM para aquello que se utilizaba mucho en nuestros cálculos y el disco para el resto de la información. La memoria es mil veces más rápida que el disco, pero también es cien veces más cara. En la actualidad, técnicamente, tampoco podemos contar con mucha memoria, lo que supone otro problema, y lo que se está desarrollando muy rápidamente es el denominado *storage class memory*. Hoy en día utilizamos en nuestros ordenadores discos sólidos, que son memorias que se han colocado donde antes había un disco. Son más rápidos, aunque más caros, y lo que se está investigando es el state *storage class memory*, que es colocar la memoria en su lugar. Cuando la comunidad científica haya solucionado este problema tendremos una capacidad de memoria equivalente a la capacidad de disco y con el tiempo esto tendrá un precio razonable. Este tipo de memoria es más económica en consumo porque no es un disco mecánico, sino que está compuesta de circuitos y, por tanto, consume menos energía, que es otra cuestión muy importante que debe tenerse en cuenta.

Las bases de datos relacionales que hasta ahora todos conocíamos y nos han explicado en las facultades ya no nos sirven para resolver grandes problemas. Están surgiendo nuevas propuestas de sistemas como los denominados "NO SQL". Podemos tener muchos datos, pero no sirven de nada porque no es información. Pero es que incluso la información no es conocimiento, y lo importante es lo que se denomina conocimiento accionable: algo que nos permite llevar a cabo una acción. Por ejemplo, no sirve que una aplicación nos informe del estado del tráfico en nuestro camino al trabajo porque el tráfico va cambiando en el tiempo que nosotros empleamos en desplazarnos: necesitamos una aplicación que haga una predicción a partir de datos actuales e históricos del tráfico, el tiempo, la hora etc., y que nos vaya indicando en tiempo real el camino para tardar el mínimo tiempo posible en llegar a nuestro destino. Los datos en sí no nos sirven. Necesitamos que generen conocimiento y esto no es trivial porque las técnicas de *machine learning* y de *data mining* sirven para miles de registros, pero no para millones, por el momento. Estamos trabajando en ello, pero actualmente no lo tenemos. Esto mismo es aplicable a otras ciencias y tenemos la suerte de contar con un centro multidisciplinar en el que colaboramos, por ejemplo, con investigadores de Ciencias de la Vida. Valorizar sus datos no es nada banal: tenemos problemas a todos los niveles, de almacenamiento, de gestión, de procesado, etc. La Nube es una oportunidad para acercar la supercomputación a todos aquellos grupos que hasta ahora no podían contar con ella.

Un universo de datos
El fenómeno *Big Data* y la Ciencia

Por Joaquín Salvachúa

Departamento de Ingeniería de Sistemas Telemáticos.
Universidad Politécnica de Madrid



Como ingeniero, mi enfoque es menos creativo y más orientado a que las cosas funcionen, aunque no siempre desde un punto de vista totalmente científico de este tipo de cosas. ¿Cómo se enfrenta uno a un proyecto de *Big Data* para intentar generar algún tipo de conocimiento nuevo o de aplicaciones o servicios nuevos a partir de los datos que se tienen? Los cambios en el mundo del *Big Data* pueden compararse al *movimiento browniano*, en el cual tenemos unas partículas, por ejemplo de polen, flotando en un fluido y generando una serie de movimientos. Hasta hace cierto tiempo, únicamente podíamos seguir el movimiento de estas partículas de polen flotantes, teníamos el resultado de una serie de movimientos característicos. Ahora es como si pudiésemos tener los datos de todas las moléculas de agua que están moviendo cada una de esas partículas. De repente se ha abierto una puerta, no solo a tener una cierta cantidad

Una de las características típicas que se muestran del Big Data *es el volumen. Son sistemas con un gigantesco volumen de datos que no están perfectamente controlados, que se generan a una gran velocidad, en muchos casos a mayor velocidad de la que somos capaces de procesar*

de datos, sino también a tener "todos" los datos posibles que están proporcionando algunos de los sistemas de medición. Evidentemente existen algunas dificultades teóricas para tener absolutamente todos los datos, pero, al menos desde el punto de vista computacional, somos capaces de leer, intentar procesar y, en muchos de los casos, almacenarlos para su posterior utilización. Esto cambia totalmente el universo de discurso al que nos estamos enfrentando. En lugar de tener una serie de datos agregados que eran más o menos fácil procesar y dar una serie de resultados, ahora tenemos todos los datos. Esto crea una serie de problemas nuevos a los que tenemos que enfrentarnos con nuevas tecnologías.

Esto está realmente en el límite de la tecnología. Es decir, actualmente muchos de estos sistemas son artesanales, se construyen específicamente para un problema, y si tenemos un problema ligeramente distinto hay que utilizar tecnologías o incluso inventar soluciones nuevas que sean capaces de resolver este tipo de problema. Por lo tanto, en muchos de estos casos las expectativas que tenemos son bastante desastrosas. No siempre es posible contar con toda la información, aunque en la televisión parezca que sí, como en algunas series en las que a partir de una fotografía con cierta resolución se obtiene información que antes no existía, y que por tanto es inventada.

Efectivamente, se están realizando muchos avances. Ahora mismo tenemos el *movimiento browniano* social, en el que antes podíamos únicamente ver ciertas cosas de las características de una persona, y ahora tenemos todo lo que hay que saber de una persona: podemos saber su ubicación, pero es que además nos cuenta lo que piensa, lo que siente…, puede tener incluso algún tipo de sensor médico, con lo cual sabemos su presión, si realmente se emociona o no se emociona; en *Smart Cities* te puede decir dónde se mueve, el coche puede ir dándote indicaciones de a dónde vas, y hay aplicaciones, como *Wise*, que según arrancas el coche ya te pregunta si vas a trabajar. Realmente, este tipo de aplicaciones pueden llegar a saber todo tipo de información de una forma muy complicada y las leyes que se aplican en cada paso dependen de la situación geográfica de la empresa. En muchos de los casos, utilizamos aplicaciones sin tener conciencia de lo que está ocurriendo con todos nuestros datos.

En cualquier caso, es un mundo fascinante para muchas investigaciones de tipo social o de otros tipos, en los cuales el mundo se está convirtiendo en una gigantesca fuente de datos que podemos analizar

Relación con *Cloud Computing*

- Despliegue de Hadoop en la Nube
- Pagas por lo que gastas
- Soporte en los distintos proveedores
- Engarce con sistemas de almacenamiento y de procesado de valores

La relación entre los sistemas de *Big Data* y *Cloud Computing* es una oportunidad y una respuesta para los pequeños grupos de investigación, permitiéndoles acceder a herramientas que hasta ahora no estaban a su alcance.

y podemos, de alguna forma, utilizar para llegar a diversas conclusiones. Muchas veces el interés radica en el procesamiento de estos datos en tiempo real o casi real, esto depende del problema que tengamos y de lo que se tenga que analizar. Tendremos una mayor cantidad de datos que, en muchos de los casos, serán muy variados. A diferencia de los sistemas que ofrecían datos más o menos procesados (como una operadora de telecomunicaciones que da una serie de datos ya agregados para interpretarlos de una forma más o menos sencilla), ahora hay muchos datos que incluso pueden estar repetidos o tomados desde puntos distintos y con distinta relevancia, o que pueden estar falseados por el propio sistema de medición. Podemos tener datos de cuya veracidad no podemos fiarnos demasiado. Por ejemplo, los datos de un GPS dependen de dónde estemos, la información que proporciona un GPS va cambiando dinámicamente y no siempre tiene el mismo radio de precisión.

Ahora contamos con la posibilidad de procesar y guardar todos los datos, lo que supone un problema muy distinto al de las bases de datos tradicionales porque aquí se realiza una sola escritura y luego múltiples tipos de lectura para procesarlos lo mejor que se pueda dentro de las capacidades que se tienen. Este fenómeno ha llevado a algunos a pensar que podemos estar ante la muerte del método científico, y es ahora mismo uno de los grandes problemas a los que se enfrenta el mundo científico. Es decir, estos datos son relativamente sencillos de procesar, se pueden ajustar con una serie de polinomios y extraer unas fórmulas que sirven para un artículo científico de un campo concreto, pero realmente no se está extrayendo conocimiento nuevo. Los grandes descubrimientos científicos se hicieron cuando se llegó a fórmulas analíticas que analizan realmente lo que estamos observando y aportan información adicional que nos permite predecir o diseñar nuevos comportamientos, como las ecuaciones de Maxwell o cualquier otro gran avance científico.

Es un problema porque es una ciencia guiada por los datos en la que los científicos investigan sobre aquellos que son más fáciles de procesar para llegar a ciertas conclusiones. Y puede ser que, como en las series de datos hay partes que no son tan buenas o no hay tantos datos, empecemos a tener agujeros en la Ciencia. Es por ello que el *Big Data* tiene una serie de riesgos posibles bastante grandes. De hecho, esto está pasando ahora mismo con Google y su experimento para predecir las olas de gripe en Estados Unidos en función de las búsquedas. La predicción les ha fallado en el último año, bien porque la gente se ha adaptado a los sistemas sociales o por otros motivos. Esto significa que el *Big Data* es una herramienta más, pero no podemos convertirlo en el centro de la Ciencia. Evidentemente, es muy interesante y en muchos campos va a ser absolutamente relevante y nos va a permitir tener muchos nuevos avances, pero tenemos que tener en cuenta que es muy peligroso y que algunos de los análisis nos pueden dar enfoques que no son adecuados o incluso hasta engañarnos.

Una de las características típicas que se muestran del *Big Data* es el volumen. Son sistemas con un gigantesco volumen de datos que no están perfectamente controlados, que se generan a una gran velocidad, en muchos casos a mayor velocidad de la que somos capaces de procesar, lo que nos llevará a disponer de diversos heurísticos y datos que pueden no ser totalmente veraces, o bien porque la captura tenga un cierto problema o porque provengan de fuentes que por alguna cuestión no sean totalmente fiables.

Respecto al volumen no se pueden aplicar las mismas soluciones que en *Business Intelligence* porque colapsan debido a ciertos problemas computacionales y a los propios algoritmos que estamos utilizando, que al estar creados para otro tipo de datos pueden afectar negativamente a nuestros resultados e impedir que sean los que realmente podemos llegar a tener.

Necesitamos enfoques que seamos capaces de almacenar y procesar. En este sentido, hasta ahora todas las soluciones van siguiendo el camino que ha marcado Google, que ha creado un sistema distribuido para analizar la web que es libre de escala y, por lo tanto, el algoritmo que se planteó era el más adecuado. El sistema de Google tiene la ventaja de ser robusto y sobrevivir aunque falle cualquier tipo de los componentes.

Los sistemas de contenido *peer-to-peer* son más sólidos en muchos casos que algunos de los servicios de pago: muchos ordenadores colaborando, dando algunas partes de sus datos con un sistema perfectamente coordinado, nos puede proporcionar en muchos casos un tipo de servicio bastante más interesante que el que podamos tener en otro tipo de sistemas. Esto representa un cambio de mentalidad en los sistemas de almacenamiento que surgió de lo publicado por Google en 2006 en una serie de artículos científicos en los que describían su infraestructura y que han permitido crear el resto.

El proyecto más importante que hay ahora mismo en este sentido, *Hadoop*, replica lo que realizó Google inicialmente para avanzar a partir de ahí, aunque hay ya algunas voces que claman la búsqueda de un nuevo paradigma. *Hadoop* diseñó una gigantesca base de datos sin estructura, al contrario de lo que ocurre con las bases de datos SQL, que permite tener múltiples discos, baratos y fungibles, y al menos tres réplicas de cada elemento para evitar que, por cuestiones estadísticas, esta información se pierda.

Por dentro, todos utilizan los mismos sistemas que los *peer-to-peer* en BitTorrent o Emule y sistemas de algoritmos denominados *Distributed Hash Tables* que permiten acceder a contenidos en función de unas claves que se generan. En lugar de ir a una parte de la memoria, va a otro ordenador que está en alguna parte del mundo. Todos estos algoritmos son fundamentales para el avance de esta cuestión. Son la base de las bases de datos no-SQL. Cada base de datos no-SQL resuelve una serie de problemas y

necesitamos ajustarlas de una manera más o menos sencilla. Tenemos los datos, se genera un *hash* o un código a partir de ellos y distribuimos esa clave por el mundo. Si tenemos suficientes participantes, esto nos proporciona una estructura de datos que es muy resistente tanto a fallos como a pérdidas.

¿Cómo se computa esto? Google publicó un artículo en 2004 que explicaba cómo lo hacían inicialmente con *MapReduce*, un modelo de programación que presupone que el ordenador no va a ser capaz de realizar por sí solo toda la computación y se sirve, por tanto, de un número variable de ordenadores disponibles, con lo cual no se necesita calcular ni preparar el programa para el tamaño que quiera tener, por lo que podrá hacerse de una manera absolutamente flexible. Esto permite que la computación avance más o menos rápida en función de los ordenadores que estén disponibles en cada momento. Si tengo, por ejemplo, un sistema de *Cloud Computing*, puedo aprovechar los valles de la demanda para que realicen este tipo de tareas y vayan avanzando. Son sistemas autorregulados y auto-configurados, por lo que nadie puede equivocarse al configurarlos y se adaptan muy bien a los distintos cambios y problemas que podamos tener.

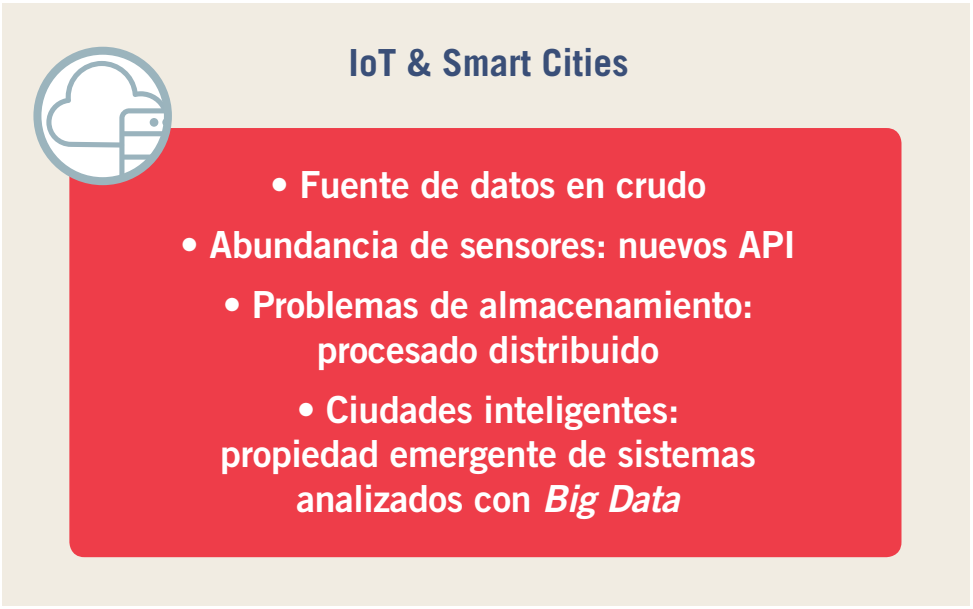
Hadoop es un proyecto de *software* libre, iniciado en el proyecto Apache, que está escrito en Java y que dispone de diversas capas que permiten un uso más sencillo. Aun así, sigue teniendo una cierta complejidad. Programar en el paradigma de *MapReduce* requiere un cambio notable de mentalidad porque son computaciones en las que se van dejando elementos temporales en los discos y se cuenta con una serie de fases en las cuales vamos agrupando los datos para poder paralelizarlos. La ventaja que tiene es que no necesitamos programar específicamente este tipo de paralelismo sabiendo números ni máquinas concretas, sino que se va autoconfigurando en función de lo que tengamos en cada momento. Evidentemente, esto tiene una relación total con el *Cloud Computing*. Los operadores que ofrecen *cloud*

ofrecen ya el servicio de *MapReduce* sobre *cloud*, permitiendo el acceso a estas funcionalidades.

Un problema muy grave es que los datos son muy variados. Muchos de los casos van a ser no estructurados, lo que quiere decir que no van a seguir una estructura, que en algunos casos pueden tener más o menos campos dependiendo de que ciertos sensores funcionen o no, ni van a estar agregados. El que los datos no estén agregados quiere decir que aunque no se haya perdido información, al manejar una gran cantidad de datos se necesita, en general, un precocinado anterior que debe ser manual. Además, como los datos pueden provenir de múltiples fuentes no disjuntas, podemos tener redundancia de datos, lo que puede llevar a que unos algoritmos engañen. Como resultado, tenemos que sobrepasar manualmente la fracasada idea de la web semántica. La promesa de que en la web semántica el proceso iba a ser automático no se ha hecho realidad, por lo que para cada caso concreto alguien tiene que hacer un curado y una unión de todos los datos que tenemos. Ésta es una de las partes más críticas para cualquier proyecto de *Big Data*.

La velocidad es otro de los grandes problemas del *Big Data* ya sea en relación a la generación, el almacenamiento, el movimiento y el procesado.

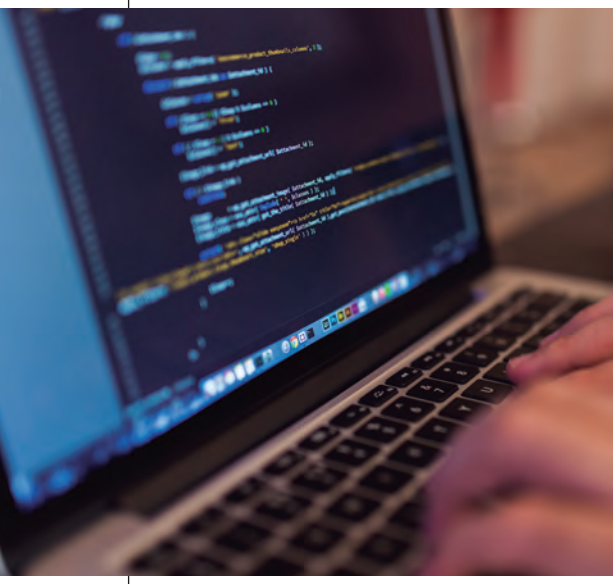
Muchos de los protocolos que se utilizaban, o se siguen utilizando para multimedia, para proporcionar flujo de datos, pueden utilizarse para todo este tipo de cosas. De hecho, muchos de estos esquemas están intentando utilizar las GPU. El problema es que los GPU son unos sistemas bastante difíciles de programar para uso general. La ventaja que tienen es que nos procesan una pantalla entera en un solo ciclo de reloj, es decir, muy rápidamente. Si conseguimos que nuestro problema case con las arquitecturas de una GPU, sí que podemos llegar a tener procesado en *streaming*: Vamos a poder procesar los datos que nos llegan desde los distintos sensores o desde los distintos sitios. De hecho, ya algunos proveedores de *cloud* empiezan a tener proveedores, no solo de CPUs sino también de GPUs, con los que podemos hacer este tipo de procesados en la Nube.



La generación de datos está creciendo de forma exponencial. La sensorización asociada al Internet de las cosas (IoT *Internet of things*) y las ciudades inteligentes (*Smart Cities*) es una de sus principales fuentes de crecimiento.

Otra parte muy importante es la visualización. Mark Twain dijo que había tres tipos de mentiras: las mentiras, las malditas mentiras y las estadísticas. La visualización es esto último, pero es un componente vital de todo análisis. Una visualización debe ser atractiva para que sea publicada y normalmente es llevada a cabo por personas que no se han dedicado a los datos. Alguien tiene que ser capaz de procesar los datos, alguien tiene que entender de computación distribuida para que esto funcione y otro distinto, con un perfil artístico, tiene que ser capaz de hacer la visualización bonita, que es lo que, al final, la gente va a ver de una investigación. Éste es un nicho con mucho futuro en el cual todavía queda bastante por hacer.

Otro punto esencial de la problemática en torno al *Big Data* es la privacidad. A día de hoy existe un cambio de esquema radical en todo este asunto y, de hecho, las nuevas generaciones tienen una perspectiva distinta de la privacidad a la de las anteriores. Las nuevas técnicas de programación utilizan un lenguaje de programación orientado a objetos, denominado *duck typing*, que se basa en el principio de que si algo grazna como un pato, anda como un pato y se mueve como un pato, lo consideramos un pato. En Internet da igual que no sepa tu nombre, pero tengo tu perfil entero. Eso quiere decir que ahora mismo gran cantidad de los sistemas que están conectados a Internet son capaces de conocer cómo nos comportamos, y esto nos puede llevar a muchos resultados. Nuestros datos están en una gran cantidad de sitios. En la obra de Orwell "1984" se planteaba ya la existencia de televisiones que eran capaces de grabar a los usuarios. Ahora, una *Smart TV* tiene HTML 5 y W3C, es capaz de grabar y de enviar la información. Ahora mismo, el Gran Hermano es posible con una *Smart TV*, el Internet de las cosas, las redes sociales y el teléfono móvil. Es un escenario bastante complicado.



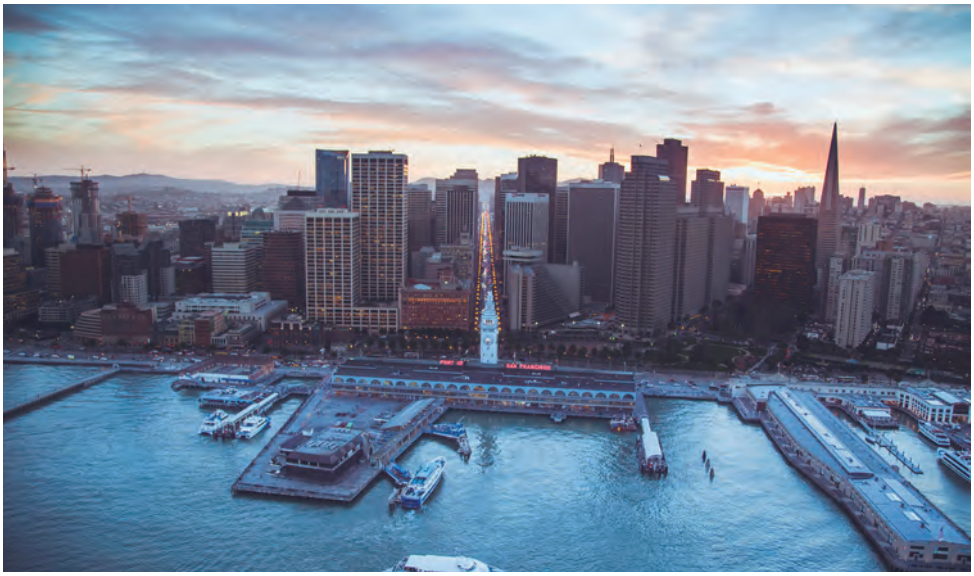
Actualmente se generan tantísimos datos que es necesario que los propios nodos de la Internet de las cosas no solo sean capaces de transmitir información, sino que sean capaz de producirlos. Para esto tenemos que diseñar nuevos algoritmos distribuidos que aún no son capaces de hacer todo este tipo de cosas, pero en el futuro las ciudades inteligentes serán propiedades emergentes de estos sistemas analizados con *Big Data*.

Otro problema es que en algunos de los casos es posible sintetizar información que está protegida por la Ley de Protección de Datos de forma no intencionada. Por ejemplo, si las tarjetas de crédito no se utilizan los sábados en absoluto se puede inferir que sus dueños son judíos, o si no se utilizan para comer durante el día, que profesa la religión mahometana. Es decir, nadie te da esos datos, pero a partir de sus datos los puedes inferir, por lo que debe aplicarse la Ley de Protección de Datos y proteger nuestras bases de datos según esta ley.

Ocurre lo mismo cuando se quiere anonimizar totalmente los datos. Es famoso el caso de Netflix, que organizó un concurso en el que proporcionaba ciertos datos y, de repente, una persona se dedicó a "desanonimizarlos" y hundió el concurso, que tuvo que ser cancelado. Por nuestra parte, estamos trabajando ahora mismo en un esquema de federación de búsqueda de resultados que manejen siempre datos agregados y eviten la posibilidad de acceder a datos concretos.

Por otro lado, también Netflix ha producido una serie, "House of Cards", que es la primera serie cuyo ritmo está producido por *Big Data*. Netflix y Amazon son los primeros proveedores que no solo saben qué películas vemos, sino que saben cómo manejamos el *stop*, el *pause* y el rebobinado; saben cómo vemos las series, a qué ritmo la vemos, y a qué horas. Toda esa información ha sido procesada con algoritmos de *Big Data* y se ha enviado a los guionistas, por lo que "House of Cards" es la primera serie en la que los guionistas obtienen realimentación sobre cómo son los patrones de las personas que consumen este tipo de series. La han estrenado hace relativamente poco y no han terminado de cerrar el bucle, pero estamos empezando a disponer de un sistema en el que las series se producen en función de los resultados que tomamos de *Big Data*, y que a su vez producirá más *Big Data* y empezará a tener un efecto, no se sabe si positivo o negativo, sobre diversos campos diferentes a la Ciencia, como el periodismo.

La importancia de ser capaces de gestionar datos no solo se aplica a científicos, sino también a cualquier periodista. Es decir, hay mucha información que parece muy interesante y muy importante para la sociedad que está disponible en iniciativas de *open data*, y que colectivos que hasta ahora no se suponía que tenían que manejar *Big Data* van a tener que hacerlo, y eso nos lleva a necesitar entornos muy sencillos para que usuarios no especializados sean capaces de extraer este tipo de información. En este sentido, los lenguajes que ahora mismo están más en boga son, por ejemplo, "R", que es el que tiene mayor extensión, comunidad y flexibilidad, aunque va un poco lento;



Julia, que es una versión con un enfoque ligeramente distinto que va bastante más rápido; y NumPy, que es una extensión numérica a Python que ha sido seleccionada precisamente por el DARPA para todos sus proyectos de *Big Data*.

Otro tipo de problemas ya existían antes. La extrapolación, situaciones como la protagonizada por Marissa Mayer, proveniente de Google y directora de Yahoo!, que ha eliminado el teletrabajo porque había analizado con *Big Data* los datos de las conexiones a las redes privadas virtuales y su conclusión fue que la mitad de la gente no trabaja. ¿Cuál es la justificación? Que lo ha dicho el *Big Data*. Esto puede dar lugar a casos horribles en muchas empresas porque el procesamiento de *Big Data* y sus extrapolaciones y conclusiones pueden ser totalmente falsas. Esto supone un riesgo muy grande que realmente podremos encontrar en múltiples aspectos de nuestra vida.

Actualmente también existen grandes problemas en redes sociales. Aquí el paradigma ha cambiado radicalmente con la ciencia de redes y es extrapolable a otros campos. Ahora mismo el grafo social es un tesoro que nadie quiere soltar, de hecho hoy las grandes guerras en Internet son sobre quién tiene el grafo social para poder analizarlo con la ciencia de redes. Sin embargo, se necesita todavía bastantes avances teóricos y existen también bastantes problemas para procesar grafos. *MapReduce* y las bases de datos SQL no pueden procesarlo y aunque Google ha publicado una respuesta, esta vez no ha acertado y la gente no lo está siguiendo. El problema de los grafos es que no tenemos un sistema adecuado para ver toda la información que se está produciendo en las redes sociales.

Por último, otro problema es el acceso. Hay mucha gente que dice que los datos son el nuevo petróleo, pero como no se sabe qué hacer con todos esos datos, de momento se tienen bajo veinte llaves. Creo que para que la Ciencia avance, en muchos de los casos van a ser necesarios enfoques abiertos y colaborativos que permitan que las personas compartan sus datos. Pueden ser datos antiguos, pero se necesitan datos para que las personas que inventan algoritmos y sistemas los puedan probar y puedan seguir avanzando y superando limitaciones teóricas, ya que el acceso a los data *sets* es bastante complicado.

BIG DATA:
DE LA INVESTIGACIÓN CIENTÍFICA A LA GESTIÓN EMPRESARIAL

BIG DATA:

DE LA INVESTIGACIÓN CIENTÍFICA A LA GESTIÓN EMPRESARIAL

INTRODUCCIÓN GENERAL

La segunda jornada organizada por la Fundación Ramón Areces en torno al mundo del Big Data *y del* Cloud Computing *tuvo lugar el 3 de julio de 2014. La jornada, coordinada por el profesor* ***José García Montalvo*** *y el consultor en TIC* ***Julio Cerezo****, fue continuación de la realizada el año anterior, titulada* El impacto de la Nube y el Big Data en la Ciencia*. La iniciativa se encuadraba dentro del interés de la Fundación por el análisis del impacto en la sociedad de las nuevas tecnologías de la información y la comunicación surgidas desde el ámbito científico, y de las implicaciones que representan su implantación y uso.*

Si en la primera jornada las materias analizadas se centraban en el ámbito de las Ciencias de la Naturaleza y de qué forma la Nube y el *Big Data* están modificando la forma de investigar en Medicina, Física o Astronomía, en esta ocasión el foco estuvo dirigido a estudiar los retos y oportunidades del "Big Data" en las Ciencias Sociales y, específicamente, en la Economía y la gestión empresarial.

El *Big Data* es uno de los fenómenos actuales de mayor transcendencia en el ámbito del desarrollo científico y tecnológico. Asociado a la gestión de gigantescos volúmenes de datos, de muy diversa naturaleza y cuyo tratamiento no se puede realizar con las herramientas y analíticas convencionales, la Ciencia de los Datos representa una nueva realidad para la sociedad en su conjunto, en distintos campos y disciplinas. Y ha sido en el mundo de la Economía y la empresa donde el impacto de las nuevas tecnologías está teniendo consecuencias disruptivas y generando una auténtica revolución en los modelos de negocio de diferentes industrias y economías.

Según el informe "Open Data in Europe", realizado por la Fundación DemosEUROPA, en 2015 la inversión total prevista en *Big Data* alcanzará los 132.000 millones de dólares. El comercio, la industria, la salud, la información, las comunicaciones, la banca, los seguros y la Administración pública son los sectores donde el aumento de la inversión será más relevante. Además, generará 4,4 millones de empleos en todo el mundo y aumentará la riqueza de la Unión Europea con un 1,9% adicional en el PIB para 2020.

Las tecnologías *Big Data* no solo ayudan a recopilar grandes cantidades de datos, sino que además permiten su almacenamiento, organización y recuperación para aprovechar todo su

valor. Y con el objetivo puesto en que su uso permita optimizar la toma de decisiones.

El objetivo de la jornada era mostrar –a partir del *Big Data* y la Computación en la Nube– el recorrido que existe entre la investigación científica y el mundo empresarial y económico. Y dar cuenta de ese camino y de los diferentes elementos –Universidad, empresa, tecnología, ciudadanos– que intervienen en el proceso.

La jornada tuvo un marcado carácter multidisciplinar, reflejo de la complejidad e interrelación que se da en la realidad. Expertos de diferentes instituciones, universidades y compañías privadas explicaron el pasado, el presente y el futuro de unas herramientas tecnológicas que en muy poco tiempo han pasado de la investigación científica al mundo de la empresa.

La jornada mantuvo la misma estructura que la primera de las celebradas. Dividida en dos sesiones, en la sesión de la mañana se analizaron los temas transversales que afectan en su conjunto a los diferentes ámbitos y sectores. En la sesión de la tarde, la atención se centró en el análisis de aspectos específicos, para ayudar a visualizar, con ejemplos concretos y experiencias reales, el alcance del fenómeno *Big Data* y cómo está abriendo las puertas a un nuevo enfoque de entendimiento de la realidad para la toma de decisiones en el ámbito empresarial.

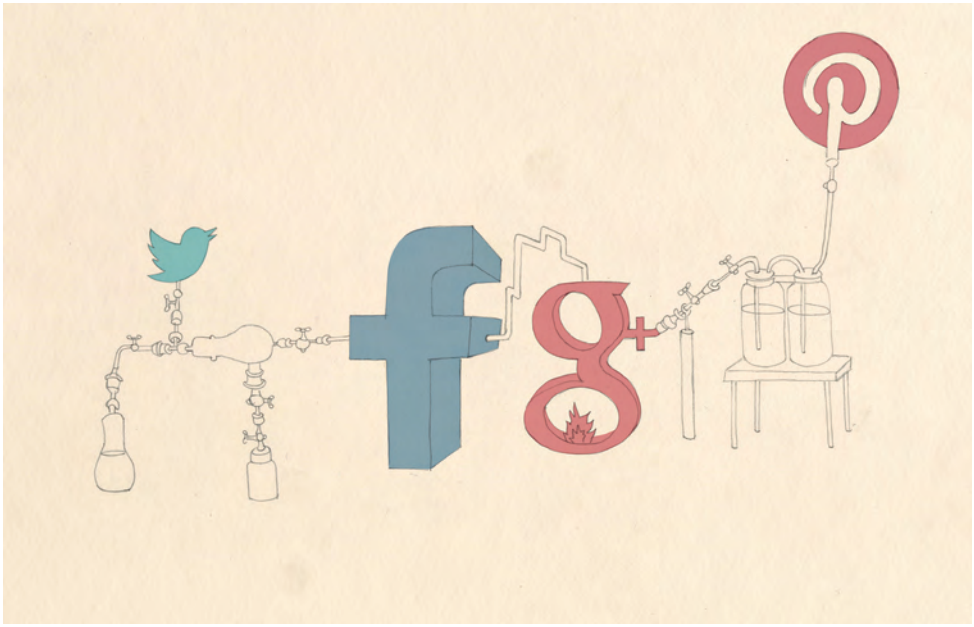
El papel de los superordenadores

Mateo Valero, director del Centro Nacional de Supercomputación del CSIC, con sede en Barcelona, fue el encargado de abrir la jornada con su ponencia *El estado del arte del Big Data & Data Science. La revolución de los datos.* Valero dedicó su tiempo a hablar de la computación, los superordenadores y cómo estos se encuentran en el centro de la innovación científica y son responsables de adelantos científicos como la secuenciación del genoma, la evolución de células cancerígenas o la identificación del bosón de Higgs en el CERN... "La simulación que ofrecen los supercomputadores es lo que permite avanzar la Ciencia. Sin esa simulación no se avanzaría", afirmó. Y destacó cómo el *hardware* ha llegado a un nivel de desarrollo en el que es capaz de discriminar de toda esa información que va almacenando "cuál va a ser realmente útil para un determinado estudio". El director del Centro Nacional de Supercomputación subrayó que las predicciones son peligrosas, que se hacen para fallarlas, y recordó que cada 10 años se multiplica por mil la velocidad de procesamiento de datos. También advirtió que "la tecnología va muy por delante de las leyes y esto es muy peligroso".

A continuación intervino **Daniel Villatoro**, perteneciente a BBVA Data & Analitycs, con su exposición sobre *Big Data, economía y organizaciones*. Villatoro, cuya experiencia en investigación se centra en tres áreas científicas –la Ciencia cognitiva, que es cómo los humanos tomamos decisiones; la Economía experimental, o cómo los individuos toman decisiones que afectan a sí mismos y a otros; y las redes sociales enfocadas al nivel de interacción y cómo el entorno y las relaciones afectan a las decisiones que tomamos–, destacó que el objetivo de su trabajo se centraba en tratar de responder a las grandes preguntas que esconde el *Big Data* desde un punto de vista científico: "¿Cómo tomamos decisiones los humanos?, ¿por qué? y ¿cómo estas decisiones podrían afectar eventualmente a la eficiencia de nuestro negocio como banco, cómo ahorrar costes o dar un mejor servicio?" Datos para entender el comportamiento de las personas, en este caso, relativo al consumo de productos y servicios.

Uso de datos y privacidad

Dentro de los elementos transversales del *Big Data*, la ética y la privacidad de los datos ocupan un lugar destacado. Para hablar sobre esta cuestión intervino **Ricard Martínez**, Data Protection Officer de la Universitat de Valencia y presidente de la Asociación Profesional Española de la Privacidad. "Cualquier persona tiene derecho a proteger la informa-



ción sobre ella y los problemas van mucho más allá de la política de protección de datos", aseguró.

Martínez explicó cómo gran parte del modelo de negocio en *Big Data* se basa en que el titular de los datos ha consentido ese uso. "El usuario no suele ser consciente de que está dando todos esos permisos. Sociólogos norteamericanos han calculado que necesitaríamos 100 días para leer y entender todos los contratos de consentimiento que aceptamos por usar *apps*, redes sociales." El presidente de la Asociación Profesional Española de la Privacidad se refirió también a los riesgos que implica el uso masivo de datos: "El análisis masivo de datos va a permitir muchos avances en Medicina, también en *marketing* o consumo –como cuando Amazon nos recomienda libros que nos interesan– pero hay que ofrecer un marco seguro, equilibrar las posiciones y no basarlo todo en un consentimiento que es falaz, evitar situaciones de cuasimonopolio y fomentar la transparencia", indicó.

Esclavos de las máquinas

La sesión de la mañana finalizó con la intervención de **Carsten Sørensen**, de la London School of Economics, *Datos y empresa: El auge de las máquinas*. Sørensen se centró en profundizar en el incipiente fenómeno del Internet de las cosas y en las diferentes formas de comunicación con las máquinas. "Las empresas buscan que nos casemos con ellas, ya no nos ofrecen productos en sí, sino emociones", señaló. "Las TIC se van a utilizar cada vez más para permitirme hacer el trabajo que tendrían que hacer las empresas por mí, como pagar yo solo sin que nadie me atienda... Ahora somos nosotros los que servimos a las máquinas, seremos sus esclavos", añadió.

Sørensen reconoció que lleva estudiando Internet desde 1993 y que desde entonces nunca ha sido capaz de predecir lo que iba a suceder en un horizonte más allá de dos años. Sin embargo, sobre el futuro de la computación y el empleo, advirtió que la sociedad camina hacia una polarización del mercado de trabajo: "A un 10% de la población le lloverán las ofertas y al resto no lo querrá contratar nadie, por lo que habrá una tremenda polarización".

La segunda sesión de la jornada se orien-

tó hacia temas más específicos del *Big Data*, como la gestión de los datos en la empresa, el *Big Data* y los servicios financieros, el análisis predictivo de las redes sociales o la relación entre opinión pública y los mercados.

Manuel Machado, socio director de Deloitte, comenzó con la ponencia sobre *Big Data y servicios financieros*, en la que analizaba las posibilidades del *Big Data* para la mejora de los servicios financieros y la experiencia del cliente, así como para aumentar la eficiencia de las corporaciones en un contexto de presión sobre la rentabilidad de las entidades financieras. La utilización de técnicas de *Big Data*, recordó, "se ha extendido a la calificación crediticia de los solicitantes de créditos o hipotecas, la detección del fraude en tarjetas, la microsegmentación o los servicios de información a los clientes".

Por su parte, **Óscar Méndez**, CEO de Stratio, una compañía con sede en Palo Alto (California) y que ha participado en proyectos de *Big Data* con muchas de las empresas incluidas en el Ibex 35, en su intervención, bajo el título *Los datos, la nueva materia prima del marketing*, destacó que "las compañías más valoradas del mundo son las que mejor usan los datos: Google, Apple, Facebook". Asimismo, recomendó a aquellas empresas que quieran empezar a trabajar con estas tecnologías que realicen un estudio de la madurez de uso de los datos que manejan.

Méndez aclaró que a las empresas hay que hablarles de resultados económicos y que el *Big Data* va en esa dirección, ofreciendo infinitas posibilidades en el área de *marketing*. A partir del análisis a gran velocidad de los datos ya almacenados de lo que ocurrió en el pasado y de lo que está pasando ahora se pueden predecir los comportamientos de los clientes en el futuro. "Por ejemplo, es muy útil para realizar un seguimiento en tiempo real de una campaña publicitaria, para comprobar si está funcionando y modificarla sobre la marcha. Sin embargo, intentar predecir algo en *Big Data* sin contar con científicos de datos es como tener el mejor avión sin piloto", insistió durante su intervención en la Fundación Ramón Areces.

Las dos últimas ponencias de la jornada se centraron en analizar un fenómeno de nuestro tiempo y muy ligado al *Big Data*: las redes sociales, que aportan un flujo incesante de datos que ofrece enormes oportunidades de negocio, tanto en términos de conocimiento de los clientes como de apertura de nuevos canales de mercado. Las redes sociales se han generalizado y la adopción por parte de los usuarios se ha universalizado. **Esteban Moro**, profesor de la Universidad Carlos III de Madrid, centró su intervención sobre *Big Data y análisis predictivo* en caracterizar esta realidad donde cada minuto en Internet se envían más de 200 millones de emails, se realizan 2 millones de búsquedas en Google o se generan 350 Gb de datos en Facebook. Para Moro, el tratamiento y estudio correcto de toda esta información es lo que posibilita el análisis predictivo, la capacidad de anticipar comportamientos o respuestas a partir de los datos previos. Entre los ámbitos donde se aplican los modelos predictivos, citó la detección de fraude y la gestión de riesgos en el sector financiero y en seguros; la adopción de nuevos productos y los servicios de recomendación en *marketing*, o los deportes. Por último, Moro señaló los riesgos que existen para los modelos predictivos, como el hecho de confundir causalidad con correlación: "Aunque ciertas variables muestren poder predictivo, eso no significa que hayamos encontrado un mecanismo que explica lo sucedido".

Por último, **Daniel Gayo-Avello**, profesor de la Universidad de Oviedo, habló de *Big Data, Twitter, opinión pública y mercados*. Su ponencia se centraba en el análisis de esta red social que, por sus características, es idónea para su estudio y análisis como canal de expresión de la opinión pública ("adversativa"), destacando también las dificultades y problemas existentes para que los modelos cumplan correctamente con su función predictiva.

El estado del arte del *Big Data & Data Science*. *La revolución de los datos*

Por Mateo Valero
Director del Centro Nacional de Supercomputación

Un país avanzado tiene que generar ideas e introducir esas ideas en productos competitivos. Pero para generar esas ideas se necesita un ecosistema que funcione muy bien. Este ecosistema está compuesto por universidades y centros de investigación, las administraciones y las empresas. Y si unimos esfuerzos, seguramente España pueda llegar a ser un país muy competitivo. Es de conocimiento común que los países más ricos hoy en día son los que más dinero han dedicado a la investigación en los últimos años, y España, desgraciadamente, no es uno de ellos. Si no dedicamos suficientes recursos para generar ideas que, a su vez, produzcan más recursos, muchos más de los que se originan en un principio, un país no tiene futuro. Es un trabajo de todos y las administraciones deberían dar recursos que sean suficientes y constantes. Además, estos recursos deben gestionarse correctamente porque a veces los recursos que se dedican son más o menos razonables pero la gestión es horrorosa y se terminan desperdiciando muchos de ellos. Ante todo, debería haber un Pacto de Estado por la Ciencia, un país que no tenga un pacto para la Ciencia no puede avanzar. En cuanto al papel de las empresas, se ha hecho muchísimo pero hay que seguir avanzando y una de las obligaciones de la universidad es producir ideas, investigar en temas punteros y hacer que esas ideas, en colaboración con las empresas, produzcan riqueza.

Para aplicar la teoría, además de los tradicionales laboratorios, a día de hoy necesitamos ordenadores que ejecuten programas y obtengan resultados muy rápido. Los supercomputadores son los aceleradores de la teoría

Supercomputación

Los supercomputadores son los ordenadores más rápidos del mundo y son el tercer pilar para la Ciencia y la ingeniería. Sin teoría, sin Matemática, sin Física no se va a ningún lado. Pero para aplicar la teoría, además de los tradicionales laboratorios, a día de hoy necesitamos ordenadores que ejecuten programas y obtengan resultados muy rápido también para que los expertos de cualquier ciencia o ingeniería puedan comprobar sus teorías y les ayuden a avanzar. Todo para que, en definitiva, se den aplicaciones prácticas, modifiquen la experimentación y modifiquen la teoría. Los supercomputadores son los aceleradores de la teoría.

Necesitamos que el conocimiento avance, pero sobre todo y aunque los supercomputadores ya han avanzado muchísimo, necesitamos máquinas mucho más rápidas. Durante los últimos 30 años, cada 10 años, la velocidad de procesamiento se ha multiplicado por mil. En 30 años se ha multiplicado por mil millones. Esto supone que el supercomputador más rápido del mundo hace 12 años hoy en día quepa en un chip. Los supercomputadores son máquinas que teóricamente nos permiten soñar, si las utilizamos adecuadamente, porque nos permiten simular cosas que sería imposible simular de otra manera; nos permiten abarcar nuevos retos que sin esa potencia de cálculo son imposibles. Por otro lado, hay que tener en cuenta la energía que utilizan estas máquinas, además de la de refrigeración. Un computador de 15 a 16 megavatios tiene un coste de 15 a 16 millones de euros al año y genera emisiones de CO_2 equivalentes a las de 400 coches circulando a 100 km/h constantemente.

Centro Nacional de Supercomputación

El origen del Centro Nacional de Supercomputación (o BSC, por Barcelona Supercomputing Center) se originó en el departamento de investigación de la Universidad Politécnica de Cataluña de arquitectura de computadores y computadores paralelos, un departamento pionero en Europa y de los mejores del mundo. Creamos el Centro Europeo de Paralelismo de Barcelona y en el año 1984 ya trabajábamos con computadores paralelos. IBM se interesó por las investigaciones llevadas a cabo en el centro y lo financió durante cuatro años para crear el CEPBA-IBM Research Institute (CIRI). Enseguida contactamos con las empresas líderes en informática para establecer centros de investigación con ellos. Si en la universidad española creamos buenos centros de investigación, sobre cualquier tema, la financiación llega, aunque no sea de Madrid o de Barcelona, porque las empresas se interesan y colaboran.

Desde los años 80 ya sabíamos que queríamos ser un grupo multidisciplinar: un grupo que fuese experto tanto en *software* (sistemas operativos, compiladores, *runtime*, etc.) como en *hardware*. Y desde entonces diseñamos computadores y fuimos ganando una gran experiencia en el uso de computadores paralelos, hasta que más tarde llegó el MareNostrum al BSC.

La entrada de España a la Unión Europea fue muy buena para aquellos grupos que podíamos competir y colaborar a nivel europeo. Desde el año 1986, en el CEPBA hemos conseguido más de 30 millones de euros para empresas y grupos de investigación españoles. Muchos proyectos para grandes marcas nos dieron una visibilidad muy grande y utilizábamos el dinero que nos sobraba de la colaboración europea para dedicarlo a la investigación básica y a la formación de nuestros doctorandos, creciendo poco a poco. Sorprendimos al

Evolution over time of the research paradigm

- **In the last millenium, science was empirical**
  - Description of natural phenomena
- **A few centuries ago opens the theoretical approach**
  - Using models, formulas and generalizations
- **In the recent decades appears computational science**
  - Simulation of complex phenomena
- **Now focuses on the exploration of Big Data (eScience)**
  - Unification of theory, experiment and simulation
  - Capture massive data using instruments of generated through simulation and processed by computer
  - Knowledge and simulation stored in computers
  - Scientist analyse databases and files on data infrastructures

Jim Gray, National Resarch Council, http://sitesnationalacademies.org/NRC/Index.htm; Computer Science and Telecommunications Board, http://sitesnationalacademies.org/cstb/Index.htm

Big Data, la última etapa en la evolución del paradigma de la investigación científica.

mundo diseñando con IBM el cuarto supercomputador del mundo, el primero que utilizó Linux (ahora todos utilizan Linux), con procesadores iguales a los que llevan los Apple. Desmitificamos aquello de que para hacer un supercomputador había que utilizar tecnología muy cara: utilizamos tecnología que ya se utilizaba para otras cosas. A nivel mundial somos pioneros, somos los únicos que han propuesto utilizar, en vez de esos *chips* con los que se construyen los supercomputadores que consumen muchísimo y son costosísimos, utilizar los *chips* de teléfonos y tabletas, que son muy baratos y consumen poco.

Como consecuencia de esto, los patronos de este centro (el Ministerio de Educación, la Generalitat de Cataluña y la Universidad Politécnica de Cataluña) decidieron crear el BSC. El BSC tiene dos objetivos: Dar servicio a todos los investigadores españoles e investigar. La investigación del centro se divide en cuatro departamentos diferentes: Ciencias de la Tierra, Ciencias de la Vida, Ciencias Informáticas y Aplicaciones Informáticas. También creamos la Red Española de Supercomputación, de manera que tenemos computadores conectados por la red en varios lugares de España, de manera que puedan ser utilizados fácilmente.

Publicar en algunas revistas y congresos siempre es fundamental, pero la buena investigación es mucho más importante. Y la buena investigación debe generar riqueza. No solo la que se publica en los mejores sitios, que tiene su valor, pero si las ideas no se llevan a la práctica, no sirven para nada. Ni Bill Gates, ni los fundadores de Facebook y Google escribieron ningún artículo, pero generan muchísima riqueza y tienen muchísima influencia en el mundo. Hemos colaborado con muchas empresas. Tenemos alianzas estratégicas con empresas generadoras del mejor *hardware* y *software* para computadores, no solamente supercomputadores, sino de computadores de altas prestaciones. Y luego, a nivel español, tenemos una alianza muy valiosa con Repsol. También hemos empezado con Iberdrola y pronto empezaremos a colaborar también con un banco español no rescatado.

Los centros de investigación que tienen un cierto nivel y quieren colaborar con empresas necesitan dinero europeo. Pero la financiación depende de la calidad. En total hemos cobrado ya, hasta diciembre del 2011, casi 60 millones confirmados, unos 30 más que vamos cobrando y desde luego los resultados son muy impresionantes. Por cada euro que nos han dado, hemos obtenido prácticamente seis de fuera de España para investigar: el 40% de estas fuentes de investigación proviene de la colaboración con empresas, el 35% proviene de proyectos europeos, y menos del 20% proviene de los presupuestos del Estado.

Y ahora voy a describir el trabajo que desarrolla cada una de las áreas de investigación del Centro.

Ciencias de la Tierra

Equipo de unas 20 personas que investigan temas como el cambio climático, realizan predicciones de la calidad del aire en tiempo real, etc. El *software* utilizado para esto último ha sido utilizado por muchas empresas para ver la influencia de, por ejemplo, construir una central térmica.

Ciencias de la Vida

Departamento de unas 100 personas, que utilizan la supercomputación para investigar genes, proteínas, medicamentos, etc. El equipo del departamento ya ha sido portada de la revista Nature y han investigado, por ejemplo, el genoma del tomate, o cómo los genes que hay en el estómago indican qué enfermedades puede tener un cuerpo. Este es un campo realmente muy importante en el que, por ejemplo, podemos secuenciar un genoma. La tecnología ha avanzado mucho durante los últimos años y cada año se multiplica por cuatro la velocidad del proceso de obtención del genoma de una persona. A día de hoy se hace en un día y la parte fundamental cuesta menos de 1.000 USD. A partir de ahí se pueden llevar a cabo muchas investigaciones. Tenemos a toda la comunidad científica española alrededor del MareNostrum, del almacenamiento de esos datos y de la creación de los mejores programas para trabajar en función de esos datos.

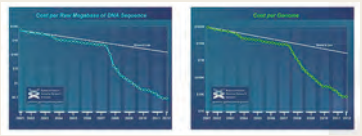
Uno de los ejemplos también publicado en Nature es el trabajo sobre medicina personalizada que están llevando a cabo desde diferentes instituciones los doctores Carlos López Utín, Elías Campo, y los profesores Modesto Orozco y David Torrents. Por primera vez se ha podido proporcionar al médico información en tiempo real para mejorar el tratamiento del paciente y ver cómo evoluciona. Esto abre unas posibilidades tremendas. Dentro de dos o tres años, gracias al avance de la tecnología (por ejemplo, Nvidia, una empresa que diseña procesadores gráficos muy potentes, ha creado un procesador de 6 cm2 de silicio con 1,5 teraflops), cualquier investigador que necesite potencia de cálculo tendrá encima de su mesa la potencia que tenía MareNostrum hace menos de diez años, lo cual permite hacer cosas impresionantes como esta, no solo en hospitales, sino también en empresas, centros de investigación, etc. Respecto a los fármacos, hemos desarrollado la base molecular de proteínas más grande de Europa, abarcando también la proteína en movimiento, lo que permite generar nuevos modelos físicos y matemáticos que permiten crear nuevos fármacos más rápidamente. Este departamento tiene alianzas muy fuertes con empresas a nivel español, europeo y mundial, como Schrödinger o Danone. También está generando *startups* en temas muy concretos.

Ciencia de los computadores

Es el departamento más grande, está originado en la UPC y trabaja en el desarrollo de *hardware* y *software*, desde en móviles hasta en supercomputadores, y pasando por centros de datos, ordenadores personales, etc., ya que la tecnología es la misma, los modelos de programación son los mismos, y los problemas de energía son los mismos, a diferente escala pero todo es igual.

Colaboramos con IBM, Microsoft,

Sequencing costs



Source: **National Human Genome Resarch Institute** (NHGRI)
http://www.genome.gov/sequencingcosts/

(1) "Cost per Megabase of DNA Sequence" – The cost of determining one megabase (Mb; a million bases) of DNA sequence of a specified quality
(2) "Cost per Genome" – The cost of sequencing a human-size genome. For each, a graph is provided showing the data since 2001.

In both graphs, the data from 2001 through october 2007 represent the cost of generating DNA sequence using Sanger-based chemistries and capillary-based instruments ('first-generation' sequencing platforms). Beginning in january 2008, the data represent the cost of generating DNA sequence using 'second-generation' (or 'next-generation') sequencing platforms. The change in instruments represents the rapid evolution of DNA sequencing technologies that has ocurred in recent years.

La tecnología ha avanzado mucho durante los últimos años y cada año se multiplica por cuatro la velocidad del proceso de obtención del genoma de una persona. Ahora se hace en un día y la parte fundamental cuesta menos de 1.000 USD.

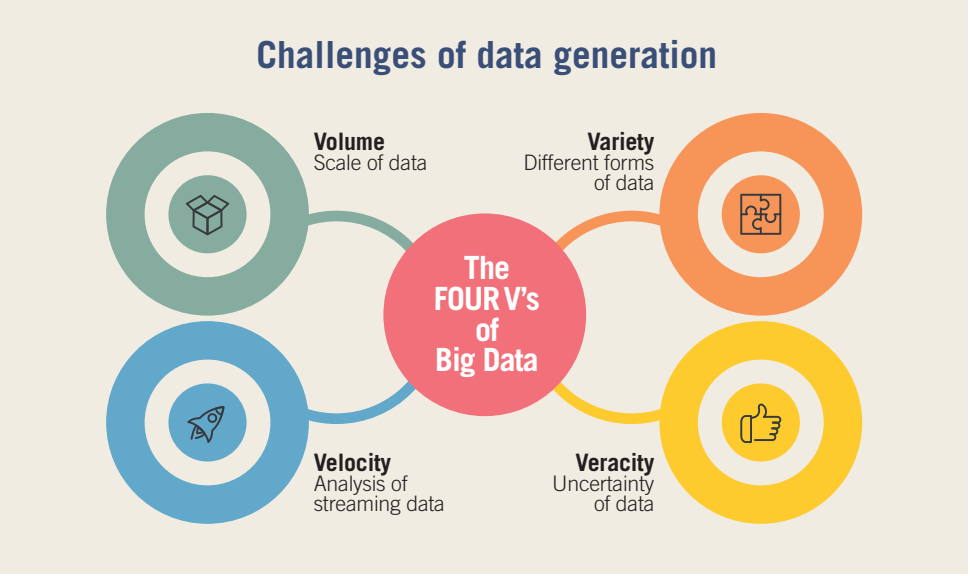
Nvidia, Intel, etc. Por ejemplo, con IBM hemos tenido hasta 40 personas trabajando en redes de interconexión, modelos de programación, aplicaciones, etc. Microsoft nos llamó en 2005, convirtiéndonos en el primer centro del mundo que trabajó con Microsoft en el diseño de *hardware*. Nuestros investigadores publican en las mejores revistas y congresos del área pero si solamente publicamos en las mejores revistas y congresos y no hacemos nada más, nos convertiremos en centros de investigación gratuitos para las multinacionales que van a copiar lo que hacemos y explotarlo económicamente.

Probablemente el BSC es el número uno a nivel mundial en modelos de programación y herramientas para ver el comportamiento de programas. También, a nivel europeo, nuestro centro, que colabora con la Agencia Europea Espacial y Airbus, es quizás de lo mejor que hay en diseño de procesadores, donde lo importante no es que sean muy rápidos, sino que lleguen a tiempo. Ante ciertas circunstancias hay que garantizar que el computador va a dar una solución y controlar aquello que controla en tiempo real (aviación, en frenos de coches, etc.).

Aplicaciones de ingeniería

Es un departamento muy multidisciplinar de unas 35 personas, que desarrolla *software* para computación de alto rendimiento (*High Performance Computing* o HPC en inglés) para proyectos de mecánica de fluidos computacional (*Computational Fluid Dynamics* o CFD en inglés), mecánica de sólidos, electromagnetismo, etc.

Los modelos de colaboración de este departamento con empresas incluyen la optimización de aplicaciones externas (Airbus) o *software* a medida, como el que hemos desarrollado para Repsol, que le ha permitido ahorrar millones de euros gracias a que hemos aumentado en un 25%-30% las probabilidades de éxito de sus sondeos, que pueden costar 100 millones de euros cada uno. También colaboramos con Iberdrola para optimizar el funcionamiento de los aerogeneradores,



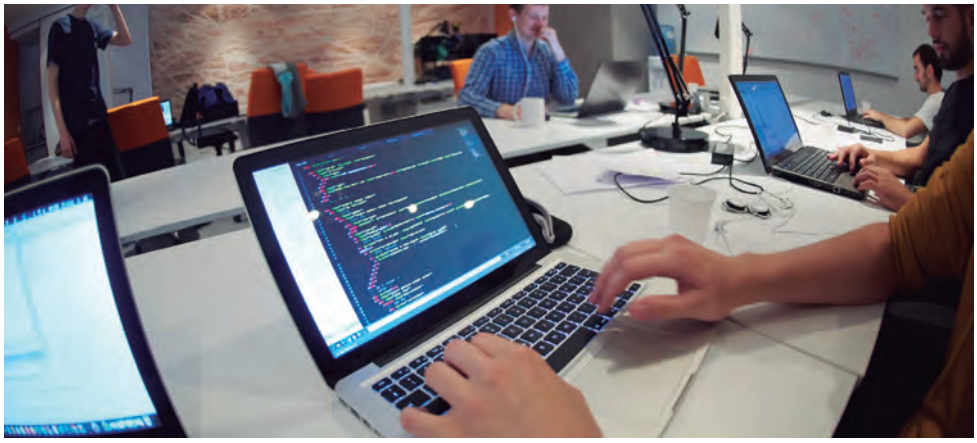
Las variables que definen los retos del *Big Data*: volumen, velocidad, variedad y veracidad.
Fuente: htpp://www-01.ibm.com/software/data/bigdata/.

y con Aerolíneas Argentinas para ayudarles a redirigir aviones de manera optimizada en función del estado de la atmósfera, especialmente en casos de erupciones de volcanes. En definitiva, ayudamos a las empresas a ahorrar muchísimo dinero con tecnología que difícilmente podrán tener, aunque sean grandes empresas. Desarrollamos tecnologías globales que, debidamente modificadas, pueden utilizarse para diferentes aplicaciones. Por desgracia, la crisis afecta a empresas españolas muy importantes que se han quedado hasta sin departamentos de investigación. En estas circunstancias es muy complicado transferir la tecnología a la empresa.

Otro proyecto es una simulación en 3D y de 3GB del funcionamiento del corazón. ¿Dónde se genera el impulso eléctrico en el corazón? Nadie lo sabe. Pero sí se sabe que los iones tienen un papel en ello, que dependiendo de dónde se genere puede dar problemas, y que los fármacos pueden cambiar el punto en el que se genera dicho impulso. Por tanto, las aplicaciones de este corazón son enormes y queremos, cuando lo mejoremos, que los médicos lo utilicen básicamente para dos cosas: para llevar a cabo operaciones antes de operar de verdad, ya que es posible hacer modelos electrónicos de cada corazón; y para observar la influencia de muchos fármacos en tiempo real cuando hay una patología.

En definitiva, en España nos queda mucho camino por recorrer pero estamos orgullosos y contentos con lo que hemos hecho hasta ahora, con nuestras colaboraciones con las mejores multinacionales y empresas del país, y con la educación. La buena investigación es la que produce riqueza en tu entorno. De nada sirve saber la influencia del botijo en la invasión de los bárbaros. Igual es un tema muy interesante pero no va a generar riqueza. Si hubiésemos patentado nuestras investigaciones de los últimos años, seríamos millonarios. Pero los centros de investigación tenemos que dedicarnos a la investigación puntera y tener buenas ideas. Y las ideas deben ser convertidas en riqueza por los empresarios. No creo que sea nuestra labor y, de hecho, muchos profesores que tienen buenas ideas no están preparados para llevar esa idea al mercado, tienen que ser las empresas.

Datos y empresa: *el auge de las máquinas*



Por Carsten Sørensen
London School of Economics

¿Por qué nos interesa el gran fenómeno del *Big Data* en este momento de la historia? Existen cinco factores principales: las empresas y cómo han cambiado en los últimos 300 años, las máquinas y cómo han ido cambiando desde hace 175 años, los materiales y cómo entendemos el esfuerzo humano en términos materiales, la innovación y cómo está cambiando y el **futuro y cómo éste será**.

En el siglo XIX, con una nueva clase media y un modelo de consumo de principios de la época Moderna, comprar un producto requería ir a una persona y encargarlo para que esta persona pudiese fabricarlo. Podía ser un carro de caballos, una peluca, un reloj o un libro, pero todos ellos eran productos hechos a mano. A día de hoy la tecnología ha cambiado pero todo lo demás sigue igual.

En el siglo XX ocurrió un fenómeno fantástico. De repente, la clase media emergió y el número de personas que podía comprar cosas se disparó: era una sociedad de masas de consumo de masas. Hoy casi todo el mundo tiene acceso a un teléfono móvil o a una red local de telecomunicaciones; de hecho, estudios contrastados muestran que incluso las personas pobres de los suburbios en India escogen tener un teléfono móvil.

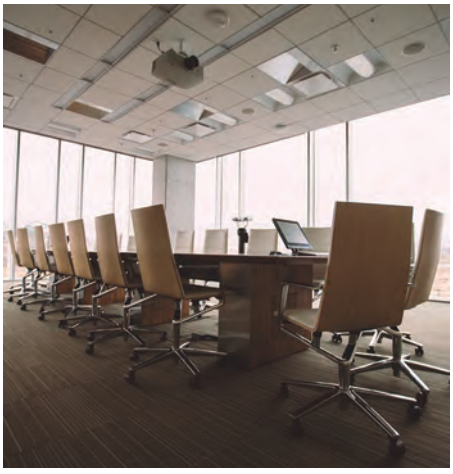
A día de hoy, en una nueva sociedad de individuos, el reto del siglo XXI es cómo conseguir que las personas se sientan animadas y felices al comprar. Así, en vez de vender productos, comenzamos a vender servicios que tienen que ser individualizados. El desafío de las empresas es cómo proporcionar relaciones de servicios, y cómo

Hemos pasado de la era de la máquina inteligente, que trataba de encuentros, a la era de la máquina generativa, que trata sobre relaciones

involucrarse en una relación emocional con el cliente. Para los clientes es como casarse con las empresas. Un buen ejemplo de ello es Apple, o lo que hizo cuando vendió el primer iPhone: Apple ofreció a sus usuarios una experiencia que no tenían con un teléfono Nokia normal que, por otro lado, era finés y aburrido.

No obstante, la única y principal diferencia entre un mundo que vende productos y uno que vende servicios es que el propietario del servicio es el prestador del mismo. Esto significa que los clientes y las empresas comienzan una relación que, como todas las relaciones, necesita de actualizaciones constantes porque las preferencias cambian con el tiempo y se mueven constantemente, y esos movimientos exigen cambios, adaptaciones y reconfiguraciones. Además, las empresas tienen que conseguir que esta interacción mutua, este mutuo compromiso con el cliente esté automatizado.

Con el fin de embarcarse en este nuevo paradigma de relaciones individuales y automatizadas, las empresas tienen que aprender a escuchar y conseguir que el cliente se suba a bordo. Tienen que escuchar porque necesitan automatizar la relación entre el cliente y el servicio con tecnología informática y crear muchos datos que necesitan ser ajustados constantemente. Además, tienen que involucrar a los clientes para que ellos mismos hagan el trabajo: ellos compran el teléfono en la tienda *online*, lo actualizan y descargan las aplicaciones.



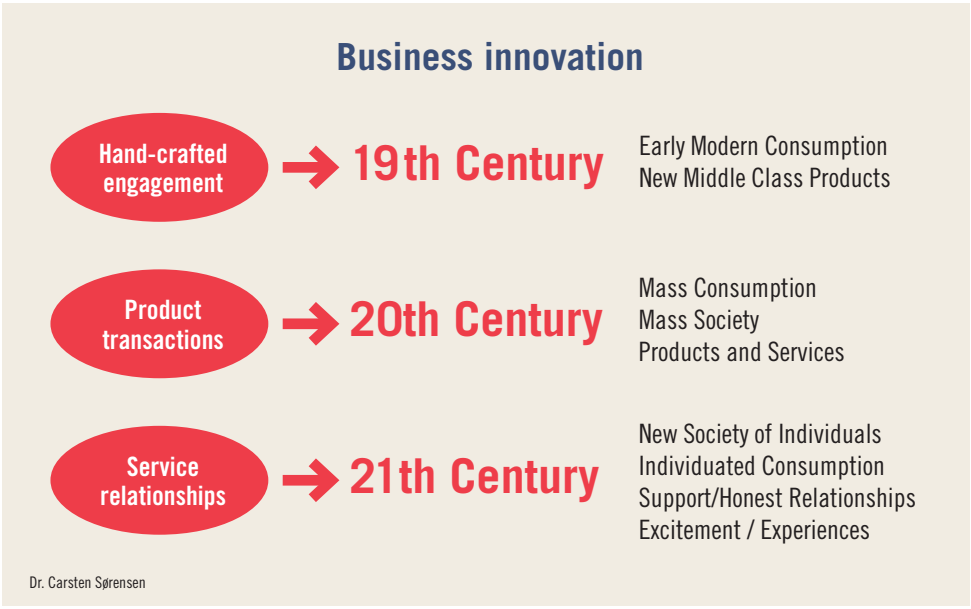
Facebook, por ejemplo, es el segundo país más grande del mundo, solo superado por China en apenas 400.000 personas, y podría desaparecer de la noche a la mañana si los usuarios dejaran de utilizarlo. Somos nosotros los que mantenemos Facebook vivo, aunque solo seamos esclavos de una gran maquinaria.

No obstante, en cualquier tipo de relación, se necesita confianza y, tras los políticos, las empresas comerciales son las entidades en las que menos se confía en el mundo. Esto ocurre, en parte, porque en las grandes empresas es imposible controlar las plantillas y los procesos, y también porque son muy rápidos a la hora de vender cualquier cosa pero muy lentos cuando hay que resolver un problema con el servicio que se ha comprado. Además, también es un problema tener que proporcionar información personal a las empresas para que actualicen sus datos.

Máquinas

Una de las cosas que nos diferencia de los primates es que construimos herramientas, y a lo largo de nuestra historia, hemos pasado por tres diferentes eras relacionadas con las máquinas: la era de la máquina mecánica, la era de la máquina inteligente y la era de la máquina generativa.

En la era de la máquina mecánica em-



En la nueva sociedad de individuos del siglo XXI, los servicios han reemplazado a los productos.

pleamos el potencial generativo del fuego para crear un desequilibrio de mercado a través de la automatización, que llevó a la Revolución Industrial. En aquel tiempo, la intensidad del capital creó mercados de capital y una integración vertical que aseguró una utilización máxima. Las máquinas también permitían y requerían la diversificación que aseguraba la utilidad de una inversión. Esto fue un gran éxito.

Durante la década de los años 50 y 60 entramos en la era de la máquina inteligente. El nuevo fuego era la información, así que la capacidad de acceder a la información y gestionarla mejor que los compañeros o la competencia ofrecía una ventaja competitiva y creaba un desequilibrio de mercado. Con el fin de lidiar con ello y gestionar la complejidad de tal sobrecarga de información, la estrategia a seguir fue la modularización, dividirlo todo en módulos para distribuirlo globalmente de una manera coordinada que, finalmente, llevó a la aparición de las jerarquías, la tercerización, la virtualización y, finalmente, a las redes de valor globales.

En el proceso de gestionar la información sucedió la digitalización. Tuvo consecuencias muy relevantes en las tecnologías de almacenamiento, procesamiento y distribución de información, que llevaron a la destrucción de productos e industrias. Y condujo a la era de la maquina generativa, una edad basada en el fenómeno de la máquina de "lo que sea", el nuevo fuego. Esta máquina de "lo que sea" es reprogramable y sus usos no vienen anticipados por sus inventores, como las aplicaciones no fueron anticipadas por Steve Jobs o el inventor de Android. La máquina de "lo que sea" permite la separación de la forma y la función, y de los contenidos y los medios, mientras que al mismo tiempo contribuye a una innovación de distribución global.

Materiales

Los grandes momentos del desarrollo de la humanidad pueden asociarse con un ma-

Age of the generative machine

- **The Turing/von Neumann Anything Machine (text your smartphone)**
- **Flexibilty reprogrammable machine is the new fire**
- **Digital technology intensively interconnected**
- **From looking at the interface to residing "within the machine"**
- **Procrastinated binding and generativity (Apps and services even Steve Jobs could not have imagined)**
- **Separation of form and function of reprogrammable universal machine**
- **Separation of contents and media**
- **Globally distributed contribution of innovation**
- **Deconstruction of products and industries (Tower Records and Blockbusters anyone?)**

Dr. Carsten Sørensen

A lo largo de nuestra historia hemos pasado por tres diferentes eras relacionadas con las máquinas: la era de la máquina mecánica, la era de la máquina inteligente y la era de la máquina generativa.

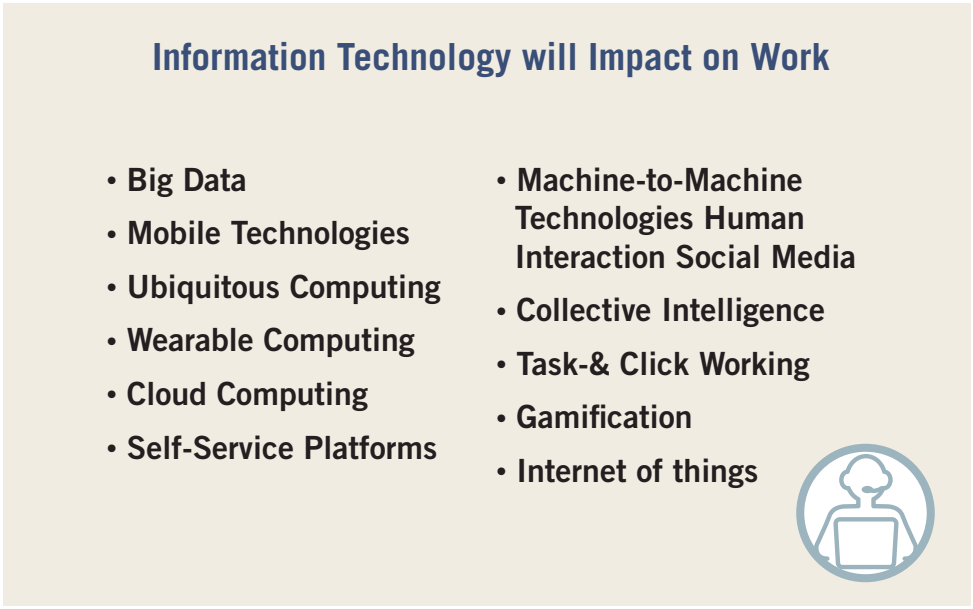
terial particular: la Edad de Piedra, la Edad de Hierro, y la Edad de Bronce. Recientemente hemos superado la edad del plástico, un material barato que puede tomar casi cualquier forma y puede ser distribuido democráticamente para cualquier propósito para el que sea necesario. No obstante, desde que podemos digitalizar las cosas, el plástico parece piedra. Los ceros y los unos son tan flexibles que no llegamos a entenderlos del todo. Como declaró una vez: Donald Rumsfeld, secretario de Defensa de Estados Unidos, "existe aquello que conocemos y que sabemos que conocemos, aquello que desconocemos y sabemos que desconocemos; pero también está aquello que no sabemos y que ignoramos que desconocemos".

En la era digital, la clave de la máquina generativa es que no sabemos lo que no sabemos: después de estudiar la Internet desde 1993, nunca he sido capaz de predecir más de un par de años y, aun así, me he equivocado.

En la era del material digital, el *Big Data* es grande porque es grande en términos de volumen, velocidad, variedad y veracidad, y es siempre grande si no se puede procesar. Sin embargo, el *Big Data* no es nada sin el *big code*, y uno de los mayores problemas en las discusiones sobre *Big Data* es que el *big code* no se menciona, porque las grandes empresas como Google o Microsoft se basan en un modelo de negocio centralizado mientras que el *big code* se basa en la descentralización.

Innovación

Existe un gran debate entre dos grupos diferentes con dos diferentes puntos de vista. Por un lado están los que dicen que estamos al final del camino. Tylor Cowen ya ha declarado que "el crecimiento de Estados Unidos está formado por los frutos maduros de personas jóvenes y energéticas que emigran a un nuevo continente y a industrias con una gran dependencia en nuevas tecnologías". Pero este ya no es el caso, la tecnología cada vez está más distribuida a lo largo del planeta.



Tecnologías de la información que impactan en el trabajo y la empresa.

Por otro lado, existen puntos de vista como el de Erik Brynjolfsson y Andrew McAfee, del MIT, que han escrito un libro que ha influenciado mucho el debate público: *La segunda era de las máquinas*. En él argumentan que la innovación es inherentemente recombinante y que ahora estamos mejor equipados para recombinar de lo que hemos estado nunca: "la innovación recombinante permite la abundancia computacional a través de la distribución global y el aumento de recombinaciones beneficiosas", aunque, como Ray Kurzweil también señaló, la recombinación se está volviendo demasiado compleja como para entenderla de manera intuitiva.

Existen varios ejemplos recientes de innovación a partir de la recombinación, como el mapeo 3D de Google, que incluso será integrado en teléfonos móviles, o la impresión 3D, que permitirá una enorme generación de innovación y distribución de fabricaciones de cuyas consecuencias totales no somos todavía conscientes. Esta idea de la robótica formando parte del día a día de la industria está acaparando cada vez más miradas y depende de nosotros decidir qué combinaciones son buenas y cuáles no lo son. Las empresas no necesitan analizar el *Big Data* para eso.

Una consecuencia natural de lo expuesto es que las máquinas generativas se utilizarán cada vez más para permitirnos trabajar para las empresas con las que estamos asociados. Ahora el cliente prefiere comprar y pasar por una caja automática o ir al cajero en vez de ir al banco, buenos ejemplos de la distribución del poder: en la era de la máquina inteligente, el ordenador nos servía a nosotros; ahora, en la era de la máquina generativa, nosotros servimos al computador.

Hemos pasado de la era de la máquina inteligente, que trataba de encuentros, a la era de la máquina generativa, que trata sobre relaciones. El siglo XX se definió para gran parte del mundo occidental por permitir la compra de productos de lujo por casi nada, entonces, la informática apoyaba a las per-

sonas: había producciones optimizadas, relaciones a través de las transacciones, intimidad a través de conexiones anónimas remotas, y la gestión científica del trabajo obrero.

Futuro

El siglo XXI trata de servicios de alta calidad individualizados por casi nada, en los que el cliente se involucra para conseguir relaciones automatizadas codificadas con una tecnología que permite una obligación aplazada y una actitud de auto-servicio. Ahora las personas apoyan a la informática y, en el futuro, solo habrá dos tipos de trabajos: los trabajos en los que inventamos nuevas relaciones automatizadas, y los trabajos en los que ayudamos a otras personas cuando estas primeras fallan, porque fallarán muchas veces.

La innovación tiene muchos desafíos por delante, y la privacidad es uno de ellos. Con Google tenemos un modelo que está roto porque está centralizado y necesitamos descentralizarlo previamente en su diseño, pasando de una Gestión de Relaciones con Clientes (CRM) a una Gestión de Relaciones con Vendedores (VRM) controlada por el cliente. No tenemos ni idea de cómo hacerlo, pero ocurrirá. Y si no, no innovaremos.

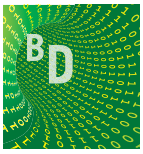
También tenemos que aprender a olvidar, saber cómo gestionar una producción distribuida muy orgánica de códigos y datos, y mantener un *middleware* abierto y estándares de infraestructuras para las plataformas propietarias y las plataformas de colaboración distribuidas y generativas.

A una escala mayor, ¿cómo creamos servicios atractivos? Contamos con tecnología que supuestamente nos ayuda a hacer las cosas, pero también tiene que ofrecer emociones y diversión, y ayudarnos a crecer como personas: un gran reto que la televisión ha intentado durante años.

Una de las cosas que seguro que va a ocurrir es que aparecerán muchos nuevos trabajos. Aprenderemos a ayudar a las máquinas, y no al contrario, y nuestros trabajos serán diferentes, serán más intensos pero más flexibles, y habrá una gran polarización de habilidades y trabajos para la que tenemos que prepararnos.

Big Data, *economía y organizaciones*

Por Daniel Villatoro
BBVA Data & Analitycs



En 2007 el editor de la revista WIRED, una de las publicaciones más conocidas en el campo de la tecnología, dijo entonces que el *Big Data* anunciaba el final de la teoría científica; teoría que años más tarde tuvieron que salir a defender científicos como Massimo Pigliucci, profesor de Ética y Filosofía de datos en la Universidad de la Ciudad de Nueva York (CUNY), para indicar que la Ciencia, a diferencia de la publicidad, no trata de encontrar patrones, sino de encontrar las explicaciones que producen esos patrones. En el Data Analytics BBVA no estamos solo preocupados en el *Big Data*, sino en las grandes preguntas que esconde el *Big Data* y queremos centrarnos realmente en las cuestiones científicas. Es útil saber a qué clientes dirigirse pero nosotros queremos entender mejor al cliente. Las preguntas científicas que nos hacemos como científicos son: ¿Cómo tomamos decisiones los humanos?, ¿por qué? y ¿cómo estas decisiones podrían afectar eventualmente a la eficiencia de nuestro negocio como banco, cómo ahorrar costes o dar un mejor servicio?

Mi experiencia en investigación se centra en tres áreas científicas: la Ciencia cognitiva, que es cómo los humanos tomamos decisiones; la Economía experimental, o cómo los individuos toman decisiones que afectan a sí mismos y a otros; y las redes sociales enfocadas al nivel de interacción –no Facebook ni Twitter sino redes sociales teóricas– y cómo el entorno y las relaciones afectan a

¿Es el Big Data *el final de la teoría científica? Definitivamente no. Necesitamos gente preparada para que, aunque tengan las herramientas de* Big Data *que responden todas las preguntas, realmente sepan hacer las preguntas adecuadas y las sepan resolver de manera válida*

las decisiones que tomamos. Aunque actualmente a esta disciplina se la denomina "Data Science", centros de investigación reconocidos como Harvard, Yale o Microsoft, están empezando a acuñar el término "Algorithmic Economics", atendiendo a esa parte de la generación de algoritmos que está muy centrada en la Economía.

El método científico se basa principalmente en la formulación de una serie de preguntas sobre las que se realiza una investigación y se construye una hipótesis para después llevar a cabo experimentos para probar esta hipótesis, analizar los resultados y extraer conclusiones, y, finalmente, redactar un informe de los resultados, y volver a empezar. Cualquier investigación científica es cíclica, siempre produce más preguntas científicas que van abriendo el camino al conocimiento.

Por ejemplo, en 2001 Fehr y Gächter, dos economistas experimentales, publicaron en Nature un artículo titulado "*Altruistic punishment in humans*" (El castigo altruista en humanos). Básicamente, su hipótesis es que el castigo altruista, que se da cuando un sujeto gasta recursos para castigar a otro sujeto cuando no hace bien las cosas, hace que la sociedad se sostenga en un entorno donde no hay una regulación central. Muchos economistas conocen el dilema del prisionero, un dilema en el que hay dos sujetos separados en salas distintas que no se pueden comunicar entre ellos. A los dos se les acusa de haber robado y tienen que decidir si confesar, y traicionar a su compañero, o no confesar. Esta situación produce cuatro posibles escenarios: si los dos se inculpan el uno al otro, los dos estarán condenados a cinco años de cárcel; si uno se mantiene callado y el otro lo inculpa, éste saldrá libre y el primero recibirá 20 años de cárcel; si los dos se mantienen callados, ambos estarían sólo un año en la cárcel. La mejor opción para ambos sería cooperar y no confesar para reducir la pena al mínimo, pero siendo egoístas, la mejor situación siempre es intentar salir libre directamente, asumiendo que el otro va a ser pro-social contigo y no va a confesar. Fehr y Gächter descubrieron un gran resultado a nivel científico: que si se añaden varias rondas a este juego y se permite que los sujetos se castiguen entre ellos, la posibilidad de que se produzca ese castigo aumenta la cooperación media. De hecho, este artículo ha generado un gran impacto científico y ha generado muchas otras investigaciones de gran interés en este mismo campo.

Los economistas experimentales se sirven de unos mandamientos que deben cumplirse en todos sus experimentos para poder ser publicados: tiene que haber dos tipos de incentivos económicos para los sujetos que participan en el experimento, tanto por asistir como por la calidad de la participación, porque dependiendo de cómo participes,

Estudio realizado por BBVA en el que se analiza el uso de las tarjetas de crédito en España durante la Semana Santa de 2011 en cuatro sectores: mercados y alimentos, bares y restaurantes, moda y gasolineras.

uno se lleva más o menos dinero; no se puede engañar a los sujetos y todas las condiciones experimentales y todas las posibles situaciones tienen que estar perfectamente descritas, no puede haber actores en el experimento, tiene que ser anónimo, y tenemos que garantizar que el sujeto que asiste al experimento no vive de asistir a experimentos y no se conoce ya todas las reglas y todos los trucos posibles.

En el experimento de Fehr y Gächter contaron solo con 240 estudiantes, todos de las universidades de Zúrich y de la Escuela Politécnica de Zúrich, con un porcentaje de mujeres del 31%, y analizaron 2.800 interacciones durante 10 sesiones, con 24 sujetos por experimento. Aunque el conjunto de datos era pequeño, aseguraron una gran significancia estadística en un entorno altamente controlado, como son estas salas de experimentos donde los sujetos se sientan y no pueden ver las decisiones de los demás y saben que, además, están siendo observados.

Obviamente, esto no es la representación de la naturaleza humana y no supone una selección de una muestra que sea representativa de la humanidad para poder hacer este tipo de afirmaciones, porque el mundo no son 240 estudiantes de un país con un alto nivel económico y una cultura pro-social altamente definida, y donde el 31% de los sujetos son mujeres. Sin embargo, el experimento está muy bien defendido a nivel científico gracias a los mecanismos de Peer Review, en los que un comité de sabios seleccionado por la revista en la que se quiere publicar evalúa la validez del trabajo antes de publicarlo. Sin embargo, un ejemplo de la mala influencia del Peer Review es un artículo publicado en Nature en el que revela que en 47 de 53 artículos específicos de un área de investigación concreta los resultados no se pueden reproducir, lo que en el mundo científico es una aberración.

Y aquí viene nuestra aproximación a *Big Data*. Básicamente, los datos se capturan para probar una hipótesis sobre el comportamiento humano, siguiendo un protocolo muy estricto que asegura que no haya externalidades y que realmente estamos probando ese comportamiento. Las preguntas que se hace cualquier economista experimental son, por ejemplo: ¿cómo se recluta el sujeto?, ¿cómo participa en el experimento?, ¿cómo son las decisiones experimentales y cómo afectan a la decisión real?, ¿cómo se incentiva a un sujeto de la manera adecuada para que se comporte de una manera fidedigna con respecto al comportamiento real que queremos observar?, etc. No obstante, nosotros en BBVA tenemos una aproximación diferente a la captura de datos. Tenemos un mundo lleno de interacciones. Nuestro mundo es muy rico e incontrolado, no tenemos una sala experimental donde podamos ver cómo se comportan los sujetos. Nosotros ya tenemos sujetos que actúan a diario y hacen miles de transacciones a distintos niveles de agregación, con distintos tipos de interacción, etc.

Un ejemplo de esto sería el vídeo "*SPRING SPREE –Spending patterns in Spain during easter 2011*", disponible en YouTube, que, aunque sea solo un simple ejemplo de visualización, da una idea de la amplitud de nuestros datos. En él se pueden observar todas las transacciones que se realizaron en España, con tarjetas o TPV del BBVA, durante la Semana Santa de 2011. En él se puede ver claramente qué días eran festivos en Cataluña pero no en el resto de España y viceversa, cómo la gente sale a cenar a bares y restaurantes por la noche, o las transacciones en gasolineras, que nos revelan los patrones de comportamiento de la gente y hacia donde se dirigían –en este caso, especialmente a la costa–.

Muchos han hablado de *Big Data* centrándose en el volumen de los datos. Es el caso de empresas como Google, Facebook, o Ebay, o el propio CERN, que manejan grandes volúmenes de datos; otros se centran en la variedad del dato, pero nosotros, además de fijarnos en el volumen y la variedad, contamos con datos muy ricos, grandes y "largos" (*long data*). Tenemos más de 26 millones de transacciones al día con distintos tipos de interacción, desde transacciones con tarjeta de crédito, transacciones de envíos de dinero de un sitio a otro, o extracciones de dinero en cajeros. Y a diferencia de los experimentos que hacen en Zúrich, encerrados en una sala donde el sujeto está observado, lo nuestro es el comportamiento real. Si uno se gasta 20€ en un determinado negocio es porque realmente uno quiere hacerlo, es la vida real y uno no se preocupa de si está siendo observado o no.

Esto lleva siempre al clásico discurso de "muy bien, lo estáis observando porque la gente se gasta dinero, pero en Facebook también podemos ver lo que a la gente le gusta y lo que no". Y nuestra posición ante esto es siempre la misma: en Facebook la gente solo habla de intenciones. Decir que quieres que llegue agua a Uganda no quiere decir que te estés jugando el pellejo por poner agua en Uganda. Decir que te gusta Bon Jovi puede ser un comportamiento que solo se produce por la presión social de tu grupo, porque queda bien decirlo. En nuestro caso, no. En nuestro caso, si te gastas 20€ en un disco de Bon Jovi o 10€ en una donación de agua a Uganda, sabemos que realmente estás interesado.

Esto lleva al famoso experimento de Facebook que últimamente está recibiendo muchas críticas y que se basa, principalmente, en observar si los usuarios que han recibido comentarios negativos de sus amigos, también publicaban mensajes negativos; y lo mismo con mensajes positivos. Esto es obvio. Naturalmente, si uno entra en Facebook y ve que un amigo suyo ha anunciado el fallecimiento de un familiar, ese día uno no pone una foto de "qué buena ha estado la paella que me he comido a mediodía", por respeto, porque probablemente no lo considere como una norma de buena educación. El resultado pasa a tener un alto grado de validez estadística porque la variación que han observado es mínima, siendo que el efecto que tiene sobre los usuarios es de 1/20. Es decir, una palabra negativa afecta que uno tenga 1/20 de probabilidades de escribir una palabra negativa. Por lo tanto, el resultado es importante, pero no tanto. Sin embargo, otro experimento de Facebook más significativo es el que realizó en las últimas elecciones estadounidenses, donde consiguieron movilizar a 61 millones de personas. Lograron observar si la presión social provocaba que la gente fuera más a votar o no. Es decir, preguntaban si ese día habías ido a votar. Los participantes estaban incluidos en dos grupos de control: en el primero, te mostraba si tus amigos habían ido a votar; en el segundo, esta información no se enseñaba. El resultado fue que en los Estados donde incentivaron el experimento, aumentaron el grado de votación en medio punto.

En BBVA, la aproximación científica que utilizamos es el método científico clásico, pero siempre centrándonos en el dato.

Tenemos que ser conscientes del dato que tenemos y hacernos preguntas en base a este dato e investigar en relación al mismo; estudiamos cómo han aproximado el problema otros científicos; construimos hipótesis, pero siempre pensando en el dato; probamos y hacemos experimentos, porque ya tenemos el dato capturado; analizamos los resultados y extraemos conclusiones sobre nuestro dato; finalmente, hacemos un informe de los resultados de los datos que hemos analizado. Como banco, nuestro negocio está en la reputación. Obviamente nos preocupamos mucho sobre la privacidad de los datos y en mantener el anonimato de los datos personales privados a los que tenemos acceso como proveedores de un servicio.

En BBVA sabemos las transacciones que se realizan con tarjetas y sabemos dónde se han realizado, por lo que podemos comparar el funcionamiento de ese negocio con respecto a otros negocios equivalentes en las proximidades, y evaluar la fidelidad del cliente y otros factores. Usando estos indicadores podemos construir una evaluación de riesgos que nos indique cómo de bien o de mal va cada negocio en cada zona de España y construir un mapa de riesgo para saber cuáles son las zonas más o menos arriesgadas de toda España, siempre basándonos en datos anónimos y agregados, y de los comercios. A nivel local podemos observar cómo, por ejemplo, consumen los ciudadanos de Sant Cugat y descubrir que las personas de Sant Cugat consumen poco en bares y restaurantes dentro de la ciudad, pero sí que lo hacen fuera, y deducir a partir de este indicador que hay que abrir un bar en Sant Cugat.

Si además podemos observar cuál es el ritmo medio de las transacciones, lo que denominamos "el pulso de la ciudad", podemos observar realmente cómo se comportan los ciudadanos en semanas normales o en semana en las que hay eventos. Esto nos lleva a una nueva herramienta que estamos desarrollando de detección de eventos. El año pasado publicamos un informe sobre los patrones de consumo de los turistas en Madrid y Barcelona usando *Big Data*, y uno de los resultados era que en Madrid lugares como, por ejemplo, el Museo Reina Sofía o el Mercado de San Antón son más interesantes para los visitantes que para los ciudadanos locales. Podemos llegar al extremo de saber realmente cómo impacta económicamente un evento dentro de la ciudad. Por ejemplo, hasta hace no mucho nadie sabía cuánto dinero llegaba a la ciudad de Madrid por la Fiesta del Orgullo. Todo eran datos aproximados porque realmente nadie tenía acceso a toda la información, pero nosotros hicimos un experimento con nuestra red de sensores distribuida (TPV) y pudimos entender un poco mejor cómo afecta el evento a las zonas convocadas y los alrededores cercanos, observar el consumo en distintos niveles de profundidad, y comparar de manera cuantificable el impacto económico de la Fiesta del Orgullo de 2011 con la de 2012.

La aplicación sobre eventos con la que gané el Innova Challenge de BBVA, un concurso en el que el banco liberó diferen-

tes tipos de datos reales para que los desarrolladores creásemos aplicaciones, se desarrolló en el Centro Tecnológico Barcelona Digital. BBVA liberó una serie de datos con una serie de requisitos que si no se cumplían impedían la publicación de ese dato concreto. Nosotros decidimos hacer correlaciones con información de Twitter, utilizando un algoritmo de detección de anomalías, del diferencial de gasto de una semana normal con respecto a la semana de un evento determinado –como puede ser un partido del Barça o el día de Sant Jordi–, para evaluar el impacto económico que tiene dicho evento en la ciudad.

Todos estos datos son muy interesantes y se pueden hacer muchos análisis con ellos, pero además tenemos que poder comercializarlos, y un punto clave de información que tenemos es el punto de venta. Cada negocio puede saber mucho sobre sus clientes, su zona y su sector; pero nosotros conocemos una parte del total, y esto nos lleva a poder dar respuestas a otros niveles en cuestiones como, por ejemplo, servicios de diversificación, venta cruzada (*cross selling*), nuevos productos, nuevos patrones de horarios, expansión de objetivos comerciales, o estudios de la lealtad de los clientes. Por ejemplo, sabemos cuál es la distancia media recorrida por los clientes de un cierto negocio en Madrid para realizar una compra. Además también podemos entender cuál es el *wallet share* (la cantidad que un cliente se gasta en negocios de una cadena por cada 100€ que gasta el cliente en un determinado negocio de esa misma cadena) para poder observar la fidelidad de ese cliente.

En definitiva, tanto los experimentos con *Small data* como aquellos con *Big Data* presentan sus problemas. Los primeros, como se veía en el experimento de Fehr y Gächter, son muy caros y requieren un buen diseño de la plataforma experimental para cumplir con todos los requisitos científicos, y además ofrecen una visión limitada y posiblemente desviada de la realidad. Por otro lado, actualmente es muy costoso entender todas las tecnologías y aplicaciones del *Big Data*, y hay pocas personas que lo hagan. El dato de por sí tiene fallos y está desviado. En los estudios con *Big Data* también hay que tener en cuenta la "*paradoja de Simpson*", que básicamente dice que si dentro de un gran conjunto de datos solo observamos una variable, puede que estemos perdiendo otra variable dentro de ese propio dato, generando relaciones espurias o conclusiones erróneas. Hay ejemplos clásicos como el de que la introducción de Internet Explorer disminuyó el tráfico de piratas marítimos en el mundo. Asimismo, en ambos casos existe el problema del *p-hacking*, del que la comunidad científica no suele ser muy consciente, pero que es muy importante porque implica que hay gente que cambia las condiciones experimentales durante el experimento para poder tener una validez estadística que sea representativa y aceptada por la comunidad científica, y poder publicar, ya que los científicos cada vez viven más de las publicaciones. A día de hoy, incluso algunas revistas quieren que se publique el *data set* junto a los análisis científicos y el código, pero eso es algo que entidades como BBVA no puede hacer, porque compartir los datos de los clientes es inviable.

¿Es el *Big Data* el final de la teoría científica? Definitivamente no. Necesitamos gente preparada para que, aunque tengan las herramientas de *Big Data* que responden todas las preguntas, realmente sepan hacer las preguntas adecuadas y las sepan resolver de manera válida. También necesitamos científicos que practique la *p-diligence* y no falseen sus experimentos: en BBVA aplicamos un TDD (*Test Driven Development*) en el que antes de comenzar el experimento planteamos cuáles son las condiciones de aceptación o rechazo del mismo. Por último, es esencial asegurar la privacidad de los sujetos y que las personas que trabajamos con *Big Data* seamos éticamente conscientes de todo lo que puede generar si esto no se cumple.

Big Data
y análisis predictivo

Por Esteban Moro
Universidad Carlos III de Madrid

Cuando la gente habla del *Big Data*, uno siempre se pregunta de qué tipo de *Big Data* estarán hablando porque, por ejemplo, la gente que trabaja finanzas y en bolsa lleva utilizando Big Data desde hace decenas de años y hoy, cuando las operaciones se reproducen con ingeniería algorítmica casi en milisegundos, todavía más. Pero el *Big Data* no solo está definido por el volumen, sino también por las variables y la velocidad del dato. Hace ya mucho tiempo que las empresas utilizan programas de ERP (*Enterprise Resource Planning*) y CRM (*Customer relationship management*), que manejan grandes cantidades de datos. Quizás, con la web 2.0, ahora tenemos un acceso a otro tipo de datos que no teníamos antes, como son los *weblogs* o las plataformas de ecommerce, pero en los últimos años han aparecido todos los tratamientos de estos datos no estructurados para los millones de vídeos que se suben todos los días a YouTube sobre diferentes temas, pero también, por ejemplo, para la predicción del tiempo.

En un minuto se habrán publicado en Twitter aproximadamente 300.000 *tweets*, lo que supone más de 64.000 líneas de Excel, pero el ritmo al cual se están publicando, el hecho de que sea un dato no estructural, o que haya que convertir mediante un procesador de lenguaje natural, a una opinión, a una queja, a un sentimiento, hace que ese tipo de datos se conviertan en *Big Data* por su volumen. Por ejemplo, cada vez que pa-

Predecir no es decir lo que va a suceder con más probabilidad, sino identificar cuál es el riesgo de que sucedan cada uno de los elementos

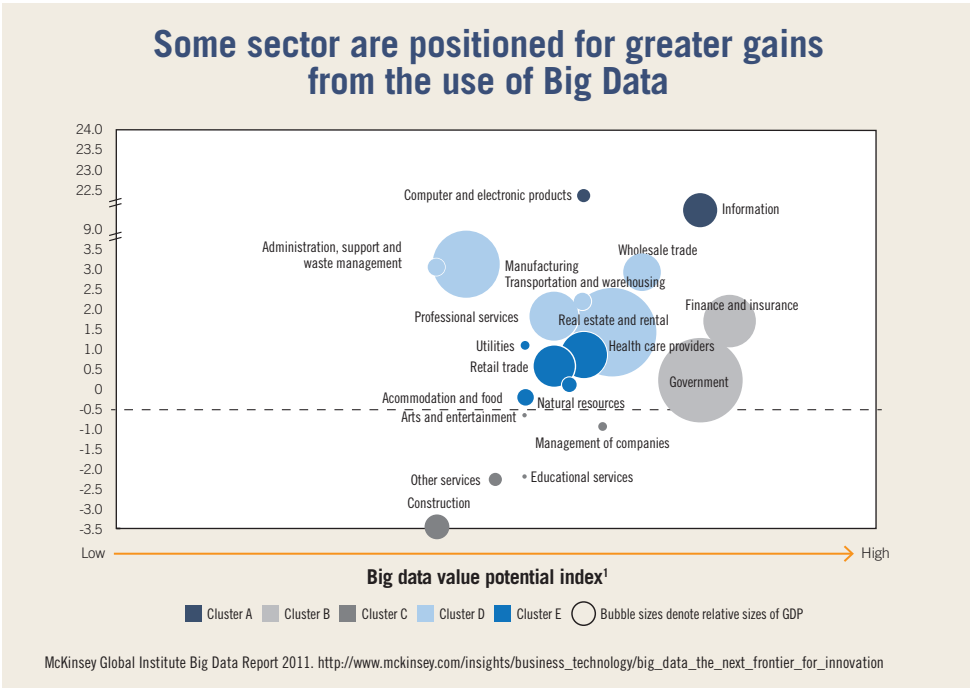
samos una tarjeta de crédito por una TPV, en el tiempo que transcurre entre que uno escribe el pin y se acepta la operación, hay una empresa que se conecta a una base de datos de 4.500 millones de transacciones, de millones de clientes, y que indica que esa transacción no es fraudulenta y que la operación está aceptada. El volumen de datos es de apenas gigabytes, pero la velocidad a la cual hay que realizarlo conlleva la utilización de una serie de tecnologías, una serie de algoritmos que constituyen para nosotros un problema de *Big Data*. Por tanto, para mí el problema del *Big Data* no es solamente un problema de volumen, sino sobre todo de velocidad y de variedad.

¿Por qué se habla ahora de *Big Data*? Quizás, no solamente porque cada vez tengamos más datos, que también, sino porque han confluido una serie de factores que han hecho que haya este interés por la analítica de datos. En primer lugar, han surgido una serie de tecnologías que permiten guardar todo ese tipo de datos y acceder a este gran volumen de datos en unos tiempos accesibles desde el punto de vista de la aplicación. Incluso en las aplicaciones en memoria, uno puede acceder casi en milisegundos. Por otro lado, están las herramientas de análisis. Hoy en día, a la vez que uno puede guardar los datos, existe también la posibilidad de tener herramientas que permiten el análisis casi en tiempo real. También existen proveedores como SAS, Revolution Analytics, SPA, Storm, S4 o Massive Learning que permiten todo este tipo de análisis. No obstante, lo más importante es que desde hace unos años hay gente que lo sabe hacer y ahora hay muchísimos sitios donde uno puede aprender a ser lo que actualmente se denomina como *Data Scientist*. Son personas que tienen capacidades transversales, que no solamente tienen formación en, por ejemplo, algoritmos, sino también saben de *hardware* o de *software*.

El flujo del valor

Lo más interesante del *Big Data* es el flujo de valor. Por ejemplo, en una empresa existen desde hace muchísimos años transacciones operacionales, pero se utilizaban solo en un departamento, hasta que se dieron cuenta de que en otro departamento esos mismos datos podían tener otra aplicación y un valor, incluso podían ofrecer servicios a terceros para que otras empresas utilicen los datos en sus procesos. En España tenemos la suerte, por ejemplo, de contar con dos grandes empresas que son pioneras en esto, Telefónica y BBVA, que han creado este tipo de servicios para que otros accedan a esos tipos de datos. Pero además, las empresas se han dado cuenta de que existen otras fuentes de datos abiertos, como las redes sociales o los servicios de meteorología, que pueden integrarse en los procesos de las empresas. Este contexto permite que, además, surjan nuevas fuentes de datos, nuevos sensores que generan datos sobre todo lo que está pasando. Empezamos a tener muchísimos datos de otro tipo, que influyen muchísimo en la gestión de campañas de *marketing* o en los motores de recomendación.

Pero lo principal de esta nueva realidad es que han aparecido tres tipos de datos que no existían antes y que, además, son precisamente los que condicionan todas nuestras acciones. El 70% de nosotros pregunta a otras personas cuando va a comprarse un producto electrónico. Es decir, los humanos somos virales y tendemos a comunicar en nuestras redes sociales, tendemos a preguntar. El otro condicionante, de los más grandes que hay en nuestras acciones, en nuestro comportamiento, es la movilidad geográfica. Somos animales de costumbres, nos move-

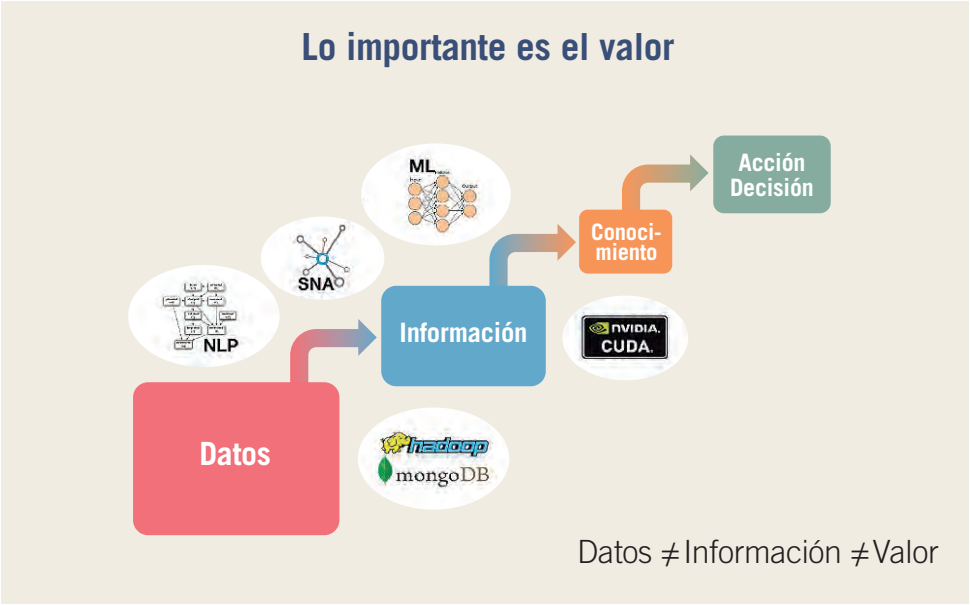


No todos los sectores empresariales se ven afectados de la misma forma por el *Big Data*. Los relacionados con la información, la venta minorista o las finanzas, incluso la Administración pública, están más predispuestos a que el *Big Data* pueda cambiar muchos de sus procesos.

mos siempre por los mismos sitios y la mayoría de nosotros, a lo largo de un mes, solo va a diez tiendas diferentes. Esto es así porque nuestra vida tiene ciertos condicionantes que limitan nuestros comportamientos a lo largo del día. Y este tipo de datos empieza a estar disponible. Por tanto, la revolución del *Big Data* tiene una de sus bases en la visión que tenemos desde que existen este tipo de comportamientos, de patrones, cuando utilizamos Facebook, Twitter o nuestros teléfonos móviles. La transfusión de esos tres tipos de datos –los que nos llegan desde las redes sociales, los datos de comportamiento, y de movilidad geográfica– es lo que ha llevado al *boom* de las aplicaciones que tienen que ver con el análisis de este tipo de datos y del *Big Data*.

No todos los sectores empresariales están igualmente afectados por este fenómeno. Por ejemplo, aquellos relacionados con la información, la venta minorista o las finanzas, incluso la Administración pública, están más predispuestos a que el *Big Data* pueda cambiar muchos de sus procesos. Sin embargo, hay otros, como la construcción, donde no se da tal predisposición.

Básicamente la cuestión no es cómo de grande o cómo de veloz es un dato, sino qué valor tiene y qué valor puede crear en una organización. Desde el punto de vista general, dentro de una organización, adoptar una analítica basada en datos y en una visión analítica de los procesos que ocurren en la compañía puede crear transparencia y reducir ineficiencias, porque a veces podemos tener un dato que nos puede ayudar a detectar que algo no se está haciendo bien, pero sobre todo puede permitir la experimentación. Teniendo este gran volumen de datos, uno puede introducir, por ejemplo, la variabilidad y el rendimiento de ciertos procesos en una plataforma de ecommerce. Esto es muy importante porque el análisis



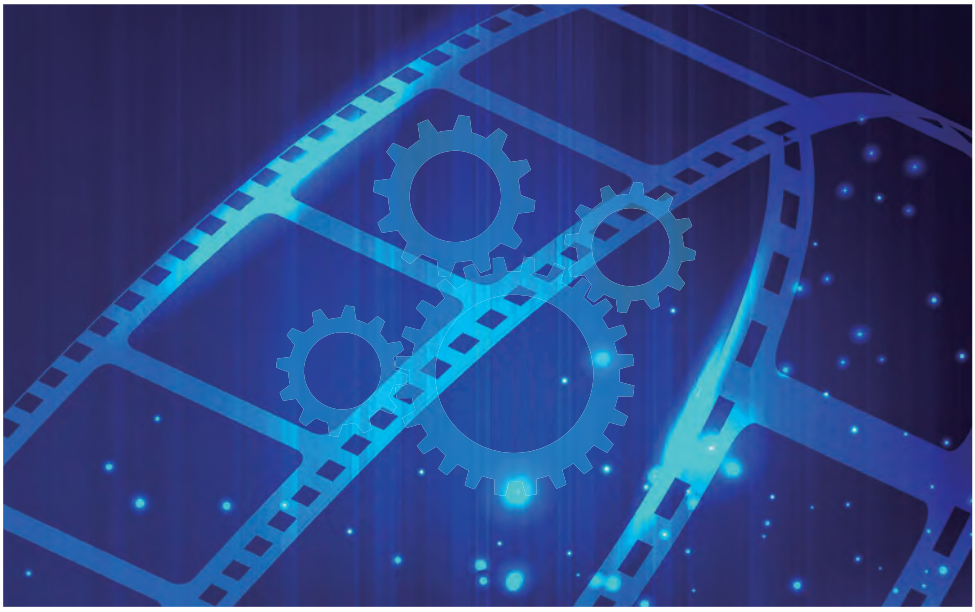
La cuestión no es cómo de grande o de veloz es un dato, sino qué valor tiene y qué valor puede crear en una organización. Y definir el proceso que permita transformar los datos en la toma de decisiones empresariales.

que se haga de los datos puede ayudar a los expertos o incluso se puede prescindir de ellos (como en el caso de las últimas campañas de Obama).

Muchas empresas como Netflix y Pandora se basan en modelos de recomendación que luego venden a otros proveedores de servicios o a empresas que utilizan otro tipo de servicios. Realmente, el valor de estos datos reside en poder llegar a generar predicciones. Cuando uno puede predecir algo, puede tomar decisiones, y es entonces cuando el dato se vuelve accionable y uno puede adelantarse a lo que va a pasar en el futuro y tomar decisiones antes de que suceda. O también puede utilizar el *nowcasting* para predecir lo que está pasando ahora sin tener datos de lo que está pasando en este preciso momento. Los modelos predictivos se aplican en la gestión y detección de fraude, en la gestión del riesgo, para mejorar campañas de *marketing*, en salud, en el mundo del deporte y las apuestas. También el *Big Data* se utiliza cada vez más en las Administraciones públicas, ya que por su naturaleza manejan muchísimos datos de todos nosotros y empieza a haber muchas posibilidades para utilizar sus aplicaciones. Por otro lado, la Ciencia lleva muchos años trabajando con *Big Data*, especialmente en el área de la Física de partículas. De hecho, muchas de las tecnologías que hoy utilizamos, como la Web o la Nube, nacieron en el CERN (el Laboratorio Suizo de Física de Partículas) porque el nivel de producción, de consumo y de análisis de datos en el CERN es brutal. El LHC genera muchísimos más datos que Facebook, Twitter y todas los demás plataformas sociales juntas por día. También, cada vez va a haber más interés en la analítica de datos de la Biología.

El dato en sí no es valor

Pero el dato en sí no es valor, ni siquiera es información, por lo que es necesario convertir ese dato en valor, en información. Luego hay que transformar esa información en conocimiento a través de series de algoritmos que nos permitan conocer información



que puede ser útil para mejorar campañas de *marketing*, predecir el gasto en una zona o en un sector comercial, identificar grupos de personas según el consumo que hagan de la información en Internet, o realizar análisis de sentimientos. Por ejemplo, nosotros contamos con un motor de análisis lingüístico, hecho por lingüistas computacionales y para el que se crean diccionarios específicos para cada tema, que puede servir a los anunciantes para saber exactamente dónde poner su publicidad.

Uno de los proyectos del laboratorio trata sobre cómo predecir el Twitter del futuro, averiguar si uno puede adelantarse en saber cuáles van a ser los contenidos más difundidos dentro de la siguiente hora. Esto es posible. Para ello hemos creado una serie de métricas que tienen que ver con la red social, con las que monitorizamos los *tweets* para intentar descubrir los factores y los grupos de influencia, y poder predecir en 45 minutos el alcance de un *tweet*. Esto no es ciencia ficción: esto se basa en una serie de modelos predictivos que tienen una serie de variables. En todo caso, hay una serie de peligros asociados al *Big Data* y a la predicción. Predecir no es decir lo que va a suceder con más probabilidad (es como predecir que en Abu Dabi no va a llover nunca y acertar con un 98% de probabilidad), sino cuál es el riesgo de que sucedan cada uno de los elementos.

Otra cuestión respecto a los modelos predictivos es que actualmente estamos construyendo nuestros modelos con datos de Twitter y Facebook. Mañana no habrá Twitter probablemente, en seis meses habrá cambiado la API, habrá cambiado la manera de gestionar la información, y los modelos que valen hoy no valdrán mañana. También hay que tener en cuenta que, por ejemplo, Twitter está muy sesgado demográficamente; y hay que comprobar las hipótesis creando modelos nulos, por ejemplo, para descartar factores. Asimismo, hay que hacer *A/B testing* para aislar algunos efectos o determinar cuáles son los efectos que causan lo que vemos en los modelos. No se puede dar por válido algo que funciona a la primera solo porque funcione. Y todo esto no es nada nuevo, al fin y al cabo los modelos predictivos están basados en el método científico, algo que nos ha funcionado durante 300 años y que ha hecho que el hombre llegue a la luna.

EL IMPACTO DEL *BIG DATA* EN LA EMPRESA

EL IMPACTO DEL

BIG DATA

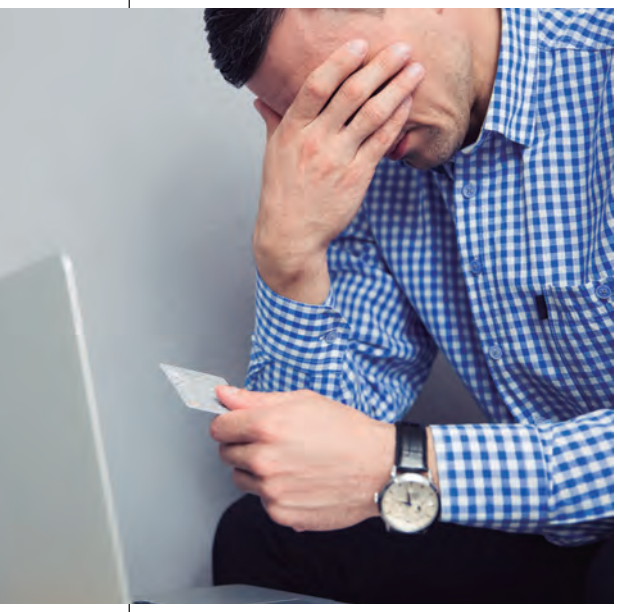
EN LA EMPRESA

INTRODUCCIÓN GENERAL

La tercera cita con el Big Data *contó en esta ocasión con la colaboración del Club Última Hora del Grupo Serra y la Universitat de les Illes Balears (UIB), junto a la Fundación Ramón Areces organizaron en noviembre de 2014 una jornada divulgativa sobre* El impacto del *Big Data* en la empresa*, que se celebró en el campus de la UIB. El acto fue inaugurado por el rector de la UIB,* ***Llorenç Huguet****; el consejero delegado del Grupo Serra,* ***Pedro Rullán****; el director de la Fundación Ramón Areces,* ***Raimundo Pérez-Hernández y Torra****; y el director general d'Educació del Govern Balear,* ***Miquel Deyá****.*

La jornada, que siguió un esquema similar a la celebrada en Madrid en el mes de julio, se enfocó hacia el estudio de algunos de los principales elementos que integran el *Big Data* en la Economía y la gestión empresarial. En la presentación, se destacó que el *Big Data* afectará a prácticamente todas las industrias y hará que muchos negocios cambien de modelo. Por ejemplo, permitirá que los servicios de una ciudad sean dimensionados en función de su demanda real; ayudará a mejorar las predicciones empresariales en todos los órdenes y procesos de negocio y ya está transformando áreas como el *marketing*, la publicidad y el comercio electrónico.

José García Montalvo, catedrático de la Universidad Pompeu Fabra y vocal del Consejo de Ciencias Sociales de la Fundación Ramón Areces, fue el encargado de abrir la jornada con la ponencia *El estado del arte del Big Data & Data Science y aplicaciones al sector financiero*. Para García Montalvo, la generalización del *Big Data* y las nuevas técnicas asociadas al tratamiento y análisis de grandes bases de datos está revolucionando tanto el trabajo científico como la gestión empresarial. Aplicaciones como las recomendaciones personalizadas de Amazon han supuesto una mejora muy significativa de la experiencia de compra de los consumidores. Analizó las posibilidades del *Big Data* para mejorar los servicios financieros y la experiencia de los clientes. La utilización de técnicas de *Big Data*, incluidos indicadores de reputación y capital social *online* –recordó– se ha extendido a la calificación crediticia de los solicitantes de



crédito, la detección del fraude en tarjetas, la microsegmentación, los servicios de información a los clientes, del cumplimiento normativo y la prevención del blanqueo de capitales y operaciones de financiación de actividades terroristas, entre otras muchas actividades del sector.

La segunda conferencia *Big Data y la toma de decisiones en la empresa*, corrió a cargo de **José Luis Flórez**, asesor internacional de Accenture en materia de Análisis Avanzado de Datos y doctor por la Universidad Europea de Madrid. José Luis Flórez señaló que si bien el cliente es importante, también lo es el contexto, dado que «las fuentes de información cambian con mucha rapidez». En este aspecto, defendió que ahora es posible «anticipar las necesidades, entender las características de cada individuo» y que el objetivo de los modelos de *Big Data* es "generar nuevo conocimiento que pueda tener impacto en el negocio".

Óscar Méndez, CEO de Stratio, habló –al igual que lo hizo en la Jornada de Madrid–sobre *Los datos, la nueva materia prima del marketing*. Méndez defendió que el futuro pasa por el análisis de millones de datos en milisegundos y animó a las empresas a incorporarse al *Big Data*: «o compites o desaparecerás», advirtió. "El *Big Data* debe entenderse como un medio, ya que como un fin en sí mismo no aporta más que una simple marquesina de publicidad».

A continuación intervino **Ricard Martínez** para hablar de *Ética y privacidad de los datos*, con los mismos argumentos expuestos en su participación anterior; al igual que **Daniel Gayo-Avello**, quien también repetía intervención: *Big Data, Twitter, opinión pública y mercados*.

La jornada se cerró con dos ponencias a cargo de **Antoni Bibiloni**, *Big Data para el Análisis de la opinión social*, y **Mario Tascón**, *Medios de comunicación y Datos*.

Antoni Bibiloni, profesor del Departamento de Ciencias Matemáticas e Informática de la Universitat de les Illes Balears, expuso el proyecto que desarrolla la Cátedra Sol Meliá, una aplicación que traduce en métricas concretas las ideas, sentimientos o intenciones que los usuarios y empresas vierten en las redes, en este caso relacionadas con el turismo. Bibiloni cedió parte de su tiempo a **Antoni Carmona**, responsable de Desarrollo de Negocios de Informática El Corte Inglés, quien explicó el proyecto del "Escaparate turístico de Baleares", que viene desarrollando para el Govern Balear.

Por último, **Mario Tascón**, socio director de Prodigioso Volcán y especialista en medios digitales y redes sociales, habló en su conferencia sobre periodismo de datos, la fusión entre el periodismo y el *Big Data*. «Con los meros datos no obtenemos información –dijo–, ésta surge cuando estos se procesan, se organizan y se construye un mensaje». Tascón se centró también en destacar la importancia de la visualización de los datos para su correcta comprensión e interpretación, y puso numerosos ejemplos de cómo lo aborda el periodismo actual y también de la "larguísima historia" que la visualización de datos ha tenido en los medios de comunicación desde su nacimiento.

Big Data, *y la toma de decisiones en la empresa*

Por José Luis Flórez

Asesor internacional de Accenture en materia de Análisis Avanzado de Datos

Llevo trabajando desde el año 95 en cuestiones relacionadas con la inteligencia artificial y en el diseño de algoritmos para la toma de decisiones, fundamentalmente en el ámbito empresarial. Mi labor en los últimos años ha estado relacionada básicamente con tratar de identificar qué se puede hacer con todo esto que llamamos el *Big Data*, cómo se puede hacer, y cuál es el programa de entrenamiento que puede hacer que un profesional o una persona con una cierta formación previa en Ingeniería, en Matemáticas, en Economía, etc., sea una persona productiva en este ámbito.

Aunque se trata de un concepto muy relativo, el *Big Data* se refiere a las cada vez mayores necesidades que tenemos de almacenamiento y procesamiento de datos. Lo que ahora consideramos grande, probablemente en dos años no lo sea, y lo que hace dos años considerábamos grande, ahora mismo probablemente no lo es. El volumen es un concepto que está cambiando continuamente. Y lo más importante no son los datos sino la forma de extraer de estos datos algo que sea valioso. Se habla habitualmente de estos conceptos vinculados al *Big Data*: el concepto de la velocidad, del volumen, de la volatilidad de la información, de la variedad de los datos. Es cierto que son cues-

La inteligencia analítica y la minería de datos se han centrado fundamentalmente en tratar de poner luz sobre aquellos elementos que sabíamos que desconocíamos

tiones importantes, pero quizás hay otra característica que no se comenta tanto y que desde Accenture tratamos de entender. ¿Cuál es el hecho sustantivo y diferencial del *Big Data*? Creemos que la gran diferencia, cuando hablamos de grandes datos, tiene que ver con el hecho de que las organizaciones se mueven en una situación en la cual tienen un cierto conocimiento de su entorno, de su negocio, de las decisiones que puede tomar, del impacto que pueden tener; hay cosas que saben, pero realmente lo que saben es poco. Hay muchas más cosas que desconocemos. De esas cosas que desconocemos, hay algunas que sabemos que desconocemos: "No sé exactamente cuál es el cliente al que tengo que ofrecer este producto en concreto". Entonces tratamos de desarrollar un modelo predictivo que identifique cuáles son los patrones, los perfiles detrás de este comportamiento para, a partir de ahí, optimizar, maximizar el rendimiento de nuestras campañas. "Sé que no conozco a priori y con certeza cuál es ese perfil, sé que lo desconozco". La Ciencia analítica, la inteligencia analítica, y la minería de datos, con todas las denominaciones que pueda haber habido en los últimos 20 años, se ha fundamentado o se ha centrado fundamentalmente en ese aspecto, en tratar de poner luz sobre aquellos elementos que sabíamos que desconocíamos.

"Lo que no sabemos que desconocemos"

No obstante, lo que marca realmente la diferencia es conocer aquello que ni siquiera sabemos que es relevante para nuestro negocio. Por ejemplo, cuando hablábamos de los sistemas de recomendación, cuando Amazon hace una recomendación está ayudando al usuario a conocer cosas que, a priori, ni siquiera sabía que desconocía; está guiando su búsqueda y eso es muy importante. Por tanto, esa es una gran diferencia, el poner el foco en todo aquello que desconocemos que no conocemos. Y esto tiene otra forma de expresarse en términos más tangibles y que tiene que ver exactamente con los datos. Cuando hablamos de estos entornos más tradicionales del análisis, lo que sucedía es que los datos que había a nuestra disposición estaban en un ambiente muy controlado, teníamos nuestros almacenes de datos, teníamos estructuras donde la información estaba bien estructurada, bien definida, bien delimitada, y donde el perímetro de información era claro. Dicho de otra manera, el continente dentro del cual se depositaban los datos era bien conocido, era estable y limitado. Ahora bien, estamos abriendo nuestros sentidos, desde el punto de vista empresarial, hacia un entorno que está cambiando dinámicamente, donde obtenemos información de Internet, a partir de los teléfonos móviles, de dispositivos y sensores que pueden tener en cuenta los biorritmos o ciertas características de nuestros clientes, cualquier tipo de información. Entonces, la situación cambia drásticamente porque ahora el continente ya no es fijo, pasa de una situación de ser un ente rígido, un ente sólido, a un ente que es gaseoso, variable y muy dinámico. Y esa situación cambia, o hace cambiar radicalmente, el enfoque analítico o metodológico que necesitamos para poder obtener valor de ese contexto.

¿Qué es lo relevante, por ejemplo, para identificar un fraude? En una situación convencional, tendríamos cierta información de cuáles son los hábitos, los comportamientos, las transacciones que se están produciendo, por ejemplo, en el uso de una

tarjeta, pero efectivamente la información, o en este caso los patrones de fraude, pueden estar cambiando dinámicamente. De hecho, los patrones cambian. En una perspectiva más tradicional, uno tendría que incorporar esos nuevos comportamientos en forma de datos, tendría que reentrenar el modelo para que pudiera captar este tipo de nuevos comportamientos. Ahora mismo la necesidad es hacerlo mucho más dinámicamente, es decir, se necesita que los propios modelos sean capaces de auto-adaptarse dinámicamente a los cambios de comportamiento, que sean capaces de adaptarse no solamente al hecho de que ciertas variables cambien su ponderación, o ciertas combinaciones de variables la cambien porque ha cambiado el patrón; tiene que adaptarse a una situación en la que los datos de entrada pueden estar también cambiando dinámicamente, con nuevas fuentes de información que surgen continuamente, lo cual constituye una gran diferencia.

Por tanto, cuando hablamos de esa situación de indefinición en la que materialmente el conjunto de información que tenemos de partida es casi infinito, cuando hablamos de una situación en la que lo que desconocemos que desconocemos es lo más importante, hay una palabra que toma un gran peso a la hora de entender hacia dónde está evolucionando este mundo analítico del *Big Data*, y es el descubrimiento. Necesitamos instaurar dentro de las organizaciones –también como ciudadanos– procedimientos, personas e infraestructuras que nos permitan estar generando nuevo conocimiento de una forma continua. Es lo que llamamos el *Owe Zone*, que significa tener la capacidad de estar continuamente aprendiendo y generando nuevo conocimiento. Esta es la gran diferencia desde nuestro punto de vista.

La orientación al cliente

Tradicionalmente las empresas se han centrado mucho en el producto, pero ya

desde la década pasada, incluso un poco antes, las organizaciones han ido centrándose cada vez más en el cliente. El producto sigue siendo importante, hay que gestionarlo adecuadamente, evolucionarlo, tiene que ser un elemento vivo y evidentemente muy alineado con el mercado, pero la empresa también quiere entender y conocer a sus clientes, quiere establecer diálogos con ellos. Y cuando se habla de *Big Data*, ya no solamente es que sea importante el cliente, es muy importante el contexto, la situación en la cual se produce la interacción con el cliente. Un cliente por la mañana no es igual, ni responde igual, ni está interesado en las mismas cosas que ese mismo cliente por la tarde, o un mes o dos meses después. Todo el elemento contextual en el que se produce la comunicación es fundamental para ofrecer el producto o el servicio adecuado en el momento adecuado.

El concepto de la interacción obliga a tener sistemas que sean capaces de captar todos esos datos en tiempo real; no solamente procesarlos y captarlos, sino también entenderlos; poder asimilar qué es lo que ese cliente está diciendo, por qué lo está dicien-

do, cómo reacciona a lo que yo le digo, y tengo que cambiar y modular mi mensaje de forma inmediata. Por lo tanto, pasamos del producto al cliente, y del cliente a la interacción, que es el elemento clave. El concepto de interacción no es necesariamente una interacción entre personas, o ni siquiera la interacción entre un sistema y una persona; en muchas ocasiones estamos hablando de interacción entre sistemas automáticos, otra cuestión muy relevante.

El entrenamiento. ¿Cuándo se entrena un modelo o un sistema que tiene que ser sensible a las señales del entorno para tomar una decisión? No vale ya con que ese entrenamiento se esté produciendo en entornos más tranquilos, una vez al mes, una vez al año, o cuando hay un cierto problema identificado, sino que es necesario que esa actualización se esté produciendo continuamente. Y la principal cuestión ya no es solamente la dinámica del mercado, el cambio de los patrones, etc., sino el hecho de que las fuentes de información cambian con mucha rapidez, una rapidez inusitada si lo comparamos con la estabilidad que teníamos hace unos años.

Otra tendencia que es imparable en el mundo analítico es la migración de lo individual a lo social. Cuando tratábamos de entender el comportamiento de un cliente, anticipar sus necesidades y, a partir de ahí, hacerle propuestas de valor que le fueran útiles, lo que hacíamos fundamentalmente era tratar de entender cuáles eran todas las características de ese individuo y tratar de buscar individuos que eran parecidos a él, o que lo habían sido en un pasado reciente. Tratábamos de entender en ese colectivo de individuos similares, con unos patrones bastante afines, qué es lo que había funcionado y lo que no, para definir el arquetipo o el paradigma del enfoque analítico. Ahora mismo, incorporamos también la perspectiva social. Nos interesa mucho encontrar relaciones entre entidades, relaciones entre productos, relaciones entre personas, relaciones entre cualquier tipo de elemento que nos pueda conducir a la toma de una serie de decisiones.

Por supuesto, antes la información empleada era fundamentalmente estructurada. Podía tener más o menos calidad, pero estaba estructurada: edad, sexo del cliente, sus ingresos, los productos, etc. Estaba perfectamente claro cuál era la información que había en cada uno de los registros. Ahora manejamos realmente cualquier tipo de información como, por ejemplo, sistemas de reconocimiento óptico, sistemas de reconocimiento de imágenes, sistemas para validar el estado de infraestructuras, sistemas térmicos o de otros tipos de sensores; sistemas de vídeo-reconocimiento de imágenes por satélite o reconocimiento de imágenes aéreas para conocer cuál es el estado de una red de distribución eléctrica o de una red de gasoductos, etc. También existen otras aplicaciones para otras áreas de actividad, como la identificación de objetos cercanos a la Tierra, lo que se llama el *Near Earth Object*: objetos, meteoritos, los satélites cercanos a la tierra que pudieran en un momento determinado suponer un problema, etc.

Otra transición importante es el paso de lo manual a lo industrializado en la elaboración de los propios modelos. Hay partes muy amplias dentro del análisis de datos avanzado, que desde algún punto de vista podríamos decir que se han convertido en un estándar. Por ejemplo, es posible automatizar en gran medida esas capacidades de análisis. ¿Quién va a ser un buen cliente para este producto? ¿Qué cliente puede dejar de serlo? ¿Cuál es el riesgo de este cliente y la probabilidad de fraude? En definitiva, todas estas cuestiones son problemas bastante parecidos desde un punto de vista técnico porque tienen siempre una serie de variables objetivo que hay que predecir, al igual que hay que optimizar la manera en la que se puede llegar a determinar esas variables. En este sentido, el grado de automatización puede ser muy grande, permitiéndo-

nos ahorrar cantidades enormes de tiempo.

Los nuevos modelos analíticos

Cuando hablamos de análisis desde una perspectiva histórica, en los años 60 Willam Fair y Earl Isaac, un ingeniero y un matemático, empezaban a crear por entonces los primeros modelos de *scoring*. Si avanzamos hasta la primera década de 2000, encontramos que el modelo tradicional analítico se corresponde a un modelo en el cual tenemos una serie de datos, denominados factores e *insights*, más o menos estructurados, conocidos, almacenados en ciertos repositorios de información, sobre los que aplicamos ciertos modelos de análisis que hemos denominado de forma genérica *Machine Learning* y que nos permiten determinar efectivamente qué sistemas pueden fallar, qué personas pueden delinquir, qué personas pueden estar interesadas en un producto, qué productos pueden interesar a estas personas, etc.

Después se genera la simulación, es decir, llevamos al mercado diferentes decisiones y en función de cuál sea el criterio de toma de decisión, el impacto económico será uno u otro. Para ello es muy interesante contar con un entorno de laboratorio, un entorno controlado donde pueda determinar qué va pasando. Luego, se toma una decisión conforme a estas simulaciones y se ejecuta. Y después vuelve a iniciarse el ciclo, con nuevos datos del impacto en el negocio y la decisión tomada. Al final, si las cosas han funcionado bien se mantienen, y si no han funcionado, se modifican.

Luego está el elemento causal. Varios profesores de la Universidad de Princeton han declarado que con el *Big Data* la causalidad muere y el modelo aristotélico ya no está tan vigente. Esto es cierto parcialmente, es decir, cuando hablamos por ejemplo de sistemas de recomendación, como los de Netflix o Amazon, lo que interesa es tomar una matriz muy dispersa donde efectivamente el K sea mucho mayor que N, factorizar la matriz,

tratar de condensar la información, buscar un criterio de medida de distancia que permita conocer cuáles son las valoraciones de ciertos productos por parte de personas relativamente cercanas a uno o que tenga gustos parecidos a los de uno en el pasado. Esto es lo que ofrece la recomendación. No obstante, después entramos en la fase en la que tenemos que controlar el proceso. Entonces ahí sí que introducimos, dentro de lo que es la metodología de *Data Discovery*, un elemento causal, porque es verdad que en analítica las capacidades computacionales que tenemos, al igual que dan mucha capacidad para analizar a gran nivel de detalle y recomendar, predecir, clasificar, etc., también ofrecen una buena capacidad para identificar causas, lo que siempre es un factor relevante para una organización empresarial.

Cuando tenemos un diagnóstico hipotético de causas podemos testarlas para contrastar las diferentes hipótesis que uno pueda tener. Es decir, tienes los datos, los filtras, generas una señal, identificas una serie de patrones, puedes tener ciertos candidatos con altas correlaciones al explicar el fenómeno, pero como la diferencia entre

correlación y causalidad muchas veces es muy tenue o muy difícil de explicar y puede conducir a muchos errores, es por lo que se requiere una fase de testeo y de generación de hipótesis.

La velocidad de los datos, la escala y el componente social

En referencia a la velocidad, hay un ejemplo muy ilustrativo de *Big Data* del que todos somos usuarios cotidianamente aunque no nos demos cuenta: el *Real Time Bidding*, las pujas en tiempo real, en este caso en el mercado publicitario. Cuando entramos en una página web, vemos contenido y también información publicitaria, como *banners*, que podrá ser más o menos interesante. Entre el momento en el que uno accede a la página web y la visualiza pasan muchas cosas: una serie de empresas recibe información de ese usuario, información que aunque probablemente no permite identificar en muchos casos al usuario, si se trata de una *cookie* pueden conocerse hasta cierto punto las páginas por las que ha pasado esta persona y es posible determinar cuáles son sus gustos, aficiones o inclinaciones. Al mismo tiempo, lo que hacemos es identificar cuál es la página a la que se está dirigiendo, y antes de que entre, la empresa analiza esa página, es decir, determina las palabras claves en esa página, los temas de los que se trata, el tono que tiene, la estructura de la página, si hay mucho o poco contenido en imágenes, los *banners* y anuncios que aparecen, etc. Y con esta información se crean métricas para saber hasta qué punto un anuncio es interesante para el usuario y para la página que visita. Finalmente, mezclando esos dos elementos, el comportamiento del usuario y la información de la página, la empresa realiza una puja por una cantidad de dinero para mostrar un anuncio determinado. Al mismo tiempo, otros sistemas automáticos, con otro inventario publicitario distinto, porque tiene acuerdo con otras marcas, hacen su cálculo y también pujan. El ganador termina por mostrar su anuncio después de analizar todos los factores anteriores y tomar una decisión en cuestión de milisegundos. La velocidad, evidentemente, tiene un impacto económico brutal. En Estados Unidos este sistema ya es la primera forma de compra de anuncios a través de la web.

El segundo punto es la escala. Actualmente las organizaciones tienen que diseñar sus productos específicamente para las necesidades de un cliente muy concreto en sus circunstancias particulares. Hay que modificar las condiciones de acuerdo a lo que el cliente dice y las empresas tienen que tener capacidades dinámicas de negociación, de fijación de precios, etc., y todos estos elementos, cada vez más orientados a la interacción, producen una explosión de modelos que solo se puede atender si se consiguen reducir los tiempos de análisis automatizando los procesos. Esa inteligencia automática es fundamental.

Y el tercer punto es el componente social. Para identificar el componente social realmente se necesitan varios pasos: primero, definir qué es una relación entre dos entidades; una vez que tenemos esa relación entre dos entidades, podemos iterar, y encontrar estructuras e identificar las denominadas comunidades (las comunidades son muy importantes desde el punto de vista del *marketing*, del fraude, del control de riesgos, etc.).

En referencia a la formación, el *data scientist* es el profesional del *Big Data*; tiene que saber de negocios, de tecnología, de matemáticas, de modelos cualitativos; es un perfil muy complicado de obtener. Nosotros creemos más en un perfil de especialización y equipos mixtos. Pero dicho esto, sí que es cierto que hay un *gap* importante, una separación entre la formación académica que recibimos y la necesaria para enfrentarse a este mundo, lo cual puede ser una magnífica oportunidad para ir reduciendo ese *gap* y crear oportunidades, tanto profesionales como de empresa para todos.

Los datos, *la nueva materia prima del* marketing

Por Óscar Méndez
CEO de Stratio

El *Big Data* es simplemente una tecnología que te permite obtener un valor adicional a partir de ciertos datos. Es una nueva manera de utilizar los datos, que combina las tecnologías anteriores de tratamiento de datos.

El *Big Data* se aplica al *marketing* porque las campañas de *marketing* tienen que gestionar todos los datos de los clientes y del mercado, lo que supone la gestión de muchísimos datos que además son muy variados. Esta variedad no solo se refiere a la naturaleza, estructurada o no estructurada, sino también a sus tiempos: el tiempo pasado, el tiempo presente y el tiempo futuro. Los datos del tiempo pasado son aquellos ya almacenados en el sistema, los datos del tiempo presente son aquellos que están entrando en nuestro sistema en ese momento y están relacionados con el mundo operacional, las aplicaciones, etc. El tiempo futuro es el mundo del análisis predictivo, de los algoritmos, de los científicos de datos; es el tiempo más importante para todo tipo de campañas.

Otro concepto importante relacionado con el *Big Data* es la velocidad, ya que en *marketing* se necesita analizar y relacionar los datos a grandísimas velocidades para ver, por ejemplo, cómo están funcionando esas campañas en tiempo real y actuar acorde a ello. En este sentido, Internet y el mundo digital permiten obtener información inme-

El límite de los segmentos socio-demográficos hoy en día no es la tecnología de Big Data *ni la de los algoritmos, que avanzan cada vez más, sino la capacidad que tienen las máquinas de generar tantas páginas o contenido como usuarios*

diata sobre el impacto de las campañas, lo que lleva a la necesidad de usar tecnologías que puedan procesar toda esa información rápidamente.

Big Data y creación de valor

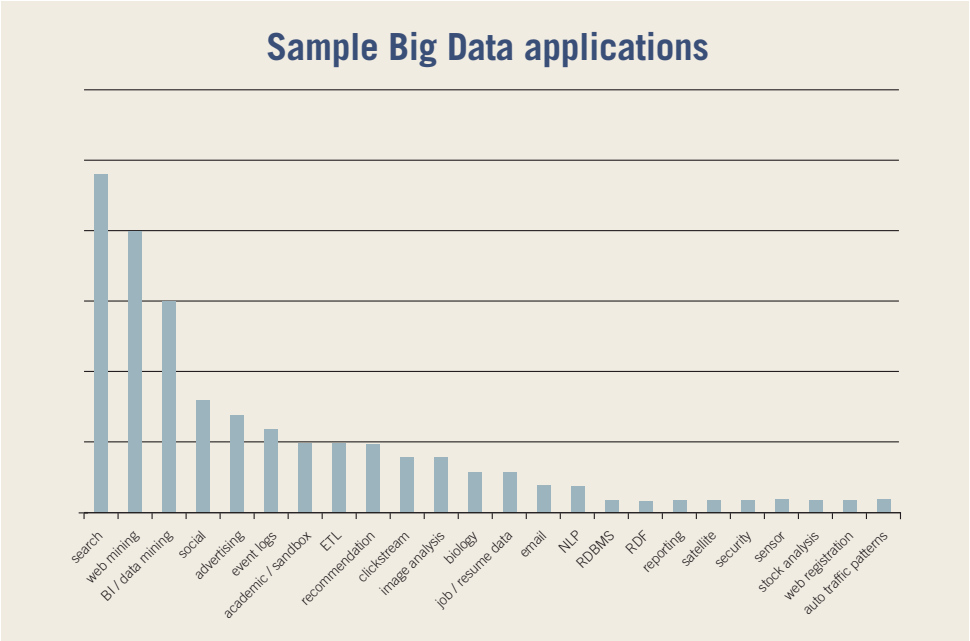
Sin embargo, la penetración del *Big Data* en España es muy baja, estimándose que apenas un 3% de las empresas realmente lo utilizan. En Europa también se utiliza muy poco. En realidad lo que se está haciendo son "pruebas de concepto" y muchos pilotos, mientras que la producción real y los proyectos están relegados a algunos casos muy concretos y a 3 ó 4 empresas en España. Nosotros organizamos "*Big Data Spain*", que es una de las mayores asociaciones en España actualmente. También hemos creado "*Big Data Hispano*", que es la organización de *Big Data* más importante de España y Latinoamérica. No se está utilizando de verdad el *Big Data* por dos motivos: primero, porque se habla de *Big Data* en términos de tecnología y no en términos de negocio, y eso es un error. Hay que hablar en términos de negocio y encontrar casos de uso en los que aporte valor, como en *marketing*, por ejemplo.

El segundo gran problema, en términos de tecnología, es poner los datos en valor. Es decir, convertir los datos en valor económico, porque los datos tienen un gran valor monetario. Solo hay que ver las cotizaciones de empresas que hacen un uso increíble de los datos para ser consciente de ello. Por ejemplo, la empresa sueca Spotify vale mucho porque tiene muchísimos datos sobre sus usuarios y sus gustos, y además tiene un trato muy bueno de estos datos y utiliza los motores de personalización y recomendación más sofisticados del mundo, que utilizan algoritmos complicadísimos. Este tipo de empresas, como Spotify, Facebook o Amazon, son claros ejemplos de que los datos tienen un valor que cotiza en bolsa, y además de que pueda ser porque sus fundadores sean unos visionarios y unos gurús, lo cierto es que hacen un uso cada vez más inteligente y más maduro de los datos. Nosotros recomendamos a las empresas que hagan un estudio de la madurez de su uso de datos y analicen si los están utilizando bien y en qué porcentaje los están utilizando.

¿Por qué son tan importantes los datos para el *marketing*? Entre otros motivos, porque Internet ya no es única, hay tantas Internets como personas que lo utilizan. Los anuncios de Amazon o los *posts* de Facebook están personalizados al igual que los buscadores como Google personalizan nuestras búsquedas, una realidad que se extiende a todos los medios que hacen un buen uso de Internet. Nosotros colaboramos con "The Guardian", que es uno de los periódicos más sofisticados en este tema y que ha determinado que aquello de "una página *online*, para todos la misma, porque tengo una línea editorial y la línea editorial la tengo que mantener" es prehistórico. Es decir, hay que generar una página *online* por cada

Social networks tracking and geolocalization

Los sistemas de visualización son muy importantes para conseguir la difusión de la información en redes sociales e Internet.

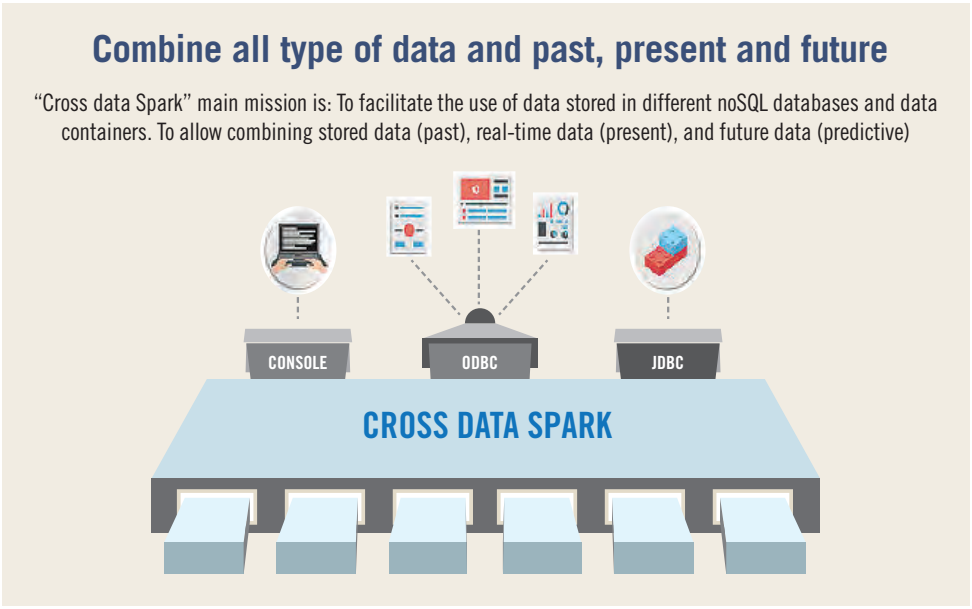


Ejemplos de tipos diferentes de aplicaciones en *Big Data*.

persona que se conecta, y esto es el marketing conocido como "*one to one*". El límite de los segmentos socio-demográficos hoy en día no es la tecnología de *Big Data* ni la de los algoritmos, que avanzan cada vez más, sino la capacidad que tienen las máquinas de generar tantas páginas o contenido como usuarios. Ese es el límite. Por eso los motores de "*behavioral customization*" que hacemos se limitan a 100, 200, 300 segmentos.

El primer paso para el *marketing* es entonces la recolección de datos (*dark Data*), y más específicamente, de los datos de navegación. Por ejemplo, Yahoo fue uno de los primeros en hacer un uso inteligente de los datos y aplicar las tecnologías de *Big Data*. Lo primero que hizo fue almacenar todos los datos de navegación de sus millones de usuarios. Para ello invirtieron decenas de millones de euros en tecnología para conseguir búsquedas mucho más personalizadas que se convertían en un mayor acceso a sus enlaces, una ventaja de un 2% respecto a sus competidores, que le produjo un retorno de 300 o 400 millones de dólares. El paso siguiente, una vez se almacenan los datos, es el cruce de los datos de múltiples canales (*omnichannel*), no solo las búsquedas y la navegación en Internet, sino también el uso del *email* o el móvil. Hay muchas posibilidades y es necesario alejarnos de los paradigmas antiguos para descubrirlas. La primera limitación somos nosotros mismos.

Otro tema importante relacionado con el *Big Data* y el *marketing* son las visualizaciones. Esta cantidad de datos, su aplicación de algoritmos y correspondiente conversión en valor no siempre se puede ver si no se cuenta con un mecanismo de visualización apropiado. De ahí que hayan surgido y estén surgiendo los mecanismos de visualización. Al fin y al cabo, en *marketing*, la difusión de información es algo esencial y los medios antiguos ya no sirven. Hoy en día hay que utilizar redes sociales y canales *online* de *marketing digital*. Para ello es muy importante entender cómo se "viraliza" la información en las redes sociales y en Internet. Twitter es



La combinación de datos del pasado, presente y futuro es un elemento clave para el éxito del *Big Data* dentro de la empresa.

la red social más rápida, pero si uno quiere que algo llegue muy rápido a otras personas en Twitter, no solo hay que generar contenido interesante, sino que este contenido tiene que llegar a un difusor (*influencer*) con mucho alcance. Y para entender todo esto, los mecanismos de visualización son importantísimos.

Igualmente, la semántica es otro aspecto muy importante del *marketing* digital. Por ejemplo, en foros como Forocoches o EnFemenino, donde no sólo se habla de coches o temas femeninos, el mejor anuncio no tiene que ser uno de coches o de un producto femenino, sino que la publicidad tiene que estar orientada al tema sobre el que estén hablando en ese momento los usuarios. Para ello se utilizan motores semánticos, aunque a día de hoy todavía tienen más fallos que la comunicación humana, es decir, no entienden bien entre un 30% y un 40% de los comentarios. Por ello, nosotros no recomendamos hacer *clipping* cuando se utilizan motores semánticos en publicidad.

Otro tema sería el seguimiento de redes sociales, siempre y cuando se haga un buen uso de ellas, porque todo aquello del cliente 360º tan de moda, que utiliza los datos del cliente internos, públicos, no estructurados, *call centers*, voz pasada a texto, etc., y las redes sociales, no aporta casi nada, no suele aportar casi nada porque cruzar los datos es muy difícil. Sin embargo, una empresa sí que puede sacar partido a Facebook y hacer un buen seguimiento de sus campañas.

Por otro lado, también está el *marketing offline*, que es aquel *marketing* para el que no tengo que estar "conectado" para que me llegue: *newsletters*, correos electrónicos, etc. En este tipo de *marketing* también se pueden personalizar las *newsletters* cruzando datos públicos con datos privados en lo que sí sería un genuino ejemplo de cliente 360º que funciona. Por ejemplo, NH recolectó todos los comentarios de Tripadvisor, Booking, y demás sitios relevantes de hoteles y los cruzó para él y para su competencia, los analizó con motores semánticos y los introdujo en un sistema de *Big Data* para poder comparar diferentes factores de cualquiera de

sus hoteles con los de su competencia, de tal manera que pudieran mejorar y dirigir mejores campañas a sus clientes. De hecho, este proyecto fue realizado por nosotros y al final cruzamos los datos privados de tal manera que comparábamos la opinión que hay en Internet de los hoteles con los ingresos de los hoteles y los beneficios, y pudimos ver la correlación exacta.

Datos del pasado, presente y futuro

El futuro de esta tecnología es la combinación y la velocidad. La combinación de cualquier dato y cualquier base de datos, NOSQL y SQL, datos en ficheros, datos estructurados, datos no estructurados, etc. Además, debemos ser capaces de combinar datos en todos sus tiempos, pasado, presente y futuro, por lo que la velocidad se convierte en un elemento fundamental, para casos, por ejemplo, como los de fraude en operaciones bancarias.

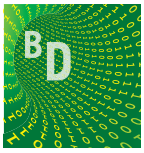
Tenemos que ser capaces de combinar de manera muy fácil los datos del pasado con los datos del presente y los datos de futuro. Estamos tan convencidos de eso, que hemos creado un nuevo sistema de tablas y de persistencia que cruza un dato que no existe, porque es del futuro, con la probabilidad de que exista. Esa serie de datos de probabilidad y de estadística, que al cruzarlos con datos del presente o almacenados, aportan valor y de manera muy sencilla. Entonces, crear los datos que todavía no existen y poder analizarlos es lo que más valor da. Y ésta es una manera para valorizar fácilmente los *scientific Data*.

Asimismo, hay que intentar que el desarrollo, la aplicación, el mantenimiento de los sistemas de *Big Data* sean muy sencillos, porque si una tecnología aporta valor pero su utilización es una pesadilla, muy pocas personas la utilizarán, que es lo que ocurre en España, donde como antes señalaba solo un 3% de las empresas la utilizan. Por tanto, hay que simplificarla, entre otras cosas, con herramientas y aplicaciones visuales. Y cuando se facilite su uso, se extenderá su uso.

Como conclusión: uno no se puede quedar mirando. Hay que arriesgarse, hay que innovar, hay que reinventarse. Y hay que hacerlo ahora, si no puede ser tarde y no hay nada más arriesgado que no arriesgarse.

Ética *y privacidad de los datos*

Por Ricard Martínez
Universidad de Valencia



El uso de los llamados datos masivos, o *Big Data*, interactúa con realidades en red, es funcional al complejo entramado en red que caracteriza no solo a las redes sociales en todas sus dimensiones, sino también a múltiples fenómenos de orden físico. Uno de los resultados determinantes de este tipo de herramientas es que es capaz de proporcionar patrones dinámicos. Y ello tanto para identificar tendencias, como desviaciones. *Big Data* no solo mira al pasado, *Big Data* se asocia a la predictibilidad y apunta al futuro. Precisamente por ello, tanto la obtención del patrón como, sobre todo, su aplicación pueden generar dudas esenciales de índole ética y jurídica.

La primera y obvia cuestión que se plantea al hablar de *Big Data* la asocia a vulneraciones de la privacidad, a la posibilidad de obtener información que afecte de modo significativo la esfera de la personalidad e incluso que sea capaz de proporcionar herramientas de control social. En este sentido, resulta fundamental recordar aquí que el derecho a la autodeterminación informativa, en su formulación germánica o a la protección de datos personales de la STC 292/2000, se asocia a la idea de control sobre nuestra información. Esta facultad de control se proyecta sobre nuestros datos, sobre todos ellos, ya que lo relevante no es su carácter público o privado sino la información que podemos extraer a partir de su tratamiento.

La primera cuestión que debemos estable-

Todavía no somos capaces de evaluar cuál será el impacto derivado del uso de los petabytes de datos que proporcionará el Internet de las cosas

cer a la hora de definir la esfera de protección que nos ofrece este derecho es si en realidad existen o no datos personales. Pues bien, un dato es «cualquier información numérica, alfabética, gráfica, fotográfica, acústica o de cualquier otro tipo concerniente a personas físicas identificadas o identificables».

Así pues la solución para eludir la aplicación de la normativa es bien simple: anonimizar, disociar los datos de manera tal que no se nos permita la identificación de un afectado o interesado. Sin embargo, la cuestión no es tan sencilla y en la práctica debemos diferenciar diferentes estadios.

Establecimiento de patrones a partir de datos anonimizados

En esta fase, el responsable del tratamiento procedería a desligar completamente los datos de sus titulares de modo que resultasen imposibles de vincular. Sin embargo, el Dictamen 5/2014 del Grupo de Trabajo del artículo 29 de la Directiva (GDT) muestra que esta no es una operación ni tan sencilla, ni precisamente banal.

En primer lugar, la anonimización constituye un tratamiento en sí misma y como tal debería ser compatible con el tratamiento original y en relación con ella contar con un fundamento, una base legal o contractual que la legitime. Ciertamente el artículo 4 LOPD cuando ordena la cancelación de oficio parece sugerir la anonimización como forma de conservación[1]. Sin embargo, esta cuestión deberá ser profundamente revisada.

En esencia la anonimización desde un punto de vista material exige:

- Que no pueda ser establecido vínculo alguno entre el dato y su titular sin un esfuerzo desproporcionado.
- Que sea irreversible.
- Que en la práctica sea equivalente al de un borrado permanente.

El problema reside en que no existe un estándar comúnmente aceptado y seguro.

Desde un punto de vista jurídico para el GdT estamos ante un tratamiento ulterior para el que sería necesario:

- Disponer de un fundamento que lo legitime, como por ejemplo el interés legítimo.
- Verificar la relación de compatibilidad entre la finalidad para la recogida inicial y un tratamiento posterior como la anonimización.
- Las expectativas del titular sobre usos posteriores.
- El impacto en el titular de los datos.
- Las cautelas adoptadas por el responsable para salvaguardar los derechos de los afectados.
- El deber de cumplir con el principio de transparencia.

En cualquier caso anonimizar no es necesariamente la mejor alternativa a cancelar los datos, pero será el camino que sin duda se seguirá y, desde el punto de vista de la protección de datos personales, en la anonimización existen riesgos:

- La persistencia de datos que permitan reidentificar.
- La posibilidad de reidentificar mediante inferencias, o por vinculación o relación (*link*) con otros paquetes de datos personales.
- Confundir pseudonimización y anonimización.
- Creer que la anonimización excluye el cumplimiento normativo sectorial.

Un ejemplo claro es el contenido por la Ley 41/2002 en relación con la investigación con datos de salud. Si atendemos a sus artículos 8 y 16 resulta que el paciente o

usuario tiene derecho a ser advertido sobre la posibilidad de utilizar los procedimientos de pronóstico, diagnóstico y terapéuticos que se le apliquen en un proyecto docente o de investigación, que en ningún caso podrá comportar riesgo adicional para su salud. Y en cuanto a los datos de su historia clínica, el acceso a la historia clínica con fines de investigación o de docencia, se rige por lo dispuesto en la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal, y en la Ley 14/1986, de 25 de abril, General de Sanidad, y demás normas de aplicación en cada caso. El acceso a la historia clínica con estos fines obliga a preservar los datos de identificación personal del paciente, separados de los de carácter clínico-asistencial, de manera que, como regla general, quede asegurado el anonimato, salvo que el propio paciente haya dado su consentimiento para no separarlos.

En resumen, anonimizar datos no será un acto de libre disposición y, además, cuando anonimización presente la menor inconsistencia cuando mediante técnicas de inferencia, de relación con otros paquetes de datos, cuando la presencia de quasi-identificadores permita la menor reidentificación operarán en bloque las garantías de la LOPD.

Pero la generación o uso de patrones puede estar también directamente sometido a la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal.

Los patrones individualizados

El primer caso, por obvio, sería cuando el patrón se basa en un individuo identificado. Esto es, no ha habido anonimización y es a partir de un universo de datos del que inferimos su patrón crediticio, de salud, etc. y derivamos una determinada consecuencia. En este caso nunca hemos salido de la esfera de tutela de la LOPD.

La aplicación de patrones

Es en este ámbito donde hay un aspecto que conviene subrayar. Cuando aplicamos un patrón a una persona identificada o identificable estamos sin duda realizando un tratamiento y debemos aplicar todas las garantías legales.

¿Dónde están los problemas?

Sin embargo, los problemas en mi opinión no se sitúan en decidir si se aplica o no la LOPD al tratamiento masivo de datos, sino en cómo lo que en algún lugar he definido las costuras de la privacidad se ven desbordadas de un modo significativo por los retos que plantea el *Big Data*.

Al referirse a la privacidad Mayer-Schönberger realiza una apropiada comparación de las leyes de privacidad como Línea Maginot. Y la comparación no es solo adecuada en términos de ingeniería militar sino también en relación con la historia. Toda Línea Maginot tiene su Bosque de las Ardenas, y el genio militar de Kleist, Guderian y Rommel capaz de embolsar y aniquilar a su enemigo… Y como en el caso de la Segunda Guerra Mundial sería muy triste que la inacción del legislador diera con nuestros huesos en las playas de Dunkerque sin más apoyo que tristes barcos pesqueros.

La falacia del consentimiento

En primer lugar, hay que referirse a lo que algunos autores han definido como falacia del consentimiento o dilema del consentimiento. El escenario más visible para el tratamiento de los datos personales es el de servicios aparentemente gratuitos de Internet cuyo modelo de negocio se basa precisamente en el de *data brokers* y que de la mano de los tratamientos masivos de datos personales, asociados a análisis comportamental y predictibilidad se erigen en los nuevos gurús capaces de predecir el comportamiento del consumidor.

Así pues en una primera fase, se ha obtenido el consentimiento para tratar una infinita cantidad de datos. No obstante, hay otros fundamentos que legitimarían un tratamiento. La reciente sentencia de 13 de mayo relativa al derecho al olvido considera que el tratamiento de datos personales por un buscador se apoya en la idea de interés legítimo, entendido este interés como el servicio que se presta al poner a disposición del usuario informaciones relevantes en sus búsquedas. La sentencia no se pronuncia, no es su objeto, sobre las condiciones de explotación de los datos que proporcionan esas búsquedas, excepto para conectar Google España con su matriz. Sin embargo la sentencia afirma con radicalidad que no puede confundirse el interés legítimo con el mero interés económico.

La calidad de los datos

En *Big Data* lo realmente relevante parece ser la suma de un cierto universo de datos de modo que la certeza de los mismos se diluye de alguna manera. Incluso lo "borroso" nos ayuda en algo. Si buscamos patrones y desviaciones se nos dice que incluso que un determinado sujeto falsee sus preferencias puede aportar valor añadido al resultado.

Pero además se produce un efecto peculiar. En la recogida directa de datos personales el responsable tiene la legítima confianza en que el afectado le facilitará datos veraces. Cuando hablamos de patrones y predictibilidad debemos confiar en que el patrón sea fiable y nos lleve a conclusiones adecuadas, ya de lo contrario el principio de veracidad de los datos peligrará.

La finalidad

Otro problema que suscita el *Big Data* reside en que su sustrato material puede quedar anticuado, pero no se agota con el uso. Por tanto, es susceptible de ser reutilizado. Y lo que resulta más peculiar, la finalidad para la que la información fue recogida si bien puede ser determinante desde un punto de vista no lo es en absoluto desde un punto de vista práctico. "Mi algoritmo se usó para estudiar patrones de consumo en un supermercado y sin embargo resultó que ofrecía elementos sustanciales en relación con las proyecciones de salud de mis clientes".

Por tanto, el concepto de finalidad puede verse por completo alterado no ya durante el uso o respecto de nuestra decisión respecto de este, sino ante resultados inesperados.

Hacia un nuevo modo de valorar la sensibilidad de los datos

En la jerga de mi profesión se tiende a llamar datos sensibles a los datos especialmente protegidos. Estos datos, desde el Convenio 108/1981 se han sometido a un régimen especial debido a su potencial uso con carácter discriminatorio. Así, un ciudadano puede ser discriminado a partir de su ideología, religión o creencias, de su raza, salud u orientación sexual.

Sin embargo, las herramientas de *Big Data* ponen en cuestión esta categorización. Primero, porque como hemos visto para establecer un perfil de orientación sexual nos basta con trazar a un sujeto en una red social o estudiar su patrón de consumo televisivo o de vídeo. También estamos en disposición de adoptar decisiones en la contratación de seguros asociados a la salud a partir de predicciones basadas en información no relacionadas con la propia salud.

En segundo lugar, porque hay datos como la geolocalización, las interacciones en redes sociales, el análisis semántico de expresiones emocionales, los hábitos sociales e Internet de las cosas, pueden aportar información relevante susceptible de ser usada con fines discriminatorios. Y también con fines de control social y policial.

Basta con citar dos sentencias del Tribunal Europeo de Derechos Humanos, Rotaru y Marper, que muestran los peligros de sociedades democráticas que conservan registros históricos de disidencia política o información genética, más allá del periodo razonable o justificable. En la misma línea el TJUE acaba de invalidar la Directiva 2006/24/CE sobre conservación de datos de tráfico en las comunicaciones. Una de las razones era que se indexa las comunicaciones de toda la población con carácter puramente preventivo.

Volvamos a imaginar un ejemplo posible de análisis masivo de los datos, ¿sería lícito analizar con fines policiales todas y cada una de las expresiones sospechosas en las redes sociales? Y, en presencia de la investigación de un delito grave, ¿podríamos vincular esa información con el terminal desde el que se realizó? Y en tal caso, ¿sería posible obtener todos los datos de geolocalización? Y ya llegamos al final, si lo relacionamos con manifestaciones con presencia de críticos con el sistema, ¿podríamos generar con ello una base de datos que indexará preventivamente a vagos, maleantes y otros sujetos caracterizados por su peligrosidad social? El mero hecho de que esta cuestión resulte posible tecnológicamente resulta sencillamente inquietante.

Por otra parte, ese potencial discriminatorio puede proyectarse a diversos sectores como la contratación laboral o de un seguro sobre la base las expectativas de salud inferidas de encuestas indirectas sobre hábitos. Por último, todavía no somos capaces de evaluar cuál será el impacto derivado del uso de los *petabytes* de datos que proporcionará el Internet de las cosas.

Dicho de otro forma, el modo tradicional de entender la sensibilidad de los datos debe abarcar ya no solo la naturaleza del dato, o la finalidad del fichero, debe incluir el escenario que deriva de la predictibilidad.

Libre elección v. libre autodeterminación

Paul Schwartz, en su artículo "Internet

privacy and the State", subraya en qué medida el manejo de información privada puede facilitar capacidad de influencia en el manejo de las preferencias sociales e individuales. En el momento de redactar estas líneas Daniel J. Solove ha publicado un *post* titulado «Facebook's Psych Experiment: Consent, Privacy and Manipulation». Al parecer el pasado fin de semana la compañía realizó un experimento que afectó a 689.000 personas. El servicio de noticias indexadas con RSS se manipuló para inducir estados de ánimo. A partir del análisis semántico se descubrió que allí donde las noticias eran positivas los *posts* de la gente eran más positivos, de idéntico modo, allí donde las noticias eran negativas, los comentarios también.

Estos hechos, aparte de confirmar mis apreciaciones sobre el valor del consentimiento nos conducen a otro territorio: el de la manipulación. ¿Puedo usar los patrones que me proporciona mi análisis de *Big Data* para manipular las preferencias del usuario? Creo que la respuesta se responde por sí sola.

El valor de los datos

Antes me referí a las compañías que tratan enormes cantidades de datos personales. En breve no nos vamos a tener que situar únicamente frente al dilema de la privacidad, sino también al de principios básicos que hasta hoy habían regido el tráfico económico. Controlar el flujo de datos equivaldrá a disponer de una posición dominante en el mercado. ¿Merece esta cuestión ser tenida en cuenta?

Pero, ¿cuál es el valor de los datos del propio usuario? Bien, la respuesta práctica es cero. Si ustedes creían que el ser peor tratado del mundo es el agricultor que vende sus berenjenas a cinco céntimos y las compra a un euro, se equivocaban. En el mundo del *Big Data*, sus datos, mis datos, nuestros datos valen 0. Curiosamente antes de la capitalización bursátil de Facebook, Garner estimó que el valor medio de un usuario era de 100 dólares. Así que ya sabe usted cuanto le cuesta usar Facebook y parece un precio razonable en términos meramente económicos.

La cuestión es que si los datos de mi consumo eléctrico son susceptibles de explotación económica adicional, y también los de la telefonía, o los movimientos de mi tarjeta de crédito, ¿cómo es que ello no repercute en mi cuenta de resultados?

En resumen, y como conclusión. Las tecnologías como el *Big Data* abren un universo de posibilidades altamente positivas en todos los ámbitos. Como comprenderán yo quiero que el análisis masivo de datos revolucione la Medicina, deseo fervientemente contar con sistemas de apoyo decisional no solo para mejorar la eficiencia de las organizaciones públicas o privadas también en mi vida personal. Y, debo confesarlo, me encanta que mi proveedor me recomiende buenos libros.

Sin embargo, los juristas a pesar de que disponemos hoy de un sólido armazón de principios, se requiere del apoyo de una regulación que descienda al detalle. Las normas de privacidad deben orientarse cada vez más al tratamiento, garantizar la anonimización y definir las condiciones que justifiquen los usos secundarios. Urge por último regular el mercado de la privacidad y de los *databrokers* y arrinconar el consentimiento a un ámbito residual. Ello es fundamental para equilibrar las facultades y poderes de negociación del individuo. Es esencial, además, que la transparencia afecte a los procesos de *Big Data*. Y parece aconsejable que los operadores en el mercado jueguen con las mismas reglas y capacidades, y se eviten situaciones de quasi-monopolio.

Se presenta ante nosotros un mundo apasionante en el que todo está por hacer, pero en el que la tierra de leche y miel que nos ofrece el *Big Data* debe ser un país donde la libertad y la autodeterminación individual no se sacrifiquen en el altar de los intereses económicos o estatales.

---

[1] «No serán conservados en forma que permita la identificación del interesado durante un período superior al necesario para los fines en base a los cuales hubieran sido recabados o registrados».

Data Science: *el futuro ha comenzado*

Por José García Montalvo

Vocal del Consejo de Ciencias Sociales, Fundación Ramón Areces

La llamada Ciencia de los Datos se caracteriza por la utilización de bases de datos masivas, lo que se suele resumir en una simple frase: la muestra es la población. Interesa tener toda información que potencialmente puede ser relevante aunque su contenido informativo pueda parecer pequeño. Una segunda característica asociada al *Big Data* es la heterogeneidad de estos datos. Las fuentes de generación de esta información pueden ser sensores, localizaciones CPS, *logs* de servidores, correos electrónicos, imágenes, voz, etc. Por tanto, dada esta enorme heterogeneidad de formatos resulta imposible, en muchos proyectos, trabajar con bases de datos SQL como ha sido la norma en el trabajo con bases de datos. Otra característica muy importante, al menos en algunas disciplinas y proyectos, es la reutilización de los datos que fueron originalmente creados para una finalidad diferente a la que fundamenta la investigación final que se realiza. En cuarto lugar, y dadas las consideraciones anteriores, los datos considerados de esta forma tienen una proporción señal/ruido muy elevada. Finalmente, gran parte de los proyectos de *Big Data* tienen como finalidad predecir y no explicar.

La causalidad pasa a ser irrelevante sustituida por la mera correlación. En este sentido la creciente utilización de técnicas de *Big Data* pone en cuestión la búsqueda de diseños cada vez más sofisticados que permitan captar la causalidad entre dos fenó-

Un aspecto interesante de la reciente importancia del Big Data *se refleja en la demanda de graduados universitarios*

menos relevantes. Por ejemplo, cuando el algoritmo *item by item* de recomendación de Amazon nos recomienda un libro, que otra persona que compró una bicicleta anteriormente también lo compró, la cuestión no es cómo explicar la causa de la relación entre la bicicleta y el libro sino la correlación observada con anterioridad entre estos dos productos.

Mitos sobre el *Big Data*

Respecto al *Big Data* existen algunos mitos que vale la pena desterrar. En primer lugar, cuando alguien habla de *Big Data* normalmente está pensando en datos producidos por Internet (Facebook, Google, etc.,) o la NSA (National Security Agency de los Estados Unidos). Sin embargo, los mayores generadores de datos son las grandes infraestructuras científicas. El Large Hadron Collider del CERN produce 600 TB/sec con sus 15 millones de sensores. Incluso después de filtrar la información se necesita almacenamiento para 25 PB/año. Y esto nos lleva al segundo punto: en la actualidad las mayores restricciones para la realización de proyectos de *Data Science* no están relacionadas con la capacidad de computación de los ordenadores sino con la capacidad de almacenamiento de información y la gran cantidad de energía que produce el tránsito de la información entre los dispositivos de almacenamiento y los procesadores. En tercer lugar, el *Big Data* requiere una visión centrada en la computación masiva en paralelo y memoria persistente, en lugar de pensar en un modelo centrado en torno a un único ordenador. Es preciso moverse a una visión de computación distribuida (escalable y computación en paralelo) y pensar en nuevos instrumentos para trabajar con bases de datos no relacionales.

Un aspecto interesante de la creciente importancia del *Big Data* se refleja en la demanda de graduados universitarios. El reciente estudio del Ministerio de Educación (2014) sobre las salidas profesionales de los universitarios muestra con claridad la influencia del *Big Data*: cuatro años después de salir de la universidad el mayor porcentaje de afiliación a la Seguridad Social se encuentra entre los graduados de Informática (78%) y Matemáticas y Estadística (72,2%). Estas son precisamente las disciplinas más vinculadas al desarrollo de la Ciencia de los Datos. Estos datos contrastan con los resultados de los titulados 15 años antes, donde la tasa de desempleo de los titulados en Matemáticas era significativamente superior a la tasa media de desempleo de los universitarios y la tasa de desempleo de los licenciados en Informática era 5 veces inferior a la tasa de los matemáticos y la mitad que los titulados en Estadística (García-Montalvo 2001).

"Without data you are just one more person with an opinion"
(anónimo)

"In God we trust; all other must bring data"
(Edward Deming)

"We are drowning in information but starved for knowledge"
(John Naisbitt)

El desarrollo del Big Data *y las técnicas asociadas al mismo están cambiando la forma de realizar investigación científica e incluso la forma en la que se enseñan disciplinas ya consolidadas*

El desarrollo del *Big Data* y las técnicas asociadas al mismo están cambiando la forma de realizar investigación científica e incluso la forma en la que se enseñan disciplinas ya consolidadas. Un ejemplo claro es la enseñanza y la aplicación de la econometría[1]. Normalmente en econometría se trabaja en el contexto de técnicas de regresión. Sin embargo, las técnicas de *machine learning*, muy ligadas a las metodologías del *Data Science*, incluyen la regresión como una técnica más en el denominado *supervised learning* junto con métodos de clasificación, árboles de decisión y redes neuronales. En nuestras clases de econometría solíamos explicar que cuando hay más variables que observaciones no se puede realizar una regresión. En los proyectos de *Big Data* sucede muy frecuentemente que existen más variables que observaciones, lo que se resuelve mediante técnicas de *shrinkage* que hacía mucho tiempo ya no explicábamos en las clases. El *software* también es muy diferente. Mientras en econometría solíamos trabajar con Stata, Mata, Gauss o Matlab los proyectos de *Big Data* usan frecuentemente Hadoop, MapReduce, Pig, Hive, ZooKeeper, Hive y R. Por tanto, la extensión del *Big Data* está transformando también las técnicas y los programas utilizados en el análisis de datos más tradicional.

Aplicaciones profesionales y comerciales

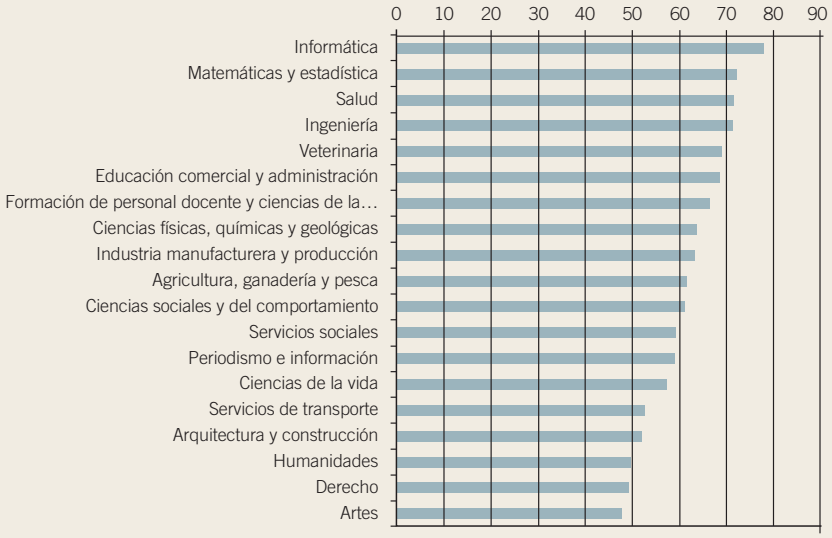
Las aplicaciones del *Data Science* se extienden desde los campos científicos hasta las aplicaciones más profesionales o comerciales. En el campo de las Ciencias Sociales y, más en concreto, en la Economía, existen multitud de ejemplos. El llamado "Billion Prices Project" del MIT utiliza millones de precios de tiendas en Internet en decenas de países para obtener un índice de precios *online* que se actualiza en tiempo real[2]. Esta tecnología utiliza la estabilidad o cambio de los componentes entre *tags* del lenguaje HTML utilizado para construir las páginas web para determinar los cambios en precios de los productos en el tiempo. Un programa puede, utilizando estos principios, identificar la información relevante sobre un producto y su precio. El URL de la página donde están indexados estos productos puede servir para clasificarlos por categorías. Una de las utilidades del proyecto es comparar la evolución de la inflación oficial y la obtenida a partir de capturas de información sobre precios de tiendas *online*. Los resultados muestran que mientras en Brasil, Chile, Colombia o Venezuela la evolución de la inflación oficial y la obtenida a partir de los precios *online* siguen patrones similares, en Argentina las diferencias son muy significativas. En media la inflación en Argentina entre 2007 y 2011 definida por el índice de precios *online* fue del 20,14% frente a la inflación oficial que era tan solo del 8,38%. Esto implica una diferencia acumulada del 65% en marzo de 2011.

Algunos investigadores utilizan Google Trends para mejorar la capacidad predictiva de modelos sobre indicadores económicos obtenidos con muy alta frecuencia. La idea consiste en complementar la información del pasado de una serie con las búsquedas presentes en algunas categorías. Por ejemplo, el Departamento de Trabajo de Estados Unidos anuncia cada jueves el número de personas que han solicitado subsidios por desempleo. Añadiendo a un modelo AR(1) de datos his-

[1] La econometría es la asignatura que he estado impartiendo los últimos 20 años en la universidad.

[2] Se almacenan 5 millones de precios de 300 tiendas en Internet en 70 países del mundo.

Proporción de graduados cotizando a la Seguridad Social cuatro años después de finalizar sus estudios



Fuente: Ministerio de Educación (2014).

tóricos la información sobre búsquedas de palabras en categorías como Jobs, Welfare, Unemployment, se mejora un 6% la capacidad predictiva en general y de los cambios de ciclo en particular. Utilizando el mismo sistema para el índice de confianza del consumidor se consigue una mejora del 9,3% en la capacidad predictiva.

La utilización de la información agregada sobre tarjetas de crédito y TPV es otra fuente importante de investigación económica en la actualidad. En una serie de artículos que han resultado muy influyentes, Mian y Sufi han utilizado la información sobre tarjetas de crédito para realizar análisis económico sobre las causas de la burbuja inmobiliaria y la crisis financiera.

Además de su influencia sobre la Ciencia y la enseñanza de las disciplinas científicas, las técnicas de *Big Data* están adquiriendo una enorme relevancia en el mundo de la empresa. Los sectores más influidos son la distribución comercial, el *marketing* y los servicios financieros aunque su influencia se extiende a casi todo el espectro de actividades empresariales.

En el campo del *marketing* y la comercialización existen muchos ejemplos, pero quizás el caso de Amazon sea de los más interesantes. Hasta 2001 Amazon utilizó docenas de críticos y editores para sugerir títulos que pudieran ser de interés para sus clientes. "Amazon voice" fue considerado en su tiempo como el crítico más influyente en Estados Unidos. A finales de los 90 Amazon puso en marcha un equipo para automatizar el procedimiento de recomendaciones de libros para sus clientes. Inicialmente se utilizaron muestras y se buscaron similitudes entre distintos compradores. Hasta que Linden propuso

una nueva solución: el llamado filtro colaborativo *item-by-item*[3]. El procedimiento utiliza algunos de los principios básicos de *Big Data*: se usan todos los datos (no se extraen muestras) y se busca capacidad predictiva y no explicativa o causalidad. La técnica de *machine learning* utilizada para realizar las recomendaciones no necesita saber por qué al comprador de El Quijote le gustaría también comprar una tostadora. Solo es necesario que exista capacidad predictiva. Cuando se compararon los dos procedimientos (críticos humanos y el algoritmo de *machine learning*) el procedimiento automatizado resultó mucho más eficiente, lo que supuso el desmantelamiento de Amazon Voice. Hoy una tercera parte de las ventas de Amazon son el resultado del sistema personalizado de recomendaciones. El sistema de Linden ha sido adoptado por muchos comercios digitales, como por ejemplo Netflix, la compañía de alquiler de películas. Este procedimiento de recomendación aumenta sin duda la satisfacción de los consumidores que pueden encontrar con facilidad productos que necesitan, les interesan y que incluso no eran conscientes de que existían.

¿Podría el futuro de los servicios bancarios discurrir por estos mismos pasos? ¿Podrían los clientes bancarios beneficiarse de sistemas que acomodaran los servicios bancarios a sus necesidades específicas y pudieran ser altamente personalizados? La utilización inteligente de la tecnología y el *Big Data* abre la posibilidad de que la banca ponga en el centro de su estrategia futura las necesidades de cada cliente de forma singularizada al igual como Amazon realiza recomendaciones personalizadas sobre productos que pueden ser de interés para cada uno de sus clientes. El objetivo debe ser mejorar la accesibilidad de familias de renta media-baja y baja a productos financieros de bajo coste adecuados a su perfil de ingresos, capacidad de pago y nivel de aversión al riesgo. De esta forma se permite el acceso a los servicios bancarios a grupos de la población que no utilizan los mismos o tiene problemas para el acceso así como se reduce los costes de los servicios. Por ejemplo, en un país financieramente avanzado como Estados Unidos se estima que existen 65 millones de personas que por no tener historial crediticio, o por su brevedad, no tienen calificación crediticia lo que les impide acceder a los servicios bancarios tradicionales. Este grupo de población es susceptible de acabar suscribiendo un *payday loan* (préstamo con un alto tipo de interés, plazo muy breve y coste entre el 20 y el 30%) o créditos informales. García Montalvo (2014) hace un recorrido por las principales aplicaciones del *Big Data* en los servicios financieros desde la generación de calificaciones crediticias de los clientes hasta el análisis y detección del fraude en tarjetas de crédito.

En conclusión, las técnicas de *Big Data* han llegado para quedarse. Las aplicaciones tanto en los campos científicos como en los empresariales se multiplican rápidamente. Será difícil entender el futuro y plantear estrategias sin comprender los métodos asociados a la Ciencia de los Datos.

Referencias

García Montalvo, José (2014), "*Big Data* y la mejora de los servicios financieros," *Papeles de Economía Española.*

García Montalvo, José (2001), *Formación y empleo de los graduados de enseñanza superior en España y en Europa.*

Linden, G., B. Smith y **J. York** (2003), "Amazon.com recomendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, 7 (1), 76-80.

Ministerio de Educación (2014), *Inserción laboral de los egresados universitarios: la perspectiva desde la afiliación a la Seguridad Social.*

[3] Linden *et al*. (2003). Este algoritmo en lugar de utilizar emparejamientos con clientes similares, empareja los ítems de las compras de los clientes a otros ítems similares para combinarlos luego en un listado de recomendaciones. En el proceso se determina el emparejamiento más similar para un determinado ítem utilizando un algoritmo que construye una lista de ítems similares que el usuario tiende a comprar juntos.

Big Data, *Ciencia y Estadística*

Por Daniel Peña

Instituto UC3M-BS de Financial Big Data y
Departamento de Estadística, Universidad Carlos III de Madrid

Desde que los británicos John Locke (1632-1704) y David Hume (1711-1776) establecieron el empirismo, los datos se han ido convirtiendo en la materia prima de conocimiento. Las ciencias experimentales han avanzado aprendiendo de las mediciones recogidas mediante observación y experimentación. La observación es un proceso lento, porque depende de la información que pueden captar nuestros sentidos. La experimentación es más eficaz, porque permite: (1) generar situaciones que ocurrirían con poca frecuencia de manera espontánea y (2) planificar la recogida de datos utilizando instrumentos de medida más precisos que nuestros sentidos. Los experimentos científicos han sido el motor del avance en el conocimiento empírico en el siglo XX, especialmente desde que R. A. Fisher, uno de los creadores de la Estadística, estableciera en 1935 los principios para diseñarlos.

En el siglo XXI se ha producido un cambio trascendental en cómo generamos datos. La digitalización de la información permite hacerlo automáticamente, y casi sin coste, mediante sensores que captan información visual, auditiva y táctil, con una precisión mucho mayor que la del ojo humano, el oído o la piel. Los avances espectaculares en la velocidad de transmisión de señales, la

Una tentación frecuente entre los científicos es pensar que al crecer la dimensión de un problema, que sabemos resolver a pequeña escala, los métodos establecidos se aplicarán con pequeños ajustes al problema de mayor dimensión

posibilidad de comunicarse sin cables, mediante wifi o telefonía móvil, y la integración de sensores en todos los dispositivos digitales, están generando masas de datos, los llamados *Big Data*, que van a proporcionar cambios de gran calado en la forma en que aprendemos, trabajamos, cuidamos nuestra salud, nos comunicamos y disfrutamos de nuestro ocio. En el siglo XX la inmensa mayoría de los datos disponibles habían sido creados por organizaciones, empresas o instituciones sociales y científicas. Actualmente, la gran mayoría (80%) se crean por la actividad diaria de las personas.

Consideremos, como ejemplo, los cambios que están apareciendo en la educación. La enseñanza *online* de finales del siglo XX se basaba en la grabación de clases y se convirtió en una alternativa más barata que la enseñanza presencial, aunque con las ventajas indudables de eliminar las distancias, las zonas horarias y las clases a horas definidas. Sin embargo, no se modificó el proceso de aprendizaje, que siguió basándose en escuchar clases magistrales, ahora grabadas en vídeo. El germen de un cambio pedagógico aparece a principios del siglo XXI, cuando en 2004 Salman Khan, un joven ingeniero del MIT, comenzó a colgar en Youtube vídeos cortos donde explicaba matemáticas a sus primos en New Orleans. Khan tuvo la intuición genial de grabar lo que vería un estudiante cuando un profesor sentado a su lado explica un concepto matemático en una hoja de papel, en lugar de mostrar el busto parlante habitual de los vídeos docentes previos. Su objetivo era hacer comprensible en pocos minutos un concepto, y hacer también al estudiante consciente de su aprendizaje poniéndolo a prueba resolviendo ejercicios y respondiendo a preguntas breves sobre este concepto. Su método tuvo un éxito inmediato y sus vídeos docentes han sido utilizados desde entonces por estudiantes de todo el mundo. Una de las claves de su éxito fue sustituir escuchar una clase magistral durante una hora por sesiones interactivas de pocos minutos, donde el estudiante invierte la mayor parte del tiempo de forma activa respondiendo a cuestiones y ejercicios.

El éxito de este enfoque impulsó las plataformas de aprendizaje gratuito y masivo (los llamados MOOCs, cursos *online* masivos y abiertos), donde los estudiantes al mismo tiempo que aprenden proporcionan información detallada sobre su proceso de aprendizaje: tiempo dedicado a cada concepto, ejercicios resueltos, partes del vídeo revisados para responder un ejercicio, etc. Estos datos permiten entender con gran detalle cómo aprende cada estudiante.

El análisis de la información proporcionada por los millones de usuarios de estos cursos va a transformar los métodos docentes. Una revolución similar se ha producido en la enseñanza de los idiomas con la aparición de Duolingo, creada por otro gran innovador, el guatemalteco Luis von Ahn, inventor de los códigos que aparecen en las páginas web para diferenciar una persona de una máquina. Esta plataforma gratuita es utilizada por más de siete millones de personas en EE.UU. para aprender idiomas, comparado con el millón y medio que asiste a clases de idiomas en el sistema de educación pública. El éxito de Duolingo, según su creador, es aprovechar la ingente cantidad de datos sobre el aprendizaje que se recogen a través de un dispositivo digital (móvil o tableta principalmente) para mejorar continuamente el aprendizaje de un idioma concreto por los nativos de otra lengua. Es conocido que para

un alemán las dificultades del inglés no son las mismas que para un español, pero este hecho no se utilizaba antes de la aparición de esta plataforma. Los datos masivos que generan los usuarios alemanes y españoles sobre su aprendizaje permiten adecuar la enseñanza del inglés a la situación de partida de cada estudiante, facilitando su rápida progresión en el idioma. Estos dos ejemplos muestran como el *Big Data* está jugando ya un papel fundamental en la modificación de los métodos de enseñanza.

La misma idea, aprovechar a los usuarios de un proceso para generar datos que lo mejoren, se ha aplicado al campo de la salud. Los sensores de los teléfonos móviles pueden recoger información para medir el ejercicio que hacemos, cómo nos alimentamos, y controlar otras variables que miden nuestra salud. Por ejemplo, con el sensor de voz se puede analizar la respiración y enviar información sobre el estado de los pulmones y pronto veremos sensores que midan la presión arterial y otras constantes vitales. De esta manera, un simple teléfono móvil puede enviar a un hospital información masiva sobre nuestra actividad, para ser analizada y comparada con la de otros pacientes y monitorizar nuestra salud. En el pasado la inmensa mayoría de la información útil para la Medicina se generaba en los centros médicos pero, en el futuro, seremos las personas las que aportaremos la mayoría de los datos relevantes para controlar la salud y prevenir la aparición de enfermedades.

Una tentación frecuente entre los científicos es pensar que al crecer la dimensión de un problema, que sabemos resolver a pequeña escala, los métodos establecidos se aplicarán con pequeños ajustes al problema de mayor dimensión. Sin embargo, es bien conocido que al aumentar la velocidad de un objeto y aproximarse a la de la luz, la Física clásica deja de ser aplicable y tenemos que utilizar las ecuaciones de la relatividad. En el mismo sentido, si descendemos a escala microscópica, hay que modificar el modelo físico habitual para englobar las nuevas fuerzas que actúan a ese nivel. En otro contexto, un medicamento que puede, en pequeñas dosis, ayudarnos a conciliar un sueño reparador, puede en dosis elevadas producirnos la muerte. El proceso científico de Hipótesis-Modelo-Experimentación-Datos-Aprendizaje, que se ha usado habitualmente para aprender de los datos, debe adaptarse a nuevas situaciones donde el punto de partida es el análisis de ingentes datos masivos generados automáticamente sobre un problema. Además, los métodos estadísticos se crearon para analizar pequeñas muestras homogéneas de una población y requieren un replanteamiento para aplicarlos a las grandes masas heterogéneas de datos actuales.

En este trabajo vamos a analizar algunas de las implicaciones del estudio de datos masivos y se organiza como sigue. En la sección 2 comentaremos cómo han ido cambiando las necesidades de almacenamiento y cálculo con los grandes bancos de datos actuales. La sección 3 analiza algunos de los cambios previsibles en la metodología de los métodos estadísticos, y cómo su interacción con la informática (Inteligencia Artificial y *Data Mining*) y la ingeniería de datos (Aprendizaje máquina o *Machine Learning*) conduce a nuevos métodos más eficaces para adquirir conocimiento del *Big Data*. La sección 4 es una llamada de atención sobre la idea, desgraciadamente muy extendida, de que al disponer de datos masivos no tenemos que preocuparnos de los problemas tradicionales de sesgos, correlaciones espurias y falsos hallazgos, estudiados en Estadística. El artículo finaliza con unas breves conclusiones.

Los Bancos de datos actuales y su tratamiento

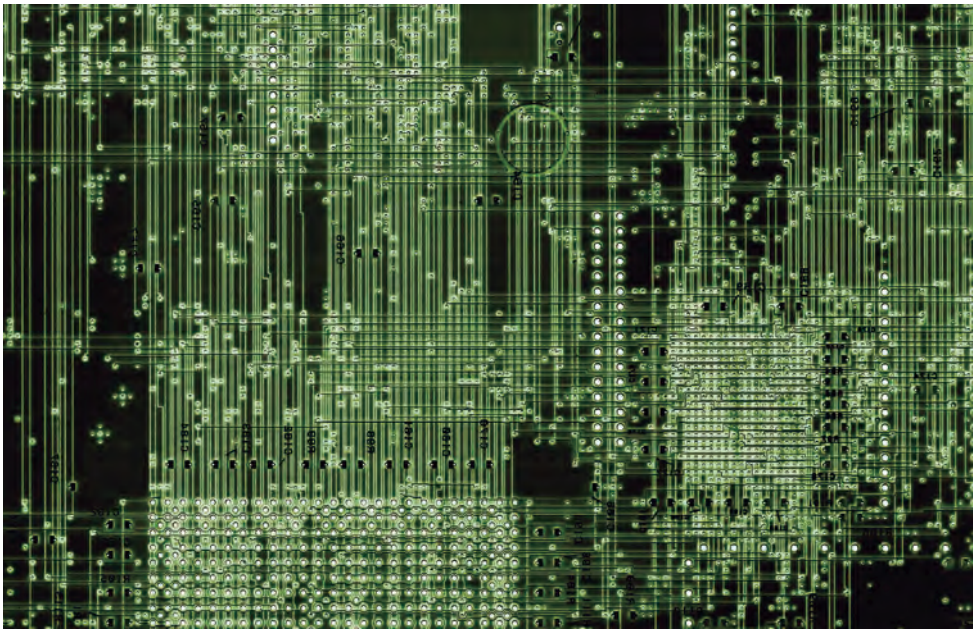
Analicemos brevemente el crecimiento de nuestra capacidad para almacenar datos y procesarlos. Recordemos que un bit (b) es la unidad mínima de almacenamiento y representa un objeto que puede solo estar en dos posiciones, por ejemplo una bombi-

lla apagada o encendida. Se representa por un dígito binario que toma los valores cero (apagado) o uno (encendido). Uniendo 8 de estos bits obtenemos un byte (B), que puede tomar 2^8= 256 valores diferentes, que son suficientes para almacenar un carácter (letra, número o símbolo). Una página escrita en A4 contiene unos 2.000 caracteres y requiere, por tanto, unos 2.000 B o 2 KB (1KB = 1.000 bytes). Las páginas de los libros impresos son más pequeñas y un libro de 350 páginas puede almacenarse en unos 400 KB = 400 x 10^3 bytes. Los primeros ordenadores personales (PC), como el XT de IBM de 1981, tenían una capacidad de almacenamiento equivalente a un libro impreso, que era la capacidad aproximada de los discos *floppy* iniciales (360 KB). Pocos años después se introdujeron en los PC los discos duros fijos, con una capacidad de 20 MB (1 megabyte = 10^6 bytes), lo que permitía almacenar unas decenas de libros. En los años 90 los PC eran capaces de guardar varias canciones de música: una canción requiere unos pocos megabytes (MB), y una sinfonía unos 80 MB. En 1992 comenzaron a instalarse discos duros de 250 MB, y desde entonces la capacidad de almacenamiento se ha incrementado el 60% cada año. En los comienzos de este siglo un PC podía almacenar unas pocas películas de hora y media, ya que cada una requiere alrededor de 1 gigabyte (GB = 10^9 bytes) y los discos duros pasaron en 2002 a 40 GB. Los teléfonos inteligentes, como el iPhone introducido en 2007, incluían 32 GB. Hoy, un PC de sobremesa puede almacenar unos pocos terabytes (TB = 10^{12} bytes), es decir, cientos de películas, miles de canciones y cientos de miles de libros, y el último iPhone incluye 128 GB, suficiente para guardar más libros de los que el dueño podría leer en su vida o canciones escuchar en un año.

Los servidores actuales se mueven en petabytes (PB = 10^{15} bytes) y muy pronto tendremos unidades de almacenamiento en exabytes (EB = 10^{18} bytes). Por ejemplo, la colección impresa de la Biblioteca del Congreso de los EE.UU. ocupa actualmente del orden de 15 terabytes, pero el World Data Centre for Climate, el WDCC (Centro Mundial de Datos para el Clima), una de las bases de datos más grande del mundo, almacena unos 400 terabytes de información sobre el clima en el planeta. Google recibe más de 100 millones de consultas al día y se supone que es capaz de almacenar cientos de terabytes de información. El CEO de Google, Eric Schmidt, estimó que la Humanidad había creado hasta 2003 una cantidad equivalente a 5 exabytes, y se estima que esta cifra se genera actualmente en 2 días.

La evolución de los sistemas operativos también ha sido muy rápida. De los programas para cálculo estadístico iniciales que realizaban una operación concreta cada vez (como en las versiones antiguas de BMDP o SPSS), se ha pasado a los programas interactivos actuales, concebidos para aplicar distintos tipos de análisis a un mismo conjunto de datos. Además, estos programas proporcionan acceso directo a los resultados intermedios, así como capacidad de programación. Al estar orientados a objetos pueden manejar indistintamente funciones, variables o gráficos. La aparición de lenguaje libre de código abierto R, en los años 90, a partir del lenguaje S+, ha creado un estándar donde cientos de investigadores de todo el mundo incorporan nuevas rutinas ampliando cada día las capacidades de análisis. Esta apertura ha dado a R una ventaja imbatible frente a otros lenguajes cerrados, que no se enriquecen continuamente por los nuevos programas escritos por miles de investigadores en todo el mundo. Además, R puede integrarse con distintas bases de datos y está evolucionando rápidamente para incorporar cálculos en paralelo, necesarios con bases de millones de datos como las actuales.

El cálculo en paralelo consiste en ejecutar conjuntos de instrucciones simultá-



neamente en varios procesadores distintos. Esto exige una programación donde, en lugar de resolver un problema secuencialmente, se descompone en partes, que pueden procesarse en paralelo con procesadores con varios núcleos, o con varios procesadores, que realizan los cálculos en paralelo y se comunican entre sí. Este sistema muestra toda su potencia cuando se conectan varios ordenadores para que trabajen conjuntamente. Puede hacerse de forma remota, donde los ordenadores no están físicamente cerca y se conectan por la web, o formando un cluster o grupo de ordenadores de potencia media, pero conectados entre sí mediante un sistema de red de alta velocidad (gigabit de fibra óptica por lo general). Además, debe existir un programa que controle la distribución de la carga de trabajo entre los equipos. Por lo general, este tipo de sistemas cuentan con un centro de almacenamiento de datos único.

Una infraestructura digital en código abierto, dentro de la licencia de la Fundación APACHE, es Hadoop, creado por Doug Cutting. Hadoop combina la computación en paralelo y distribuida permitiendo desarrollar tareas muy intensivas de computación dividiéndolas en pequeñas parte y distribuyéndolas en un conjunto tan grande como se quiera de máquinas. A diferencia de las soluciones anteriores para datos estructurados, la tecnología Hadoop introduce técnicas de programación nuevas y más accesibles para trabajar en almacenamientos de datos masivos con datos tanto estructurados como no estructurados.

En resumen, la capacidad de cálculo ha seguido aumentado de acuerdo con la ley de Moore, que predice que, aproximadamente, cada dos años se duplicará el número de transistores en un microprocesador. La capacidad de almacenaje también ha crecido a un ritmo muy fuerte, con aumentos del 80% cada año. En la actualidad, siguen apareciendo continuamente nuevos avances para mejorar nuestra capacidad de almacenar y procesar el *Big Data*.

Big Data y los cambios en la Estadística

La Estadística nace como la disciplina científica que se ocupa del análisis de datos

en Inglaterra, a principios del siglo XX, bajo el impulso de K. Pearson y R. A. Fisher. Inicialmente los datos considerados eran variables numéricas, o bien continuas, como la medida de la temperatura, o discretas (o de atributo) como el color del pelo. Estos datos se obtenían de muestras aleatorias pequeñas de poblaciones cuyas características desconocidas se deseaba estimar. Durante la mayor parte del siglo XX los modelos utilizados y los métodos de inferencia y optimalidad han correspondido a este esquema, aunque en los últimos años la Estadística se está transformando a gran velocidad para adaptarse a los datos masivos. Por ejemplo, un tema central en el trabajo de Pearson fue contrastar si los datos habían sido generados siguiendo un modelo de distribución determinado (la distribución normal o Gaussiana, por ejemplo), y una propiedad fundamental de un buen estimador, descubierta por R. A. Fisher, es ser suficientes, en cuyo caso aprovecha toda la información existente en la muestra. Estos dos temas, centrales en el pasado, tienen poca importancia con datos masivos: cualquier hipótesis de que unos datos reales han sido generados por una distribución fija será rechazada por los contrastes habituales y la suficiencia pierde importancia con datos heterogéneos frente a otras propiedades, como la robustez. Por otro lado, los datos ya no son solo variables aisladas o en conjuntos (multivariantes) sino también imágenes, vídeos, textos, sonidos o funciones. Es cierto que sobre estos objetos pueden definirse variables: por ejemplo una imagen en color está formada por tres matrices, cada una de un color, RGB (por sus siglas en inglés: *red, green, blue*) y los píxeles de cada matriz indican la intensidad del color en una escala que va del 0 al 255 (ya que se codifica con un byte, que recordemos tiene 256 valores posibles). Cada color se define por tres números enteros entre paréntesis, por ejemplo, el rojo es (255,0,0), y el amarillo (255,255,0). Los vídeos, textos o sonidos pueden también tratarse como datos estructurados incluyendo variables temporales, pero este campo se encuentra todavía en sus inicios.

Los datos pueden contener frecuentemente valores atípicos, consecuencia de errores de medición o cambios en las condiciones de observación. En los cuarenta últimos años se han introducido en Estadística los métodos robustos, que proponen nuevos estimadores que se vean poco afectados por unos pocos valores atípicos. Es previsible que la robustez tenga una importancia creciente en el desarrollo de métodos automáticos para *Big Data*, donde es importante asegurarse que las conclusiones no dependen de unos pocos datos erróneos. Los datos masivos están también sujetos a problemas de heterogeneidad más generales, como la presencia de distintos tipos de observaciones que forman conglomerados o, con datos temporales, cambios de modelos en el tiempo. Existen muchos algoritmos de aprendizaje no supervisado, o cluster, para encontrar grupos, pero, en general, no están adaptados para las grandes masas de datos actuales.

Un problema central para el futuro es cómo combinar información diversa: distintas personas, países, instituciones, momentos de tiempo, en datos de distinto tipo: funciones, gráficos e imágenes. Para ello, habrá que desarrollar nuevos métodos de Meta Análisis, que surgió precisamente para combinar información de pacientes de distintos centros. En general los métodos Bayesianos son más flexibles para manejar distintos tipos de información, por lo que es esperable su crecimiento, aunque como complemento de los métodos clásicos o frecuentistas.

Los métodos automáticos irán teniendo cada vez más peso por las necesidades de procesar rápidamente los datos. Hasta la introducción del criterio de Akaike en 1973 los estadísticos han confiado en el trabajo artesanal de construcción de modelos como la mejor forma de aprender de los datos. Sin embargo, las necesidades de procesar grandes masas de datos han hecho cada

vez más populares los métodos automáticos. Por ejemplo, el éxito de los programas TRAMO y SEATS desarrollados por Gómez y Maravall (1996) para el análisis de series temporales y la desestacionalización, es una muestra de la enorme demanda en todo el mundo por buenos métodos automáticos capaces de extraer en pocos segundos la información de un conjunto de datos temporales.

La Estadística ha ido gradualmente incorporando métodos de análisis desarrollados en otras áreas. Por ejemplo, los métodos cluster surgieron primero en las ciencias de la computación y las redes neuronales en el campo del aprendizaje máquina o *Machine Learning*, y los estadísticos han tardado en incorporar estos avances en sus estructuras de trabajo. Los investigadores en *Data Mining* e Inteligencia Artificial han propuesto nuevos métodos de clasificación, como las técnicas de Máquinas de vectores soporte (*Support vector machines*), de búsqueda de grupos, de reducción de la dimensión y de visualización de datos en muchas dimensiones. Todos estos avances, unidos a los cambios en los métodos estadísticos tradicionales, cristalizan en una interdisciplinaria Ciencia de los datos, con contribuciones de estadísticos, matemáticos, ingenieros e informáticos. Un texto que presenta una visión unificada de estos nuevos enfoques es Hastie, Tibshirani and Friedman (2011).

Sin embargo, se ha avanzado muy poco en los métodos de *Big Data* para variables dinámicas. Este es un campo donde es esperable que se produzcan avances muy importantes en los próximos años y donde los métodos estadísticos no tienen todavía ninguna alternativa eficaz desde otros campos alternativos para el análisis de *Big Data* con dimensión temporal.

Riesgos en el análisis de datos masivos

Un punto de vista frecuente entre las personas que se acercan al fenómeno de *Big Data* es suponer que un análisis puramente empírico de los datos masivos será suficiente para proporcionar los conocimientos del futuro. Este punto de vista puede ser peligroso si se olvidan algunos principios fundamentales del aprendizaje estadístico. Vamos a re-

	SOLICITUDES	ADMISIONES	PROPORCIÓN DE ADMISIÓN EN %
Mujeres	2.000	1.136	56,80
Hombres	2.000	955	47,75
Total	4.000	2.091	52,27

Tabla 1. Resultados agregados de admisión en una universidad.

visar brevemente algunos de los riesgos que pueden aparecer en un análisis no reflexivo de datos masivos.

Confundir asociación con causalidad y, por ello, generar malas previsiones

Todo científico bien informado conoce la diferencia entre una asociación positiva entre dos variables, es decir, que los valores altos en una se presentan en general con valores altos de la otra y viceversa, y la causalidad entre ellas, que implica que si una aumenta producirá en la otra también un aumento, en promedio. Por ejemplo, el número de matrimonios en un mes y su temperatura están asociados en España, porque los matrimonios en verano son los más frecuentes, pero no existe causalidad: una ola de calor en julio no hará aumentar el número de matrimonios en ese mes. Sin embargo, con frecuencia recibimos mensajes de correlaciones entre variables ligadas a nuestra salud que parten de una asociación para hacer previsiones que suponen una relación causal. Por ejemplo, de una correlación observada entre el consumo intenso de carne procesada y la frecuencia de cáncer no podemos deducir que comiendo más (o menos) carne aumente (disminuya) nuestro riesgo de cáncer. La asociación encontrada puede ser debida a que las personas con consumo intenso de carne tienen otros hábitos de vida que son los responsables del aumento en el riesgo de cáncer, o a un aditivo añadido a ciertos tipos de carne procesada, pero no a todos, que es cancerígeno. Si no entendemos la cadena causal, que solo podemos deducir mediante una bien planificada experimentación, las correlaciones encontradas pueden ser engañosas, como explicó hace casi 50 años con gran maestría el genial estadístico George Box (1966).

Sin embargo, el olvido de estos principios estadísticos básicos ha llevado recientemente a uno de los fracasos más conocidos en el análisis de *Big Data*: las predicciones de Google de los contagios de gripe (https://en.wikipedia.org/wiki/Google_Flu_Trends). Una estimación inicial realizada al detectar una asociación entre el número de contagios de la gripe y el número de consultas realizadas en el buscador sobre esta enfermedad, condujo a un gran éxito inicial en la predicción de la gripe, seguida de predicciones desastrosas en los años siguientes. Véase Lazer, Kennedy, King and Vespignani (2014) para un análisis de las causas de este fracaso. Estos autores concluyen que el análisis automático de *Big Data* puede complementar, pero nunca reemplazar, los métodos estadísticos tradicionales de recoger datos y analizarlos.

Encontrar relaciones inexistentes entre variables independientes

Supongamos que tenemos una base de datos con 1.000 variables que en realidad son independientes. Para buscar relaciones se calculan los ([1.000]/2) = 499.500 coeficientes de correlación por parejas y se consideran ver-

	SOLICITUDES	ADMISIONES	PROPORCIÓN DE ADMISIÓN EN %
Hum-m	800	560	70
Hum-h	300	225	75
Ing-m	200	36	18
Ing-h	700	140	20
Eco-m	1.000	540	54
Eco-h	1.000	590	59
Total	4.000	2.091	52,27

Tabla 2. Resultados desagregados de admisión por centros en una universidad.

daderas las correlaciones que son significativas al 99%, es decir, que solo una vez de cada mil aparecerán como ciertas cuando no existen. Entonces, el número esperado de relaciones falsas encontradas será 0,001 x 499.500 = 499,5 por lo que podemos estar seguros de que, con muchas variables, con seguridad encontraremos muchas relaciones inexistentes. En los últimos años se ha desarrollado la teoría de falsos descubrimientos (*False Discovery Rate*), para modelar y comprender estas situaciones. Podemos concluir que cuando existen muchas variables y se hacen en consecuencia muchas comparaciones hay que ser extremadamente cauto y riguroso para evitar que concluyamos con muchas falsas relaciones entre las variables.

Olvidarnos de los sesgos presentes y generar malas predicciones

Si un banco tiene datos abundantes sobre los gastos con tarjetas de crédito de una parte de sus clientes es tentador utilizar esta información para predecir los gastos futuros de todos. Sin embargo, si los usuarios de las tarjetas no son representativos del total, las conclusiones pueden ser muy equivocadas. Hay una tendencia creciente a generalizar a una población más amplia los resultados de analizar mensajes de las redes sociales pero si los usuarios de Twitter o Facebook difieren en aspectos importantes de la población general, lo encontrado puede no ser aplicable a la población española. Las técnicas de muestreo y de diseño de experimentos pueden ayudar a investigar si una muestra, grande o pequeña, tiene sesgos sistemáticos respecto al conjunto de la población.

Es importante recordar que si los datos no se han obtenido por procedimientos aleatorios sino por suministro de los usuarios, como en las redes sociales, un tamaño de datos grande no asegura una buena representatividad. Por ejemplo, una correlación entre dos variables con 100.000 observaciones puede ser creada por un solo dato, que puede además ser un error de observación. Es importante, por tanto, no olvidar los controles estadísticos necesario para generalizar de los datos a una población.

Ignorar la heterogénidad puede llevar a falsas conclusiones

Ilustraremos este importante resultado primero con variables cualitativas. La Tabla 1 muestra los resultados agregados de admi-

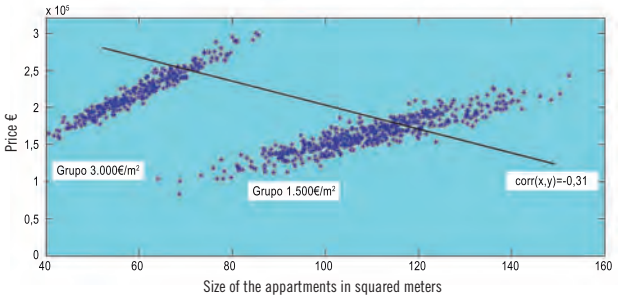


Fig. 1. Two groups of different appartments in Madrid

sión en una universidad: ingresan más mujeres que hombres y podemos concluir que las clases tendrán, en su mayoría, preponderancia femenina. La Tabla 2 desglosa esta misma información por facultades y muestra que esta conclusión es falsa: en las tres facultades se admiten más hombre que mujeres y las clases tendrán con certeza preponderancia masculina. Este fenómeno se conoce en Estadística como la paradoja de Simpson y puede resumirse así: si mezclamos unidades heterogéneas, las conclusiones obtenidas en el agregado pueden ser opuestas a las deducidas con los datos desagregados.

Los riesgos de mezclar poblaciones heterogéneas son todavía mayores en el caso de variables continúas. Por ejemplo la Figura 1 muestra la relación entre el tamaño y el precio de un piso en un barrio heterogéneo de una ciudad. En ese barrio existen apartamentos nuevos de alta calidad junto con pisos más grandes, pero en peor estado, y cuyo precio por metro cuadrado es sustancialmente menor que en los apartamentos nuevos. Dentro de cada grupo existe una relación clara y fuerte entre el precio y el tamaño del apartamento, pero al mezclar todos los pisos, la relación que aparece entre tamaño y precio es negativa, con una correlación de -0,3. De nuevo, vemos como una relación que se manifiesta en un conjunto de grupos heterogéneos puede cambiar de dirección en los datos agregados.

Conclusión

El *Big Data*, analizado con los métodos adecuados, va a proporcionarnos una gran oportunidad de avanzar nuestro conocimiento. Por un lado, pone a nuestra disposición datos con una precisión y grado de detalle y desagregación que nunca han existido en la historia. Por otro, las necesidades de análisis con datos masivos van a requerir un enfoque más interdisciplinario, con la Estadística en una posición central, pero con aportaciones fundamentales de las Ciencias de la Computación y del Aprendizaje Máquina. Además, cualquier análisis debe enmarcarse en el conocimiento ya adquirido y contrastado de la disciplina concreta que estudia en cada caso los datos analizados. En esta tarea, la creación de institutos de investigación interdisciplinarios sobre *Big Data* y *Data Science* facilitará la cooperación de estos científicos y que las herramientas más eficaces desarrolladas en unos campos de aplicación se trasladen a otros. Es importante no caer en particularismos y defensas gremiales para asegurar la unidad del método científico, que ha sido la mejor garantía de los avances pasados y lo será, indudablemente, de los futuros.

Referencias

Box G. E. P. (1966). Use and abuse of regression. Technometrics, 8, 4.

Fisher, R. A. (1935). The Design of Experiments, Edinburgh: Oliver and Boyd.

Gómez, V. and **Maravall, A.** (1996). Programas TRAMO and SEATS. Documento de Trabajo, Banco de España. SGAPE-97001.

Lazer, D. Kennedy, R. King, G. and **Vespignani, A.** (2014). The Parable of Google Flu: Traps in Big Data Analysis Science, 343, 6176, pp. 1203-1205.

Hastie, T., Tibshirani, R. and **Friedman, J.** (2011). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2th edition. Springer Series in Statistics.

Big Data en el *Retail:*
Ciencia y tecnología al servicio del consumidor

Por Juan Andrés Pro Dios

Director de Sistemas de Información del Grupo El Corte Inglés, S.A.

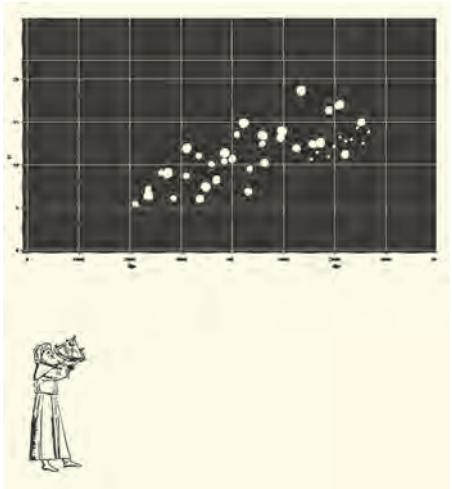
Big Data significa cosas variopintas para distintas personas. Convengamos que con el término *Big Data* nos referimos a la tendencia en el avance de la tecnología y en la reducción drástica de sus costes, que ha abierto las puertas hacia un nuevo enfoque de comprensión de la información y de toma de decisiones mediante el uso intensivo de la estadística y de la investigación operativa, que es utilizada para describir enormes cantidades de datos que llevaría demasiado esfuerzo cargar para su análisis en una base de datos relacional. De esta manera, el término *Big Data* se aplica para toda aquella información que no puede ser procesada o analizada utilizando sistemas y herramientas informáticas tradicionales. Para el comercio minorista, *Retail* en el anglicismo más extendido, su rápido desarrollo supondrá la transformación completa de su industria.

El comercio minorista es intensivo en el uso de la información. Su explotación para extraer conocimiento de clientes y mercancía siempre ha sido un hecho diferenciador. Desde los años sesenta del siglo pasado los *retailers* hemos utilizado modelos de análisis multivariante para segmentar a los clientes, evaluar el riesgo de las operaciones o evitar el fraude, modelos para prever los flujos de tesorería o aquellos encaminados a opti-

La velocidad que exige la economía global y la rapidez del cambio demográfico y social que experimentamos, hacen que el entorno del Retail *sea muy sensible al tiempo*

mizar las operaciones. Ya a finales del siglo pasado la aplicación de redes neuronales y de otras técnicas propias de la inteligencia artificial para resolver problemas cotidianos del negocio había dejado de ser exclusiva de centros de investigación para difundirse, entre otras, a la industria del *Retail*.

Conforme avanza la digitalización de la economía, la adopción masiva de tecnología por parte del consumidor y las experiencias de compra multicanal se han convertido en norma y han disparado exponencialmente el volumen de información y la velocidad con la que ésta se genera. Los datos se han convertido en un elemento crítico para competir. Son como el petróleo, la materia prima de nuestro tiempo imprescindible para sostener el negocio de los *retailers*, que depende ahora de la capacidad de cada uno de éstos para gestionar, integrar, analizar y comprender el gran volumen de información que genera tanto su actividad comercial como sus clientes y el resto de consumidores en el desempeño de su vida privada y profesional. La sostenibilidad del negocio depende, pues, de la capacidad de cada *retailer* para explotar analíticamente su *Big Data*.



La velocidad que exige la economía global y la rapidez del cambio demográfico y social que experimentamos, hacen que el entorno del *Retail* sea muy sensible al tiempo. De ello se deriva la necesidad de esta industria de analizar, comprender y predecir tendencias o comportamientos en tiempo real. Sin duda este hecho diferenciará a unos *retailers* de otros. Para las tecnologías de la información y las comunicaciones (TIC) supone una diferencia sustancial en el soporte que venían haciendo de los métodos y modelos estadísticos y de investigación operativa durante los últimos cincuenta años: hemos pasado de describir, inferir y predecir resultados sobre conjuntos de datos de manera diferida a tener la necesidad de hacerlo instantáneamente.

El tiempo se ha convertido en factor crítico a lo largo de toda la cadena de valor del *Retail*. El análisis de la demanda, la definición de la oferta, la compra de la mercancía, los planes de surtido, la gestión del inventario, la logística y la distribución, la determinación del precio, el *marketing*, la promoción, las ventas, el servicio al cliente, los pagos, las devoluciones, las finanzas, los empleados y otras muchas actividades requieren decisiones cada vez más próximas al tiempo real. Todas ellas deben ser ejecutadas inmediatamente, buscando la satisfacción del cliente y la minimización del coste para incrementar el beneficio.

Según el estudio "*Analytics: The real-world use of Big Data*", del IBM Institute for Business Value y la Said Business School de la Universidad de Oxford, el 100% de los *retailers* disponen en sus *Big Data* de información derivada de sus transacciones de *back-office*; un 67%, de los registros que permiten la trazabilidad completa de su

actividad comercial; un 57%, de la información generada por sus terminales punto de venta, escáneres y RFID; un 43%, de los datos capturados en las redes sociales; un 40%, de los datos obtenidos a través de sensores y el mismo porcentaje de ellos, de los correos electrónicos y de datos provenientes de fuentes externas (climatología, estadísticas oficiales, etc…). Según describe el estudio citado, cabe destacar que un 25% de los *retailers* ya disponen de información geoespacial, de audio y de vídeo incorporada a sus *Big Data*.

El detonante de todo ello ha sido el desarrollo de las TIC, en especial el de la telefonía móvil inteligente y el de las redes sociales, la consolidación del *Cloud* como modelo de entrega más eficiente de los servicios TIC y la explosión del Internet de las cosas. Son las fuerzas que, junto al *Big Data*, alentarán la transformación de la industria del *Retail* en los próximos años.

La mayor parte de esta industria comenzó el uso del *Big Data* con un enfoque pragmático, usando métodos estadísticos tradicionales o bayesianos y arquitecturas tecnológicas híbridas que incorporan bases de datos en memoria comerciales y herramientas *ad-hoc* de *software* libre. Pero la velocidad del cambio, la incertidumbre y la complejidad que caracterizan a nuestra era son enemigos de los sistemas de computación analítica tradicionales, máxime cuando las primeras se manifiestan en un ambiente donde las fuentes y los formatos de información son dispares, el volumen de los datos se incrementa exponencialmente y hay carencia de profesionales expertos en el análisis estadístico de los mismos. Se hace necesario entonces explorar la aplicación de la computación cognitiva a las analíticas del *Retail*.

Esta tecnología, *hardware* y *software*, se compone de sistemas que infieren, predicen y de alguna manera piensan, aplicando sobre el *Big Data* algoritmos de inteligencia artificial y *machine learning*. Los sistemas de computación cognitiva interpretan datos estructurados y no estructurados de contexto (audio e imagen) y aprenden por experiencia de la misma manera que lo hacemos los seres humanos. Sus capacidades marcarán el futuro del *Retail*, una industria en la que la Ciencia y la tecnología se aúnan al servicio del consumidor.

BIG DATA Y CAMBIO CLIMÁTICO

Big Data para el estudio del *cambio climático y la calidad del aire*

Por Francisco J. Doblas-Reyes
Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona.
Earth Sciences Department,
Barcelona Supercomputing Center-Centro Nacional de Supercomputación
(BSC-CNS), Barcelona

Francesco Benincasa y Pierre-Antoine Bretonnière
Earth Sciences Department,
Barcelona Supercomputing Center-Centro Nacional de Supercomputación
(BSC-CNS), Barcelona

Más que nunca en la historia de las ciencias en general, y de las Ciencias de la Tierra en particular, los investigadores se enfrentan al problema de tratar con rápida disponibilidad de cantidades ingentes de datos heterogéneos que crecen continuamente a un ritmo que hace de su procesamiento y diseminación con métodos convencionales todo un reto. Por ejemplo, Overpeck *et al.* (2011) muestra que mientras que los datos de modelos, satélites u observaciones *in situ* a nivel global podrían alcanzar los 50 PB en 2015, la proyección para 2030 es alrededor de siete veces superior. Los datos proceden de fuentes muy diferentes y distribuidas, comprendiendo desde plataformas satelitales hasta sensores de todo tipo, pasando por simulaciones con modelos con capacidades muy distintas o reanálisis del pasado. Estos autores afirman que, por ejemplo, los datos climáticos están creciendo de forma dramática tanto en volumen como en complejidad, de la misma forma en que sus usuarios aumentan en número y diversidad. Sugieren un nuevo paradigma de acceso a los datos meteorológicos y climáticos más rápido y sencillo para los usuarios, de manera que la sociedad pueda reducir su vulnerabilidad a la variabilidad del clima y el cambio de origen antropogénico, explotando al mismo tiempo las oportunidades técnico-científicas que puedan surgir. Este es un ejemplo perfecto de las famosas tres "Vs" que determinan un problema de *Big Data*: volumen, velocidad y variedad.

El volumen principal de los datos climáticos y meteorológicos se crea con modelos basados en procesos del sistema terrestre o procede de imágenes satelitales. Las mejoras en los modelos incluyen un número creciente de fenómenos físicos, que crean al mismo tiempo modelos que requieren ordenadores más grandes y complejos, así como un mayor número de variables que analizar y diseminar. Al mismo tiempo, estos modelos aumentan regularmente su resolución espacial para incorporar mejor en la simulación la influencia de fenómenos que ocurren a escalas más pequeñas o con frecuencias temporales más altas. Esto implica que incluso si la complejidad de los modelos no aumentara, el volumen de los datos generados aumentaría con el tiempo, convirtiéndose en un problema mayor de lo que es en la actualidad.

Además de la necesidad de tener capacidad computacional cerca de los grandes archivos de datos climáticos, la gran cantidad de datos que almacenar también implica tener que considerar problemas como la compresión, la diseminación eficiente (datos y documentación), la conservación, la energía, la replicación, la gobernanza de los metadatos y el acceso seguro y sencillo para un espectro amplio de usuarios

Para hacer las cosas aún más complicadas, todos los modelos numéricos usados (ya sea en simulaciones en meteorología, clima o calidad del aire) han adoptado la metodología de predicción por conjuntos, que consiste en realizar simulaciones paralelas que se distinguen solo por pequeñas perturbaciones introducidas en las condiciones iniciales (con el objetivo de aprovechar la bien conocida sensibilidad a la incertidumbre en las condiciones iniciales en estos sistemas). Los superordenadores aseguran una fuente de recursos de cálculo para realizar estas simulaciones, pero se han dedicado proporcionalmente menos recursos a la manera en la que se gestiona el resultado de las simulaciones, tanto dentro del superordenador como una vez que los datos generados se almacenan para su análisis o diseminación.

Más allá de las salidas generadas al realizar una simulación meteorológica, climática o de calidad del aire, hay una serie de fases críticas adicionales que requieren el uso de conceptos de *Big Data*. Entre ellas están el "posprocesado", el "data mining" o la diseminación orientada al usuario de los conjuntos de datos complejos (multifuente, multiagencia) generados alrededor del mundo con formatos diferentes. Las descargas múltiples y las transferencias redundantes ocupan gran parte del tiempo de la red, los servicios y los usuarios en estas fases cuando se usan métodos convencionales, por lo que uno de los objetivos obvios es la búsqueda de métodos que permitan reducir el tráfico de datos.

La gran cantidad de datos producidos en una situación típica se puede ilustrar con ejemplos de los volúmenes de datos generados con modelos meteorológicos y climatológicos típicos de las configuraciones actuales, así como las configuraciones esperadas para los próximos años. Las Tablas 1 y 2 muestran el tamaño de las salidas generadas de dos modelos diferentes utilizados en el Departamento de Ciencias de la Tierra del Barcelona Supercomputing Centre-Centro Nacional de Supercomputación (BSC-CNS[1]) en dos contextos diferentes. Los números de la Tabla 2 son más impresionantes que los de la Tabla 1, aunque debe tenerse en cuenta la diferencia en los tipos de experimentos realizados. El modelo usado para ilustrar los volúmenes en problemas de calidad del aire se usa en un contexto operativo en el que se hace una predicción cada día. Las estimaciones corresponden a una operación normal, teniendo en cuenta que la configuración que se usa actualmente es la de resolución estándar, durante un año. Este tipo de operación permite que haya suficiente tiempo para "posprocesar", almacenar y diseminar las predicciones de cada día antes de tener que tratar las del día siguiente. El factor crítico en este caso reside en poner las predicciones en manos de los usuarios lo antes posible para que puedan tomar las decisiones relevantes antes de que las predicciones pierdan su valor. En el caso de la Tabla 1 se ofrecen las estimaciones para una única simulación, que frecuentemente es de muchos años. Además, las simulaciones climáticas se realizan usando el método por conjuntos, lo que implica realizar varias simulaciones en paralelo con un patrón de generación de datos (frecuencia, tamaño) prácticamente idéntico, lo que requiere una gestión delicada del flujo de datos en el su-

[1] https://www.bsc.es/earth-sciences.

	Resolución horizontal (atmósfera/océano)	Tamaño de las salidas de un año de simulación, campos globales (en NetCDF, ficheros de restart no incluidos)
Resolución estándar	T255/ORCA1 60km/100km	26 GB
Alta resolución	T511/ORCA025 40km/25km	120 GB
Muy alta resolución	T1279/ORCA012 25km/12km	1 TB

Tabla 1. Tamaño de las salidas de diferentes configuraciones del modelo global de clima EC-Earth.

	Resolución horizontal	Tamaño de las salidas de un año de simulación, campos globales (incluyendo la meteorología, aerosoles y química gaseosa)
Resolución estándar	10 km	2.3 Pb
Alta resolución	4 km	9.1 Pb
Muy alta resolución	1 km	36.5 Pb

Tabla 2. Tamaño de las salidas de diferentes configuraciones del modelo de calidad del aire NMMB/BSC-CTM.

perordenador. En este caso el énfasis se pone en la extracción de los datos de la simulación del superordenador lo suficientemente rápido pero usando estructuras de datos a la hora de almacenarlos que permitan su descubrimiento a posteriori cuando se vaya a realizar su análisis y diseminación.

Estos ejemplos sencillos ilustran algunos de los retos relacionados con el *Big Data* en una pequeña parte de las Ciencias de la Tierra, una parte en la que las predicciones climáticas y de la calidad del aire tienen que abordar múltiples problemas con prioridades distintas en una misma plataforma de computación. Los conjuntos de datos meteorológicos, climáticos y de calidad del aire tienden a compartir las mismas soluciones *hardware* y *software* a pesar de que las necesidades son a menudo diferentes, lo que implica que hace falta llegar a compromisos para poder satisfacer a todos los usuarios, que es necesario explorar tecnologías y soluciones diferentes y que hay que implicar perfiles técnicos de amplio espectro. Este último aspecto ilustra la dificultad para atraer y retener suficientes recursos humanos con la experiencia apropiada, aunque una descripción de este problema está más allá del objetivo de esta contribución.

Las predicciones de la calidad del aire pueden usarse para ilustrar otros retos con los que se enfrentan la comunidad de meteoro-

logía, climatología y calidad del aire. El BSC-CNS proporciona regularmente servicios públicos de calidad del aire. Para la mejora de la credibilidad del modelo usado se lleva a cabo una validación en tiempo real que consiste en comparar tanto cualitativa (usando diferentes tipos de mapas y herramientas gráficas) como cuantitativamente (usando medidas de calidad) las predicciones con las mejores observaciones disponibles. Las observaciones proceden de múltiples fuentes tales como las estaciones AERONET, una iniciativa de NASA, las observaciones satelitales de la ESA o EUMETSAT, e incluso de redes de estaciones de medida meteorológica y de la contaminación operadas por municipios, comunidades autónomas o entidades estatales. Todos estos datos se reciben en el BSC-CNS tan pronto como se producen, momento en el que se procesan teniendo en cuenta la variedad de datos y formatos, la diferente calidad y cantidad de metadatos, la falta de información suficiente sobre su error asociado y la gran cantidad de canales por los que se reciben. A continuación se usan en la validación de las predicciones y finalmente se almacenan de manera que la información sobre las predicciones y su calidad esté disponible para los usuarios. Aunque este proceso es suficientemente complejo, la situación está cambiando con la llegada de nuevos sensores donde datos útiles de emisión de contaminantes, variables meteorológicas y exposición de las poblaciones (sobre todo urbanas) comienzan a captarse a través de dispositivos móviles, en los coches y barcos, contadores inteligentes y otras muchas vías. Esta nueva generación de fuentes observacionales promete crecer exponencialmente con el paradigma del Internet de las cosas, creando nuevas oportunidades y dificultades.

El BSC-CNS opera en este momento dos sistemas de predicción de la calidad del aire. Uno de estos sistemas es CALIOPE (http://www.bsc.es/caliope), el cual proporciona predicciones de calidad del aire para Europa y, a mayor resolución, España (Figura 1). Tradicionalmente, la mayoría de sistemas operativos han diseminado sus resultados a través de páginas web o en forma numérica usando formatos estándar. Sin embargo, un nuevo reto ha surgido con la necesidad de llegar mejor a los usuarios a través de plataformas móviles e inteligentes (Figura 2). La aplicación móvil de CALIOPE ofrece una mejor interacción con el usuario, proporcionando predicciones de la calidad del aire para Europa en general o para sitios específicos con un horizonte temporal de dos días. La aplicación usa la capacidad de posicionamiento del dispositivo para encontrar la localización del usuario y envía una petición a la base de datos de CALIOPE, mostrando la calidad del aire para las estaciones más cercanas, incluyendo una clasificación en cinco categorías (buena, aceptable, deficiente, mala y muy mala). Esta aplicación ofrece por primera vez una información individual a aquellos usuarios vulnerables al estado de la calidad del aire por su sensibilidad a, por ejemplo, las enfermedades cardiorrespiratorias. El desarrollo de esta aplicación ilustra algunos de los retos con los que nos enfrentamos (que no son muy diferentes a los de otras comunidades): el desarrollo de un *workflow* lo suficientemente flexible como para incluir la producción de las predicciones (un problema computacional intenso y extenso) y la diseminación de la información a través de un dispositivo inteligente en el menor tiempo posible, mientras que se usa el mismo dispositivo para recoger la mayor cantidad de información medioambiental local posible.

El segundo sistema operativo de calidad del aire gestionado por el BSC-CNS es el Barcelona Dust Forecast Center (BDFC, http://dust.aemet.es), una iniciativa conjunta con AEMET, la Agencia Española de Meteorología, en colaboración con la Organización Mundial de Meteorología. Este es el primer centro regional que proporciona de manera rutinaria predicciones de polvo mineral atmosférico para el norte de África, Oriente Medio y Europa (Figura 3). El sistema uti-

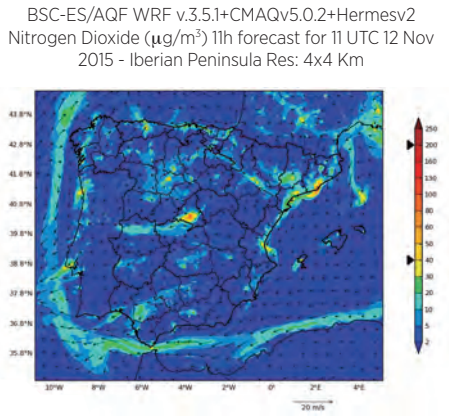


Fig. 1. Predicción de la concentración de dióxido de nitrógeno para el 12 de noviembre de 2015 realizada por el sistema de predicción de calidad del aire CALIOPE. Sobresalen los altos valores en torno a Madrid y Barcelona que reflejan el primer pico otoñal asociado a un sistema de altas presiones sobre el suroeste de Europa. Imágenes similares están disponibles en la página web de CALIOPE.

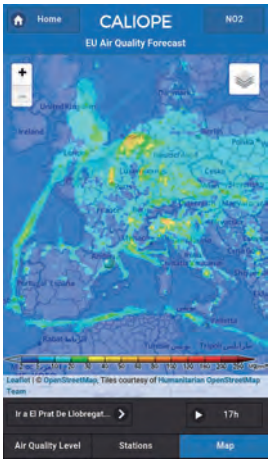


Fig. 2. Predicciones de la concentración de dióxido de nitrógeno realizadas por el sistema de predicción de calidad del aire CALIOPE vistas a través de sus aplicación para móvil. Existen aplicaciones para Android con predicciones para Europa y, a mayor resolución, para España.

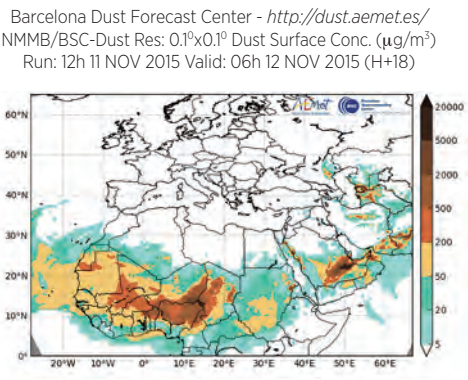


Fig. 3. Predicción de la concentración de polvo en superficie realizada por el Barcelona Dust Forecast Center para el 11 de noviembre de 2015.

liza las predicciones del modelo NMMB/BSC-Dust que se desarrolla y ejecuta con una resolución de 0.1º x 0.1º (aproximadamente 10 km) en el BSC-CNS y proporciona predicciones cada tres horas con un horizonte temporal de 72 horas para seis variables diferentes.

Debido al gran alcance regional de las predicciones del BDFC, los gestores del sistema necesitan obtener información sobre su uso para asegurar su utilidad. Google Analytics ofrece una solución preliminar a este problema (Figura 4). Sin embargo, este servicio ofrece información limitada ya que no está construido específicamente para este problema. Se necesitan otras soluciones que permitan captar una información cualitativa de los usuarios que pueda ser analizada por científicos sociales para mejorar la utilidad del servicio de acuerdo a las necesidades de un gran número de países con culturas diferentes.

Los dos sistemas de predicción de calidad del aire operativos descritos, CALIOPE and BDFC, ilustran el rango de los retos asociados con la credibilidad y el alto nivel de servicio que se espera de este tipo de sistemas. Las tecnologías de *Big Data* existentes y los nuevos conceptos asociados pueden sin ningún tipo de duda ofrecer el apoyo necesario para un servicio más eficaz.

Uno de los problemas más complejos para la aplicación de las soluciones de *Big Data* en el campo de la meteorología, la climatología y la calidad del aire lo ilustra el *Coupled Model Intercomparison Project* (CMIP; Meeh *et al.*, 2014). El objetivo de CMIP es el desarrollo del mejor sistema de información sobre el cambio climático para el pasado y el

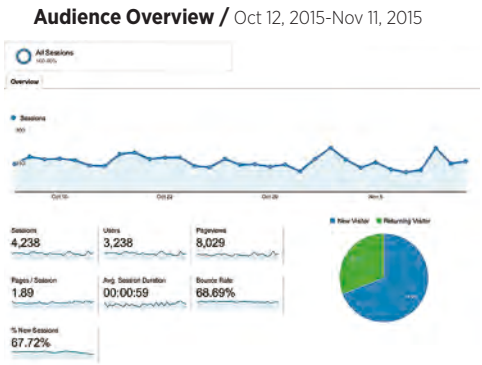


Fig. 4. Google Analytics results for the Barcelona Dust Forecast Center over the period 12th October to 11th November 2015.

futuro que apoye los esfuerzos de mitigación y adaptación promovidos por los tratados internacionales. CMIP requiere la ejecución de una gran variedad de experimentos, muchos de ellos con simulaciones de varios siglos de duración, con modelos climáticos globales por parte de varias instituciones que realizan investigación climática alrededor del mundo. En su primera realización (CMIP en 1996) el tamaño de los datos generados (1 GB) no fue un obstáculo en el momento de compartir los experimentos. Sin embargo, el último ejercicio de CMIP hasta la fecha (CMIP6, que comenzará en 2016)[2] debería generar petabytes de datos, a menudo con altas frecuencias temporales (cada seis horas de la simulación) y para un gran número de variables que presentan niveles de interés y prioridad diferentes (lo que implica un número de descargas esperadas muy distinto según la variable) por parte de los usuarios (Tabla 3).

Dejando aparte el enorme reto de la reducción de los resultados de los modelos durante las simulaciones de CMIP6 (en las que los modelizadores guardan solo aquellas variables o diagnósticos que se espera que tengan un valore real para científicos y usuarios) para disminuir el tráfico de datos antes de que los resultados se almacenen, uno de los aspectos en los que CMIP ha concentrado sus esfuerzos es en la diseminación de los datos. La documentación de todos los experimentos ha sido estandarizada para asegurar una conservación apropiada. El acceso a los datos se realiza usando una serie de portales distribuidos alrededor del mundo y gestionados por la Earth System Grid Federation (ESGF) en el que cada productor de datos los ofrece usando los mismos criterios. Además, algunas instituciones clave replican la mayoría de los datos facilitados por otros centros en otros continentes para disminuir el tráfico de datos a larga distancia y facilitar un acceso más rápido y eficaz, todo ello financiado por fondos públicos y abierto a cualquier uso que se quiera realizar. Se espera que el nivel de replicación que puede alcanzarse aún no sea suficiente para algunos usuarios, lo que ha abierto el debate de la importancia de "llevar la computación a los datos". Esto significa que los nodos de diseminación de la federación deberían también ofrecer un servicio para reducir los datos de acuerdo con las necesidades de cada usuario. Los miembros de la federación podrían ofrecer plataformas con mucha memoria y una capacidad computacional media (un problema clásico de *Big Data*) para realizar algunos cálculos previos básicos.

Además de la necesidad de tener capacidad computacional cerca de los grandes archivos de datos climáticos, la gran cantidad de datos que almacenar también implica tener que considerar problemas como la compresión, la diseminación eficiente (datos y documentación), la conservación, la energía, la replicación, la gobernanza de los metadatos y el acceso seguro y sencillo para un espectro amplio de usuarios. Otro problema en común con los sistemas operativos descritos anteriormente son la conversión de volúmenes masivos de datos de procedencias muy diferentes en un producto que usuarios de sectores distintos puedan utilizar en su toma de decisiones. Una solución posible consiste en incluir herramientas de análisis de datos

[2] https://www.wcrp-climate.org/index.php/wgcm-cmip/wgcm-cmip6.

	CMIP (1996)	CMIP2 (1997)	CMIP3 (2005-2006)	CMIP3 (2010-2014)
Número de experimentos[3]	1	2	12	110
Centros participantes	16	18	15	24
Número de modelos diferentes	19	24	21	45
Núm. de simulaciones (modelos x expts)	19	4	211	841
Tamaño total del conjunto de datos	1 GB	540 GB	36 TB	3.3 PB
Descargas totales			1.2 PB	(aún en crecimiento)
Número de artículos científicos publicados		47	595	1.015 (aún en crecimiento)

Tabla 3. Algunas características de las simulaciones realizadas en las distintas fases del *Coupled Model Intercomparison Project* (CMIP). Las estimaciones de CMIP6 aún no están disponibles, pero se espera que sean un orden de magnitud superior a las de CMIP5.

dirigidas por los usuarios, así como visualización avanzada, de manera que ellos mismos puedan extraer un mensaje útil de los datos.

En el fondo, la extracción de un mensaje significativo y orientado a la acción de la masa de datos heterogéneos que tenemos y seguimos produciendo es el interés principal del paradigma del *Big Data*. La meteorología, climatología y calidad del aire ofrece retos específicos como la naturaleza operativa de muchas de sus actividades, que implica reunir y compartir información con calendarios muy estrictos, o la necesidad de extraer información de conjuntos de datos inmensos por parte de usuarios que seguramente no son conscientes de las limitaciones de esos datos. En un contexto revolucionario como el que vivimos es importante tener en cuenta que esta comunidad tiene la particularidad de estar muy estructurada alrededor del mundo, tener una larga experiencia en el uso de la estadística y la computación y está muy adaptada a la definición y el uso de estándares. Estas características únicas hacen de ella un objetivo interesante a la hora de probar algunos de los desarrollos recientes que se realizan sobre *Big Data* en otras comunidades (Bourne *et al.*, 2015).

Bibliografía

Bourne, P.E, J.R. Lorsch y **E.D. Green** (2015). Sustaining the big-data ecosystem. Nature, 527, S16-S17, doi: 10.1038/527S16a.

Meehl, G. A., R. Moss, K. E. Taylor, V. Eyring, R. J. Stouffer, S. Bony y **B. Stevens** (2014). Climate Model Intercomparison: Preparing for the next phase. Eos, Trans. AGU, 95, 77.

Overpeck, J.T., G.A. Meehl, S. Bony y **D.R. Easterling** (2011). Climate data challenges in the 21st Century. Science, 331, 700-702, oi:10.1126/science.1197869.

[3] Se entiende por experimento el estudio de un proceso físico, escala temporal o técnica numérica. Un experimento puede incluir varios modelos y simulaciones.

Big Data y servicios climáticos: *un caso de estudio*

Por Fernando Belda

Director de Producción e Infraestructuras AEMET

Uno de los grandes retos que tienen los Servicios Meteorológicos en la presente década es tener la capacidad suficiente para dar productos y servicios climáticos con valor añadido útiles para la correcta toma de decisiones en tiempo "casi" real. Información meteorológica procedente de observaciones, modelos numéricos, satélites, radares, cámaras, etc., estamos hablando de la gestión de gran cantidad de información y el desarrollo de herramientas eficientes para la extraccion de información y del conocimiento.

Problemas como el almacenamiento y la definición de estándares, el análisis de la información desde diferentes puntos de vista de una forma rápida, el diagnóstico de cada uno de los casos y por tanto la correcta construcción de modelos, son algunas de las dificultades que nos encontramos cuando abordamos gran cantidad de información.

Sistemas de información, minería de datos o *Big Data* son conceptos que hacen referencia al manejo de grandes cantidades de datos y a los procedimientos y herramientas utilizadas para encontrar patrones repetitivos que nos sirvan para generar modelos predictivos que faciliten la generación de productos requeridos por la sociedad y de fácil uso (plataformas web, informes, estadísticas...).

Sistemas de información, minería de datos o Big Data *son conceptos que hacen referencia al manejo de grandes cantidades de datos y a los procedimientos y herramientas utilizadas para encontrar patrones repititivos que nos sirvan para generar modelos predictivos que faciliten la generación de productos requeridos por la sociedad y de fácil uso*

La correcta predicción y detección de los fenómenos meteorológicos adversos, la elaboración de eficientes sistemas de alerta temprana conlleva el manejo de una gran cantidad de información que debe ser analizada correctamente. El presente artículo intenta exponer de una forma sencilla un caso de uso para uno de los fenómenos con un impacto creciente en nuestras latitudes, la sequía.

La sequía es un fenómeno recurrente del clima europeo de especial influencia en las regiones mediterráneas. Este fenómeno necesita la definición de un marco adecuado para poder describirlo. La sequía afecta a una amplia variedad de sectores, su diversidad geográfica y distribución temporal, y la demanda de agua para uso humano hacen difícil establecer una definición única. Es posible definir la sequía en términos de las condiciones meteorológicas, hidrológicas, agronómicas y/o socio-económicas dominantes, razón por la cual existen un gran número de índices y parámetros asociados a ella (WMO, 1975).

En este caso nos referimos al concepto de sequía meteorológica, a saber, condiciones meteorológicas que provocan ausencia o reducción de la precipitación durante un período prolongado de tiempo (semanas, meses, años). Desde el punto de vista meteorológico es necesario el estudio de las sequías cortas (importantes para la agricultura) o muy prolongadas (relevantes para evaluar la disponibilidad de agua subterránea, la escorrentía y los niveles de reservas de agua).

La precipitación y la evapotranspiración son los principales factores que controlan la aparición y persistencia de las condiciones de sequía. Dificultades históricas para la cuantificación de la evapotranspiración han sugerido la definición de esquemas de clasificación utilizando solamente la precipitación. En este sentido, índices basados solamente en la precipitación han sido comparados con índices meteorológicos-climatológicos más complejos (Oladipio, 1985). En el presente caso, utilizamos el índice SPI (McKee *et al.*, 1993) que ha sido contrastado frente a índices de cálculo más complejo (Lloyd-Hughes and Saunders, 2002).

Para la realización del caso de estudio se han utilizado técnicas para la búsqueda y extracción de información y conocimiento a partir de grandes cantidades de datos almacenados. En la Figura 1 se muestran los pasos generales del proceso de descubrimiento del conocimiento utilizados (Penadés, 2005).

A partir de los repositorios de datos (información disponible) que puede estar almacenada en cualquier formato y soporte, se realiza un proceso de limpieza e integración de la misma, seleccionándose y transformándose los datos si fuera necesario. Posteriormente se construye el almacén de datos como una coleccción de datos orientados a temas, integrados, historiados y no volátiles que sirven de apoyo al proceso de toma de decisiones (Inmon, 1996). A partir de aquí empieza la evaluación de patrones y presentación del conocimiento, en este caso aplicamos la tecnología OLAP-Mining. En Han (1997) se propone OLAP-Mining como un mecanismo que integra técnicas propias de la tecnología OLAP (Codd, 1993) con las de minería de datos (Fay, 1996). Esta integración facilita la búsqueda de patrones o conocimiento interesante de forma multidimensional y a varios niveles de abstracción, puesto que las herramientas de análisis trabajan directamente sobre un cubo de datos construido a partir del almacén de datos.

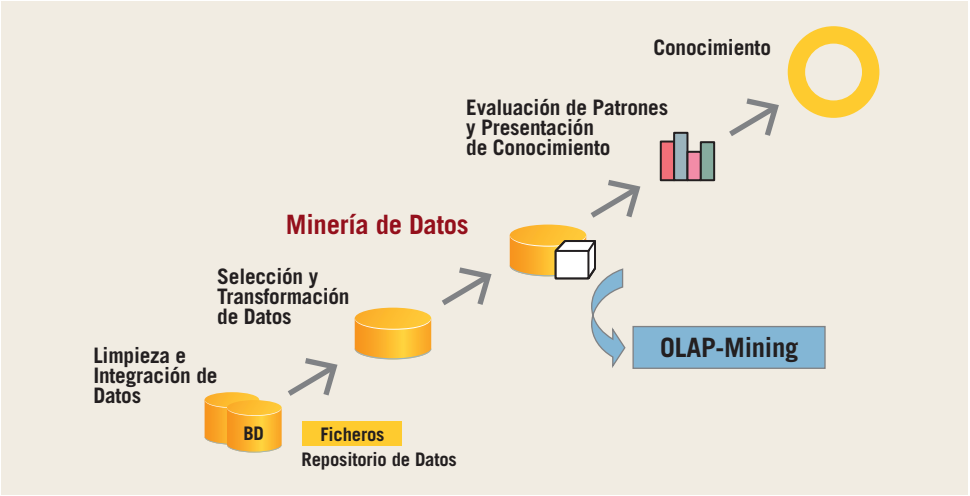


Fig. 1. Pasos del proceso de descubrimiento de conocimiento. (*Knowledge discovery in databases.*)

La Figura 2 muestra la arquitectura de 3 niveles del almacén de datos que tomamos como referencia para nuestro análisis. Como puede observarse, las herramientas de consulta e informes, de análisis y/o minería de datos, para la exploración y visualización de los datos del almacén se encuentran en el tercer nivel.

Con la definición del modelo y procedimientos el siguiente paso es aplicarlo a la monitorización y cuantificación de la sequía en diferentes áreas de España. En la Figura 3 se muestran los actores del modelo a desarrollar y el flujo de información.

Los datos climatológicos utilizados proceden de la red termopluviométrica de la AEMET (Agencia Estatal de Meteorología). A partir de los datos de precipitación se calcula el índice SPI siguiendo el método definido por McKee *et al.* (1993). Se consideran diferentes patrones sinópticos utilizando los reanálisis de los campos de 500 hPa y 850 hpa del NCEP/NCAR (Kistler *et al.*, 2001) (Figura 4).

A partir de los datos climatológicos se generan *grids* mensuales de precipitación, temperatura y SPI a diferentes escalas (García-Haro *et al.* 2008). Se utilizan entre 2.000 y 5.000 (dependiendo del período) estaciones distribuidas por todo el territorio con datos desde 1950 hasta la actualidad.

Se van incorporando al cubo de datos información procedente de diferentes instrumentos de teledetección (MERIS, MODIS, SEVIRI,…). Se considera fundamentalmente series de tiempo de FVC (*Fraction Vegetation Cover*) y LAI (*Leaf Area Index*) desde 2000 hasta 2008 (1 km, 8 días). Finalmente, se incorpora la imagen de tipos de vegetación (Figura 5).

Este es un ejemplo de un modelo de datos multidimensional y su gestión automatizada en una arquitectura de tres niveles.

La aplicación de esta metodología en el campo de la meteorología y la climatología es incalculable, se pueden incorporar cualquier tipo de información directa o indirecta (teleconexiones). El meteorólogo puede definir las condiciones y reglas de asociación según las características del estudio que se esté realizando. Debido a la incorporación de gran cantidad de información, este método debe ser introducido gradualmente con mínimos cambios.

Es de vital importancia la correcta y óptima parametrización de la base de datos. La

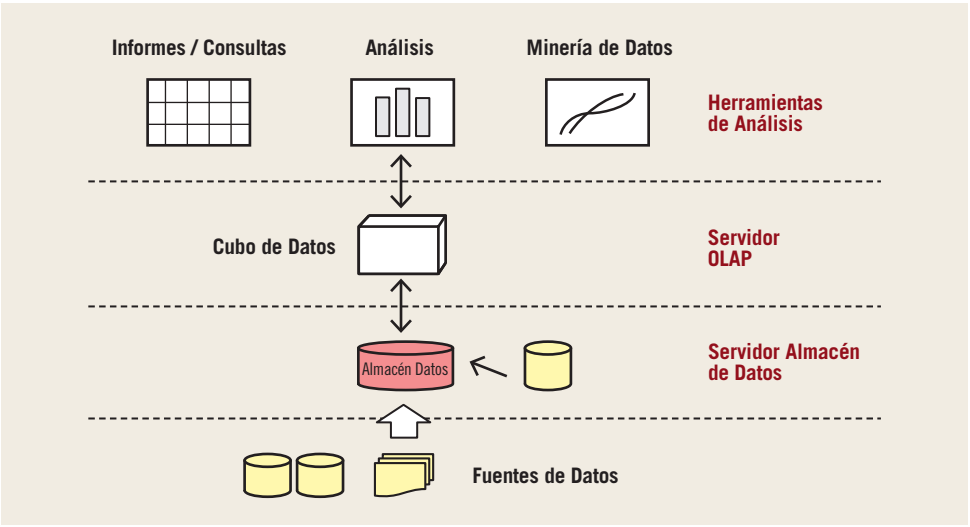


Fig. 2. Arquitectura de 3-niveles del almacén de datos. (3-*tier architecture in a data warehouse.*)

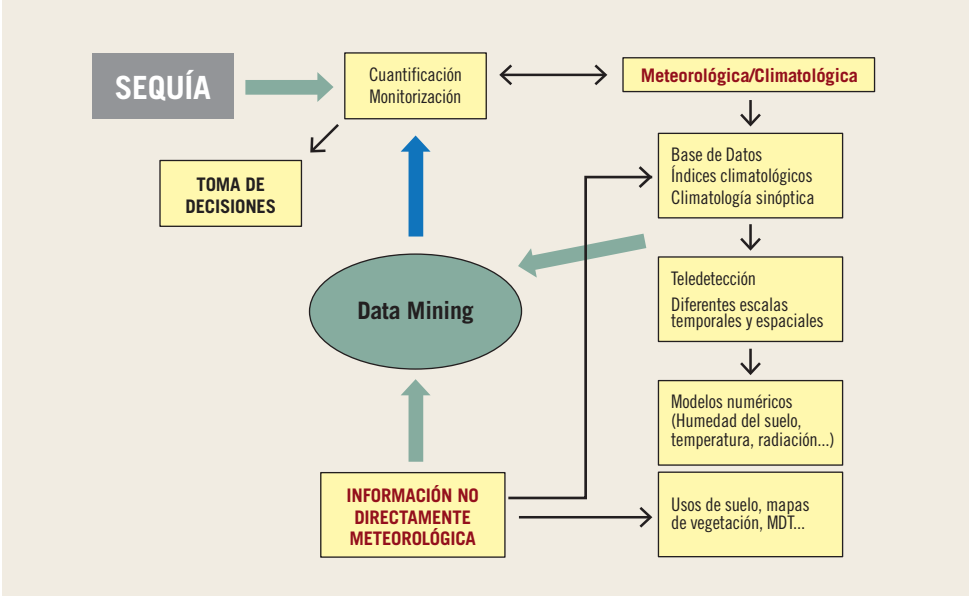


Fig. 3. Modelo de procesos.

técnica será mucho más eficiente si los datos son de una alta fiabilidad y de una máxima precisión. Este modelo de datos multidimensional nos permitirá de una forma eficiente y sencilla introducir parámetros oceánicos, más estacionarios, que afecten a la circulación general de la atmósfera, así como incorporar reanálisis del ECMWF. De esta forma podremos encontrar, por ejemplo, períodos de sequía precedidos por determinados valores del SOI, MEI, PNA, NAO.

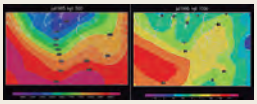


Fig. 4. Reanálisis NCEP/NCAR de 1.000 y 500 hgt correspondientes a julio de 1995.

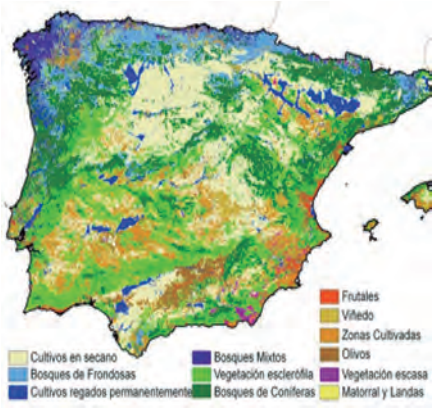


Fig. 5. Tipos de vegetación.

Bibliografía

Belda, F. (1997) "Climatología y teledetección en zonas forestales de la provincia de alicante. Aplicación a zonas incendiadas". Tesis Doctoral. Servei de publicacions de la Universitat de Valencia. ISBN: 84-370-3206-7.

Belda F. and **M.C. Penades**. (2010) "Applying Data-Minig techniques to study drought periods in Spain". 10th Annual Meetings of the EMS/8th ECAC. Vol. 7. EMS2010-444.

Codd, E.F., Codd, S.B., Salley, C.T. (1993) "Beyond Decision Support", Computer World, **27**.

Fayyand, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996) "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press.

García-Haro, F.J. Belda, F. and **Poquet, D.** (2008). "Estimation of climatological variables in Spain during 1950-2008 period using geostatistical techniques", 8th Annual Meetings of the EMS/7th ECAC. Abstracts. A-00319.

García-Haro, F. J., Belda, F., Gilabert Navarro, M.A, Meliá, J., Moreno, A., Poquet, D., Pérez-Hoyos, A., Segarra, S. (2008b), "Monitoring drought conditions in the Iberian Peninsula using moderate and coarse resolution satellite data", In *Proc. of the '2nd MERIS / (A)ATSR User Workshop*', ESA SP-666, European Space Agency, Noordwijk, The Netherlands, ISBN 978-92-9221-230-8, 7 pp.

Han, J. (1997) "OLAP-Mining: An Integration of OLAP with Data Mining", In Proc. IFIP Conference on Data Semantics, Leysin, Switzerland, 1-11.

Han, J. (2001) "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers.

Hayes, M., Svoboda, M., Wilhite, D.,A. and **Vanyarkho** (1999): "Monitoring the 1996 drought usuing SPI". *Bulletin of American Meteorology Society*, **80**, 429-438.

Inmon, W.H. (1996) "Building the Data WareHouse", John Wiley & Sons.

Lloyd-Hughes, B. and Saunders, M.A. (2002): "A drought climatology for Europe". *International Journal of Climatology*, **22**, 1571-1592.

Kistler R., Kalanay, E., Collins, W., Saha, S., White, G., Woollen, J.,Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., Van den Dool, H., Jenne, and **Fiorino, M.** (2001). "The NCEP-NCAR 50 year reanalysis: Monthly means CD-Rom and documentation". *Bulletin of the American Meteorology Society*, **82**, 247-267.

McKee, TB., Doesken, NJ. and **Kliest, J.** (1993): "The relationship of driught frecuency and duration to time scales". *Proceedings of the 8th Conference of Applied Climatology, 17-22 January, Anaheim*, CA. American Meteorological Society: Boston, MA; 179-184.

Oladipio, EO. (1985): "A comparative performance analysis of threee meteorological drought indices". *International Journal of Climatology*, **5**, 655-664.

Penadés, M.C. (2002) "Una Aproximación Metodológica al Desarrollo de Flujos de Trabajo". Tesis Doctoral. Universidad Politécnica de Valencia. Editorial: ProQuest. Information and Learning España.
I.S.B.N.: 0-493-82722-6, 264 pp.

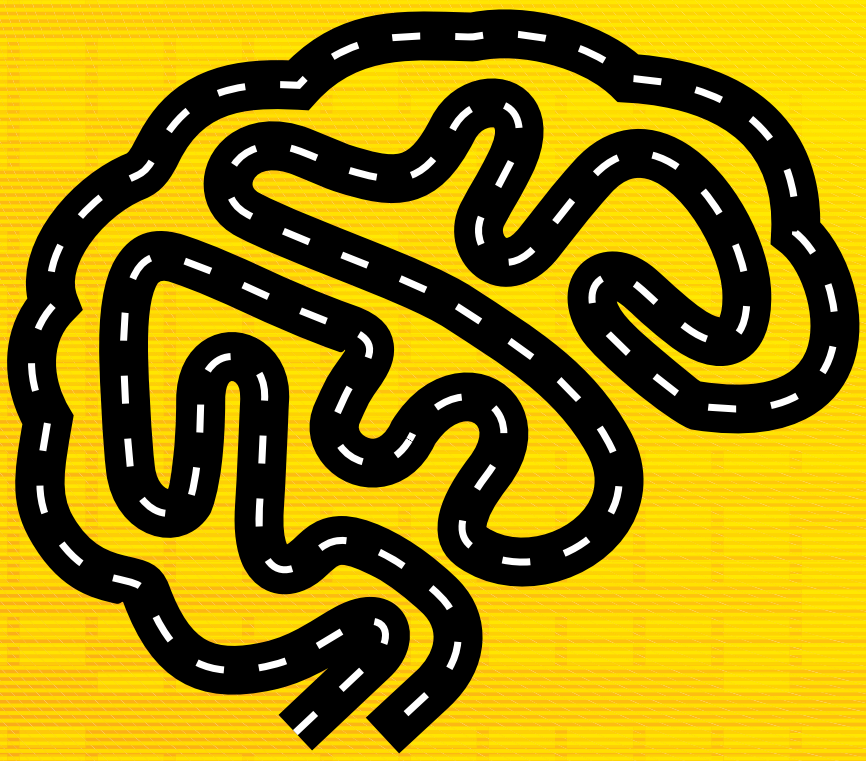
Penadés, M.C. (2005) "Workflow Mining". Minería de Datos: Técnicas y Aplicaciones. Ediciones de la UCLM, 187-212.

Thorn, H.C.S. (1966): "Some methods of Climatological Analysis". *WMO Technical Note*. nº. **81**,116-22.

WMO. (1975): "Drought and Agriculture". Technical Note. Nº 138. WMO – Nº 392.

Young, K.C. (1992): "A three-Way Model for Interpolating for Monthly Precipitation Values". *Monthly Weather Review*, **120**, 2561-2569.

fundacionareces.tv



Más de 2.000 conferencias magistrales de expertos en Salud, Innovación, Nuevas Tecnologías, Nanociencias, Astronomía, Biotecnología, Ciencias del Mar, Energía, Cambio Climático, Big Data, Economía, Economía de la Educación, Cambio Demográfico, Bioeconomía, Historia Económica...

FUNDACIÓN RAMÓN ARECES

Compartimos el conocimiento

Vitruvio, 5
28006 Madrid
España

www.fundacionareces.es
www.fundacionareces.tv

FUNDACIÓN
RAMÓN ARECES