Manual sobre utilidades del **big data** para bienes públicos

Entinema

Manual sobre utilidades del **big data** para bienes públicos

GOBERNA
Escuela de Política y Alto Gobierno
INSTITUTO UNIVERSITARIO DE INVESTIGACIÓN ORTEGA Y GASSET

Entinema

© Entimema, 2017
Fuencarral, 70
28004 Madrid
Tel.: 91 532 05 04
Fax: 91 532 43 34
www.entimema.com

Manual sobre utilidades del *big data* para bienes públicos

ISBN: 978-84-8198-983-0
IBIC: UNC
Depósito legal: M-20359-2017

"La máquina tecnológicamente más eficiente
que el hombre ha inventado es el libro".
Northrop Frye

Índice

Prólogo

Big data para bienes públicos. El sector público frente a la revolución de datos

Tamara Dull*

"¿Se encuentra la ciudad de Los Ángeles retrasada respecto al movimiento del *big data*?", preguntó un miembro de la audiencia. "Sí", respondió el primer panelista en el evento 2014 #TechLA. Francamente, estaba aturdida. Este panelista era un muy conocido, un aclamado experto en la industria tecnológica. Pero él no era de L.A, y ni siquiera era de California, así que, ¿cómo lo sabía?

Tuve que responder. Me sentí obligada a responder. Yo era de L.A., y estaba en el mismo panel sobre *big data*, que también incluía a un jefe de policía, un representante de la ciudad y un CEO de una empresa de seguridad de *software*.

"A pesar de que 'Frank' [el primer panelista] remarca algunas cosas interesantes", me dije, "no estoy de acuerdo con que L.A. llega tarde al *big data*, y aquí está por qué". Mientras hablaba, volví a enfatizar tres ideas claves que había planteado a lo largo de la discusión del panel, a saber: (1) el *big data* no es nuevo; (2) el *big data* no es fácil; y (3) hay espacio en el fenómeno de *big data* para todos.

Tanto si estás en Los Ángeles como en Latinoamérica, las tres ideas que compartí en el evento #TechLA en 2014 siguen siendo válidas hoy. Voy a presentar brevemente cada idea a continuación, pero luego le invito a leer este libro para aprender cómo las comunidades de América Latina (y L.A. de nuevo) están tratando con estas mismas realidades y cómo están utilizando el *big data* para mejorar tanto su eficacia organizativa como la participación de la comunidad.

› Realidad # 1: el *big data* no es nuevo

Si has estado trabajando con datos desde hace tiempo, podrías sentirte sorprendido al observar cómo el *big data* apareció sobre el año 2010-2011 como de la nada. Este súbito estallido nos dejó a muchos preguntándonos, "¿Qué es el *big*

* Directora de Tecnologías Emergentes en el Instituto SAS.

data y cómo es de diferente respecto a los datos que hemos estado tratando durante años?

Consideré lo siguiente: dicen que el 20% de todos los datos que se generan en estos días son datos estructurados de tipo transaccional, el tipo que hemos estado recolectando durante décadas y almacenado en nuestras bases de datos relacionales y almacenes de datos. El otro 80% es lo que llaman datos semiestructurados y no estructurados. También se conocen estos datos porque se ha estado trabajando con ellos durante décadas, es decir, lo que hemos estado almacenando en nuestras hojas de cálculo, documentos de Word, PDF, fotos, vídeos, datos de redes sociales, *weblogs* y la lista puede continuar.

Entonces, ¿qué ocurrió realmente en 2010-2011 para que hubiera tanto alboroto? La respuesta breve es que no se trataba tanto de los datos como de las nuevas y grandes tecnologías de datos que estaban asolando al sector privado. Muchas de estas tecnologías se han desarrollado utilizando *software* de código abierto, lo que hace que estas tecnologías sean gratuitas para cualquier organización, pública o privada.

Apache Hadoop, uno de los proyectos de código abierto más populares, se convirtió en el primer mensaje para el movimiento *big data*, ya que dio a las organizaciones la capacidad de almacenar y procesar todo tipo de datos estructurados, semiestructurados y no estructurados a un costo mucho menor. Los muros alrededor de nuestros datos habían comenzado a reducirse.

Hemos recorrido un largo camino en los últimos siete años. Hadoop sigue siendo una de las grandes tecnologías de datos en el núcleo de muchas iniciativas basadas en datos, pero no es el único "chico" en la ciudad. Las grandes tecnologías de datos continúan multiplicándose y evolucionando, año tras año, a medida que los proveedores de tecnología y la comunidad de código abierto trabajan juntos para apoyar las necesidades de los sectores público y privado.

› Realidad # 2: *big data* no es fácil

Uno de los primeros errores adoptados por las tecnologías *big data* fue suponer que, dado que el *software* era gratuito para descargar, sería fácil de implementar. Por supuesto, la instalación de un único proyecto de código abierto, como Apache Hadoop, es relativamente simple, pero como las organizaciones aprendieron rápidamente, Apache Hadoop no es una solución independiente. A menudo

requiere docenas de proyectos de código abierto adicionales (y productos de código cerrado) para ayudar a cumplir con los requisitos funcionales de cualquier iniciativa basada en datos.

Más allá de la implementación de estas tecnologías, una pregunta a menudo hecha por muchas organizaciones sobre *big data* es. "¿Por dónde empezar?". Para los recién llegados y expertos experimentados por igual, no hay atajo para encontrar la "señal" en el "ruido" actual del fenómeno del *big data*. Esta no es una tarea fácil. Es lioso y es complicado, y no dejes que nadie te diga lo contrario.

Para aquellos que apenas comienzan su gran viaje en el fenómeno del *big data* o para aquellos que ya han comenzado y están buscando alguna corrección en el rumbo, aquí están algunas prácticas recomendadas de alto nivel que las organizaciones han aprendido (a veces de la manera más difícil) en los últimos años:

1. Haga su tarea. Aprenda lo que otros han hecho y están haciendo con *big data* en los sectores público y privado. Asistir a grandes conferencias centradas en datos. Hable con otros que han tenido éxito y fracasado en sus iniciativas *big data*. Pista: aprenderás más de aquellos que han fracasado.
2. Identificar y priorizar las oportunidades. A medida que se desarrolle su comprensión del *big data*, comenzará a ver oportunidades en todas partes, dentro y fuera de su organización. Concentre su atención en aquellas oportunidades que proporcionarán un valor real a su organización y comunidad. Luego priorice sistemáticamente estas oportunidades, teniendo en cuenta el valor que aporta a la comunidad, el esfuerzo requerido y la dificultad técnica.
3. Obtener el liderazgo interno. Muchos proyectos vinculados al *big data* fallan porque se saltan este paso. Si el proyecto no se ajusta a las prioridades de liderazgo de su organización, su supervivencia puede verse comprometida. Para evitarlo, muestre su lista de oportunidades prioritarias (del paso 2) a sus funcionarios de alto nivel y obtenga su retroalimentación y apoyo.
4. Construir el equipo. Una vez que tenga la aprobación para el esfuerzo inicial, es hora de construir un equipo multidisciplinario. Este equipo puede incluir a: actores oficiales, expertos en la materia, profesionales de TI y recursos externos, según sea necesario. Esta es una gran oportunidad también para incorporar a los voluntarios de la comunidad. El tamaño y el alcance del proyecto ayudarán a dimensionar el equipo.

5. Implementar el proyecto inicial. Asegúrese de que el esfuerzo seleccionado esté bien definido. Se debe empezar en pequeño. Usted está buscando una victoria rápida para demostrar que el *big data* es bueno para la organización.
6. Compartir los resultados. Enlistar a las partes interesadas en la presentación de los resultados de su esfuerzo. Esto es cierto si el proyecto ha tenido éxito o ha fracasado. Es bueno ser considerado responsable, y es muy probable que obtenga más apoyo en el camino.
7. Haga un proyecto *post-mortem*. Al final del proyecto, reunir al equipo para llevar a cabo un *post-mortem* del proyecto. Identificar lo que funcionó bien, lo que no funcionó y las lecciones aprendidas. Identificar oportunidades de mejora para proyectos futuros.
8. Refinar los procesos y medir el valor. Sobre la base de la retroalimentación del equipo *post mortem* y el equipo de liderazgo del proyecto, perfeccionar los procesos y ajustar el equipo del proyecto en consecuencia. Dependiendo del tipo de proyecto, comience a medir el valor anticipado.
9. Inicie el siguiente proyecto. Ahora, con un proyecto *big data* en su haber, inicie el siguiente proyecto y repita el ciclo de compartir resultados (paso 6), procesos de refinación y valor de medición (paso 8).

Una iniciativa* big data *no se completará en un solo proyecto. Será una serie de proyectos que variarán en alcance y recursos cada vez.

› Realidad # 3: hay espacio en la gran mesa del *big data* para todos

Echemos un vistazo más profundo al paso 4 de la sección anterior: construya el equipo. Al considerar los desafíos que las organizaciones enfrentan en el mundo digital en evolución, los líderes y los representantes públicos han llegado a esperar que el *big data* impulse grandes cambios. Y probablemente lo hará. Pero será una progresión continua; un viaje que requiere un replanteamiento de las normas culturales.

Si bien el proceso y la tecnología ciertamente desempeñan un papel en esta evolución, el cambio en la cultura dependerá únicamente de las personas. Desde la estructura de la organización y los roles y responsabilidades definidas hasta los equipos que se convocarán y disolverán a medida que las iniciativas van y vienen, el componente personal del *big data* es lo que más importa.

Cuando construya un equipo con garantías de éxito para afrontar proyectos *big data*, tenga en cuenta estas cinco características:

1. No tiene que ser un trabajo interno. Su equipo *big data* debe incluir a profesionales de datos tanto internos como externos, tanto del sector público como privado.
2. Cruzará la división funcional. Su equipo *big data* tendrá una mezcla saludable de profesionales de negocios y técnicos. *Big data* no es un proyecto de TI disfrazado; es un proyecto organizacional con soporte de TI.
3. No es estático. Su equipo *big data* continuará cambiando, es decir, diferentes personas y número de personas debido a la rápida evolución de la industria del *big data* y a las oportunidades que ofrece.
4. El patrocinador del equipo cambiará. Cada iniciativa *big data* necesita tener uno o más patrocinadores sénior. El alcance de su iniciativa te servirá para analizar quiénes serán los mejores patrocinadores para el actual equipo *big data*.
5. Será diversa. Su equipo *big data* requiere de una amplia gama de habilidades y experiencia. Debería incluir personas de todas las edades, género y condición socioeconómica.

Esta última característica que se centra en la diversidad es la más importante. Cuando se considera la gran cantidad de datos involucrados en una iniciativa particular de *big data*, especialmente en el sector público, es fácil entender que los datos en sí serán diversos y representarán a ciudadanos de todas las edades, género y nivel socioeconómico. Un equipo diverso ayudará a darle voz a esos datos.

En la construcción de un equipo diverso, asegúrese de incluir un montón de mujeres. No estoy diciendo esto por ser mujer y amar el campo de la tecnología. Lo digo porque es lo correcto. Las mujeres no solo representan aproximadamente la mitad de la población, sino que son también las que están usando y comprando la tecnología que estamos creando para hacer todas nuestras vidas más fáciles. ¿Y no es este el objetivo final de la tecnología?

Otra razón para incluir a las mujeres es que tienen opiniones (¡a veces fuertes!) que merecen ser escuchadas. Se acercan a los problemas de manera diferente a los hombres, y ven el mundo a través de una lente diferente. ¿Qué mejor campo que la tecnología para que las mujeres participen?

Creo que todos somos conscientes de que las mujeres están subrepresentadas en el sector tecnológico y que, aunque se están haciendo esfuerzos por conseguir que las mujeres jóvenes se involucren y educen a través de oportunidades de STEM (ciencia, tecnología, ingeniería y matemáticas), ese esfuerzo hay que continuar impulsándolo.

Mientras tanto, os invito a que, al leer esta gran guía de *big data*, mantengáis vuestros ojos y oídos abiertos a ideas que puedan animar el interés de las jóvenes en vuestra vida. Este es un momento emocionante en la tecnología, y creo que esta gran revolución ha jugado un papel importante en acercar los datos de manera cercana y personal para todos nosotros. Todos tenemos un papel que desempeñar para mantener el impulso y ayudar a la próxima generación de hombres y mujeres jóvenes a abrazar y extender lo que ya está en movimiento. Hagamos nuestra parte.

Prólogo

Manuel Acevedo*

El tiempo pasa, las buenas conversaciones avanzan y las tecnologías vuelan. Hace ya más de diez años conocí a un jovencísimo y entusiasta Pablo Díaz. Comenzaba su participación en el programa UNITeS (United Nations Information Technology Services), anunciado en 2001 en el *Informe del Milenio* del entonces secretario general de la ONU Kofi Annan, y que gestionábamos con mi equipo en la agencia UN Voluntarios. UNITeS involucraba a voluntarios con conocimientos avanzados sobre las TIC en proyectos de desarrollo por todo el mundo. Pablo fue uno de los primeros voluntarios españoles de UNITeS, y empezamos por aquel entonces nuestra conversación sobre el impacto de la brecha digital y cómo poner las TIC al servicio del progreso de la humanidad. Nuestras conversaciones han seguido desde entonces, siendo para mí fuente de aprendizaje, estímulo y de satisfacción como comprobación de que hay mucho talento aplicado a solucionar los grandes problemas de la humanidad. Ahora a principios de 2017 Pablo no solo me cuenta sobre el *big data,* una de las herramientas más novedosas en el campo de las TIC para el desarrollo, sino que ha organizado la elaboración de una publicación en nuestro idioma tan interesante como necesaria sobre este relativamente desconocido tópico en el desarrollo internacional.

Los datos son uno de los pilares para la toma de decisiones. Sin datos no podemos conocer el nivel de pobreza en una sociedad o cuántas mujeres han muerto víctimas de la violencia machista. Sin embargo, en el umbral de la nueva agenda de desarrollo definida por los Objetivos de Desarrollo Sostenible de 2030, menos de la mitad de los países mantienen estadísticas fiables sobre algo tan elemental como nacimientos y muertes, y solo un 40% lo hacen con la violencia contra las mujeres. Por otra parte, muchos de los indicadores de desarrollo existentes provienen de trabajosas encuestas domésticas, con lo que a menudo las políticas públicas se basan en datos desactualizados —dos, tres y hasta cinco años atrás—. Como puedo constatar en mi trabajo[1], las carencias de estadísticas públicas confiables no solo ocurren en los países más pobres, sino también

* Consultor y experto en cooperación al desarrollo.

1. En tareas de planificación estratégica y evaluación de programas de desarrollo.

inclusive en países de desarrollo medio-alto que en principio poseen los medios para generar datos actualizados y confiables.

Por eso el nuevo fenómeno del *big data* resulta de gran valor para el sector del desarrollo internacional. En el ecosistema digital en el que nos movemos, se generan volúmenes fenomenales de datos todos los días. Alimentamos nuestra huella digital al usar el teléfono móvil, al pagar con una tarjeta bancaria, cuando hacemos una búsqueda por Internet, al añadir canciones a nuestras *playlists online*, cuando comentamos una noticia en redes sociales, al comprar un libro en una tienda virtual o cuando nos reclinamos en el sofá para terminar plácidamente el día con una de nuestras series favoritas en Netflix. Y no es exclusivo de los países ricos o las clases medias: los teléfonos móviles de miles de millones de habitantes se están convirtiendo en pequeñas computadoras conectadas a Internet, generando sus estelas digitales por todo el planeta como si fueran una flota gigante de minisatélites a ras de suelo. La gran pregunta es: ¿se puede aprovechar estas mareas de *bits* y *bytes* para el progreso de la humanidad? Y si es así, ¿cómo?

Este libro ayuda a entender las posibilidades reales del uso del *big data* para mejorar y extender los bienes públicos. También a relativizar los entusiasmos utópicos que han caracterizado otras olas tecnológicas (¿recuerdan cuando se decía que Internet democratizaría al mundo...?). Aparece en un buen momento, al inicio de la Agenda 2030 y cuando en ámbitos externos al desarrollo ya se han generado experiencias significativas sobre las que basarse. Creo que la lectura de este manual, resultará útil y didáctica a cualquiera que le interese o bien la temática del desarrollo o la de la ciencia de datos, aun cuando tenga experiencia en uno de los dos campos. Yo desde luego he aprendido mucho y he registrado algunas herramientas nuevas para futuros trabajos.

Como señala Paula Hidalgo Sanchís, directora del laboratorio de innovación Global Pulse de Naciones Unidas en Uganda, la revolución de los datos que se quiere incorporar ahora al desarrollo ya prospera en el sector privado desde hace más de una década[2] . El desafío consiste en trasladar y emplear ese *know-how* durante los próximos 15 años para ayudar a solucionar los problemas más acuciantes de la humanidad. Por eso me pareció muy relevante que el capítulo introductorio describa los tipos de usos del *big data* en el sector privado, como trampolín a los capítulos siguientes que exploran sus aplicaciones solidarias y para el desarrollo. Estos usos son: (i) para aprender de los usuarios e iterar el diseño de productos,

2. http://elpais.com/elpais/2017/01/16/planeta_futuro'1484566096_928831.html
Pulse Lab Kampala - http://www.unglobalpulse.org/kampala

(ii) para personalizar los productos a las necesidades de cada usuario, (iii) para personalizar la publicidad (léase "comunicación") a los intereses de cada usuario, (iv) para diseñar mejor los mercados, y, muy importante, (v) para controlar algoritmos y automatismos.

Al recorrer esta lista, la imaginación del lector seguramente se ha activado y empezado a visualizar cómo se adaptarían estos tipos de usos a entornos y organizaciones de desarrollo. Vamos por el buen camino. En los capítulos restantes el lector encontrará un tratamiento riguroso de los temas centrales sobre *big data* (como estándares, fuentes o el sensible aspecto de la ciberseguridad) junto con ejemplos y casos reales en varios países de la región (Chile, Colombia, Costa Rica) y de algunas organizaciones (como Telefónica o el Banco Mundial). Creo que supondrá una experiencia rica y amena de aprendizaje.

En mi caso, a lo largo de la lectura surgieron dos atributos del *big data* que encuentro particularmente relevantes y diferenciados para el desarrollo sostenible y que suplen notorias carencias en nuestro trabajo de planificación e iniciativas para el mismo: el valor de los datos obtenidos/accesibles **en tiempo real** y la exigencia de **no dejar a nadie atrás**. La velocidad de estas nuevas herramientas y fuentes de datos permiten acceder a datos "frescos" que no solo ayudan a planificar mejor, y no con información de 2 o 3 años atrás, sino que permiten un monitoreo que lleva a corregir el rumbo de políticas, programas y proyectos cuando sea necesario y sin esperar meses o inclusive años. Por otro lado, la amplia capilaridad de estas metodologías (recogiendo datos de millones de individuos en un país determinado) permite generar información sobre minorías y grupos vulnerables que suelen estar infrarrepresentados en los datos colectados y disponibles, y por ello son virtualmente invisibles para las políticas de desarrollo (como las personas con discapacidad o poblaciones indígenas, que conjuntamente superan el 20% de la población en Latinoamérica). Como afirma el informe de un grupo de expertos para la ONU sobre la revolución de datos, "Nunca más debería ser posible decir 'no lo sabíamos'. Nadie debería ser invisible"[3].

Si bien es necesario aprovechar el potencial del *big data* para el desarrollo, me recordaba hace unos días un amigo que ha trabajado en la CEPAL sobre la temática de sociedad de la información que lo que realmente importa es contar con los datos necesarios para guiar las políticas de desarrollo, su implementación y su seguimiento, sea cual sea la metodología. Al respecto, este pasado enero se

3. Un mundo que cuenta:
http://repositorio.cepal.org/bitstream/handle/11362/37889/1/UnMundoqueCuenta.pdf

llevó a cabo en Cape Town el primer Foro Mundial de Datos de la ONU, con un lema claro: "No podemos lograr lo que no somos capaces de medir". En esa bella ciudad sudafricana, así como en muchos otros foros durante los últimos 3 o 4 años, actores de desarrollo de todo el mundo han expresado la imperiosa necesidad de disponer de datos adecuados cualitativa/cuantitativamente para poder medir los 230 indicadores en los que se desagregan los 17 Objetivos de Desarrollo Sostenible (ODS). Parafraseando a Bill Clinton, "¡son los datos, estúpido!".

En cualquier caso no debemos suponer que la revolución de los datos será beneficiosa para el desarrollo sin más. Solo lo será si se convierte en una revolución abierta y extendida. Así como cada paquete de *big data* es una acumulación de muchos *bytes* de datos, es necesario que cada actuación de desarrollo se responsabilice en generar datos usando metodologías comprobadas y fiables. Y también que las capacidades para la obtención y el análisis de datos se multipliquen acorde a las necesidades previstas para la implementación de la Agenda 2030. Entonces sí podremos decir que el *big data* realmente contribuye al *big* desarrollo.

Nota del coordinador del manual

Pablo Díaz*

Ha sido un inmenso honor coordinar esta obra coral dedicada a descifrar y aclarar el fenómeno tecnológico del *big data* en nombre de la Fundación Ortega Marañón. Concentrar tanto talento en un mismo manual no ha sido tarea fácil, si bien, la sencillez y humildad en el trato, así como el esfuerzo y dedicación de cada uno de los coautores ha facilitado mi labor sobremanera. Mi más sincero agradecimiento a todos ellos por hacerme partícipe de este libro.

Cuando desde la Fundación Ortega-Marañón decidimos diseñar y coordinar un manual sobre el fenómeno del *big data* nos fijamos dos objetivos claros:

1. Debía ser escrito por expertos de primer nivel en la materia y por protagonistas que tuvieran o hubieran tenido responsabilidades o cargos públicos vinculados con el ámbito tecnológico.
2. El estudio de casos debía tener una presencia destacada en el manual.

Considero que los dos objetivos se han logrado de manera satisfactoria y son estos mismos fines los que definen la estructura de esta obra.

Contamos con un primer bloque introductorio donde se analizan y aclaran aspectos de primer orden a la hora de desarrollar iniciativas o proyectos *big data*. Como ocurre con muchas modas tecnológicas, el *big data* ha generado cierta oscuridad en su definición, lo cual puede llevar a confusión en su aplicación en el sector público, provocando incertidumbre y dudas cuando se trata de entender cómo estos conceptos pueden beneficiar a una organización. En este apartado, los expertos proponen las claves para entender perfectamente este fenómeno, la importancia de transmitir una proposición de valor en cualquier tecnología que quiera introducirse en las organizaciones públicas y la necesidad de una mayor gobernabilidad en el control de los datos, la calidad de los mismos y su coherencia semántica.

* Colaborador de Goberna América Latina.

Dentro de este primer bloque, también se dan a conocer a los líderes y responsables públicos los aspectos metodológicos y las herramientas que deben tener en cuenta a la hora de incorporar el *big data* a sus organizaciones, dando asimismo respuesta a cuestiones relacionadas con la seguridad, la privacidad, los estándares, la madurez organizativa, los nuevos puestos de trabajo generados por esta industria y otros elementos que no dejarán indiferente al lector.

El segundo bloque de esta obra está compuesto, en su totalidad, por el estudio de casos a lo largo y ancho del espacio iberoamericano, combinando experiencias desde lo público (ejemplos de Chile, Costa Rica, Colombia, Guatemala o la ciudad de Los Ángeles) y desde lo privado (LUCA o Telefónica), así como también desde organismos de tercer sector (NESTA o Data Pop Alliance). Cada uno de estos casos está redactado desde su propia experiencia, de modo que el lector podrá afrontar su lectura en el orden y prioridad que desee, e incluso, si su conocimiento sobre *big data* ya es amplio, iniciar la lectura de este manual por alguno de los casos que se describen.

Sin duda alguna, el futuro de la prestación de servicios públicos se presenta apasionante. Las organizaciones públicas tienen ante sí la oportunidad de predecir y dar respuesta a las necesidades y demandas de sus ciudadanos de manera mucho más eficaz, eficiente y proactiva gracias a tecnologías como el *big data*. Asimismo, la ingente cantidad de información y datos que seremos capaces de almacenar, procesar y analizar dará lugar a nuevas estructuras organizativas y perfiles profesionales que ya, hoy en día, pueden considerarse una realidad.

No me queda más que agradecerte que tengas este manual entre tus manos (ya sea en papel o en formato digital), y esperar que disfrutes con su lectura.

Bloque I
Introducción al *big data*

Como ocurre con muchas modas tecnológicas, el *big data* ha generado cierta oscuridad en su definición, lo cual puede llevar a confusión en su aplicación en el sector público, generando incertidumbres y dudas cuando se trata de entender cómo estos conceptos pueden beneficiar a la organización. En este bloque, los expertos propondrán las claves para entender perfectamente este fenómeno, la importancia de transmitir una proposición de valor en cualquier tecnología que quiera introducirse en las organizaciones públicas, y la necesidad de una mayor gobernabilidad en el control de los datos, la calidad de los mismos y la coherencia semántica.

En este bloque, también daremos a conocer a los líderes y responsables públicos los aspectos metodológicos y las herramientas que deben tener en cuenta a la hora de incorporar el *big data* a sus organizaciones. A lo largo de este bloque daremos respuesta a cuestiones tan importantes como, ¿aporta un valor cuantificable la incorporación de *big data* que justifique el esfuerzo y los recursos invertidos en el desarrollo de esta tecnología?, ¿cómo definirías indicadores o métodos de valoración para su medición?, etc. Asimismo, se darán a conocer herramientas que cubren todo el ciclo de vida de un proyecto de *big data*.

Capítulo 1

La revolución del *big data* en el sector privado y sus implicaciones para el sector público

Miguel Socías*

Durante los últimos 20 años hemos presenciado una verdadera revolución digital, primero a través de la expansión de Internet, y luego a través de la masificación de los teléfonos inteligentes. Se estima que un 42% de la población mundial está conectada a Internet y que alrededor de un tercio tiene un teléfono inteligente. Y ambas cifras —el número de personas conectadas a Internet y el número de usuarios de teléfonos inteligentes— siguen creciendo a una tasa cercana al 10% al año[1].

Esta revolución digital ha tenido profundas consecuencias en todo ámbito de nuestras vidas. Servicios digitales coordinan hoy en día nuestras comunicaciones (Skype, WhatsApp, etc.), nuestro transporte (Uber, Didi, etc.), nuestros entretenimientos (Netflix, YouTube, Minecraft, etc.), nuestras compras (Amazon, Alibaba, etc.), nuestras amistades (Facebook, Instagram, Snapchat, etc.), nuestras vidas profesionales (LinkedIn) y nuestro conocimiento (Google, Baidu, etc.), entre otros. Cada una de estas plataformas digitales se enfoca en un dominio específico, pero todas ellas comparten una característica común: el uso de big data en sus operaciones diarias.

En este primer capítulo introductorio analizamos cómo el big data está siendo utilizado en el sector privado para informar las tomas de decisiones, los elementos que pueden ser adoptados en el sector público y los desafíos involucrados en este proceso. Capítulos posteriores abordan en mayor profundidad diferentes aspectos mencionados en este primer capítulo.

* Científico de datos en eShares y Phd en Educación Internacional Comparada por la Universidad de Standford.

1. Ver informe *Internet Trends 2016* de Kleiner Perkins Claufield Byers (http://www.kpcb.com/Internet-trends).

¿Qué es el big data?

La primera pregunta que surge al mencionar este nuevo concepto es cómo se define el big data. ¿Qué grande debe ser una base de datos o el número de usuarios de una plataforma para ser considerado big data? A pesar de que no existe una respuesta única a esta pregunta es importante tener presente que el orden de magnitud de big data se mide en *terabytes* y *petabytes*, no en *gigabytes*. Es decir, big data es un orden de magnitud mayor a lo que antes de la revolución digital considerábamos grandes bases de datos.

Otra estrategia para identificar big data es a través de las herramientas utilizadas para analizar estos datos. *Big data* ha traído consigo un sinnúmero de nuevas tecnologías que han tenido que ser desarrolladas para analizar estas grandes bases de datos. Un hito histórico en este dominio fue la publicación del *paper* sobre MapReduce de Google en 2004, en el cual esta empresa compartió con el público general una de sus estrategias para analizar grandes bases de datos[2]. Además de MapReduce, otras tecnologías importantes para manejar datos de estas dimensiones incluyen Presto (de Facebook)[3] y Spark (de la Fundación Apache)[4], por mencionar solo tres ejemplos. Es decir, un buen indicador de que estamos en presencia de big data es cuando los métodos tradicionales como bases de datos centralizadas y hojas de cálculo no son capaces de procesar estos datos, y es necesario usar tecnologías como MapReduce, Presto y/o Spark.

¿Cómo surge el big data?

El big data surge principalmente por tres razones. La primera es el gran número de usuarios de estas plataformas. Como se mencionó anteriormente, más de un 40% de la población mundial tiene acceso a Internet, lo que se traduce en un mercado potencial enorme para estas plataformas. A fines de 2015 las principales redes sociales presentaban las siguientes cifras:

- Facebook: más de 1.500 millones de usuarios.
- YouTube: más de 1.000 millones de usuarios[5].
- WhatsApp: más de 900 millones de usuarios.
- Facebook Messenger: más de 700 millones de usuarios.

2. https://research.google.com/archive/mapreduce.html
3. https://prestodb.io
4. https://spark.apache.org
5. https://www.youtube.com/yt/press/en-GB/statistics.html

- Instagram: más de 400 millones de usuarios.
- LinkedIn: más de 380 millones de usuarios.
- Twitter: más de 320 millones de usuarios.
- Skype: más de 300 millones de usuarios.
- Snapchat: más de 100 millones de usuarios.

Además del gran número de usuarios, la intensidad de uso que se le dan a estas plataformas es igualmente impresionante. A continuación, se listan algunas cifras de uso:

- El usuario promedio de Facebook usa esta plataforma por más de 15 horas al mes[6].
- Un 60% de los usuarios de Instagram usan esta plataforma todos los días[7].
- Cada día se comparten alrededor de 58 millones de fotos en Instagram[8].
- Cada minuto se suben más de 400 horas de vídeo a YouTube[9].

Respecto a Latinoamérica, las personas conectadas a Internet en esta región del mundo son particularmente ávidos usuarios de las redes sociales. Más del 95% de la población con conexión a Internet en México, Brasil y Argentina usa redes sociales[10]. De hecho, estos tres países son también los países con el mayor número de horas de uso de redes sociales del mundo después de Filipinas, con 3,3 horas de uso al día en Brasil y 3,2 horas en México y Argentina[11].

Estas cifras nos ayudan a dimensionar el tamaño de estas plataformas y la intensidad de uso de estas. Y todas las fotos, vídeos y comentarios se van acumulando minuto a minuto, generando lo que ahora conocemos como big data.

La segunda razón por la que surge el big data es por la gran concentración de mercado de estas plataformas. Si existieran 1.000 Facebooks o YouTubes en el mundo, estas grandes bases de datos estarían distribuidas en un número mayor de empresas y cada una de ellas tendría bases de datos de menor tamaño. Pero dado que se trata de medios digitales, el mercado en el que se compite es global y las plataformas ganadoras tienden a ser mundiales. No importan fronteras entre países o continentes, ni barreras de idioma. Una plataforma digital como Facebook compite con posibles redes locales que puedan surgir en Filipinas o

6. Ver informe *Internet Trends 2016* de Kleiner Perkins Claufield Byers.
7. http://blog.instagram.com/post/146255204757/160621-news
8. http://www.statisticbrain.com/instagram-company-statistics/
9. http://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/
10. http://www.comscore.com/content/download/13029/267601/
11. http://www.slideshare.net/wearesocialsg/digital-in-2016

Indonesia. Y estas plataformas globales como Facebook, WhatsApp o YouTube —entre otras— son formidables competidores con grandes habilidades técnicas en el manejo de datos. Ante esta realidad, la estrategia de plataformas locales ha sido enfocarse en mercados menos atractivos para estos grandes competidores o posicionarse estratégicamente desde un comienzo a desarrollar un producto para luego ser adquirido por alguna de estas grandes empresas. El análisis dinámico de mercado que surge es el de unas pocas empresas con gran concentración de los usuarios mundiales en el largo plazo.

Pero además de tratarse de un mercado global, el "efecto de red" (*network effect*, en inglés) también contribuye de manera importante a este equilibrio a largo plazo de unas pocas empresas ganadoras. El efecto de red surge cuando el beneficio de ser parte de una red aumenta a medida que crece el tamaño de la red[12]. Es por ello que ser parte de Facebook tiene un beneficio mayor para cada usuario que ser parte de una red idéntica pero mucho más pequeña. Es decir, más allá de las capacidades técnicas de una plataforma, el solo hecho de que la gran mayoría de nuestras familias y amigos sean parte de esa red incrementa nuestra probabilidad de unirnos a ella. Es por ello que redes más grandes tienden a crecer incluso más que redes más pequeñas, acaparando el mercado global a largo plazo.

Una excepción a esta concentración global ha sido el caso de China, en el que una mezcla de políticas estatales que favorecen la economía doméstica, altas capacidades técnicas de empresas chinas y un mercado local muy grande ha erosionado las ventajas inherentes de las plataformas globales. Empresas como Uber, Google y Facebook han sido considerablemente menos exitosas en el mercado chino que en otros mercados, y han decidido enfocar sus recursos en otras economías emergentes como India e Indonesia. China es claramente un caso en el que los efectos de red locales suficientemente grandes, políticas gubernamentales y talentos técnicos han sido capaces de crear un nuevo equilibrio a largo plazo con un mayor número de empresas ganadoras a nivel mundial.

La tercera razón por la que surge el big data se debe a la importancia que tienen hoy en día los datos para el aprendizaje. En la siguiente sección analizamos este fenómeno.

12. La "Ley de Metcalfe" propone que el valor de una red —en particular una red de telecomunicaciones— es igual al cuadrado del número de sus usuarios (https://en.m.wikipedia.org/wiki/Metcalfe%27s_law).

¿Cómo se aprende hoy en día?

Experimentar en el mundo real es muy caro. ¿Cuánto aumentaría el flujo vehicular si en lugar de construir una carretera por el trayecto A utilizamos el trayecto B? En el mundo real no sería económicamente viable construir ambas carreteras para poder comparar las cifras de vehículos. Es decir, en el mundo real es muy costoso aprender de la experimentación. Pero el costo de experimentar en el mundo digital es marginal. En una plataforma digital podemos construir rápidamente una versión alternativa del producto y testear su efectividad. Dado este bajo costo de la experimentación, el desarrollo de productos se lleva a cabo hoy en día de manera iterativa y aprendiendo de la reacción de los usuarios finales. Todas estas plataformas globales se han convertido en verdaderos laboratorios de ciencia, formulando un sinnúmero de hipótesis simultáneas, introduciendo cambios en sus productos a grupos aleatorios de usuarios y finalmente sacando conclusiones de los datos de uso de sus clientes. Estas plataformas son ahora primordialmente organizaciones de aprendizaje.

Dado el rol central que juegan los datos en la mejora de los productos y servicios, estas plataformas digitales guardan cada interacción de los usuarios con sus páginas web y aplicaciones móviles. Cada página visitada y cada *click* se guarda para después ser usado en análisis estadísticos. Al igual que en un laboratorio, cada uno de estos usuarios ha estado expuesto a una cierta intervención (como el diseño A *versus* el diseño B del producto) y su interacción en la plataforma es usada para identificar el efecto que tiene esa intervención en aumentar la probabilidad de venta o el tiempo de uso de la plataforma, entre otros objetivos. Dada esta experimentación, la dimensionalidad de los datos almacenados es aún mayor, ya que no existe una sola versión del producto, sino que diferentes usuarios han usado un producto diferente en distintos momentos del tiempo. Todas estas diferentes versiones —con el grupo de usuarios expuestos a cada una de ellas— tienen que ser almacenadas para su análisis posterior. Es por ello que big data también es consecuencia del modelo de aprendizaje a través de la experimentación[13].

13. Para poder hacer inferencias sobre el efecto causal de una variable sobre otra —como, por ejemplo, el cambio de diseño de un producto en la intensidad de uso de este— es necesario poder asignar el nuevo diseño a un grupo aleatorio de usuarios. De lo contrario la relación entre ambas variables puede ser meramente correlativa, producto de la autoselección de diferentes tipos de usuarios a diferentes versiones.

Los otros usos del big data

El lector atento habrá notado en el párrafo anterior que para *linkear* (o relacionar) usuarios a versiones específicas del producto hay que poder identificar en primer lugar a cada usuario en la plataforma. Si se expone a un usuario a un cierto diseño y ese usuario vuelve a interactuar dos días más tarde con la plataforma, ¿cómo sabemos de qué usuario se trata y a qué diseño fue expuesto? El guardar *cookies* en el *browser* de los usuarios (o usar sus "huellas digitales" como su dirección IP, modelo de computador, tipo de *browser*, estado de la batería del teléfono, etc.) para su posterior identificación es una práctica ampliamente adoptada por empresas. En aplicaciones móviles los usuarios son identificados simplemente con sus credenciales de acceso.

Identificar a los usuarios no solo es necesario para realizar análisis estadísticos de impacto, sino también para personalizar los productos. Los nuevos libros que Amazon nos muestra dependen de nuestra historia de libros adquiridos anteriormente en esa plataforma. Esto mismo sucede con las películas recomendadas por Netflix, la música en Spotify y los posibles nuevos contactos en LinkedIn. Estos productos son hoy en día altamente personalizados y moldeados a los intereses de cada uno de nosotros.

Además de la necesidad de identificar usuarios para poder evaluar los diferentes diseños y personalizar productos, muchas de estas plataformas también identifican a sus usuarios porque sus modelos de negocio así lo exigen. Plataformas de información que usan publicidad *online* como su modelo de negocio (como Google y Facebook, por ejemplo) son particularmente ávidas a este proceso de identificación personal. Al igual que los productos, la publicidad también está siendo personalizada a la medida de nuestros intereses, y personas de diferentes características están siendo expuestas a publicidad de distintas empresas. Esta publicidad personalizada ha tomado la forma de remates *online* en tiempo real. En cada momento se analizan los usuarios conectados a estas plataformas y sus características, y se rematan los espacios de publicidad disponibles. Es decir, el big data que surge del almacenamiento de características personales también se debe en parte al modelo de negocio basado en publicidad de muchas de estas plataformas.

Nótese la importancia que tiene el big data en el diseño de nuevos mercados. El uso de big data en el mercado de la publicidad *online* es solo un ejemplo. Uber, Lyft y Didi han creado un mercado en el que el precio de cada servicio de transporte se ajusta a las condiciones de mercado en cada momento del tiempo (es

decir, depende del número de conductores y clientes en una zona geográfica específica en un momento determinado). Otros ejemplos de "mercados spot" (es decir, en tiempo real) son el transporte aéreo, el alojamiento en casas y departamentos en Airbnb, el remate de productos en eBay y ahora incluso productos en Amazon cuyos precios estamos viendo variar día a día. El uso de big data es fundamental para diseñar correctamente estos mercados, ya que los precios dinámicos deben ser fijados de tal manera que eliminen cualquier exceso de demanda u oferta en cada momento del tiempo.

Ya hemos mencionado cuatro usos de big data en el sector privado: a) para aprender de los usuarios y seguir iterando el diseño de los productos, b) para personalizar los productos a los intereses de cada usuario, c) para personalizar la publicidad a las características e intereses de cada usuario, y d) para diseñar correctamente la estructura de los diferentes mercados *online* y fijar sus precios dinámicos. Pero ahora existe además una quinta razón, el desarrollo de algoritmos e inteligencia artificial. En la siguiente sección analizamos este nuevo fenómeno.

La automatización de las decisiones

El primer sistema de piloto automático apareció en la aviación en 1912[14]. Después de este hito hemos visto que esta idea de automatizar decisiones se ha masificado a todo tipo de industrias y trabajos durante los últimos cien años. Google, Uber y la gran mayoría de los fabricantes de automóviles están implementando sistemas de piloto automático con la idea de liberarnos del tedioso trabajo de conducir. El piloto automático también está siendo introducido en el transporte público. Además de trenes de metro sin conductor, buses urbanos están comenzando a operar sin chóferes[15]. En agricultura, tractores con piloto automático ya están trabajando nuestros campos sin intervención humana directa[16].

La automatización de la conducción de un vehículo terrestre, marítimo o aéreo es por supuesto solo un ejemplo de lo que estamos viendo en todas las industrias y trabajos. Desde pronósticos médicos[17] a recomendaciones de inversión[18] y servicios de asesoría legal[19], todo tipo de decisiones están siendo automatizadas. Al

14. https://en.wikipedia.org/wiki/Autopilot
15. http://yle.fi/uutiset/helsinki_rolls_out_driverless_bus_pilot/9099541
16. http://www.trimble.com/Agriculture/autopilot.aspx
17. https://www.ibm.com/watson/health/
18. https://investorjunkie.com/35919/robo-advisors/
19. http://arstechnica.co.uk/tech-policy/2016/08/donotpay-chatbot-lawyer-homelessness/

igual que la revolución industrial hizo más eficiente el trabajo manual usando líneas de producción, la revolución digital está haciendo más eficiente el trabajo cognitivo a través del aprendizaje automático (*machine learning*, en inglés). Y esta gran automatización de decisiones ha sido gracias a la explosión de los datos disponibles no solo para testear hipótesis, sino también para crear nuevos algoritmos de decisión que surgen de los mismos datos. Es decir, en estricto rigor, ya no es necesario definir *a priori* cómo un sistema debe tomar decisiones, sino que estas pueden adaptarse en el tiempo al flujo de datos sin intervención humana directa (lo que se conoce como "aprendizaje sin supervisión", *unsupervised learning*). El big data es el fenómeno responsable, detrás del nuevo impulso de inteligencia artificial, el cual ha surgido en la ausencia de un descubrimiento revolucionario en inteligencia artificial a nivel simbólico[20].

Los emprendedores son los agentes clave de nuestro tiempo encargados de implementar los sistemas de automatización de decisiones. La gran mayoría de los emprendimientos en Silicon Valley, Berlin, Shanghai y el resto del mundo consisten en un mecanismo de captura de datos masivos, un algoritmo que transforma estos datos en una decisión automatizada y un modelo de negocio que financia este proceso. Waze, por ejemplo, toma datos masivos de tránsito de cada uno de nosotros y entrega decisiones de conducción. Netflix toma nuestros datos de uso para decidir qué nuevas series de televisión y películas producir en el futuro. La gran fortaleza de este proceso es que estos sistemas son tremendamente eficientes, y solo mejoran en el tiempo (gracias al aprendizaje sin supervisión). La desventaja —a juicio personal— es que la intervención humana directa se hace menos necesaria y se fortalece la tendencia de la desigualdad del ingreso en la población. La siguiente sección aborda esta importante discusión.

› El sector público en los tiempos del big data

Los grandes temas que debe abordar el sector público en Latinoamérica frente al big data son básicamente dos: a) cómo utilizar estas nuevas herramientas desarrolladas durante los últimos años en el sector privado y b) cómo el big data afecta el rol del Estado.

La primera pregunta es relativamente simple de abordar. Los gobiernos latinoamericanos deben adoptar estas nuevas tecnologías para diseñar políticas

20. Actualmente no existe consenso en el área de la inteligencia artificial sobre si el nuevo auge de la disciplina representa un cambio cualitativo de nuestro entendimiento de los procesos subyacentes a la inteligencia humana.

públicas basadas en evidencia empírica, iterar el diseño de estas políticas para perfeccionarlas en el tiempo, e incluso personalizar el servicio público ofrecido a sus ciudadanos. Por ejemplo, el Ministerio de Vivienda debiera estar capacitado para usar el big data para predecir futuros crecimientos urbanos para ampliar la oferta de vivienda en ciertas zonas geográficas específicas. El Ministerio de Salud debería —si no ha comenzado aún— integrar los datos de todos sus hospitales y clínicas para analizar la efectividad de intervenciones médicas. Análisis predictivos también debieran ser usados en el Ministerio de Educación para identificar de manera temprana a estudiantes en riesgo de abandono escolar y que necesiten de un mayor apoyo académico y emocional de sus escuelas.

En Estados Unidos y Europa análisis predictivos están siendo desarrollados usando big data en el área de la seguridad pública, donde indicadores líderes ayudan a identificar posibles fraudes y crímenes, incluso antes de que estos sean llevados a cabo. Ciertamente estamos entrando en una época que hace pocos años considerábamos solo ciencia ficción. Y debido a estas nuevas capacidades técnicas, cada país latinoamericano deberá establecer sus propios límites que quiera tener en cuanto al uso del big data, por una parte, y al derecho a la privacidad de sus ciudadanos, por otra. Este debate ya se ha llevado a cabo en otros países y la tendencia en el tiempo ha sido hacia un mayor —no menor— monitoreo de los ciudadanos a cambio de una promesa de mayor seguridad pública. A medida que los países de Latinoamérica adopten sistemas de big data, esta misma discusión pública tendrá que ser abordada por su ciudadanía.

El sector público también cuenta con un monopolio en la recaudación de los impuestos. ¿Se están usando sistemas de big data y *data science* en países de América Latina para analizar los patrones de recaudación de impuestos con tal de disminuir la evasión tributaria? ¿Se está utilizando esta nueva herramienta para modernizar el código tributario de modo que se minimice el impacto negativo de los impuestos en la creación de empleo? *Big data* no solo debiera aumentar la productividad de los emprendedores de Silicon Valley, sino también debe convertirse en una nueva herramienta a disposición de los funcionarios públicos de América Latina.

Pero además de la adopción del big data en el sector público para modernizar las políticas públicas, el segundo tema relevante es cómo debemos abordar el desafío futuro que afrontará el mercado laboral mundial[21]. La explosiva

21. Por supuesto que es imposible vislumbrar todas las consecuencias que tendrá el uso del big data en nuestra sociedad. Pero en nombre del mismo sistema iterativo que vemos hoy en día en

productividad que tendrán trabajadores altamente calificados junto a grandes bases de datos y algoritmos de decisión tendrá un impacto brutal sobre el empleo en industrias, comenzando probablemente con los coches autónomos y la pérdida de trabajo de taxistas y camioneros alrededor del mundo. Y si dado que big data está afectando todo tipo de empleos e industrias, cabe la pregunta de cuántos puestos de trabajo se necesitarán en el futuro, incluso a nivel mundial.

Es por ello que el tema tributario estará en el centro del debate público del big data a medio plazo y, en especial, el debate de un "ingreso básico" (*basic income*, en inglés) que puedan recibir todos los ciudadanos de un país. A modo de ejemplo, Y Combinator, una de las incubadoras de emprendimiento más importantes del mundo, está llevando a cabo actualmente un experimento entregando un ingreso básico a familias de escasos recursos en Oakland, California[22]. Otro experimento de *basic income* está siendo llevado a cabo en la actualidad en Finlandia. La idea central es estimar empíricamente las consecuencias de entregar un ingreso básico a familias de escasos de recursos. Si en el futuro el trabajo de unos pocos pudiese financiar la vida de muchos, el sistema tributario pudiera en un principio considerar un ingreso básico financiado con los frutos de los datos de todos los ciudadanos.

Como señala el *Harvard Business Review*, "un país no es una empresa"[23]. De hecho, un país no es siquiera una industria, ni un conjunto de industrias. El espacio público de un país debe cuidar y respetar los intereses de todos sus ciudadanos, para lo cual es necesario tener una mirada de equilibrio general, no parcial. Y esta visión global es aún más desafiante en nuestros tiempos, ya que personas del sector público deberán adelantarse a los temas y desafíos que surgirán en el medio y largo plazo como producto de las transformaciones tecnológicas que estamos viendo hoy en día.

› Palabras finales

Las oportunidades y desafíos que presenta el big data deben ser estudiados y analizados con mucho cuidado. Es fácil caer en simplificaciones, y asegurar confiados que esta nueva revolución tecnológica es similar a la modernización ya

emprendimientos en todo el mundo, es importante llevar a cabo este ejercicio a pesar de ser imperfecto. Sí es necesario iterar a futuro para mejorarlo.

22. https://blog.ycombinator.com/basic-income
23. https://hbr.org/1996/01/a-country-is-not-a-company

vista en el sector agrícola y la revolución industrial, transformaciones que liberaron recursos humanos para ser usados en tareas más creativas y de mayor desafío cognitivo. La gran diferencia de esta revolución digital es que esta vez no hay sector que se escape del uso de big data y la automatización de decisiones. No está claro qué sector podrá absorber a futuro toda esa mano de obra disponible. O al menos, no está claro hoy en día[24].

El objetivo de este libro no es entregar respuestas simples, sino abrir una importante conversación en referencia a los grandes desafíos que afrontarán los gobiernos de América Latina debido al surgimiento del big data. Frente a grandes desafíos es difícil mantener una postura neutral, sobre todo cuando estos involucran cambios sociales de gran envergadura. A nuestro entender, las oportunidades y desafíos que abre el big data solo son comparables a los desafíos relacionados al cambio climático. Y ante este desafío es por lo que presentamos en este libro un gran abanico de análisis y propuestas.

24. Algunas referencias para esta discusión se encuentran en:

- Rise of the Robots: Technology and the Threat of a Jobless Future (https://www.amazon.com/dp/0465059996).
- The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies (https://www.amazon.com/dp/0393239357).
- The Zero Marginal Cost Society: The Internet of Things, the Collaborative Commons, and the Eclipse of Capitalism (https://www.amazon.com/dp/B00HY09XGQ).
- Throwing Rocks at the Google Bus: How Growth Became the Enemy of Prosperity (https://www.amazon.com/dp/1617230170).

Capítulo 2

El trinomio dato-información-conocimiento

Eduard Martín Lineros*

Algunas veces contemplando aquel ventanal que mostraba la belleza de la ciudad de Barcelona, y esa magnífica zona destinada a ser el centro neurálgico de la eclosión de las TIC en la capital catalana como punto de referencia para el mundo, uno sentía la sensación de intentar comprender cómo era posible la aparente armonía que se avistaba.

Porque ciertamente nuestras ciudades, más allá del urbanismo aplicado como una ciencia, lo que buscan es la armonía, que no quiere decir ineludiblemente uniformidad, aburrimiento, aplicación de cánones más o menos clásicos... Me refiero a la armonía entendida como el funcionamiento lento pero a ritmo claro y conciso de todo lo que sucede.

En esta etapa nos preocupaba la consecución de una ciudad autosuficiente, hiperconectada, ecologicamente sostenible y de barrios que funcionasen a velocidad humana dentro de una metrópolis a velocidad global (Vicente Guallart, *La ciudad Autosuficiente*, RBA Libros S.A, 2012). Una clara manifestación de la necesidad de la consecución del concepto de sostenibilidad urbana. Sostenibilidad extendida no solo a la eficiencia energética, o la eficiencia en el sistema de transportes y movilidad, sino sostenibilidad en mayúsculas para que el funcionamiento de todos los sistemas verticales de acción urbana alcanzaran la máxima precisión a través de la orquestación. Un equilibrio armónico donde un sistema aprovecha las capacidades del sistema vecino para enriquecer sus potencialidades, de una manera bidireccional y natural.

En definitiva, una ciudad concebida como un universo más o menos ordenado. Su componente básico —su materia prima— es su población. Personas que deciden agruparse para ofrecerse mejores posibilidades de supervivencia. Las personas circunscriben espacios de existencia en un lugar físico: el territorio. La ciudad, por tanto, es la simbiosis de personas en un determinado espacio de un

* Director Estrategia Sector Público Sopra Steria y exdirector de Innovación, Sociedad del Conocimiento y Arquitecturas TIC del Ayuntamiento de Barcelona.

territorio que reunidas establecen estándares de convivencia de los que surgen los sistemas de acción —y después de gestión— urbana.

Figura 1. Los sistemas urbanos, en el eje territorio-población

Población-territorio es el binomio que crea economía, y de esta economía aparecen los sistemas que la dinamizan: la energía —cualquiera que sea— para alimentarla, la movilidad para posibilitar la amplificación de las relaciones, el transporte para vehicular la movilidad, los suministros para construir vidas saludables, los servicios avanzados que permiten asistir a las personas, protegerlas y también, por qué no, perseguirlas.

Y en este contexto, que no es más que el contexto urbano de la realidad misma de la existencia del ser humano social, ¿qué papel ha de jugar la tecnología de los datos?

Según la Real Academia de la Lengua Española, un dato es la "Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho". En una segunda acepción un dato es un "documento, testimonio, fundamento". Según la misma Academia la información es la "acción de informar" e informar es "enterar o dar noticia de algo". Podríamos deducir pues que cuando hablamos de información estamos dando cuenta de los datos que conocemos.

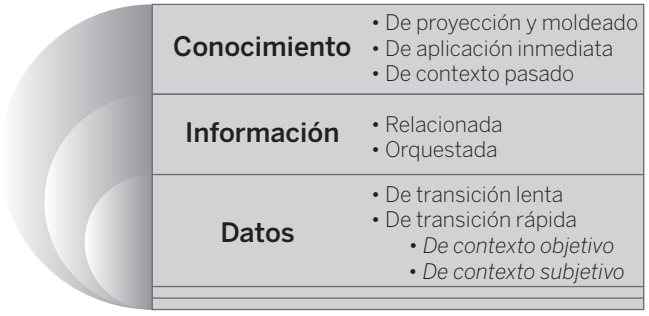
¿Pero, y el conocimiento? Según una de las acepciones que da esta misma Academia, el conocimiento es la "noción, saber o noticia elemental de algo". Parece pues que datos, información y conocimiento no son exactamente lo mismo, aunque todos estos términos hacen referencia a una realidad evidente: la constatación de alguna realidad.

Obviamente la traslación directa de estos conceptos al ámbito de las tecnologías de la información tiene sus propias características. Y resulta evidente que la distinción entre estos tres conceptos es fundamental para el desarrollo de los sistemas computacionales que trabajan, en definitiva, con entradas para obtener resultados (*input / output*).

El trinomio dato-información-conocimiento puede observarse desde diversos puntos de vista, pero en mi opinión constituyen una jerarquía construida desde lo más concreto, pequeño y constatable a lo más abstracto. Podríamos asimilarlo al sistema computacional: mis entradas o *inputs* son datos, el tratamiento los convertirá en información válida, y quizá las salidas o *outputs* serán conocimiento.

¿Cómo se aplica este trinomio a la utilización de las TIC en la gestión de las nuevas ciudades? La importancia del tema es enorme en las ciudades consideradas "inteligentes". Básicamente porque la inteligencia aplicada a un territorio deviene únicamente de sus habitantes: de la gente. Deberíamos hablar con más propiedad de *smart citizens* trabajando con una herramienta como son las TIC, para conseguir territorios más sostenibles, eficientes y habitables. Y este trabajo se realiza en buena forma obteniendo datos, trabajando su contenido para convertirlos en información, y utilizando está información hacia atrás y en proyección, para amasar conocimiento.

Figura 2. El trinomio y su desarrollo



Datos

El paradigma de los datos aparece en nuestros días como el elemento fundamental para la consecución de unos sistemas computacionales que permitan mejorar servicios urbanos. Ahora bien: ¿cuántos datos somos capaces de obtener?, ¿cuál

es la procedencia de los mismos?, ¿qué valor tiene cada dato?, ¿cuáles hemos de guardar, y cuáles hemos de rechazar? Estas y otras preguntas son las que intentaré abordar seguidamente.

Una de las grandes revoluciones que se están y se seguirán produciendo es la capacidad humana a través de la tecnología de obtener datos de una manera masiva. En nuestros días los datos ya no solo provienen de la comunicación hablada o de la simple observación de los fenómenos naturales, ni mucho menos de la observación en general del mundo. En nuestros tiempos podemos obtener datos sobre aspectos que con la simple observación no es posible; en nuestros tiempos podemos obtener datos imperceptibles...

En el tiempo del "Internet of Things", donde "todo" es susceptible de ser conectado, direccionado y accedido a través de la red, la obtención y clasificación de datos es fundamental para la utilización correcta de los *inputs* necesarios para la obtención de los resultados deseados.

Además de su proliferación podríamos estructurar los datos en dos grandes familias (como hacen muchos autores):

a) Los datos "tradicionales" o de transición lenta: que son los datos que tradicionalmente hemos obtenido de la observación directa, y que desde el punto de vista tecnológico hemos convertido en datos "relacionales". Son datos que se pueden poner en relación con otros a través de relaciones simples y que ciertamente transicionan poco en el tiempo. Un ejemplo podría ser el nombre de las personas. Se obtiene fácilmente por la observación, la pregunta directa, o simplemente por la consulta de un registro. El nombre transiciona poco o nada (es más o menos constante en el tiempo) y es fácilmente relacionable con otros datos (otros nombres, los apellidos...). Es un dato "fácil de reconocer" y aporta evidencias sobre algunos temas (los nombres relacionados constituyen las jerarquías familiares...). Podemos hablar de él como un dato básico, tradicional... reconocible y de transición lenta.

b) Los datos "nuevos" o de transición rápida: que son los datos que no podemos obtener fácilmente a través de la observación humana, ya sea por física pura, ya sea por desconocimiento o incapacidad. Muchas veces estos datos mutan rápidamente, son difícilmente relacionables fuera de su propio contexto, y son susceptibles de "amontonarse" en grandes cantidades. Estos datos, además, no se pueden poner fácilmente en relación, y son susceptibles de grandes cambios. Por ejemplo, pensemos en el índice exacto

de polución en una determinada coordenada gps, en un momento determinado del día. La exactitud de este dato no puede ser percibida por la simple observación, se ha de obtener a través del uso de tecnología (TIC o no TIC), a través del uso de alguna herramienta especializada. Este índice es altamente volátil (en un determinado minuto es de una magnitud, en otro minuto de otra). Además podemos tener en muy poco espacio de tiempo diferentes índices de polución tomados en periodos de tiempo muy cortos entre sí.

Por tanto, en nuestro "mundo" datos de diferente índole se encuentran a nuestra disposición de manera amplia y variada. La irrupción de todo tipo de sensores hace que la riqueza de *inputs* que poseemos se convierta en realidad en una grave problemática de gestión de los mismos, y de contextualización de las magnitudes que incorporan.

Utilizando los ejemplos anteriores, pensemos en el axioma: "Juan es alérgico. Ha de tener cuidado con los índices de polución". Fijémonos que, en el fondo, esta afirmación podría desencadenar la necesidad de poner en contacto el nombre "Juan" —con todo lo que significa— (dato de transición lenta), con el "índice de polución" —con todo lo que conlleva— (datos de transición rápida). La contextualización de esta relación puede no ser sencilla, ni simple, al margen de los *outputs* que se puedan obtener.

Así pues "datos rápidos" y "datos lentos". Nuestra actual sociedad con sus avances tecnológicos posibilita que los datos lentos, tradicionales, puedan tener un tratamiento excepcionalmente eficaz. Ya desde hace muchos años los sistemas gestores de base de datos relacionales han tratado este tipo de datos con eficiencia. Nuestras ciudades y poblaciones tienen la "lista" de sus habitantes (los "nombres") mecanizada en estos sistemas y su gestión es eficiente y eficaz. La lista de nombres "relacionada" con la lista de las calles —con el modelo organizativo territorial—, constituye el verdadero padrón de habitantes. Sobre esta base dato poblacional-dato territorial, hemos construido desde antiguo el contexto de las ciudades y poblaciones. De esta relación aparece la clasificación económica del territorio, la ubicación y desarrollo de las actividades, los sistemas de movilidad y de transporte, el suministro eléctrico y de los otros servicios básicos (agua, gas...).

Pero ¿cómo afecta la proliferación de los "datos rápidos"? La capacidad tecnológica de obtener datos de este tipo hace que las relaciones casi perfectas que se establecen entre los datos de transición lenta, se vea modificada. De la misma

manera la relación entre los datos de transición rápida con otros datos de transición rápida, precisa de consideraciones diferentes.

Pensemos en la telemetría de un vehículo de competición. Los datos objetivos de su construcción —las partes que lo componen, las piezas del motor, cada uno de sus engranajes, depósitos y componentes— se deben relacionar con el conjunto de información de contexto que los mil y un sensores —generadores de datos de transición rápida— para conseguir *outputs* deseados (por ejemplo el comportamiento diferente de uno de estos componentes). Además estos *outputs* han de ser inmediatos en el tiempo: no se pueden demorar. Los "actuadores" —elemento contrapuesto y complementario al sensor— aplicarán estos *outputs*. Es obvio que la aplicación práctica de la relación entre los datos de transición lenta, con la gran cantidad de datos de transición rápida, en un vehículo de estas características, puede modificar en unos segundos su comportamiento. La telemetría nos permite observar y gestionar estas relaciones y, por tanto, observar los datos no aislados, sino en el contexto adecuado.

Pero no siempre es tan evidente la relación, ni tan fácil el encuentro del contexto. Apliquemos el ejemplo del vehículo a una ciudad.

La primera dificultad, obvia, será determinar qué datos deben considerarse, aparte de los obvios, como datos de transición lenta, y qué datos deben considerarse como datos de transición rápida. Aceptando el paradigma que todos los datos almacenados en sistemas de gestión relacional son datos de transición lenta, y que todos los datos almacenados en sistemas de gestión distribuida y altamente volátil son datos de transición rápida, obtendremos que el problema de los datos en una ciudad es un problema no solo de relación entre unos y otros, no siempre evidente, sino también de capacidad y volumen.

Pero es realmente necesario obtener las claves en la relación entre unos y otros. Solo la relación —el "cruce"— de unos datos de un tipo con otros de otro tipo, conduce a la contextualización, es decir, conduce a la consideración de obtener información objetiva —basada en datos— sobre un determinado aspecto.

Pensemos en el ejemplo anterior: Juan es alérgico y, por lo tanto, debe vigilar los niveles de polución. Juan vive en esta ciudad, por tanto a priori forma parte de la lista de habitantes, y como reside en una determinada vivienda, de un determinado edificio, de una determinada calle, de un determinado barrio, el cruce de su información personal con la territorial permite identificarlo unívocamente respecto de los otros habitantes. Ocupa una posición en el territorio. La

contextualización de los niveles de polución en esa determinada posición territorial, condicionará la vida de Juan en relación a su problema alérgico. La forma, frecuencia y sistemática en la obtención de los datos de transición rápida, a través de los sensores medioambientales, y su relación con la ubicación territorial de Juan, contextualizará su problema, pero en ningún caso lo habrá resuelto, porque Juan se mueve. La acción de moverse o no modificará la relación, nuevamente, entre datos de transición lenta y datos de transición rápida.

Por lo tanto, el establecimiento de una orquestación adecuada entre los datos tradicionales —desde el punto de vista de su obtención— con los datos cuyo origen proviene de los nuevos sistemas desarrollados a partir de la eclosión de las nuevas tecnologías es básico para la construcción de soluciones basadas en la observación realmente objetiva de la realidad.

La aparición de los datos de contexto —que no quiere decir que no sean tan importantes, como los datos tradicionales— es determinante para el desarrollo de las ciudades y los territorios de gestión avanzada —me resisto a llamarlos inteligentes—. El contexto que aportan estos datos es vital para la adaptación de soluciones, para la generación de una nueva economía o para la gestión ajustada de los diferentes sistemas urbanos. Y es básica para la construcción de lo que consideramos información.

Es más, no me he olvidado de los datos de contexto que aportan las redes sociales. En este sentido, la especial trascendencia de los flujos informacionales que la red nos aporta, hace más complejo el escenario planteado. Sobre todo en el desafío que supone la consideración de datos de velocidad extremadamente rápida, y además de perspectiva subjetiva: los datos aportados por los usuarios de los mecanismos de relación en Internet (redes sociales y también portales, blogs, noticias, foros... etc.).

› Información

Coherentemente con lo expuesto, la información, entendida en los términos académicos, no se hallaría en el dato concreto. La información es en realidad la salida, el "output" deseado por todos. Información basada en la combinación, orquestación y relación razonable de datos sobre realidades (los datos de transición lenta) y los datos de contexto (los datos de transición rápida).

Pero el problema no es realmente la consecución de la información, sino la metodología de cómo conseguir información, en el sentido de fundamento

para la toma de decisiones, válida, objetiva y exenta del examen subjetivo de la realidad.

Uno de los grandes paradigmas es la definición de las reglas de relación, coordinación y orquestación de los datos concretos (lentos y rápidos). La correcta adaptación de estas reglas de relación determinará la objetividad de la información obtenida. Información que estará compuesta de datos tradicionales más su contexto.

Este es uno de los grandes problemas en la definición, diseño y construcción de soluciones realmente "inteligentes" —en el sentido de estar basadas en la objetividad del dato concreto, sea cual sea su naturaleza— respecto de las soluciones basadas en criterios de oportunidad, por interés o, en definitiva, criterios políticos. La definición correcta de las reglas de juego que permitan relacionar todos los datos concretos de una manera eficaz para la mejora de los servicios públicos sin estar sometido el resultado final a ningún interés. Todo un reto.

Imaginemos un caso concreto en el que jugaremos con los dos tipos de datos expuestos (lentos y rápidos) pero también con la especifidad de los datos obtenidos de las fuentes "subjetivas": los datos de contexto rápidos procedentes de Internet.

A una determinada hora del día se produce una recepción de datos a través de los sensores de detección de gases de un edificio público, que indica niveles anómalos de oxígeno. Paralelamente a través de la redes sociales se extiende la alarma de una fuga de gas en la zona del edificio, a la misma hora.

Evidentemente estos datos tomados singularmente podrían dar lugar a *outputs* diferentes, según la interpretación de contexto que cada observador realizara respecto de los mismos. Seguramente el tener a mano los datos sobre el número de personas que se hallan a esa hora en el edificio, el número de habitantes de la zona donde se halla, datos sobre la circulación de vehículos en los alrededores, y datos sobre las condiciones climatológicas, podrían ayudar a realizar un análisis objetivo inicial tendente no solo a resolver la emergencia ordenadamente, sino también a avanzar en el análisis de las causas y la toma de decisiones respecto de las condiciones de todos los sistemas de gestión urbana afectados.

Ahora bien: ¿cómo podemos conseguir una correcta orquestación de toda esta información para que el conjunto de *outputs* resultantes sea coherente? He aquí uno de los grandes retos de los sistemas de información llamados "inteligentes".

No basta con establecer relaciones, no basta con establecer normas de orquestación. Las relaciones y las normas de orquestación deben adaptarse a la idiosincrasia y la cultura del territorio.

Sin duda los sistemas de información como ingenios de la ciencia computacional pueden ser los mismos para todos los territorios, ciudades, etc., pero lo que no puede copiarse o trasladarse de forma simple, son las reglas de relación y orquestación, que deben de contener una información de contexto más: la cultura, la tradición, las reglas propias de convivencia de aquella comunidad.

En un futuro muy cercano, o en un presente, los sistemas de información de gestión urbana inteligente, que están orientados al procesamiento de datos para obtener información veraz y objetiva para la resolución de problemas o simplemente para el moldeado, planificación o estructuración de soluciones, serán sistemas ordinarios, incrustados en los sistemas de gestión urbana tradicionales (lista de habitantes y lista de calles, plazas, avenidas..., que dan como resultado el sistema tributario e impositivo). Estos sistemas además serán exportables de ciudad a ciudad, de territorio a territorio, y esto solo será posible si los sistemas son independientes de las reglas de "negocio" (las relaciones y orquestaciones basadas en la información de contexto local) que los hagan funcionar.

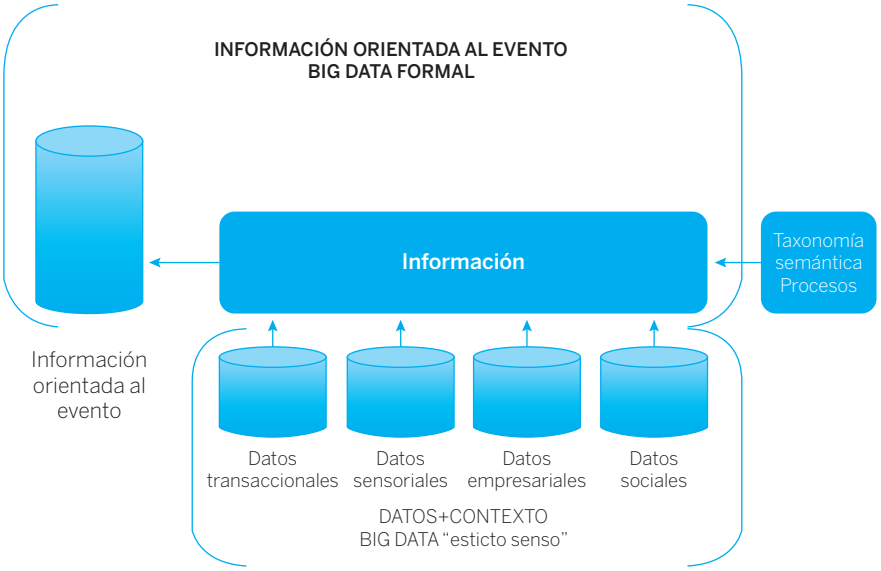
En la definición de estas reglas de "negocio" se halla la verdadera esencia de la información necesaria para el desarrollo de soluciones donde la tecnología se encarga de automatizar las entradas (datos) y las salidas (soluciones a los problemas reales).

Por tanto, nuestra ecuación se complica un poco más: datos lentos, datos rápidos de contexto objetivo procedentes de los nuevos orígenes de datos (básicamente sensores), datos rápidos de contexto subjetivo procedentes de la red o las redes, y ahora, el contexto económico, social y cultural del territorio donde nos hallamos.

La consecución de información a partir de los datos o la "realización" completa de los mismos en contenidos completos de información es algo que no solo lo visualizamos para una mejor gestión de los servicios urbanos aplicados a ciudades. En el conjunto de las Administraciones públicas —ya sean de índole estatal, local o de cualquier otra organización gubernamental— la disposición orquestada de datos capaces de representar realidades concretas pero matizadas y ajustadas por su contexto es una necesidad real en un mundo tremendamente complejo.

La realidad del que conocemos como *big data* es la capacidad de manejar grandes cantidades de datos de diferente índole como hemos visto, pero también la potencia de poder ponerlos en conexión —independientemente de su naturaleza física— para que nos aporten "información" de valor añadido. Evidentemente que los procesos *smart* —simplemente *smart* porque son los propios a los avances tecnológicos de nuestro tiempo— son más fáciles de entender asociados a la gestión del *big data*, pero no por ello los procesos tradicionales de las Administraciones públicas no podrían ser mucho más eficientes y eficaces si el ejercicio de contextualizar los datos concretos para obtener información de valor añadido se pusiese en práctica de forma generalizada.

Figura 3. Datos que orquestados constituyen información en su contexto



En este sentido durante estos últimos años ha aparecido el concepto de la "digitalización de las Administraciones públicas", como contrasentido al ya muy desgastado término de *smart city* o *smart government*. En el fondo de la cuestión subyace la misma necesidad y los mismos objetivos: obtener sistemas de información suficientemente orquestados (antes se imploraba al término "interoperabilidad") y relacionados para tener información sobre los asuntos de la cuestión pública, no solo objetiva sino convenientemente contextualizada. El término "digitalización" hace referencia al medio para conseguir este objetivo, que obviamente pasa, en nuestros días por la utilización de las TIC. Digitalización, *smart*... en el fondo de la cuestión se halla la gestión racional de los datos

para que se conviertan en información realmente útil y debidamente contextualizada.

La definición de las reglas analíticas que han de servir para la orquestación, digital, y, por supuesto, inteligente (*smart*), de los datos con el objetivo de consolidar una *smart city*, o con el objetivo de "digitalizar" una Administración pública, son la piedra angular de todo proceso real de modernización de la Administración pública. Ahora bien, el descubrimiento de estas reglas, que recordemos son las únicamente diferenciales entre los distintos territorios, ciudades... o estados, pasa primero por el reconocimiento de un contexto y la separación definitiva de la conceptualización de los datos, de los procesos que los gestionan.

En la definición de estas reglas (recordemos que son reglas simples que después darán lugar a reglas complejas por la relación de las mismas), deberemos abordar el problema de la clasificación de todos los elementos unitarios que las compondrán, y el significado de cada uno de ellos: **la taxonomía y la semántica.**

No es objetivo de este capítulo tratar temas tan complejos y extensos como determinar la manera o la forma en la que se ha de construir la taxonomía de una ciudad o un territorio. Este trabajo es un trabajo arduo, que todavía ha de ser completado por los propios territorios, y en el cual la aparición de estándares de "ciudad" o "territorio" se me antoja un elemento imprescindible. Tampoco podemos abordar cómo deben ser interpretados estos elementos, construyendo una verdadera semántica, para una comprensión universal de estas reglas, pero una aproximación seria a la consecución de sistemas de información completos deberá abordar, sin duda alguna, la definición de esta semántica en el contexto del *big data* —entendido en este caso como grandes cantidades de datos—.

› Conocimiento

Si pudiésemos definir universalmente el paper de las Administraciones públicas, independientemente de los diferentes contextos políticos y económicos que, desafortunadamente, inciden en demasía en el funcionamiento de estas Administraciones, podríamos asegurar que uno de sus mayores objetivos es la creación y mejora del bien público (algunos hablan de lo "común"). Bien público, riqueza pública para un desarrollo en condiciones de igualdad, de la calidad de vida de los ciudadanos en sentido amplio.

Las Administraciones públicas ejercen su actividad recaudatoria y de control de acuerdo a determinadas estrategias políticas, pero siempre deberían hacerlo para favorecer la mejor prestación de los servicios públicos, la redistribución de la riqueza y el tratamiento en condiciones de igualdad de los ciudadanos.

Sin duda alguna, una información basada en la toma de datos de múltiples orígenes, de diferente naturaleza, con un prisma objetivo, contextualizados y puestos en relación puede dotar a las Administraciones de la mejor información en cada momento. Información sujeta a la taxonomía propia de cada territorio y a la semántica que hace comprensible esta información. Proyectada hacia el "pasado" esta información se convierte en un valor incalculable para comprender la evolución del bien público, la comprensión de errores o la aplicación de estrategias erróneas. La proyección hacia el futuro de esta misma información podrá prevenir errores detectados, podrá planificar introduciendo mecanismos de cooperación, cocreación y coinnovación ciudadana, cómo el territorio, cómo las ciudades, pueden mejorar esta prestación de servicios en condiciones de igualdad.

El conocimiento, por lo tanto, lo interpretaremos como "el saber", la noción cierta, en este caso sustentada en información objetiva debidamente contextualizada, para poder mejorar la calidad de vida de las personas que habitan un cierto territorio.

En este sentido, obviamente, el conocimiento ha existido, existe y existirá, de acuerdo al funcionamiento tradicional de los gobiernos y las Administraciones públicas. La introducción de las tecnologías de la información y la comunicación como elemento diferenciador de nuestra época, lo que aporta, es la mayor capacidad de captación de datos, los mejores sistemas para su relación y orquestación, y la posibilidad de aplicar reglas de comportamiento sobre ordenaciones vastas de gran cantidad de información. Con la ventaja de tener reglas ciertas de interpretación (semántica).

Por lo tanto, la aplicación de las tecnologías digitales para la conversión de información y conocimiento es obvia, como es obvio que el uso de esas tecnologías y avances debe estar orientado a la mejor prestación de servicios públicos y al incremento de la calidad de vida de los ciudadanos. La generación de valor público a través de la introducción de la innovación —en este caso, la eclosión del tratamiento de los datos para conseguir información de valor añadido— es una realidad que cada vez más se hará tangible en las Administraciones públicas.

La capacidad de modelación de los territorios, de capacidad de análisis previo para la mejora de los procesos administrativos permite y permitirá en un futuro

tener Administraciones públicas más sostenibles económicamente, más transparentes de cara al ciudadano —que podrán tener acceso a información y su contexto—, más colaborativas respecto de los trabajadores públicos, y sobre todo una Administración más eficiente y más efectiva.

Las claves de la transformación

Hoy en día el paradigma, como ya he apuntado, no es hablar de *smart cities* o de *smart territorios*. Ciertamente el término *smart* ha perdido su fuerza de atracción, y muchas veces ha sido rechazado por determinadas connotaciones políticas. La irrupción del concepto del "Internet de las cosas o de todo" (IoT o IoE) no es más que una parte de la conceptualización de la realidad que se quería expresar con el concepto *smart*. Hablar de transformación digital también nos parece un poco obvio: las tecnologías de principios del siglo XXI se basan en el concepto de almacenamiento digital, con todas las connotaciones que lleva.

Ciertamente esta revolución de las TIC en su conjunto es la eclosión lógica del avance de las ciencias de la computación, las telecomunicaciones y en general de la aplicación de las matemáticas y la física al mundo real. Ingeniería en estado puro. Por lo tanto, no deja de ser un proceso lógico, razonable y consecuente con los avances de la humanidad. El acto de comunicación, magnificado, mejorado y globalizado por la tecnología es una oportunidad que todos debemos de aprovechar. En cierto modo son lógicos los miedos, las preocupaciones, la necesidad de "seguridad" que nos aborda. La tecnología por sí misma no suele ser perjudicial; el uso que se hace de la misma sí que puede ser el aspecto a tener en cuenta.

Las claves para el pleno aprovechamiento de los innumerables *inputs* que tendremos disponibles pasará por un uso responsable de la tecnología, ya al alcance, para hacer que se conviertan en *outputs* positivos. En este sentido, y a mi modo de ver, las claves para una transformación de toda esta potencia tecnológica, en valor positivo para las Administraciones públicas, han de pasar por los siguientes puntos:

a) Un uso responsable, consciente y honesto de las capacidades de sensorización y actuación automáticas. Hemos de sensorizar aquello que sea necesario para un mejor funcionamiento de los territorios, con el fin de reequilibrar sus capacidades, mejorar servicios y contribuir a una mejor redistribución de los recursos.

b) Un uso transparente de los datos singulares —en todas sus variantes— para conseguir construcciones de relación y orquestación beneficiosas para el conjunto de la población.

c) Una construcción de "reglas" de funcionamiento basada en la cooperación, cocreación, coinnovación y corresponsabilidad con los actores (emisores y receptores) de las mismas.

d) Un uso del conocimiento resultante en pro del bien general (algunos lo llaman común), para un mejor desarrollo de los territorios y las comunidades que los habitan.

Un gran reto.

Capítulo 3

¿Datos, información o conocimiento?

Óscar Corcho*

Cuando hablamos de *big data* nos referimos normalmente a las características principales de los datos con los que tenemos que trabajar, con un foco especial en las V's (volumen, velocidad, variedad, veracidad). Sin embargo, en muchas ocasiones nos olvidamos de prestar atención a la esencia de la materia prima con la que tenemos que trabajar, los datos. En este capítulo vamos a reflexionar sobre qué entendemos por datos, qué tipos de datos son los más adecuados para trabajar de manera efectiva en el mundo del *big data* y cómo podemos dotar a los datos de significado para hacer más fácil su utilización.

Como el lector habrá podido observar en el capítulo anterior, si comenzamos analizando lo que entendemos por dato, podríamos partir de su acepción etimológica en latín, en la que *datum* se refiere a "lo que se da". Pero claro está, utilizar una palabra procedente del latín para referirse a un concepto tan recientemente acuñado como el de *big data* puede parecer un poco inapropiado. Continuemos pues nuestro análisis con la definición dada en el capítulo anterior y que podemos encontrar en el Diccionario de la Real Academia Española[1], donde se define como (1) "Información sobre algo concreto que permite su conocimiento exacto o sirve para deducir las consecuencias derivadas de un hecho", (2) "Documento, testimonio, fundamento" y (3) "Información dispuesta de manera adecuada para su tratamiento por una computadora". Claramente, la tercera de las acepciones parece la más adecuada en el contexto en el que hablamos de *big data*, por el enlace que se hace al hecho de que los datos van a ser tratados mediante computadoras. Sin embargo, es interesante también tener en cuenta de estas definiciones el hecho de que se incluye entre las acepciones el hecho de que los datos pueden ser documentos y testimonios, y algo que resulta también bastante interesante si queremos ser precisos en la definición de lo que consideramos como un dato, que es el hecho de que se utiliza en dos de las acepciones la palabra "información". En la primera de las secciones de este capítulo vamos a discutir con más detalle sobre todas las palabras que se han utilizado habitualmente para referirse a estos conceptos.

* Catedrático en la Universidad Politécnica de Madrid y cofundador de Localidata.

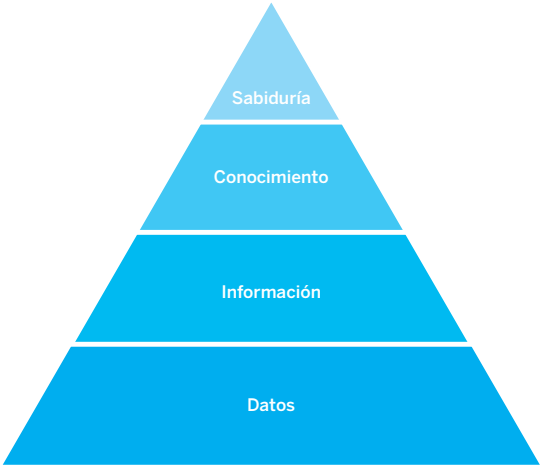
1. http://dle.rae.es/

La pirámide del conocimiento

En el capítulo anterior se introducían tres términos que de manera informal se suelen tratar de manera indistinta, pero que de una manera más formal son claramente distinguibles: datos, información y conocimientos.

La figura 1 representa la que se conoce como "la jerarquía o pirámide del conocimiento". Existen múltiples referencias bibliográficas en las que se tratan estos conceptos y se comienza a hablar de las relaciones entre ellos, fundamentalmente en la literatura sobre gestión de conocimientos (por ejemplo, Zeleny, 1987 y Ackoff, 1989). También se pueden encontrar extensas revisiones sobre las definiciones de estos conceptos y sus interrelaciones, como por ejemplo en Rowley, 2007 o Zins, 2007.

Figura 1. La pirámide del conocimiento (basado en Ackoff, 1989)



En nuestro caso, y a diferencia del capítulo anterior, nos vamos a centrar en proporcionar algunas definiciones informales sobre todos estos conceptos, desde una perspectiva práctica, a través de ejemplos, y con el foco en su uso en el contexto *big data*.

Datos

De manera general, en *big data* nos referimos al uso de datos de todo tipo: datos que proceden de bases de datos de organizaciones públicas y privadas, y que en ocasiones pueden estar disponibles como datos abiertos, datos que se encuentran

disponibles de distintas maneras en la web, descargables en formatos estructurados o entrelazados con textos en páginas web, datos que podemos captar a partir de sensores de todo tipo, etc. Vamos a ver un ejemplo concreto, procedente del portal de datos abiertos del Ayuntamiento de Zaragoza (http://datos.zaragoza.es/ o http://www.zaragoza.es/ciudad/risp/). Concretamente, nos vamos a centrar en el conjunto de datos de calidad del aire[2], que proporciona datos sobre la calidad del aire que se mide periódicamente en distintas estaciones de medición situadas en la ciudad. Al navegar por las páginas que describen este conjunto de datos se puede llegar a distintos enlaces que nos proporcionan acceso a los datos específicos de cada estación, a distintas horas del día y en distintos días del año. Por ejemplo, en la figura 2 podemos ver la visualización de los datos correspondientes a la estación de El Picarral para el día 16 de febrero de 2015.

Figura 2. Datos correspondientes a las mediciones de calidad del aire en la estación de El Picarral, el día 16/02/2015

Fuente: Ayuntamiento de Zaragoza.

2. http://www.zaragoza.es/ciudad/risp/detalle_Risp?id=131

Asimismo, estos datos están disponibles para su descarga en formatos XML, CSV o JSON, entre otros, de tal manera que puedan ser utilizados fácilmente por cualquier persona u organización. La figura 3 ofrece un extracto de estos datos en formato JSON, con campos específicos que representan la fecha de las mediciones, el nombre de la estación, los valores de los distintos componentes medidos por la estación, etc. Estos podrían considerarse como los elementos primarios con los que vamos a trabajar normalmente. De hecho, en ocasiones se habla de la distinción entre *raw data* (o datos en crudo) y datos que han sido ya procesados de alguna manera, refiriéndose al hecho de que han podido ser agregados, limpiados, etc. Estos datos podrían considerarse como datos en crudo, mientras que los datos correspondientes a la media de las mediciones diarias durante todo un mes podrían ser considerados como datos agregados.

Figura 3. Extracto del código JSON correspondiente a los datos de la figura 2


```
- {
    id: "id:medicionaire-1_16-02-2015",
    uri: "http://www.zaragoza.es/ciudad/medioambiente/atmosfera/redconta/detalle_RedConta?id=1&fecha=16/02/2015",
    title: "Datos para la estación El Picarral del 16-02-2015",
    language: "es",
    texto_t: "",
    x_coordinate: 677292.84,
    y_coordinate: 4615441.97,
    coordenadas_p_0_coordinate: 41.66918284323936,
    coordenadas_p_1_coordinate: -0.8715528012249274,
    coordenadas_p: "41.66918284323936,-0.8715528012249274",
    pm10_d: 5,
    no2_d: 21,
    co_d: 0.2,
    o3_d: 50,
    sh2_d: 2,
    fecha_dt: "2015-02-15T23:00:00Z",
    tipocontenido_s: "historico",
    last_modified: "2015-02-17T09:52:42Z",
  - text: [
        "Datos para la estación El Picarral del 16-02-2015",
        "El Picarral"
    ],
  - content_type: [
        "HTML"
    ],
  - category: [
        "Calidad del aire"
    ],
  - estacion_smultiple: [
        "El Picarral"
    ]
}
```


› Información

Los datos anteriores no tienen mucho valor si no se les puede otorgar un significado concreto. De nada sirve tener atributos y sus valores si no se comprende qué es lo que significa cada uno de ellos y cómo se puede interpretar. La utilización de nombres de atributos como *x_coordinate* e *y_coordinate* no nos aportaría nada si

no supiéramos que existe una forma de representar el punto en el que se están haciendo las mediciones (donde se encuentra la estación de medida) que se puede representar mediante coordenadas x e y de una proyección espacial concreta, y que estos puntos son aproximadamente los mismos (con algún pequeño margen de error) que los que se representan con los atributos *coordenadas_p_0_coordinate* y *coordenadas_p_1_coordinate,* que representan la latitud y longitud de dicho punto.

Algo similar ocurre con los valores de *pm10_d, no2_d, co_d, o3_d, sh2_d,* que se corresponden con las concentraciones en el aire de los distintos tipos de compuestos que se están midiendo, y que tienen sus propias unidades de medida, las cuales ni siquiera se están expresando en estos datos en crudo. De hecho, será habitual para un periodista que quiera hacer una noticia sobre la calidad del aire en Zaragoza, o para un científico que quiera reutilizar estos datos, o para un técnico que tenga que tomar una decisión sobre si se cierra al tráfico una parte de las vías de Zaragoza entender qué significa cada uno de estos atributos en su contexto.

En esta situación es cuando podemos comenzar a hablar de información. Es decir, la información se puede definir como la combinación de los datos y su contexto, lo que permite que estos datos puedan ser interpretados. El contexto puede variar desde la simple descripción textual de lo que significan los encabezados de las columnas de un CSV, o los nombres de los atributos de un fichero JSON, hasta la descripción más detallada de la metodología que se ha seguido para la obtención del conjunto de datos, la forma de identificar posibles errores en los datos, la descripción de valores frecuentes y valores anómalos, las unidades de medida de cada uno de los valores, cuando estas tienen sentido, etc. También se puede considerar información cualquier tipo de información adicional que se puede extraer a partir de los datos (por ejemplo, una noticia en prensa donde se habla sobre dichos datos y se discuten diferentes perspectivas sobre la calidad del aire en la ciudad).

Un aspecto a destacar en esta diferenciación entre datos e información es que en ocasiones lo que para una persona puede ser información para otra pueden ser datos, y viceversa. Un ejemplo se puede encontrar precisamente en el caso de la información textual, como la comentada anteriormente sobre la noticia en prensa. Mientras que hemos considerado en esta descripción que este sería un caso claro de información, un investigador o profesional trabajando en realizar minería de textos para comprender, por ejemplo, cómo se suelen escribir este tipo de noticias consideraría estos textos como sus datos primarios. Es decir, la frontera entre datos e información, y la decisión sobre la cantidad de contexto necesaria para convertir datos en información, no es una frontera estricta.

Conocimientos

El siguiente escalón en la pirámide de nuestra figura 1 es el del conocimiento, aunque una buena parte de la literatura, y yo mismo, preferimos utilizar el término "conocimientos", en plural. Por conocimientos entendemos la mezcla de la información y el *know-how* que nos permite entender mejor los datos y la información que manejamos, abstraer los patrones que aparecen de manera regular en dichos datos e información, comparar datos, buscar conexiones y predecir nuevos datos, y finalmente, de manera más general, aplicarlos en nuestra organización.

En el caso de nuestro ejemplo, podríamos combinar nuestros datos de calidad del aire y datos adicionales de meteorología para conseguir derivar patrones habituales. Por ejemplo, para saber que durante los días de lluvia la concentración de dióxido de nitrógeno es mucho más pequeña, o que la generación de ozono se produce a unas determinadas horas del día y con unas características de radiación solar concretas. Los datos también pueden permitir derivar nuevos conocimientos (por ejemplo, mediante la aplicación de técnicas de minería de datos), validar conocimientos existentes (mediante el contraste de los datos esperados con los datos reales observados) o refutar hipótesis (demostrando que los datos no están en consonancia con lo esperado según el conocimiento existente).

Finalmente, en la parte superior de la figura 1 se habla también del concepto de sabiduría (en otros casos también se utiliza el término inteligencia), que nos permite referirnos a la recolección de una cantidad de conocimientos sobre un dominio particular que sea suficiente para permitirnos tomar siempre la decisión adecuada en todo momento. Normalmente este concepto es difícil de articular de una manera clara y representar en una base de conocimientos o memoria institucional, dada la complejidad de todos los concimientos que se utilizan dentro de una organización.

¿Qué es lo que más necesitamos? ¿Datos, información o conocimientos?

Cuando hablamos de *big data* parece que el foco está siempre puesto en la disponibilidad de los datos, a partir de los cuales se pueden hacer todos los análisis necesarios para derivar información y conocimientos útiles para nuestros propósitos. En su charla TED de 2009[3], Sir Tim Berners Lee hacía mención a la importancia de que todas las organizaciones (con especial énfasis en las públicas) hicieran sus datos crudos disponibles en la web, bajo el lema "*Raw Data Now*". En estos

3. https://www.ted.com/talks/tim_berners_lee_on_the_next_web

momentos, ya se había propuesto el concepto de *web linked data* (también denominada web de datos enlazados o vinculados) como una de las alternativas más adecuadas para la representación de datos en la web. En todo el mundo ha surgido una gran cantidad de iniciativas relacionadas con los datos abiertos[4], fundamentalmente lideradas por Administraciones públicas que han legislado (o se han visto forzadas por la legislación existente en sus entornos) a publicar los datos de los que disponen mediante licencias abiertas, para facilitar la reutilización de los datos, la rendición de cuentas y la generación de valor (económico y social).

También han sido muchas las organizaciones del ámbito privado que han comenzado a hacer disponibles sus datos (de manera pública o mediante algún tipo de pago) con el objetivo de que haya terceros (individuos u organizaciones) que puedan generar valor añadido con dichos datos. Ejemplos incluyen buscadores como Google a través de las Google APIs, redes sociales como Facebook o Foursquare, plataformas de microblogs como Twitter, etc. En casi todos estos casos, estas organizaciones han hecho disponibles APIs de acceso a los datos para que puedan ser utilizadas, bajo unos términos y condiciones que dependen de cada caso y del tipo de utilización que se quiera realizar.

Aunque se podría caer en la tentación de pensar que los datos son suficientes (sin ellos se podría hacer poco, y el valor de las técnicas de *big data* se ve claramente aumentado cuando la cantidad de datos disponibles empieza a ser muy grande y variada), no se debe olvidar que también es importante el contexto que rodea la publicación de estos datos (es decir, lo que los convierte en información y facilita su entendimiento y posterior procesamiento reduciendo errores) y los conocimientos que se puedan haber derivado en cualquier organización y que puedan hacerse disponibles para que otros los utilicen (por ejemplo, en nuestro caso, las reglas que definen cómo se comporta la concentración de los distintos compuestos del aire ante distintas situaciones meteorológicas). Por tanto, todos los elementos recogidos en la pirámide son igualmente relevantes y será importante disponer de todos ellos para poder generar valor añadido en nuestro posterior procesamiento con técnicas de *big data*.

› De datos no estructurados a información útil

Ya hemos discutido en la sección anterior sobre el coste de oportunidad de disponer de una buena cantidad de fuentes de datos, para poder así realizar nuestras tareas de generación o validación de conocimientos, de acuerdo con la

4. https://okfn.org/opendata/

estructura de la pirámide. Sin embargo, no hemos hablado aún de la importancia que también tiene que los datos estén disponibles en formatos que sean fáciles de procesar.

Atendiendo a la tercera de las acepciones recogidas en la entrada correspondiente al término "dato" en el Diccionario de la Real Academia Española, los datos deberían ser procesables por computadoras. En principio, esto quiere decir que para que podamos hablar de datos normalmente consideraremos que estos datos están ya recogidos y almacenados en algún formato electrónico, y disponibles a través de algún medio telemático, preferiblemente por Internet (en la web, en alguna base de datos accesible, a través de un *web socket*, etc.). De hecho, la definición de datos abiertos, por ejemplo, hace hincapié en la necesidad de que los datos estén disponibles preferiblemente en medios como la web.

Sin embargo, existen multitud de formatos en los que los datos pueden hacerse disponibles, lo que puede añadir bastante complejidad a su tratamiento. Atendiendo al ejemplo que hemos utilizado en este capítulo, los datos se hacían disponibles de manera gráfica, para su consumo visual fundamentalmente, así como en formatos XML, CSV y JSON. Este se podría considerar como un ejemplo de buenas prácticas para la publicación de datos, dado que permite su consumo de manera relativamente sencilla por parte de desarrolladores, científicos de datos y reutilizadores en general.

Otras posibles opciones con mayor dificultad para su reutilización podrían haber sido la publicación de estos datos en un fichero PDF, insertados dentro de un documento, en algún formato propietario para la representación de datos de sensores, o su puesta a disposición en la web intercalados con otros textos en HTML, en forma de tablas HTML. En todos estos casos se hacen necesarias etapas de preprocesado previo de los datos para intentar conseguirlos en un formato más fácil de tratar, con herramientas que pueden dar lugar a errores en la extracción de los datos y, por tanto, a errores posteriores en su procesamiento.

Por estas razones, siempre se recomienda que los datos estén disponibles en formatos de fácil tratamiento, puesto que las labores de preprocesamiento serán generalmente menos costosas, así como bien documentados, para que se puedan comprender bien los procesos seguidos para su adquisición, generación, mantenimiento, curación, etc., así como el significado y las posibles interpretaciones de cada uno de ellos. Por ejemplo, en el caso del conjunto de datos de calidad del aire de Zaragoza la ficha de datos[5] proporciona información sobre todos estos aspectos.

5. http://www.zaragoza.es/ciudad/risp/detalle_Risp?id=131

Ontologías

El último de los aspectos a tener en cuenta a la hora de ingestar y tratar datos procedentes de cualquier fuente, y sobre todo en el caso de utilizar fuentes heterogéneas, es el hecho de que si se comparte una estructura común entre todos los que publican datos similares, y se comparten las definiciones de los campos y atributos utilizados, se facilitará aún más el tratamiento posterior de los datos, así como que se evitarán en general los errores que se pueden cometer al realizar interpretaciones posiblemente incorrectas sobre los mismos.

Anteriormente se ha comentado que en el caso de los datos de calidad del aire de Zaragoza, las coordenadas en forma de latitud y longitud se exponían utilizando unos atributos denominados *coordenadas_p_0_coordinate* y *coordenadas_p_1_coordinate*. Asimismo, los datos correspondientes a las concentraciones de distintos compuestos se hacían disponibles mediante un atributo con el nombre corto del compuesto seguido de *_d,* y sin indicación de la unidad de medida. Sin embargo, tanto las coordenadas geográficas como los compuestos del aire ya están descritos de manera formal en otros lugares en la web, con su descripción correspondiente (tanto formal como en lenguaje natural), una indicación de las unidades de medida utilizadas por defecto, etc.

Estas descripciones a las que se hace referencia en el párrafo anterior están disponibles en forma de ontologías (entendidas como "representaciones formales y explícitas de conceptualizaciones compartidas" [Studer *et al.*, 1998]). Por ejemplo, la representación de la latitud y la longitud de un punto geográfico puede hacerse utilizando la ontología Geo[6] del W3C. Y la representación de los datos correspondientes a la calidad del aire pueden hacerse utilizando la ontología de calidad del aire[7] recomendada en la norma técnica UNE178301:2015[8] sobre datos abiertos para ciudades inteligentes, que a su vez reutiliza la ontología Semantic Sensor Network del W3C [Compton *et al.*, 2012].

Este es el caso de los datos que se exponen en Zaragoza, que también se hacen disponibles en su endpoint SPARQL de acuerdo con las ontologías anteriormente mencionadas, tal y como se describe en la ficha de descripción del conjunto de datos, cuya URL ha sido indicada anteriormente, en la pestaña SPARQL. De hecho, los datos visualizados en la figura 2 han sido obtenidos a partir de este endpoint SPARQL. De este modo, si otras ciudades deciden comenzar a publicar sus datos de calidad del aire en formatos semánticos, pueden reutilizar estos vocabularios y exponer los datos de manera similar, lo que facilitará el trabajo a los reutilizadores.

6. https://www.w3.org/2003/01/geo/
7. http://vocab.linkeddata.es/datosabiertos/def/medio-ambiente/calidad-aire
8. http://www.aenor.es/aenor/normas/normas/fichanorma.asp?tipo=N&codigo=N0054318

Conclusiones

En este capítulo hemos analizado algunos de los fundamentos básicos que debemos tener en cuenta cuando hablamos de *big data*, como es la diferenciación entre datos, información y conocimientos. Asimismo, también hemos discutido sobre la importancia de que los datos estén disponibles en formatos que sean lo más amigables posible para su posterior tratamiento por los que van a hacer uso de ellos (desarrolladores, científicos de datos, integradores, etc.), pudiendo llegar a hacer uso de ontologías, entendidas como vocabularios formalmente descritos que proporcionan estructuras de datos consensuadas y bien definidas en una comunidad de práctica.

Los datos son la materia prima fundamental de todos los trabajos que podemos hacer en el contexto de *big data* y, por tanto, una buena calidad de los mismos será una condición indispensable para poder conseguir una buena calidad en los resultados de la aplicación de nuestras técnicas de *big data*. Esto lo debemos tener en cuenta para las descripciones que se harán en los próximos capítulos.

Referencias bibliográficas

Ackoff, R. L. (1989). "From Data to Wisdom". *Journal of Applied Systems Analysis,* 16: 3-9.

Compton, M., Barnaghi, P., Bermúdez, L., García-Castro, R., Corcho, O., Cox, S., Graybeal, J., Hauswirth, M., Henson, C., Herzog, A., Huang, V., Janowicz, K., Kelsey, W. D., Le Phuoc, D., Lefort, L., Leggieri, M., Neuhaus, H., Nikolov, A., Page, K., Passant, A., Sheth, A., Taylor, K. (2012). "The SSN ontology of the W3C semantic sensor network incubator group". *Web Semantics: Science, Services and Agents on the World Wide Web,* 17: 25-32.

Rowley, J. (2007). "The wisdom hierarchy: representations of the DIKW hierarchy". *Journal of Information Science,* 33(2): 163-180.

Studer, R., Benjamins, V. R., Fensel, D. (1998). "Knowledge engineering: Principles and methods". *Data & Knowledge Engineering,* 25(1-2): 161-198.

Zeleny, M. (1987). "Management Support Systems: Towards Integrated Knowledge Management". *Human Systems Management,* 7(1): 59-70.

Zins, C. (2007). "Conceptual Approaches for Defining Data, Information, and Knowledge". *Journal of the American Society for Information Science and Technology,* 58(4): 479-493.

Capítulo 4

El reto *big data* para la estadística pública

Alberto González Yanes*

El viejo-nuevo problema *big data* en la estadística pública

La sociedad de finales de siglo XX y principios del siglo XXI está cambiando rápidamente en muchos aspectos, entre ellos, los vinculados al mundo de la información. Vivimos en la época **SMAC** (Social, Mobile, Analytics, Cloud) donde las personas, muchas de ellas denominadas nativas digitales, no conciben su vida sin un dispositivo móvil a través del que se relacionan con el mundo. Este estilo de vida, al que ya se llama digital, genera un tsunami de cambios y una verdadera montaña de datos en flujo constante.

A esto se suma lo que Kevin Ashton denominó **"Internet de las cosas"** (IoT, por sus siglas en inglés), concepto que se refiere a la interconexión digital de objetos cotidianos con Internet. La idea subyacente es que los objetos se equipan con sensores, que generan datos que se comunican por Internet. La "Internet de las cosas" tiene un fuerte demandante de equipos conectados en las ciudades inteligentes, en las que los sistemas de iluminación, la señalización y otros servicios públicos automatizados representarán millones de objetos conectados a Internet.

El nacimiento de estos nuevos fenómenos es producto del advenimiento de las computadoras, que trajo consigo equipos de medida y almacenaje que hicieron sumamente más eficiente el proceso de datificación. La incorporación de ordenadores a las empresas y a las Administraciones públicas extendió el almacenamiento y tratamiento de datos durante los años ochenta y noventa del siglo pasado, dando lugar a la inteligencia de negocios aplicada tanto a la empresa como al sector público ***(business intelligence)***, y dando lugar también a implicaciones en la estadística pública con el surgimiento de la estadística basada en registros administrativos[1]. Este avance también se facilitó gracias al tratamiento y análisis

* Matemático y jefe de Estadísticas Económicas del Instituto Canario de Estadística (ISTAC) del Gobierno de Canarias, España.

1. Wallgren, Anders y Wallgren, Britt. *Register-Based Statistics: Administrative Data for Statistical Purposes. Wiley Series in Survey Methodology.* Chichester, England; Hoboken, NJ: John Wiley & Sons Ltd, 2007.

matemático de datos, permitiendo descubrir su valor oculto y dando lugar a términos comerciales como ***data mining*** que describe el uso de la estadística y de métodos matemáticos en el análisis de los datos empresariales.

Por lo tanto las empresas vienen desarrollando desde hace años sistemas de extracción, tratamiento y análisis de datos de sus sistemas de gestión. Además con el tiempo se ha extendido el acceso y la disponibilidad de datos, convirtiéndose en la base de nuevos modelos de negocio más allá del negocio tradicional —como por ejemplo el proyecto Smart Step de Telefónica[2]—, de negocios nuevos basados en datos a cambio de servicios —Google sería el ejemplo paradigmático— o de negocios fundamentados en datos abiertos de origen público o privado —PriceStat[3] sería un buen ejemplo—.

De acuerdo con la Wikipedia, *big data* es un concepto que hace referencia a la acumulación masiva de datos y a los procedimientos usados para identificar patrones recurrentes dentro de esos datos. También según la Wikipedia, la disciplina dedicada a los datos masivos se enmarca en el sector de las tecnologías de la información y la comunicación. Esta disciplina se ocupa de todas las actividades relacionadas con los sistemas que manipulan grandes conjuntos de datos. Las dificultades más habituales vinculadas a la gestión de estas cantidades de datos se centran en la captura, almacenamiento, búsqueda, compartición, análisis y visualización.

Este tipo de definiciones, con una perspectiva tecnológica, describen al *big data* como la gestión y el análisis de enormes volúmenes de datos que no pueden ser tratados de manera convencional, ya que superan los límites y capacidades de las herramientas de *software* habitualmente utilizadas para la captura, gestión y procesamiento de datos.

Sin embargo, el problema de gestión de grandes volúmenes de datos es un problema al que ha tenido que enfrentarse la estadística pública desde hace muchos años. Por ejemplo, a medida que la población de Estados Unidos crecía, la US Census Bureau fue buscando continuamente estrategias para mejorar la velocidad y la precisión del proceso de levantamiento de censos. Cuando los Estados Unidos contaban con 3,9 millones de residentes en el primer censo en 1790, el gran volumen de trabajo de tabular a mano los resultados era uno de los mayores retos. A medida que el país creció, también lo hizo el desafío. Cuando contar los resultados del censo se hizo tan largo que casi duraba una década, la

2. http://dynamicinsights.telefonica.com/blog/488/smart-steps-2
3. http://www.pricestats.com/

búsqueda de soluciones condujo necesariamente a la creación de la moderna tecnología de procesamiento de datos. El primer dispositivo para acelerar el conteo del censo fue creado en 1872 por el Oficial Mayor del Censo Charles W. Seaton. La máquina utilizaba rodillos para sumar las pulsaciones de teclado introducidas manualmente. Sin embargo, incluso con la máquina Seaton el censo se llevó casi la década completa para su procesamiento. Unos años después, en 1880 comenzó a realizarse nuevamente el censo en EE. UU. y debido a la cantidad de personas que lo formaba tardó ocho años en finalizarse, incluso hubo variables que no se llegaron a tabular. Por este motivo, la US Census Bureau llevó a cabo un concurso en 1888 para encontrar un método más eficiente para procesar y tabular el gran volumen de datos que recogía. El resultado fue la tabuladora de Herman Hollerith, génesis de la fichas perforadas y de la empresa IBM[4].

Tal como señala Caballero en el libro *Las bases de big data*[5] el almacenamiento y procesamiento de datos ha sido una de las tareas asociadas a los ordenadores desde su aparición. El primer ordenador comercial, UNIVAC I, construido en 1951, fue adquirido por la Oficina del Censo de Estados Unidos para tratar la ingente cantidad de información obtenida en los censos, a la que había que sumar los datos que comenzaban a recopilarse a través de muchas otras fuentes: hospitales, escuelas, etc. Pronto, UNIVAC reveló su potencia a la hora de realizar cálculos y predicciones estadísticas imposibles hasta el momento. Uno de sus mayores éxitos fue la predicción del resultado de las elecciones presidenciales en 1952. A partir de un recuento de tan solo un 1% del total de votos, UNIVAC predijo que el siguiente presidente sería Eisenhower, mientras que la mayoría de los comentaristas políticos daban por ganador a su rival, el hoy olvidado Stevenson.

Entonces, como diría el Bugs Bunny traducido *"¿Qué hay de nuevo, viejo?"*. En 2001 Douglas Laney[6] propuso tres características que distinguían a lo que ahora denominamos *big data*: **volumen, velocidad y variedad.** Tradicionalmente, como hemos visto, las oficinas de estadísticas se han enfrentado a los problemas de volumen, pero en la actualidad aparecen dos elementos nuevos: la velocidad y la variedad. Siguiendo esta dirección, el primer documento (UNECE, 2013) que estudia el problema *big data* en la estadística pública *What Does "Big Data" Mean for Official Statistics?*[7] lo define como una variante de la propuesta de Douglas Laney:

4. http://www.datosconinteligencia.blogspot.com.es/2015/09/el-viejo-problema-del-big-data-en-la.html
5. Caballero Roldán, Rafael, y Martín Martín, Enrique. *Las bases de big data.* Madrid: Los Libros de la Catarata: Universidad Complutense de Madrid, 2015.
6. http://www.gartner.com/analyst/40872/Douglas-Laney
7. Conference of European Statisticians. "What Does 'Big Data' Mean for Official Statistics?" UNECE, March 10, 2013.

> "*Big data* son las **fuentes de datos** que generalmente pueden ser descritas como de alto volumen, velocidad y variedad, que requieren formas rentables e innovadoras de procesamiento con el fin de mejorar los análisis y de apoyar las tomas de decisiones".

Por lo tanto, para la estadística pública el problema *big data* se aborda como un problema de nuevas fuentes de datos. En esa dirección el problema se enfrenta considerando que estas fuentes de datos podrían complementar o sustituir las fuentes tradicionales utilizadas en la estadística pública, las encuestas y los registros administrativos, pero con algunas características peculiares:

1. La propiedad sobre las fuentes de datos generalmente no es pública, con los problemas derivados para el acceso, uso y mantenimiento de las fuentes.
2. La fuentes de datos no están pensadas para fines estadísticos con los problemas derivados de conceptualización y sesgos.

En el documento anteriormente citado se enumeran algunos de los retos derivados de las características señaladas: (1) **legislativo**, p.e. respecto al acceso y uso de los datos; (2) **privacidad**, p. e., gestión de la confianza pública para la aceptación del uso de esas fuentes y su enlace con otras fuentes de datos; (3) **financiero**, p. e., coste-beneficio potencial de acceso a las fuentes de datos; (4) **gestión**, p. e., políticas y directivas sobre la gestión y protección de los datos; (5) **metodológico**, p. e., calidad de los datos e idoneidad de los métodos estadísticos y (6) **tecnológico**, p. e., temas relacionados con la tecnología de la información.

› La estadística pública en la sociedad datificada

La estadística pública en la encrucijada de la pérdida de hegemonía

El cambio de contexto en el mundo de la información, que hemos señalado en el apartado anterior, tiene implicaciones directas en las oficinas de estadística. Entre ellas encontramos muchas cuestiones prácticas, pero también existe una importante y estratégica: **¿Qué posición quieren ocupar las oficinas de estadística en el futuro sociedad de la información?**

Hasta alrededor de la década de los 80 los datos fueron esencialmente un bien escaso por el alto precio de su adquisición. Antes de la era de la datificación,

mucha información no estaba disponible y debía ser recogida para un propósito particular. La información estadística oficial, basada fundamentalmente en datos de encuestas o censos, tenía un valor único pues simplemente no había otra alternativa. Por ejemplo, los datos de los censos de población, recogidos puerta a puerta, eran inmensamente valiosos para los responsables políticos, investigadores y otros usuarios.

A partir de las década de los 90, los datos recogidos por las Administraciones públicas fueron cada vez más accesibles para fines estadísticos, como consecuencia de la informatización de sus procedimientos. En este escenario, la recopilación de datos estadísticos por medio de cuestionarios se complementó, e incluso se sustituyó, por fuentes de datos administrativas[8], con el fin de reducir costes y reducir la carga sobre los encuestados. Hoy en día algunos países no llevan a cabo amplios estudios poblacionales, y realizan su censo mediante la combinación y el análisis de datos de varias fuentes administrativas. Aún en este contexto, la información proporcionada por las oficinas de estadística seguía siendo única. Esta posición se reforzaba ante la posibilidad de combinar los datos de diferentes fuentes, ya que en muchos países no hay otra organización autorizada para realizar esas combinaciones.

Sin embargo, la datificación está cambiando el entorno de las oficinas de estadística, dando lugar a que la escasez de datos se convierta en un problema menor. Para las oficinas de estadística hay beneficios potenciales en estas nuevas fuentes de datos, de las que surgen nuevas posibilidades tanto en la reducción de cargas a los encuestados y costes de producción, como en la producción de nueva información. Pero también da lugar a la **pérdida de la hegemonía de sus datos**, ya que otros jugadores en el mercado de la información pueden empezar, y de hecho han comenzado a hacerlo, a producir estadísticas que hasta el momento solo ejecutaban las oficinas de estadística.

Por ejemplo, el *Billion Prices Project*[9] del Massachusetts Institute of Technology (MIT) dirigido por Alberto Cavallo y Roberto Rigobon, que en la actualidad se ha convertido en una propuesta comercial a través de la empresa PriceStats[10], nació como una iniciativa académica que utilizaba los precios recogidos diariamente en cientos de tiendas en línea de todo el mundo para llevar a cabo una investigación económica. Este proyecto se fundamentó en la tesis doctoral de Cavallo,

8. Wallgren, Anders, and Britt Wallgren. Register-Based Statistics: Administrative Data for Statistical Purposes. Wiley Series in Survey Methodology. Chichester, England ; Hoboken, NJ: John Wiley & Sons Ltd, 2007.
9. http://bpp.mit.edu/
10. http://www.pricestats.com/

A (2009)[11] en la Universidad de Harvard, y también dió lugar en 2007 a la aparición del proyecto *InflacionVerdadera.com* creado para proveer índices de precios alternativos a los oficiales en Argentina, publicados por el Instituto Nacional de Estadística y Censos (INDEC).

Desde 2007 hasta 2012 se publicó un índice de alimentos y bebidas y otro de la canasta básica alimentaria, utilizando los precios diarios en dos grandes supermercados de Buenos Aires y utilizando las mismas metodologías del INDEC. Los resultados del trabajo, cuyo objetivo era demostrar la manipulación de las estadísticas oficiales en Argentina, fueron publicados en el artículo académico "Online and official price indexes: Measuring Argentina's inflation"[12]. En agosto de 2012 reemplazaron los índices originales de *InflacionVerdadera.com* por un Índice de precios al consumidor producido por PriceStats, comparable al IPC general del INDEC. El índice es publicado semanalmente en la revista *The Economist*[13] como alternativa a las estadísticas oficiales del INDEC.

Este es un claro ejemplo de cómo las fuentes *big data* pueden ser un instrumento al servicio del control externo del cumplimiento de los principios y valores de las oficinas estadísticas reconocidos internacionalmente. En esta nueva situación surgen diversas **cuestiones fundamentales** para una oficina de estadística y **el futuro de la estadística pública**:

1. ¿Cómo garantizar que las oficinas de estadística aportan valor añadido único en el futuro? y en ese sentido ¿las oficinas de estadística deben seguir haciendo estadísticas para las que existe una alternativa de mercado?
2. ¿Pueden las oficinas de estadística asumir nuevas funciones o capacidades, en base a su posición institucional y a los conocimientos que han acumulado? Por ejemplo, ¿se puede garantizar el acceso a fuentes de datos de propiedad privada?
3. ¿Sería mejor cambiar el papel de las oficinas de estadística pasando de la producción de información estadística hacia la validación de la información producida por los demás?

Desde un punto de vista práctico también surgen preguntas importantes respecto al uso potencial de las fuentes *big data*:

11. Cavallo, A. Scraped Data and Sticky Prices: Frequency, Hazards, and Synchronization [recurso Electrónico]. Harvard University, 2009.
12. Cavallo, A. "Online and Official Price Indexes: Measuring Argentina's Inflation." *Journal of Monetary Economics 60,* no. 2 (2013): 152-165. doi:http://dx.doi.org/10.1016/j.jmoneco.2012.10.002.
13. http://www.economist.com/node/21548242

1. ¿En qué medida son útiles las fuentes *big data* para la producción y la mejora de las estadísticas públicas actuales? y ¿qué nueva información puede producir una oficina estadística mediante el uso de estas nuevas fuentes de datos?
2. ¿Cuál debe ser el marco jurídico de acceso a las fuentes *big data* para fines estadísticos?, y si se tuviera acceso, ¿cuáles son los riesgos de usar datos sobre los que no se controla su generación por parte de las Administraciones públicas? además de ¿cómo asegurar la seguridad y confidencialidad de dichos datos?
3. ¿Cuáles son los requerimientos metodológicos y tecnológicos para el uso de fuentes *big data*?
4. El uso de estas fuentes ¿significa cambios de procedimientos?, ¿es necesario aumentar la velocidad de producción/difusión de datos para aprovechar una de las principales características de estas fuentes?

Misión, principios y valores de la estadística pública en el contexto *big data*

Naciones Unidas reconoce a las estadísticas oficiales como un elemento indispensable en el sistema de información de una sociedad democrática pues proporcionan a los gobiernos, a la economía y a la ciudadanía datos de la situación económica, demográfica, social y ambiental de un país o de una región. En ese sentido considera que la información estadística es esencial para el desarrollo, pero también para el conocimiento mutuo y el comercio entre los Estados y los pueblos del mundo. Con este fin, NN. UU. indica que las oficinas de estadística han de compilar y facilitar, de forma imparcial, estadísticas oficiales de comprobada utilidad práctica para que los ciudadanos puedan ejercer su derecho a mantenerse informados.

Pero para que los ciudadanos confíen en las estadísticas oficiales, los organismos estadísticos deben contar con un conjunto de valores y principios fundamentales. De acuerdo con Naciones Unidas, los principios generales son la (1) independencia, (2) la pertinencia o relevancia, (3) la credibilidad, así como (4) el respeto a los derechos de los informantes. Estos principios han sido desarrollados en los principios fundamentales de las estadísticas oficiales[14].

En coherencia con las líneas trazadas por Naciones Unidas, en el ámbito europeo, el Reglamento (CE) nº 223/2009, del Parlamento Europeo y del Consejo de

14. Documentos Oficiales del Consejo Económico y Social, 1994, suplemento nº 9 (E/1994/29), cap. V. Para más información véase el apéndice II o consúltese el sitio web <http://unstats.un.org/unsd/statcom/doc94/s1994.htm>

11 de marzo de 2009 relativo a la estadística europea, señala los siguientes principios en su artículo 2: (1) Independencia profesional, (2) imparcialidad, (3) fiabilidad, (4) secreto estadístico y (5) rentabilidad. Estos principios estadísticos se desarrollaron posteriormente en el Código de Buenas Prácticas de la Estadística Europea, que tiene por finalidad garantizar la confianza de la población en las estadísticas europeas mediante la determinación de la forma en que deben desarrollarse, elaborarse y difundirse las estadísticas con arreglo a los principios estadísticos europeos y a las mejores prácticas internacionales.

Las normas citadas desempeñan un papel vital en la obtención de la confianza en las estadísticas oficiales. A su vez estas normas se refuerzan con los códigos éticos de los estadísticos, destacando la *Declaración sobre Ética Profesional* del Instituto Internacional de Estadística (ISI), que además se complementan con diferentes códigos éticos elaborados por los distintos sistemas estadísticos nacionales.

En ese sentido, la pregunta que nos debemos hacer en un principio desde la estadística oficial es cómo el nuevo contexto *big data* encaja dentro de la misión, principios y valores que guía nuestra actividad pública. Para ellos vamos a realizar una revisión sintética a partir de la agrupación de los principios en tres grandes bloques:

1. *Big data* y los principios asociados a las fuentes de datos para fines estadísticos.
2. *Big data* y los principios asociados al derecho de acceso y la protección de la intimidad.
3. *Big data* y los principios de objetividad política y científico-técnica.

Big data *y los principios asociados a las fuentes de datos para fines estadísticos*

En la Resolución sobre los Principios Fundamentales de las Estadísticas Oficiales aprobada por la Asamblea General de NN. UU. el 29 de enero de 2014, indica que los datos para fines estadísticos pueden obtenerse de todo tipo de fuentes, ya sea encuestas estadísticas o registros administrativos. Sorprende que no haya mención explícita a las fuentes *big data*, siendo una resolución del año 2014, pero de la esencia del principio podríamos extraer que la intención es establecer que la estadística pública pueda realizarse no solo a partir de encuestas, sino con cualquier tipo de fuente de datos útil para sus fines.

Esta propuesta de pluralismo de fuentes se ordena en el principio mencionado, indicando que estas se deben seleccionar considerando: su calidad, oportunidad, costo y carga que impondrá a los encuestados. Los criterios de oportunidad, costo

y carga a los encuestados son también considerados en el Código de Buenas Prácticas de las Estadísticas Europeas y son fácilmente comprensibles; sin embargo, el criterio de calidad necesita ser explicitado cuando se trabaja con datos no recopilados con fines estadísticos como pueden son los datos administrativos o las fuentes *big data*. En ese sentido debemos referenciar una propuesta sobre marco de calidad para el uso de fuentes *big data* en la estadística pública, elaborada por UNECE Big Data Quality[15] e inspirada en el documento *Checklist for the Quality Evaluation of Administrative Data Sources*[16]. Este marco se estructura en tres hiperdimensiones, cada una con sus dimensiones de calidad, que a su vez se organizan en factores a considerar.

Dimensiones del marco de calidad para el uso de fuentes *big data*

Hiperdimensión	Dimensiones de calidad	Factores a considerar
Fuente	Entorno institucional	Sostenibilidad de la entidad proveedora de datos Confiabilidad general de los datos Transparencia e interpretabilidad de la entidad proveedora y de los datos
	Privacidad y seguridad	Legislación que afecta a los datos Restricciones de privacidad, seguridad y confidencialidad Percepción ciudadana sobre el uso de los datos
Metadatos	Complejidad	Restricciones técnicas Datos estructurados, semiestructurados o no estructurados Legibilidad de los datos Presencia de jerarquías y anidamientos
	Completitud	Metadatos disponibles, interpretables y completos
	Usabilidad	Recursos adicionales necesarios para el tratamiento de los datos Análisis de los riesgos
	Tiempo	Oportunidad Periodicidad Cambios a través del tiempo

15. UNECE Big Data Quality Task Team. "A Suggested Big Data Quality Framework." UNECE, December 2014.
16. Piet Daas, Saskia Ossen, Rachel Vis-Visschers, and Judit Arends-Tóth. "Checklist for the Quality Evaluation of Administrative Data Sources." Discussion Paper. The Hague/Heerlen: Statistics Netherlands, 2009.

Dimensiones del marco de calidad para el uso de fuentes *big data* (cont.)

Hiperdimensión	Dimensiones de calidad	Factores a considerar
Metadatos	Enlazamiento	Presencia y calidad de variables de enlace Niveles al que se puede realizar enlazamiento
	Coherencia y consistencia	Estandarización Disponibilidad de metadatos para variable clave
	Validez	Transparencia de métodos y procesos Solvencia de métodos y procesos
Datos	Exactitud y selectividad	Error total de la muestra Datos de referencia con los que comparar Selectividad. Problemas de cobertura
	Enlazamiento	Calidad de las variables de enlace
	Coherencia y consistencia	Coherencia entre los metadatos y los datos
	Validez	Coherencia de los procesos y métodos con los datos observados

Big data *y los principios asociados al derecho de acceso y la protección de la intimidad*

El Código de Buenas Prácticas de las Estadísticas Europeas indica claramente, en su principio sobre recogida de datos, que las autoridades estadísticas deben tener un mandato jurídico claro para recoger la información destinada a la elaboración de estadísticas. Asimismo señala que a petición de las autoridades estadísticas, **se puede obligar por ley** a las Administraciones, las empresas, los hogares y el público en general a permitir el acceso a los datos destinados a la elaboración de estadísticas europeas o a facilitar dichos datos.

En el apartado primero de este capítulo señalamos que una de las características peculiares de las fuentes *big data* es que generalmente la propiedad sobre las mismas no es pública. Asimismo, tal como veremos más adelante, muchas empresas han encontrado en estos datos un nuevo nicho de mercado que hasta el momento no habían explotado; donde los clientes potenciales identificados son tanto el sector privado como el sector público. Estos nichos de mercados se definen no tanto como acceso a datos sino como acceso a servicios a partir de datos, así tenemos, por ejemplo, el proyecto Smart Step de Telefónica[17].

17. http://dynamicinsights.telefonica.com/blog/488/smart-steps-2

Evgeny Morozov, investigador sobre estudios políticos e implicaciones sociales de la tecnología, en una entrevista[18] de presentación de su último libro *La locura del solucionismo tecnológico*[19] en el diario *El País* señala que "los datos son una de las más preciadas mercancías". A lo largo de la entrevista, Morozov insiste en que en las últimas cinco décadas los datos se han convertido en una de las más preciadas mercancías:

> "Tu seguro quiere saber qué posibilidades tienes de enfermar; tu banco quiere saber qué probabilidades tienes de no pagar tu hipoteca. Hay un mercado gigante de la venta de datos, no solo de tipo digital: si no miras lo que firmas cuando ofreces datos, es más que posible que acaben siendo agregados en una base administrada por un puñado de firmas norteamericanas".

¿Y qué es lo que se debería hacer con ellos?, Evgeny Morozov plantea tres opciones:

1. Una es el *statu quo*: que un par de monopolios, Google y Facebook, continúen recopilando aún más información sobre nuestra vida para que pueda ser integrada en dispositivos inteligentes: mesas inteligentes, termostatos inteligentes; cualquier cosa que tenga un sensor generará un dato. Google Now es el paradigma de un sistema que intenta hacer acopio de todos esos datos para hacer predicciones y darte ideas. Si sabe que vas a volar te recuerda que hagas el *check in* y te dice el tiempo que te va a hacer en destino, como un asistente virtual. Es el discurso de Google en términos de movilidad social: dar a los pobres los servicios que los ricos ya reciben.
2. La segunda es seguir a los disruptores. Hay compañías que chupan nuestros datos y los convierten en dinero. Una solución es que cada cual capture sus propios datos y los integre en un perfil, dando acceso a quien quiera y cobrando por ello. De ese modo, cada persona se convierte en un empresario.
3. Y la tercera opción aún no está muy articulada, pero debería ser perseguida según Morozov. Los datos, en un buen marco político, económico y legal, pueden llevarnos a servicios fantásticos. El único futuro del transporte público es una combinación de datos, algoritmos y sensores que determinan dónde está la gente y adónde quiere ir.

18. http://elpais.com/elpais/2015/12/17/eps/1450358550_362012.html
19. Morozov, Evgeny. La locura del solucionismo tecnológico. Madrid; Móstoles, Madrid; Buenos Aires: Clave Intelectual ; Katz, 2015.

En ese sentido, Evgeny Morozov indica que habría que oponerse a que el paradigma de la propiedad privada se extienda a los datos:

> "Ha habido esfuerzos de comercializar hasta el aire, y hay que oponerse. Los datos, sin la capacidad de analizarlos, no son gran cosa. Hoy en día solo algunas grandes empresas son capaces de estudiarlos. Esa información debería estar bajo un control público, que no significa un control del Estado, sino de los ciudadanos. La reciente fascinación en Europa por esa idea del común, que no tiene nada que ver con la de los comunes, es un marco sano. La gente podría ceder esos datos voluntariamente, pero siendo propietaria de estos".

Esta perspectiva contrasta con la aportaciones del European Big Data Value Partnership[20] en sus informes *European Big Data Value Strategic Research & Innovation Agenda*[21] en los que se ponen en valor la potencialidad económica de las fuentes *big data* y se define una agenda estratégica de investigación e innovación europea para su desarrollo.

Como vemos, hay un debate intenso sobre los datos, su propiedad y el derecho de acceso para fines públicos. Si bien la legislación estadística puede obligar a facilitar el acceso a las oficinas estadísticas, esta capacidad tendrá que convivir en la tensión de intereses público-privado contrapuestos; tensión que necesitará de espacios de cooperación con los proveedores. En esa línea se sitúan seminarios como el *Joint OECD-PARIS21 Workshop-Access to New Data Sources for Statistics: Business Models for Private-Public Partnerships*[22] para cuya preparación se elaboró el informe *Public-Private Partnerships for Statistics (*Klein, Jütting, and Robin, 2016) que es un buen análisis sobre el problema aquí planteado.

Por otra parte, el *Manual de organización estadística* nos recuerda que la potestad que confiere la legislación a las oficinas de estadística para recabar información no es de mayor utilidad, a menos que todos los sectores de la sociedad estén dispuestos a cooperar. En ese sentido es importante señalar que la confidencialidad de la información individual es, probablemente, la mayor preocupación de los informantes; especialmente cuando se trata de gran acumulación de datos por parte del Estado, datos que en un principio han sido generados por los ciudadanos para otros fines distintos a los estadísticos.

20. http://www.bdva.eu/
21. Big Data Value Europe. *European Big Data Value Strategic Research & Innovation Agenda.* Big Data Value Association, January 2016.
22. http://www.oecd.org/std/oecd-paris21-workshop-access-to-new-data-sources-for-statistics.htm

Ante lo expuesto es importante señalar que existe el peligro de que entre la sociedad se genere una visión de las oficinas de estadísticas como instituciones orwellianas. Por ejemplo, tras la publicación del artículo denominado "Las operadoras seguirán el rastro de tu móvil para alimentar el censo de 2021"[23] en el que se hace público por parte del Instituto Nacional de Estadística (INE) de España el uso de datos de telefonía móvil para los estudios de movilidad del Censo de 2021, se desató un amplio debate en Menéame[24] contrario a su uso. Paralelamente se publicaron varios artículos en blogs especializados sobre la legalidad de la medida, como por ejemplo el artículo titulado "La ilegalidad de usar los datos del móvil para completar el censo"[25].

Big data *y los principios de objetividad política y científico-técnica*

El *Manual de Organización Estadística* elaborado por NN. UU. advierte que para tener credibilidad y desempeñar su función es preciso que las oficinas de estadística tengan una posición de independencia ampliamente reconocida. Sin la credibilidad derivada de un alto grado de independencia, los usuarios perderán la confianza en la exactitud y la objetividad de la información del organismo y quienes le proporcionan los datos estarán menos dispuestos a cooperar con él. Esta credibilidad se desarrolla en varios principios fundamentales de las estadísticas oficiales:

1. *Relevancia, imparcialidad y acceso equitativo:* las estadísticas oficiales constituyen un elemento indispensable en el sistema de información de una sociedad democrática y proporcionan al gobierno, a la economía y al público datos acerca de la situación económica, demográfica, social y ambiental. Con este fin, los organismos oficiales de estadística han de compilar y facilitar en forma imparcial estadísticas oficiales de comprobada utilidad práctica para que los ciudadanos puedan ejercer su derecho a la información pública.
2. *Patrones profesionales, principios científicos y ética:* para mantener la confianza en las estadísticas oficiales, las oficinas de estadística han de decidir con arreglo a consideraciones estrictamente profesionales, incluidos los principios científicos y la ética profesional, acerca de los métodos y procedimientos para la reunión, el procesamiento, el almacenamiento y la presentación de los datos estadísticos.
3. *Responsabilidad y transparencia:* para facilitar una interpretación correcta de los datos, las oficinas de estadística han de presentar información conforme

23. http://www.eldiario.es/hojaderouter/tecnologia/moviles/censo-2021-INE-big_data-operadoras_0_493100796.html
24. https://www.meneame.net/m/tecnolog%C3%ADa/operadoras-seguiran-rastro-tu-movil-alimentar-censo-2021
25. http://derechoynormas.blogspot.com.es/2016/03/la-ilegalidad-de-usar-los-datos-del.html

a normas científicas sobre las fuentes, métodos y procedimientos de la estadística.

4. *Uso de patrones internacionales:* la utilización por las oficinas de estadística de cada país de conceptos, clasificaciones y métodos internacionales fomenta la coherencia y eficiencia de los sistemas estadísticos a nivel oficial.

El Código de Buenas Prácticas de las Estadísticas Europeas es más exhaustivo respecto al conjunto de principios relacionados con la objetividad política y científico-técnica:

1. *Independencia profesional*. La independencia profesional de las autoridades estadísticas frente a otros departamentos y organismos políticos, reguladores o administrativos, y frente a los operadores del sector privado, garantiza la credibilidad de las estadísticas europeas.
2. *Imparcialidad y objetividad*. Las autoridades estadísticas desarrollan, elaboran y difunden estadísticas europeas respetando la independencia científica y de forma objetiva, profesional y transparente, de modo que todos los usuarios reciben el mismo trato.
3. *Metodología sólida*. Las estadísticas de calidad se apoyan en una metodología sólida, que requiere herramientas, procedimientos y conocimientos adecuados.
4. *Procedimientos estadísticos adecuados.* Las estadísticas de calidad se apoyan en procedimientos estadísticos adecuados, aplicados desde la recogida de los datos hasta la validación de los mismos.
5. *Precisión y fiabilidad.* Las estadísticas europeas reflejan la realidad de manera precisa y fiable.
6. *Coherencia y comparabilidad.* Las estadísticas europeas son consistentes internamente a lo largo del tiempo y comparables entre regiones y países; es posible combinar y utilizar conjuntamente datos relacionados procedentes de fuentes diferentes.

Revisando los principios y considerando que las fuentes *big data*, tal como hemos señalado anteriormente, en buena medida son de origen privado y que además no están diseñadas para fines estadísticos, se pueden dar algunos problemas que las oficinas estadísticas deben saber abordar. Por ejemplo:

1. Desconfianza de la ciudadanía en los resultados estadísticos, como producto de su desconfianza en las empresas cedentes de los datos y en la no manipulación de los mismos por parte de dichas empresas a favor de sus

intereses económicos, o la ruptura de los acuerdos de cesión si los datos no les son favorables. En definitiva, no es más que una nueva figura de desconfianza sobre la independencia profesional de la oficinas estadísticas frente a los operadores del sector privado.

2. Dificultad para armonizar distintas fuentes con diferentes objetivos, con la finalidad de poder proporcionar datos comparables entre regiones y países; y consistentes internamente a lo largo del tiempo.
3. Problemas metodológicos no triviales, al estar habitualmente ante grandes volúmenes de datos que no son datos censales, sino en todo caso muestras de una población o más genéricamente de eventos de una población. Por lo tanto nos encontramos ante la suma de las dificultades metodológicas producto de muestras no probabilísticas, a las que se deben sumar los problemas habituales de las estadísticas basadas en registros administrativos.

› Los retos a los que se enfrenta la estadística pública

La respuesta de la estadística pública

En 2010, la Oficina de la Conferencia de Estadísticos Europeos[26] creó el Grupo de Alto Nivel Modernisation of Statistical Production and Services (HLG)[27] para supervisar y coordinar el trabajo internacional sobre modelos de negocio dentro de las oficinas de estadística. Dentro de este grupo se formó un equipo de trabajo de expertos, coordinados por la Secretaría de la UNECE[28], con el objetivo de producir un documento que explicara los problemas relacionados sobre el uso del *big data* por las oficinas estadísticas.

El grupo de trabajo publicó en marzo de 2013 el documento *What Does "Big Data" for Official Statistics?*[29] que es el primer documento estratégico que analiza los principales desafíos en materia de legislación, privacidad, cuestiones financieras, gestión, metodologías y tecnología y que además ofrece algunas recomendaciones básicas para las oficinas de estadística. El documento señala desde un primer momento que la recolección de datos de fuentes *big data* y su

26. http://www.unece.org/stats/cesbureau.html
27. http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Official+Statistics
28. http://www.unece.org/info/ece-homepage.html
29. Conference of European Statisticians. "What Does 'Big Data' Mean for Official Statistics?" UNECE, March 10, 2013. http://www1.unece.org/stat/platform/download/attachments/58492100/Big+Data+HLG+Final.docx?version=1&modificationDate=1362939424184.

incorporación al proceso de producción de estadísticas no es tarea fácil, y en ese sentido intenta abordar dos cuestiones elementales:

1. En qué conjuntos de datos deben centrar su atención las oficinas de estadística.
2. Cómo una oficina de estadística puede utilizar la fuentes *big data* y los retos asociados a su uso.

En esta dirección, la Comisión de Estadística de Naciones Unidas acordó, en su 45ª sesión de marzo de 2014, crear el Grupo de Trabajo Global (GTG)[30] sobre *big data* y Estadísticas Oficiales. Este grupo de trabajo nació tras la celebración del seminario previo a la 44ª sesión de la Comisión de Estadística en 2013 sobre *Big Data for Policy, Development and Official Statistics*[31]. En este seminario oradores del sector privado y de oficinas de estadística llegaron a la conclusión de que las fuentes *big data* constituyen una fuente de información que no puede ser ignorada por la estadística pública y que los estadísticos oficiales debían organizarse y tomar medidas urgentes para explotar las posibilidades y abordar los retos asociados con eficacia.

Con la aprobación del grupo de trabajo, la comunidad estadística internacional reconoce el potencial de las fuentes *big data* para las estadísticas oficiales. Las labores del grupo y sus comisiones de trabajo se han ido complementando con seminarios y conferencias internacionales, hasta la fecha han sido los siguientes:

- 2014 Octubre (Beijing)-International Conference on Big Data for Official Statistics[32].
- 2015 Marzo (Nueva York)-Big Data Seminar at the 46th UN Statistical Commission[33].
- 2015 Octubre (Abu Dhabi)-2nd Global International Conference on Big Data for Official Statistic[34].
- 2016 Septiembre (Dublin)-3rd Global International Conference on Big Data for Official Statistics[35].

30. http://unstats.un.org/bigdata/
31. http://unstats.un.org/unsd/statcom/statcom_2013/seminars/Big_Data/default.html
32. http://unstats.un.org/unsd/trade/events/2014/beijing/default.asp
33. http://unstats.un.org/unsd/statcom/statcom_2015/seminars/big_data/default.html
34. http://unstats.un.org/unsd/trade/events/2015/abudhabi/default.asp
35. http://unstats.un.org/unsd/bigdata/conferences/2016/default.asp

El nacimiento del Grupo de Trabajo Global de Naciones Unidas vino precedido por el Scheveningen Memorandum[36] sobre *Big Data and Official Statistics* adoptado por el European Statistical System Committee (ESSC)[37]. Los acuerdos incluidos en el memorándum son los siguientes:

1. *Reconocimiento.* Reconocer que el *big data* representa nuevas oportunidades y desafíos para las estadísticas oficiales, y por lo tanto animar al Sistema Estadístico Europeo y sus socios a examinar el potencial del *big data* en ese sentido.
2. *Necesidad de estrategia*. Reconocer que el *big data* es un fenómeno que está afectando a muchos ámbitos. Por tanto, es esencial desarrollar una "Estrategia de estadísticas oficiales basadas en *big data*" y examinar el lugar y las interdependencias de esta estrategia en el contexto más amplio de una estrategia global del gobierno a nivel nacional, así como a nivel de la UE.
3. *Legislar el acceso de datos*. Reconocer las implicaciones del *big data* en la legislación de protección de datos y derechos de las personas (por ejemplo, acceso a fuentes de datos en poder de terceros), implicaciones que deben ser abordadas apropiadamente como un asunto prioritario.
4. *Compartir experiencias*. Tener en cuenta que varios institutos nacionales de estadística están iniciando actualmente o considerando los diferentes usos del *big data* en un contexto nacional. Es necesario compartir las experiencias obtenidas en los proyectos *big data* concretos y colaborar dentro del Sistema Estadístico Europeo y a escala internacional.
5. *Formación.* Reconocer que el desarrollo de las capacidades y habilidades necesarias para explorar con eficacia los *big data* es esencial para su incorporación en el Sistema Estadístico Europeo. Esto requiere esfuerzos sistemáticos, con cursos de formación adecuados y el establecimiento de comunidades de intercambio de experiencias y buenas prácticas.
6. *Cooperación.* Reconocer el carácter multidisciplinar del *big data*, lo que requiere sinergias y asociaciones entre los expertos y las partes interesadas de diversos dominios, incluyendo gobierno, universidades y titulares de las fuentes de datos privadas.
7. *Innovación metodológica y tecnológica.*. Reconocer que el uso de grandes volúmenes de datos en el contexto de las estadísticas oficiales requiere nuevos desarrollos metodológicos, de evaluación de la calidad y de abordaje de los problemas tecnológicos relacionados. El Sistema Estadístico Europeo debería hacer un esfuerzo especial para apoyar esos desarrollos.

36. http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13
37. http://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/essc

8. *Plan de acción*. Los directores coinciden en la importancia de dar seguimiento a la implementación del memorando, por lo consideran que es necesario adoptar un plan de acción y plan de trabajo del Sistema Estadístico Europeo para el uso de fuentes *big data*.

La *European Statistical System-Vision 2020*[38] es una respuesta estratégica común del Sistema Estadístico Europeo a los desafíos a lo que se enfrentan las estadísticas oficiales. En ella se identifica como uno de los elementos clave para el sistema la incorporación de nuevas fuentes de datos, en ese aspecto el sistemas se visiona para el 2020 de la siguiente manera:

> "Basamos nuestros productos y servicios estadísticos en encuestas tradicionales y nuevas fuentes, incluyendo datos administrativos, geoespaciales y, cuando sea posible, fuentes *big data*. Las nuevas fuentes de datos complementan las ya existentes y nos ayudan a mejorar la calidad de nuestros productos. Vamos a trabajar juntos para conseguir el acceso a nuevas fuentes de datos, crear métodos y encontrar la tecnología adecuada con el fin de utilizar nuevas fuentes de datos para elaborar estadísticas europeas de una manera fiable".

La necesidad de elaborar un plan de acción fue uno de los elementos considerados en la convocatoria del *2014 ESS Big Data Event: Big Data in Official Statistics*[39]. Posteriormente en la 22ª Sesión del European Statistical System Committee (ESSC)[40] se aprobó el documento *Big Data Action Plan and Roadmap 1.0*[41] en el que se se plantea una visión para 2020 y post-2020.

Los retos identificados por la estadística pública

Como resumen podríamos decir que es evidente que el desafío del uso de datos de fuentes *big data* dentro de la estadística pública significa necesariamente la modernización de las oficinas estadísticas. Ese desafío requiere el abordaje de diferentes retos, que sintéticamente podemos resumir en los siguientes puntos:

38. http://ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255
39. https://ec.europa.eu/eurostat/cros/sites/crosportal/files/Big%20Data%20Event%202014%20-%20Technical%20Final%20Report%20-finalV01_0.pdf
40. http://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/essc
41. Eurostat Big Data Task Force. "Big Data Action Plan and Roadmap 1.0." Eurostat. Accessed February 28, 2016. https://ec.europa.eu/eurostat/cros/sites/crosportal/files/ESSC%20doc%2022_8_2014_EN_Final%20with%20ESSC%20opinion.pdf.

- **Estrategia:** es necesario definir cómo integrar las nuevas fuentes *big data* en la actividad de las oficinas estadísticas. Esta estrategia puede estar dirigida tanto a la integración de las nuevas fuentes en la producción habitual de las oficinas, como en la identificación de nueva información estadística basada en dichas fuentes.
- **Acceso:** existe un debate intenso sobre los datos, su propiedad y el derecho de acceso para fines públicos. Si bien la legislación estadística puede obligar a facilitar el acceso a las oficinas estadísticas, esta capacidad tendrá que convivir en la tensión de intereses público-privado contrapuestos; tensión que necesitará de espacios de cooperación con los proveedores.
- **Privacidad:** la datificación de buena parte de nuestras vidas genera actitudes diversas en la opinión pública sobre el derecho a la intimidad. Sin embargo cuando se trata de gran acumulación de datos por parte del Estado la confidencialidad, proporcionalidad y fin de los mismos pasan a ser una importante preocupación ciudadana. En ese sentido existe el peligro de que entre la sociedad se genere una visión de las oficinas de estadísticas como instituciones orwellianas.

  Por otra parte, la generación de gran cantidad de datos a gran velocidad pone sobre la mesa nuevos retos tecnológicos para cumplir el mandato del deber de secreto estadístico, que impide que a través de la información publicada por las oficinas estadísticas se pueda identificar directa o indirectamente a las unidades de observación.
- **Calidad:** la dimensiones de evaluación de la calidad de las fuentes *big data* para su integración en la actividad de las Oficinas Estadísticas deben ser identificadas, especialmente debido a que son datos recopilados para fines no estadísticos.
- **Metodología:** con las fuentes *big data* nos encontramos ante la dificultad de datos recopilados para fines no estadísticos, por lo tanto estamos ante problemas similares a los planteados con los registros administrativos, al menos en lo que respecta a los conceptos usados en la recolección de datos y su relación con las definiciones internacionalmente armonizadas. Además muchas de las fuentes *big data* son muestras, con el problema añadido de ser muestras no probabilísticas y posiblemente sesgadas por el método o por las cuotas de mercado del agente recolector.
- **Tecnología:** la incorporación de fuentes *big data* a la actividad estadística requerirá de la incorporación de tecnología *big data* a las oficinas estadísticas. Definir arquitecturas, *hardware* y *software* requeridos es uno de los retos que debe ser abordado.
- **Formación:** el desarrollo de las capacidades y habilidades necesarias para explorar con eficacia los *big data* es esencial para su incorporación a la actividad

de la Oficinas Estadísticas. Esto requiere esfuerzos sistemáticos, como cursos de formación adecuados y el establecimiento de comunidades de intercambio de experiencias y buenas prácticas.

Referencias bibliográficas

Big Data Value Europe (2015). "European Big Data Value Strategic Research & Innovation Agenda". Big Data Value Association, January 2015. http: //www.bdva.eu/ sites/default/files/europeanbigdatavaluepartnership_sria__v1_0_final. pdf#overlay-context=downloads%26page%3D1%3Fq%3Ddownloads%26p age%3D1.

Big Data Value Europe (2016). "European Big Data Value Strategic Research & Innovation Agenda". Big Data Value Association, January 2016. http: //www.bdva.eu/ sites/default/files/EuropeanBigDataValuePartnership_SRIA__v2.pdf.

Borgman, Christine L. (2015). *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Massachusetts: The MIT Press.

Caballero Roldán, R., Martín Martín, E. (2015). *Las bases de Big Data*. Madrid: Los Libros de la Catarata- Universidad Complutense de Madrid.

Cavallo, A. (2009). *Scraped Data and Sticky Prices: Frequency, Hazards, and Synchronization [recurso Electrónico]*. Harvard University. https: //books.google.es/ books?id=3r_-ZwEACAAJ.

Cavallo, A. (2013). "Online and Official Price Indexes: Measuring Argentina's Inflation". *Journal of Monetary Economics,* 60(2): 152-165. doi: http: //dx.doi.org/10.1016/j.jmoneco.2012.10.002.

Conference of European Statisticians. "What Does 'Big Data' Mean for Official Statistics?". UNECE, March 10, 2013. http: //www1.unece.org/stat/platform/download/attachments/58492100/Big+Data+HLG+Final.docx?version=1&modificationDate=1362939424184.

Eurostat (2016). "European Statistical System-Vision 2020". *Eurostat. Accessed November* 6. http: //ec.europa.eu/eurostat/documents/10186/756730/ESS-Vision-2020.pdf/8d97506b-b802-439e-9ea4-303e905f4255.

Klein, T., Jütting, J., Robin, N. (2016) "Public-Private Partnerships for Statistics: Lessons Learned, Future Steps". *OECD Development Co-operation Working Papers*, february 29. http: //www.oecd-ilibrary.org/ development/public-private-partnerships-for-statistics-lessons-learned-future-steps_5jm3nqp1g8wf-en.

Letouzé, E. (2012). "Big Data for Development: Challenges & Opportunities". UN Global Pulse, May 2012. http: //www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf.

Maeztu, D. (2016). "La ilegalidad de usar los datos del móvil para completar el censo. Del derecho y las normas". Accessed May 4. http: //derechoynormas.blogspot.com.es/2016/03/la-ilegalidad-de-usar-los-datos-del.html?spref=tw.

Mayer-Schönberger, V., Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Houghton Mifflin Harcourt.

Morozov, E. (2015). *La locura del solucionismo tecnológico*. Madrid; Móstoles, Madrid; Buenos Aires: Clave Intelectual; Katz.

Piet, D., Ossen, S., Vis-Visschers, R., Arends-Tóth, J. (2009). "Checklist for the Quality Evaluation of Administrative Data Sources". Discussion Paper. The Hague/Heerlen: Statistics Netherlands. http: //ec.europa.eu/ eurostat/documents/64157/4374310/45-Checklist-quality-evaluation-administrative-data-sources-2009.pdf/24ffb3dd-5509-4f7e-9683- 4477be82ee60.

Reimsbach-Kounatze, Ch. (2015). "The Proliferation of 'Big Data' and Implications for Official Statistics and Statistical Agencies". OECD Digital Economy Papers, January 12. http: //www.oecd-ilibrary.org/science-and-technology/ the-proliferation-of-big-data-and-implications-for-official-statistics-and-statistical-agencies_5js7t9wqzvg8-en.

Struijs, P., Braaksma, B., Daas, Piet J. H. (2014). "Official Statistics and Big Data". *Big Data & Society,* 1(1) (June 10). doi: 10.1177/2053951714538417.

UNECE (2014). Big Data Quality Task Team. "A Suggested Big Data Quality Framework". UNECE, December.

Unión Europea. *Directiva 2013/37/EU del Parlamento Europeo y del Consejo de 26 de junio, por la que se modifica la Directiva 2003/98/EC relativa a la reutilización de la información del sector público*. Accessed February 20, 2016. http: //eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ: L: 2013: 175: 000 1: 0008: ES: PDF.

United Nations (2005). *Manual de Organización Estadística. El Funcionamiento y la Organización de una Oficina de Estadística.* New York: United Nations Publications. http: //www.cepal.org/publicaciones/xml/7/15497/lcw6e.pdf.

Wallgren, A., Wallgren, B. (2007). *Register-Based Statistics: Administrative Data for Statistical Purposes*. Wiley Series in Survey Methodology. Chichester, England ; Hoboken, NJ: John Wiley & Sons Ltd, 2007.

Capítulo 5

Acceso, privacidad y ética pública en la era del *big data*

Noemí Brito*

› Los datos como piedra angular de la actual economía y sociedad del conocimiento

Introducción

Desde hace ya algunos años, se aboga por la idea de construir **una nueva economía mundial basada en el adecuado procesamiento a escala de los datos e información que fluye a nivel global**[1].

Las decisiones más apropiadas, ya sean públicas o privadas, se nutren de la mejor información de base posible convirtiéndose esta en la nueva moneda de cambio[2], en el elemento más preciado para todo y por todos, elevándose a la categoría de bien o recurso esencial para el crecimiento económico, la creación de empleo y, en general, para el progreso social y económico[3]. Tal y como adelantaba el propio Francis Bacon, "la información es poder" y, ahora, el "poder ansía como nunca información".

Sin embargo, resulta curioso cómo **hoy en día la mera información ya no parece ser suficiente por sí misma**. En tal sentido, esta reporta mayor valor si es

* Abogada digital. Directora de Derecho Digital en LEGISTEL e IT GRC en COMTRUST. Vocal Junta Directiva de ENATIC.

1. Es interesante conocer el proyecto pionero que desarrolla el Observatorio Nacional de las Telecomunicaciones y de la Sociedad de la Información (ONTSI), en colaboración con la Universidad Carlos III de Madrid, para determinar las posibilidades sobre reutilización de información de Internet como fuente de datos (Internet as Data Source-IaD): http://www.red.es/redes/sala-de-prensa/noticia/el-ontsi-desarrolla-un-proyecto-pionero-sobre-la-viabilidad-de-usar-Internet-

2. Según la propia Comisión Europea, el valor de la economía de datos de la UE se estimó en 257.000 millones de euros en 2014, es decir, el 1,85% del PIB de la UE. Esto aumentó a 272.000 millones de euros en 2015, es decir, el 1,87% del PIB de la UE. La misma estimación prevé que si se establecen las condiciones políticas y jurídicas para la economía de datos en el tiempo, su valor aumentará a 643.000 millones de euros para 2020, lo que representa el 3,17% del PIB total de la UE.

3. Remítase a la Comunicación de la Comisión Europea titulada "BUILDING A EUROPEAN DATA ECONOMY". COM (2017) 9 final: https://ec.europa.eu/digital-single-market/en/towards-thriving-data-driven-economy

parametrizada, trazada y cruzada con otras tantas fuentes y elementos informativos de interés a través de la aplicación de diversas técnicas y algoritmos, sometiéndola por consiguiente a un examen y análisis inteligente y, en conclusión, realizando una gestión eficaz de la misma. Y es en este punto en el que las tecnologías *big data* aportan soluciones precisas para recabar la mejor información disponible en cada momento y para caso sobre cuestiones, situaciones y sectores de actividad muy concretos.

Por ello, no es de extrañar que el ***big data* haya pasado a un primer plano en el debate mundial por el desarrollo**[4].

A través del *big data* se pretende generar nueva información que permita tanto a públicos, como a privados adoptar posicionamientos específicos y depurados en torno a posibles programas, iniciativas y proyectos que deseen lanzar u activar en un sinfín de campos, como son el sanitario, el medioambiental, el energético, el industrial, los servicios, el comercial, los sistemas de logística y de transporte, el desarrollo de las ciudades y demás proyectos inteligentes, etc. Incluso, el *big data* podría ofrecer la información necesaria de partida para la elaboración de las nuevas normas y reglamentaciones que se prevean.

En definitiva, su utilización es capaz de impulsar la competitividad, la calidad y eficacia de los servicios públicos, la reducción del gasto público y, en general, suponer una mejora significativa de la vida de los ciudadanos.

De forma adicional, tampoco se debe olvidar que **la generación de esta información igualmente permite, como valor agregado, el desarrollo de otras tantas oportunidades de negocio e iniciativas empresariales a través de sistemas de Reutilización de la Información** asociada al Sector Público[5] (RISP)[6]

4. Resulta de interés conocer las publicaciones y el trabajo de estandarización impulsado por la UIT (Unión Internacional de las Telecomunicaciones): http://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx

5. La actual normativa española de Reutilización de la Información del Sector Público que incorpora las previsiones de la Directiva sobre este tema del año 2003 puede consultarse a través de esta URL: https://administracionelectronica.gob.es/pae_Home/pae_Estrategias/pae_Gobierno_Abierto_Inicio/pae_Reutilizacion_de_la_informacion_en_el_sector_publico.html#.WJdg5tLJzIU. Por otra parte, recientemente, se ha publicado un informe europeo sobre el estado actual de la transposición por los distintos Estados miembros de la Directiva de Reutilización de Información por el Sector Público: https://ec.europa.eu/digital-single-market/en/news/transposition-psi-directive-state-play-and-discussion-charging-criteria

6. Es interesante conocer las recomendaciones de la AEPD contenidas en el documento titulado "Orientaciones sobre Protección de Datos para la Reutilización de la Información del Sector público": http://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/2016/Orientaciones_proteccion_datos_Reutilizacion.pdf

y privado[7]. Hablamos de las posibilidades relatadas en relación al llamado "sector infomediario"[8]. En esta línea, el preámbulo de la Ley 18/2015, de 9 de julio, por la que se modifica la Ley española 37/2007, de 16 de noviembre, sobre Reutilización de la Información del Sector Público (en lo que sigue, Ley 37/2007) reza lo que sigue:

> "[...] En la sociedad, [...] se ha producido una **creciente concienciación del valor de la información pública, y como consecuencia, ha aumentado el interés por la reutilización con fines comerciales** y no comerciales [...]. [...] la Ley incorpora la obligación prevista en la Directiva de **fomentar el uso de licencias abiertas**, de tal forma que **las licencias para la reutilización de la información del sector público planteen las mínimas restricciones posibles** [...]"[9].

Por este motivo, y siendo plenamente consciente de tales potencialidades[10], así como de que la economía de los datos exige un elevado nivel de confianza, es por lo que la Comisión Europea adoptó el pasado 10 de enero de 2017 una estratégica **comunicación sobre la construcción de la "Economía Europea de Datos"** o **"Mercado Único de los Macrodatos"**[11], sobre la base de las conclusiones contenidas en su comunicación previa del año 2014 titulada "Hacia una economía de los datos próspera[12]".

Asimismo, esta institución europea ha lanzado en paralelo una consulta pública sobre el tema que permanecerá abierta hasta el 26 de abril de 2017 para conocer la opinión y entablar un diálogo abierto con todos los agentes socioecónomicos y partes interesadas sobre el particular[13].

7. Se recomienda consultar los siguientes enlaces web: http://datos.gob.es/ y http://datos.gob.es/es/noticia/la-accesibilidad-de-la-informacion-en-poder-de-las-administraciones-publicas-conforme-los
8. Puede accederse a más información sobre este sector a través del enlace siguiente: http://www.red.es/redes/sala-de-prensa/noticia/la-reutilizacion-de-datos-publicos-genera-un-volumen-de-negocio-proximo-los-5
9. Las referencias en negrita las realiza la autora.
10. Se estima que la economía de los datos supuso en la Unión 272.000 millones de euros en 2015 (crecimiento anual del 5,6 %) y que podría dar empleo a 7,4 millones de personas para 2020. Remítase al siguiente artículo: http://europa.eu/rapid/press-release_IP-17-5_es.htm
11. El término "macrodatos" se refiere a una gran cantidad de diferentes tipos de datos producidos a alta velocidad a partir de un gran número de diversos tipos de fuentes. Para manejar los conjuntos de datos muy variables y en tiempo real de hoy en día, se necesitan nuevas herramientas y métodos, como *software*, algoritmos y procesadores de gran potencia (véase la pág. 5 de la COMUNICACIÓN DE LA COMISIÓN AL PARLAMENTO EUROPEO, AL CONSEJO, AL COMITÉ ECONÓMICO Y SOCIAL EUROPEO Y AL COMITÉ DE LAS REGIONES. COM (2014) 442 final: http://ec.europa.eu/transparency/regdoc/rep/1/2014/ES/1-2014-442-ES-F1-1.Pdf).
12. COMUNICACIÓN DE LA COMISIÓN AL PARLAMENTO EUROPEO, AL CONSEJO, AL COMITÉ ECONÓMICO Y SOCIAL EUROPEO Y AL COMITÉ DE LAS REGIONES. COM (2014) 442 final.
13. Acceso al contenido de esta consulta pública: https://ec.europa.eu/digital-single-market/en/news/public-consultation-building-european-data-economy

En ambas comunicaciones **los aspectos normativos resultan clave**. Entre otros, destacan los relativos a la **privacidad y a la protección de los datos personales**, la anonimización y la seudonimización de los datos, la **seguridad/ciberseguridad de la información**, la **estandarización**, la **interoperabilidad**, la **portabilidad**, el **acceso y la transferencia de la información**[14], los aspectos relacionados con los **consumidores** y la mercadotecnia, la **responsabilidad**, etc.).

Ello explica la necesidad esbozada por la propia Comisión de dotarse de nuevos enfoques legales que favorezcan la eclosión del **"mercado europeo de los datos"** —íntimamente relacionado con el mercado único digita—, al tiempo que permitan combatir firmemente frente a las restricciones injustificadas a la libre circulación de datos a través de las fronteras de los diversos Estados miembros. O dicho de otro modo, tal y como ha declarado la actual comisaria responsable de Mercado Interior, Industria, Emprendimiento y Pymes de la UE **"**[...] **en lugar de construir fronteras digitales debemos centrarnos en construir una economía de los datos europea plenamente integrada y competitiva dentro de la economía de los datos mundial** [...]**"**[15].

Asimismo, deben considerarse los **aspectos éticos** concurrentes en lo que concierne, por ejemplo, a la transparencia de la información generada como consecuencia de procesos *big data*, así como al uso y aplicación concreta de la misma. Y es que los posibles desequilibrios o desviaciones informativas de partida que pudieran concurrir podrían conllevar, a su vez, la toma de decisiones poco o nada adecuadas, éticas, equitativas y/o justas. Estas decisiones, sin lugar a dudas, pueden igualmente incidir en el devenir de una determinada sociedad, de ciertos individuos, o de grupos concretos en los que estos se integren. Es evidente que existe un claro riesgo de que tales decisiones o posicionamientos finales puedan suponer el trazado de perfiles o de categorías o grupos definidos de ciudadanos/individuos y ello pueda afectarles jurídicamente o a cualquier otro nivel.

Por último, se destaca que desde finales del año 2011 se ha venido reforzando la estrategia para **"hacer fructificar los datos de las Administraciones públicas"**, en particular, a través de sendas iniciativas, ya tengan carácter supranacional o nacional, de datos abiertos[16]. Como manifestó la propia Comisión Europea, las

14. La amplia utilización de datos no personales generados por máquinas puede propiciar grandes innovaciones, empresas de nueva creación y nuevos modelos de negocio nacidos en la UE.
15. Puede accederse al contenido de esta declaración desde este enlace web: http://europa.eu/rapid/press-release_IP-17-5_es.htm
16. Es interesante consultar esta página web: https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

Administraciones públicas europeas tienen una **"auténtica mina de oro de potencial económico sin aprovechar"**, aludiendo a los grandes volúmenes de información recogidos por los numerosos poderes y entidades públicas[17].

¿Qué papel ostenta el sector público en relación al sector del *big data*?

Las Administraciones públicas y, en general, el sector público adquiere un **especial protagonismo en el sector del *big data*** por tres motivos principales, a saber:

a) Por la posibilidad de que estas entidades "autoexaminen" los grandes volúmenes de información vinculados al legítimo desarrollo de sus funciones y competencias, fundamentalmente, de forma anonimizada y/o disociada, extrayendo valiosas conclusiones que podrían, a la postre, ayudar a reducir de forma significativa el gasto público y a hacerlo más eficiente, así como a planificar y elaborar políticas, normas o planes más ajustados a las necesidades o intereses generales. Entre otras iniciativas recientes, por ejemplo, destaca una propuesta del Gobierno de España para tratar la información de los turistas que visitan nuestro país a partir de los datos asociados al *roaming* con apoyo de las operadoras (*big data* turístico)[18]. Desde la perspectiva sanitaria, el proyecto británico Care.Data[19] representa un auténtico desafío a la gestión tradicional de la información sanitaria en los sistemas públicos de protección, y que ya comienzan a plantearse igualmente en el marco de sistema sanitario español, al menos, en algunas comunidades autónomas[20]. Sin duda, las posibilidades de reutilización por terceros de esta información también plantean retos particulares desde la óptica de la privacidad y la seguridad de los datos personales implicados en este tipo de iniciativas.

b) Por la posibilidad legal que asimismo ostenta el sector público de recabar y analizar más información que cualquier otro sector, no solo porque podrá

17. Remítase, a modo de ejemplo, a la siguiente URL para ampliar información sobre estas materias: https://ec.europa.eu/digital-single-market/pillar-i-digital-single-market/action-107-proposals-strengthen-data-industry-europe.

18. De esta noticia se han hecho eco diversos medios de comunicación social, entre otros: http://www.eleconomista.es/tecnologia/noticias/8102539/01/17/El-Big-Data-del-turismo-Gobierno-y-operadoras-recabaran-datos-de-turistas-a-traves-del-roaming.html y http://cincodias.com/cincodias/2017/01/23/empresas/1485197572_288367.html

19. Se trata de un controvertido programa de *big data* sanitario sobre el que se puede recabar más información a partir de los siguientes enlaces: https://www.england.nhs.uk/ourwork/tsd/care-data/ y https://ico.org.uk/for-the-public/health/

20. Un ejemplo reciente es el de la Generalitat catalana: http://www.eldiario.es/catalunya/sanitat/nuevo-Big-Data-sanitario-catalan_0_607189931.html

tratar y analizar de forma inteligente aquella información que maneje de forma directa, en ejercicio de competencias y fines públicos, sino también porque podría, si así lo planteara, acceder a información en manos del sector privado, bajo determinadas condiciones, para el cumplimiento de fines estadísticos y de análisis. Así por ejemplo lo permite y promueve la Disposición Adicional Cuarta de la Ley española 56/2007, de 28 de diciembre, de Medidas de Impulso de la Sociedad de la Información (en adelante, Ley 56/2007)[21].

21. Según se establece en la Disposición Adicional Cuarta de la Ley 57/2006 titulada "Requerimientos de información para fines estadísticos y de análisis":

"[...] 1. La Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información, y los órganos estadísticos de las Comunidades Autónomas con competencias en materia de estadística, podrán requerir de los fabricantes de productos y proveedores de servicios referentes a las Tecnologías de la Información, a la Sociedad de la Información, a los contenidos digitales y al entretenimiento digital la información necesaria para el ejercicio de sus funciones para fines estadísticos y de análisis. La Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información podrá dictar circulares que deberán ser publicadas en el Boletín Oficial del Estado, en las cuales se expondrá de forma detallada y concreta el contenido de la información que se vaya a solicitar, especificando de manera justificada la función para cuyo desarrollo es precisa tal información y el uso que pretende hacerse de la misma. No obstante lo señalado en el párrafo precedente, el Ministerio de Industria, Turismo y Comercio podrá en todo caso realizar requerimientos de información particularizados sin necesidad de que previamente se dicte una circular de carácter general. La Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información podrá realizar las inspecciones que considere necesarias con el fin de confirmar la veracidad de la información que en cumplimiento de los citados requerimientos le sea aportada. Los datos e informaciones obtenidos por la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información en el desempeño de sus funciones, que tengan carácter confidencial por tratarse de materias protegidas por el secreto comercial, industrial o estadístico, solo podrán ser cedidos a la Administración General del Estado y a las Comunidades Autónomas en el ámbito de sus competencias. El personal de dichas Administraciones públicas que tenga conocimiento de estos datos estará obligado a mantener el debido secreto y sigilo respecto de los mismos. Las entidades que deben suministrar esos datos e informaciones podrán indicar, de forma justificada, qué parte de los mismos consideran de trascendencia comercial o industrial, cuya difusión podría perjudicarles, a los efectos de que sea declarada su confidencialidad respecto de cualesquiera personas o entidades que no sean la propia Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información, la Administración General del Estado o las Comunidades Autónomas, previa la oportuna justificación. La Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información decidirá, de forma motivada, sobre la información que, según la legislación vigente, esté exceptuada del secreto comercial o industrial y sobre la amparada por la confidencialidad.

2. Son infracciones de la obligación de cumplir los requerimientos de información establecida en el apartado anterior las conductas que se tipifican en los apartados siguientes. Las infracciones establecidas en la presente disposición adicional se entenderán sin perjuicio de las responsabilidades civiles, penales o de otro orden en que puedan incurrir los titulares de las entidades que desarrollan las actividades a que se refieren.
3. Las infracciones administrativas tipificadas en los apartados siguientes se clasifican en muy graves, graves y leves.
4. Son infracciones muy graves:
   a) La negativa reiterada a facilitar a la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información la información que se reclame de acuerdo con lo previsto en la presente Ley.

En aplicación de esta norma, la actual Secretaría de Estado para la Sociedad de la Información y la Agenda Digital, así como los órganos estadísticos de las comunidades autónomas con competencias en materia de estadística podrán realizar requerimientos específicos de información a los fabricantes de productos y proveedores de servicios referentes a las tecnologías de la información, a la sociedad de la información, a los contenidos digitales y al entretenimiento digital, solicitando la información necesaria para el ejercicio de sus funciones desde la perspectiva estadística y de análisis en estas materias, y que pueden tener carácter general (sector específico) o particularizados

b) Facilitar intencionadamente a la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información datos falsos.

5. Son infracciones graves:
   La negativa expresa a facilitar a la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información la información que se reclame de acuerdo con lo previsto en la presente Ley.
6. Son infracciones leves:
   No facilitar a la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información los datos requeridos o retrasar injustificadamente su aportación cuando resulte exigible.
7. Por la comisión de las infracciones señaladas en los apartados anteriores, se impondrán las siguientes sanciones:
   a) Por la comisión de infracciones muy graves tipificadas en el apartado 4, multa desde 25.000 euros hasta 50.000 euros.
   b) Por la comisión de infracciones graves tipificadas en el apartado 5, multa desde 5.000 euros hasta 25.000 euros.
   c) Por la comisión de infracciones leves tipificadas en el apartado 6, multa de hasta 5.000 euros.
   En todo caso, la cuantía de la sanción que se imponga, dentro de los límites indicados, se graduará teniendo en cuenta, además de lo previsto en el artículo 131.3 de la Ley 30/1992, de 26 de noviembre, de Régimen Jurídico de las Administraciones públicas y del Procedimiento Administrativo Común, lo siguiente:
   a) La gravedad de las infracciones cometidas anteriormente por el sujeto al que se sanciona.
   b) La repercusión social de las infracciones.
   c) El beneficio que haya reportado al infractor el hecho objeto de la infracción.
   d) El daño causado.
   Las sanciones impuestas por infracciones muy graves podrán ser publicadas en el "Boletín Oficial del Estado" una vez que la resolución sancionadora tenga carácter firme.
8. La competencia para la imposición de las sanciones muy graves corresponderá al ministro de Industria, Turismo y Comercio y la imposición de sanciones graves y leves al secretario de Estado de Telecomunicaciones y para la Sociedad de la Información.
   El ejercicio de la potestad sancionadora se sujetará al procedimiento aplicable, con carácter general, a la actuación de las Administraciones públicas.
9. Las estadísticas públicas que elabore la Secretaría de Estado de Telecomunicaciones y para la Sociedad de la Información relativas a personas físicas ofrecerán sus datos desagregados por sexo, considerando, si ello resultase conveniente, otras variables relacionadas con el sexo para facilitar la evaluación del impacto de género y la mejora en la efectividad del principio de igualdad entre mujeres y hombres.
10. En caso de que la información recabada en ejercicio de las funciones establecidas en esta disposición adicional contuviera datos de carácter personal será de aplicación lo dispuesto en la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal y en su normativa de desarrollo. [...]".

(empresa/entidad particular). Es más, incluso se prevé la posibilidad de que se puedan realizar aquellas inspecciones que se consideren necesarias con el fin de confirmar la veracidad de la información que fuera aportada y, en su caso, la posible imposición de multas económicas de hasta 50.000 euros en caso de no atender los citados requerimientos de información. Eso sí, se establecen específicas limitaciones en relación al acceso a datos o materias protegidas como consecuencia del secreto comercial, industrial o estadístico. De igual forma, en el caso de que la información recabada en ejercicio de tales funciones contuviera datos de carácter personal, sería de aplicación lo dispuesto en la vigente normativa protectora de datos personales.

Curiosamente, esta posibilidad de acceso a la información en manos de privados en atención a intereses públicos se erige como una de las grandes propuestas legislativas que plantea la Comisión Europea en su última comunicación relativa al mercado europeo de los datos y en el Documento de Trabajo que se acompaña a la misma[22].

c) Por la posible reutilización por terceros de la información del sector público, con el objetivo de propiciar la generación de nuevos modelos de negocio, la creación de empresas y de empleo. Todo ello en coherencia con la normativa de reutilización que resulte de aplicación.

Tarifas, precios y tasas públicas asociados a la reutilización de la información derivada de procesos *big data*

Como ya se ha comentado, el impulso de procesos *big data* por entidades públicas propicia la existencia de nueva información pública que es susceptible de ser reutilizada por terceros conforme dispone la legislación aplicable y, además de las ventajas que ello comporta en sí mismo, poder generar recursos e ingresos públicos adicionales que pueden revertir de vuelta en beneficio de la sociedad derivado del posible establecimiento de tarifas, tasas o precios públicos[23] por reutilización. Esta posibilidad ya se contempla en el artículo 7 de la Ley 37/2007, pudiendo resultar de obligada aplicación, por ejemplo, cuando se trata de organismos del sector público a los que se exija generar ingresos para cubrir una parte sustancial de sus costes relativos a la realización de sus misiones de servicio público.

En estos casos, los organismos deberán calcular el precio total conforme a criterios objetivos, transparentes y comprobables, que serán fijados mediante la

22. Véase el Documento titulado "COMMISSION STAFF WORKING DOCUMENT on the free flow of data and emerging issues of the European data economy Accompanying the document Communication Building a European data economy (COM (2017) 9 final)". Pág.32 y siguientes.
23. En estos casos, será aplicable el régimen jurídico correspondiente en estos casos, en particular, la Ley 8/1989, de 13 de abril, de Tasas y Precios Públicos, y demás normativa tributaria concordante.

normativa que corresponda y podrán aplicarse tarifas diferenciadas según se trate de reutilización con fines comerciales o no comerciales. Además, los ingresos totales de estos organismos obtenidos por suministrar documentos y autorizar su reutilización durante el ejercicio contable apropiado no deberán superar el coste de recogida, producción, reproducción y difusión, incrementado por un margen de beneficio razonable de la inversión.

Las Administraciones y organismos del sector público, en estos casos, vienen obligados a publicar por medios electrónicos —preferentemente a través de su respectiva sede electrónica—, y siempre que sea posible y apropiado, las tarifas fijadas para la reutilización de documentos que estén en su poder, así como las condiciones aplicables y el importe real de los mismos, incluida la base de cálculo utilizada.

Renovarse o morir: los nuevos escenarios exigen nuevas soluciones jurídicas

Protección de datos personales y privacidad en la era del *big data*

No toda la información que circula por la Red tiene carácter personal, es decir, afecta o se refiere a personas físicas identificadas o identificables en el sentido jurídico del término[24]. Ahora bien, **cuando esta tiene carácter personal, sin duda, resulta de aplicación la normativa protectora de datos personales** lo que implica ciertas restricciones en el tratamiento de aquella por afectar al ejercicio de derechos fundamentales (derecho a la protección de los datos personales[25]).

Sin embargo, **ello no supone que exista una limitación absoluta al despliegue de las tecnologías *big data* puesto que, como cualquier otro derecho humano, el mismo no tiene carácter absoluto** cediendo ante otros derechos e intereses igualmente protegibles como son, por ejemplo, los de la libertad de expresión e información, el de igualdad y no discriminación, entre otros.

24. Vid. artículo 4.1 del REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO de 27 de abril de 2016 (RGPD): https://www.boe.es/doue/2016/119/L00001-00088.pdf
25. Recogido como tal en el artículo 8 de la Carta Europea de Derechos Fundamentales: http://www.europarl.europa.eu/charter/pdf/text_es.pdf

Del mismo modo, la privacidad, sobre todo, en lo concierne a las comunicaciones electrónicas, tampoco puede suponer un freno a la innovación, la competencia[26] y al crecimiento económico, erigiéndose el principio a la libre circulación de los datos en uno de los principios rectores incluidos tanto en el nuevo Reglamento General europeo de Protección de Datos (RGPD)[27], como en la Directiva 2002/58/CE del Parlamento Europeo y del Consejo de 12 de julio de 2002 relativa al tratamiento de los datos personales y a la protección de la intimidad en el sector de las comunicaciones electrónicas (Directiva sobre la Privacidad y las Comunicaciones Electrónicas)[28] y en su reciente propuesta de revisión[29].

Por lo tanto, la clave está en **balancear innovación, competencia y protección de los derechos de las personas**. En este ámbito, resulta muy interesante bucear entre la doctrina y textos emanados tanto del Supervisor Europeo de Protección de Datos (EDPS), como del Grupo de Trabajo del Artículo 29 (GT29)[30] en torno a la construcción doctrinal del concepto del *big data protection*[31].

En tal sentido, al hilo de tales documentos y, en particular, de la Opinión EDPS 8/2016, sobre una aplicación coherente de los derechos fundamentales en la era del *big data*[32], resulta de interés esbozar las siguientes premisas, a saber:

26. Es muy interesante en este punto consultar este texto de la OECD: http://www.oecd.org/daf/competition/big-data-bringing-competition-policy-to-the-digital-era.htm
27. Documento normativo accesible desde la siguiente URL: https://www.boe.es/doue/2016/119/L00001-00088.pdf. Así conforme establece su Considerando 6: "[...] La rápida evolución tecnológica y la globalización han planteado nuevos retos para la protección de los datos personales. La magnitud de la recogida y del intercambio de datos personales ha aumentado de manera significativa. La tecnología permite que tanto las empresas privadas como las autoridades públicas utilicen datos personales en una escala sin precedentes a la hora de realizar sus actividades. Las personas físicas difunden un volumen cada vez mayor de información personal a escala mundial. La tecnología ha transformado tanto la economía como la vida social, y ha de facilitar aún más la libre circulación de datos personales dentro de la Unión y la transferencia a terceros países y organizaciones internacionales, garantizando al mismo tiempo un elevado nivel de protección de los datos personales [...]"..
28. Documento accesible desde la url: https://www.boe.es/doue/2002/201/L00037-00047.pdf
29. Texto accesible desde el siguiente enlace web: https://ec.europa.eu/digital-single-market/en/news/proposal-regulation-privacy-and-electronic-communications
30. Que pueden consultarse a través del siguiente enlace web: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/index_en.htm y http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp221_en.pdf
31. Resulta de interés consultar esta URL: https://secure.edps.europa.eu/EDPSWEB/edps/cache/offonce/Consultation/big_data
32. Documento accesible desde este enlace web: https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/EDPS/Events/16-09-23_BigData_opinion_EN.pdf. El mismo parte de una opinión preliminar, también del propio Supervisor Europeo de Protección de datos del año 2014: https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2014/14-03-26_competitition_law_big_data_EN.pdf

- **La protección de los datos personales no supone un obstáculo para el desarrollo de proyectos *big data*, pero debe tenerse muy en cuenta[33]:** es necesario contextualizar y relativizar la protección de los datos personales y la privacidad por relación a la protección de otros derechos igualmente protegibles para evitar desequilibrios y la posible conculcación de los mismos.
- **El tratamiento de los datos personales objeto de proyectos *big data* debe fundamentarse en sólidas bases legítimas:** la privacidad es un parámetro que debe formar parte de la calidad de los servicios y productos ofrecidos a terceros, lo que se refuerza a la vista del nuevo principio de privacidad desde el diseño contenido en el RGPD. Por lo tanto, cuando se ofrecen servicios "gratis" a terceros a cambio de sus datos, como una fórmula de rápido acceso al combustible que propulsa el *big data*, se debería considerar el anterior principio, por aplicación de la normativa actual, así como justificar su legítimo tratamiento (consentimiento del interesado, ley aplicable, interés legítimo, etc.).
- **La evaluación del impacto en la protección de los datos personales en proyectos *big data*:** el artículo 35 del RGPD plantea que cuando sea probable que un tipo de tratamiento, en particular si utiliza nuevas tecnologías que por su naturaleza, alcance, contexto o fines entrañe un alto riesgo para los derechos y libertades de las personas físicas, el responsable del tratamiento (ya sea público o privado) debe realizar, antes del tratamiento proyectado, una evaluación del impacto de las operaciones de tratamiento en la protección de datos personales (EIPD). Ahora bien, matiza que sí será necesario realizar esta EIPD en los casos de:

  a) evaluación sistemática y exhaustiva de aspectos personales de personas físicas que se base en un tratamiento automatizado, como la elaboración de perfiles, y sobre cuya base se tomen decisiones que produzcan efectos jurídicos para las personas físicas o que les afecten significativamente de modo similar;
  b) tratamiento a gran escala de las categorías especiales de datos a que se refiere el artículo 9, apartado 1, o de los datos personales relativos a condenas e infracciones penales a que se refiere el artículo 10, u;
  c) observación sistemática a gran escala de una zona de acceso público.

  Del mismo modo, la autoridad de control que corresponda podrá prever y publicar una lista de los tipos de operaciones de tratamiento que requieran una EIPD.

33. A estos efectos puede resultar interesante consultar la nueva *Guía de Buenas Prácticas en Protección de Datos para Proyectos Big Data*, editada por la Agencia Española de Protección de Datos (AEPD) e Isms Forum Spain, del que la autora es co-redactora: https://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/2017/Guia_Big_Data_AEPD-ISMS_Forum.pdf

A estos efectos, la aplicación de tecnologías *big data* aconseja la realización de este tipo de evaluaciones de riesgo e impacto previas al tratamiento de datos que se pretenda. Un buen punto de partida desde la perspectiva práctica es considerar las orientaciones y pautas claras propuestas por la Guía de Buenas Prácticas para la realización de EIPD publicada por la Agencia Española de Protección de Datos (AEPD)[34].

- **No basta con aplicar la ley, se debe adoptar una actitud y política responsable:** si bien la normativa aplicable, sobre todo a nivel europeo[35], ha propiciado una garantía uniforme de los derechos a la protección de datos y a la privacidad en toda la Unión Europea (UE), en consonancia con la actual revisión de la normativa europea de consumo[36], resulta igualmente necesario que los agentes de tratamiento de la información adopten una especial actitud responsable y diligente en este ámbito en cumplimiento del llamado "principio de accountability". Eso se traduce en la necesidad de que demuestren que han acometido los pasos y actuaciones concretas para el adecuado cumplimiento normativo en estos casos (*compliance*). En este ámbito, por ejemplo, la adopción y/o la adhesión por los mismos a específicos Códigos de Buenas Prácticas o de Conducta en Protección de Datos dirigidos a proyectos de *big data*[37] es una buena opción desde la perspectiva del *compliance* y de la minimización de los riesgos existentes y de la posible responsabilidad legal que pudiera concurrir según los casos[38].
- **Reutilización de datos personales disociados:** cuando se trata de proyectos de *big data* que conciernan a datos de carácter personal, desde la perspectiva

34. Publicación accesible desde esta URL: http://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/Guia_EIPD.pdf
35. Destacan en este ámbito, el Reglamento europeo General de Protección de Datos y la Directiva ePrivacy: https://ec.europa.eu/digital-single-market/en/online-privacy, http://ec.europa.eu/justice/data-protection/reform/index_en.htm y https://ec.europa.eu/digital-single-market/en/proposal-eprivacy-regulation
36. Para mayor información, puede consultar estos enlaces dispuestos a continuación: http://ec.europa.eu/consumers/consumer_rights/index_en.htm y http://ec.europa.eu/consumers/consumer_rights/rights-contracts/directive/index_en.htm
37. Recientemente, la Asociación Española de Fomento de la Seguridad de la Información (ISMS Forum Spain) https://www.ismsforum.es/iniciativas/index.php?idcategoria=4 ha presentado en el IX Foro de la Privacidad que organiza, el primer Código de Buenas Prácticas en materia de Protección de Datos para Proyectos de Big Data. Un documento orientador elaborado por diversos expertos y académicos, en estrecha colaboración con la AEPD y que se pretende se convierta en el primer código de conducta español aprobado por esta autoridad de control: http://www.expansion.com/juridico/actualidad-tendencias/2017/02/01/58922b5f46163f63308b4597.html
38. La adhesión a Códigos de Conducta en los términos descritos por el RGPD puede considerarse una circunstancia atenuante desde la perspectiva de las infracciones y sanciones aplicables en materia de protección de datos personales (véase el Considerando 148 del RGPD): https://www.boe.es/doue/2016/119/L00001-00088.pdf

de su posible reutilización, se deben aplicar las técnicas adecuadas a fin de garantizar la irreversibilidad de los procesos de anonimización y la correcta disociación de la información.

Teniendo en cuenta la importancia y, a veces, la dificultad de aplicar de forma correcta las actuales técnicas de anonimización de datos personales es por lo que la AEPD ha publicado recientemente orientaciones sobre cómo desarrollar con ciertas garantías legales y técnicas tales procesos[39], las cuales están a su vez en consonancia con los trabajos previos del GT29, en concreto, con lo dispuesto por el Dictamen 5/2014 sobre las técnicas de anonimización[40] de datos personales[41].

Lo anterior no es baladí puesto que, según se desprende de la normativa aplicable, los principios de protección de datos no deben aplicarse a la información anónima, es decir, a la información que no guarda relación con una persona física identificada o identificable, ni a los datos convertidos en anónimos de forma que el interesado no sea identificable o deje de serlo[42].

El principio de la libre circulación de los datos como criterio interpretativo básico

Con carácter adicional a lo indicado, se debe partir del hecho de que el "principio de la libre circulación de los datos" en el seno de la Unión es uno de los principales pilares sobre el que se sostiene el Mercado Único Digital (Digital Single

39. El documento digital relativo a tales orientaciones es directamente accesible desde este enlace web que se indica a continuación: http://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/2016/Orientaciones_y_garantias_Anonimizacion.pdf

40. El Grupo de Trabajo reconoce el valor potencial de la anonimización, en particular como estrategia para permitir a las personas y la sociedad en su conjunto beneficiarse de los "datos abiertos" al mismo tiempo que se mitigan los riesgos para los interesados, apuntando a la aleatorización y la generalización como las principales técnicas de anonimización sin que estas tengan carácter exclusivo mu excluyente (la norma no impone ninguna técnica concreta en este ámbito). En todo caso, la idea subyacente es que el resultado de la anonimización, entendida esta como una técnica aplicada a los datos personales, debe ser, de acuerdo con el actual estado de la tecnología, tan permanente como el borrado. En otras palabras: debe garantizarse que es imposible tratar los datos personales o su posible reversibilidad a estos efectos. Y es que la anonimización puede definirse como el resultado de un tratamiento de datos personales realizado para impedir de forma irreversible la identificación del interesado .De forma adicional, también apunta a que sabiendo que la anonimización y la reidentificación son campos de investigación activos en los que se publican con regularidad nuevos descubrimientos, es por lo que recomienda que los procesos de anonimización no se contemplen como un procedimiento esporádico, animando a los responsables a evaluar regularmente los riesgos de reidentificación existentes en este ámbito.

41. Este Dictamen puede consultarse a través de esta URL: http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_es.pdf

42. Así de claro se prevé en el Considerando 26 del RGPD.

Market (DSM)), por lo que se deberían remover los obstáculos que impiden la misma, en especial, los de tipo legal, y es que resulta habitual que las divergencias en el nivel de protección de los derechos y libertades de las personas físicas, en particular respecto al derecho a la protección de los datos de carácter personal, o las medidas e iniciativas públicas tendentes a la localización territorial y/o geográfica de cierta información, constriñen o pueden impedir injustificadamente la eclosión del nuevo mercado de datos europeo no permitiendo el libre flujo informativo a nivel europeo[43].

Este principio es aplicable tanto si la información tiene carácter personal como si no lo tiene, de hecho, el RGPD en lo relativo a los datos de carácter personal se refiere al mismo tanto en su propio título[44], como en diferentes apartados de esta norma en los siguientes términos, a saber:

– "[...] Considerando 6: La rápida evolución tecnológica y la globalización han planteado nuevos retos para la protección de los datos personales. La magnitud de la recogida y del intercambio de datos personales ha aumentado de manera significativa. La tecnología permite que tanto las empresas privadas como las autoridades públicas utilicen datos personales en una escala sin precedentes a la hora de realizar sus actividades. Las personas físicas difunden un volumen cada vez mayor de información personal a escala mundial. La tecnología ha transformado tanto la economía como la vida social, y ha de facilitar aún más la libre circulación de datos personales dentro de la Unión y la transferencia a terceros países y organizaciones internacionales, garantizando al mismo tiempo un elevado nivel de protección de los datos personales. [...]"[45].

– "[...] Considerando 9: "[...] Las diferencias en el nivel de protección de los derechos y libertades de las personas físicas, en particular del derecho a la protección de los datos de carácter personal, en lo que respecta al tratamiento de dichos datos en los Estados miembros pueden impedir la libre circulación de los datos de carácter personal en la Unión. Estas diferencias pueden constituir, por lo tanto, un obstáculo al ejercicio de las actividades económicas a nivel de la Unión, falsear la competencia e impedir

43. Lo que ha determinado, entre otros motivos, la necesidad de dotar de una mayor uniformidad a los criterios legales aplicados por los distintos Estados miembros de la Unión a través de Reglamentos de directa aplicación como ocurre, por ejemplo, con el RGPD.
44. REGLAMENTO (UE) 2016/679 DEL PARLAMENTO EUROPEO Y DEL CONSEJO, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a ***la libre circulación de estos datos***.
45. El subrayado lo realiza la autora.

que las autoridades cumplan las funciones que les incumben en virtud del Derecho de la Unión. [...]"[46].

–"[...] Considerando 13. "[...] El buen funcionamiento del mercado interior exige que la libre circulación de los datos personales en la Unión no sea restringida ni prohibida por motivos relacionados con la protección de las personas físicas en lo que respecta al tratamiento de datos personales [...][47]".

"[...] Artículo 1.3 RGPD: "[...] La libre circulación de los datos personales en la Unión no podrá ser restringida ni prohibida por motivos relacionados con la protección de las personas físicas en lo que respecta al tratamiento de datos personales. [...][48]".

"[...] Artículo 51.1 RGPD: "[...] Cada Estado miembro establecerá que sea responsabilidad de una o varias autoridades públicas independientes (enadelante 'autoridad de control') supervisar la aplicación del presente Reglamento, con el fin de proteger los derechos y las libertades fundamentales de las personas físicas en lo que respecta al tratamiento y de facilitar la libre circulación de datos personales en la Unión. [...][49]".

El libre flujo de la información en la UE también requiere el reforzamiento del carácter interoperable de los sistemas de información, así como de las medidas de seguridad que resulten de aplicación, lo que ha provocado la reciente aprobación de importantes normas a tal fin como es la Directiva (UE) 2016/1148 del Parlamento Europeo y del Consejo, de 6 de julio de 2016 relativa a las medidas destinadas a garantizar un elevado nivel común de seguridad de las redes y sistemas de información en la Unión (Directiva NIS)[50].

Más allá de la privacidad: otras posibles limitaciones al libre acceso y transferencia de los datos

No solo la normativa protectora de datos personales puede plantear concretos límites al libre acceso y a la circulación de los datos, también la normativa protectora de los derechos de propiedad intelectual e industrial determina ciertas restricciones a tal acceso y transferencia.

46. El subrayado lo formula la autora.
47. El subrayado lo realiza la autora.
48. El subrayado lo realiza la autora.
49. El subrayado lo formula la autora.
50. Se puede acceder al texto íntegro de esta Directiva a través del enlace web dispuesto a continuación: https://www.boe.es/doue/2016/194/L00001-00030.pdf

De igual forma, garantizar los secretos comerciales y el *know-how* empresarial puede justificar limitaciones específicas al libre acceso y a la circulación de cierta información. Al respecto, así parece corroborarlo la reciente Directiva (UE) 2016/943 del Parlamento Europeo y del Consejo, de 8 de junio de 2016, relativa a la protección de los conocimientos técnicos y la información empresarial no divulgados (secretos comerciales) contra su obtención, utilización y revelación ilícitas[51].

Consideraciones jurídicas adicionales de interés

Sin perjuicio de lo apuntado con anterioridad, parece de interés abordar otras cuestiones que se están planteando por la Comisión Europea en estos momentos a través de las comunicaciones anteriormente señaladas, esto es:

El nuevo ''derecho del productor de los datos''

Se esboza el posible reconocimiento de un nuevo "derecho del productor de datos", es decir, el derecho a utilizar y autorizar el uso de datos no personales o anónimos por parte del propietario o usuario a largo plazo del dispositivo. Este enfoque plantea la posibilidad de que los usuarios utilicen sus datos y contribuyan así al desbloqueo de datos generados por máquinas en un ecosistema marcado por el "Internet de las cosas" (IoT)" y donde estas tendrán que comunicarse de forma inexorable entre las mismas. Todo ello al objeto de mejorar la negociabilidad de los datos no personales o anónimos de la máquina como un bien económico.

Dicho derecho podría calificarse jurídicamente, según indica la Comisión, como un derecho de propiedad que comprendería el derecho exclusivo a utilizar determinados datos y el derecho a conceder posibles licencias de uso. Alternativamente, en lugar de diseñar este derecho de productor de datos como un derecho real, se plantea concebirlo como un conjunto de derechos puramente defensivos.

Sin embargo, habría que clarificar de forma transparente las posibles excepciones aplicables al ejercicio de este derecho. A modo de ejemplo, los organismos del sector público podrían esgrimir un interés legítimo en obtener acceso a determinados datos como los estadísticos o de análisis. Además, en consonancia con la política de la Comisión en materia de ciencia abierta y acceso abierto, podría igualmente considerarse una excepción que garantice el acceso a datos privados pertinentes para los científicos que realizan investigaciones enteramente o mayoritariamente financiadas con recursos públicos.

51. Texto normativo accesible desde esta URL: https://www.boe.es/doue/2016/157/L00001-00018.pdf

La previsión de un nuevo régimen de licencias para los datos generados por máquina

Esta posibilidad se prevé respecto a aquellos datos que tengan el carácter de anónimos, generándose un escenario potencialmente basado en ciertos principios clave, tales como términos justos, razonables y no discriminatorios (FRAND), para que los titulares de datos, como fabricantes, proveedores de servicios u otras partes para facilitar el acceso a los datos que detenten contra una posible retribución previa la anonimización de los mismos. Podría preverse la consideración de diferentes regímenes de acceso para diferentes sectores y/o modelos de negocio a fin de tener en cuenta las especificidades de cada sector.

El lanzamiento de mecanismos de incentivación a las empresas y otras entidades para la compartición de datos

Todo ello a fin de favorecer el acceso y la compartición a la información que detenten con pleno respeto a los derechos que pudieran concurrir en cada caso.

La previsión de instrumentos regulatorios que favorezcan la portabilidad de los datos

En particular, en relación a aquellos que no tengan carácter personal, dado que los datos de carácter personal ya cuentan con un marco jurídico propio y diferenciado, al menos en Europa, a través del artículo 20 del RGPD. Se trataría de plantear un régimen jurídico de portabilidad similar al de los datos de carácter personal pero para datos que no tengan este carácter y en el que las cuestiones de estandarización e interoperabilidad de los sistemas y procesos adquieren una especial relevancia.

En todo caso, la Comisión prevé consultar a todas las partes y agentes interesados sobre estos temas a fin de recabar su opinión acerca del funcionamiento de los mercados de datos por sectores y, de esta forma, explorar las posibles y mejores alternativas.

› Conclusiones finales

Como se comprueba, cuando se habla de *big data*, son diversos los retos y desafíos legales que se plantean en la actualidad. Desde la perspectiva europea, la Comisión y el Parlamento Europeo son especialmente conscientes de la necesidad de analizar y remover los obstáculos jurídicos que frenen el pleno desarrollo de la nueva economía de datos ya que, hacer lo contrario, afectaría gravemente a la posición de la UE en el contexto de la economía mundial de los datos y, por

consiguiente, al crecimiento económico y social de esta región en comparación con otras, principalmente, con los EE. UU.

Por lo tanto, las instituciones de la Unión están abocadas a la promulgación de normas eficaces respecto a la protección de los datos personales y la seguridad de la información y de las redes que la canalizan fomentando, al mismo tiempo, el libre acceso y la circulación de los datos. En otras palabras, deben focalizar sus esfuerzos en la aprobación de normas que contemplen la debida protección de todos los derechos e intereses en juego y que, asimismo, incluyan mecanismos correctores ante las posibles desviaciones que pudieran producirse en este ámbito.

El *big data* para el sector público —al igual que para el privado— ofrece posibilidades inmensas de actuación y de aplicación práctica en ejercicio y en coherencia con las competencias públicas que concurran en cada caso. Ahora bien, es muy importante conocer la normativa aplicable, según el tipo de información de que se trate, e igualmente adoptar aquellos criterios y protocolos de cumplimiento que resulten pertinentes.

En orden a facilitar el cumplimiento de la normativa aplicable en este ámbito, resultaría de interés que el sector público pudiera promover Códigos de Conducta propios en materia de Protección de Datos para Proyectos Big Data que se pudieran ajustar a sus específicas particularidades jurídicas y que también atiendan a los procesos de reutilización de la información bajo su poder que deben igualmente promover de conformidad con la vigente legislación.

Referencias bibliográficas

Código de buenas prácticas en protección de datos para proyectos de Big Data", publicado por la AEPD e ISMS Forum Spain en 2017, del que Noemí Brito es coautora: https://www.agpd.es/portalwebAGPD/revista_prensa/revista_prensa/2017/notas_prensa/news/2017_05_11-ides-idphp.php y https://www.agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/Guias/2017/Guia_Big_Data_AEPD-ISMS_Forum.pdf

Commission Staff Working Document on the free flow of data and emerging issues of the European data economy Accompanying the document Communication Building a European data economy. COM (2017) 9 final). Consultado a febrero de 2017 <https: //ec.europa.eu/digital-single-market/en/news/staff-working-document-free-flow-data-and-emerging-issues-european-data-economy>

Comunicación de la Comisión al Parlamento Europeo, al Consejo, al Comité Económico y Social europeo y al Comité de las Regiones, titulada "Hacia una economía de los datos próspera". COM (2014) 442 final. Consultado a febrero de 2017 <http: //ec.europa.eu/transparency/regdoc/rep/1/2014/ ES/1-2014-442-ES-F1-1.Pdf>

Comunicación de la Comisión Europea al Parlamento Europeo, al Consejo, al Comité Económico y Social europeo y al Comité de las Regiones titulada "Building a European Data Economy». COM (2017) 9 final. Consultado a febrero de 2017 <https: //ec.europa.eu/digital-single-market/en/towards-thriving-data-driven-economy>

Dictamen 5/2014 sobre las técnicas de anonimización de datos personales. GT29. 2014. Consultado a febrero de 2017 <http: //ec.europa.eu/ just i ce/ data-protection/article-29/documentation/opinion-recommendation/ files/2014/wp216_es.pdf>

Gil, E. (2016). *Big Data, Privacidad y Protección de Datos. Accésit 2015. XIX Edición del Premio Protección de datos Personales de Investigación*. Ed. AEPD y Agencia Estatal BOE.

Informe Europeo sobre la Transposición de la Directiva de Reutilización de la Información del Sector Público (Transposition of the PSI Directive- state of play and discussion on the charging criteria) (2016). Consultado a febrero de 2017 <https: //ec.europa.eu/digital-single-market/en/news/ transposition-psi-directive-state-play-and-discussion-charging-criteria>

Informe ITU (2013). "Big Data: Big today, normal tomorrow". 2013. Consultado a febrero de 2017 <http: //www.itu.int/dms_pub/itu-t/ oth/23/01/ T23010000220001PDFE.pdf>

Informe OCDE (2016). "Big Data: Bringing Competition Policy to the Digital Era". Consultado a febrero de 2017 <https: //one.oecd.org/document/ DAF/ COMP(2016)14/en/pdf>

Martínez Martínez, R. (2014). "Ética y Privacidad de los Datos". *XXV Jornadas Técnicas de RedIRIS. Big data y Gobierno Abierto*. Ed. Fundación Ramón Areces. Consultado a febrero de 2017 <file: ///C: /Users/nbrito. LEGISTEL/Downloads/ jt2014-jt-sesion4b_big_data-a15b3c1.pdf> y <http: // www.fundacionareces. tv/watch/bigdata?as=53d296758d85927a508b46 dc> y <http: //www. fundacionareces.es/fundacionareces/cargarAplicacion AgendaEventos.do? verPrograma=1&identificador=1675>

Mayer-Schönberger, V. (2015). *Big Data. La revolución de los datos masivos.* Ed. Turner.

Orientaciones y Garantías en Procesos de Anonimizacion de Datos. AEPD. 2016. Consultado a febrero de 2017 <http: //www.agpd.es/portalwebAGPD/ canaldocumentacion/publicaciones/common/Guias/2016/Orientaciones_y_ garantias_Anonimizacion.pdf>

Orientaciones sobre Protección de datos para la Reutilización de la Información del Sector público. 2016. Consultado a febrero de 2017. http: //www. agpd.es/portalwebAGPD/canaldocumentacion/publicaciones/common/ Guias/2016/Orientaciones_proteccion_datos_Reutilizacion.pdf

Puyol, J. (2015). *Aproximación Jurídica y Económica al Big Data*. Ed. Tirant lo Blanch.

Rallo Lombarte, A. *et al.* (2015). *Hacia un Nuevo Derecho Europeo de Protección de Datos*. 1ª ed. Ed. Tirant lo Blanch.

Tercer Informe de caracterización del Sector Infomediario en España. ONTSI. Red.es. Entidad pública empresarial adscrita al Ministerio de Energía, Turismo y Agenda Digital (MINETAD). 2015. Consultado a febrero de 2017. http: //www.red.es/redes/sala-de-prensa/noticia/la-reutilizacion-de-datos-publicos-genera-un-volumen-de-negocio-proximo-los-5

Capítulo 6

Ciberseguridad y *big data*

Enrique Ávila*

El ciberespacio ya es la parte más importante de nuestras vidas. En las sociedades avanzadas del siglo XXI, la interacción entre el mundo físico y el mundo virtual es evidente. No obstante lo antedicho, la línea divisoria entre ambos dominios es muy brumosa para la mayoría de la población.

Hacemos uso de la tecnología y nos beneficiamos de la misma pero no tenemos un conocimiento profundo de qué es lo que, en realidad, estamos haciendo con ella. Nos solventa problemas, de hecho, el grado de evolución al que hemos llegado en las sociedades avanzadas no sería posible, en ningún modo, sin la existencia del ciberespacio. Todo ello lo sabemos, lo hemos leído en multitud de ocasiones desde diversas fuentes. Nos bombardean a diario con este incontrovertible hecho... Nos INFOXICAN...

La integración entre el mundo virtual es, cada minuto que pasa, cada algoritmo que es desarrollado y, sin ser nosotros conscientes, implantado en un sistema o servicio para que tome decisiones, en tiempo real, sobre nuestra existencia física, más estrecha.

El dominio de ciberespacio, con la eclosión de las nuevas tecnologías, pero no esas que pensamos sino las que no son tan evidentes en nuestro entorno. Se está desarrollando a una velocidad sin precedentes y generando condiciones para que el ser humano viva un momento disruptivo en su propia evolución como especie.

Muchos de los conflictos que sufrimos en la actualidad, si son analizados desde una perspectiva multidimensional, de forma epistemológica, vienen inducidos por la resistencia de grandes grupos sociales a integrarse en un mundo globalizado, complejo, dependiente de la tecnología y en el que en ningún caso es evidente qué grupos toman decisiones sobre el mismo o qué órganos de gobernanza y toma de

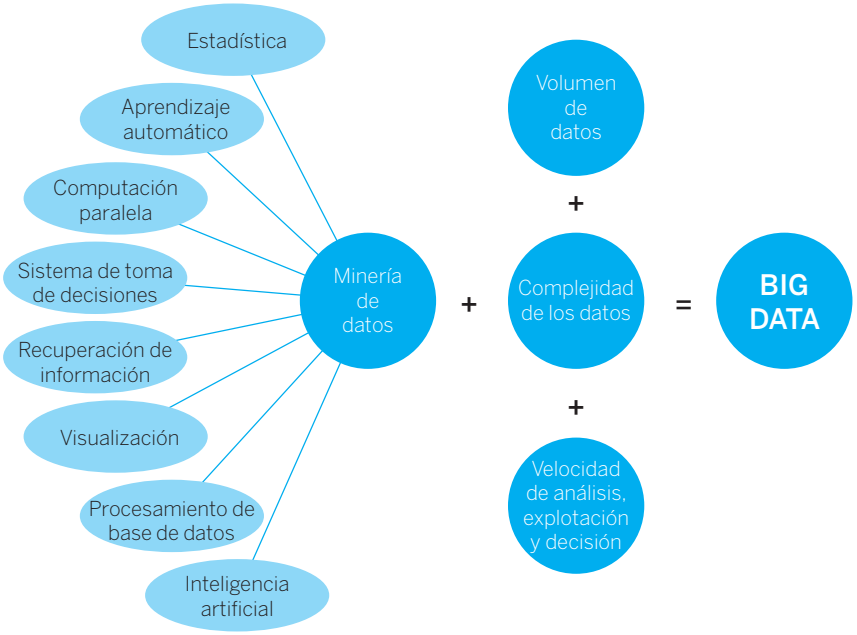
* Licenciado en Derecho y director del Centro Nacional de Excelencia en Ciberseguridad de España.

decisión se encuentran preparados o dominan el devenir de estas sociedades fuertemente tecnologizadas...

El reto al que nos enfrentamos como especie es, como podemos observar, inmenso.

El modelo conceptual que aún manejamos, en el mundo de la ciberseguridad, bajo nuestro punto de vista, es completamente erróneo. Aún estamos anclados al ordenador personal. Cuando hablamos de nuestro terminal de datos móvil aún nos referimos a él como nuestro teléfono. El uso de esta terminología, que no hace sino hacer referencia al modelo de conocimiento subyacente, implica que estamos cometiendo un grave error de apreciación al evaluar el riesgo. Nuestro perímetro clásico de seguridad, completamente estático, ha desaparecido como por ensalmo.

Pero es que el problema conceptual es aún mayor si nos detenemos a reflexionar en profundidad sobre este asunto. Aún personalizamos nuestros dispositivos interconectados. Los referenciamos a un determinado ser humano. Usamos expresiones tales como "nuestro ordenador", "nuestro teléfono móvil", "nuestro reloj inteligente"...



Fuente: INCIBE.

Siendo cierto que estos dispositivos tienen capacidad de cómputo y de almacenamiento y que se encuentran conectados a la Red, no son ya sino una infinitesimal minoría de los dispositivos conectados. Y en franco decrecimiento. Las nuevas "Redes SCADA" conformadas en el "Internet de las cosas" que, básicamente, lo serán TODO, se configuran como la principal fuente de información con la que hemos de enfrentarnos en los próximos años. Configuran ya un inmenso "BIG DATA" no comparable con el que explotábamos hace no más de cinco años. Generador de riesgos y amenazas y, cómo no, de enormes oportunidades de negocio si es que somos capaces de embridar tan ingente cantidad de información tanto desde el punto de vista técnico como desde el aún más importante, punto de vista legal ya que, reconozcámoslo, no somos capaces de hacer prospectiva de las consecuencias que tendrán para los ciudadanos el tratamiento masivo de la información obtenida de tan variopintas fuentes.

Los modelos productivos y de distribución de servicios y de mercancías están mutando. La introducción de la inteligencia artificial en todos los procesos involucrados abre una nueva dimensión en el uso de los datos y su explotación con fines diversos.

Tal y como expone Rifkin en su obra *La sociedad de coste marginal cero*, al reintroducir en la ecuación productiva el concepto de "procomún colaborativo", a partir del cual edifica un nuevo modelo social y económico basado en el incontrovertible hecho de que la mayoría de los productos y servicios terminarán considerándose *commodities* y el coste marginal de su producción y distribución tenderá a cero, minimizando los beneficios obtenidos y, con ello, rompiendo el modelo clásico de generación de valor proveniente de la primera y segunda revolución industrial, induce, decimos, con este nuevo modelo conceptual, la necesidad de modificar varios modelos que considerábamos como "ciertos" y no sujetos a discusión.

La denominada "transformación digital" de nuestras sociedades supone un verdadero cambio de paradigma en los modelos clásicos que solo ha sido posible gracias al uso masivo de la tecnología. Esta transformación supone poner de manifiesto la necesidad de generar nuevas conceptualizaciones que impactan de lleno en el plano de la ciberseguridad. Un término que, rápidamente, ha sido superado por el que, bajo mi punto de vista, es mucho más adecuado: el de ciberinteligencia.

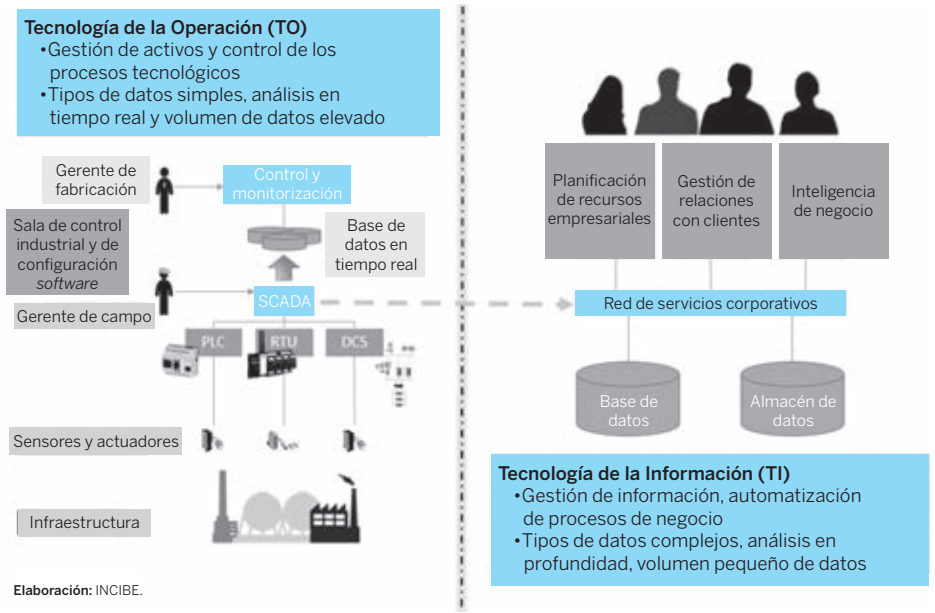
¿Por qué ciberinteligencia? Porque en el dominio del ciberespacio el concepto de seguridad clásico se desvanece en un océano de complejidad. La norma moral

del mundo físico. El bien y el mal, esos escurridizos conceptos, devienen, en el ciberespacio, en oportunas interpretaciones de una serie de hechos fuertemente mediatizados por cuestiones de plazo, resiliencia y continua revisión de objetivos. Y no es que en el mundo físico no deban aplicarse similares criterios. La necesidad, por ejemplo, de reevaluación de las políticas públicas se ha demostrado como un arma necesaria para evitar ineficacias (no confundamos un objetivo con un criterio, que sería el de eficiencia. Las políticas públicas han de ser siempre eficaces y pretender, en la medida de lo posible, la eficiencia, es decir, conseguir los objetivos prefijados al menor coste posible pero nunca renunciar a la consecución de los objetivos prefijados). El problema añadido, en el dominio del ciberespacio, son los tiempos de respuesta. Nuestro cerebro, caracterizado por sus capacidades de computación masivamente paralela a baja frecuencia, ha competido con ventaja, hasta el momento, con sistemas de alta frecuencia de procesamiento pero de baja paralelización. Es una ventaja que el ser humano va perdiendo poco a poco y que, además, en el ciberespacio, deviene, en muchas ocasiones, en irrelevante.

El ciberespacio necesita decisiones rápidas y, en la mayoría de los casos, automatizando la respuesta frente a determinadas condiciones. Así es el comportamiento de los algoritmos. Y son estos los que, en definitiva, gobiernan la mayoría de las decisiones del ciberespacio.

A partir de este hecho, desde nuestro punto de vista incontestable, aparece un nuevo horizonte de sucesos con el que, desde el punto de vista de la ciberseguridad (mejor ciberinteligencia), hemos de lidiar: ¿Cómo protegemos toda la cadena de producción y tratamiento de información de este dominio?

Generemos una imagen mental del sistema en su conjunto. Partamos desde la capa de perímetro más exterior: los sensores. No se limitan a su labor de sensorización sino que, en la mayoría de los casos, disponen de capacidad de cómputo autónoma y de capacidad de almacenamiento. ¿Cómo podemos proteger de forma eficaz los trillones de sensores que van a ser puestos en funcionamiento en los próximos meses o años? Cada uno de ellos con un sistema operativo y unas aplicaciones que han sido creados en función del beneficio económico inmediato y no como una pieza de un enorme puzle de seguridad en el que todas las piezas se encuentran, de algún modo, interconectadas en una red neuronal-fractal.



Elaboración: INCIBE.

¿Cómo haremos para parchear todos estos dispositivos de perímetro en caso de que se produzca un incidente de seguridad que les afecte? Y, volvemos a reiterar, no pensemos en estos dispositivos como asociados a una determinada persona. Se trata de inmensas redes SCADA que están ofreciendo una miríada de información a un gigantesco BIGDATA del que es posible extraer conocimiento e inteligencia ilimitados.

Relacionado con la anterior aseveración, creo que ya todos somos conscientes de que el registro de toda esta información, en nuestras sociedades avanzadas, supone un riesgo para estas mismas sociedades por la acumulación de poder y la fuerte tendencia hacia la conformación de oligopolios que adquieren más poder aún que la mayoría de los Estados actualmente existentes.

Subamos un peldaño en nuestro análisis. Hemos dejado atrás el nivel de sensorización y entramos en el mundo del análisis local de los datos obtenidos. Este análisis será la base sobre el que los algoritmos de gobierno de las infraestructuras críticas, que son las que nos proveen de servicios esenciales, tomarán las decisiones más cercanas a los usuarios de las mismas. La masiva ingesta de datos en un *big data* centralizado tiene un límite a su crecimiento. Por ello, no todos los datos son subidos a este sino que muchos de ellos son analizados en

anillos exteriores. Será este análisis el que, para ser consolidado en entidades de mayor nivel de decisión (también algorítmica), llegarán al *big data* central.

El escenario es, como podemos observar, complejo. El perímetro de seguridad expuesto, aterradoramente amplio. La mayoría del mismo, además, inaccesible por la inteligencia humana.

Si, citando a INCIBE, recordamos el concepto básico de las 4 uves del *big data*:

- Volumen: la cantidad de datos a analizar es elevada.
- Variedad: las fuentes de información no tienen conexión entre ellas y presentan los datos de forma desestructurada.
- Velocidad: el tiempo entre la recogida de la información, su procesado y la toma de decisiones debe ser mínimo.
- Veracidad: tanto los datos obtenidos como su posterior tratamiento deben ser veraces para no alterar la toma de decisiones.

Y que estos parámetros también son aplicables a los entornos industriales, podremos imaginar que, en los antedichos entornos, tanto la velocidad como la veracidad se configuran como fundamentales. Tanto una como otra habrán de ser gestionadas a través de algoritmos avanzados dejando para el operador humano una actividad residual no de última decisión sino, probablemente, de parada del sistema. Una especie de "dispositivo del hombre muerto" que únicamente tendría entre sus funciones la de tener la actividad de un algoritmo que, por mala implementación o por "hackeo", por ejemplo, deja de cumplir su función correctamente, poniendo en riesgo la actividad del propio sistema.

Capítulo 7

Aspectos a tener en cuenta a nivel organizativo

Miguel Quintanilla*

La gestión pública de sistemas *big data*

Según los investigadores del MIT, McAfee y Brynjolfsson, las organizaciones que desean obtener un beneficio de la implantación de soluciones de *big data* deben gestionar el cambio de forma efectiva en los siguientes cinco aspectos críticos:

- **El liderazgo.** Las Administraciones públicas que destaquen por el uso de soluciones *big data* no lo harán simplemente porque tengan más o mejores datos, sino porque hayan sido capaces de crear equipos bien liderados que se plantearán las preguntas correctas, fijarán objetivos y definirán métricas. Las soluciones *big data* no van a sustituir bajo ningún concepto la necesidad de disponer de líderes con talento, creatividad y visión en las organizaciones públicas. Las Administraciones más eficientes serán aquellas que hayan contado con el liderazgo suficiente para afrontar con éxito la implantación de soluciones *big data*.
- **El talento.** La aparición de las nuevas soluciones *big data* lleva aparejada no solo la aparición de nuevas tecnologías de gestión y de analítica de datos, sino también el nacimiento de nuevas profesiones y roles dentro de las organizaciones. Acaso existían hace 20 años *chief data officer* (CDO)[1], analistas de *big data*, responsables de SEM y SEO o científicos de datos. Un informe de la consultora McKinsey alerta sobre la escasez de profesionales con estos perfiles y de la necesidad creciente de su reclutamiento en las organizaciones. Es este rol, el del científico de datos, el que va a resultar clave para que las Administraciones públicas puedan desarrollar con éxito sus iniciativas de *big data*.
- **La tecnología.** Las herramientas adquiridas para desarrollar una estrategia basada en la explotación de grandes volúmenes de datos deben cumplir con el modelo 4V (volumen, variedad, velocidad y veracidad). Deben estar

* Ingeniero de Telecomunicaciones, Executive MBA.

1. Un *chief data officer* (CDO) es un director corporativo responsable de la gobernanza y utilización de la información, entendida esta como un activo de la organización, por medio de su procesado, análisis, minería y otros procesos. Los CDOs reportan generalmente al director general de la organización (CEO).

preparadas para gestionar grandes volúmenes de datos, soportar distintos formatos y protocolos de acceso a la información, gestionar los datos a gran velocidad y disponer de mecanismos que permitan descartar datos aparentemente incorrectos. Para hacer frente a estos nuevos retos ha aparecido, en el sector de las bases de datos, nuevas arquitecturas basadas en esquemas no relacionales, mucho más adecuadas para gestionar grandes volúmenes de datos con agilidad.

- **La toma de decisiones.** El primer paso para tomar decisiones acertadas es hacernos las preguntas adecuadas, es decir, definir la pregunta a la que queremos dar respuesta con precisión. El siguiente paso es identificar las fuentes de información que pueden contribuir a responder esa pregunta, sean internas o externas (hay que tener siempre presente que existen grandes volúmenes de datos públicos no explotados). Por último, definir los modelos de análisis de esos volúmenes de datos que generan respuestas sencillas a la pregunta inicial. En muchos casos, el resultado puede ser incluso binario (bien o mal).
- **La cultura.** La implantación de soluciones *big data* en una organización pública precisa de un cambio cultural profundo. Debe imponerse una cultura analítica, basada en la utilización de indicadores, orientada al resultado y a la mejora continua de los procesos. Esta cultura de organización es el sustrato óptimo para poder afrontar innovaciones basadas en la utilización de soluciones de *big data*.

No cabe duda de que las decisiones controladas por datos son mejores. La utilización de tecnología *big data* en las Administraciones públicas va a facilitar a los gestores públicos decidir sobre la base de la evidencia en lugar de tener que seguir su intuición. Por ese motivo, se considera que el principal potencial de la tecnología *big data* es su capacidad para revolucionar la gestión del sector público.

Retos organizativos a la implantación de *big data*

La gestión de los datos en una organización pública es un reto horizontal de la organización, que no implica necesariamente un reto tecnológico. Es un error muy habitual vincular aspectos como la modernización, la innovación o la gestión de datos con el área de tecnología, cuando el mayor reto en la implantación de soluciones para la modernización de una organización pública es la gestión del cambio, la innovación puede estar referida a procesos y en la gestión de datos son más importantes las preguntas que nos hacemos que las respuestas a las que llegamos.

Es decir, el gobierno de los datos requiere de una visión transversal de la organización, de la definición de políticas de adquisición, transporte y almacenamiento de datos y debe estar fuertemente acoplada con el contexto estratégico de la organización. Por lo tanto, el responsable de esta tarea debe tener dependencia directa del presidente de la organización.

Las principales fuentes de riesgo, que suponen retos a la implantación de soluciones *big data*, son la falta de confianza, la incertidumbre acerca de su potencial real, la aparición de errores en la analítica y la escasez de talento. A continuación, se analiza en detalle cada uno de estos retos organizativos.

- La falta de confianza es un factor que puede cambiar la situación actual en términos de acceso y uso de información, hay una creciente conciencia de las personas sobre los datos que son recolectados acerca de todos los aspectos de sus vidas y la gente podría empezar a ser más sensible sobre su privacidad en el mundo electrónico. Adicionalmente, las prácticas de algunas organizaciones públicas, que han sido reveladas recientemente, como las de la Agencia Nacional de Seguridad de Estados Unidos y otros escándalos similares, contribuyen a esta falta de confianza.
- La novedad de *big data* crea incertidumbre sobre el potencial real de sus aplicaciones (por ejemplo, algunos conjuntos de datos son considerados "anonimizados" porque no contienen datos personales, pero no hay garantía de que la identidad de las personas no pueda ser revelada a partir de una combinación de varios conjuntos de datos "anonimizados"). Igualmente, dado que *big data* puede habilitar el descubrimiento de hechos o relaciones en diversidad de situaciones, no está claro hasta qué punto esto puede incluir, por ejemplo, el descubrimiento de información sensible para la seguridad nacional o la violación de la privacidad a través del uso de datos que no son compartidos por las personas de manera intencional. La incertidumbre relacionada con la falta de confianza y los usos impredecibles de los datos generan un grupo de riesgos que pueden afectar la evolución del aprovechamiento de *big data*. Algunos supuestos de Gartner plantean que, para el año 2020, las empresas y gobiernos van a fracasar en la protección de información sensible y van a garantizar el acceso público a la misma.
- El mal uso de la analítica, como origen de riesgo en la explotación de *big data*, no se limita al uso con intenciones criminales, sino que también incluye interpretaciones y decisiones erradas cuando no son considerados algunos elementos (por ejemplo, datos generados de manera activa como el caso de *crowdsourcing*, blogs, foros en periódicos digitales y redes sociales pueden contener información falsa y pueden ser afectados por contenidos virales).

En el mismo sentido, World Economic Forum menciona que los datos basados en teléfonos móviles no siempre son confiables para individualizar comportamientos porque algunas veces las tarjetas SIM son usadas por varias personas y una persona puede usar más de una tarjeta SIM.

- La falta de talento humano capacitado podría llevar a errores de análisis surgidos de pasar por alto la diferencia entre percepciones y hechos o la selección inadecuada de muestras. Otro factor que podría producir errores de uso de *big data* es la falsa ilusión de que una enorme cantidad de datos es garantía para controlar sistemas complejos; en ese sentido, la falla en los modelos analíticos usados por los bancos y agencias calificadoras, que eran alimentados con grandes volúmenes de datos, fue una de las causas de la crisis financiera de 2008.

Factor humano y resistencia

¿Cuál es el papel de las personas en todo esto? ¿Cómo influye su predisposición a favor o en contra de contribuir al avance del *big data*? Los hechos demuestran que hay motivos para desconfiar a la hora de proporcionar datos a terceros. A pesar de ello se da la paradoja de que esta desconfianza resulta aligerarse cuando al usuario le interesa mucho acceder a una tecnología concreta o por el efecto de arrastre de la masa. Es el caso, como hemos visto, de las redes sociales o de los servicios de telefonía móvil.

Es importante que tengamos presente que las personas tienen un doble impacto sobre el desarrollo de soluciones *big data*. Por un lado, porque son los impulsores, diseñadores y ejecutores de las iniciativas *big data*, por otro lado, porque intervienen en algún momento de la definición, grabación o acceso a las fuentes de datos seleccionadas. Este doble rol requiere un análisis del factor humano desde dos perspectivas totalmente distintas, en primer lugar, desde un punto de vista de las capacidades, en segundo lugar, desde un punto de vista de los intereses.

La disponibilidad de talento humano, con conocimiento y actitudes específicas, es identificada como una condición clave para el aprovechamiento de *big data*, en organizaciones públicas como privadas. Es posible identificar tres niveles de caracterización de este requerimiento:

- **Para la operación de datos** se requieren conocimiento en desarrollo, implementación y mantenimiento de herramientas de *software* y *hardware* usadas para *big data* incluyendo la gestión de clústeres para manejo de volumen,

velocidad y variedad de datos, así como "bodegas" de datos lógicas y herramientas como Hadoop. Generalmente, las Administraciones públicas que no disponen de este conocimiento lo externalizan.

- **Para el análisis de datos** se requiere conocimiento en estadística, *machine learning* y ciencia de los datos; por ejemplo, se necesitan personas que manejen la correlación como concepto estadístico usado para análisis predictivo y que tengan conocimientos actualizados que van más allá de técnicas de estadística clásica para asegurar un adecuado tratamiento de nuevas tipologías de datos. Además de eso, dichas personas deben tener conocimiento de la estrategia global de la organización y una mentalidad orientada hacia los datos.
- **Para el uso de datos** se requiere también mentalidad orientada hacia los datos en los niveles de toma de decisiones lo cual implica conciencia del valor de los datos como un activo organizacional valioso y curiosidad intelectual orientada a cerrar los vacíos de información con el uso de herramientas innovadoras.

Analizado el factor humano desde el punto de vista de las capacidades, llega el momento de poner el foco en los intereses. Especialmente aquellos que pueden poner en riesgo el éxito en la implantación de tecnología *big data*, que se fundamentan generalmente en miedos, y que se pueden resumir en:

- Miedo al mayor control y fiscalización de las funciones. Es especialmente habitual que cuando el origen de los datos con el que queremos dar respuesta a las preguntas iniciales es el mismo que el de las preguntas tengamos que recurrir a personal de ese departamento, área o institución para poder acceder a la información. Cuando el trabajador percibe que del acceso y posterior análisis de los datos que va a facilitar va a suponer un mayor control de su trabajo o una fiscalización más exhaustiva de sus funciones existe un elevado riesgo de que no facilite dicha información o de que trate de manipularla a su interés antes de dar acceso a ella.
- Miedo a la pérdida de poder. Es habitual encontrar en organizaciones públicas y privadas personas que ostentan poder, no por lo que saben hacer, sino por la información que custodian (por no decir, retienen contra su voluntad) y la importancia de esta para la organización. Mantienen y perpetúan silos de información, porque de eso depende su poder en la organización. Este tipo de riesgos es el que mayor impacto puede causar en la implantación de tecnología *big data*, ya que generalmente las personas que secuestran datos en una organización son plenamente conscientes de ello, y cualquier intento por devolver esa información a la organización va a contar con todo tipo de resistencias por su parte, resistencias que difícilmente se superan con pedagogía.

- Miedo a la pérdida de funciones. Es muy habitual también que algunos trabajadores perciban las soluciones *big data* como una automatización de funciones que hasta ese momento venían desarrollando personas (en ocasiones incluso muchas personas). Es cierto que en algunas ocasiones la implantación de tecnología *big data* puede suponer una menor carga de trabajo en determinadas posiciones, en estos casos es especialmente relevante haber diseñado un plan de transición, en el que queden perfectamente delimitadas las tareas de cada puesto antes y después de la implantación de la tecnología *big data*. Plan que debe ser compartido con todas las personas alcanzadas por el proyecto, para que visualicen de antemano cómo será su transición hacia el nuevo contexto tecnológico.

Habilidades y recursos

Las iniciativas de *big data* pueden ser decepcionantes cuando las organizaciones carecen de las habilidades y capacidades adecuadas. Como se mencionó anteriormente, el valor de *big data* está muy relacionado con la analítica y algunas técnicas y tecnologías que soportan la analítica son originarias de otros campos de estudio, diferentes a las TIC, como estadística, matemáticas aplicadas, economía y genética. En este escenario, el talento humano surge como una restricción de gran relevancia para el aprovechamiento de *big data* en diferentes niveles organizacionales.

En el nivel directivo, se requiere conocimiento práctico y habilidad para consumir información, lo cual significa que sean personas capaces de hacer las preguntas correctas, analizar, interpretar y cuestionar los resultados con una visión crítica para tomar las decisiones apropiadas. Asimismo los gerentes requieren mejorar su habilidad para identificar qué información está haciendo falta y cómo se puede conseguir. Sin embargo, la situación parece ser distinta en las organizaciones, una encuesta realizada por Gartner entre 720 directivos de diferentes sectores (incluyendo al gobierno) en Norteamérica, Europa, Medio Oriente, África, Asia Pacífico y Latinoamérica muestra que el 56% de ellos consideran un gran reto saber qué fuentes de datos analizar y qué problemas de negocio se pueden resolver con *big data*.

Además de la carencia de habilidades para el consumo de información, el nivel directivo podría presentar limitaciones que son más actitudinales y culturales. Dichas limitaciones podrían resumirse como una mentalidad poco orientada hacia los datos, esta es mencionada como un factor clave para el aprovechamiento

de *big data* y podría ser entendida como la comprensión de los beneficios de la toma de decisiones basada en datos. Esta restricción también está relacionada con la consideración de que el fenómeno de *big data* está cambiando la idea de "trabajar con lo que tenemos" y crea escenarios para "trabajar con lo que podamos conseguir/encontrar". En esos nuevos escenarios se requieren personas con pensamiento creativo y curiosidad intelectual. Con respecto a la carencia de esta mentalidad y del liderazgo necesario para iniciativas de *big data*, destaca la realidad de que algunos ejecutivos están más acostumbrados a sustentar sus decisiones más en su experiencia e instinto que en datos.

En el nivel analítico de las organizaciones, se requieren personas con capacidades técnicas en estadística y *machine learning*. Ellas deberían estar capacitadas para analizar grandes volúmenes de datos y generar conocimiento de negocio. En este nivel, también es importante contar con personas que tengan conocimiento práctico y que tengan la habilidad de consumir información que es requerida por el nivel directivo.

En el nivel operativo y de soporte, se requieren profesionales dedicados a desarrollar, implementar y mantener *hardware* y *software* requerido para el uso de *big data*. Aquí se esperan algunos cambios (por ejemplo, los expertos en infraestructura TI necesitan cambiar su foco del volumen y la velocidad hacia la gestión e integración de gran variedad de fuentes de datos). Adicionalmente, algunas técnicas y tecnologías requieren ser implementadas por expertos. Los líderes de analítica en grandes organizaciones públicas exitosas contribuyen a la identificación de retos relacionados con el uso de datos. Surgen para cubrir la necesidad de contar con "traductores", personas cuyo talento haga de puente entre disciplinas como TI, ciencias de los datos, analítica y el ambiente del negocio en el que trabajan.

Se requieren tres roles clave; el primero es el estratega de los datos que combina conocimiento en TI y experiencia en la toma de decisiones del negocio, este rol debería ser quien defina requerimientos de datos para generar valor con analítica. El segundo rol es el científico de datos, una combinación de vasta experiencia en analítica y conocimiento en TI; este rol debería ser el que desarrolle modelos y algoritmos de análisis de datos. El tercer rol es el consultor de analítica, una combinación de conocimiento práctico del negocio y experiencia en analítica; este rol debería contribuir a la formulación estratégica de iniciativas de *big data*. Además de estos, algunos autores consideran la necesidad de nuevos roles como el *chief data officer* o el *chief analytics officer*, así como el estratega de la información o el gerente de productos de información.

En el mismo sentido, otros autores consideran la interdisciplinariedad como un reto para la conformación de equipos de *big data*. Con respecto a la interdisciplinariedad, además de los datos, este fenómeno involucra el uso creativo de varios tipos de conocimiento humano, como pueden ser la sicología del comportamiento, la antropología social, la analítica del comportamiento y la lingüística cuando el objetivo es hacer análisis de sentimientos a través de datos de redes sociales. Esos conocimientos sumados a habilidades "suaves" como comunicación, colaboración, liderazgo y pasión por los datos son especialmente relevantes.

Otro aspecto a tener en cuenta en la evaluación de restricciones relacionadas con el talento humano es que tener personas con conocimiento profundo en estadística, matemáticas aplicadas y ciencia de los datos puede llevar mucho más tiempo que crear conciencia sobre el valor de los datos o mejorar el conocimiento en operación de tecnologías. Debido a esto, se prevé una carencia de talento humano para iniciativas de *big data* a nivel global y para todos los sectores. Adicionalmente, World Economic Forum plantea que la gente con esas capacidades podría estar más interesada en usar su conocimiento para sus propias iniciativas de emprendimiento en vez de trabajar para organizaciones públicas o privadas.

La sostenibilidad de las soluciones *big data*

Como ya se ha mencionado anteriormente, el primer paso que debemos dar para diseñar una solución basada en *big data* es hacernos las preguntas a las que queremos dar respuesta. Y esas preguntas deben estar, lógicamente, alineadas con los objetivos estratégicos de la organización. El sector público, a diferencia del sector privado, no tiene una cuenta de resultados, no debe dar rentabilidad económica a sus accionistas, no tiene, en definitiva, ánimo de lucro. Pero eso no significa que no deba velar por la eficiencia y calidad en sus operaciones, no solo en los servicios públicos que presta, sino también en su propio funcionamiento interno.

La irrupción de nuevas herramientas de comunicación social, la tendencia global a la implantación de soluciones *smart* y, en general, el desarrollo exponencial de la sociedad de la información y de las infraestructuras de telecomunicaciones (que son las autopistas para que toda esa información viaje, casi, en tiempo real) hacen que el volumen de datos que se está generando crezca de forma exponencial. A modo de ejemplo, **el volumen de datos generados se ha duplicado en los últimos dos años,** es decir, hemos generado tantos datos los últimos dos

años como en todos los años anteriores. Existe una extraordinaria oportunidad de mejora en el sector público por medio de la introducción de soluciones basadas en *big data*, especialmente en la gestión de servicios públicos como el transporte de pasajeros, la gestión del tráfico, la recogida de residuos sólidos, la gestión de las emergencias, la sanidad o la educación.

Para asegurar la sostenibilidad de las soluciones *big data* a implantar, las Administraciones públicas tienen la responsabilidad de asegurar un equilibrio coste-beneficio. Es decir, el coste de adquirir los datos, transportarlos, almacenarlos y procesarlos hasta generar información que repercuta en una mejor prestación de los servicios públicos no puede ser superior al beneficio que reporta dicha información. El beneficio puede (y debe) ser medido en múltiples dimensiones, pero si queremos asegurar la sostenibilidad del sistema a largo plazo, deberá ser traducido a unidades económicas para asegurar que el impacto económico que genera es superior a los costes que introduce, incluso con margen suficiente como para asegurar la amortización de la inversión inicial en un plazo razonable (5 años es el plazo estándar).

A modo de ejemplo, les citaré la implantación de la plataforma de inteligencia turística de Las Palmas de Gran Canaria, que analiza, a partir de los datos de una entidad financiera y de un operador de telecomunicaciones, el comportamiento de los turistas y sus patrones de gasto, por países de origen, puntos de entrada a la ciudad, sexo y edad. **El objetivo de esa plataforma no es otro que el de entender mejor esos patrones de comportamiento con el objetivo de diseñar acciones de promoción turística que contribuyan al aumento de la facturación anual por turismo de la ciudad.** En ningún caso, el coste de adquisición, transporte, almacenamiento y procesamiento de los datos que nutren la plataforma puede superar el incremento anual de facturación vinculada al turismo, en ese caso estaríamos poniendo en serio riesgo la sostenibilidad del proyecto.

Otro elemento que impacta directamente en la sostenibilidad de una solución *big data* es el grado de automatización, especialmente en los procesos de adquisición y tratamiento de los datos. En la medida en que alguno de estos procesos requiera intervención humana, el riesgo para que el sistema se degrade o quede en desuso se dispara. La adquisición del dato debe estar automatizada, puede ser un sensor, un sistema de información de la organización, el ciudadano u otra administración, y la conexión con el sistema de *big data* debe ser directa, mediante acceso directo a la fuente del dato (conexión a base de datos, intercambio de ficheros, servicios web...). Es habitual encontrar en el sector público

iniciativas de *big data* que se lanzan, con gran impulso inicial, pero que a los pocos años han caído en el olvido, especialmente como consecuencia de la no automatización en la adquisición de los datos.

Gestión urbana basada en *big data*

Las ciudades generan mucha información de diferente naturaleza, mucha más de lo que cualquier ser humano o sistema informático es capaz de analizar. Cada día se producen 2,5 quintillones de datos y solamente el 5% de estos datos están estructurados.

Disponen de tres fuentes principales de información urbana: datos generados por sensores de varios tipos, datos de informes accesibles en plataformas de acceso abierto y redes sociales. Por lo general, cada organismo ha sufrido una evolución sustancial en la cantidad de datos que procesa a diario, pero esta evolución se ha producido bajo el modelo de "silo", lo que quiere decir que es capaz de integrar y gestionar más información en sus sistemas, pero de forma aislada con respecto al resto de sistemas del organismo, ciudad o región. En la actualidad, las ciudades que están aprovechando la potencia del *big data* están trabajando para construir sistemas que logren integrar o conectar todos los sistemas de registro de datos de los que se dispone.

Algunos de los ejemplos de cómo las ciudades explotan los datos para otorgar valor añadido a su gestión y además mejor atención al ciudadano son:

- Mapas de salud por zona o barrio, donde puede acceder a información sobre las enfermedades, ratios de nacimiento y muertes.
- Mapas de energía, donde consultar el nivel de consumo de energía por barrio, con la intención de ayudar a los vecinos a comparar entre barrios afines y consensuar mejoras o consejos rápidos para ahorrar energía.
- Mapas de construcción de bloques, donde consultar el avance del desarrollo urbanístico y las zonas que están siendo construidas o derribadas. Información muy útil para gestionar alquileres o arranques de nuevos negocios.
- Mapa de cierre y acceso a escuelas públicas, donde los padres pueden consultar qué escuela está cerrando o a punto de cerrar y cuáles son las alternativas cerca de la zona.

Otras tareas que llevan a cabo las ciudades aprovechándose del estudio de datos masivos son:

- **Eficiencia y atención ciudadana.** Por ejemplo, se están cruzando los datos municipales de contabilidad, asistencia a eventos y equipamientos urbanos con información de sensores que instalamos para medir la humedad, el tráfico, la densidad de población, la climatología, etc., para hacer un uso más eficiente de los sistemas de riego, de la gestión de residuos y del transporte público o para facilitar la organización de eventos. O se están comenzando a conectar los indicadores y sensores de los diferentes cuerpos de seguridad para generar un mapa global de situación donde se reflejan alertas o avisos con diagnósticos producidos por algoritmos que integran datos de sensores de toda la ciudad, como preparación de la población en núcleos muy poblados, frente a grandes contingencias, como pueden ser terremotos, maremotos o grandes tormentas.
- **Seguridad.** El gobierno de Estados Unidos, por ejemplo, ha creado un Centro de Excelencia en NYC que es capaz de gestionar miles de fuentes de información dispersa, habilitar la conexión a diferentes redes y subredes de datos de forma transparente para el operador del Centro de Mando y Control (cámaras de vigilancia, semáforos, sistemas industriales, sensores de humedad, sensores de presencia, sistema de detección de intrusos, sistemas de seguridad de acceso, móviles, ordenadores, etc.) y aportar sensores virtuales que proporcionan nuevos tipos de información, todo compartido en tiempo real.
- **Gestión de eventos.** A través del registro de transacciones de tarjetas de crédito en los comercios de la ciudad, el *big data* proporciona recomendaciones objetivas del impacto económico de la celebración de un evento. Gracias a esta información, no solo puede identificar qué evento genera más ingresos, sino que, además, puede producir información sobre cómo se comportan los visitantes en relación con los comercios de la zona, identificando las áreas más activas y las más afectadas económicamente.
- **Tráfico**. Las ciudades son conscientes de la velocidad del crecimiento demográfico y, por consiguiente, del aumento en el uso del transporte para moverse por la ciudad y entre ciudades. Esto contribuye a que los vehículos sean una de las principales fuentes de contaminación urbana (las emisiones de gases de efecto invernadero, la calidad del aire local y el ruido), que afecta directamente a la salud de los ciudadanos y a su bienestar. El objetivo es la búsqueda de un transporte urbano sostenible con el medio ambiente, garantizando al mismo tiempo la competitividad, y abordar las preocupaciones sociales, como son los problemas de salud o las necesidades de las personas con movilidad reducida. Todo ello es un desafío común y urgente de las principales ciudades de Europa.

Riesgos del uso de fuentes de datos "sociales"

Big data promete el acceso a cantidades ingentes de información en tiempo real de fuentes públicas y privadas que deberían permitir una mejor comprensión en las preferencias de comportamiento, las opciones políticas y los métodos para la mejora de los servicios públicos. El sector privado lleva años sacando partido a las soluciones *big data*, especialmente en el campo del *marketing*. En el sector público las Administraciones son menos sensibles y ágiles en sus interacciones en tiempo real, y no están sacando provecho aún a esta tecnología, por lo que su potencial de aplicación es, si cabe, mayor. Se trata de comprender mejor a su "cliente", el ciudadano.

Sin embargo, se plantean varias preocupaciones importantes con respecto a depender de grandes volúmenes de datos como una herramienta de decisión y formulación de políticas públicas. Mientras que en teoría las soluciones *big data* son exhaustivas y completas, en la práctica, la versión actual de algunas soluciones *big data* tiene características que hacen que los gestores públicos y los académicos se mantengan reticentes. En primer lugar, una parte de lo que consideramos *big data* es realmente *data exhaust*, es decir, los datos recogidos para fines distintos de las operaciones del sector público. Los conjuntos de datos accesibles públicamente a partir de redes sociales como Facebook o Twitter fueron diseñados por razones puramente técnicas. El grado de acierto con el que estas fuentes de datos pueden dar respuesta a preguntas relativas al sector público es absolutamente residual. **El uso de fuentes de datos para propósitos para los que no fueron concebidas puede generar resultados impredecibles.** Un buen ejemplo es el intento de Google de predecir la gripe en base a los términos de búsqueda.

En segundo lugar, hay cuestiones éticas que pueden surgir cuando se utilizan los datos que fueron captados como un subproducto de la interacción de los ciudadanos entre sí o con una cuenta de medios de comunicación social de una Administración pública. Los ciudadanos no son capaces de comprender o controlar cómo se usa su información y no han dado su consentimiento para el almacenamiento y reutilización de sus datos. En este punto resulta esencial asegurar el anonimato en los datos, es decir, la información debe agregarse hasta tal punto que no sea posible identificar un indicador concreto con una persona específica. Una regla no escrita es que cualquier conjunto de datos debe contener, al menos, diez muestras para unos criterios de filtrado definidos. No debería ser posible obtener grupos de muestras inferiores a diez con ninguna combinación de filtros, en ese caso el sistema no debería devolver el dato.

Por último, el *big data* que proviene de herramientas de comunicación social representa solo al sector de población de las personas que en algún momento han tenido actividad en esa herramienta social. Sabemos que ciertos segmentos de la sociedad optan por una vida *online* (mediante el uso de las redes sociales o los dispositivos conectados a la red), otros optan por una vida *offline* (a sabiendas o sin ser conscientes), y otros simplemente no tienen los recursos para poder decidir. La población de Internet en este aspecto es relevante. Por ejemplo, las Administraciones públicas tienden a utilizar los datos de Twitter por la sencillez de su API, que permite la recogida de datos, pero rara vez tienen en cuenta que esos usuarios de Twitter no son representativos de la población general.

En resumen, queda claro el inmenso potencial del *big data* a partir de fuentes de comunicación social y su utilidad para el sector público, pero es importante tener presente que esos datos deben ser enriquecidos de manera eficaz con otros datos recopilados por la propia Administración para poder aportar valor a la mejor gestión pública. **Cada vez más, los gestores públicos deben saber cómo recoger, gestionar y analizar grandes volúmenes de datos, pero al mismo tiempo deben estar plenamente familiarizados con las limitaciones y su potencial de uso indebido.**

Capítulo 8

Innovación basada en ciencia de datos, modelos y tecnologías

Diego May*, Frans van Dunné**

› Introducción

En la última década se fue haciendo cada vez factible obtener valor de los datos ya que era más barato generarlos, almacenarlos y procesarlos. Hace solo veinte años (1996) se hizo más rentable guardar los datos en computadoras que en papel (Morris y Truskowski, 2003). Las grandes empresas de tecnología están creando importantes ventajas competitivas a través del uso intensivo de datos. Los gobiernos están abriendo *datasets* relevantes que los desarrolladores y las empresas utilizan cada vez más. Y hay cada vez más soluciones y plataformas para procesar y crear valor a partir de los datos.

Las oportunidades de innovación a través de los datos se hacen cada vez más claras y los modelos, *frameworks* y prácticas van madurando en tanto estos se aplican a empresas de distintos tamaños en diferentes segmentos de mercado.

Esta tendencia que comenzó con las grandes empresas de tecnología como Google, Facebook, Twitter, AirBnB (por nombrar algunas) está llegando ya a otros niveles de empresas y en los próximos diez años será masivo. Todas las organizaciones contarán con infraestructura y tecnología para recolectar y almacenar datos, y deberán entender y aplicar ciencia de datos para sacar provecho de forma adecuada.

Además de nuevas oportunidades, este desarrollo también presenta nuevas necesidades ya que va a requerir profesionales que a todos niveles, desde los más técnicos hasta gerentes y líderes, puedan no solo entender el valor de los datos sino también interpretar los resultados y saber cómo actuar de acuerdo con los mismos.

* Cofundador de ixpantia (Ciencia de Datos) y de Junar (Datos Abiertos). MBA por el MIT Sloan School of Management.
** Cofundador de ixpantia. PhD en Biología de la Universidad de Amsterdam.

El objetivo de este capítulo es cubrir algunos de estos aspectos básicos sobre infraestructura y ciencia de datos. Los temas que se abordarán son los siguientes:

- Modelos generales sobre innovación basada en datos.
- El ciclo de innovación en ciencia de datos y un modelo que describe puntos de entrada.
- Un modelo para caracterizar los datos y según esto definir cómo afrontar distintos tipos de problemas y preguntas en ciencia de datos.
- Algunos comentarios acerca de plataformas, tecnologías y herramientas.

Este capítulo debiera dar una buena introducción a la temática y ofrecer algunas herramientas para aquellas organizaciones que están buscando desarrollar iniciativas de innovación basada en datos, así como también a profesionales que están interesados en esta temática.

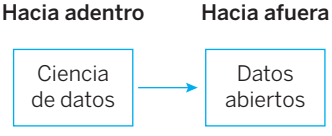
Dos aproximaciones a la innovación basada en datos

Cada vez es más fácil guardar mayores volúmenes de datos y hacerlos disponibles de formas adecuadas para que estos puedan transformarse en información relevante. Pero también es cierto que con el mayor volumen de datos existentes el análisis de los mismos se hace más complejo y es importante encontrar las formas adecuadas para generar dicha innovación a partir de los datos.

En esta sección hablaremos de dos grandes paradigmas a considerar a la hora de evaluar alternativas de innovación en base a datos. Las mismas son:

1. Innovación hacia fuera de la organización o alineada con el movimiento de *datos abiertos*.
2. Innovación hacia dentro de la organización o alineada con la analítica, *big data* y ciencia de datos.

Imagen 1. Modelos de innovación



Hacia afuera, innovación basada en datos abiertos

La innovación basada en datos abiertos se ve en general en el sector. En estos casos, los gobiernos (nacionales, regionales o municipales) abren datos a través de portales con la motivación de generar: transparencia, colaboración, participación, eficiencias e innovación.

Los gobiernos cuentan con muchos datos valiosos que podrían generar un alto impacto para los ciudadanos. Por otro lado, estos datos que tienen los gobiernos son recolectados y mantenidos gracias a los impuestos que paga la ciudadanía, por lo cual se considera que estos datos deberían estar disponibles de manera abierta, siempre que no se comprometa la privacidad o seguridad de los individuos.

Imagen 2. Etapas *open data*

El paradigma de colaboración e innovación en estos casos es simple: los gobiernos tienen los datos, los ciudadanos y el sector privado en general pueden contribuir a generar soluciones innovadoras estos datos. Y estas soluciones pueden suponer un impacto directo en la ciudadanía (*apps*, nuevos servicios) y en el gobierno (*apps*, mayor eficiencia), o contribuir al desarrollo de nuevas empresas o a la mejora de productos que tal vez más indirectamente impactan positivamente a la sociedad.

Para implementar estos programas se suelen seguir los siguientes pasos:

1. **Mapeo de los datos.** Típicamente existe una unidad (tecnologías de la información o desarrollo económico por dar un par de ejemplos) con la responsabilidad de realizar esta apertura de datos en el gobierno. Esta unidad trabaja con el resto de la organización para entender cuáles son los conjuntos de datos que los diferentes departamentos generan y mantienen.

Según este mapeo y al entender los pedidos de información y datos de la ciudadanía se genera una lista completa de la información que podría abrirse y después de pasar los filtros legales generarse un plan gradual de apertura de datos (*open data roadmap*).

2. **Creación de un portal de datos abiertos.** Existen diferentes formas de implementar un portal de datos abiertos. Son muy pocos los gobiernos que actualmente optan por desarrollar y mantener una solución pero existen algunos de estos casos. Típicamente los gobiernos optan por implementar una solución de *software* en la nube (Socrata, Junar, *open data soft*) o implementar alguna solución de código abierto que luego la organización puede mantener (*ckan, dkan, junar*). Cualquiera sea el portal de datos abiertos desarrollado por la organización, el mismo deberá permitir el acceso a los datos publicados considerando las características solicitadas por las distintas audiencias: ciudadanos (encontrar y entender datos a través de visualizaciones), academia (poder acceder a los datos crudos), desarrolladores y periodistas (poder tener acceso sistemático vía APIs).
3. **Generar programas de comunicación, promoción y utilización de los datos.** No basta con haber dado los pasos 1 y 2 para tener éxito en generar innovación. Es clave que las organizaciones que publican los datos, tanto en el sector público como privado, hagan esfuerzos para promover iniciativas que permitan la difusión y que incentiven la utilización de estos datos. Más allá de hacer prensa y difusión sobre los programas, se logra mucho cuando los gobiernos generan *hackathones*. Estos eventos que atraen a desarrolladores, diseñadores, *hackers* cívicos y al sector privado en general buscan que en periodos cortos de tiempo se generen algunas soluciones a problemas existentes en la ciudad, de acuerdo con desarrollos de aplicaciones fundadas en los datos que abren las instituciones de gobierno.

Son muchos los ejemplos de ciudades que han logrado generar impacto e innovación a través de programas de datos abiertos. En EE. UU. grandes ciudades como Chicago, San Francisco y New York y ciudades más pequeñas como Mesa (en Arizona) y Palo Alto (en California) han desarrollado programas muy completos. En Latinoamérica hay casos relevantes en Chile (Providencia, Puente Alto), en Perú (Miraflores, San Isidro) y en Argentina (Ciudad de Buenos Aires, Ciudad de Bahía Blanca) como también en México, Colombia y Brasil.

Los casos de éxito tienen en común:

- Realizan una apertura inicial bien promocionada.
- Generan algún evento de innovación o *hackathon*.

- Dan seguimiento a iniciativas interesantes y apoyan a emprendedores e innovadores.
- Tienen un *roadmap* de apertura datos que incluye datos en tiempo real (valiosos para innovar).

Lo que ha generado impacto en el sector gobierno ya está permeando al sector académico y al sector de impacto social (ONG, fundaciones) y existen casos de apertura de datos para generar innovación en el sector privado (competencias Kaggle, Netflix Prize). Y esto es solo el comienzo.

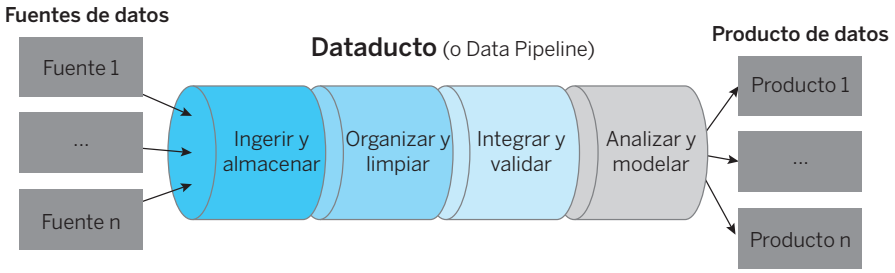
Hacia adentro, innovación basada en ciencia de datos

En las próximas secciones de este capítulo hablaremos en mayor detalle acerca de esta innovación basada en ciencia de datos. Lo que nos interesa ahora es poder dejar claro tanto la distinción como la relación entre innovación hacia dentro e innovación hacia fuera.

Cuando las organizaciones están generando innovación hacia dentro, deben antes que nada contar con recursos humanos que sepan utilizar técnicas, procesos y metodologías para interpretar y sacar valor de los datos que idealmente se deben encontrar alojados en una infraestructura que permita el fácil acceso a los mismos.

El científico de datos tendrá preguntas específicas que contestar y según esto podrá tener los *datasets* en una infraestructura de fácil acceso (un *lake* o lago de datos) y los hará interoperables a través de procesos de integración. También podrá ver qué datos externos a la organización pueden enriquecer el análisis con el objetivo de responder las preguntas. Finalmente podrá hacer el análisis y generar resultados, reportes y productos de datos. Todo esto se verá en próximas secciones y se puede ver esquematizado en la imagen 3.

Imagen 3. El proceso de ciencia de datos

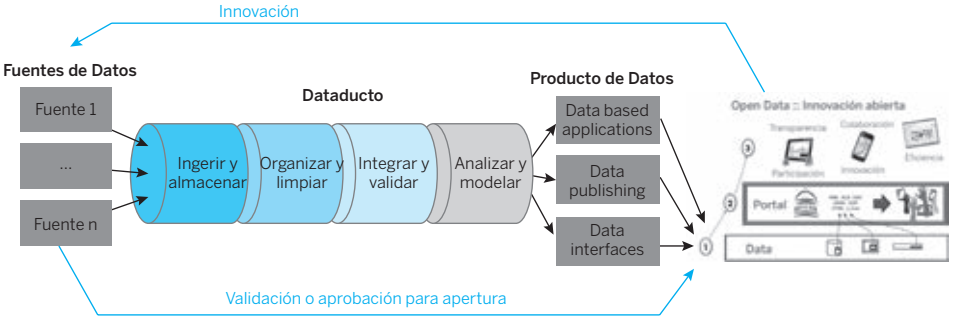


Los equipos que suelen trabajar en prácticas de ciencia de datos en las organizaciones son diferentes a los tradicionalmente conocidos como BI (*business intelligence*) o *inteligencia de negocios*. Principalmente porque en ciencia de datos se buscan equipos que operen con un método científico para tratar de explorar caminos no transitados y responder a preguntas nuevas, mientras que tradicionalmente el equipo de BI se dedicaba a interactuar con las bases de datos y sistemas para generar reportes con resultados específicos esperados y mayormente conocidos.

Por este motivo, según el tipo de organización y de la complejidad de los datos, así como de la importancia que se da a la innovación a través de estos, se definen distintos esquemas de trabajo. Los tres esquemas más conocidos son: equipos centralizados, equipos distribuidos o equipos híbridos. No es objetivo de este capítulo entrar en detalle en estos modelos, pero sí al menos describirlos para que los lectores puedan luego profundizar. Algunas empresas optan por tener un equipo de ciencia de datos central que atiende las distintas unidades de negocios. Otras organizaciones prefieren un equipo distribuido con científicos de datos operando en las distintas unidades de negocios. Finalmente, existen modelos híbridos donde se busca obtener los beneficios de un equipo centralizado de ciencia de datos que se retroalimenta y que permite alta innovación, junto con la ventaja que tienen los equipos distribuidos de operar de forma cercana a las unidades de negocios y entender con mayor profundidad las problemáticas que tienen que resolver estos grupos.

Es interesante ver que ambos modelos de innovación basada en datos (hacia dentro y hacia fuera) pueden interactuar y complementarse. Existirán "preguntas" que deberán responderse internamente o que al menos requerirán un preproceso a lo interno de la organización aplicando metodologías de ciencia de datos. Como resultado se pueden generar nuevas preguntas que podrán ser adecuadas para programas de innovación de base a datos abiertos llegando a poblaciones externas a la organización (ciudadanía, empresas, academia, empresarios, innovadores). Esta interacción está esquematizada en la imagen 4.

Imagen 4. Interacción entre innovación basada en datos hacia adentro (ciencia de datos) y hacia afuera (datos abiertos)



El ciclo de innovación basada en ciencia de datos

Diferentes modelos de innovación se han descrito con diversos modelos tales como lineales, cadenas interconectadas o circulares (Kline y Rosenberg, 1986). También existen modelos con diferentes enfoques como gestión (Bassiti y Ajhoun, 2013) o el proceso creativo (Buchanan, 1992). En esta sección miramos un ciclo de innovación que propone integrar el método científico en la aproximación. Esto no solo sirve para asegurar una metodología replicable y eficiente en producir preguntas que se dejan responder con los datos disponibles, sino también para plantear las pautas bajo las cuales podemos identificar prioridades en las propuestas de innovación.

El ciclo que proponemos para la innovación basada en datos se puede ver en la imagen 5. Como modelo para empezar a generar ideas y para identificar oportunidades de innovación se puede comenzar desde cualquier punto del ciclo. Comúnmente empezamos desde la búsqueda de una ventaja competitiva identificada o desde observaciones. Daremos ejemplos al describir los pasos uno por uno.

Pasos del ciclo de innovación basada en ciencia de datos

Cada uno de los pasos descritos a continuación puede ser el puntapié inicial para un proyecto de innovación a través de datos o simplemente una parte del proceso para gestionar dicha innovación.

Imagen 5. Ciclo de Innovación basado en datos

Búsqueda de ventajas competitivas

Mucha de la innovación que se genera en las organizaciones se basa en esta búsqueda constante de diferenciación. Tener una ventaja competitiva permite a las organizaciones ofrecer mayor valor al mercado a un menor costo, o cuando podemos pedir un mayor precio porque nos diferenciamos favorablemente de nuestros competidores. Los datos ofrecen múltiples formas para encontrar ventajas ya que permiten incrementar:

- la velocidad de las decisiones,
- la calidad de las decisiones,
- la velocidad de respuesta a clientes y a desarrollos en el mercado,
- la eficiencia de nuestros procesos de negocio o de producción.

Según estas búsquedas de ventajas competitivas se puede comenzar el proceso de innovación en base a datos para posteriomente generar las preguntas que serán respondidas con datos.

A partir de observaciones

En muchos casos el puntapié inicial para generación de innovación es el contar con una observación que proviene de alguna de las áreas de negocios. Aquí un par de ejemplos de este tipo de observaciones:

- "integrando estos procesos seríamos más eficientes",
- "segmentando los mercados seríamos más efectivos con nuestras ofertas".

Según estas observaciones se puede comenzar un proceso que nos lleve a buscar los datos que permitan validar (o no) la observación y así innovar. Por ejemplo, si hemos decidido que necesitamos incrementar la velocidad de responder a quejas de clientes, necesitamos observar dónde o cuándo se quejan los clientes. El mejor caso es cuando llaman al *call center* o a nuestros representantes de ventas. Pero muchas empresas han visto que frecuentemente las quejas se hacen de forma indirecta, a través de medios de comunicación sociales como Facebook o Twitter.

El ejemplo anterior también pudo haber sido nuestro punto de comienzo: alguien en la organización observa que las quejas no van a través de canales ya identificados, sino a través de medios sociales.

Preguntas

Las preguntas que se hace la organización pueden también ser una forma de comenzar estos procesos de innovación en base a datos. Debajo algunos ejemplos de preguntas:

Las observaciones nos llevan a formular preguntas que nos indican dónde, cuáles y qué tipo de datos necesitamos colectar. Siguiendo el ejemplo de arriba te podrías imaginar preguntas como:

- ¿Qué medios reciben más quejas?
- ¿Qué medios reciben las peores quejas?
- ¿Dónde se ven más quejas?
- ¿Dónde se ve más discusión sobre quejas?
- ¿Cómo influye una queja el sentimiento sobre nuestra marca?

Estas preguntas nos llevan luego a explorar los datos que nos llevarían a las respuestas y así comenzar un nuevo proyecto de ciencia de datos en la organización.

Datos

En muchos casos las organizaciones llegan a la conclusión de que cuentan con muchos datos valiosos y se dan cuenta de que es posible que estos datos, bien utilizados, permitan responder preguntas y generar mayores eficiencias y ventajas competitivas. Este *approach* es cada vez más relevante dado que son muchas las organizaciones que al ver el advenimiento de *big data* y ciencia de datos sienten que tienen que hacer algo.

El primer paso en estos casos es entender los datos con que se cuenta y la arquitectura de la información en la organización. De acuerdo con esto, después tiene

sentido tratar definir alguna pregunta puntual que pueda permitir el desarrollo de un primer proyecto de ciencia de datos.

Traducir preguntas en la necesidad de tener datos ya empieza a explicar por qué hablamos de ciencia de datos. Se acerca más al diseño experimental porque las decisiones sobre cuáles, cuántos y qué tipos de datos recogemos tienen un efecto directo sobre las decisiones a las cuales podemos dar soporte.

Al trabajar los datos también hay que tener en cuenta los posibles análisis y formas que estos pueden tener.

Productos de datos

Aunque los productos de datos suelen ser un resultado de un proyecto de innovación a través de datos, dada la experiencia previa existente en inteligencia de negocios en las organizaciones suele suceder que un resultado de un cubo o un reporte puede generar un nuevo proyecto de ciencia de datos.

En muchos casos, el proyecto de innovación en base a datos habrá comenzado desde alguna pregunta u observación y esto concluirá en el desarrollo de un producto de datos. Para mejorar una parte de la organización, por ejemplo, el proceso de negocio llamado "atención al cliente", necesitamos traducir lo que aprendimos de los datos en un producto. Hablamos de un producto en un sentido muy amplio porque puede incluir cosas como:

- Una gráfica o series de gráficas que nos ayudan a ejecutar un trabajo o tomar decisiones.
- Una fuente de datos trabajados que son leídos por otro proceso automático para mejorarlo.
- Un informe que dirige acciones entre personas.
- Una alarma que le indica a una persona o a una máquina que ha de tomar acción.

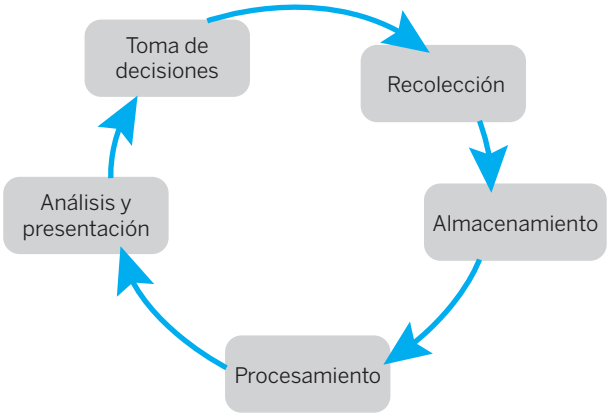
No podríamos dar una lista completa de ejemplos, ya que muchos de estos aún están por descubrirse. La innovación basada en datos está en movimiento.

Ciclo de valor IBD (innovación basada en datos)

En el ejemplo anterior llegamos a la siguiente iteración del ciclo cuando seguimos buscando dónde y cómo innovar con base en datos. Por otro lado, los productos que hemos generado entran al ciclo de valor de innovación basada en datos (*DDI Value Cycle*) como se presenta en la imagen 6. En este ciclo mejoramos

el producto de datos de forma continua y tenemos un marco de referencia para medir la calidad de cada paso.

Imagen 6. Ciclo de valor de datos (basado en Hayden *et al*., 2015)



Recolección

La recolección de datos, una vez que se ha llevado el producto a producción (implica que ya está operativo y utilizado por usuarios reales), incluye todos los pasos desde la fuente (que puede ser un cliente, un agente de ventas o una máquina) hasta el momento en que lo almacenamos. Esto incluye los posibles controles de disponibilidad y calidad que hayamos definido.

Almacenamiento

El almacenamiento de datos tiene que cumplir varias funciones importantes para asegurar que tenemos los datos disponibles en el plazo necesario:

- La forma de acceso, al nivel de agregación necesario para el consumidor.
- El gobierno del acceso (*data governance*) para que solo los indicados tengan acceso a los datos.
- La calidad del almacenamiento (si lo necesitamos en 30 años, ¿cómo nos aseguramos de que no hay cambios o borrado de datos inesperados?).
- La calidad de *retrieval*, incluyendo la gestión de metadatos, la auditoría de uso y aplicación.

Procesamiento de datos

Los datos crudos (no procesados), si bien tienen mucho valor latente, difícilmente están listos para ser consumidos y generar valor directamente. El valor se

añade en el momento de procesarlos, y la ventaja competitiva de nuestras empresas de cara al futuro está en gran parte en nuestros activos de procesamiento de datos: algoritmos y dataductos (del inglés *data pipelines*).

Es muy importante que en el procesamiento de datos se apliquen las reglas de gobierno de datos también, para poder validar que en el momento de procesamiento no se rompen las reglas.

Análisis y presentación

En este paso se puede ver cómo empacar los datos o presentarlos para que sean consumidos fácilmente por los distintos actores o *stakeholders*.

Toma de decisiones

La toma de decisiones es quizá el momento que se deja medir más fácilmente de cara al negocio. Si la calidad se incrementa porque se mide mejor, hay mejor conversión en ventas o una mejora en indicadores operativos, sabemos que nuestros productos de datos están cumpliendo con sus objetivos.

No basta con medir la mejora de toma de decisiones una sola vez y declarar un producto de datos como un éxito. Hay muchos puntos dentro del ciclo de valor DDI que pueden cambiar la calidad de la toma de decisiones. Quizá la decisión más importante es ver si aún estamos evaluando las preguntas correctas, para lo cual hay que regresar al ciclo de innovación basada en datos.

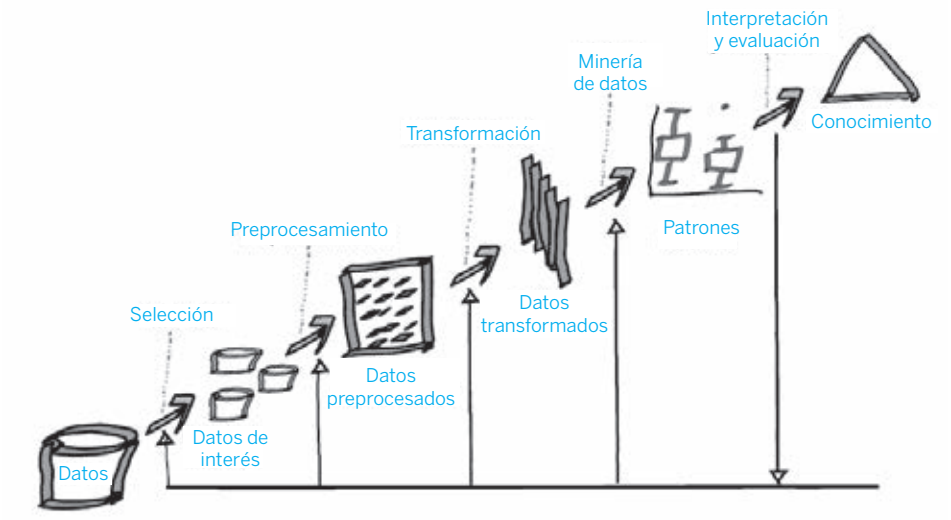
Es importante aclarar a estas alturas que en algunas empresas (StitchFix es un buen ejemplo) los productos de datos pueden ser composiciones sofisticadas de algoritmos y procesos. En estos casos, la mejora continua de estos productos puede implicar la interacción entre algoritmos y el criterio experto de los humanos (en el caso de estas empresas, los diseñadores que toman las recomendaciones de las máquinas o incluso las devoluciones de los mismos clientes).

› ¿Tengo datos, ahora qué?

En la práctica descrita arriba operan científicos de datos y consultores de negocio mano a mano para poder identificar las prioridades de innovación. Pero un momento clave en todo el proceso es el momento en que hay datos disponibles y en coro suena la pregunta: "¿Tengo datos, ahora qué?".

Quizá el modelo más conocido para llegar desde datos (crudos) a conocimiento (el producto final de datos) es el de Fayyad *et al.* (1996) llamado "descubrimiento de conocimiento en bases a datos". Los pasos que describe este modelo se pueden aplicar en casi todas las circunstancias y ayudan a entender el proceso (imagen 7).

Imagen 7. El proceso de descubrimiento de conocimiento en bases de datos (basado en Fayyad *et al*,. 1996)



Lo primero que necesitamos identificar es dónde están los datos. En un contexto organizativo esto puede ser un ejercicio de mapear dónde hay datos, para lo cual nos debemos asegurar de incluirlos todos y no solamente aquellos que están mapeados y se manejan a través de un protocolo o método de gestión establecido. En otras palabras, podemos tener acceso a datos en bases de datos "oficiales" dentro de la organización, pero probablemente hay datos en bases de datos no oficiales o semioficiales. Hay datos en diferentes estados de validez en hojas Excel y en documentos. Además, y cada vez con más relevancia, hay bases de datos externas que son de interés, ya sean datos abiertos o datos relevantes para la industria en la cual estamos.

De todos estos datos necesitamos identificar cuáles son los de mayor interés, por ejemplo, usando el ciclo de innovación basada en datos descrito anteriormente. Al identificarlos ponemos en marcha un proceso técnico para seleccionarlos

(con SQL, R-dplyr, MapReduce, etc.). Esto no da la selección de datos, probablemente tenemos que procesarlos para que se puedan usar para análisis. Por ejemplo, necesitamos incrementar el valor de los datos sacando todos los errores de ortografía en una categorización de interés. O tenemos datos en dos o más bases de datos separadas que queremos integrar.

Estos pasos del preprocesamiento también se traducen a código y pasos operativos en tecnología. Y con ellos estamos construyendo el siguiente paso en nuestro dataducto, a través del cual vamos a asegurarnos de que podemos repetir todos los pasos. Esto lo hacemos no solamente para protocolizar de forma automática los pasos para ingerir datos nuevos de forma rápida, sino también para poder implementar correcciones y mejoras en el proceso a medida que vamos aprendiendo más de cada iteración de nuestro ciclo de innovación y ciclo de valor.

Los datos preprocesados sirven para un análisis exploratorio que nos va a indicar qué transformaciones tenemos que hacer en los datos para que se puedan analizar para reconocer patrones y construir modelos predictivos. Por ejemplo, necesitamos convertir pies a metros para tener una sola dimensión de distancia en el conjunto de datos completo. O tenemos datos de interés sobre una cantidad que aún están incluidos como texto.

Es laborioso llegar al punto en que los datos están listos para aplicarles métodos de estadísticas y *machine learning*. Una forma de evitar malgastar recursos en limpiar datos que no sirven tanto como pensamos es definir pruebas de conceptos sobre un subconjunto de los datos para valorar el resultado y el impacto que tendrá sobre nuestro negocio.

El paso que Fayyad llama "minería de datos" (imagen 7) es descrita por Wickham y Grolemund (2016) como un ciclo que requiere que sepamos cómo tratar los datos de interés. Aquí entra la necesidad de saber sobre estadísticas y *machine learning* para poder definir qué métodos se pueden aplicar, tanto de acuerdo a lo que los datos permiten, como a lo que necesitamos para obtener respuesta a la pregunta que hemos formulado. Los pasos de transformar, modelar y visualizar datos en un ciclo estrecho e iterativo lleva cada vez a mayor conocimiento sobre el fenómeno bajo estudio.

El objetivo es poder comunicar de forma efectiva, y esta comunicación es lo que anteriormente hemos llamado producto de datos. El objetivo del proceso total es identificar la metodología, el modelo, el algoritmo o la visualización que puede ser la base para el producto de datos que estamos construyendo.

Esto nos lleva a las tres preguntas esenciales que hay que considerar para definir el modelo, algoritmo o visualización que va a formar la base para el producto de datos. Las presentamos en la imagen 8. En resumen son:

1. ¿Qué quiero con mis datos?
2. ¿Qué características tienen los conjuntos de datos?
3. ¿Cómo son las variables de mis datos?

Imagen 8. Un esquema para marcar las preguntas y respuestas que determinan qué métodos se pueden usar para analizar datos

¿Qué quiero con mis datos?

Lo principal es saber qué es lo que quieres o en otras palabras: ¿a qué pregunta quieres dar respuesta? A grandes rasgos vemos que comúnmente lo que queremos hacer con los datos es comparar, agrupar, predecir, reconocer o asociar.

- **Comparar.** Cuando queremos comparar datos necesitamos saber de antemano qué grupos tenemos. Esto parece lógico, pero muchas veces pasa que queremos comparar los tres grupos meta más importantes cuando lo que realmente queremos es "identificar nuestros tres grupos meta más importantes". Casos más claros son, por ejemplo, la comparación del desempeño de mercados regionales. Es importante notar que la decisión de qué método aplicar para poder comparar grupos es en la mayor parte una razón estadística (hay métodos paramétricos, no paramétricos y de rangos por dar algunos ejemplos) que depende de la tercera categorización que se describe más adelante, como son las variables de mis datos.

- **Agrupar.** Otro popular objetivo es agrupar. Es una forma natural de enfrentar la diversidad en nuestro entorno, donde nuestro cerebro busca rasgos comunes para categorizar objetos. Así reconocemos un tomate naranja, aunque solamente hemos estado expuestos a tomates rojos y verdes antes. Al introducir formas de agrupar en nuestros negocios es importante saber qué pasos en el tiempo queremos revisar en la agrupación que tenemos. ¿Es algo para revisar de forma continua?, ¿con qué periodicidad?, ¿cada segundo, cada mes, cada año? La respuesta tendrá poco que ver con datos y mucho que ver con objetivos de negocio.
- **Predecir.** Con predicción nos referimos a la construcción de modelos con datos que tenemos disponibles, para poder decir algo sobre situaciones de los cuales no tenemos (todos) los datos disponibles aún. Y predicción se puede hacer a diferentes niveles de exactitud y de precisión según el contexto. Si vendo helados quisiera saber "más o menos" qué caluroso va a ser el año entrante. Pero si tengo un modelo para predecir cuándo reemplazar un componente de un motor de avión, el "más o menos" ya no es suficiente, tiene que ser con un muy alto grado de precisión.
- **Reconocer**. Cuando queremos reconocer patrones queremos contar con la ayuda de un computador para, por ejemplo, reconocer la cara de las personas al entrar a un estadio o estación de metro. O queremos reconocer el tamaño y la forma de una piña para identificar estos datos como indicadores calidad de exportación. Hay un sinnúmero de oportunidades de negocio en la industria en esta temática, sobre todo, en el control de calidad de sistemas de producción.
- **Asociar**. Como último caso están las asociaciones, como, por ejemplo, las correlaciones. Son relaciones numéricas entre variables que no necesariamente indican una causa y efecto, pero indican una tendencia. Pueden ser una buena base para una observación con la cual arrancar nuestro ciclo de innovación basado en datos, por ejemplo. Pero también se usan con frecuencia para dar soporte a decisiones sobre mercadeo y optimización de procesos.

¿Qué características tienen los conjuntos de datos?

Sabiendo lo que queremos con nuestros datos, el siguiente paso es identificar las características principales de los datos. ¿Cuál es el volumen de los datos? ¿Qué grande es la variedad de los datos? ¿Se producen o hay que ingerirlos a cierta velocidad? ¿Qué veraces son y cuál es su valor?

- **Volumen**. Cuando hablamos del volumen de los datos, buscamos indicar cuántos puntos de medición hay o cuántos *bytes* si se tratan de texto o

imágenes. Los problemas se dan cuando hay más datos de lo que cabe en la memoria del computador que estás usando, o peor, más de lo que cabe en el disco duro de tu servidor. Por el otro lado, también hay problemas cuando el volumen es muy bajo y tienes pocos datos y muchas preguntas.

- **Variedad**. Cuanto más homogéneas sean las variables, más fácil será trabajar los conjuntos de datos, y esto marcará la variedad de los datos. Por ejemplo, si solo tengo datos numéricos, o mejor aún solo datos numéricos continuos, tengo más posibles metodologías que si tengo 100 GB de texto con comentarios en palabras en 40 lenguajes. O si tengo una mezcla de datos numéricos, texto, vídeo e imágenes. Cuanto más diferentes son los tipos de variables en mi conjunto de datos, más alta es la variedad.
- **Velocidad**. Cuando los datos se generan continuamente y fluyen con cierta velocidad (*streaming data*) se requieren soluciones especiales. Cuando la velocidad es baja, podríamos, por ejemplo, recalcular el promedio de una variable cada ciertos intervalos de tiempo. Pero cuando la velocidad del flujo de información incrementa muy pronto, ya no tenemos la velocidad de cálculo disponible para recalcular con todos los datos en tiempo real.
- **Veracidad**. También es importante saber qué datos reflejan hechos reproducibles. En otras palabras, qué veraces son los datos que tenemos. Por ejemplo, si uso datos experimentales, voy a tener mayor confianza en mis datos que si hago el experimento yo mismo. Si les pido a 20 personas al azar repetir el experimento, es posible que la confiabilidad (a priori) de los datos sea más baja. Algunas personas quizá no entiendan el experimento o se inventen resultados para poder entregar más rápido. Igual pasa en ciencia de datos, sobre todo en datos de empresas grandes, donde a veces ya nadie sabe de dónde vienen los datos, o se trata de encuestas donde no todos los que responden tienen interés en responder con precisión.
- **Valor y costo**. Por último, vale la pena pensar en el costo o valor de los datos que tengo disponibles o que necesito conseguir. Hay conjuntos de datos que han costado millones de dólares reunir (valor en dinero). O donde un solo punto extra, cuesta días, meses o más en conseguir (valor en tiempo). También hay análisis donde un error puede suponer perder millones.

¿Cómo son las variables de mis datos?

Por último, y ya a un nivel de detalle del conjunto de datos específico que vamos a incluir en un modelo, algoritmo o visualización, es si los datos son continuos, ordinales, nominales, texto o imágenes. Por lo general este es el primer capítulo de todo libro de introducción a estadísticas porque tiene impacto directo en qué

métodos puedo aplicar y cuáles no. Cada tipo de datos se comporta de una forma propia al someterlos a análisis.

- **Continuos**. Datos continuos son los que se pueden dividir sin límite. Por ejemplo, en una escala de temperatura se puede medir en grados Celsius o fracciones hasta tal punto que pueda obtener precisión de mi instrumento de medir.
- **Ordinales**. Datos ordinales tienen un orden, por ejemplo, grande, pequeño y mediano, pero no los puedo fraccionar. No hay una medida entre grande y pequeño.
- **Nominales**. Datos nominales son los que tienen un nombre, pero no un orden predeterminado. Son datos que podríamos tener en el CRM de la empresa, tales como el oficio de las personas, el género o el tipo de dispositivo móvil que usan.
- **Palabras**. Cada vez más se ven palabras como variable. El contenido de documentos, de e-mails, de mensajes en Twitter u otros medios sociales. Carecen de una estructura previa y tienen su propia familia de metodologías para analizarlos.
- **Imágenes**. Por último, tenemos imágenes, ya sean estáticas (fotos) y dinámicas (vídeos). Estos también son relativamente nuevos y tienen menos historia de análisis estadístico que los demás. Pero quizá tienen más historia de análisis en *machine learning* e inteligencia artificial.

Resumiendo

Obtener datos automáticamente nos obliga a pensar en qué hacer con ellos. Aun cuando hemos pensado en cómo analizarlos antes de recolectarlos es probable que encontremos sorpresas que nos permitan dar respuestas a preguntas que no habíamos anticipado, o que encontremos límites a la aproximación que habíamos planteado. Es indispensable estar claros sobre lo que se quiere obtener, cómo se caracterizan los conjuntos de datos e identificar los tipos de variables que nuestros conjuntos de datos contienen.

Plataformas, tecnologías y herramientas

En gran parte el auge de la ciencia de datos se debe no solo a la mayor disponibilidad de datos sino también a la disponibilidad de infraestructura a bajo costo que nos permite almacenar, procesar, analizar los datos y presentar los resultados de tal forma que humanos y máquinas pueden tomar mejores decisiones más rápidamente.

Es un mundo en sí hablar del desarrollo de tecnología. De entre un tema tan amplio vamos a destacar los siguientes aspectos: infraestructura, lenguajes y algoritmos, y Data Ops. Esta sección busca dar una visión amplia para tener una base que permita entender lo que se encuentra en el mercado, tanto a nivel de necesidades como de soluciones.

Infraestructura

No es muy difícil recolectar tantos datos que ya no caben en una hoja de cálculo. Para algunos, ¡eso ya es suficiente razón para pensar que tienen un *big data*! En la práctica, definir *big data* en términos de volumen no es muy preciso. ¿Es lo que no cabe en un solo disco duro o es lo que requiere almacenamiento en clústeres de servidores? ¿O quizá es lo que no cabe en una cantidad pagable de memoria RAM, digamos para hoy en día más de 512 GB, y nos obliga a paralelizar nuestros algoritmos de análisis?

Los datos vienen en tres formas: estructurados, semiestructurados y no estructurados. Es más, una buena definición de *big data* no toma como base el volumen (la cantidad), sino la gran diversidad y falta de estructura (la calidad) de los datos para dar una definición (Letouzé, 2015). Este cambio de énfasis de cantidad a calidad se refleja en las soluciones tecnológicas que se han desarrollado tanto a nivel de almacenamiento y gestión de los datos, como en el procesamiento de datos y la gestión de dataductos.

Gestión de datos en big data

Quizá cuando pensamos en datos lo primero que viene a la mente son sistemas de archivos (las carpetas y las archivos con los que trabajamos a diario en nuestros ordenadores) y bases de datos. Ambas formas de almacenar información llegan a un límite cuando estamos usando un solo servidor, y para ambas hay métodos para extender el almacenaje a un conjunto (un clúster) de servidores, de tal forma que podemos acceder a los datos como si estuvieran en un solo sistema. Hay coherencia.

Un ejemplo conocido de sistemas de archivos distribuidos es el sistema de archivos Hadoop (HDFS), el componente de almacenaje por defecto de Hadoop. Encima de esto, Hadoop da varias formas de obtener acceso, y el más conocido es MapReduce. Mapear y reducir datos es necesario cuando hemos distribuido los archivos en múltiples servidores, porque cualquier búsqueda que hago en los datos, lo tengo que repetir en todos los servidores que son parte de mi clúster en forma paralela. Y el resultado tiene que ser un resultado agregado de los resultados de cada servidor.

HDFS y MapReduce representan una forma de gestionar muchos datos de gran diversidad. Pero la desventaja es que es costoso de implementar y relativamente lento. Cualquier búsqueda se necesita repartir sobre múltiples servidores. Otros sistemas de almacenaje distribuido que solo mencionamos son GFS (Google File System), MapR FS, CEPH, Lustre y Terragrid, entre otros.

Por otro lado está el desarrollo de las bases de datos NoSQL. A diferencia de bases de datos como Microsoft SQL, MySQL y PostgreSQL, estas no consisten de tablas de datos relacionadas entre sí con columnas llave. Las bases de datos NoSQL consisten de objetos con una llave única para identificar el objeto. Esto puede ser un documento JSON, en los *document stores.* Puede identificar un dato único, como en los *key-value stores,* puede ser una fila de un *column store* o un nodo único que está conectado con otros nodos a través de bordes en un *graph database*.

La gran ventaja que tienen las bases de datos NoSQL es que brindan más velocidad de acceso a los datos (dada una implementación correcta) y soluciones especializadas para aplicaciones geográficas. Además muchos tienen una mayor tolerancia a errores, aunque esa tolerancia tiene un precio en términos de consistencia. Casi todas las bases de datos NoSQL permite despliegue distribuido, lo que da la posibilidad de trabajar con datos a gran escala.

Por último, las base de datos SQL, que no han perdido su popularidad por varias razones. Porque se trabaja con ellas desde hace décadas y, por lo tanto, son fáciles de implementar y dan valor a un costo conocido y bajo. Además, por su estructura relacional imponen un orden, una calidad de los datos, lo cual tiene beneficios a largo plazo en términos de gestión de calidad y a corto plazo porque es más fácil acceder los datos para finalmente procesarlos y analizarlos.

Como regla simple es bueno tratar de buscar almacenar datos estructurados en forma estructurada y si no podemos evaluar trabajar con datos semiestructurados en una base de datos NoSQL apropiada. Los datos no estructurados (por ejemplo, el contenido de libros, grabaciones de vídeo, imágenes, contenido de e-mails, etc.) tienen retos más grandes para hacer preguntas que añaden valor.

Para todas estas soluciones tenemos servicios disponibles en plataformas de IAAS (infraestructura como servicio) de proveedores como Amazon, IBM, Microsoft y Google.

Los lenguajes de datos

Hace diez años la lengua franca de datos era SQL como se refleja en su nombre: es la abreviación de *standard query language*. Los analistas eran estadísticos que usaban aplicaciones y entornos de análisis especializados y costosos como SAS y SPSS. Pero en los últimos diez años hemos visto un proceso impresionante de democratización de análisis impulsado por dos lenguajes R y Python. Tanto que recientemente R superó a SPSS en su uso en publicaciones académicas (Muenchen, 2016).

Para tratar de explicar la popularidad de R y Python, y poner a ambos en el contexto de *big data* vamos a tratarlos por separado. Y al hacerlo somos muy conscientes de que cualquier organización que usa Hadoop va a tener Java como lenguaje de desarrollo principal. Para optimizar algoritmos en producción vamos a necesitar C++ o Julia. Pero el objetivo aquí es no entrar en detalles especializados pero describir el desarrollo a grandes rasgos. Y el punto de entrada para análisis para la mayoría de nosotros es R.

El desarrollo de R comenzó en 1997 cuando John Chambers comenzó a automatizar el uso de unos algoritmos escritos en FORTRAN. Como punto de referencia usó la descripción del lenguaje S, del cual R es una implementación. Hasta el día de hoy vemos que R es un lenguaje para agilizar el uso de algoritmos ya escritos por otros. Muchos de los algoritmos usados en R hasta el día de hoy están escritos en FORTRAN (muchos otros en C++). Y esa también es parte de la razón de su popularidad.

R es un lenguaje específico para estadísticas (*domain specific language*) y está enfocado en el uso interactivo para trabajar con conjuntos de datos. Esto es de enorme ventaja para el que no tiene un enfoque de programación, porque la sintaxis refleja la forma de trabajar y hacerse preguntas de alguien interesado en datos y no en programación. Es por esto que muchas empresas (AirBnB, StackOverflow) usan R para diseminar algoritmos dentro de sus empresas.

Por otro lado esta Python, un lenguaje general que tiene su lugar y fama dentro del mundo de la ciencia de datos por unas bibliotecas reconocidas como NumPy, SciPy, Pandas y SciKit. La razón principal para escocer Python para análisis es porque ya se conocía el lenguaje previamente, o porque hay razones puntuales de integración o de la disponibilidad de un algoritmo.

Más común es la posición de Python como el lenguaje para la construcción de dataductos. Por ejemplo, para construir flujos de trabajo automatizados tenemos

varios paquetes en Python disponible que son capaces de orquestar y monitorear tareas de ETL y análisis. Proyectos como Dask, Airflow, Pinball y Luigi son conocidos en este contexto.

DataOps

Hay otro desarrollo importante, que trae su nombre del desarrollo de *software*. DataOps es un término nuevo que busca describir el conjunto de mejores prácticas para mantener el desarrollo, testeo y despliegue de productos de datos dentro del margen de responsabilidad del autor. En otras palabras, a la hora de querer poner un producto de datos en producción no queremos tener que pasar por múltiples capas de protocolos para realizar el testeo y despliegue. Ser flexible y rápido en este desarrollo es un indicador importante de éxito.

Una parte importante del mundo de DataOps es trabajar bajo una arquitectura de microservicios. Esto significa que permitimos que la solución que tenemos operando en cualquier momento sea una colección fluida y mutable de soluciones puntuales. Todos los servicios están unidos de una forma que se denomina acoplamiento ligero. Si uno de los servicios deja de funcionar genera un aviso de que hay un error, pero el servicio como total sigue funcionando.

Una forma de imaginar esto es en un dataducto donde tenemos múltiples fuentes de datos, que cada uno pasa por diferentes pasos de preprocesamiento y transformación antes de alimentar un modelo predictivo. Podemos identificar la carga computacional de cada paso para asignar infraestructura a medida, y de esta forma optimizar el uso. Todas las plataformas IAAS de IBM, Microsoft y Amazon brindan servicios ya prefabricados que toman parte del dataducto y lo ejecutan. Sobre todo en el área de análisis hay una diversidad creciente de servicios disponibles, ya sean generales como *machine learning* (por ejemplo, Azure ML, Bluemix Predictive Analytics, Amazon ML) o específicos como análisis de sentimientos o reconocimiento visual.

Estos servicios los podemos integrar dentro de nuestro dataducto con servicios propios, por ejemplo, desplegados en contenedores Docker. Un contenedor Docker contiene el mínimo absoluto de carga de sistema operativo y permite crear servicios donde no instalamos todo un servidor en una máquina virtual, pero solo el *software* mínimo para ejecutar una tarea. Esto nos ahorra infraestructura en términos de tiempo de computación y uso de memoria. Además, nos permite crear dataductos resilientes de los cuales podemos incrementar la escala de aquellos servicios que requieren incremento.

Conclusiones y recomendaciones

En la última década hemos visto cómo los titanes de tecnología como Google, Facebook, Twitter, AirBnB y otras de este calibre han ido incorporando el análisis de datos de forma masiva para generar ventajas competitivas o incluso nuevos negocios.

También se están dando modelos como el de la empresa Stitch Fix (por poner un ejemplo), donde se parte de un modelo de negocios en el cual el uso de datos y la aplicación de algoritmos es el diferencial principal y la base sobre la cual se construye el negocio. En este caso, los datos permiten hacer recomendaciones que finalmente son curadas por humanos en un proceso iterativo hombre-máquina.

Además hemos visto cómo las grandes empresas y bancos han comenzado a incorporar a sus tradicionales equipos de *business intelligence* algunos procesos y unidades para descubrir cosas nuevas en base a datos, y así generar innovación y diferenciadores.

Esta es una tendencia que comenzó con las grandes empresas de tecnología y los *startups* tecnológicos enfocados en la temática. Ahora está siendo también incorporada por los grandes bancos y conglomerados. Pero está muy claro que en breve esta tendencia y forma de trabajo con los datos irá permeando hacia otros actores del sector privado incorporando empresas medianas, luego pequeñas, ONG y otros.

Las empresas que no comiencen a evaluar formas de generar valor en base a sus datos no serán competitivas en 2025. Nuevamente por si no quedó claro: ¡las empresas que no comiencen a evaluar formas de generar valor en base a sus datos no serán competitivas en 2025!

Por todo lo mencionado, los profesionales deben comenzar a educarse en datos. La tecnología es accesible y es clave que los profesionales puedan definir estrategias basadas en datos, implementarlas, operar equipos que cuentan con analistas, y asegurarse de que sus unidades interactúan con los equipos de ciencia de datos que si no existen ya en sus organizaciones, definitivamente se harán presentes en próximos años.

Comenzar no es complicado, es cosa de revisar temas de estadística, jugar con Excel y luego pasar a entender un poco de R. Esto además de prestar atención a lo que está pasando en distintos verticales y mercados con un ojo estadístico puede ser un buen primer paso.

Referencias bibliográficas

El Bassiti, L., R. Ajhoun (2013). "Towards an Innovation Management Framework". *International Journal of Innovation, Management and Technology,* 4(6): 551-559.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3), 37-54. Retrieved from http: //dx.doi.org/10.1609/aimag.v17i3.1230

Glass, H., Livesay, A., Preston, D. (2015). "Data driven innovation in new zealand". Innovation Partnership.

Kline, Stephen J., Rosenberg, N. (1986). "An Overview of Innovation". *The Positive Sum Strategy: Harnessing Technology for Economic Growth*. National Academies Press.

Morris, R. J. T., Truskowski, B. J. (2003). "The Evolution of Storage Systems". *IBM Systems Journal,* 42(2): 205-217.

Muenchen, B. (2016). "R Passes SAS in Scholarly Use (Finally)". Blog. *r4stats*, June 8.

Wickham, H., Grolemund, G. (2016). *R for Data Science*. O'Reilly.

Capítulo 9

La necesidad de normalizar el *big data*

Ray Walshe*, Jane Kernan**

Alrededor del mundo, las normas son utilizadas por todos los seres humanos en la vida cotidiana. Las normas hacen posible llevar a cabo nuestras actividades diarias ya que estas impactan en las comunicaciones, la tecnología, los medios, el cuidado de nuestra salud, la construcción, los elementos e incluso la energía.

Algunas normas han pasado la prueba del tiempo[1], estando presente por cientos o incluso miles de años. Por ejemplo, las normas que rigen el ancho de las vías de los ferrocarriles se basan en las ranuras hechas en la superficie de la tierra por las ruedas de los carruajes romanos. Siglos después, los dueños de las carretas descubrieron que el viaje era más cómodo si las ruedas se acoplaban a esas ranuras. Este enfoque se alargó y fue adoptado como una norma para los primeros carruajes y carretas, donde el espacio entre las vías se determinó por el tamaño de las ruedas. Esto ahorró mucho tiempo, dinero y esfuerzo al no tener que inventar o reinventar una nueva "norma" de hacer las cosas.

Las normas nos pueden servir en aspectos como:

- Confiabilidad. Aplicar las normas nos ayudan a garantizar la seguridad, la confiabilidad y el cuidado ambiental. Estos tipos de productos y servicios son percibidos como más seguros, y este sentimiento aumenta la confianza de los usuarios, las ventas y la adopción de nuevas tecnologías.
- Políticas gubernamentales y apoyo legislativo. Las normas se usan por los reguladores y legisladores para proteger los intereses del consumidor y para apoyar las políticas del gobierno. Las normas juegan un papel fundamental en las políticas de la Unión Europea en la generación del mercado único.
- Interoperabilidad. Los productos y servicios compatibles con las normas establecidas permiten a los dispositivos trabajar en conjunto.
- Beneficios de Negocios. Las normas proveen una base sólida en la cual se desarrollan nuevas tecnologías y se realzan las prácticas existentes.

* Profesor en Dublin City University (DCU) y Chairman del grupo ISO en estándares *big data*.
** Big Data Standards Researcher, Dublin City University.
1. http://www.etsi.org/standards/why-we-need-standards

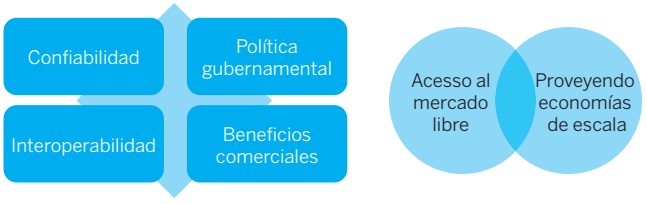
Las normas sirven específicamente para:

- Abrir el acceso al mercado.
- Proveer economías de escala.
- Motivar la innovación.
- Incrementar la consciencia de desarrollos e iniciativas técnicas.
- La elección del consumidor. Las normas proveen la base para una mayor variedad de nuevos productos con nuevas especificaciones y opciones.

Sin las normas tendríamos:

- Productos que no podrían funcionar o que serían peligrosos.
- Productos de calidad inferior o incompatibles con otros.
- Los clientes podrían ser cautivos de un solo proveedor.
- Los fabricantes inventarían sus propias normas incluso para los problemas más sencillos.

Figura 1. Las normas nos ayudan a acceder a nuevos mercados



La necesidad de una normalización internacional en la provisión de bienes y servicios a los consumidores debería ser obvio debido a los puntos previamente detallados. Esta determinación se basa en hechos y en muchos ejemplos de éxito sobre el desarrollo de las normas ya establecidas.

La tecnología de comunicación móvil GSM™ y sus sucesores (3G, 4G), los cuales fueron liderados por el Instituto Europeo de Normas para Telecomunicaciones (ETSI), es un buen ejemplo de las normas ya establecidas. El servicio de GSM idealmente se originó como una solución de telecomunicaciones para Europa, pero las tecnologías fueron rápidamente adoptadas y desplegadas por todo el mundo. Gracias a las normas establecidas, los viajeros internacionales pueden comunicarse y usar estos servicios comunes en cualquier parte del mundo.

Normalización en tecnologías de la información en la UE

La UE apoya un marco efectivo y coherente de normalización, el cual asegura que las normas son desarrolladas de una manera que apoyan las políticas de la UE y la competitividad dentro del mercado global.

Las regulaciones[2] en la normalización europea establecen el marco legal dentro del cual los diferentes actores en el sistema de normalización puedan operar. Estos actores son la Comisión Europea, las organizaciones europeas de normalización, la industria, las pequeñas y medianas empresas (pymes) y los actores sociales.

La Comisión está empoderada para identificar las especificaciones técnicas[3] de la tecnología de información y comunicación (TIC) que son aptas para usar de referencia en la contratación pública. Las autoridades públicas pueden por ende hacer uso del rango completo de las especificaciones cuando compran equipamiento informático, *software* o servicios tecnológicos, logrando así una mayor competencia y reduciendo el riesgo de quedar atrapados con sistemas propietarios. La Comisión apoya el trabajo de las tres organizaciones europeas de normalización, ETSI, CEN y CENELEC de forma económica.

ETSI, Instituto Europeo de Normas para las Telecomunicaciones

ETSI, el Instituto Europeo de Normas para Telecomunicaciones, produce normas[4] de aplicación global para las tecnologías de información y comunicación (TIC), incluyendo la red fija, móvil, radio, tecnologías de transmisión e Internet. Estas normas habilitan las tecnologías de las cuales tanto los negocios y la sociedad dependen. Las normas ETSI para el GSM™, DECT™, tarjetas inteligentes y firmas electrónicas han contribuido a revolucionar la vida moderna alrededor del mundo.

ETSI es una de tres organizaciones de normas europeas reconocida de manera oficial por la Unión Europea. Es una organización mundial, sin fines de lucro, con más de 800 miembros alrededor del mundo, procedentes de 66 países y cinco continentes.

2. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32012R1025
3. https://ec.europa.eu/growth/sectors/digital-economy/TIC-standardisation/TIC-technical-specifications_en
4. http://www.etsi.org/about/

ETSI está al frente de las tecnologías emergentes, intentando resolver los problemas técnicos que conduce la economía del futuro y mejorará la vida de las personas para la siguiente generación.

CEN, el Comité Europeo de Normalización[5]

CEN, el Comité Europeo de Normalización, es una asociación que componen los órganos de normalización nacional de 33 países europeos. CEN es también una de las tres organizaciones europeas de normalización (junto con CENELEC y ETSI) que han sido oficialmente reconocidas por la Unión Europea y por la Asociación Europea de Libre Comercio (AELC) como la responsable de desarrollar y definir las normas a nivel europeo.

CEN provee una plataforma para el desarrollo de las Normas Europeas y otros documentos técnicos en relación a varios tipos de productos, materiales, servicios y procesos. Apoya las actividades de normalización en relación a un amplio margen de campos y sectores, incluyendo: aire y espacio, químicos, construcción, productos del consumidor, defensa y seguridad, energía, el ambiente, comida y alimentación, salud y seguridad, cuidados de la salud, TIC, maquinaria, materiales, equipos de presión, servicios, vida inteligente, transporte y sistemas de paquetería.

CENELEC, el Comité Europeo de Normalización Electrotécnica[6]

CENELEC es el Comité Europeo de Normalización Electrotécnica y es responsable de la normalización en el campo de ingeniería electrotécnica. CENELEC prepara normas, las cuales ayudan a facilitar el comercio entre países, crear nuevos mercados, reducir costes de litigios y apoyar el desarrollo de un mercado europeo único. CENELEC crea un acceso al mercado a nivel europeo y también a nivel internacional, adoptando normas internacionales donde sea posible, mediante sus colaboraciones con la Comisión Electrotécnica Internacional (CEI)[7], bajo el Acuerdo de Dresden.

En la economía global, CENELEC fomenta la innovación y competitividad, haciendo que la tecnología esté disponible para las industrias mediante la producción de normas. Los miembros de CENELEC, sus expertos, las federaciones de industrias y consumidores ayudan a crear las Normas Europeas para fomentar

5. https://www.cen.eu/Pages/default.aspx
6. http://www.cenelec.eu/
7. https://www.cenelec.eu/aboutcenelec/whoweare/globalpartners/iec.html

el desarrollo tecnológico. Esta labor asegura la interoperabilidad y garantiza la seguridad y salud de los consumidores y provee protección para el medio ambiente. Al designarse como una Organización de Normas Europeas por la Comisión Europea, CENELEC es descrita como una organización técnica sin fines de lucro establecida bajo la ley belga. Fue creada en el año 1973 como resultado de la fusión entre dos organizaciones europeas previas: CENELCOM y CENEL.

La Plataforma Multilateral Europea de Normalización de las TIC

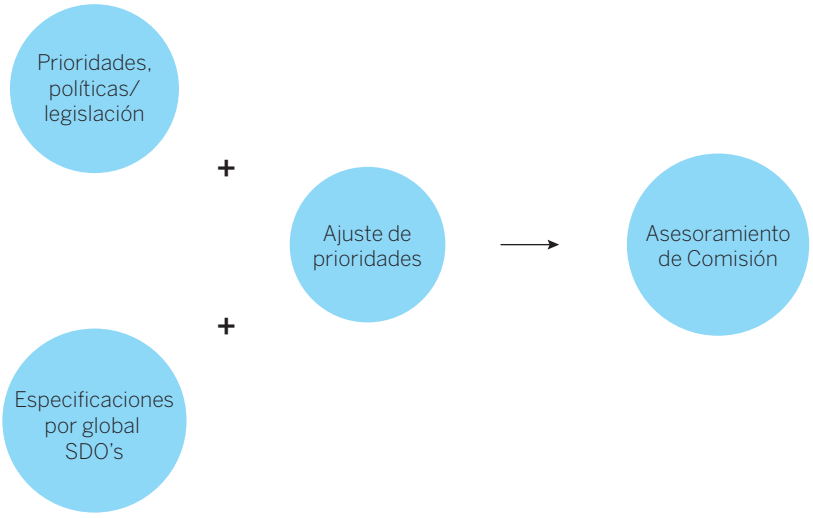
La Plataforma Multilateral Europea (MSP)[8] de Normalización de las TIC fue establecida en 2011. Esta organización asesora a la Comisión sobre las cuestiones de aplicación de la política de normalización de las TIC. Esto incluye la fijación de prioridades en apoyo de la legislación y las políticas, y la determinación de especificaciones elaboradas por las organizaciones mundiales de normalización de las TIC. La Plataforma Multilateral aborda los siguientes temas:

- El futuro potencial de las necesidades de normalización de las TIC.
- Las especificaciones técnicas para adquisiciones públicas.
- La cooperación entre organizaciones que establecen normas TIC.
- Una perspectiva general multianual de las necesidades de actividades de normalización preliminar o complementarias a las TIC en apoyo a las actividades políticas de la UE (Plan Continuo de Normalización de las TIC[9]).

La MSP está compuesta por representantes de las autoridades nacionales de los Estados miembros de la UE y países AELC. Esto incluye los cuerpos europeos e internacionales de normalización de las TIC y organizaciones interesadas que representan a la industria, empresas pequeñas y medianas, consumidores, etc. La MSP se reúne cuatro veces por año y es codirigida por la Comisión Europea de Dirección General del Mercado Interior[10], Industria, Emprendimiento, PYMES y CONNECT[11].

8. http://ec.europa.eu/digital-agenda/european-multi-stakeholder-platform-TIC-standardisation
9. https://ec.europa.eu/digital-single-market/en/rolling-plan-TIC-standardisation
10. http://ec.europa.eu/growth/about-us/index_en.htm
11. http://ec.europa.eu/dgs/connect/

Figura 2. Comisión Europea: Plataforma Multilateral



La Plataforma también asesora en la elaboración e implementación del Plan Continuo de Normalización de las TIC[12].

El Plan Continuo (PM) provee una perspectiva general multianual de las necesidades de las actividades de normalización de las TIC, tanto preliminares o complementarias en apoyo a las actividades políticas de la UE.

El Plan Continuo:

- Coloca la normalización dentro del contexto de la política.
- Identifica las prioridades de las políticas de la UE en actividades de normalización.
- Cubre las infraestructuras y normalizaciones horizontales TIC.

El Plan Continuo de Normalización de las TIC de 2016[13] cubre todas las actividades que pueden apoyar a la normalización y da prioridad a las acciones para la adopción e interoperabilidad de las TIC.

12. https://ec.europa.eu/growth/sectors/digital-economy/TIC-standardisation_en#rolling_plan_TIC_standardisation
13. http://ec.europa.eu/DocsRoom/documents/15783/attachments/1/translations

El Plan ofrece detalles en los contextos internacionales para cada política:

- Desafíos sociales: *e-Health* (salud en la red), accesibilidad de productos y servicios TIC, accesibilidad a la red, e-*Skills* (competencias electrónicas) y e-*Learning* (educación electrónica), comunicaciones de emergencia y *e-Calls* (llamadas electrónicas).
- Innovación para el mercado único digital: contratación electrónica, e-facturación, Internet y pagos móviles, lenguaje de informes empresariales extensibles (XBRL) y resolución de disputas en línea (ODR, por sus siglas en inglés).
- Crecimiento sostenible: redes inteligentes y medidores inteligentes, ciudades inteligentes, impacto medioambiental de las TIC, Servicio Europeo de Telepeaje (EETS) y Sistema de Transporte Inteligente (STI).
- Facilitadores de clave y seguridad: programación en la nube, datos (abiertos), E-*government* (gobierno electrónico), identificación electrónica y servicios de confianza incluyendo *e-Signatures* (firmas electrónicas), Identificación por radio frecuencia (RFID), "Internet de las cosas" (IoT) y e-Privacy (privacidad en los sitios eeb).

Casos de uso de *big data*

En junio de 2013, el Grupo de Trabajo Público de Big Data del Instituto Nacional de Normas y Tecnología (NIST) comenzó a formar una comunidad entre las partes interesadas de todos los sectores, incluyendo la industria, el mundo académico y el gobierno para desarrollar un consenso sobre las grandes definiciones de datos, taxonomías, arquitecturas de referencia seguras, requisitos de seguridad y privacidad y, en última instancia, un mapa de las normas establecidas. Parte del trabajo realizado por el grupo de trabajo identificó los Casos de uso de *big data* en la Plataforma de Interoperabilidad de Big Data del NIST: "Volumen 3, Casos de uso y requisitos generales", que servirían como ejemplos para ayudar a desarrollar una arquitectura de referencia para *big data* (NBDRA).

El NBD-PWG (Big Data Public Working Group) definió un caso de uso como "una aplicación típica declarada a un alto nivel con el propósito de extraer requisitos o comparar usos entre campos". Comenzaron recopilando casos de uso de información pública disponible para varios ejemplos de arquitectura de *big data*. Este proceso devolvió 51 casos de uso en nueve áreas amplias (es decir, dominios de aplicación). Esta lista no pretende ser exhaustiva y se considerarán otros dominios de aplicación. Cada ejemplo de la arquitectura de *big data* constituyó un caso de uso. Los nueve dominios de aplicación fueron los siguientes:

- Operación de Gobierno.
- Comercio.
- Defensa.
- Salud y Ciencias de la Vida.
- Aprendizaje Profundo y Medios Sociales.
- Ecosistema para la Investigación.
- Astronomía y Física.
- Ciencias de la Tierra, Medioambiente y Ciencia Polar.
- Energía.

Figura 3. Casos de uso de *big data*

Resúmenes de casos de uso

El enfoque inicial del subgrupo de Casos de Uso y Requisitos del NBD-PWG (Big Data Public Working Group) era formar una comunidad de interés de la industria, la academia y el gobierno, con el objetivo de desarrollar una lista consensuada de *big data* entre todas las partes interesadas. Esto incluyó la recolección y la comprensión de varios casos de uso de diversos dominios de aplicación. Las tareas asignadas al subgrupo incluían:

- Reunir aportes de todas las partes interesadas sobre los requerimientos de *big data*. Un objetivo que se convirtió en recopilación de casos de uso.
- Analizar / priorizar una lista de requerimientos generales difíciles derivados de casos de uso que pueden retrasar o impedir la adopción de *big data*.
- Desarrollar una lista completa de requerimientos de *big data*.

El informe fue producido a través de un proceso abierto de colaboración que incluyó conversaciones telefónicas semanales e intercambio de información utilizando el sistema de documentación NIST. Los casos de uso se organizaron en los nueve sectores / áreas (dominios de aplicación) descritos anteriormente (incluyendo entre paréntesis el número de casos de uso por área):

- Operación de gobierno (4): Administración Nacional de Archivos y Registros, Oficina del Censo.
- Comercio (8): Finanzas en la Nube, Respaldo en las Nubes, Mendeley (Citaciones), Netflix, Búsqueda Web, Materiales Digitales, Envío de Carga (ejemplo: UPS).
- Defensa (3): Sensores, Vigilancia de Imagen, Evaluación de la Situación.
- Salud y Ciencias de la Vida (10): Registros Médicos, Gráficos y Análisis Probabilísticos, Patología, Imagen Biológica, Genómica, Epidemiología, Modelos de Actividad Personal, Biodiversidad.
- Aprendizaje Profundo y Medios Sociales (6): Autoconducción de Automóviles, Georreferenciar Imágenes, Twitter, Crowd Sourcing, Redes de Ciencia, conjuntos de datos de referencia del NIST
- Ecosistema para la investigación (4): Metadatos, Colaboración, Traducción de Idiomas, Experimentos con Fuentes de Luz.
- Astronomía y Física (5): Encuestas de Cielo (y comparaciones con la simulación), Gran Colisionador de Hadrones (GCH) en la Organización Europea para la Investigación Nuclear (CERN), Belle II Acelerador de Partículas Japonés.
- Ciencia de la Tierra, Medio Ambiente y Polar (10): Dispersión de Radar en la Atmósfera, Terremoto, Océano, Observación de la Tierra, Dispersión de Radares de Hielo, Cartografía de Radar Terrestre, Conjuntos de Datos de Simulación Climática, Identificación Turbulenta Atmosférica, Biogeoquímica Subsuperficial (microbios a cuencas hidrográficas), la red AmeriFlux y el conjunto de datos de Gas FLUXNET.

La evolución de las normas de *big data*

Para lograr los objetivos del *big data* establecidos por las empresas y los consumidores se requerirá de la intercomunicación de múltiples sistemas, tecnologías,

legados y nuevos organismos. La integración tecnológica requiere normas para facilitar la interoperabilidad entre los componentes de valor del *big data*[14]. Por ejemplo, UIMA, OWL, PMML, RIF y XBRL son patrones clave de *software* que soportan la interoperabilidad del análisis de datos con un modelo de información sin plataforma, ontologías para modelos de información, modelos predictivos, reglas de negocio y un formato para reportes financieros. La comunidad de normas ha lanzado varias iniciativas y grupos de trabajo sobre el *big data*. En 2012, Cloud Security Alliance (Alianza de seguridad de la Nube) estableció un gran grupo de trabajo de datos con el objetivo de identificar técnicas escalables para la seguridad centrada en datos y problemas de privacidad. Se espera que la investigación del grupo aclare las mejores prácticas de seguridad y privacidad en *big data*, y también guíe a la industria y al gobierno en la adopción de mejores prácticas. El Instituto Nacional de Normas y Tecnología de Estados Unidos (NIST) inició sus actividades en *big data* con un taller en junio de 2012 y un año más tarde lanzó un grupo de trabajo público. El grupo de trabajo del NIST[15] tiene la intención de apoyar la adopción segura y efectiva del *big data* mediante el desarrollo de un consenso sobre definiciones, taxonomías, arquitecturas de referencia seguras y una hoja de ruta tecnológica para técnicas analíticas en *big data* e infraestructuras tecnológicas.

› Grupos públicos de trabajo de *big data* (NIST)

NIST desarrolló una Plataforma de Interoperabilidad de Big Data[16] que consta de siete volúmenes, cada uno de los cuales aborda un tema clave específico, resultante del trabajo del NBD-PWG (Big Data Public Working Group). Los siete volúmenes son los siguientes:

Volumen 1. Definiciones

El volumen Definiciones aborda los conceptos fundamentales necesarios para entender el nuevo paradigma para aplicaciones de datos, conocidas colectivamente

14. http://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000220001PDFE.pdf
15. https://www.nist.gov/el/cyber-physical-systems/big-data-pwg
16. Big Data Definitions: http://dx.doi.org/10.6028/NIST.SP.1500-1
Big Data Taxonomies: http://dx.doi.org/10.6028/NIST.SP.1500-2
Big Data Use Cases and Requirements: http://dx.doi.org/10.6028/NIST.SP.1500-3
Big Data Security and Privacy: http://dx.doi.org/10.6028/NIST.SP.1500-4
Big Data Architecture White Paper Survey: http://dx.doi.org/10.6028/NIST.SP.1500-5
Big Data Reference Architecture: http://dx.doi.org/10.6028/NIST.SP.1500-6
Big Data Standards Roadmap: http://dx.doi.org/10.6028/NIST.SP.1500-7

como *big data* y los procesos analíticos conocidos colectivamente como ciencia de datos. *Big data* ha generado muchas definiciones, pero en este volumen se concluye que se produce cuando la escala de los datos lleva a la necesidad de contar con un conjunto de recursos de computación y almacenamiento para proporcionar una gestión de datos efectiva a nivel de costes. La ciencia de datos combina diversas tecnologías, técnicas y teorías de diversos campos, principalmente relacionados con la informática y la estadística, para obtener conocimiento accionable a partir de los datos.

Volumen 2. Taxonomías

Las taxonomías fueron preparadas por el Subgrupo de Definiciones y Taxonomía del Grupo de Trabajo Público de Big Data del NIST (NBD-PWG - Big Data Public Working Group) para facilitar la comunicación y mejorar el entendimiento entre los interesados en *big data* al describir los componentes funcionales de la arquitectura de referencia de *big data* del NIST. Las funciones de nivel superior de la taxonomía son orquestadores de sistemas, proveedores de datos, proveedores de aplicaciones de *big data*, proveedores de marcos de *big data*, consumidores de datos, seguridad, privacidad y Administración. La taxonomía NBDRA tiene como objetivo describir los nuevos problemas en los sistemas de *big data*, pero no es una lista exhaustiva.

Volumen 3. Casos de Uso y Requerimientos Generales

El volumen 3 fue preparado por el Subgrupo de Casos de Uso y Requerimientos del Grupo de Trabajo Público de Big Data del NIST (NBD-PWG) para recopilar casos de uso y extraer requisitos. El Subgrupo desarrolló una plantilla de casos de uso con 26 campos que fueron completados por 51 usuarios en las siguientes áreas generales:

- Operación de Gobierno (4).
- Comercio (8).
- Defensa (3).
- Salud y Ciencias de la Vida (10).
- Aprendizaje Profundo (*deep learning*) y Medios Sociales (6).
- Ecosistema para la Investigación (4).
- Astronomía y Física (5).
- Ciencias de la Tierra, Medioambiente y Ciencia Polar (10).
- Energía (1).

Volumen 4. Seguridad y privacidad

Seguridad y privacidad fue preparado por el Subgrupo de Seguridad y Privacidad del Grupo de Trabajo Público de Big Data (NBD-PWG) para identificar problemas de seguridad y privacidad que son específicos de *big data*. Los dominios de aplicación de *big data* incluyen atención médica, descubrimiento de fármacos, seguros, finanzas, venta al por menor y muchos otros en los sectores privados y públicos. Entre los escenarios dentro de estos dominios de aplicación se encuentran los intercambios de salud, ensayos clínicos, fusiones y adquisiciones, telemetría de dispositivos, mercadotecnia y antipiratería internacional.

Volumen 5. Encuesta de Arquitectura (*White Paper*)

Encuesta de Arquitecturas (*White Paper*) fue preparado por el Subgrupo de Arquitectura de Referencia del Grupo de Trabajo Público de Big Data del NIST (NBD-PWG) para facilitar la comprensión de las complejidades operativas en *big data* y servir como una herramienta para desarrollar arquitecturas específicas del sistema usando un marco de referencia común. Este esfuerzo reveló una consistencia notable de la arquitectura de *big data*. Los temas más comunes que se citan a través de las arquitecturas estudiadas se describen a continuación.

- Manejo de *big data*:
  - Datos estructurados, semiestructurados y no estructurados.
  - Velocidad, variedad, volumen y variabilidad.
  - SQL y NoSQL.
  - Sistema distribuido de archivos.
- Analíticas de *big data*:
  - Descriptiva, predictiva y espacial.
  - Tiempo real.
  - Interactiva.
  - Analítica de lotes.
  - Reportes.
  - DashBoards.

- Infraestructura de *big data*:
  - Redes de datos en memoria.
  - Base de datos operacionales.
  - Base de datos analíticas.
  - Base de datos relacionales.
  - Archivos planos.
  - Sistema de gestión de contenidos.
  - Arquitectura horizontal escalable.

Volumen 6. Arquitectura de Referencia

El Subgrupo de Arquitectura de Referencia del Grupo de Trabajo Público de Big Data del NIST (NBD-PWG- Big Data Public Working Group) preparó este marco de interoperabilidad de *big data* del NIST para proveer un modelo conceptual neutro de tecnología y examinar las cuestiones relacionadas con el mismo. El modelo conceptual, denominado NIST Arquitectura de Referencia de Big Data (NBDRA), fue elaborado mediante el examen de las arquitecturas de *big data* públicamente disponibles que representan diversos enfoques y productos. Los *inputs* de los otros subgrupos NBD-PWG (Big Data Public Working Group) también se incorporaron en la creación de la NBDRA. Es aplicable a una variedad de entornos empresariales, incluyendo sistemas empresariales estrechamente integrados, así como industrias verticales que dependen de la cooperación entre partes interesadas independientes. La NBDRA captura las dos cadenas de valor económico conocidas en *big data*: información, donde el valor se crea mediante la recopilación de datos, la integración, el análisis y la aplicación de los resultados a servicios basados en datos y tecnología de la información (TI), donde el valor es creado a través de redes, infraestructura, plataformas y herramientas en apoyo de aplicaciones verticales basadas en datos.

Volumen 7. Hoja de Ruta de Normas

La Hoja de Ruta de Normas resume los resultados de los otros subgrupos NBD-PWG (Big Data Public Working Group) (presentados en detalle en los otros volúmenes de esta serie) y presenta los trabajos del Subgrupo de Hoja de Ruta Tecnológica de NBD-PWG (Big Data Public Working Group). En la primera fase del desarrollo, el Subgrupo de la Hoja de Ruta Tecnológica de la NBD-PWG (Big Data Public Working Group) investigó las normas existentes que se relacionan con *big data* e identificó categorías generales donde estas normas tenían ciertas brechas.

Comité de Administración de Datos y Normas de Intercambio (SC32) de la ISO/IEC JTC1

El Comité de Administración de Datos y Normas de Intercambio (SC32) de la ISO/IEC JTC1 realizó un estudio sobre la próxima generación de análisis y *big data*[17]. El W3C ha creado varios grupos sobre diferentes aspectos del *big data*.

En el Plenario de SC32 de junio de 2012 en Berlín, Jim Melton, el presidente de SC32, nombró un comité *ad hoc* para los cuatro grupos de trabajo SC32: WG1 Negocios Electrónicos, WG2 Metadatos, WG3 Lenguajes de Bases de datos y WG4 Multimedia.

Dado que la solicitud original de JTC1 hizo referencia a un informe de Gartner Group, es útil revisar su visión de tecnologías estratégicas para 2012 referente a analítica y *big data*:

- Analíticas de la próxima generación
  - Las analíticas están creciendo junto a tres dimensiones claves:
    - Desde la analítica tradicional sin conexión a la analítica en línea. Este ha sido el enfoque de muchos esfuerzos en el pasado y seguirá siendo un enfoque importante para la analítica.
    - Desde el análisis de datos históricos para explicar lo que sucedió al análisis de los datos históricos y en tiempo real de múltiples sistemas para simular y predecir el futuro.
    - En los próximos tres años, los análisis madurarán a lo largo de una tercera dimensión, desde datos estructurados y sencillos analizados por individuos hasta el análisis de información compleja de muchos tipos (texto, vídeo, etc.) de muchos sistemas que apoyan un proceso de decisión colaborativo que atrae a múltiples personas para analizar y proponer un sin número de ideas y tomar decisiones.
- *Big data*: el tamaño, la complejidad de los formatos y la velocidad de entrega superan las capacidades de las tecnologías tradicionales de gestión de datos; se requiere el uso de nuevas tecnologías simplemente para manejar el volumen de la información. Muchas nuevas tecnologías están surgiendo, con el potencial de ser disruptoras (por ejemplo, DBMS, sistema de gestión de bases de datos). La analítica se ha convertido en una importante aplicación para

17. http://www.jtc1sc32.org/doc/N2351-2400/32N2388b-report_SG_big_data_analytics.pdf

el almacenamiento de datos, con el uso de MapReduce "programación de Google para la computación distribuida" fuera y dentro del DBMS, y el uso de almacenes de información (*data-marts*) de autoservicio. Una implicación importante de *big data* es que en el futuro los usuarios no podrán poner toda la información útil en un solo almacén de datos. Los almacenes de datos lógicos que reúnen información de múltiples fuentes según sea necesario reemplazarán al modelo de almacén de datos únicos.

Normas de Trabajo de Big Data ISO

Normas de Trabajo de Big Data ISO IEC JTC 1 WG9

Los ecosistemas normativos se requieren para cumplir con el procesamiento analítico independientemente de las necesidades del conjunto de datos en relación con las características de las V (volumen, velocidad, variedad, etc.), las plataformas de computación subyacentes y cómo se implementan las herramientas y técnicas de análisis de *big data*. La arquitectura de plataforma de datos unificados apoyará la estrategia de *big data* a través de la gestión de información, el análisis y la tecnología de búsqueda.

Un ecosistema basado en las normas proporciona plataformas neutras de proveedores, tecnología e infraestructura que permitirán a los científicos e investigadores de datos compartir y reutilizar herramientas y técnicas de análisis interoperables. El WG 9 trabaja con académicos, con la industria, con el gobierno y con otras partes interesadas para comprender las necesidades y fomentar un ecosistema de normas *big data*.

El WG 9 tiene un enfoque técnico de tres vías para lograr definir y desarrollar las normas de este ecosistema:

a) Identificar las normas de la arquitectura de referencia de *big data* (RA): este enfoque ya ha sido desarrollado en ISO / IEC 20547 para identificar los componentes de la RA generales y sus descripciones de interfaces.

b) Identificar la normativa en arquitecturas de referencia para el *big data*: este sería un nuevo proyecto para investigar cómo fluyen los datos entre los componentes de la RA y definir normas de interfaces para dichas interacciones. El objetivo es utilizar estas interfaces de normas validadas para crear aplicaciones de *big data*.

c) Identificar las normas de herramientas de administración de *big data*: este sería otro nuevo proyecto, para investigar cómo la recopilación de herramientas analíticas y recursos de cómputo pueden ser administrados

eficiente y eficazmente para permitir el desarrollo de normativa informática empresarial a nivel de *big data*.

El WG 9 está activo y actualmente reclutando a expertos de *big data* y promoviendo el desarrollo de la normalización de *big data* de JTC 1 organizando talleres antes de las reuniones sobre las normas del WG 9. Algunas de estas normas que se están desarrollando quedan recogidas en esta tabla.

Título	Editor jefe	Editores asociados
ISO/IEC TR 20547-1, Information technology. Big Data Reference Architecture. Part 1: Framework and Application Process	David BOYD (US)	Suwook HA (KR), Ray WALSHE (IE)
ISO/IEC TR 20547-2, Information technology. Big Data Reference Architecture. Part 2: Use Cases and Derived Requirements	Ray WALSHE (IE)	Suwook HA (KR)
ISO/IEC 20547-3, Information technology. Big Data Reference Architecture. Part 3: Reference Architecture	Ray WALSHE (IE)	David BOYD (US), Liang Guang (CN), Toshihiro Suzuki (JP)
ISO/IEC TR 20547-5, Information technology. Big Data Reference Architecture. Part 5: Standards Roadmap	David BOYD (US)	Toshihiro SUZUKI (JP), Abdellatif Benjelloun TOUIMI (UK)

Tendencias y proyecciones futuras de normativa *big data*[18]

Información del sector público, datos abiertos y big data

Con la creciente cantidad de datos (a menudo referidos bajo la noción de *big data*) y la creciente cantidad de datos abiertos (*open data*), la interoperabilidad se convierte cada vez más en un problema clave para aprovechar el valor de estos datos. La normalización a diferentes niveles (como los esquemas de metadatos, los formatos de representación de datos y las condiciones de concesión de licencias de *open data*) es esencial para permitir una amplia integración de datos, intercambio de datos e interoperabilidad con el objetivo general de fomentar la innovación sobre la base de los datos. Esto se refiere a todos los tipos

18. http://ec.europa.eu/DocsRoom/documents/15783/attachments/1/translations/en/renditions/pdf

de datos (multilingües), incluyendo tanto datos estructurados como no estructurados, así como datos de diferentes dominios tan diversos como datos geoespaciales, datos estadísticos, datos meteorológicos y datos de investigación, por citar sólo algunos.

Actividades normativas que se desarrollan actualmente en Europa

En general, debe fomentarse la aplicación de formatos y protocolos normalizados y compartidos para recopilar y procesar datos de diferentes fuentes de manera coherente e interoperable entre sectores y mercados verticales. Por ejemplo, esto se aplica en proyectos de Investigación, Desarrollo e Innovación I+D+i, en el Portal de Open Data de la UE y el Portal Europeo de Open Data. Los estudios realizados en nombre de la Comisión Europea muestran que las empresas y los ciudadanos se enfrentan a dificultades para encontrar y reutilizar la información del sector público. En su comunicado sobre *open data* del 12 de diciembre de 2011, la Comisión Europea afirma que la disponibilidad de la información en un formato legible por máquina, así como la posibilidad de generar una capa de metadatos consensuados, podría facilitar el intercambio de datos, la interoperabilidad y el valor para su reutilización. Surge así la iniciativa del Vocabulario de Catálogos de Datos (DCAT) en colaboración con FIWARE (Plataformas de Middleware para el Internet Futuro). DCAT ha sido desarrollado como un proyecto común del Programa ISA (Soluciones de interoperabilidad para las Administraciones públicas europeas), la Oficina de Publicaciones (OP) y la Dirección General de Redes de Comunicación, Contenido y Tecnología - Comisión Europea (DG CONNECT) para describir los catálogos de datos sectoriales y conjuntos de datos, y para promover la especificación que se usará en los portales de datos por toda Europa. Al acordar un perfil común de aplicación y promoverlo a los Estados miembros, se incrementará sustancialmente la interoperabilidad entre los catálogos de datos y el intercambio de datos entre los Estados miembros. El DCAT-AP (Perfil de Aplicación para Portales de Datos en la Unión Europea) es la especificación que utilizará el Portal Paneuropeo de Open Data, que forma parte de la infraestructura del mecanismo "Connecting Europe". FIWARE CKAN es una solución de código abierto para la publicación, gestión y consumo de *open data*. FIWARE NGSI es una interfaz de programación de aplicaciones (API) que proporciona un medio simple y ligero para recopilar, publicar, consultar y suscribirse a información contextual[19].

19. http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/final_version_study_psi.docx
http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive_proposal/2012/open_data.pdf
http://ec.europa.eu/digital-agenda/overview-2003-psi-directive
http://ec.europa.eu/information_society/policy/psi/rules/eu/index_en.htm

El rastreo de las normas pertinentes y existentes para una serie de áreas grandes de datos sería beneficioso. Además, podría ser útil identificar a los grupos de industrias europeas que sean lo suficientemente homogéneas en sus actividades para desarrollar normativa de datos. En particular, en el contexto de *open data* es necesario abordar los temas de la procedencia de los datos y la concesión de licencias (por ejemplo, el potencial de las licencias legibles por máquina) tal como lo indica la directiva de la Oficina de Información del Sector Público (OPSI). Esta directiva (2013/37/UE) fomenta el uso de licencias normalizadas que deben estar disponibles en formato digital y procesarse electrónicamente (artículo 8 (2)). Además, la directiva fomenta el uso de licencias abiertas disponibles en línea, que eventualmente se convertirán en una práctica común en la UE (Preámbulo 26). Además, para ayudar a los Estados miembros en la transposición de las disposiciones revisadas, la Comisión adoptó unas directrices que, entre otras cosas, recomendaban el uso de tales licencias abiertas para la Reutilización de la Información del Sector Público (RISP).

http://ec.europa.eu/information_society/policy/psi/docs/pdfs/report/final_version_study_psi.docx for an overview and
http://ec.europa.eu/information_society/policy/psi/docs/pdfs/opendata2012/open_data_communication/en.pdf

Capítulo 10

Administración predictiva y proactiva vinculada al *big data*

Juan Muñoz*

Introducción

En este capítulo se hace una revisión del concepto de *big data* y su utilidad para el establecimiento de metas y la toma de decisiones bajo los enfoques de administración predictiva y proactiva. También se analizan algunos aspectos que deben ser tomados en cuenta para obtener un valor informativo de estos tipos de fuentes de datos, finalmente se proporcionan algunos ejemplos de aplicaciones de *big data* que pueden servir como detonadores de nuevas ideas para su aplicación.

El valor de los datos para la administración predictiva y proactiva

La administración predictiva y proactiva son dos estrategias complementarias. La primera basa sus métodos en proyecciones sobre el futuro planeando en consecuencia; la segunda monitorea los posibles cambios que habrán de ocurrir en el entorno para anticiparse y minimizar los riesgos que pueden materializarse por dichos cambios.

La administración predictiva utiliza modelos matemáticos para formular posibles escenarios del futuro. Dentro de un contexto de administración predictiva, la información que arrojan los modelos se utiliza para realizar la planeación estratégica y establecer metas para apoyar a la misión de la organización. A partir del establecimiento de las metas, el paso siguiente es el diseño de estrategias y proyectos para alcanzarlos realizando posteriormente la fase de implementación.

* Doctor en Ciencias de la Computación y director de Planeación y Normatividad Informática en el Instituto Nacional de Estadística y Geografía (INEGI, México).

Los escenarios son dinámicos, los valores de los parámetros que sirven como entrada de dichos modelos pueden cambiar a lo largo del tiempo conforme se presentan diferentes situaciones. Los modelos matemáticos reaccionan acorde a los datos que se ingresan como parámetros y sus resultados dependerán de los valores que reciban.

De acuerdo al diccionario Merriam-Webster (n. d.), los datos son la base de los cálculos o de las mediciones. Son hechos o información utilizada para calcular, analizar o planear algo. Los datos permiten determinar hechos expresados como cantidades que se localizan en algún momento dado en cierto lugar y con base en ellas tomar decisiones.

Tomar los datos como puntos aislados en el tiempo permite reaccionar y hacer ajustes para corregir situaciones que ya han ocurrido, pero tiene muchas limitaciones cuando tratamos de emplear un enfoque proactivo. Realizar varias mediciones de un mismo hecho conforme transcurre el tiempo permite generar modelos en los que se describen patrones y tendencias.

Si algún hecho muestra un comportamiento regular, entonces es factible reconocer patrones y predecir el comportamiento que tendrá en el futuro. Si los factores (datos) que se toman como parámetros continúan con el comportamiento descrito por el patrón, entonces podremos obtener no solamente la magnitud actual, sino calcular la magnitud esperada en el futuro. De esta manera la planeación predictiva plantea sus objetivos para cierto horizonte en el tiempo anticipándose a hechos que probablemente ocurrirán.

La anticipación es un elemento fundamental de la administración predictiva. Antes de construir alguna obra, desarrollar algún negocio o implementar algún servicio se debe conocer la demanda esperada, es decir, las personas que van a hacer uso, consumirlo o simplemente beneficiarse. También deben tomarse en cuenta los recursos con que se cuenta y se espera contar durante el desarrollo del proyecto, así como las restricciones que habrán de afrontarse.

Los datos representan esos hechos que deben ser considerados al tomar decisiones, bajo el enfoque de la administración predictiva no solamente se tomará el valor actual de un hecho, sino la proyección de su valor futuro.

Para entender mejor este tipo de razonamiento podemos considerar como ejemplo la construcción de cierta obra. Si la población de una ciudad es de un

millón de habitantes y sabemos que el crecimiento promedio anual durante los últimos diez años ha sido del 1,4%, una obra que tenga una vida útil de 30 años y que pretenda atender a un diez por ciento de la población deberá considerar ser útil para atender a un poco más de 150.000 usuarios, si la tasa de crecimiento y otras variables relacionadas se mantienen constante (ver tabla 1: Ejemplo de proyección de capacidad requerida).

Tabla 1. Ejemplo de proyección de capacidad requerida

Tasa de crecimiento anual	Año	Habitantes	Capacidad requerida
	0	1.000.000	100.000
1,4%	5	1.071.988	107.199
Porcentaje de población a atender	10	1.149.157	114.916
	15	1.231.883	123.188
	20	1.320.563	132.056
10,0%	25	1.415.627	141.563
	30	**1.517.535**	**151.753**

El ejemplo anterior no solamente se aplica a la demanda en relación al crecimiento demográfico, sino a hechos sociales y económicos como el nivel de educación de la población, el valor de la moneda, etc. Si el comportamiento de un hecho de interés es conveniente para los objetivos que han sido establecidos, probablemente convendrá cuidar que los parámetros continúen más o menos igual, y en caso de que no lo sea, se buscará alterarlo modificando sus factores de entrada. Esto hacen los gobiernos cuando deciden establecer alguna política pública.

Una vez que la administración predictiva ha establecido las metas, la administración proactiva instituye una vigilancia sobre los datos que se utilizaron como parámetros de la planeación. Al detectar cualquier cambio en los parámetros que pueda traer alguna consecuencia en los resultados esperados, la administración proactiva realiza acciones para movilizar recursos y hacer ajustes para reducir riesgos o aumentar las ventajas que faciliten conservar la viabilidad de conseguir los objetivos establecidos.

Bajo la lógica de la administración proactiva, los datos se constituyen como parámetros sobre los que se debe decidir y actuar. Los datos deben tener ciertas características para que puedan ser útiles en el proceso de toma decisiones:

- **Desagregación**. Deben tener un nivel de detalle suficiente (pero no excesivo) para satisfacer las necesidades de conocimiento. Esta característica generalmente se relaciona con la granularidad, la cual trata del nivel al que se pueden detallar los componentes que integran un ente de mayor tamaño, por ejemplo, un país se puede dividir en estados, municipios, localidades, etc.; una mayor granularidad permite una visión con mayor detalle e implica que hay una mayor desagregación.
- **Accesibilidad**. Deben encontrarse al alcance de quien requiere utilizarlo. Esta disponibilidad va más allá de que se conozca su existencia; implica también que los costos, conocimientos, capacidades técnicas, recursos requeridos para su uso, etc., sean aceptables y disponibles para quien los requiere.
- **Relevancia**. Deben tener una relación con el hecho o materia sobre la que se está tomando una decisión. Una mayor relevancia implica que un cambio, aun cuando sea pequeño, tendrá un afecto que podrá ser percibido.
- **Precisión**. La cantidad que representan debe ser clara y lo más cercana posible a la realidad. También implica contar con toda la información necesaria interpretarlos correctamente, estos son sus metadatos estructurales. También es necesario contar con los metadatos de referencia que permitan conocer los métodos y otros aspectos relevantes sobre la forma en que fueron producidos; en especial deberá presentarse información adicional suficiente para determinar los errores de medición a los que estén sujetos.
- **Oportunidad**. Idealmente deben presentar el estado del hecho que miden lo más cercanamente posible a su ocurrencia y al tiempo en que se requiere tomar la decisión. Cabe mencionar, que la oportunidad es afectada también por la variabilidad, en general cuando un dato conserva su valor por un tiempo largo y sus variaciones no son abruptas, entonces es posible flexibilizar un poco más los tiempos de medición sin que se pierda oportunidad.
- **Comparabilidad**. Deben ser comparables con otros datos similares, esto implica en muchos casos que exista una estandarización en su conceptualización, las escalas y los métodos de medición. La comparabilidad permite además que puedan combinarse con otros datos para producir nueva información.
- **Autonomía**. Deben medir fielmente el hecho de manera independiente a cualquier interferencia o tipo de interés (económico, político, etc.).

La calidad del dato generalmente se relaciona con el grado en que satisface cada una de las características mencionadas, pero depende de muchos factores, entre otros: las capacidades y habilidades de quien lo produce y de quien lo recibe, las metodologías e instrumentos utilizados para su medición y recolección, las fuentes de las que se obtiene, el medio a través del cual se comunica, la forma en

que se presenta, etc. Un conjunto de datos de buena calidad reduce la incertidumbre en la toma de decisiones.

Con el tiempo, las tecnologías con que se producen, recolectan, integran, procesan, visualizan, almacenan y transportan los datos han evolucionado favoreciendo un aumento en su calidad al incidir en el grado de cumplimiento de una o más de las características ya mencionadas.

Tradicionalmente la producción de datos de calidad se percibe como una actividad exprofeso que debe controlarse desde el momento en que se conceptualiza el dato. Pero, las nuevas tecnologías aparejadas con la conexión de dispositivos y sistemas de información a Internet han creado un universo digital en expansión con millones de datos que pueden ser explotados con diferentes fines. Aprovechar este universo requiere de nuevos enfoques y de la atención de varios aspectos conceptuales, metodológicos, matemáticos, tecnológicos, éticos, legales, culturales, etc.

El Grupo Asesor Independiente de Expertos Sobre una Revolución de Datos para el Desarrollo Sostenible del secretario general de las Naciones Unidas (IEAG) introduce el concepto llamado "revolución de los datos" (IEAGE, 2014), definido como la explosión en el volumen de datos ocasionada por: la velocidad con que se producen; la cantidad de nuevos productores; la facilidad para su diseminación y la variedad de cosas en que hay datos que proceden de nuevas tecnologías (tales como teléfonos móviles, "Internet de las cosas" y otras fuentes, así como los datos cualitativos y de percepciones generados por los ciudadanos).

El IEAG encuentra que la "revolución de los datos" representa una oportunidad para mejorar los datos que son esenciales para la toma de decisiones, la rendición de cuentas y la resolución de retos de desarrollo.

La importancia de *big data* en la generación de información para la toma de decisiones

Cox y Ellsworth (1997) describieron por primera vez el término de *big data* como un problema en el cual los grandes conjuntos de datos no caben en la memoria principal o en los discos locales o remotos de sistema. La cantidad de grandes conjuntos de datos va incrementándose con el tiempo conforme se desarrollan capacidades para producir información digital y se incorporan a Internet nuevos sistemas y dispositivos que alimentan con datos un universo digital en expansión.

Desde el año 2000 Leyman y Varian (2003) comenzaron a publicar estudios en los que se hace la estimación del crecimiento anual de la nueva información contenida en medios físicos (principalmente digitales) de almacenamiento. Sus estimaciones arrojaron que el crecimiento de la información entre 1999 y 2002 tuvo una tasa anual del 30%, representando un total de 5 *exabytes* al final de ese periodo. Tomando en cuenta una población de 6.300 millones de habitantes, se estima que en promedio cada persona produciría 800 MB de información. Aunque los resultados de estos estudios puedan ser discutibles y sujetos a diferentes acotaciones e interpretaciones, no queda duda que la información digital que se produce es cada vez mayor.

Después de la conceptualización inicial de *big data* realizada por Cox y Ellsworth, han aparecido nuevas definiciones para el término de *big data*. La definición de *big data* realizada por Gartner ha sido ampliamente utilizada en el medio de las tecnologías de información, describe el concepto como un conjunto de activos de información que se caracterizan por un gran volumen, una alta velocidad de producción y una extensa variedad de formatos que exige formas rentables e innovadoras para su procesamiento y que permiten una visión más amplia: una toma de decisiones mejorada; así como una automatización de procesos. En realidad, aún no existe un consenso generalmente aceptado sobre la definición de este concepto.

Parte de la dificultad para llegar a un acuerdo sobre qué significa el término *big data* se debe a que hace referencia a diferentes tipos de datos con diferentes características. Por ejemplo, si se toma como base la fuente a partir de la cual se producen y la forma en que deben ser procesados, es posible llegar a una clasificación como la siguiente:

- **Datos producidos por medidores y sensores inteligentes**. Esto incluye cámaras de tráfico, dispositivos GPS, medidores de consumo eléctrico, relojes inteligentes, teléfonos inteligentes, etc. Aquí se encuentran datos que se encuentran más o menos estructurados y que en lo general se generan con gran velocidad, por lo que, aunque el tamaño individual de la información que representan no ocupa mucho espacio, el conjunto de observaciones que se producen puede llegar a ser de miles o millones de registros para un corto periodo de tiempo y es común que las mediciones presenten errores o interferencias.
- **Interacciones sociales.** Corresponde a conversaciones y publicaciones en redes sociales y otros sistemas que facilitan la interacción y colaboración entre personas. Las actividades en este tipo de medios sociales electrónicos

producen datos cualitativos que pueden incluir opiniones, posturas, comentarios, expresiones de estado de ánimo, etc. La naturaleza libre tanto en estructura como en temática de estos contenidos agrega cierta dificultad para estudiarlos e interpretarlos y generalmente requieren el empleo de métodos de análisis de lenguaje natural en conjunto con otras técnicas de análisis de contenidos pertenecientes al área de la ciencia de datos. Asimismo, el análisis de metadatos de las estructuras en las que se encuentran estos contenidos puede arrojar información valiosa.

- **Transacciones de negocios.** Comprende los movimientos que se dan por efecto de la realización de actividades económicas soportadas por medios electrónicos, como los relacionados utilizando tarjetas de crédito o débito, las ventas procesadas con cajas registradoras, los registros empleados para efectuar la facturación de teléfonos celulares, etc. Estos datos en lo general presentan estructuras más o menos uniformes para los registros que produce alguna unidad de negocios. La frecuencia con que se producen cada una de las transacciones individuales que registran puede ser muy variable y describir patrones regulares o irregulares que en sí mismos pueden significar información valiosa.
- **Archivos electrónicos.** Se trata de documentos que se encuentran disponibles en formatos electrónicos como archivos PDF, páginas de Internet, vídeos, audios, imágenes de satélite, medios de difusión digitales, etc. Su velocidad de producción puede no ser muy alta, pero representan un gran reto para su recolección y procesamiento ya que sus contenidos y estructuras son muy diversas y pueden llegar a ocupar volúmenes importantes de almacenamiento.
- **Medios de difusión.** Contempla las corrientes de datos que generan las fuentes de vídeo y audio digital que se producen en tiempo real. Actualmente su análisis representa grandes retos debido a que:

  - En general, las estructuras que se utilizan para contenerlos no se orienta a facilitar el análisis de sus contenidos sino a su presentación visual o auditiva.
  - Se genera una gran cantidad de información en periodos de tiempo muy cortos, que idealmente debería ser procesada conforme es producida.
  - La velocidad y riqueza de información requeridas para hacer la representación fiel del sonido y el vídeo que contienen provoca que el volumen que ocupan sea muy alto.

Aunque existen muchas dificultades para su procesamiento, es el tipo de datos que está creciendo con mayor velocidad, por lo que diferentes comunidades interesadas en la explotación de *big data* se encuentran desarrollando tecnologías y metodologías para su procesamiento y análisis.

Esta diversidad de tipos presenta una rica gama de posibilidades en la obtención de información valiosa, así como retos traducidos principalmente como dificultades técnicas y metodológicas para hacer posible su utilización.

Aunque inicialmente, los grandes conjuntos de datos fueron percibidos como un problema, diferentes instituciones de los sectores público y privado comenzaron a reconocer el potencial de *big data* para producir valor a través de su procesamiento. Las oficinas de diferentes naciones y organismos internacionales encargadas de la producción de estadísticas oficiales utilizadas para la toma de decisiones y la definición de políticas públicas reconocen en el documento: *What Does "Big Data" mean for official statistics* (UNECE/ECE, 2013), que *big data* representa el potencial de producir estadísticas más oportunas que las fuentes tradicionales (censos, encuestas y registros administrativos) de estadísticas oficiales.

En general, existe una brecha entre lo que se conoce de un dato y el valor informativo que este puede llegar a representar. El valor informativo que se determina para el dato en lo general se relaciona con los objetivos que se persiguen, así como con el enfoque, los conocimientos y las técnicas que se utilizan para analizarlos.

Para convertir un conjunto de datos en conocimiento es necesario contar con recursos tecnológicos físicos y lógicos (*hardware* y *software*), técnicas, metodologías y expertos adecuados para la tarea. Debido a que *big data* representa retos muy específicos, cada uno de los elementos mencionados deben ser desarrollados de manera especializada. La ciencia de datos es el área en la que se ubican los conocimientos y técnicas necesarios para trabajar con *big data*.

La ciencia de datos es un campo interdisciplinario que cubre ramas como: procesamiento de señales, modelos de probabilidad, aprendizaje maquinal, aprendizaje estadístico, minería de datos, bases de datos, ingeniería de datos, reconocimiento de patrones, aprendizaje de patrones, analítica predictiva, modelado de incertidumbre, *data warehousing*, compresión de datos, programación computacional, cómputo de alto desempeño, geolocalización y georeferencia.

Los equipos de trabajo con enfoque en la administración predictiva y proactiva son equipos multidisciplinarios que requieren incluir personal con conocimientos en ciencia de datos o combinaciones de especialistas en diferentes campos relacionados con esta área del conocimiento para poder explotar las fuentes de datos conocidas como *big data*. Como puede observarse, más allá de contar con

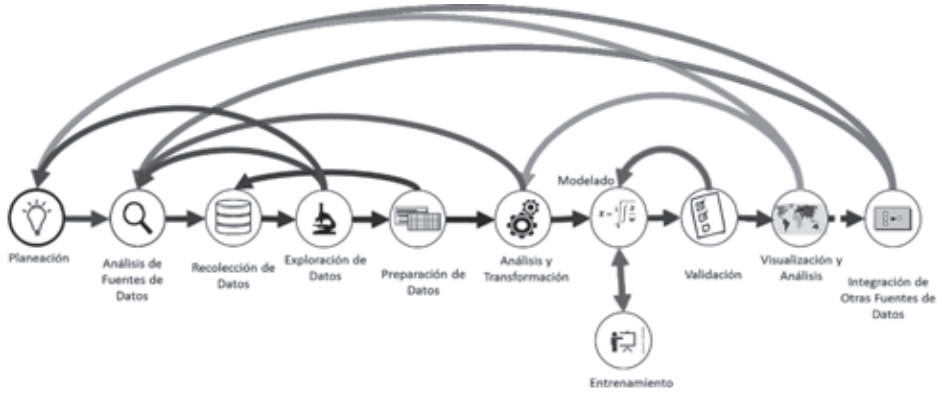
herramientas de *software* especializadas que faciliten la recolección, integración, tratamiento, análisis de visualización de los datos, es necesario contar con especialistas que generen modelos para extraer el valor de los datos que proporcionen información útil a quienes toman las decisiones.

Existen actualmente una diversidad de propuestas de procesos para extraer información proveniente de fuentes de *big data*. Una propuesta de proceso que toma en cuenta varias de estas propuestas y se complementa con la experiencia práctica obtenida al explotar fuentes de *big data* para producir información comprende los siguientes pasos (ver figura 1: Proceso para la extracción de información de fuentes de *big data*):

1. **Planeación**. Hacer una conceptualización inicial y establecer los objetivos que se desean cubrir.
2. **Análisis de fuentes de datos**. Revisar las fuentes de datos existentes y a partir de sus características seleccionar aquellas que presenten un mayor potencial de servir para satisfacer nuestras necesidades.
3. **Recolección de datos**. Obtener y almacenar los datos provenientes de las fuentes seleccionadas tomando en cuenta las características de la fuente. El uso de técnicas de filtrado adecuadas puede ayudar a mitigar la recolección de datos que no serán útiles para los propósitos que se persiguen.
4. **Exploración de datos**. Analizar tanto los contenidos como las estructuras que los contienen ayudará a entender los datos que se pueden obtener de las diferentes fuentes.
5. **Preparación de datos**. Una vez que se han comprendido los datos (tanto en lo referente a sus estructuras como a sus contenidos) es necesario ordenarlos y/o sintetizarlos para facilitar su análisis.
6. **Análisis y transformación**. Se realizan los análisis necesarios para establecer los procesos de transformación que servirán para extraer la información de las fuentes de *big data*.
7. **Modelado**. Se generan modelos para representar la información y extraer los datos, para ciertos tipos de información se deberá hacer uso de técnicas de aprendizaje por lo que será necesario realizar algunas tareas de entrenamiento como parte de la conformación del modelo.
8. **Validación**. Se revisa y refina el modelo comparando los resultados que se obtienen de él con otros datos conocidos y extraídos de fuentes que tradicionalmente se han considerado confiables. Con este paso buscamos maximizar la confiabilidad y la capacidad de predicción del modelo.

9. **Visualización y análisis**. Los resultados deben presentarse de manera que quien toma las decisiones pueda entenderlos fácilmente para que pueda usarlos como elementos en su análisis para llegar a conclusiones.
10. **Integración con otras fuentes de datos**. Un análisis más complejo encontrará que es necesario combinar los resultados de más de una fuente para obtener toda la información requerida para la toma de decisiones.

Figura 1. Proceso para la extracción de información de fuentes de *big data*



Del proceso descrito, se deben destacar dos implicaciones que son relevantes con respecto a su aplicación para hacer uso de *big data*:

1. **Hay un proceso de descubrimiento y aprendizaje implícito**. Al estudiar los datos es posible que se descubran características y aplicaciones para las fuentes de datos (solas o combinadas con otras) que inicialmente no eran percibidas.
2. **Se crean procesos y subprocesos iterativos**. La naturaleza de descubrimiento y aprendizaje del proceso hará necesario realizar iteraciones en algunos de los pasos o del proceso completo ya que con frecuencia se encontrarán ajustes y consideraciones que inicialmente no hubiera sido posible detectar o suponer.

Conforme se aplica el procedimiento a nuevos casos la organización adquirirá mayores habilidades para descubrir aplicaciones del valor informativo de los grandes conjuntos de datos que se clasifican como *big data*.

Para comprender mejor el potencial del valor informativo que se puede extraer de las fuentes de *big data* para apoyar a los enfoques de administración

predictiva y proactiva podemos recurrir a algunos ejemplos que muestran su posible uso.

Tomemos como base los datos que se pueden extraer de las publicaciones que se hacen en la red social Twitter. Una publicación en Twitter se conoce como un *tuit* y consiste en una estructura en un formato llamado JSON que además del contenido de la publicación (lo que quiere expresar quién lo publica) contiene otros datos, entre los que podemos encontrar: la identificación del usuario que hizo la publicación, la fecha y hora en que fue realizada, el lugar, etc.

Mediante técnicas de análisis y minería de textos, modelos de clasificación y aprendizaje maquinal se pueden hacer análisis para establecer si el contenido de un *tuit* expresa un sentimiento positivo (a favor), negativo (de disgusto) o neutro. Si se obtienen solamente las publicaciones realizadas por los usuarios de alguna región sobre cierto tema a través de la aplicación de filtros, es posible entender si ese tópico provoca apoyo o aversión.

Se debe recordar que la publicación de *tuits* ocurre en tiempo real, así que es posible monitorear inmediatamente el efecto de las acciones que se han ejecutado y hacer ajustes oportunamente, lo que puede ayudar a minimizar el impacto de aquellas decisiones que no han sido adecuadas; sin embargo, debe considerarse que por la velocidad con que se propaga la información a través de estos medios electrónicos, los efectos de una decisión pueden ser magnificados al producirse casi de manera inmediata.

El análisis de sentimientos requiere del desarrollo de modelos que analicen el lenguaje natural y que permitan interpretar las emociones que se infieren de una frase, pero no es indispensable para extraer información útil de los contenidos de los *tuits*. El simple hecho de determinar la cantidad de usuarios diferentes que hacen publicaciones sobre algún tópico (por ejemplo, una noticia) en el tiempo permite obtener la magnitud y permanencia del impacto que tiene entre la comunidad que utiliza dicha red social y actuar en consecuencia.

De las estructuras que contienen los *tuits* se puede obtener información. Si para un grupo de usuarios se cuenta con suficientes *tuits* a través de un periodo de tiempo relativamente largo es posible hacer análisis de los patrones de movilidad que describen. Si un usuario emite la mayor parte de sus *tuits* desde cierta ciudad y en algún periodo de tiempo se observa que los *tuits* son emitidos desde otra, entonces es posible deducir que realizó un viaje a dicha ciudad.

Asociando la información de movilidad para un conjunto de usuarios con otra información de contexto, como, por ejemplo, los días feriados, entonces se puede tener información sobre el impacto de esos días en la cantidad de visitantes que recibe una ciudad y prepararse en consecuencia para convertir ese hecho en una ventaja que aproveche una organización en la consecución de sus objetivos o la Administración pública para maximizar los efectos positivos y disminuir los negativos de la afluencia de personas.

Posiblemente para algunos tomadores de decisiones que desean hacer análisis de movilidad, los *tuits* presenten cierto sesgo o representen un conjunto demasiado pequeño de la población. En algunos casos, es posible utilizar otras fuentes que comparten similitudes para realizar este tipo de análisis. Por ejemplo, se pueden sustituir los *tuits* por registros de llamadas de usuarios de teléfonos celulares que también contengan información que permita relacionar un usuario con una ubicación en el tiempo. El uso de una u otra fuente dependerá entonces de otros factores como la disponibilidad de los datos, el alcance y la precisión requerida, etc.

El valor de la información conceptualizada como *big data* que principalmente es producida por entes privados representa oportunidades de negocio que comienzan a ser aprovechadas por sus productores. Con la finalidad de que esos datos puedan ser empleados para bien de la sociedad es necesario cuidar que las legislaciones reconozcan su importancia como bienes públicos.

Aplicaciones de *big data* en el contexto de las oficinas estadísticas de algunos países

Para ejemplificar algunas de las posibles aplicaciones de *big data* para apoyar la toma de decisiones en los diferentes sectores de la sociedad y la definición de políticas públicas, se presentan a continuación algunos casos que muestran proyectos que buscan explotar estas fuentes de información:

- El Reino Unido planea reemplazar con medidores inteligentes los medidores actuales de gas y electricidad con que cuentan a más tardar en 2020 (UK Gov; n. d.). La información que generarán permitirá conocer los patrones de consumo de gas y electricidad de todos los abonados. Con esta información es posible desarrollar tarifas flexibles que incentiven o desincentiven el uso de energía en ciertas horas del día; incrementar la eficiencia y confiabilidad del sistema eléctrico (detectando fallas, insuficiencias o cortes); determinar

si existen aumentos o disminuciones en la demanda por parte de ciertos sectores (como, por ejemplo, las industrias, con lo que se puede prever que existirá una aceleración o desaceleración en la actividad industrial, lo que repercutirá a su vez en el desempeño de los indicadores económicos).

- Los países bajos tienen desplegada una red de más de 60.000 sensores (que combinan tecnologías de inducción, cámaras, *bluetooth*) en sus carreteras que les permiten determinar los flujos vehiculares, calcular los tipos de vehículos en circulación (conforme a su tamaño) y los patrones de tráfico (Puts, n. d.). Con estos sensores pueden detectar remotamente y de manera oportuna interrupciones en el tráfico que pueden ser ocasionadas por accidentes, incidentes naturales, fenómenos sociales, etc., y actuar oportunamente; también con la información de los patrones de tráfico pueden planear acciones como la programación de rutas de transporte público o la construcción de obras que permitan disminuir los atascos viales; la información sobre los flujos vehiculares permite también determinar cuándo será conveniente realizar mantenimiento a las carreteras a través del cálculo de su desgaste.
- La Oficina Estadística de Europa (Eurostat) está desarrollando un proyecto para generar estadísticas a partir del uso de las redes sociales por parte de las empresas (Eurostat, n. d.). En su estudio ha encontrado que el 39% de las empresas en Europa utilizan redes sociales y que de ellas el 79% las utiliza para propósitos de negocio, como fortalecer su imagen, interactuar con sus clientes y manejar sus relaciones con ellos, etc.; combinados con otras fuentes de datos se podrá medir el impacto de las estrategias de las empresas en las redes sociales para incrementar su participación o favorecer su supervivencia de acuerdo al sector en que participan.
- El INEGI en México (INEGI, n. d.) está realizando estudios de movilidad y de análisis de sentimiento utilizando información de la red social Twitter para obtener datos sobre el origen de turistas nacionales que visitan los pueblos mágicos, el comportamiento de desplazamientos de personas entre diferentes estados, la movilidad de las personas que habitan en la frontera norte del país, patrones de visitas a ciertos tipos de giros de negocio durante la semana en diferentes horas del día, el posible impacto en el estado de ánimo debido a ciertas noticias relevantes, detección de uso de lenguaje discriminatorio o misógino, etc.
- El banco BBVA (BBVA, n. d.) tiene un proyecto para generar valor con los datos que genera su actividad cotidiana. Con estos datos realizan estimaciones de riesgo al proporcionar un crédito a un cliente, determinar el comportamiento de compra de los visitantes a ciertas ciudades, el impacto en la derrama económica de algún evento.

- La Oficina de Estadísticas de Nueva Zelanda (Krsinich, 2015) está utilizando los datos de los registros de ventas para medir las variaciones en los precios de doce productos electrónicos que incluyen computadoras, televisores, teléfonos inteligentes, cámaras, tarjetas de memoria, etc. La información obtenida les permite obtener además de los patrones de variación en los precios, la sensibilidad de la demanda a estas variaciones, la posible relación de la demanda de estos productos con el desempeño general de la economía, etc.
- El Buró de Estadísticas de Australia (Marley *et al.*, 2014) está trabajando en el análisis de datos de satélite para hacer la medición remota de áreas cultivadas. Con la información obtenida se busca obtener la estimación de producción de diferentes tipos de cultivos, con datos obtenidos en diferentes temporadas se pueden establecer patrones de crecimiento o decrecimiento de la producción y asociarlos a factores climatológicos y económicos provenientes de otras fuentes.
- La Oficina de Estadísticas de la República de Eslovenia trabaja en la extracción de estadísticas de vacantes laborales obtenidas de páginas web especializadas en bolsas de trabajo (Nikic, 2015). Esta información permite obtener información oportuna sobre el comportamiento de los mercados laborales de manera que además de ayudar a medir el desempleo y la demanda, al cruzar la información con otras fuentes se pueden hacer estimaciones como el probable comportamiento económico de algún sector o las necesidades de capacitación en alguna especialidad.

Conclusiones

Las fuentes reconocidas actualmente como *big data* facilitan la implementación de los enfoques de administración predictiva y proactiva al proveer información para desarrollar modelos que permiten anticipar necesidades y hechos para establecer metas más acertadas, así como atender oportunamente a situaciones que están desarrollándose para aprovecharlas al máximo en caso de ser positivas o reducir sus efectos en caso contrario, lo que se traduce en un uso más óptimo de recursos y en acciones que pueden reflejarse en una mayor satisfacción de los usuarios.

Referencias bibliográficas

Álvarez Nebreda, Carlos C. (1998). *Glosario de Términos para la Administración y la Gestión de los Servicios.* Ediciones Díaz de Santos.

Cox, M. I., Ellsworth, D. (1997). "Application-controlled demand paging for out-of-core visualization". *Proceedings of the IEEE 8th conference on Visualization.* IEEE, USA, 1997.

Eurostat. Obtenido en septiembre de 2016, de http: //ec.europa.eu/eurostat/statistics-explained/index.php/Social_media_-_statistics_on_the_use_by_enterprises

IEAG (Grupo Asesor Independiente de Expertos Sobre una Revolución de Datos para el Desarrollo Sostenible del Secretario General de las Naciones Unidas) (2014). "A World That Counts", Independent Expert Advisory Group Secretariat. Data Revolution Group; noviembre 2014.

INEGI. obtenido en septiembre de 2016, de http: //unstats.un.org/unsd/bigdata/conferences/2016/presentations/day%202/Juan%20Mu%C3%B1oz.pdf

Krsinich, F. (2015). "Implementation of consumer electronics scanner data in the New Zealand CPI". Statistics New Zealand, New Zealand.

Lyman, P., Varian, Hal R. (2003). "How Much Information". University of California at Berkeley, USA, 2003; recuperado de http: //www.sims.berkeley. edu/how-much-info-2003 en septiembre de 2016.

Marley, J., Elazar, D., Traeger, K. (2014). Research Paper: Methodological Approaches for Utilising Satellite Imagery to Estimate Official Crop Area Statistics (Methodology Advisory Committee), ABS (Australian Bureau of Statistics); Australia.

Merriam-Webster (n. d.). Obtenido en septiembre de 2016, de http: //www.merriam-webster.com/dictionary/data

Nikic, B., Spech, T., Klune, Z. (2015). "Usage of new data sources at SURS"; UNECE. *Conference of European Statisticians*. USA.

Puts, M. "Big Data and Road Sensors. Implementing a Big Data Statistics"; obtenido en septiembre de 2016, de https: //ec.europa.eu/eurostat/cros/system/files/marco_puts_ess_traffic_loops_big_data_-_final.pdf_en

UK Gov. Obtenido en septiembre de 2016, de https: //www.gov.uk/guidance/smart-meters-how-they-work

UNECE/ECE (2013). *What Does "Big Data" Mean for Official Statistics?* Comisión de Estadísticos Europeos de la Comisión Económica para Europa de las Naciones Unidas, Suiza.

› Bibliografía recomendada

Berman, J. J. (2013). *Principles of Big Data. Preparing, Sharing and Analyzing Complex Information*. Morgan Kufman, Elsevier, USA.

Feinleib, D. (2014). Big *Data Bootcamp*. Apress, USA.

Grus, J. (2015). *Data Science from Scratch*. O'Reilly, USA.

Myatt, Glenn J., Johnson, Wayne, P. (2014). *Making Sense of Data I. A Practical Guide to Exploratory Data Analysis and Data Mining*. John Wiley & Sons, USA.
Ojeda, T., Murphy, Sean P., Bengfort, B., Dasgupta, A. (2014). *Practical Data Science Cookbook*, Packt Publishing, USA.
Pierson, L. (2015). *Data Science for Dummies.* John Wiley & Sons, USA.
Schutt, R.; O'Neil, C. (2014). *Doing Data Science*. O'Reilly, USA.
Stanton, J. (2012). *An Introduction to Data Science*. Syracuse University, USA.
UNSD (2016). *Report of the 2015 Big Data Survey.* United Nations Statistical Division. Statistical Commission. Forty-Seven Session, USA.

Capítulo 11

Nuevas estructuras organizativas y tecnologías emergentes en las organizaciones *data-driven*

Antonio Moneo*

Gobiernos, empresas, universidades y organizaciones sin ánimo de lucro de todo el mundo buscan en los datos nuevas oportunidades. Los datos se han convertido en "El Dorado" del siglo XXI: un tesoro que todos quieren conquistar. El paradigma de las ciudades inteligentes (*smart cities*) representa un modelo ágil y moderno que ahora muchas organizaciones tratan de emular para mejorar su productividad y competitividad. Muchos creen que este modelo para reducir ineficiencia que lastran los negocios tradicionales y permitirá generar nuevas oportunidades de negocio.

El paradigma *data-driven* describe un mundo eficiente, automático y de nuevas oportunidades, en el que las personas son capaces de utilizar la tecnología para generar impactos globales y atacar grandes problemas estructurales. Facebook, Google, Amazon y otras compañías tecnológicas se han convertido en ejemplos sobre el valor de comprender las preferencias y necesidades de las personas. Lo que no se ajuste a las necesidades de las personas, desaparecerá.

Los datos masivos o *big data* son una pieza fundamental de este paradigma. El cruce de variables sociodemográficas con datos comerciales o urbanísticos puede ayudar a entender cómo las personas determinan sus preferencias; y esto es clave para explicar patrones de consumo o incluso de voto. El potencial de los datos es prometedor para cualquier industria, e instituciones líderes como Forrester, Gartner o IDC afirman que el *big data* representa una oportunidad en expansión que alcanzará los 46.000 millones de dólares en 2019. Aunque no exista consenso sobre cómo cuantificar el potencial de esta llamada revolución, sí parece claro que el *big data* ha venido para quedarse. Pero ¿cómo puede una empresa o un gobierno aprovechar al máximo esta oportunidad?

Este artículo presenta algunos conceptos clave que cualquier organización debería considerar al plantear su estrategia para utilizar los datos como una fuente

* Especialista en Gestión de Conocimiento en BBVA Data & Analytics.

de cambio para su negocio. Los conceptos que se presentan han sido capturados desde la práctica profesional y no pretenden ser exhaustivos. El objetivo del artículo es ordenar el debate para que el lector pueda formularse preguntas clave y continuar investigando.

La economía API

La "economía API", como la llaman algunos, podría definirse como el conjunto de bienes y servicios que se generan a partir de la búsqueda de una mayor interoperabilidad entre sistemas de información para la mejora de la eficiencia organizativa o la identificación de oportunidades. Este concepto es relevante porque permite objetivar el contexto de este debate.

Este nuevo sector económico surge de la digitalización de las industrias tradicionales, así como de la producción y explotación de datos generados por sistemas informáticos y sensores inteligentes. Más que de una revolución, que pretende subvertir un orden, sería más conveniente hablar del surgimiento de una industria que pretende transformar una materia prima, los datos, en bienes y servicios digitales.

La industria de los datos está en proceso de formación, y es altamente dinámica y cambiante. Nuevos competidores y tecnologías disruptivas emergen cada día y todavía es pronto para saber quién la dominará. En esta industria coexisten y compiten actores de diversa naturaleza y tamaño, y pueden encontrarse especializaciones en producción, recolección, análisis e interpretación de los datos. Entre los actores tradicionales, destaca el rol de los individuos, quienes generan gran cantidad de datos a través de sus teléfonos móviles, compras en Amazon o mensajes en las redes sociales.

A pesar de su potencial estimado, hay quienes temen a una nueva burbuja tecnológica como la que ocurrió en 2000 con las *puntocoms*. PWC advierte de los problemas que muchas organizaciones encuentran para trabajar con datos masivos. No es tan fácil como parece: el correcto uso de datos masivos tiene altos costes de implementación derivados de la instalación de infraestructura, la contratación de personal especializado y de la creación de una cultura organizacional adecuada capaz de incluir el análisis de datos en sus rutinas.

En esta sección se revisarán algunos hitos de la industria tecnológica que permitirán entender algunos conceptos básicos de esta nueva industria.

Los datos como mercancía

Los datos generados por los dispositivos y aplicaciones informáticas son la pieza clave de esta industria. Cada *click* en una página web o aplicación de teléfono móvil, sea un correo, un *tuit*, una compra electrónica o una búsqueda en Google, queda registrado en listados de eventos, conocidos como *log*, que además de la acción realizada capturan información sobre el usuario que la realiza. Con esos datos no agregados se pueden identificar patrones de comportamiento y establecer hipótesis sobre cualquier tema. Por ejemplo, las compañías de seguros agrícolas pueden usar datos meteorológicos para determinar el riesgo de un cultivo en una zona específica, y las compañías inmobiliarias pueden usar las redes sociales para conocer el nivel adquisitivo y los intereses de las personas en un determinado barrio.

Estos datos tienen un valor muy alto en el mercado. En octubre de 2014, Facebook pagó 16.000 millones de dólares por la aplicación de mensajería Whatsapp, 40 euros por usuario, y marcó una referencia del valor que los datos pueden llegar a tener en el mercado. El caso de Whatsapp hace pensar que el acceso a este tipo de datos determina el valor de una compañía, incluso por encima del valor de su infraestructura. Muchos comparan los datos con el oro y el petróleo porque entienden que son una materia prima que se puede vender y transformar para generar otros bienes y servicios.

Para determinar el valor de los datos es necesario entender su naturaleza. Pueden diferenciarse según su grado de estructuración (no estructurados y estructurados), su origen (personal, comercial o gubernamental), su grado de accesibilidad (cerrados y abiertos). El grado de accesibilidad es el más determinante para el Open Data Institute, quien diferencia cinco grandes categorías según su grado de apertura. En un lado del espectro se encuentran los datos que contienen información confidencial y solo son accesibles desde dentro de una organización. En el lado opuesto, los datos abiertos, que son accesibles por cualquiera para cualquier propósito.

La infraestructura y el *software* como servicio

La industria de los datos se apoya sobre dos grandes pilares: la infraestructura tecnología (servidores) y los programas (*software*) para el procesamiento de datos.

Amazon, Google y Microsoft han lanzado ya plataformas como Amazon Web Services, Google Cloud y Microsoft Azure, que permiten contratar servidores

para el almacenamiento de información y el despliegue de herramientas de *cloud computing*. Estas plataformas dan acceso a tecnologías de última generación y usan precios flexibles que se determinan en función del consumo del servicio. En octubre de 2015, Apple adquirió un terreno de 80 hectáreas en Oregón, que muchos interpretaron como un paso para ampliar su centro de datos y competir en la oferta de servicios de infraestructura.

Por otro lado, grandes empresas como IBM, Microsoft y SAP han desarrollado potentes herramientas de gestión y análisis de datos que ofrecen soluciones integrales para la recogida, procesamiento y análisis de datos. Herramientas como Watson Analytics, Power BI y Hana tratan de facilitar el análisis de datos para los usuarios con menor capacidad técnica. La fuerte inversión de IBM en el desarrollo de Watson Analytics, que supera ya los 1.000 millones de dólares, y la reciente adquisición de compañías especializadas en la gestión de datos masivos, como The Weather Channel por unos 2.000 millones de dólares, hace pensar un futuro dinámico en el sector.

Además de los casos mencionados, hay un interesante compendio de compañías emergentes que están experimentando un fuerte crecimiento. Tableau, una plataforma para la visualización de datos, generó unos ingresos de 653 millones de dólares en 2015 y ha sido recientemente adquirida por HyPer Technology, una compañía de infraestructura de datos. Domo, una plataforma de gestión de datos, recibió en julio una inyección de 235 millones de dólares y ha sido valorada en dos mil millones de dólares. También destacan otras *startups* con un fuerte crecimiento en los últimos meses como Mapbox (52 millones de dólares en junio de 2015) y Carto (23 millones de dólares en septiembre de 2015), especializadas en la visualización de datos geoespaciales, pero que más recientemente han comenzado a adquirir tecnologías complementarias para ampliar su capacidad. Por otro lado, la reciente adquisición de Pentaho, una plataforma de integración y gestión de datos, por Hitachi (500 millones de dólares), sugiere la posibilidad de la formación de nuevos competidores de gran tamaño. En este aspecto, también deben observarse de cerca los movimientos en compañías del sector de la telefonía, la defensa o la salud.

› El mercado de los datos

La compra y venta de datos está dando lugar a un mercado muy particular en el que surgen nuevos servicios que conectan los procesos de producción, análisis, publicación y reutilización de datos.

A lo largo de los últimos años, muchas empresas han desarrollado aplicaciones que permiten recolectar información de los usuarios ofreciendo servicios gratuitos. Empresas como Google, Facebook, Whatsapp o Twitter utilizan la información recopilada a través de sus aplicaciones para crear canales de publicidad personalizada. Algunas de ellas, como Netflix, Google y Amazon, también aprovechan los datos recopilados para comprender las preferencias de los usuarios y desarrollar nuevos servicios.

A las empresas de producción de datos ya mencionadas habría que sumar también otras compañías telefónicas y financieras que, a través de sus dispositivos (telefonos móviles, tarjetas de crédito), generan datos que pueden ayudar a comprender patrones de movimiento de las personas, de compra. Estos datos están resultando ser interesantes para comprender patrones de comportamiento en la población, como hicieron las ciudades de Valencia o Zaragoza, que utilizaron datos de teléfonos móviles para optimizar las rutas de transporte público.

Muchos gobiernos de todo el mundo se han sumado a esta tendencia y han comenzado a hacer sus datos accesibles al público. Gobiernos de Reino Unido, Estados Unidos, Francia, España o Alemania han liberado ya alrededor de un millón de bases de datos de distinta naturaleza y más de 600.000 son accesibles desde el Portal Europeo de Datos. Estos datos son la base de grandes emprendimientos en áreas como la agricultura, el transporte o la salud, y comienza a demostrar su potencial en áreas como la educación o la justicia.

El rasgo más característico y novedoso de esta industria es el papel que desempeñan los individuos, cuya actividad registrada a través de distintos medios genera la información que todas las compañías tratan de capturar. La regulación sobre el derecho a la privacidad y propiedad de los datos generados por los individuos es un debate multifacético y complejo, como demuestra el debate sobre la regulación del derecho al olvido en la Unión Europea.

La cesión de los datos personales se realiza muchas veces a cambio de un servicio gratuito en las plataformas de Facebook, Twitter o Google. Iniciativas como DataDonors, DonorsChoose o mPower permiten a los usuarios donar sus datos para fines médicos o sociales, y otras como Quandl permiten comercializar datos personales. Los casi 17 millones de dólares en financiación recibidos por Quandl hacen pensar que próximamente aparecerán más plataformas para la compraventa de datos personales.

Nuevos perfiles profesionales

El surgimiento de la industria de los datos está dando lugar a nuevos perfiles profesionales, y es importante que las organizaciones tomen decisiones estratégicas sobre el talento que necesita incorporar o contratar para poder maximizar sus oportunidades. En esta sección, se comentarán cinco nuevos perfiles profesionales: evangelizadores, científicos, gestores, expertos en protección de datos y periodistas de datos.

Evangelizadores de datos

En Estados Unidos, se ha extendido rápidamente el perfil de *data evangelist* para describir las funciones de quienes promueven la transición hacia un modelo organizacional basado en los datos. No es raro encontrar profesionales que se presentan ya como *data evangelists*. Estos perfiles son fundamentales para articular el cambio cultural que requiere la economía digital.

Su misión principal es capacitar y orientar a la organización en el proceso de adopción de nuevas tecnologías para la reutilización de datos. Su papel es parecido al de un comercial, en tanto que tratan de identificar *early-adopters* dentro de las organizaciones, es decir, personas dispuestas a adoptar nuevas tecnologías antes que sus compañeros. Por ello, requieren un fuerte conocimiento del negocio de la institución y un conocimiento avanzado de la tecnología. Para lograr su objetivo, trabajan con clientes potenciales, ayudándoles a definir sus necesidades, documentan casos de éxito y participan en conferencias de sensibilización y eventos de capacitación explicando los factores críticos de éxito y colaboran en el desarrollo de herramientas para reutilizar datos.

Científicos de datos

El científico de datos (*data scientist*) es uno de los perfiles más demandados en la actualidad. Se entiende como una evolución del perfil de analista de datos, porque combinan las técnicas de análisis estadístico tradicional con nuevas técnicas para la recolección y análisis de grandes volúmenes de datos (*big data*). Los científicos de datos están a medio camino entre la parte operativa y la corporativa, y tratan de reutilizar la información disponible para generar nuevas oportunidades de negocio. Diseñan algoritmos que permiten automatizar el procesamiento de distintas fuentes de datos, e intervienen en el diseño de herramientas

de aprendizaje automático (*machine learning*) que permiten mejorar los resultados del análisis.

Las ofertas de trabajo normalmente solicitan candidatos con curiosidad, autónomos, apasionados y orientados a la acción; idealistas capaces de pensar en grande y sin aceptar las inercias de las organizaciones. Dado que estamos ante un área nueva para muchas empresas, se espera que los científicos de datos sean capaces de guiar a las organizaciones en la incursión a una nueva forma de trabajar. Resulta ejemplificador que entre los siete mejores *data scientists* del mundo según Forbes, muchos trabajan para el sector público, como D. J. Patil (científico de datos de la Casa Blanca); Elizabeth Warren (senadora de Massachusetts) y Todd Park (Departamento de Salud y Servicios Sociales de EE. UU.).

Gestores de datos

Los gestores de datos (*data managers*) son responsables del diseño, mantenimiento y reutilización de los datos de una organización. Podrían considerarse como una evolución del perfil de los gestores documentales y archiveros, aunque su labor es más especializada porque además de definir protocolos y estándares para garantizar la calidad e interoperabilidad de los datos, diseñan herramientas para facilitar la captura de nuevas de fuentes de información no estructuradas, como el contenido de las redes sociales. Asimismo, también aplican el uso nuevas tecnologías para facilitar la visualización de la información.

Estos profesionales interactúan con los departamentos de tecnología y los departamentos operativos y se encargan de asegurar la calidad, exactitud y solidez de las bases de datos. Buscan de manera proactiva que la información almacenada se reutilice. Para ello, realizan estudios periódicos del contenido de las bases de datos y ayudan a los empleados de las organizaciones a encontrar la información necesaria. Estos perfiles requieren conocimientos informáticos avanzados sobre el manejo de información, ontología y tecnologías de la información, pero exige además un conocimiento profundo de las áreas de operación de los clientes.

Responsables de protección de datos

El responsable de protección de datos es un experto encargado de prestar asesoramiento sobre la recogida, uso, acceso o tratamiento de datos personales

dentro de una empresa. Su labor es la de crear, difundir y verificar el cumplimiento de las políticas corporativas de tratamiento de datos personales, así como la de vigilar la seguridad y el cumplimiento de los sistemas para que dichos datos personales se traten conforme a la normativa aplicable. Es de especial relevancia la consolidación de la figura del *data protection officer* en el ámbito europeo mediante la publicación del nuevo reglamento de protección de datos.

Entre sus diversas funciones, interactúan con las unidades de decisión y dirección de las compañías coordinándose con las unidades tecnológicas para determinar escenarios de riesgo y adaptar las políticas corporativas a fin de asegurar la actualización constante de las medidas de seguridad y la mitigación de posibles amenazas ante el tratamiento inadecuado de datos de carácter personal. Estos perfiles requieren conocimientos jurídicos en materia de protección de datos, incluyendo medidas técnicas y organizativas, además de tener un conocimiento específico de la industria en la que desempeñan sus funciones. Idealmente, es necesario tener conocimientos tecnológicos en temas relacionados con la ciberseguridad y la gestión de riesgos. La norma ISO/IEC 27001, el estándar para la seguridad de información, recoge más información al respecto.

› Periodistas de datos

El periodismo de datos es una especialización surgida del mundo del periodismo que se caracteriza por la producción de información a partir de datos. Es una pieza clave en la cadena de valor de los datos porque se encarga de generar historias basadas en datos y de divulgarlas en los medios de comunicación y redes sociales.

Son un perfil frecuente en muchos medios de comunicación y, cada vez más, en otros tipos de organizaciones. Interactúan con las unidades de negocio, a través de sus expertos en datos y las audiencias externas, y su función principal es alinear los mensajes institucionales con los resultados de los análisis. Establecen alianzas internas y externas para asegurar que la información precisa y contundente llegue a las audiencias clave. Este perfil requiere una gran capacidad de redacción y análisis, así como de conocimientos avanzados de *marketing* y diseño web. Exigen también un conocimiento preciso de la misión organizacional. Se puede obtener más información sobre este perfil profesional en el *Manual de periodismo de datos*.

La cadena de valor de los datos

En la jerga de la industria de los datos, la expresión *"Garbage in, garbage out"* ("Basura dentro, basura fuera") aparece frecuentemente en todos los debates. Esta gráfica expresión resume el desafío central de esta industria: si los datos de los que se dispone no son relevantes, las conclusiones tampoco lo serán. El proceso de trabajo con datos es costoso y debe ordenarse para para evitar ineficiencias. A continuación se explican cinco factores críticos de éxito para desarrollar una estrategia de explotación de datos.

Recolección

El proceso de recolección de datos consiste en reunir y sistematizar la información relevante con un objetivo determinado. Tradicionalmente, ha sido un proceso manual, altamente costoso y frecuentemente inexacto. Según un estudio de la Dra. Yacyshyn de la Universidad de Alberta (Canadá), el coste de construir un censo de población en Estados Unidos era de 23,29 dólares por habitante en el año 2000 (6.500 millones de dólares) y de 17,06 en Canadá (511 millones). Estos censos, que se construyen cada cierto número de años, son fundamentales para que los gobiernos puedan hacer sus cálculos macroeconómicos y, sin embargo, son defectuosos en los países menos desarrollados. Un estudio de la compañía Market Research Inc. afirma que el coste medio de un estudio de mercado oscila entre los 30.000 y 50.000 dólares y, aunque las cifras puedan variar, queda claro que la recolección de información supone un coste importante para cualquier organización.

Muchos ven en la tecnología y los datos procedentes de las herramientas digitales un potencial enorme para abaratar los costes de recolección, y tienen razón para hacerlo. Es posible que, a través de los datos accesibles de Twitter, una organización pueda conseguir información a bajo coste sobre su audiencia, pero ¿cuánto cuesta almacenar esa información?

Almacenamiento

La información digital ofrece enormes ventajas y facilidad de uso y transformación pero, con frecuencia, se subestima el coste de almacenamiento. Varias estimaciones afirman que se producen y procesan alrededor de 2,5 *exabytes* (2.500 millones de *gigabytes*) al día, y cada persona en Estados Unidos consume alrededor de 60 megabytes al día a través de su teléfono móvil. Al final del año 2016

se estima que el volumen de datos aumente hasta los 3,77 *zetabytes*. Las dimensiones son tan grandes que muchos desconocen el significado de las nuevas magnitudes, pero basta con decir que el tamaño medio de un disco duro ha venido duplicándose cada 12 o 18 meses. Un ejemplo de este crecimiento son las tarjetas microSD que muchos teléfonos incorporan como memoria adicional. En 2005, la capacidad media era de 128 MB y, diez años más tarde, de 128 GB.

Una organización que quiera trabajar con datos debe tener en cuenta el precio del almacenamiento y valorar decisiones como la contratación de servicios de almacenamiento en la nube, donde es posible encontrar servicios elásticos en los que se paga por volumen almacenado, en vez de por dispositivo. Los nuevos servicios en la nube tienen grandes ventajas porque evitan el problema de la obsolescencia de los dispositivos y permiten pagar solo por la infraestructura utilizada. Sin embargo, aunque esto pueda parecer un ahorro, no hay que olvidar que el volumen de datos es exponencialmente creciente y el coste de almacenamiento podría dispararse. Es necesario determinar qué información debe almacenarse y cuál puede descartarse.

› Limpieza de datos

El proceso de limpieza de datos es un proceso complejo y fundamental que consiste en eliminar errores e inconsistencias en las bases de datos, completar información cuando sea posible, estandarizar la información y, en muchos casos, anonimizar la información. Es fundamental que los datos sean acertados, completos, consistentes y uniformes para poder realizar luego las tareas de análisis de datos. Unos datos equivocados pueden tener terribles consecuencias en el futuro.

La anonimización de los datos es un ejemplo de las transformaciones más necesarias y comunes que deben realizarse. Si una organización utiliza una fuente de datos que contiene información personal y se actualiza constantemente, deberá desarrollar un proceso para asegurar que en cada actualización se conserven los mismos principios de privacidad.

› Interpretación

La interpretación de los datos es quizá la parte más visible de esta industria y es el proceso fundamental para conectar los datos con los objetivos empresariales. La interpretación de los datos puede realizarse mediante distintas herramientas, como el conocido *software* de código abierto R, que permite realizar análisis

estadísticos computacionales y gráficos, o plataformas como la ambiciosa Watson Analytics, que pretende universalizar el análisis de datos.

La visualización de los datos a través de gráficas y mapas interactivos se ha convertido en un componente crucial de cualquier estrategia de datos y, en muchos casos, se ha convertido en una herramienta fundamental de comunicación corporativa. La visualización de datos está permitiendo desarrollar poderosas herramientas de comunicación para las redes sociales que sirven para transmitir mensajes que las personas comparten rápidamente en la red, como la conocida visualización sobre el aumento de temperatura entre 1850 y 2016.

Apertura

La apertura de información es un proceso que implica documentar, licenciar y publicar la información para que pueda ser consumida por otros. En muchos casos, la apertura no se entiende como un proceso asociado al trabajo con datos, pero estudios como los de McKinsey revelan que los datos abiertos tienen un enorme potencial multiplicador. Un famoso reportaje, publicado en 2014, afirmaba que los datos abiertos representaban una oportunidad de negocio de unos tres billones de dólares americanos al año en sectores como el transporte, la educación o la salud. Un informe del European Data Portal afirmaba en una línea similar que la apertura de datos de los gobiernos podía generar una oportunidad de 325.000 millones y 100.000 empleos en los 28 Estados de la Unión Europea entre 2016 y 2020. Grandes multinacionales como Orange han dado algunos pasos para abrir parte de sus datos a fin de buscar soluciones a problemas de desarrollo. Por otra parte, Telefónica cedió datos anonimizados para la realización de un estudio sobre el terremoto del 16 de abril de 2016.

La apertura de los datos es una decisión compleja para muchas organizaciones porque implica una decisión ejecutiva, pero como algunas instituciones gubernamentales ya han adoptado la apertura de datos como un estándar (*open by default*), es fundamental tener en cuenta este proceso.

Principios de gobernanza

Una estrategia de datos debe estar guiada por unos principios claros. Teniendo en cuenta lo explicado en este artículo, esta sección resumirá los principios de la Carta Internacional de los Datos Abiertos, para que sean tomados como guía.

Abiertos por defecto

Impulsados por el deseo de mejorar la transparencia, muchos gobiernos se han comprometido a hacer sus datos abiertos por defecto. Esto significa que a menos que exista una razón fundada, como la privacidad de los individuos o la seguridad, la información debería ser abierta y estar disponible para cualquier uso. Por ejemplo, en mayo de 2013, el gobierno de Estados Unidos firmó una orden ejecutiva para obligar a todas las agencias gubernamentales e instituciones públicas a publicar todo aquello que hubiera sido financiado con recursos públicos.

Oportunos y exhaustivos

La oportunidad y exhaustividad de los datos es un principio que puede guiar los procesos de recolección, almacenamiento y limpieza. Es importante que los datos utilizados estén completos, sean correctos y se ofrezcan con las mínimas manipulaciones posibles. Este principio hace énfasis también en la documentación necesaria que debe acompañar a la descripción de los datos. Para garantizar esto, es conveniente desarrollar protocolos internos que describan los procesos de publicación, las personas responsables de los datos, la información contextual sobre las bases de datos (metadata) y las licencias de uso que se van a aplicar en cada caso.

Accesibles y utilizables

Los conjuntos de datos (tablas) deben considerarse como un producto de conocimiento y, como tal, deben contar con una licencia de uso que deje claro qué usos se les pueden dar a los datos. Actualmente existen licencias Creative Commons que regulan específicamente los tipos de permisos que se pueden conceder. Si el objetivo es compartir los datos, es importante observar la aclaración de Creative Commons sobre el uso de licencias no derivativas y no comerciales, porque pueden limitar la reutilización de los datos publicados para fines de investigación o comerciales.

Comparables e interoperables

Para que el flujo de información sea más rápido y evitar costes en los procesos de transformación y estandarización de los datos, es crucial que se usen estándares

internacionales. Los datos deben estar preparados para que sean legibles por aplicaciones informáticas e ir acompañados de la documentación necesaria para que los expertos comprendan la naturaleza de los datos. Actualmente existen algunos estándares como DCAT, el cual ha sido adoptado por la Unión Europea para estructurar el proceso de recolección y publicación de datos de los 28 países.

Consejos para el desarrollo de una estrategia de datos

Los datos masivos representan una oportunidad para la transición hacia modelos organizativos *data-driven*. Esta transición debe basarse necesariamente en una estrategia sólida que permita maximizar la oportunidad que se presenta. Gobiernos y empresas, colegios, hospitales y cualquier tipo de organización encontrarán en esta nueva industria posibilidades para mejorar y hacer crecer su oferta de productos y servicios. El trabajo con datos modificará los procesos de negocio y generará oportunidades rentables solo si se consideran estratégicamente los elementos mencionados. A continuación se presentan algunos consejos a modo de conclusión.

Desarrollar una teoría del cambio para la organización

El primer paso para trabajar con datos es reconocer la voluntad de cambio y sus consecuencias. El trabajo con datos afecta a toda la organización, tanto a su funcionamiento interno como a los productos y servicios que ofrece, y es necesario desarrollar una estrategia que permita enfocar la dirección del cambio. Es importante que los miembros de la organización comprendan la visión estratégica y cómo puede modificar su trabajo diario. Para hacerlo, resultan útiles los modelos de madurez como el que recomienda IBM, que permiten medir el progreso de la organización y establecer una visión de largo plazo.

Incluir los datos en la identidad corporativa

Es necesario comprender la información como un activo y calcular con precisión la inversión necesaria para transformarla en una herramienta para la toma de decisiones. No hacerlo podría implicar costes sobrevenidos que podrían hacer

inviable la transición hacia modelos organizativos inteligentes. Los nuevos procesos deben ser capaces de conectar distintas áreas de negocio y, especialmente, mejorar la conexión entre la organización y los beneficiarios. El éxito de compañías tecnológicas como Facebook o Google y la continua apuesta de gobiernos de todo el mundo por introducir una cultura de trabajo basada en datos ilustra la esperanza de muchos de crear un mundo de oportunidades con menos ineficiencias y más oportunidades de negocio.

› Crear un esquema de alianzas estratégicas

Las alianzas son una herramienta importante para mitigar los riesgos y maximizar los resultados de la transición hacia un modelo basado en datos. La industria de los datos atrae a organizaciones de distinta naturaleza, con necesidades similares y capacidades complementarias. Actualmente existen varias alianzas globales como el Global Partnership for Sustainable Data, que trata de llevar el potencial de los datos a los países en desarrollo; o Data-Pop Alliance, que promueve una visión del *big data* como una oportunidad para mejorar la vida de las personas. Existen también alianzas sectoriales, como Global Open Data for Agriculture and Nutrition, que trata de profundizar en el potencial de los datos para solucionar un tema concreto.

› Adoptar estándares internacionales

La estandarización de la información mejora la interoperabilidad y, en consecuencia, incrementa su valor: cuanto más estandarizada sea la información, más posibilidades existirán de que pueda ser combinada con otras fuentes de datos. La Organización Internacional para los Estándares (International Standards Organization-ISO) publica regularmente estándares como la norma ISO/IEC DIS 20802-2, que regula el protocolo OData para la publicación de datos abiertos; o la norma ISO/IEC 27001 que regula los estándares de seguridad de información. Otros organismos como AENOR también publican normas similares como UNE 178301:2015, que define los estándares para los datos abiertos en ciudades inteligentes.

Capítulo 12

El doble reto: *open data* & *big data*

Lourdes Muñoz*

La demanda social creciente del *open data*

Gobierno abierto u *open government*

Apostar por el gobierno abierto no es solo una oportunidad para aprovechar las TIC, supone adecuar las Administraciones públicas a la dinámica de la nueva sociedad en red y en consecuencia a las nuevas demandas de la ciudadanía.

Un gobierno abierto u *open government* exige una actitud y unas acciones basadas en los principios de transparencia, colaboración y participación. Una institución transparente fomenta y promueve la rendición de cuentas ante la ciudadanía y proporciona información sobre sus actuaciones. Una institución colaborativa implica y compromete a diversos agentes en sus acciones Unas instituciones participativas favorecen el derecho de las personas a participar activamente en la conformación de políticas públicas.

Aplicar el gobierno abierto implica instaurar medidas en tres ámbitos:

- Transparencia: publicación de datos e información que pueda ayudar a la comprensión, escrutinio y análisis de la función pública.
- *Open data*: publicación de información del sector público en formatos que permitan su reutilización por parte de terceros para la generación de nuevo valor, lo que se conoce por RISP, Reutilización de la Información del Sector Público.
- *Open process:* la creación de canales de interacción a través de los cuales la ciudadanía pueda opinar, instar, solicitar, aportar y colaborar.

El **objetivo último del gobierno abierto es hacer realidad la idea original de la democracia porque establece canales de comunicación y contacto directo entre la ciudadanía y la Administración gracias a las utilidades que ofrece**

* Ingeniera técnica en Informática, cofundadora de la Iniciativa Barcelona Open Data.

Internet, porque la idea originaria de la democracia es gobernar entre todas las personas que forman parte de una comunidad.

Open data, una de las patas de la gobernanza abierta

El concepto de datos abiertos u *open data* se refiere a la exposición pública de información, siempre de una forma adecuada para su acceso y libre reutilización por la ciudadanía, empresas u otros organismos. Los datos abiertos son aquellos que pueden ser utilizados, reutilizados y redistribuidos libremente por cualquier persona, y que se encuentran sujetos, como mucho, al requerimiento de atribución y de compartirse de la misma manera en que aparecen.

Una estrategia de *open data* de una institución debe cumplir una serie de criterios:

- Aplicar *open data* por defecto, de este modo, toda la información pública debe ser expuesta públicamente por defecto y permitir su reutilización, con excepción de la información personal o datos que vulneren la seguridad.
- La información debe publicarse con una descripción sobre su calidad y la metodología para facilitar su reutilización. Deben ofrecerse los datos primarios de forma adecuada para ser útiles.
- Optar por publicar la mayor cantidad y variedad de conjuntos de datos para posibilitar las máximas opciones como recurso.
- Datos usables de forma universal. La información debe ser accesible para cualquier persona o colectivo al que le sea de interés, permitiéndose el uso gratuito o al costo mínimo de reproducción de la misma, sin establecer barreras físicas, administrativas ni burocráticas.
- Datos sin restricciones de uso. Las formas de distribución no estarán sujetas a restricciones de uso, de forma que cualquier persona pueda interpretar y utilizar los recursos a través de herramientas gratuitas o de uso común.

Los datos abiertos suponen una mejora democrática ya que su publicación hace "real" su carácter público y ponen al alcance de toda la ciudadanía información útil y utilizable que, a su vez, es un excelente recurso para el desarrollo de aplicaciones y servicios con un elevado valor social.

El impacto positivo que los datos abiertos tienen en la sociedad puede resumirse en cuatro pilares: la mejora de los gobiernos; su aportación para la resolución de problemas públicos; el empoderamiento de la ciudadanía y la generación de oportunidades económicas más equitativas.

Las estrategias de datos abiertos mejoran a los gobiernos ya que aportan información útil en la lucha contra la corrupción, aumentan la transparencia y mejoran los servicios públicos y la asignación de recursos.

Los datos abiertos están jugando un papel cada vez más importante en la solución de grandes problemas públicos. Nos aportan nuevas formas de evaluación basada en datos y permiten una participación más directa y colaborativa basada en esos datos.

Empoderar y formar a la ciudadanía es otra de las consecuencias de la apertura de datos. Por esta razón es importante considerar como un derecho el acceso a los datos públicos de las instituciones.

Los datos abiertos acompañados de nuevas formas de comunicación y acceso a la información permiten la toma de decisiones a una ciudadanía más y mejor informada y fomenta o permite nuevas formas de movilización social.

Además del impacto positivo de los datos abiertos sobre la gobernanza, tienen un gran potencial económico. La utilización de los datos abiertos son la base de una gran cantidad de actividades innovadoras y de productos y servicios que se pueden crear lo que genera crecimiento económico y de empleo.

La conciencia en el valor de datos y la demanda *open data*

> Existe un boom informativo sobre la importancia del big data y el análisis de datos aplicados a la estrategia de las empresas, sobre el valor de los datos que las personas ceden a los servicios de Internet. "Los datos son la nueva utility de las empresas" afirman diversas personas expertas en la materia.

Quienes defienden el *open data* argumentan que la información que la Administración genera con los recursos públicos debe ser expuesta para poder ser consultada o reutilizada. Así, aportar datos de calidad desde las instituciones a la red para construir la sociedad del conocimiento significa tener información disponible y accesible de manera fácil para todas las personas. Cuando el primer ministro Gordon Brown presentó el portal de datos abiertos del Reino Unido (data.gov.uk) en 2009 explicó que esta iniciativa tenía el objetivo de abrir oportunidades para las empresas, aumentar la transparencia y otorgar poder a los y las consumidoras.

Existe una conciencia social creciente sobre el valor de los datos —"los datos son el petróleo del XXI"— que va a hacer crecer la demanda social del *open data*,

del derecho al acceso a la información y reutilización de información pública, generada con los recursos de toda la ciudadanía.

Gestión DATA en instituciones: *open data* & *big data*

"La Casa Blanca nombra al *primer chief data scientist* de EE. UU.", así titularon algunos medios de comunicación el nombramiento de D. J. Patil, una apuesta por la innovación pública. Obama continuaba marcando un hito en la innovación publica, como ya había hecho nombrando el primer *chief information officers* (CIO) a nivel nacional en la historia de Estados Unidos en 2009.

Fue una noticia relevante para el sector de los datos ya que mostró cómo estos estaban ocupando una posición cada vez más importante dentro de la toma de decisiones en las organizaciones más importantes del mundo. Además el perfil del elegido, Patil habia trabajado en varias de la grandes compañías que más están sacando partido de los macrodatos: PayPal, LinkedIn, eBay y Skype.

Este nombramiento supone la evolución de la Admnistración Obama, que lanzó y se comprometió con el concepto de *open goverment*, que por supuesto ha implicado iniciativas pioneras en *open data*. Se produjo un cambio cuando en 2009 se lanzó la web de publicación de los datos del gobierno: Data.gov y un compromiso público por la publicación sistemática de datos abiertos. En 2013 la orden ejecutiva de gobierno abierto, por la que "la información del Gobierno por defecto se publicará en abierto y lectura automática", ha abierto mas de 138.000 bases de datos. A final de su primer mandato decidió utilizar *big data* para su reelección en 2012, creó en su equipo de campaña un departamento de analítica de datos. Así incorpora el análisis de datos para la toma de decisiones inteligentes al gobierno de EE. UU.

En este sentido aparecen figuras emergentes y con un papel creciente en las Administraciones públicas como el *chief data officer,* una figura directa responsable de la gestión integral de los datos, con la responsabilidad de impulsar las políticas de *open data*, como de aplicar análisis de datos o *big data* para la toma de decisiones inteligentes. La gestión global DATA o de los datos confluyen y deben tener una gobernanza única y gestión común, ya que los procesos también confluyen.

› Cambios organizativos

Digitalización de las empresas. Empresa 2.0 & gobierno abierto & instituciones *open big data*

En el proceso de asunción de las TIC por parte de las empresas en la primea fase se produjo la digitalización u automatización de algunos procesos, con la revolución de Internet apareció un cambio cultural, la denominada empresa 2.0. En el caso de la Administración pública se transita por la e-administración y la adopción de las redes sociales: los primeros pasos del gobierno abierto.

Actualmente la transformación digital coincide con la revolución de los datos. En esta aparecen las empresas *data driven* y las instituciones *open big data*, que asumen el *open data* y se convierten en proveedoras de datos, que aplican el análisis de datos masivos o *big data* para las decisiones públicas inteligentes.

Empresa 2.0

Del mismo modo que **Tom O'Reilly introdujo el término web2.0**, el primero en nombrar ese concepto aplicado a la empresa fue **McAffe**, en la primavera de 2006, y lo denominó empresa 2.0 (Cook, 2008). Macfee (2006) acuñó el término empresa 2.0 para referirse al uso de las plataformas emergentes de *software* social en el interior de las empresas o entre empresas y sus socios y clientes.

Emergen nuevas formas de organizar la actividad empresarial como resultado de la interacción entre tecnologías digitales y cambio organizativo. Es decir, la integración de las TIC en la actividad productiva ha permitido la aplicación de nuevos conocimientos e informaciones sobre aparatos de generación de conocimiento y procesamiento de la información y de comunicación (Castells, 1997). Esta visión de la empresa supera los conceptos tradicionales como los niveles jerárquicos, la responsabilidad o la división funcional, y potencia los procesos de externalización de actividades y funciones de acuerdo con criterios económicos y estratégicos (Lonsdale y Cox, 2000 citado por Torrent, 2008b). Las empresas están utilizando las nuevas tecnologías de la web para transformar las prácticas de negocios y reinventar la dirección.

Democratización-Proconsumidor

El entorno de la empresa 2.0 se caracteriza no solo por el incremento de usuarios sino por la participación creciente de estos en la gestión de los contenidos. Aparece el concepto de *prosumer*: un consumidor proactivo, informado y experto, que ejerce una gran influencia sobre los demás. Existen diversos niveles de

participación de los clientes y en los niveles máximos, las comunidades de usuarios participan en el el codiseño y la cocreación de productos y servicios.

Las seis características de las empresas 2.0

1. La empresa red solo es posible a partir de un cambio cultural interno. Requiere una cultura empresarial interna que sitúe el trabajo en red en el centro de su propia definición.
2. La empresa red combina activos especializados, frecuentemente intangibles, bajo un control compartido. **La integración estratégica de los proveedores y clientes** con una visión global de todos los recursos para el logro de objetivos, bajo una cultura empresarial común.
3. La empresa red se fundamenta en una **toma de decisiones basada en el conocimiento y no en la jerarquía**. Sitúa el conocimiento específico del puesto de trabajo en el epicentro de las decisiones, sustituyendo progresivamente las relaciones jerárquicas. Transforma la relación contractual del trabajador con un nuevo modelo de contraprestación centrado en el control de las actividades y en la toma de decisiones.
4. La gestión de la información y del conocimiento en la empresa red se basa en unas **comunicaciones directas, que incluyen el conjunto de todos sus nodos**.
5. La empresa red se organiza en **equipos de trabajo multidisciplinarios** de configuración variable. La especialización basada en el conocimiento y las comunicaciones directas permiten la configuración de grupos de trabajo multidisciplinarios, variables y específicos para cada proyecto, lo que rompe las tradicionales barreras de las áreas función nacionales. Una vez alcanzados los objetivos estratégicos de los proyectos, estos **equipos se reubican flexiblemente** en otros proyectos.
6. Las relaciones de los integrantes de la empresa red superan las tradicionales vinculaciones contractuales basadas en el precio, las características funcionales y el nivel de servicio. El elevado grado de integración estratégica definido por la empresa red hace insuficientes los tres elementos anteriores, que caracterizan la vinculación entre dos empresas. Nuevas variables, como la capacidad de adaptación a diferentes culturas empresariales y la confianza para compartir información relevante, se configuran como significativas en las relaciones entre las diferentes unidades de negocio en red.

Instituciones que adoptan open data*, proveedoras de datos*

Tras repasar diversos manifiestos, conclusiones de jornadas gubernamentales a nivel internacional y organizaciones que promueven el gobierno abierto en general,

y el open data en particular, podríamos decir que estas son las características de una institución *open data*.

- Disponer de políticas publicas de datos abiertos, establece las acciones de la institución para hacer posible la publicación y el acceso de las personas a la información pública.
- Establecen por defecto la apertura de datos, apostando por la divulgación proactiva de información pública —por medio de Internet— y en formatos abiertos.
- Tener prevista la publicación de nueva información específica y la actualización de la información disponible periódicamente.
- Estipular una política de datos abiertos para entidades semipúblicas.
- Determinar la publicación de datos abiertos en contratos para manejar, investigar o generar datos.
- Disponer de una governanza de los datos y de responsables del gobierno de los datos en la organización.
- Custodiar apropiadamente la información sensible.
- Requerir que la excepción de publicación de ciertos datos sea por razón de estar considerados en disposiciones que reconozcan el interés público en determinar si la información será compartida o no.
- Remover las restricciones al acceso a Información.
- Disponer de un portal web especifico de datos abiertos, para facilitar la distribución de estos actuando como un punto de búsqueda de fácil acceso.
- Generar un entorno de fácil entendimiento de la información, que permita a la ciudadanía decidir qué datos, en qué formato y qué tipo de visualización quieren.
- Incorporar la colaboración de la ciudadanía en la recolección de datos, en aquellos tipos de datos que sea posible.
- Establecer acciones para facilitar el acceso y reutilización de los datos por parte de la ciudadanía.
- Crear procesos para asegurar la calidad de los datos.
- Generar alianzas público-privadas para disponer de una variedad importante de datos y fomentar su reutilización.

Empresas orientadas a datos u organizaciones data driven

Las empresas *data driven* o empresas orientada a datos son empresas basadas en la información, empresas donde los datos son el epicentro de los procesos y toma de decisiones. Una empresa *data driven* precisa de un control absoluto de la información, lográndolo mediante la fusión y homogeneización de los datos procedentes de las distintas fuentes de información que maneja. Una empresa

data driven necesita de métodos de análisis intuitivos y muy visuales para optimizar y mejorar la toma de decisiones.

Las caracteristicas de la empresa *data driven*:

1. Se asienta sobre una política de analítica de datos muy potente con una cultura orientada al dato. Se definen los procesos pensando en indicadores e informar al personal para que pueda trabajar con objetivos.
2. Datos centralizados y organizados. Para que los procesos funcionen, los datos han de ser de calidad, esto implica, entre otras cosas, estar actualizados constantemente. Para ello, es imprescindible que la información esté organizada y sea de fácil acceso.
3. Establece una gobernanza de acceso a los datos. La gobernanza de la información ha de ser sofisticada, de modo que el *self service* no esté reñido con la seguridad y haya diversos tipos de acceso.
4. Dispone de herramientas de análisis de datos integradas. La capacidad de análisis ha de ser capaz de integrarse en las plataformas más habituales, incluso tradicionales de la organización. Así se garantiza agilidad, calidad de datos y que el personal esté dispuesto a participar en la cultura del dato.
5. Adoptar estas características tiene como consecuencias otros cambios estructurales, así encontramos que: la dirección y los trabajadores tienen una comunicación directa, la evaluación de trabajadores se realiza según indicadores y objetivos cumplidos y la toma de decisiones se produce de forma rápida.

Las instituciones open big data

Las instituciones *open big data* como productoras de datos deberían asumir las características de una institución *open data* y como consumidoras de datos, deberían adoptar las características de una organización *data driven* enfocada a tener datos para la governanza inteligente. En resumen podríamos resumirlo en que:

1. Trabaja en red internamente.
2. Integra en los procesos a diferentes agentes: sociedad civil, empresas, ciudadanía.
3. Toma de decisiones basadas en conocimiento y datos.
4. Comunicaciones con todas las personas trabajadoras.
5. Comunicaciones abiertas a la ciudadanía.

6. Incorpora una cultura del dato.
7. Gobierna sus datos.
8. Dispone de herramientas de análisis de datos para las decisiones.
9. Publica sus datos abiertos y reutilizables.
10. Facilita el acceso y reutilización de los datos por parte de la ciudadanía.

Herramientas instituciones *open big data*

El impulso de una iniciativa *open data* por parte de una institución debe ir acompañada de la capacidad de dotarse de una serie de herramientas, que hagan sostenible y viable esta opción. Una iniciativa *open data* habitualmente supone un cambio de paradigma, además de afectar a las relaciones internas y a las relaciones externas. A las relaciones internas, debido a la necesaria alineación de la organización para proveer de datos que hagan posible la publicación; y a las relaciones externas ya que el portal de datos abiertos precisamente está dirigido al exterior.

Normativa open data

Es necesario adoptar una normativa que avale la apertura de datos. Se trata de una herramienta marco e impulsora del proceso, que legitime las acciones necesarias y que implique impulso y compromiso político.

Estrategia o plan open data *de la institución*

La elaboración de una estrategia y planificación global que establezca claramente nuestros objetivos debe establecer unas pautas claras sobre qué datos deben ser publicados y de qué manera. Para conseguir que la publicación de datos sea incremental, se debe fijar un calendario de los datos que se publican. Además se deberá indicar quién será la salvaguarda de dichos datos considerándolo como fuente legítima.

Los planes *open data* deben contener un aprovisionamiento de los recursos humanos y económicos necesarios para poder llevarse a cabo. Además de un dimensionamiento de la tecnología (*software* y *hardware*), que incluya cuestiones de seguridad y escalabilidad.

Aunque el objetivo final de una iniciativa de datos abiertos siempre es **publicar todos los datos disponibles** en función de los recursos disponibles y el tiempo, se deben establecer las prioridades sobre la información más valiosa. La priorización debe adecuarse a cada caso, pero existen algunos criterios generales que pueden servir: las exigencias de la normativa y leyes aplicables, la importancia

de los datos para la sociedad, la importancia de los datos para la economía, la actualidad de los datos, la cantidad de datos y el nivel de detalle.

Debe contemplar de forma específica un **plan de integrabilidad**, cuya meta principal consiste en lograr que todo aquel organismo que sea fuente auténtica de algún dato sea capaz de proveerlo al resto de los organismos que lo requieran, en lugar de solicitárselo. De este modo se pueden conseguir sistemas de gestión eficientes, para proveer información correcta, completa, vigente y segura.

Herramientas de participación

Una iniciativa de este tipo debe dotarse de un proceso y un espacio de participación en el que deberían formar parte proveedores de datos, consumidores de datos, activistas protransparencia y *open data*.

Además la propia institución debería participar en redes y proyectos junto con otras instituciones homólogas para compartir experiencias en planes de esta materia.

Plataformas de publicación de datos es una pieza clave para dar el soporte tecnológico a una estrategia de apertura. Se debe realizar una selección que nos permita llevar a cabo nuestros objetivos. Idealmente, se debería desarrollar una plataforma personalizada que se ajuste completamente a nuestras necesidades, pero también existen productos previamente "empaquetados" y preparados para su uso inmediato que podrían ser una opción más que razonable en algunos casos.

A continuación se detallan algunas plataformas *tecnológicas* para crear catálogos de datos y visualizar la información.

CKAN

Plataforma *open source* desarrollada por Open Knowledge. Utiliza muchas soluciones de *software* libre probadas y maduras en un solo paquete integrado que permite con poco trabajo tener un portal de datos abiertos funcional. CKAN integra herramientas como Apache, PostgreSQL, Json, AJAX u OpenStreetMap. Se integra fácilmente con otras herramientas como WordPress y Drupal, que permiten agregar a una plataforma de datos otro tipo de herramientas como blogs o páginas corporativas. Opera en local o si algún proveedor lo diera, en modo servicio. La organización que lo adopte debe alojarlo y parametrizar sus funcionalidades.

Entre sus usuarios más destacados están los gobiernos de Brasil y Reino Unido.

SOCRATA

Sócrata proporciona una plataforma comercial para agilizar la publicación de datos, la gestión, el análisis y la reutilización. Ofrece varios módulos adicionales que permiten realizar gran variedad de acciones a los usuarios para acceder, visualizar y compartir los datos. Opera en *cloud*. A todos los conjuntos de datos alojados en Sócrata se puede acceder mediante la API REST. Sócrata tiene la capacidad de personalizar el conjunto de datos de metadatos de acuerdo a las necesidades de las personas que lo consultan.

Muy utilizado en EE. UU., por ejemplo, los portales de Chicago y Nueva York.

OPEN DATA SOFT

Plataforma comercial especializada en la gestión de portales de datos abiertos, su análisis y reutilización. Ofrece grandes prestaciones respecto a la posibilidad de personalizar el conjunto de datos según las demandas. Opera en cloud. Tiene algún desarrollo como *soft* libre pero no el core de la herramienta. **Utilizado en Francia.**

› Nuevas necesidades: organizativas & profesionales

Nuevas necesidades organizativas

Una vez definida la estrategia o plan de *open data*, la institución debe adaptar sus procedimientos y protocolos para adoptar las medidas estructurales, orgánicas, técnicas y legales planificadas a corto y largo plazo en el plan de datos abiertos.

La institución debe aplicar una serie de acciones internas que deberían incluir, entre otros aspectos: la formación interna del personal involucrado en el proyecto y la designación de responsables, la creación de órganos con competencias relativas a la gestión y publicación de la información pública, aplicación del cambio en las políticas sobre la gestión y publicación de dicha información —ya que la adopción de *open data* implica un cambio de paradigma respecto a la gestión tradicional de la información pública—.

Una estrategia de datos abiertos incluye una serie de pasos, que por su propia naturaleza suponen la implicación transversal de la organización y la adopción de algunos cambios en los procesos para realizar y mantener una estrategia de datos abiertos (productor de información pública abierta), como son: la identificación y selección de la información potencialmente reutilizable, diseño de un

procedimiento para la preparación y generación de conjuntos de datos y documentos, establecimiento de las condiciones de reutilización, creación de un espacio web específico, publicación y mantenimiento de los datos y documentos, que serán catalogados y accesibles, definición de una estrategia de comunicación y formación, definición de una estrategia de evaluación y mejora.

Realizaremos algunos apuntes sobre las implicaciones en los procesos respecto a las principales acciones de una estrategia de datos. El desarrollo y mantenimiento del portal de datos abiertos supone un procesos de recopilación y publicación de los datos y documentos, que requiere la identificación en los sistemas internos y responsables de que la información esté autentificada. La promoción de la reutilización de la información para instituciones implica diseño de formación al personal y actividades divulgativas dirigidas a los colectivos reutilizadores.

Una vez puesta en marcha la iniciativa se debe realizar un análisis que contemple la monitorización y ajuste de forma periódica.

Formación para las nuevas necesidades profesionales

Nuevas necesidades profesionales

Los retos de la adopción de una cultura de datos pasan por una figura directiva responsable de los procesos de datos como es el *chief data officer* o director de datos. Esta persona es la líder estratégica de la información que debe ejercer el liderazgo en el cambio y ser una pieza clave para el CEO de la organización. Debe elaborar e implementar la estrategia de datos, estándares, procedimientos, políticas y gestionar los equipos de expertos de datos a nivel corporativo.

Sus funciones pasan por definir la estrategia de datos y liderar los planes que hagan de su institución una capaz de identificar, combinar y administrar múltiples fuentes de datos. Construir modelos para predecir y optimizar resultados, siempre acompañado de otras personas directivas que colaboren para hacer una institución data driven. También debe potenciar la información como activo estratégico del negocio y como motor de ingresos y ahorros, asimismo, sin perder de vista la capacidad de convertir las métricas en objetivos.

Un aspecto de gestión compleja para alinear el negocio con los datos masivos es la existencia de silos de datos en diferentes departamentos (ventas, *marketing*, RR. HH., etc.), cada uno de ellos restringido y controlado. Hay razones para que los silos de datos existan, pero si estos no están disponibles para la persona adecuada, se están estableciendo barreras incluso antes de empezar a resolver el problema.

Existen otras figuras a parte del *chief data officer* que deberían formar parte del equipo data de la organización. Es el caso del científico de datos, que recopila data, la analiza y crea modelos de predicción; el *marketing* de datos que lidera y ejecuta la estrategia de *big data marketing* interaccionando con la clientela basándose en métricas concretas; el *data visualization*, que busca, interpreta, contrasta y compara datos para permitir un conocimiento en profundidad para transformarlo en información comprensible para las personas; y el periodista de datos, que recaba y analiza grandes cantidades de datos mediante *software* especializado y hace comprensible la información a la audiencia.

Pero si hay una figura importante en el manejo de datos y que en unos meses será muy común en las instituciones públicas y empresas es la del *data protection chief* o jefe de protección de datos. Se trata de una nueva obligación que recoge el nuevo reglamento de la Unión Europea (Reglamento (UE) 2016/679 del Parlamento Europeo). Este nuevo reglamento supone una garantía adicional a la ciudadanía europea y, por lo tanto, un mayor compromiso de las organizaciones, tanto públicas como privadas, con la protección de datos y obliga a muchas instituciones y empresas a incorporar una nueva figura: el delegado de protección de datos, lo que podemos denominar como *data protection chief*.

Las funciones del *data protection chief* serán asegurar el cumplimiento normativo de la protección de datos haciendo compatible el funcionamiento de la organización. Además, para poder ejercer deberá acreditar formación y conocimientos especializados en materia de protección de datos.

Formación para nuevas necesidades profesionales

Formación universitaria superior

Existen diferentes categorías formativas por lo que se refiere a formarse en *big data*. Pese que muchos programas no solo se centran en un ámbito concreto de actuación, según los ámbitos de aplicación principal de cada máster o curso universitario superior los podemos clasificar en las siguientes áreas de conocimiento: modelos analíticos y estadísticos, aplicación *big data* a *business* y arquitecturas y tecnologías para *big data*.

Modelos analíticos y estadísticos

Su objetivo es formar expertos en el uso y gestión de grandes volúmenes de datos, tanto desde un punto de vista analítico como tecnológico. Basan su formación en modelos y técnicas avanzadas para el análisis de formación, de manera

que están dirigidos a ingenieros, graduados en matemáticas, física, TIC..., que quieran desarrollarse profesionalmente como científicos de datos.

Máster	Universidad	Perfil	Modalidad
Modelos analíticos y estadísticos			
Máster de Inteligencia de Negocio y Big Data (ITINERARIO ANÁLISIS DE DATOS)	Universitat Oberta de Catalunya (UOC)	Técnicos e ingenieros informáticos o de telecomunicación Analistas de datos en departamentos de control de gestión u otros Matemáticos o candidatos con una experiencia profesional equivalente	*Online*
Master de Fonaments de la Ciència de Dades	Universitat de Barcelona (UB)	Graduados en ingeniería informática, matemáticas, física y estadística	Presencial Barcelona
Introduction to Data Science and Big Data	Universitat de Barcelona (UB)	Graduados de informática, ciencias aplicadas, matemáticas, estadísticas e ingenierías	Presencial Barcelona
Master Executive en Big Data Science	Universitat Internacional de Catalunya (UIC)	Analistas de datos y estadísticos Programadores o personas con conocimiento de programación Personas con conocimientos de bases de datos: bases de datos relacionales (SQL) y de sistemas operativos	Presencial Barcelona
Master on Intelligent Interactive Systems (track big data)	Universitat Pompeu Fabra (UPF)	Graduados en informática, matemáticas, física o ingenierías	Presencial Barcelona
Master in Big Data Analytics	Universidad Carlos III de Madrid (UC3M)	Graduados e ingenieros en informática, telecomunicaciones, estadística, matemáticas, física o ingeniería industrial	Presencial Madrid
Programa en SAS Data Scientist	Madrid School of *Marketing* (MSMK)	Estudios: ingenierías, matemáticas, informática economía, estadística Profesiones: *project manager,* programación, estadística, ingeniería, consultoría	Presencial Madrid
Master en Data Science & Business Discovery	Madrid School of *Marketing* (MSMK)	Estudios: Ingenierías, matemáticas, informática economía, estadística Profesiones: *project manager,* programación, estadística, ingeniería, consultoría	Presencial Madrid

Aplicación *big data* a la empresa o *business*

Dirigido a un abanico más amplio de profesiones, este tipo de programas potencian la carrera de sus estudiantes en áreas concretas de las organizaciones incorporando el uso del *big data*, para la toma de decisiones. Es lo que llamamos aplicar el big data al *management*.

Máster	Universidad	Perfil	Modalidad
Aplicación *big data* al negocio			
Master of Science in Management (Business Analytics)	Universitat Pompeu Fabra (UPF)	Economía y administración de empresas Ingenieros, matemáticos y físicos	Presencial Barcelona
Postgrado en *Marketing* Intelligence & Big Data	School of Continuing Education – UB	Tecnológico (ingenierías, departamentos de IT, *business intelligence*, etc.) Analíticos (estadistas, sociólogos, matemáticos, físicos, etc.) Empresa y economía, *marketing* o ventas, CRM, *Marketing Research*, desarrollo de mercado, etc. Otros ámbitos como la consultoría, agencias de comunicación y *marketing* relacional, etc.	*Online*
Global Executive Master in Business Intelligence	INSA Business, *Marketing* & Communication School	Recién licenciados y profesionales Empresas que busquen especialización de su personal Titulados en Formación Profesional Grado Medio o de Grado Superior que buscan una preparación práctica para entrar en el mundo laboral.	Presencial Barcelona
Master Big Data & Data Intelligence	INSA Business, *Marketing* & Communication School	Recién licenciados y profesionales Empresas que busquen especialización de su personal. Titulados en Formación Profesional Grado Medio o de Grado Superior que buscan una preparación práctica para entrar en el mundo laboral	Presencial Barcelona
Master Open Big Data Management	Euncet Business School	Personas directivas de pymes que quieran actualizarse en la dirección basada en datos Directivos y directivas de instituciones que quieran especializarse en dirigir procesos de gobierno abierto Personas tituladas universitarias que deseen completar su formación con un conocimiento profundo de la dirección inteligente basada en datos	Presencial Barcelona

Máster	Universidad	Perfil	Modalidad
Aplicación *big data* al negocio			
Master en Business Intelligence y Big Data	Escuela de Organización Industrial (EOI)	Técnico (ing. de cualquier rama de las TIC) Estadístico (matemáticos o profesionales de diferentes campos científicos) Negocio (profesionales de cualquier área que quieran formarse como analista de datos y analistas de negocio)	*Online*
Master of Science in Management (IT Management)	Universitat Pompeu Fabra (UPF)	Ingenieros IT: informática y telecomunicaciones Matemáticas y física Ciencias sociales: economía y administración de empresas	Presencial Barcelona
Programa Ejecutivo en Big Data & Business Analytics	Escuela de Organización Industrial (EOI)	Directivo: directivos comerciales y de *marketing*, financieros, de IT y operaciones, gerentes de consultoría	Presencial Madrid
Master en Data Management e innovación tecnológica	OBS Business School	Directivos y profesionales (*data manager*)	*Online*

Arquitecturas y tecnología de *big data*

Este tipo de formación cuenta con un componente tecnológico elevado que va desde el diseño de *datacenters* escalables, hasta el análisis de datos en plataformas web y móviles, *software* para procesar y almacenamiento escalable de información incluidas en el ecosistema Hadoop. Este tipo de programas van dirigidos a carreras técnicas como son las ingenierías informáticas.

Máster	Universidad	Perfil	Modalidad
Arquitecturas y herramientas de Business Intelligence y *big data*			
Máster de Inteligencia de Negocio y Big Data (ITINERARIO SISTEMAS DE INFORMACIÓN)	Universitat Oberta de Catalunya (UOC)	Técnicos e ingenieros informáticos o de telecomunicación Analistas de datos en departamentos de control de gestión u otros Matemáticos o candidatos con una experiencia profesional equivalente	*Online*
Posgrado en Sistemas de inteligencia de negocio y Big Data	Universitat Oberta de Catalunya (UOC)	Técnicos e ingenieros informáticos o de telecomunicación Analistas de datos en departamentos de control de gestión u otros Matemáticos o candidatos con una experiencia profesional equivalente	*Online*

Máster	Universidad	Perfil	Modalidad
Arquitecturas y herramientas de Business Intelligence y *big data*			
Postgrado Big Data Management and analytics	Universitat Politécnica de Catalunya (UPC)	Profesionales informáticos : desarrollador, arquitecto, analista de datos y administrador de sistemas	Presencial Barcelona
Máster en Big Data	La Salle	Ingenieros y graduados en Informática y profesionales Titulados universitarios de Grado, Máster, Ingenierías Técnicas y Superiores Profesionales provenientes de ciclos formativos con experiencia en redes, programación y cálculo	Presencial Barcelona
Master in Data Mining and Business Intelligence	Universitat Politécnica de Catalunya (UPC)	Graduados e ingenieros en informáticas, matemáticas y estadística (buen nivel de inglés)	Presencial Barcelona
Postgrado Big Data Management and analytics	Universitat Politécnica de Catalunya (UPC)	Profesionales informáticos : desarrollador, arquitecto, analista de datos y administrador de sistemas	Presencial Barcelona

Master Open Big Data Management

El Master Open Big Data Management está impulsado por Euncet Business School con la idea de acercar los conocimientos de *big data* a las pequeñas y medianas empresas y ofrecer unos estudios en *big data* orientado al management, o dicho de otra forma, el management basado en datos, lo que denominamos *management* en *open data* y *big data*.

De este modo este máster tiene como objetivo formar a personas directivas capaces de tomar decisiones inteligentes basadas en datos, especialmente, datos abiertos, y está impregnado de la cultura *open* y del trabajo en red.

El programa académico se concreta en dos itinerarios y titulaciones; el Postgrado en Gobernanza Abierta centrado en los procesos de gobierno abierto, específicamente *open data*; y el Master Open Big Data Management enfocado a la toma de decisiones inteligentes basadas en datos, especialmente, abiertos y/o gratuitos.

Salidas laborales según los tipos de estudios descritos

Orientados al análisis de datos	Científicos y analistas de datos, tratamiento de datos, visualización de datos
Aplicación de *big data* al negocio	*Marketing* de datos, datos para la estrategia, *marketing* para la dirección, *chief data officer* o gobernanza de datos
Herramientas tecnológicas *big data*	Programadores en *software* de analistas y tratamiento de grandes volúmenes de datos, diseñadores de *datasets* escalables, diseñadores de plataformas web para gestión de datos

Ofertas de formación *online*

Una de las opciones que está proliferando es la formación *online*, con un enfoque profesional, que realiza unas ofertas cortas y modulares, adaptándose así a las necesidades de las personas. Podríamos hablar de ofertas microcontenidos, la mayoría de veces estos módulos o microcontenidos se pueden ir completando con otros para conseguir unas competencias más globales de una materia. Por tanto se trata de una formación más flexible, no solo por las características tradicionales de la formación *online*: tiempo y el momento lo elige el usuario, asincrónica, sino que también en cuanto a los contenidos, es la persona quien elige a medida cuál es su "carrera", qué certificados quiere obtener. Así una de las cuestiones más populares son los MOOC —es el acrónimo en inglés de *massive online open courses* (o cursos *online* masivos y abiertos)—. Permiten hacer un curso gratis en Harvard o Stanford y conectar con gente que estudia en las principales universidades del mundo.

Algunas de las plataformas donde se pueden encontrar MOOCS de universidades de referencia, y en las que encontrarás formación de calidad sobre *big data* (no repasamos cada curso, ya que la oferta es muy dinámica, y más en los estudios TIC), son: https://miriadax.net/, https://es.coursera.org/ o el caso de http://www.ucatx.cat/, plataforma de las universidades catalanas.

Bloque II
Estudio de casos

Capítulo 13

Big data para el bien social: oportunidades y retos

Nuria Oliver*

Vivimos en un mundo de datos, de *big data*, cantidades ingentes de datos que crecen de manera exponencial, lo que dificulta nuestra capacidad para entender de cuántos datos estamos hablando. Solo en 2016, se estima que el tráfico de Internet alcanzará los 1.3 *zettabytes* de datos, según un estudio de Cisco[1], es decir, 10 a la potencia de 21 *bytes* o un millón de millones de *gigabytes*. Y esto son solo los datos que viajaran por Internet...

¿Pero de dónde vienen todos estos datos? Esta explosión de datos es fruto, por una parte, de la digitalización del mundo físico, sobre todo impulsada por la ubiquidad de los móviles y el crecimiento exponencial de todo tipo de sensores (muchos de ellos son parte de lo que se conoce como "Internet de las cosas" o IoT), de manera que nuestras acciones en el mundo físico generan una *huella digital* (por ejemplo, sensores de tráfico, transacciones de compras con tarjeta de crédito, etc.). Por otra parte, también son resultado de la cada vez más intensa actividad en el mundo digital, fomentada por la adopción masiva de *smartphones* conectados a Internet, con miles de aplicaciones móviles que también dejan un rastro digital como resultado de su uso.

Aunque parte de estos datos son datos generados por actividad sin vinculación directa a la actividad humana (por ejemplo, aceleradores de partículas, telescopios espaciales, mediciones del medioambiente, etc.), una parte importante de estos datos son generados como resultado de nuestra actividad, la de cada uno de nosotros.

De manera que, por primera vez en la historia de la humanidad desde que existimos como especie, somos capaces de medir el comportamiento humano a gran escala, impulsando el nacimiento de una nueva disciplina: las ciencias sociales

* Chief Data Scientist, DataPop Alliance, director of Research in Data Science, Vodafone.

1. http://www.cisco.com/c/dam/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf

computacionales[2]. Dónde vamos, qué consumimos, quiénes son nuestros amigos, cuánto dinero gastamos y en qué... son algunos de los aspectos de nuestro comportamiento que tienen un reflejo digital y, por tanto, son susceptibles de ser modelados computacionalmente. Estos modelos a partir de datos permiten *personalizar* nuestra interacción con la tecnología o con servicios concretos. Gracias a técnicas de personalización entrenadas a partir de nuestros datos recibimos recomendaciones[3] de películas relevantes en Netflix, libros potencialmente interesantes en Amazon, cupones de descuento para compras en nuestro supermercado favorito, anuncios y *banners* en Facebook o Google, etc.

Sin duda, el potencial de impacto de *big data* en el negocio es claro. Según el reciente informe de Accenture llamado "Gran éxito con grandes datos[4]", un 89% de los participantes en la encuesta han implementado un proyecto que utiliza *big data* en el negocio, creen que el *big data* revolucionará las operaciones del negocio y un 85% cree que cambiará dramáticamente cómo se hacen los negocios. La medicina, la banca, el periodismo, el comercio, el transporte, la producción industrial, la agricultura, el deporte, la ciencia...son algunas de las disciplinas que sin duda ya están aprovechando el *big data*. La firma de investigación de mercado IDC predice[5] un 50% de aumento en los ingresos de la venta de *software*, *hardware* y servicios relacionados con el *big data* y el análisis de datos entre el 2015 y el 2019, alcanzando los 189.000 millones de USD en 2019.

Pero hay una oportunidad del *big data* mucho menos conocida a pesar de su inmenso potencial: el *big data* para el bien social[6], que es el foco de este capítulo.

La existencia de datos de comportamiento humano a gran escala, datos anonimizados (es decir, preservando la privacidad de las personas) y agregados (es decir, analizados en su conjunto, no individualmente) nos permiten, por primera vez desde que existimos como especie, modelar y medir migraciones, cuantificar el impacto de desastres naturales y ayudar en la toma de decisiones a gobiernos e instituciones relevantes, anticipar la progresión de pandemias o determinar automáticamente el desarrollo socioeconómico de una región.

2. http://science.sciencemag.org/content/323/5915/721
3. http://www.cs.ubbcluj.ro/~gabis/DocDiplome/SistemeDeRecomandare/Recommender_systems_handbook.pdf
4. https://www.accenture.com/us-en/insight-big-data-research
5. http://www.idc.com/getdoc.jsp?containerId=IDC_P33195
6. http://telefonicacatalunya.com/wp-content/uploads/2014/09/ActualidadEconomica_Una_mina_de_datos_y_bunas_acciones.pdf

Los datos están presentes en los 17 Objetivos de Desarrollo Sostenible establecidos por Naciones Unidas a finales de 2015, tanto para permitir medir el progreso hacia la consecución de dichos objetivos como para conseguirlos.

Un artículo reciente publicado por United Nations Global Pulse[7] presentaba los retos y oportunidades planteados por el *big data* cuando se utiliza para el bien social y proponía una taxonomía con tres niveles de uso: para la "concienciación en tiempo real", para ayudar los sistemas de "aviso temprano" y para proporcionar "*feedback* en tiempo real". Un artículo posterior[8] sobre *big data* para la prevención de conflictos distinguía entre tres funciones distintas: descriptivas (por ejemplo, mapas), predictivas (por ejemplo, para generar predicciones) y prescriptivas (para poder llevar a cabo inferencias causales. La mayor parte de las aplicaciones de *big data* para el bien social pertenecen hasta ahora a las dos primeras categorías, aunque ejemplos de casos de uso de la función prescriptiva seguramente crezca en el futuro.

Durante mis más de ocho años como directora científica en Telefónica I+D (2007-2016), creé y lideré un área de investigación dedicada al *big data* para el bien social, en colaboración con investigadores de mi equipo y con instituciones externas como Naciones Unidas, el Gobierno de México, el Programa de Alimentos Mundial, MIT, FBK y Data-Pop Alliance, donde soy *chief data scientist*. Durante estos años hemos validado el valor de datos capturados por la red de telefonía móvil para detectar las zonas más afectadas por inundaciones[9], para cuantificar el impacto de intervenciones gubernamentales ante un riesgo de pandemia[10], para inferir niveles socioeconómicos de una región[11] o para predecir crímenes[12] en la zona metropolitana de Londres, proyecto ganador del "Datathon para el Bien Social"[13] organizado por Telefónica en colaboración con MIT, la Campus Party de Londres y el Open Data Institute en 2013. En este capítulo resumiré dos de estas experiencias, incluyendo mis reflexiones derivadas de nuestro trabajo en estos proyectos.

Antes de explicar los proyectos en detalle, es importante destacar que afortunadamente, no somos los únicos. Orange ha organizado en dos ocasiones

7. http://unglobalpulse.org/projects/BigDataforDevelopment
8. https://www.ipinst.org/2013/04/new-technology-and-the-prevention-of-violence-and-conflict
9. http://unglobalpulse.org/sites/default/files/Mobile_flooding_WFP_Final.pdf
10. https://www.youtube.com/watch?v=H5_FeuuS-zs
11. http://www.vanessafriasmartinez.org/uploads/ictd2012Frias.pdf
12. https://arxiv.org/pdf/1409.2983.pdf
13. http://news.o2.co.uk/?press-release=telefonica-the-open-data-institute-and-the-mit-set-data-challenge-for-campus-party-2013

(2013 y 2015) un reto mundial para el análisis de datos de móviles con fines humanitarios llamado "Datos para el Desarrollo" o D4D[14] donde ha dado acceso a datos agregados y anonimizados de Costa de Marfil y Senegal a decenas de grupos de investigación a nivel mundial. El resultado han sido numerosos proyectos[15] donde se ha utilizado el *big data* para entender el transporte y las ciudades, mejorar las estadísticas oficiales, contribuir a la salud pública, identificar patrones relevantes entre el comportamiento humano y los desastres naturales, las necesidades energéticas o la agricultura, entre otros.

Telecom Italia también ha organizado un reto mundial de análisis de datos[16] con más de 1.000 participantes de 20 países diferentes que presentaron más de 100 ideas innovadoras que utilizaban el *big data* para varios fines, incluyendo fines sociales. Otras empresas de telecomunicaciones como Telenor y Vodafone, donde actualmente soy directora de investigación en ciencias de datos, también están desarrollando proyectos con el fin de tener impacto social positivo.

A nivel institucional, además de distintas unidades dentro de Naciones Unidas (en particular, United Nations Global Pulse y UNICEF, con quienes hemos colaborado), cabe destacar el trabajo de Flowminder[17], una organización sin ánimo de lucro, basada en Estocolmo (Suecia), que lleva desde 2008 trabajando en la mejora de la salud pública y el bienestar en países con niveles socioeconómicos medios y bajos a través del análisis de datos móviles agregados y anonimizados, datos de satélites y datos del censo; y el trabajo de Data-Pop Alliance[18], una organización sin ánimo de lucro con el respaldo de MIT, Harvard, el Instituto para el Desarrollo (ODI) y Flowminder, cuyo objetivo es promover una revolución centrada en las personas y en el *big data*.

En este capítulo, ilustraré el valor del *big data* —y en particular de los datos de la red de telefonía móvil— en el contexto del bien social a través de la descripción de dos proyectos en dos ámbitos distintos y de impacto: los desastres naturales y la criminalidad.

14. http://www.d4d.orange.com/en/Accueil
15. http://www.d4d.orange.com/fr/content/download/43453/406503/version/1/file/D4DChallengeSenegal_Book_of_Abstracts_Scientific_Papers.pdf
16. http://www.telecomitalia.com/content/tiportal/en/innovazione/tutte-le-news/big-data-challenge.html
17. http://www.flowminder.org/
18. http://datapopalliance.org/

El valor de *big data* ante los desastres naturales

El impacto económico de los desastres naturales a nivel mundial es inmenso. En los diez años posteriores al huracán Katrina en EE. UU., ha habido en el mundo una media de 260 desastres naturales, con un coste económico medio anual de 211.000 millones de dólares, 63.000 millones de dólares en pérdidas de seguros y 76000 vidas perdidas, según el informe de Aon sobre las catástrofes globales causadas por el clima[19].

Uno de los retos cuando se produce un desastre natural en una región consiste en identificar las zonas y estimar el número de personas afectadas, para poder dimensionar la ayuda enviada y determinar dónde enviarla. En este contexto, el *big data* y en particular los datos agregados y anonimizados procedentes de la red de telefonía móvil pueden ser de gran ayuda[20].

En esta sección presentaré un resumen de un proyecto que realizamos en colaboración con UN Global Pulse y la Universidad Politécnica de Madrid (UPM) con el objetivo de explorar el uso de *big data* como herramienta para acción humanitaria —asesorados por expertos del Gobierno de México y el Programa Mundial de Alimentos (PMA)—. Este resumen ha sido publicado[21] en el blog del Banco Iberoamericano para el Desarrollo (BID) y el trabajo fue publicado en el congreso internacional del IEEE GHTC 2014[22].

El estado de Tabasco (México) sufrió entre 2007 y 2011 numerosas inundaciones que generaron daños y pérdidas por más de 57.000 millones de pesos, según el PNUD. Considerando la inundación de finales de 2009 como caso de estudio, el objetivo del proyecto era averiguar si mediante *big data* es posible crear nuevos mecanismos de alerta temprana y gestión de inundaciones. Los resultados del proyecto demostraron que, aunque haya que tomar con cautela la posibilidad de anticipar una inundación real, sí es posible localizar las zonas más afectadas rápidamente y medir los intervalos de tiempo entre las lluvias y los efectos y acciones que desencadenan para la mejora de la gestión.

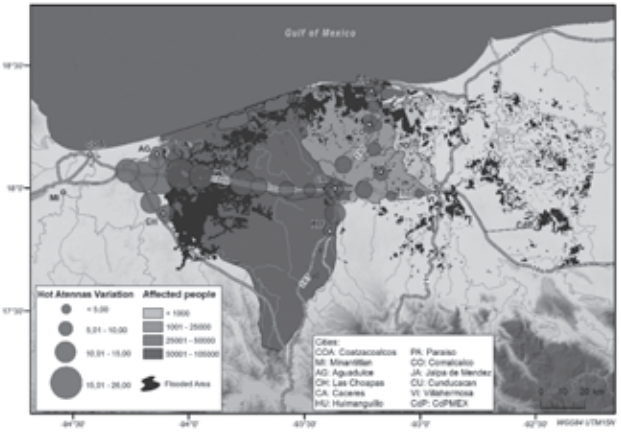
19. http://thoughtleadership.aonbenfield.com/sitepages/display.aspx?tl=460
20. http://journal.frontiersin.org/article/10.3389/fpubh.2015.00189/full
21. http://blogs.iadb.org/abierto-al-publico/2016/02/25/como-el-big-data-ayuda-en-la-gestion-de-desastres-naturales-el-caso-de-tabasco-en-2009/
22. "Flooding through the lens of mobile phone activity". Pastor- Escuredo, D., Morales-Guzmán, A. *et al.*, IEEE Global Humanitarian Technology Conference, GHTC, 2014.

El valor de datos de fuentes múltiples para abordar un problema real

En el estudio comparamos la reacción social (detectada a partir de patrones de comportamiento de las antenas de telefonía móvil —*big data*—), la actuación institucional (en particular, el aviso de protección civil) y el patrón de lluvias (factor externo). Para medir la reacción social analizamos registros anonimizados de llamadas de móviles, asegurando la privacidad de sus usuarios. La actuación institucional se recuperó mediante noticias, registros públicos y actas de protección civil sobre las inundaciones de 2009 y 2010 en México. Para medir el avance de las inundaciones utilizamos datos abiertos de satélites de la NASA.

Datos oficiales del censo verificaron que la distribución geográfica de los usuarios formaba una muestra representativa en la región. Los datos de teléfonos móviles, anonimizados y agregados, registraron la posición de la antena más cercana durante una llamada y como resultado convirtieron a las antenas en sensores de actividad y movilidad. De este modo, pudimos generar una medida de anomalías en el volumen de llamadas y crear mapas de movilidad. Un análisis de imágenes LANDSAT nos permitió detectar zonas inundadas con detalle. A su vez, datos del proyecto TRMM-NASA ofrecieron información diaria de las precipitaciones que sirvieron para comparar la evolución de la inundación con los patrones en los datos de móviles. Mediante análisis SIG, análisis de redes y series temporales fue posible alinear las fuentes y cuantificar la inundación y su impacto.

Figura 1. Mapa de impacto de las inundaciones de 2009 en Tabasco



El mapa muestra el día más crítico con los valores más altos de variación en los niveles de actividad en las torres celulares.

Conclusiones obtenidas a partir del análisis de los datos

Según los registros, Protección Civil envió una señal de alerta coincidiendo con el día de máximas precipitaciones. Sin embargo, el tráfico de llamadas no reveló ninguna variación sincronizada con esta alerta, sino días más tarde en zonas muy concretas como carreteras durante los peores efectos de la inundación. Este hallazgo es importante porque demuestra que, en caso de inundaciones, el nivel de concienciación ciudadana puede no incrementarse a tiempo, ignorando alertas institucionales, ya que incluso con fuertes lluvias, es difícil anticipar una inundación real. Así, a partir de nuestra experiencia en este proyecto, concluimos que este análisis puede ayudar a (1) mejorar la actuación localizando rápido las zonas más afectadas y dónde se condensa la población y (2) rediseñar y evaluar mecanismos de prevención.

Consecuencias de este estudio

El ejemplo de Tabasco destaca el potencial del uso de grandes datos y de datos abiertos agregados y anonimizados, siempre preservando la privacidad de las personas, para mejorar la respuesta a desastres naturales. Los datos pueden mejorar sustancialmente cualquiera de las fases de respuesta a un desastre natural: (1) al comienzo del desastre, con alerta temprana de una posible reacción social que pueda ser clasificada como anomalía antes del impacto físico del desastre; (2) durante el propio curso del desastre, para monitorizar a la población y optimizar la intervención humanitaria y la difusión de información; (3) tras la toma de decisiones para realimentar y adaptar estas decisiones y maximizar su eficacia (4) en retrospectiva, evaluando los hechos para el posterior diseño de políticas para la mejora de mitigación, capacidad de preparación y recuperación.

Para introducir estas innovaciones, es necesario desarrollar nuevos modelos de compartir datos de manera responsable que aumenten su disponibilidad en tiempo real. A su vez, datos abiertos y plataformas abiertas de visualización y análisis que aceleren el tiempo de respuesta para asesorar sobre los planes de intervención y al mismo tiempo preserven la privacidad de las personas.

En el siguiente apartado, describiré otro ejemplo del valor que el *big data* puede aportar para mejorar la calidad de vida, en este caso con relación a la criminalidad en las ciudades.

El valor de *big data* para modelar la criminalidad urbana

A pesar de su gran impacto en la calidad de vida y el desarrollo económico de una región[23] el crimen no ha sido estudiado extensivamente dentro del área de *big data for social good*, salvo algunos ejemplos[24, 25, 26].

Arquitectos y urbanistas han estudiado la relación entre la dinámica de una ciudad, las características del entorno urbano y el crimen. La activista urbana Jane Jacobs[27] enfatizó en los años 60 el papel de lo que ella denomina la *vigilancia natural* como un elemento clave para reducir el crimen: conforme las personas se mueven en un área, actúan como *ojos en la calle* capaces de observar lo que sucede a su alrededor. Por tanto, Jacobs sugiere que una alta diversidad en la población y un gran número de visitantes contribuyen a aumentar la seguridad de un área resultando en menor incidencia de crimen. Por el contrario, Newman[28] propuso unos años más tarde que un alto nivel de diversidad en la gente genera la anonimidad necesaria para el crimen. Por tanto, según Newman, una diversidad poblacional baja, un número reducido de visitantes y un alto porcentaje de residentes son algunas de las características necesarias para conseguir que una zona sea segura. Varios estudios han intentado encontrar evidencia que ratifique alguna de estas dos teorías contrapuestas. Felson y Clarke[29] propusieron a finales de los 90 la Teoría de la Actividad Rutinaria que investiga cómo distintas situaciones y variaciones en el estilo de vida afectan las oportunidades para el crimen. Encontraron que lugares como bares y pubs suelen atraer el crimen.

En los últimos años, los criminólogos han comenzado a investigar en detalle concentraciones significativas de crimen en ciertas áreas geográficas, dando lugar al concepto de *puntos calientes* (*hotspots)* de crimen. En el año 2008, el criminólogo David Weisburd[30] propuso cambiar el paradigma de crimen centrado en las personas a un paradigma del crimen centrado en el espacio.

23. http://pricetheory.uchicago.edu/levitt/Papers/CullenLevittCrimeUrban1999.pdf
24. https://arxiv.org/abs/1404.1295
25. http://link.springer.com/chapter/10.1007/978-3-319-13734-6_29#page-1
26. http://web.mit.edu/rudin/www/docs/WangRuWaSeECML13.pdf
27. https://en.wikipedia.org/wiki/The_Death_and_Life_of_Great_American_Cities
28. https://en.wikipedia.org/wiki/Defensible_space_theory
29. http://www.popcenter.org/library/reading/pdfs/thief.pdf
30. https://www.iadlest.org/Portals/0/Files/Documents/DDACTS/Docs/Place-Based%20Policing.pdf

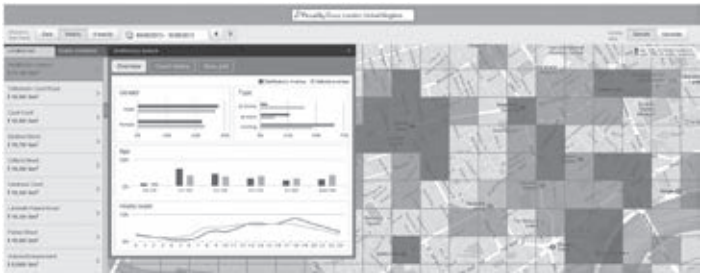
A partir de este trabajo previo, nosotros enmarcamos el problema de la predicción del crimen a partir de *big data* desde un punto de vista centrado en el espacio. En particular, investigamos el valor predictivo de la dinámica de las personas —dinámica inferida según los datos de la red móvil de telefonía y de información demográfica— para predecir si una determinada zona geográfica tiene probabilidad de concentrar crímenes o no[31].

En el proyecto presentado en esta sección, utilizamos conjuntos de datos que fueron compartidos durante una competición pública —un "Datathon para el Bien Social"[32]— organizado por Telefónica Digital, el Open Data Institute y el Media Lab del Massachusetts Institute of Technology (MIT) en septiembre de 2013. Entre otros, se compartieron con los participantes de este *datathon* los siguientes conjuntos de datos:

1. *Datos agregados y anomizados sobre la demografía y actividad humanas*. Nos referiremos a estos datos como *smartsteps* ya que los datos sobre la actividad humana proceden de un producto de Telefónica llamado *smartsteps*[33], ilustrado en la figura 2.
   La granularidad espacial para estos datos son celdas cuadradas que corresponden a una división del espacio en un *grid* con cuadrados de lado inversamente proporcional a la densidad de torres celulares en una determinada localización geográfica, es decir, conforme hay más torres celulares, menor es el lado del cuadrado. Para cada celda, se proporcionaron una serie de variables calculadas cada hora durante 3 semanas del 9 al 15 de diciembre de 2012 y del 23 de diciembre al 5 de enero de 2013. En particular, las variables incluyen: (1) una estimación del número total de personas en la celda calculada a partir de la actividad en la red móvil y extrapolando para la población en general teniendo en cuenta la cuota de mercado; (2) una estimación del género, edad y porcentaje de personas para las que dicha celda en ese momento es su hogar, su lugar de trabajo o son visitantes de la misma. El género y la edad proceden de estimaciones de GFK, una firma de análisis de mercado. La edad está representada en divisiones de 0-20 años, de 21-30 años, de 31-40 años, etc.

31. Una descripción completa del proyecto (en inglés) puede encontrarse en http://online.liebertpub.com/doi/abs/10.1089/big.2014.0054?journalCode=big
32. http://dynamicinsights.telefonica.com/2013/08/29/using-big-data-for-social-good/
33. http://dynamicinsights.telefonica.com/smart-steps/

Figura 2. Ejemplo de visualización de la información proporcionada por el producto *smartsteps*



Combinando datos agregados y anonimizados de movilidad junto con datos demográficos, podemos obtener información espacio-temporal sobre la dinámica de la ciudad, de gran utilidad para el escenario de interés: la predicción del crimen.

2. *Datos abiertos geolocalizados*, incluyendo datos de criminalidad, del censo, de climatología y de transporte. En este proyecto utilizamos los datos de criminalidad y los del censo.

    a) *Datos de criminalidad:* incluyen la geolocalización de todos los crímenes reportados en el Reino Unido con una granularidad temporal de mes y año. Los datos proporcionados para la competición correspondían a diciembre de 2012 y enero de 2013. Utilizamos los datos de diciembre para entrenar los modelos y los de enero para evaluarlos. La granularidad espacial es lo que se conoce como Lower Layer Super Output Area (LLSOA) que es la unidad censal utilizada en el Reino Unido. Los crímenes están clasificados en 11 tipos (por ejemplo, comportamiento antisocial, robo, crimen violento...).
    b) *Datos del censo:* los datos oficiales del censo de Londres contienen 68 variables que caracterizan la población en cada LLSOA, incluyendo número de hogares, proporción de la población que trabaja, proporción de la población mayor de 65 años, proporción de inmigrantes, etnicidad, lengua materna, salario medio, precio medio de la vivienda, número de coches por hogar, porcentaje de niños en familias sin trabajo, niveles de felicidad, esperanza de vida, etc.

Modelo predictivo del crimen

Formulamos el problema de la predicción del crimen como un problema de clasificación binario aplicado a cada celda del espacio. Dado que el *dataset* de crímenes

no es balanceado, utilizamos la mediana para determinar si una celda es un *hotspot* de crimen: le asignamos un 0 si el número de crímenes en dicha celda en un mes es igual o menor a la mediana (=5 en nuestro caso) y le asignamos un 1 si el número de crímenes es superior a 5 y, por tanto, se considera un *hotspot*.

Utilizamos el 80% de las celdas para entrenar el modelo y el 20% de las celdas para validar el modelo.

Un paso previo fundamental requiere georeferenciar todos los datos en un mismo sistema de referencia geográfico. Las celdas de *smartsteps* son cuadradas mientras de las LLSOA tienen formas arbitrarias. Para unificar los sistemas de representación, asignamos los datos de criminalidad y del censo a la celda de *smartsteps* con el centroide más cercano.

Estudios anteriores han encontrado que variables que captan la diversidad y la regularidad del comportamiento humano son importantes para la caracterización del mismo. En particular, el concepto de *entropía* ha sido utilizado para estimar las características socioeconómicas de lugares y ciudades[34], la movilidad y los patrones de consumo. Por ello, para abordar el problema de la predicción del crimen, calculamos la media, la mediana, la desviación estándar y los valores mínimos y máximos de la entropía de Shannon de los datos de *smartsteps*. Además, para tener en cuenta relaciones temporales, calculamos las variables estadísticas anteriores (*i.e.* media, mediana, etc.) sobre ventanas temporales de longitud variable (1 hora, 4 horas y 1 día).

En el caso del censo, usamos las 68 variables originales sin transformaciones.

Dado que disponíamos de 68 variables del censo, tras un proceso de selección de características utilizando el coeficiente de Gini, redujimos el número de variables extraídas a partir del *dataset* de *smartsteps* a 68 también. De esta manera, podemos comparar directamente el comportamiento de dos modelos, uno construido a partir de las variables de *smartsteps* (modelo basado en la dinámica de la población) y otro a partir de las variables del censo (modelo basado en variables tradicionales).

Con relación a la clasificación, utilizamos Random Forests[35] ya que este tipo de modelos satisfacen la propiedad de máximo margen, no requieren el tuneado de parámetros y sobre todo no requieren una especificación del espacio de características requerido por otros modelos comunes incluyendo Support Vector Machines.

34. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.468.5423&rep=rep1&type=pdf
35. http://link.springer.com/article/10.1023/A:1018054314350

Evaluamos tres modelos, además de un modelo de base basado en un clasificador de "mayoría" que retorna la clase mayoritaria. Los tres modelos comparados son: (1) un modelo que utiliza las variables del censo; (2) un modelo que utiliza las variables de *smartsteps*; y (3) un modelo conjunto, que combina variables del censo y de *smartsteps*.

Para todos los modelos, calculamos la precisión y el score F1 que es la media armónica entre la precisión y el *recall*, así como el área bajo la curva ROC (AUC). La tabla 1 resume los resultados de evaluación de los modelos y la figura 3 ilustra la clasificación de las distintas celdas de la zona metropolitana de Londres en puntos calientes de crímenes o no.

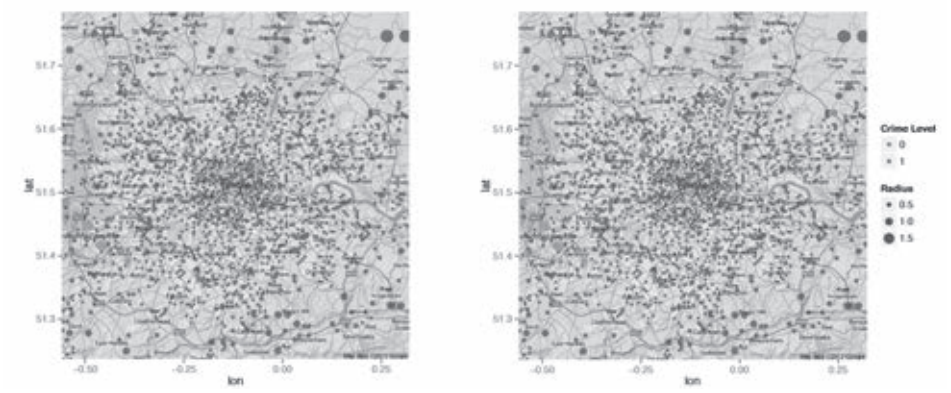
Tabla 1. Resultados de evaluación de cuatro modelos en la predicción de puntos calientes de crimenes

Modelo	Precisión	Precisión con intervalo de confianza, 95%	Score F1	AUC
Base	53.15	(0.53,0.53)	0	0.50
Smartsteps	68.04	(0.66,0.69)	67.66	0.63
Censo	62.18	(0.60,0.64)	61.72	0.58
Smartsteps+censo	68.83	(0.67,0.70)	68.52	0.63

Como puede observarse en la tabla, los modelos que incluyen variables sobre la dinámica de la población (*smartsteps* y *smartsteps*+censo) obtienen mejores resultados de clasificación que el modelo que utiliza las variables tradicionales de censo, así como el modelo base de comparación.

Mirando en detalle el rendimiento de los modelos para la clase correspondiente a un número "alto de crímenes" (valor 1 en el modelo), si nos fijamos en el índice de positivos verdaderos (*true positive rate*), es decir, la proporción de celdas que alto nivel de crímenes que han sido correctamente identificadas, y en el índice de negativos verdaderos, es decir, la proporción de celdas con un número reducido de crímenes que son correctamente detectadas. Los modelos usando datos de *smartsteps* y *smartsteps*+censo obtienen unos buenos resultados de índice de positivos verdaderos con un 74,2% y 73,9%, respectivamente. El modelo que utiliza los datos del censo solo alcanza un índice de positivos verdaderos del 68,81%, más de un 5% por debajo que los modelos que utilizan los datos de *smartsteps*.

Figura 3. Puntos calientes de crimen (izquierda) y predicciones hechas por el modelo (derecha)



Respecto al índice de negativos verdaderos, todos los modelos obtienen peores resultados: 63,07% el modelo de *smartsteps*+censo; 61,06% el modelo solo de *smartsteps* y 54,66% el modelo solo del censo. Cabe destacar que nuestro modelo (*smartsteps*) obtiene un índice de positivos verdaderos (*true positive rate*) aproximadamente un 10% mejor que el índice de negativos verdaderos (*true negative rate*). Es decir, los modelos detectan mejor las celdas con alto nivel de crímenes que las que tienen bajo nivel de crímenes. En nuestro caso de uso es más importante obtener buenos resultados en la clase de "alto nivel de crímenes", ya que desde un punto de vista social, es menos peligroso asignar a una celda como de alto nivel de crímenes erróneamente que al contrario. Tal y como muestran los resultados, el modelo propuesto aporta ventajas significativas en la tarea de detección de *hotspots* o puntos calientes de crimen.

Implicaciones

Los resultados presentados en la sección anterior muestran que los datos sobre el comportamiento humano agregado así como datos demográficos básicos con una escala temporal diaria y mensual aportan valor significativo para la predicción de lugares con alta concentración de crímenes, con unos resultados de clasificación superiores a los modelos que solo utilizan datos del censo (demografía, etnicidad, datos de empleo, etc.). Los datos del censo suelen ser muy detallados, pero son muy costosos de obtener y tienen una granularidad temporal elevada (normalmente el censo se calcula cada 10-12 años). Dada la baja frecuencia de obtención de este tipo de datos y dado el carácter dinámico de las ciudades, el valor de los mismos para la predicción de crimen es limitado en

nuestra experiencia. Los datos dinámicos extraídos a partir de la red de telefonía móvil, combinados con datos demográficos también dinámicos, aunque son mucho menos detallados que los datos del censo, proporcionan una mejor resolución temporal y espacial que parecen ser importantes para el problema de estudio.

Si estudiamos las 20 variables con mayor valor predictivo en nuestro modelo, observamos:

- En el modelo conjunto (*smartsteps*+censo), no aparece ninguna variable del censo entre las 20 variables con mayor valor predictivo.
- Las variables calculadas con una ventana temporal diaria tienen mayor valor predictivo que las variables calculadas con una ventana temporal mensual.
- Variables relativas al porcentaje de personas para las que una determinada celda es su hogar tanto a nivel diario como mensual parecen tener mucha importancia. De hecho, 11 de las 20 variables más predictivas están relacionadas con el concepto de *hogar*. La teoría de Newman del "espacio defendible" postula la relevancia de las variables relativas al número de residentes en un lugar. Sin embargo, a diferencia de la teoría de Newman, encontramos correlaciones positivas entre las variables de hogar y el crimen. Por tanto, nuestros experimentos no corroboran la teoría de Newman. Un trabajo reciente realizado por Traunmueller[36] ha encontrado resultados similares.
- La entropía de las variables también juega un papel importante ya que 8 de las 20 variables más predictivas incluyen la entropía. La entropía captura la estructura predecible de un lugar con relación a los tipos de personas que hay en ese lugar a lo largo del día. Un lugar con alta entropía tendría mucha variedad con relación a los tipos de personas que lo visitan diariamente; mientras que un lugar con baja entropía se caracterizaría por una gran regularidad en los patrones del tipo de personas que lo visitan. Tanto nuestros resultados como los de Traunmueller *et al.* corroboran la teoría de Jacobs de la vigilancia natural, según la cual una alta diversidad de personas (tanto a nivel de edad como de género) actúan como "ojos en la calle" que monitorizan el entorno y reducen la criminalidad.

La metodología propuesta en esta sección podría tener implicaciones prácticas informando a los departamentos de policía y los gobiernos de las ciudades sobre cómo y dónde invertir esfuerzos y cómo reaccionar a eventos criminales.

36. http://link.springer.com/chapter/10.1007/978-3-319-13734-6_29#page-1

Conclusión

El potencial de estos datos para mejorar el mundo es inmenso. Aunque estamos en los comienzos de esta revolución, es una revolución sin duda necesaria y de gran importancia. Hay una serie de retos de diversa índole que será preciso superar para poder realizar este potencial, incluyendo retos:

- *Regulatorios* ya que la regulación no está actualizada para contemplar casos de uso como los descritos en este capítulo con datos que han sido capturados para otros fines. Además, no hay principios claros definidos con relación al análisis y compartición de datos agregados y anonimizados de manera segura y con fines humanitarios. Cabe destacar en este sentido esfuerzos como OPAL[37], una plataforma de algoritmos abiertos que pueden ejecutarse de manera segura en las premisas de los dueños de los datos para la extracción de indicadores clave en el contexto del Bien Social.
- *Sociales.* Es importante tener en cuenta potenciales consecuencias no esperadas que puedan resultar del análisis de los datos, aun cuando el objetivo del análisis es tener impacto social positivo. Asimismo es fundamental ser conscientes del riesgo de crear una brecha entre quienes tienen acceso a los datos y tienen el conocimiento necesario para poder analizarlos y extraer información valiosa de los mismos, y los que no lo tienen. Por ello, para poder realizar el potencial del Big Data para el bien social debemos invertir en programas de alfabetización en ciencias de datos (data literacy);
- *Técnicos*, de diversa índole. Al ser un área de investigación emergente, los retos técnicos son numerosos. Es necesario entender como de representativos son los datos, que potenciales sesgos tienen y hasta que punto podemos generalizar las conclusiones alcanzadas a través del análisis de los mismos. Para abordar problemas reales, es crítico combinar datos procedentes de distintas fuentes. Conseguir realizar los análisis en tiempo real es primordial es algunos casos de uso. Sin embargo, todavía estamos lejos de poder conseguirlo. En muchos ejemplos no hay ground truth de manera que es difícil validar y cuantificar el impacto de los resultados sin hacer intervenciones derivadas del análisis de los datos. La transparencia de los algoritmos que se usan para tomar decisiones basadas en datos es otra característica primordial. Finalmente, gran parte de los estudios hasta ahora se centran en encontrar correlaciones significativas entre diversas variables de interés. Sin embargo, algunas oportunidades importantes surgen cuando se puede atribuir una relación de causalidad entre las variables, algo que hasta ahora ha sido difícil de demostrar.

37. http://www.data4sdgs.org/dc-opal/

- *De privacidad y seguridad*. Los riesgos de privacidad deberían ser minimizados y cuantificados[38]. La transparencia, el control de acceso, la minimización de posibles sesgos, la seguridad y la trazabilidad de los datos son fundamentales[39]. También es necesario definir y actuar de acuerdo con un código de conducta y unos principios éticos con relación al manejo y análisis de los datos.
- *De modo de trabajo*. Los casos de uso para el Bien Social son por definición multi-disciplinares y multi-institucionales, ya que requieren el trabajo conjunto de expertos en diferentes áreas (e.g. análisis de datos, salud pública, gestión de crisis, etc...). Para conseguir que los proyectos se conviertan en una realidad, es importante hacer una buena gestión de equipos multi-disciplinares en instituciones diversas, a veces con agendas no totalmente alineadas. Además, estos proyectos suelen requerir la firma de acuerdos de colaboración entre diversas instituciones lo que puede retrasar la ejecución de los proyectos.

Afortunadamente, ninguna de estas barreras es infranqueable. La oportunidad del uso de Big Data para el Bien Social es demasiado grande y necesaria para no unir esfuerzos e intentar superar los retos. Por ello, cuando pensamos en Big Data, les animo a pensar no solo en una oportunidad de negocio, sino también en su valor para mejorar el mundo. Es sin duda lo que me motiva a investigar y trabajar en este apasionante campo

Coautores

Los proyectos descritos en este capítulo son el fruto del trabajo de colaboración con las siguientes personas: Dr. Vanessa Frías-Martínez, Dr. Enrique Frias-Martínez, Dr. Alex Pentland, Dr. Bruno Lepri, Dr. Jacopo Staiano, Dr. Emmanuel Letouze, Dr. Andrey Bogomolov, Dr. Miguel Luengo-Oroz, Dr. Alex Rutherford, Jong Gun Lee, Liudmyla Romanoff, Carlos Castro-Correa, Amit Wadhwa, Jean-Martin Bauer, Dr. Yolanda Torres-Fernández, Dr. Alfredo Morales-Guzmán, Dr. David Pastor-Escuredo y Dr. Fabio Pianesi.

38. http://openscholar.mit.edu/sites/default/files/bigdataworkshops/files/draft_modelsformobilephonedatasharing.pdf
39. "The Tyranny of Data?: The Bright and Dark Sides of Data-driven Decision-making for Social Good" in "Transparent data mining for Big and Small data" Springer, 2016

Capítulo 14

Big data en la Administración pública chilena: oportunidades para la gestión de políticas públicas

Pedro Huichalaf*

› Introducción

La cantidad de información que actualmente generan las organizaciones ha aumentado exponencialmente. Tan solo en los últimos dos años se ha creado el 90% de toda la información en el mundo y el 80% de la misma es información no estructurada, según datos de IDC-IBM. Y esta tendencia sigue firme, ante el surgimiento de tecnologías y herramientas que permiten almacenar y procesar terabytes de información a costos cada vez menores y velocidades cada vez mayores. El sector público no está ajeno a esta realidad. Analizaremos los casos del Ministerio de Educación, Servicio de Impuestos Internos y del Ministerio de Transportes y Telecomunicaciones del Chile, que han utilizado técnicas para abordar problemas de gestión con el uso de estas tecnologías.

Sabemos que los grandes volúmenes de información o *big data* provenientes de redes sociales, dispositivos móviles, sensores de máquinas o incluso tecnologías tradicionales como el correo electrónico o los sistemas transaccionales pueden ser procesados de forma muy eficiente con nuevas tecnologías, a través de técnicas de cómputo distribuido. Esta tendencia ha llevado a las organizaciones públicas y privadas a reconocer el valor de la información como insumo para la optimización de procesos y decisiones del negocio.

Esta posibilidad hace que surjan oportunidades, principalmente asociadas al análisis de dicha información, que antes eran impensables y que pueden brindar a las organizaciones en los próximos años una ventaja competitiva muy grande, debido a su carácter innovador.

Para definir *big data*, haciendo un esfuerzo importante de simplificación, podría resumirse al menos en tres variables básicas: volumen, variedad y viscosidad, o calidad de los datos; a partir de ello surgen una serie de interpretaciones que

* Abogado, magíster (c) Derecho Informático y Telecomunicaciones y exviceministro de Telecomunicaciones de Chile.

incluso se cruzan con la pertinencia para que un dato sea oportuno en la toma de decisión, en el lugar y momento indicados (tiempo real o casi tiempo real).

El *big data* es hoy por hoy una de las mayores tendencias a nivel mundial en el desarrollo tecnológico, lo que permite el procesamiento de grandes volúmenes de datos, de forma distribuida, y que se haya aumentado el procesamiento de la información en un menor tiempo de uso de los procesadores.

Ahora podemos procesar lo que antes tardábamos horas o días en solo un par de segundos. Es, sin duda, un gran avance tecnológico, que las Administraciones tanto públicas y privadas debemos aprovechar.

Sin embargo, el desafío es cómo tratar la información estructurada y no estructurada; sabemos que la primera se puede sistematizar y procesar de mejor manera, es el caso de los documentos tipo Excel, cvs y bases de datos; sin embargo, la segunda, que son esencialmente vídeos, imágenes y sonidos, presentan en la gran mayoría un desafío no menor para la gestión de los datos y la construcción de la información de valor.

La sociedad del Gigabit se caracterizará porque el "nuevo oro" serán precisamente los datos, pero estos deben ser oportunos, administrables y tratables, sin lo anterior no podremos hacer nada con cantidades inconmensurable de datos.

Visión estratégica

La economía digital es y será la economía de los datos, por lo tanto, los países deberemos asegurar que la mayor cantidad de datos estén a disposición de **toda la población**, con eso desarrollaremos un nuevo modelo económico basado en la **colaboración, participación y la inclusión social y democrática**, que maximizará los beneficios de la sociedad del conocimiento para todos y todas.

En esto último, los gobiernos como en otros pocos terrenos avanzamos a mayor velocidad que el sector privado, ofreciendo gran cantidad de datos en formato *open data*, debido a que somos grandes actores tanto en el consumo como en la creación de *data*, sobre todo por los esfuerzos transversales de poner a disposición información de forma transparente a los ciudadanos. Teniendo en cuenta que analizamos información y no los datos personales de las personas, con el fin de determinar o modificar algunas políticas públicas necesarias para el

bien común. Sin embargo, se hace cada vez más necesario que el sector privado avance en esta misma dirección, dando señales de transparencia y compromiso con el desarrollo de las economías digitales.

Los gobiernos hemos avanzado significativamente en ofertar información para la ciudadanía, la política de datos abiertos es una ventana al conocimiento y a la generación de valor en nuestras economías y sociedad. Cabe destacar el esfuerzo del Gobierno de Chile en el portal de gobierno, datos.gob.cl, donde se ofrece información de los distintos ministerios para que la ciudadanía pueda utilizar dicha data.

Un gobierno abierto fomenta la transparencia y la participación ciudadana, transformando a los ciudadanos en actores claves para el desarrollo de una mejor sociedad; no obstante, todavía faltan esfuerzos en la tarea de poner a disposición de los ciudadanos una mayor cantidad de datos abiertos, con criterios que resguarden la privacidad de la información individual de las personas y permitan ver tendencias.

Los países deberemos al menos consensuar estándares en interoperabilidad y datos abiertos, lo que será la clave del éxito para alcanzar economías de escala que propicien el desarrollo de "Internet de las cosas", inteligencia artificial y *big data*, disponilizando y/o reutilizando los datos para que fomenten la innovación.

Los gobiernos por ejemplo, necesitaremos información de salud de los ciudadanos para políticas públicas de salud que se obtendrán de sensores en las personas. En este mismo sentido, una de las medidas de la agenda digital de Chile 2020 incluye iniciativas como la ficha médica electrónica, que contará con altos estándares en materia de protección de datos personales.

El tema de la gobernanza de los datos en el contexto del *open data* también representa un importante desafío; para ello, el Gobierno de Chile ha recibido de muy buena forma la recomendación de la OECD, en el documento titulado "Fortalecimiento del marco institucional y gobernabilidad para el gobierno digital", lo que es una guía práctica para desarrollar una institucionalidad a nivel de agencia o subsecretaría, donde una de sus funciones sea la gobernanza en materias de *open data*.

De tal manera, la regulación, bajo un ambiente de derechos de propiedad intelectual y respeto a la privacidad de las personas, debe fomentar la interoperabilidad entre las plataformas y los servicios de manera que se asegure el libre flujo de la información.

Ahora *big data* es una muy buena herramienta para conocer las necesidades reales de los ciudadanos, y con ello tratar de maximizar la función-objetivo de cualquier gobierno, que debe ser aumentar el bienestar social, y para ello es fundamental aplicar estas técnicas en beneficio directo de los ciudadanos y así aumentar el estándar de vida de las personas a través del desarrollo de políticas públicas adecuadas.

A pesar de que en ninguno de los cinco ejercicios de agenda digital que ha realizado el país aparece este concepto tácitamente, debemos entender que tangencialmente, en aspectos como la mejor gestión de los servicios del Estado y las políticas de datos abiertos, tienen implícitas estas técnicas para el mejor uso y tratamiento de la información pública.

› Casos prácticos en el Gobierno de Chile

Ministerio de Educación del Gobierno de Chile: ¿cómo contribuyen los datos masivos a la toma de decisiones en educación?

Chile está impulsando una reforma educativa, cuyos principales objetivos son: priorizar y fortalecer la educación pública, transfiriendo la administración desde los municipios a los Servicios Locales de Educación (SLE) Públicos; regular la provisión de la oferta de los establecimientos particulares subvencionados: eliminación de selección, lucro y copago; y fortalecer la profesión docente.

De tal manera, para apoyar el eje de la creación de los Servicios Locales de Educación Pública, que sustituirán a los municipios en la administración, se desarrolló el estudio "Apoyando la formulación de políticas públicas y toma de decisiones en educación utilizando técnicas de análisis de datos masivos: el caso de Chile", que es un proyecto conjunto entre la Universidad de Chile y la Universidad Adolfo Ibáñez, y que está financiado por el fondo de fomento de proyectos científicos y tecnológicos Fondef de Conicyt, dependiente del Ministerio de Educación de chile.

Ahora los nuevos servicios locales de educación van a agrupar los territorios de una forma más amplia que los actuales municipios, sin embargo, estos nuevos sistemas locales de educación afrontan el desafío de un país diverso, que tiene muchas diferencias geográficas y sociodemográficas, diferencias en la accesibilidad urbana, existe concentración en los centros de empleo y segregación económica; aunque el sistema educativo no opera en el vacío sino sobre un contexto

de estas relaciones que funcionan en el territorio. Sin embargo esta relación muchas veces no es evidente y normalmente se ignora en la toma de decisiones y en la elaboración de políticas públicas.

La pregunta que surge entonces es cómo podemos visibilizar estas relaciones para integrarlas en la mesa de toma de decisiones. Para ello, este proyecto plantea la generación de modelos explicativos y predictivos que permitan estudiar y entender ciertos fenómenos que ocurren en el ámbito de la educación, lo primero que se requiere es formular las preguntas correctas a partir del estudio de ciertos fenómenos.

Pero, la diversidad geográfica y sociodemográfica de Chile conlleva desafíos invisibles en diseño e implementación de políticas públicas en educación, que deben tomarse en cuenta, ya que el desempeño del sistema escolar también depende de otros factores no educativos.

Usando técnicas de ciencia de datos, empleando como insumos en su mayoría datos abiertos provistos por el Gobierno, se generó evidencia para medir aspectos del sistema educativo previamente invisibles como la falta de equidad espacial en el acceso a oferta educativa y que además cumpla estándares mínimos de aprendizaje, o la posibilidad de predecir y prevenir fenómenos de alto impacto social, como son la deserción y el abandono escolar. Así, es posible evidenciar la necesaria coordinación y diálogo interagencias en el diseño de políticas públicas y toma de decisiones. El uso de indicadores espaciales permite visualizar la inequidad de los territorios, dando origen a una nueva generación de políticas públicas que supere la lógica y limitaciones de las divisiones político-administrativas de los territorios.

Cuando hablamos de la mejora de la educación, y especialmente de las brechas entre el sector público y el privado, pensamos en los estudiantes, los docentes, los directores, y hasta quizá en la infraestructura de las escuelas: pupitres, salones, materiales, computadores, accesibilidad a Internet, entre otros.

Asumimos que las diferencias existen solo por lo que ocurre dentro de la escuela. Es decir, que los estudiantes son "iguales" y las escuelas producen cambios.

Sin embargo, la evidencia muestra que las supuestas diferencias entre escuelas se explican por el nivel socioeconómico y sociocultural de las familias de los estudiantes. Así, las escuelas desarrollaron mecanismos para "capturar" estudiantes de un cierto nivel. Es decir, las diferencias se producen por lo que ocurre

fuera del colegio. Entonces, terminamos con un sistema donde las escuelas eligen a los estudiantes compitiendo por los de mejor nivel socioeconómico bajo métricas de "calidad" reducidas al resultado de pruebas estandarizadas.

Así para mejorar un sistema educativo es necesario también entender el contexto en el cual opera, como la dimensión territorial que revela la heterogeneidad del país, evidenciando relaciones con factores no educativos previamente ocultos para las políticas públicas.

Por ejemplo, si analizamos el desplazamiento diario de estudiantes en Santiago, encontramos que un 30% asiste a una escuela fuera de su comuna. Solo entre las comunas de Puente Alto y La Florida más de 11.000 estudiantes viajan diariamente. Esto produce congestión vehicular, contaminación ambiental y presiona los requerimientos del transporte público, ya que en la primaria los niños viajan acompañados.

Gracias a técnicas analíticas de modelación encontramos que el 32% de los estudiantes no tiene un colegio a menos de 10 minutos caminando, y que la distribución espacial de los colegios se concentra en sectores de mejores ingresos con escasa cobertura en la periferia, que es donde viven más estudiantes con peor situación socioeconómica. Los desplazamientos se producen desde las afueras hacia el centro de la ciudad, impulsados por la mejor cobertura existente en áreas centrales. Este fenómeno se repite en el acceso a escuelas con buenos resultados en las pruebas estandarizadas, donde existen territorios en los cuales los estudiantes están "atrapados" por su falta de recursos para movilizarse, lo que empuja a un círculo vicioso de fracaso escolar.

Estos análisis permiten visualizar la inequidad territorial, determinando dónde y en qué zonas existen las mayores brechas. Esto permite desarrollar políticas públicas específicas para los territorios dependiendo de su geografía y sociodemografía, haciendo explícita la necesaria coordinación y diálogo interagencias para resolver problemas que son multisectoriales.

De este modo, se puede avanzar hacia una inteligencia de valor público (equivalente social de la inteligencia de negocios) que genere evidencia para las decisiones gubernamentales usando los propios datos que el mismo Estado recopila y genera.

Esta iniciativa ocupó el primer lugar en el *call for papers* "Nuevos debates, datos para el desarrollo" del BID.

Transantiago como fuente de datos los datos pasivos puede ayudarnos a hacer una mejor gestión de la ciudad

En el Departamento de Ingeniería Civil (DIC) de la Universidad de Chile vieron esa oportunidad y en 2008 comenzamos a trabajar con nuestro sistema de transporte público, que si bien se pueden obtener mediante mediciones, su costo es muy elevado, y su nivel de cobertura espacio-temporal muchísimo menor, dado que existen recursos limitados para realizar este tipo de análisis. Por ejemplo, las matrices origen-destino de viajes tradicionalmente se obtienen de encuestas origen-destino que se realizan a una muestra de la población.

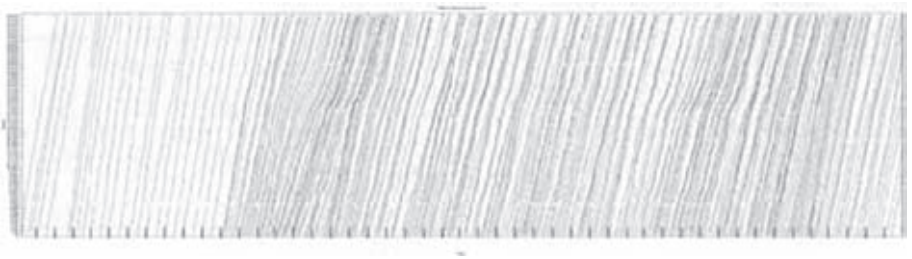
En Santiago se realiza una cada diez años, la última se realizó en 2012-2013, con una muestra de alrededor de 18.000 hogares y un costo aproximado de 700 millones de pesos, incluyendo la realización de las encuestas y mediciones complementarias. Los resultados aún no están disponibles, pues con posterioridad al proceso de recolección de datos se requiere un detallado posproceso para filtrar errores y realizar la expansión a la población. Con los datos de *transacciones bip*! se obtiene información detallada del 80% de los viajes en transporte público, que corresponden a más cuatro millones de viajes diarios de más de dos millones de usuarios (*tarjetas bip*!), logrando una cobertura espacio-temporal que es imposible alcanzar con datos de encuestas.

Mediante posprocesamiento de los datos de transacciones se distingue los transbordos de los destinos de viaje donde los usuarios realizan actividades (Devillaine *et al.,* 2012], y para los usuarios frecuentes (aquellos que viajan al menos cuatro veces a la semana) se identifica la zona de residencia, observando la posición de la primera transacción del día en todos

La llegada de Transantiago como sistema integrado de transporte público de Santiago, Chile, a partir de febrero de 2007, fue polémica, algo traumática..., en fin, podríamos decir mucho, y ya se ha dicho mucho sobre el tema, pero hay algo sobre lo que no se ha hablado tanto, que es el beneficio colateral de la generación constante de enormes cantidades de datos, los datos en un proyecto PBCT (Programa Bicentenario de Ciencia y Tecnología). El primer desafío fue limpiar y procesar datos de posicionamiento de los más de 6.000 buses del sistema, que cuentan con dispositivos GPS que emiten una señal de posición cada 30 segundos, generando del orden de 80 millones de registros a la semana.

Estos permiten observar el movimiento de los buses con un nivel de cobertura y precisión que nunca antes había estado disponible, y generar perfiles de velocidad (figura 1) [Cortés *et al.,* 2011).

Figura 1. Trayectoria espacio-tiempo de los buses de un servicio



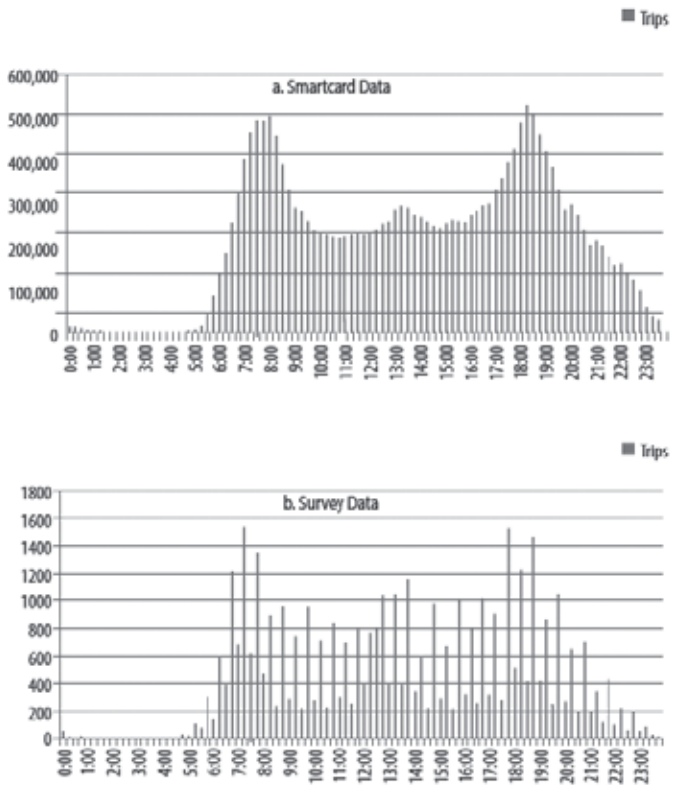
Por otra parte, están los datos de *transacciones bip!,* del orden de 35 millones a la semana, que de por sí contienen información valiosa, al mostrar la distribución temporal de la demanda, pero que además al cruzarlos con la base de datos de posicionamiento hace posible asignar posición a los registros de subida (figura 2). Esto hace posible obtener la distribución espaciotemporal de la demanda. Mediante una metodología desarrollada en el DIC, que se basa en observar la secuencia de transacciones de una misma tarjeta [Munizaga y Palma, 2012], se estima el paradero o estación de bajada como aquel más conveniente para acceder a la posición de la siguiente subida, dentro de un radio de 500 m. Esto se logra para sobre el 80% de las transacciones, abriendo la puerta a una variada gama de análisis, que incluyen construir matrices origen-destino de viajes en transporte público, construir perfiles de carga de buses y metro (Gschwender *et al.,* 2012), construir indicadores de calidad de servicio, etc. Todas estas son valiosas herramientas para quienes están encargados de realizar la planificación y gestión los días en que la tarjeta es observada.

Si estas tienen coincidencia espacial, se estima que esa zona corresponde al lugar de residencia del usuario (tarjeta). Asimismo, hay otro tipo de análisis que sería interesante realizar, como, por ejemplo, observar los patrones de viaje y completar información faltante.

Una segunda etapa en el desarrollo de este proyecto es la validación, porque si bien se logra realizar estimaciones de paradero de bajada para sobre el 80% de las transacciones, identificar destinos de actividades para esos viajes, e incluso estimar zona de residencia para los usuarios frecuentes del sistema, se requiere información exógena para comprobar que esas estimaciones sean correctas. Para realizar la validación hasta ahora se ha contado con una pequeña muestra de validación que

proviene de la encuesta origen-destino de viajes en metro, realizada en 2010, en que a una muestra de usuarios se les registra el número identificador de la *tarjeta bip!,* además de aplicárseles la encuesta origen-destino. Los resultados son positivos, pero insuficientes debido al reducido tamaño muestral (882 encuestas). Sin embargo, próximamente contaremos con los resultados de la EOD 2012-2013, que es una muestra representativa de la población de Santiago, que podrá ser usada para validación, dado que en esta también se incluyó una pregunta que pide registrar el número identificador de la o las tarjetas que el encuestado utiliza para viajar.

Figura 2. Comparación del nivel de precisión obtenido con datos de *transacciones bip!* y con datos de encuesta



Otra línea de desarrollo interesante a partir de este proyecto es la modelación. Lo que se obtiene de las transacciones y GPS de los buses es una observación de la realidad actual, algo así como una foto de alta definición, pero para poder aportar al proceso de toma de decisiones, requerimos más que una foto, necesitamos

elaborar modelos de comportamiento que nos permitan predecir qué va a suceder con el sistema si aplicamos cambios en él. Por ejemplo, ¿qué va pasar con las velocidades de los buses y los tiempos de viaje de los usuarios si construimos corredores segregados para los buses?

¿Qué efecto tendrá en la demanda mejorar la regularidad de los intervalos entre pasadas de buses de un mismo servicio? Dado que ya contamos con información de casi ocho años consecutivos, con amplia cobertura espacio-temporal, es posible observar esos cambios y deducir leyes de comportamiento.

Esta es una etapa fascinante en una disciplina que tradicionalmente se ha enfrentado a la escasez de datos. Hay factores relevantes para las decisiones de los usuarios como, por ejemplo, la variabilidad del tiempo de viaje o el tiempo de espera, que difícilmente han sido analizados con propiedad debido a la escasez de datos. Ahora cambiamos de paradigma, pasamos de la escasez de datos a la abundancia abrumadora de ellos, y el desafío es utilizarlos adecuadamente en beneficio de la sociedad.

El gran desafío es utilizar los datos que se generan automáticamente mediante la operación del sistema, para contribuir a mejorar la gestión de la ciudad, haciendo que todos sus sistemas funcionen de forma más eficiente, amable y sustentable.

La invitación está abierta a quienes quieran realizar investigación en esta área, porque aún hay mucho por hacer, y las más diversas disciplinas pueden contribuir a ello, incluyendo por cierto a la ciencia de la computación.

Ministerio de Hacienda, Servicio de Impuestos Internos: estrategia e integración de datos *big data*, instrumento de apoyo para optimizar la operación

El Servicio de Impuestos Internos es el organismo recaudador de los impuestos de Chile, este servicio depende del Ministerio de Hacienda y fue creado por el Decreto con Fuerza de Ley nº 7, de Hacienda, de 30 de septiembre de 1980, y la designación de su director es una facultad privativa del presidente de la república.

A nivel de desarrollo digital, podemos mencionar que el Servicio de Impuestos Internos representa un éxito en la implementación de sistemas digitales de nuestro país, junto con el sistema de compras públicas, y ha sido un caso referente a nivel internacional.

Algunos datos del servicio:

- 1.227 millones de visitas al año (sii.cl).
- 8,7 millones de usuarios con clave.
- 416 millones de DTE.
- 230.000 inicios de actividades (Internet), 81,9% del total.
- 84.000 términos de giro (Internet), 90,3% del total.
- 17,5 millones de boletas honorarios electrónicas.
- 3,36 millones de declaraciones en AT2016.
- 2,78 millones de contribuyentes con solicitud de devolución.
- 99,7% de las declaraciones de la declaración renta fueron presentadas por Internet.
- 2,38 millones de propuestas de F-22 76,2% de propuestas no modificadas.
- 986.000 formularios recibidos en promedio mensualmente, 82,8% por Internet.
- 175.000 cupones de pagos de IVA en promedio mensualmente, 14,8% del total de declaraciones F29.

El Servicio de Impuestos Internos describe su estrategia: considera el *big data* como un instrumento para optimizar la operación, pero son cuidadosos al hablar de *big data vs*. analítica de data u otras técnicas o conceptos, finalmente entienden estos como elementos que apoyan al negocio.

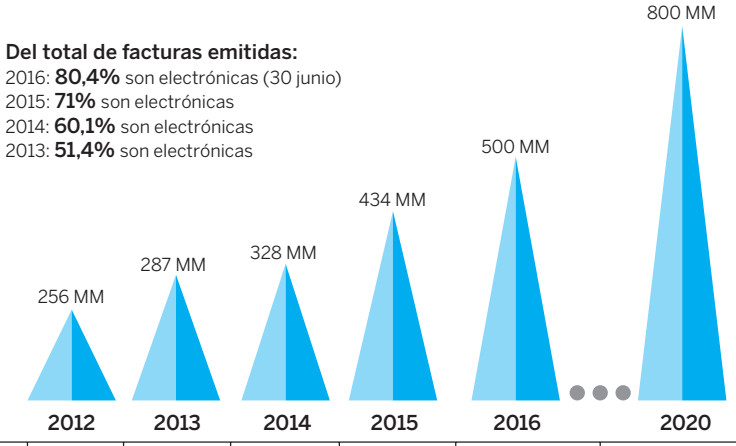
Uno de los principales desafíos a los que se ha enfrentado el Servicio de Impuestos Internos en los últimos años ha sido la implementación de la reforma tributaria, iniciativa llevada a cabo por la presente Administración para mejorar la calidad de la educación de nuestro país.

En ello se modificaron aspectos como:

- Postergación pago de IVA.
- Impuesto a las emisiones de vehículos motorizados.
- Repatriación de capitales.
- Nuevo Régimen Simplificado de Tributación 14 Ter.
- Consulta pública de normativa.
- Actualización de sistemas, gormularios e instrucciones.
- Notificaciones por correo.
- Término de giro por parte del SII.
- FUT histórico y retiros en exceso.
- Sustitución de multas por cursos de capacitación.
- Notificación por página web.
- Norma general antielusión.

- Restricción al uso del IVA por compras en supermercados.
- Modificaciones al crédito especial para empresas constructoras.
- Derogación del artículo 14 bis y 14 quáter de la LIR

También la normativa particular referente a la **masificación de la factura electrónica** ha sido un reto para el servicio, en el sentido que muestra el gráfico:ÓNICA, ha sido un reto para el servicio, en el sentido que como muestra el grafico:



DTE	222MM	247MM	283MM	367MM	215MM (*)	600MM
F22, F29, F45, F50, SIFM, DJ, BHE	34MM	40MM	45MM	67MM	35MM (*)	-

(*) Al 15 Julio

Experiencia

Fiscalización del IVA con tecnología *big data*

Mejorar los procesos de control del IVA, incluyendo nuevas fuentes de control de procesos masivos

Objetivo general

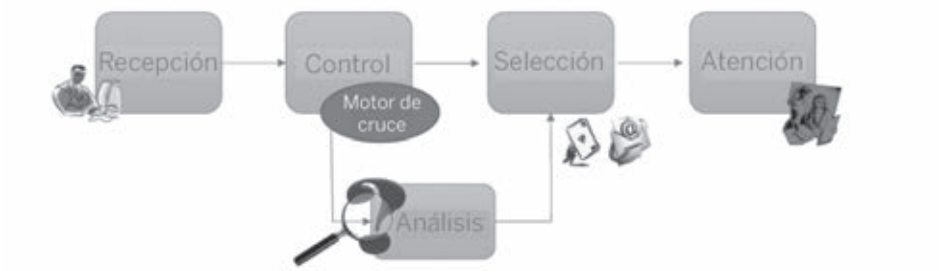
mejorar los procesos de control del IVA, incluyendo nuevas fuentes de control de procesos masivos.

Objetivos específicos

- Aportar información para el control del IVA.
- Incluir nuevos algoritmos de validación sobre grandes volúmenes de datos.

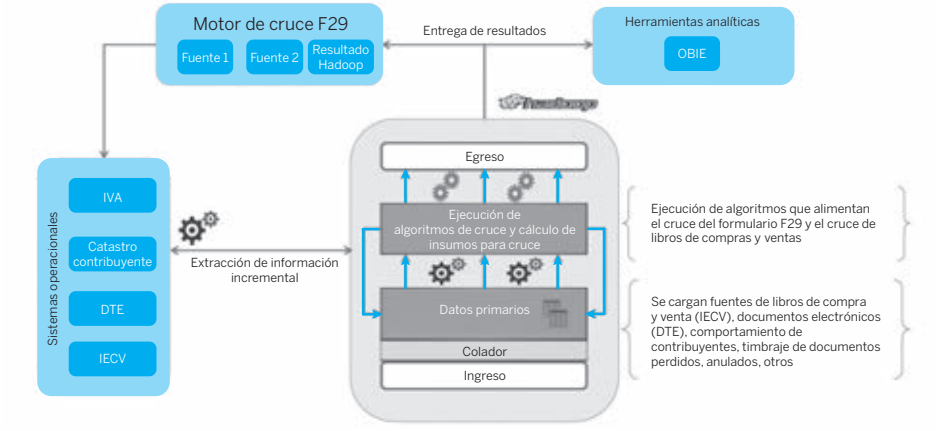
- Otorgar alternativas de reprocesos de información masiva en tiempos reducidos.
- Masificar el uso de las tecnologías *big data* para la mejora en la fiscalización.

Modelo de control masivo del IVA (F29)



- Contribuyente ingresa formulario 29.
- Se ejecuta proceso de control en línea (cruce de información con otras fuentes de datos).
- Se genera resultado de cruce de IVA.
- Se definen planes de atención (presencial y a distancia) para formularios con anomalías, en base a los resultados del cruce.
- Se realiza la atención.
- Contribuyente acude ante una notificación.
- Contribuyente rectifica por Internet su declaración de IVA o sus documentos electrónicos (libros de compras o ventas).

Modelo de procesamiento masivo de data utilizando Big Data Hadoop



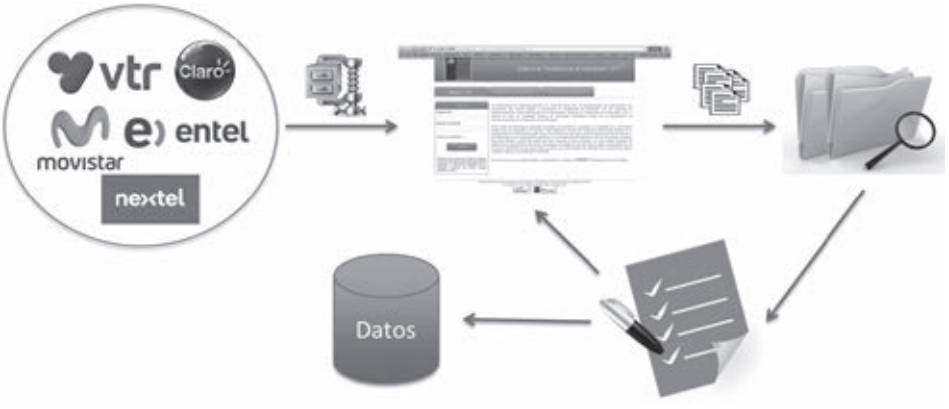
Subtel: los datos que genera del sector de telecomunicaciones

El Sistema de Transferencia de Informacion (STI) es un sistema que utiliza la Subsecretaria de Telecomunicaciones (Subtel) para recibir información por parte de las empresas de telecomunicaciones, donde entre otros temas se recibe información de fallas, cortes e incluso información comercial, que luego es procesada por Subtel para determinar cargos (multas) o información estadística a partir de esta data.

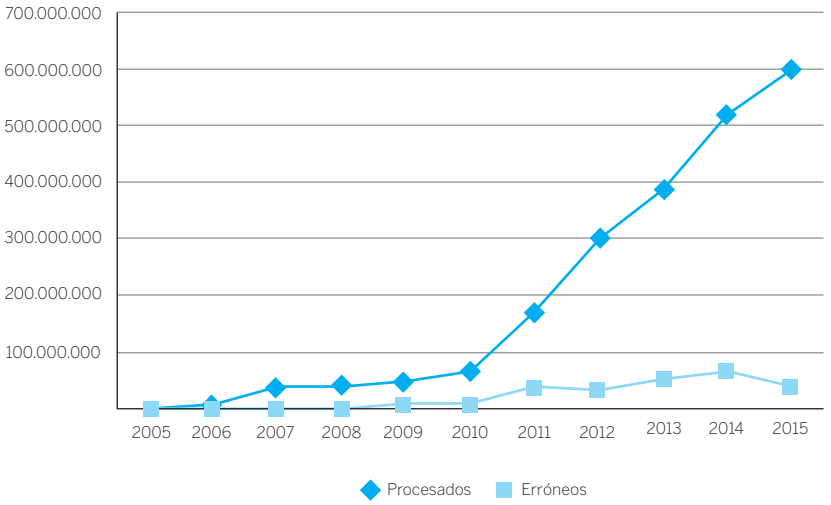
El STI fue donado por la Superintendencia de Electricidad y Combustibles (SEC) a Subtel el año 2004, donde después de un periodo de pruebas, en el que se realizaron las adecuaciones y personalizaciones necesarias, el STI es lanzado con un proceso; posteriormente ese mismo año se agregan otros cuatro, finalmente fue puesto en funcionamiento en el año 2005.

Posteriormente en el año 2007 se realiza la primera gran mejora al STI, la versión 2.5 de la plataforma que es que se usa actualmente.

El STI posee tres procesos principales: **recepción, validación y carga**. Los operadores ingresan la información al portal donde es cargado (copiado) y validado; si todo resulta bien se cargan los datos a la base de datos institucional. Adicionalmente de todo esto se informa a los operadores mediante el portal STI.



El sistema STI es un sistema implementado el año 2005 y, por lo tanto, los volúmenes de información que está diseñado para manejar son muy diferentes a los manejados actualmente.



Además el número de procesos ha cambiado en el tiempo.

Año	Número procesos
2005	5
2015	27

Tiempos de procesamiento promedio por proceso del actual STI..

Proceso	Registro promedio procesamiento	Tiempo	Procesamiento	Nº empresas
TM Calidad de red móvil	3,9 MM (2 archivos)	0,8 días	21 días	6
OT Interrupciones y descuentos	82 M (3 archivos)	1,5 días	22 días	32
OT Productos y tarifas V2.0	40 M (10 archivos)	20 segundos	3 minutos	56
OT Indicadores de calidad de atención telefónica	2 (1 archivo)	10 segundos	2 minutos	38
OT Red	20 (2 archivos)	10 segundos	2 minutos	32
OT Cargos por acceso	230 (2 archivos)	10 segundos	3 minutos	56
VI Archivos mensuales suscriptores	8 (1 archivo)	10 segundos	2 minutos	12
OT Reclamos	65 M (1 archivo)	45 segundos	3 minutos	40
TL Tráfico	5 M (3 archivos)	35 segundos	2 minutos	20
LD Tráfico	2 M (4 archivos)	45 segundos	3 minutos	25
TM Tráficos	3 M (4 archivos)	50 segundos	3 minutos	15
TL Líneas en servicios	1,5 M (2 archivos)	30 segundos	2 minutos	21
TM Abonados móviles 30 y 90 días	30 (1 archivo)	10 segundos	3 minutos	15

Proceso	Registro promedio procesamiento	Tiempo	Procesamiento	Nº empresas
OT Conexiones 30 y 90 días	200 (1 archivo)	20 segundos	3 minutos	39
TV Archivos mensuales	300 (1 archivo)	20 segundos	3 minutos	13
TL Demanda horaria v2.0	300 M (1 archivo)	20 segundos	2 minutos	12
OT Calidad disponibilidad enlace	6 M (1 archivo)	50 segundos	3 minutos	19
OT Calidad red Internet	40 M (11 archivos)	45 segundos	3 minutos	18
TM Bloqueo y desbloqueo	10 M (1 archivo)	30 segundos	2 minutos	8
TM CBS SAE	33 M (1 archivo)	35 segundos	2 minutos	8
TM Demanda horaria v2.0	22 M (1 archivo)	30 segundos	2 minutos	19
TM Numeración asignada v2.0	1 (1 archivo)	10 segundos	2 minutos	15
TM Recargas prepago	2 (1 archivo)	10 segundos	2 minutos	13
TM SAR SAE	200 (1 archivo)	20 segundos	2 minutos	13
VI Archivos trimestrales (tráfico)	100 (1 archivo)	15 segundos	2 minutos	12
OT Densidad potencia	200 (300 archivos)	20 segundos	2 minutos	24
OT Inversión y empleo	20 (2 archivo)	15 segundos	2 minutos	72

El gran volumen de datos que maneja actualmente el sistema, junto con las reaperturas de procesos, hacen que el STI no sea capaz de procesar adecuadamente todos estos datos.

Actualmente el cuello de botella de la aplicación son dos procesos en particular "calidad de red móvil" e "interrupciones y descuentos", esto dado que son los procesos más complejos y con mayor volumen de datos.

Solución

A fin de solucionar este problema se ha pensado en revitalizar el STI incorporando varias mejoras en el proceso de validación y carga de datos.

- Nuevo motor de BD especial para Big Data Open Source.
- Nuevo orquestador de procesos.
- Procesos de validación generados a medida para un mayor rendimiento.

Primera etapa: evitar el hundimiento

La primera etapa consiste en evitar que el STI colapse completamente, para esto se migró los 2 procesos más "pesados" a la nueva arquitectura informática.

Segunda etapa: mirar al futuro

La segunda etapa contempla migrar el resto de los procesos del STI a fin de actualizar completamente la plataforma.

Adicionalmente se evalúa la posibilidad de realizar mejoras profundas en la forma de funcionar del STI, incorporándose entre otras mejoras:

1. Transferencia electrónica de datos.
2. Disminuir los tiempos de recepción de los datos y fraccionar la carga.
3. Aportar al negocio con información más que con datos.

La evolución de la plataforma STI

Futuro del STI (nuevo)

Futuro en los tiempos de respuesta

En DDT se han realizado pruebas para medir el rendimiento de esta nueva estructura y compararlo con el STI actual.

Para estas pruebas se utilizó el proceso "**Calidad de red móvil**" migrado y validado por DDT.

Proceso	Empresa	Periodo	Nro. Registros	STI 2.5	STI 3.0
Calidad de red móvil	Claro	01/2014	3.227.675	13:52:15	00:24:32
Interrupción y descuento	Movistar	03/2015	557.280	49:24:15	00:05:24

Los tiempos fueron obtenidos en condiciones muy similares, sin embargo, estos pueden varias en ambiente productivo.

› Conclusiones

Podemos concluir que el desafío del *big data* no solo tiene la complejidad de derribar barreras técnicas, sino que también debe derribar barreras culturales, el

big data tiene que ver con las personas y los procesos en los cuales están involucradas. Sin embargo, no debemos sesgar el análisis a una variable técnica, sino que debemos amplificar el análisis a múltiples variables que tienen que ver con la gobernanza y la toma de decisiones, ya que factores políticos y culturales afectarán los resultados. Sería absurdo pensar que *big data* resuelve todos los problemas, pero el análisis debe ser certero en encontrar incentivos correctos para la incorporación de elementos técnicos a la gestión del Estado.

Referencias bibliográficas

Analytics: el uso de big data en el mundo real. IBM Institute for Business Value (2012).

Fuentes, M. (2016). *La Tercera Plataforma y su Impacto en el Negocio.* IDC.

Marinkovic, E. (2016). *Estrategia e Integración de Datos Big Data, Instrumento de Apoyo para Optimizar la Operación*. SII.

Munizaga, M. (2014). *Big data en Chile: desafíos y oportunidades*. Universidad de Chile.

Rodríguez, P. (2016). *Apoyando la formulación de políticas públicas y toma de decisiones en educación utilizando técnicas de análisis de datos masivos: el caso de Chile.* Universidad de Chile y Universidad Adolfo Ibáñez.

Capítulo 15

La política de *big data* en Colombia: una apuesta conjunta del sector público, el sector privado y la academia

María Isabel Mejía Jaramillo*

Introducción

Durante los últimos cuatro años tuve la fortuna de ser la viceministra de Tecnologías y Sistemas de la Información en momentos en que el mundo se prepara para la "Cuarta Revolución Industrial", un concepto acuñado por el profesor Klaus Schwab en la última versión del Foro Económico Mundial en Davos, Suiza (Schwab, 2016). Podíamos ver cómo se empezaban a masificar algunas tendencias tecnológicas como el *big data*, el "Internet de las cosas", la robótica y la inteligencia artificial, tecnologías intensivas en conocimiento.

Colombia históricamente no se ha distinguido por ser un *early adopter* de la tecnología y generalmente ha sido más un consumidor que un productor o generador de estas innovaciones. Sin embargo, teniendo en cuenta el gran potencial en términos de talento humano que tiene nuestro país, la apuesta que como gobierno habíamos hecho para la formación de talento en tecnologías de información, y las grandes oportunidades de generación de valor para la industria, para el sector público y para la sociedad en general que vimos en estas tendencias tecnológicas, decidimos que esta vez no íbamos a llegar tarde; que nos íbamos a subir en la ola de inmediato y que nos íbamos a convertir en un "hub" de *big data* para la región, y muy pronto en referente no solo para América Latina sino a nivel mundial.

El siguiente paso era crear las capacidades en los actores del ecosistema TIC del país para estar a la altura de semejante reto y desde el Ministerio TIC decidimos poner una semilla: lideramos la conformación de dos centros de excelencia y apropiación: el de Big Data y Data Analytics y el de "Internet de las cosas". Estos centros de investigación aplicada se conformaron como alianzas entre las más prestigiosas universidades del país, los líderes mundiales

* Directora ejecutiva de Info Projects, exviceministra TI de Colombia.

de la industria de tecnologías de información y empresas líderes de otros sectores de la economía.

Este es solo un ejemplo de las acciones que se están desarrollando en Colombia para masificar el uso de *big data*, pero lo más importante de este esfuerzo es la colaboración y articulación que se logra entre los diferentes actores del ecosistema, que para mí es uno de los factores de éxito para la innovación en países como el nuestro.

En los siguientes apartados de este capítulo mostraré los avances de Colombia en términos de la política pública de *big data*, ilustraré algunos casos de éxito del sector público, y otros del sector privado, resaltaré el rol de las empresas de tecnología y finalmente los aportes de la academia.

Finalmente, quiero agradecer a todas las personas que dedicaron su tiempo para contarme sus experiencias, logros, retos, incertidumbres y certezas, pero sobre todo agradecer su compromiso con este gran sueño que ya está mostrando resultados concretos.

La política pública de *big data* en Colombia

Departamento Nacional de Planeación

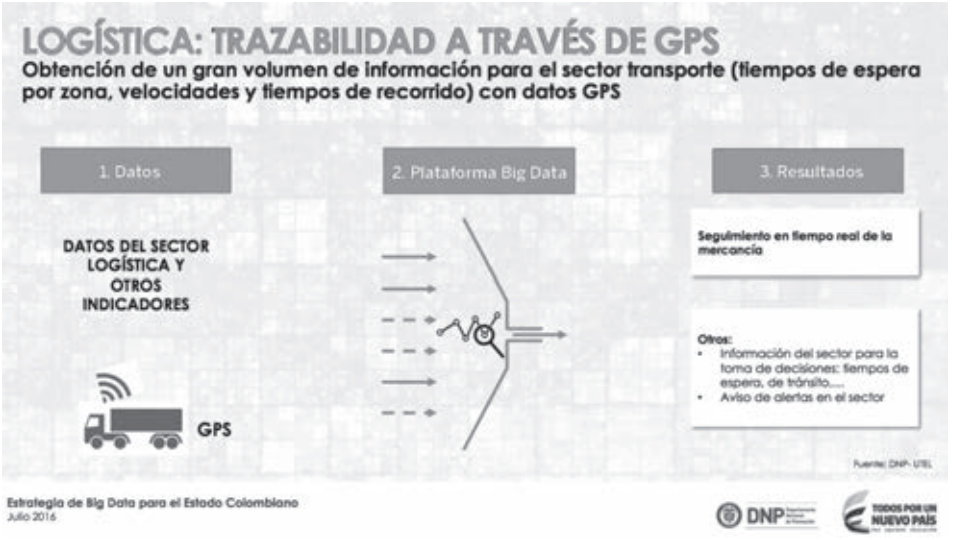
La Ley 1753 del 9 de junio de 2015 a través de la cual se expide el Plan Nacional de Desarrollo 2014-2018 "Todos por un nuevo país", le asigna al Departamento Nacional de Planeación —DNP— la función de liderar la estrategia de *big data* para el estado colombiano. El DNP como estructurador de la política pública colombiana y gestor de la inversión en la nación, ve en el *big data* un instrumento para tomar decisiones más asertivas, mejor informadas, basadas en análisis cruzados y considerando las externalidades (positivas y negativas) en los diferentes sectores. Es considerada como una herramienta para aumentar la productividad del Estado y del sector privado, aportar al crecimiento económico y generar valor público (DNP, 2016).

El DNP ha desarrollado algunos proyectos piloto con el fin de fortalecer el proceso de formulación de la política pública de *big data*, entre los cuales cabe mencionar los siguientes:

- Observatorio de logística: el objetivo de este proyecto piloto es mantener actualizada la perspectiva completa de las necesidades y avances en logística del país, con base en información sobre el desempeño y las necesidades

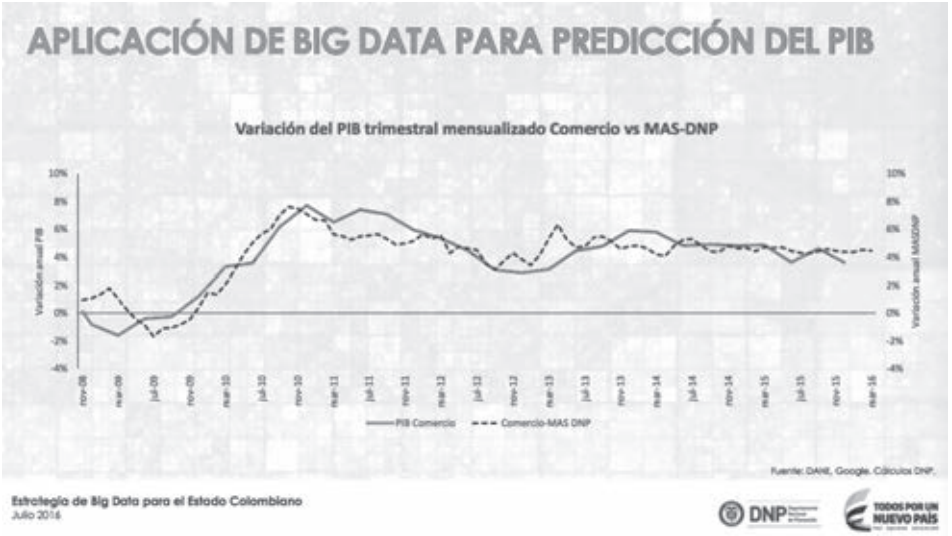
empresariales, las tecnologías e innovación, las necesidades de la cadena de aprovisionamiento, información sobre la infraestructura, los mercados, las políticas y las necesidades regionales. Uno de los principales ámbitos de evaluación del desempeño logístico se concentra en la trazabilidad y el hecho de conocer en cada momento donde se encuentra el *stock*. En el gráfico 1 se presenta el esquema de funcionamiento del observatorio.

Gráfico 1. Esquema de funcionamiento del Observatorio de Logística



- MAS-DNP: el Modelo Anticipado Sectorial es la herramienta que construyó el DNP con base en Google Trends y variables económicas observadas para anticipar el comportamiento del PIB utilizando la técnica "NOWCASTING" —modelo de factores dinámicos—. Es una herramienta útil, rápida y económica para el análisis de fenómenos sociales y económicos. En el gráfico 2 se puede ver la variación del PIB trimestral mensualizado del sector comercio calculado por el Departamento Administrativo Nacional de Estadística *versus* el que se calcula con el MAS-DNP. Dentro del análisis económico Google Trends ha sido utilizado como índice de la actividad económica real dado el mínimo rezago con que se puede obtener la información.

Gráfico 2. Aplicación de *big data* para predicción del PIB



Ministerio de Tecnologías de la Información y las Comunicaciones

De acuerdo con la Ley 1341 del 30 de julio de 2009 el Ministerio de Tecnologías de la Información y las Comunicaciones tiene dentro de sus objetivos formular políticas, planes, programas y proyectos del sector TIC con el fin de contribuir al desarrollo económico, social y político de la nación, y elevar el bienestar de los colombianos. En cumplimiento de este objetivo, en el año 2010 se definió el Plan Vive Digital, el cual busca impulsar la masificación del uso de Internet en Colombia para reducir la pobreza, generar empleo y aumentar la competitividad del país.

En el marco de la iniciativa de "Investigación, Desarrollo e Innovación de TIC", una de las iniciativas del Plan Vive Digital, el Ministerio TIC y Colciencias[1] en el año 2015 abrieron una convocatoria pública para crear un Centro de Excelencia y Apropiación en *big data* y Data Analytics para generar soluciones innovadoras apalancadas en TIC y para agregar valor a los sectores estratégicos del país a través del análisis, la ciencia y la ingeniería de los datos. La propuesta ganadora de esta convocatoria fue CAOBA[2], una alianza conformada por cuatro de las mejores universidades del país: la Universidad de los Andes, la Universidad

1. Colciencias es el Departamento Administrativo de Ciencia, Tecnología e Innovación de Colombia.
2. El Ministerio TIC aportó 3.313 millones de pesos y los miembros de la alianza aportaron 5.975 millones de pesos para el primer año de operación de CAOBA.

Javeriana, la Universidad EAFIT y la Universidad ICESI; tres de los más reconocidos proveedores de tecnología a nivel mundial: IBM, EMC y SAS; una entidad de incubación y fomento de empresas de base tecnológica: CREATIC; una entidad de gobierno estratégica en el manejo de grandes volúmenes de información y líder de la política de *big data* del país: el Departamento Nacional de Planeación; y dos empresas líderes que aportan su experiencia, conocimiento y problemáticas para el fortalecimiento del centro: Bancolombia y Nutresa (Caoba, 2015).

CAOBA (http://alianzacaoba.co) es un Centro de Excelencia y Apropiación en Big Data y Data Analytics, comprometido con la innovación y la generación de productos y servicios que promuevan el desarrollo y competitividad del país y de la región. CAOBA toma los principios y la visión de sus fundadores y los integra para generar valor a partir del desarrollo de sus actividades que se concentran en cuatro ejes:

- Investigación aplicada fundamentada en problemas relevantes para el país y desarrollada con la rigurosidad científica requerida para trascender en el área de conocimiento y posibilitar la transferencia tecnológica de las soluciones creadas.
- Consultoría que favorezca la apropiación de tecnologías, metodologías y herramientas de análisis y procesamiento de grandes volúmenes de información en el país y la región.
- Formación que integre los saberes y experiencia de las partes que la conforman para contribuir de manera efectiva en la apropiación de las temáticas asociadas a *big data* y *data analytics* en el país.
- Apoyo de iniciativas de emprendimiento y *spin-offs* colombianos cuya propuesta de valor esté fundamentada en la generación de soluciones alrededor de las temáticas concernientes a CAOBA.

Otra de las iniciativas adelantadas por el Ministerio TIC es la estrategia de Fortalecimiento de la Industria de Tecnologías de la Información —FITI—, la cual tiene como propósito contribuir a la transformación de la industria de TI en un sector competitivo y de clase mundial. Con el fin de contar con herramientas para monitorear la industria y contar con información que permita no solo definir los lineamientos de política pública sino evaluar los resultados e impactos de la misma, el Ministerio TIC en conjunto con la Federación Colombiana de la Industria del *Software* y Tecnologías Informáticas Relacionadas —Fedesoft—, implementaron el "Observatorio TI" http://observatorioti.co/ y realizaron estudios de caracterización que han generado importantes hallazgos entre los cuales vale la pena destacar los siguientes:

- De acuerdo con el "Informe de Caracterización de la Industria de *Software* y Tecnologías de la Información" en Colombia se han creado nuevos perfiles de empleo de acuerdo con el desarrollo de la industria y se prevé que surgirán más. En este momento los perfiles más demandados y claves para el sector son: "gerente de infraestructura", "analista de inteligencia de negocios", "científico de datos", "ingeniero de seguridad de la información", "analista de seguridad", "ingeniero de *big data*", "modelador de datos y auditor TI" (MINTIC, 2015).
- Según el "Estudio de Salarios del Sector de *Software* en Colombia 2015" para el 2020 la demanda de talento digital será de 123.918 profesionales, aumentando a un ritmo vertiginoso su importancia relativa frente al total del empleo calificado del país (MINTIC, 2015).
- El "Estudio de Caracterización de la Brecha de Talento Digital en Colombia-2015" calculó la brecha de profesionales TI en la industria de tecnologías de la información para el 2016 en 5.986, para 2017 esta llegaría a 7.242 y para 2020 alcanzaría 12.098 profesionales. El acumulado de la brecha 2016-2020 sería de 44.267 plazas de la industria TI colombiana sin cubrir por la oferta probable (MINTIC, 2015).
- Una de las líneas de acción de FITI es la de talento TI la cual se ha enfocado en cerrar la brecha de profesionales TI para la industria, que es una brecha no solo en lo cuantitativo sino en lo cualitativo. El Ministerio TIC ha adelantado campañas para motivar a los jóvenes a estudiar carreras TI; ha destinado recursos para financiar el estudio de carreras TI, a nivel técnico, tecnológico, universitario y maestría, a través de créditos condonables hasta en un 100%, a través de los cuales se han beneficiado más de 9.200 personas. Con el fin de ayudar a cerrar la brecha cualitativa, el Ministerio TIC ha realizado inversiones para que los profesionales TI puedan adquirir las competencias técnicas y las competencias transversales que requiere la industria.

Específicamente en el campo de Big Data Analytics se formaron y certificaron con Microsoft 60 profesionales y actualmente se encuentran en proceso de formación con la firma BD Guidance más de 1.000 profesionales los cuales se certificarán como Big Data Scientist, Big Data Architect, Big Data Consultant y Big Data Engineer.

Casos de éxito del sector público

Secretaría Distrital de Movilidad

La Secretaría de Movilidad de Bogotá viene adelantando un proyecto a través del cual instaló 350 sensores en las vías de la capital, los cuales ya están generando datos en tiempo real de ubicación y tiempo y pueden ser visualizados en el Centro de Gestión de Tráfico.

Esta información sirve para saber por dónde se mueve la población, dónde vive, dónde trabaja y dónde realiza otras actividades. La forma tradicional de tomar las decisiones sobre la movilidad en la ciudad era a través de encuestas; ahora se podrá hacer gestión pública a través de los datos y no de las percepciones. Alejandro Forero, gerente del Sistema Inteligente de Transporte de la Secretaría Distrital de Movilidad, dice que antes de finalizar el 2016 empezará a publicar todos los datos en la nube, en cumplimiento de la política de datos abiertos promovida por el Estado colombiano, con el fin de que otros actores externos a la Secretaría de Movilidad, como desarrolladores, empresas de la industria TI, emprendedores, entre otros, puedan analizar los datos, proveer soluciones para mejorar la movilidad y en general la calidad de vida de los ciudadanos.

UIAF

La Unidad de Información y Análisis Financiero —UIAF— https://www.uiaf.gov.co recibe información de más de 15.000 entidades del sector financiero, casas

de cambio, notarías, empresas de compra venta de oro, entre otras, en su función de detección y prevención del lavado de activos y financiación del terrorismo.

Empezaron a aprovechar a esta información hace varios años con técnicas de minería de datos y ahora están migrando a una plataforma de *big data*. El hecho de poder procesar en línea esos grandes volúmenes de información es estratégico para la entidad por la oportunidad. Asimismo, se han podido integrar datos de fuentes diversas, de formatos variados y de mayor magnitud, con gran facilidad para escalar.

Según Luis Alejandro Navas, subdirector de Informática, uno de los factores de éxito de este tipo de proyectos es la buena definición del caso de uso, de los límites, del alcance. Otro factor es el talento, que para este tipo de proyectos es costoso y difícil de conseguir. En la UIAF crearon un grupo de analítica dentro de la entidad, conformado por un grupo interdisciplinario de funcionarios que cumplen el rol de científicos de datos, con una característica que ha sido excelente para el éxito de los proyectos y es que todos saben programar.

Casos de éxito del sector privado

Nutresa[3]

Había identificado hace varios años a *big data* como una tendencia tecnológica clave para apalancar las estrategias del grupo. En un ejercicio de prospectiva a 2030 identificaron *big data*, "Internet de las cosas" y computación cognitiva como tecnologías clave para alcanzar las metas a largo plazo y como parte del proceso de transformación digital. En ese momento apareció la oportunidad de participar en la alianza Caoba la cual ofrecía un escenario perfecto para aprender y contar con una estrategia más estructurada.

La estrategia de Nutresa pretende mostrar casos reales del uso de *big data*, mostrar que es una realidad y no ciencia ficción. "*Big data* no es el Oráculo de Delfos, no contesta todo, requiere una metodología clara, es un proceso de aprendizaje, se requiere el desarrollo de capacidades nuevas y perfiles nuevos dentro de la organización", dice Juan Mauricio Montoya, gerente de Servicios Financieros y TI de Servicios Nutresa. En este momento cuentan con un equipo de personas enfocadas en el análisis de información en las áreas de mercadeo,

3. Grupo Nutresa, S.A., es la compañía de alimentos procesados líder en Colombia y uno de los jugadores más importantes en el sector en América Latina http://www.gruponutresa.com/quienes-somos/?lang=en

ventas y servicio al cliente. Se encuentran desarrollando casos de uso enfocados en el conocimiento del cliente, del consumidor y del comprador.

Con CAOBA están desarrollando un proyecto de segmentación de consumidores basado en información no estructurada proveniente de redes sociales. Paralelamente Nutresa, dentro de su proyecto de estrategia digital, tiene un capítulo de *big data* orientado a crear las capacidades internas, desarrollar la cultura de la analítica dentro de la organización, basada en datos y en evidencias estrictas, definiendo cuál debe ser la estructura organizacional que soporte la estrategia: si deben contar con centros especializados o descentralizados, cuáles son las tecnologías que se van a adoptar (*hardware* y *software*) y cuáles serán sus aliados tecnológicos.

Por ahora el liderazgo es del área de tecnología y trabajan por proyectos con las áreas usuarias, entre las cuales se encuentran la de logística y la de inteligencia de mercados que es el aliado natural y uno de los más adelantados dentro de la compañía.

Finalmente, Juan Mauricio comenta que Colombia está arrancando bien, que hay una apuesta muy grande y visionaria del Gobierno y desde el Ministerio TIC y el DNP, pero se necesita talento humano. Ese es el reto más grande. Hay que lograr una masa crítica con todos los actores conformando una red de conocimiento alrededor del mundo de los datos.

Bancolombia[4]

Bancolombia desde hace muchos años venía adelantando procesos de analítica en diferentes áreas de la organización (riesgos, mercadeo, etc.) pero en el año 2013, después de un proceso de planeación estratégica denominado "Visión 2020", decidió abordar el tema de analítica como una estrategia corporativa, transversal para todas las áreas del banco. Según Pablo Arboleda, gerente de Capacidades Analíticas de Bancolombia, en el banco más que hablar de *big data* hablan de una estrategia analítica, como un "proceso riguroso, científico, repetible, de transformar datos en información, con dos propósitos: tomar mejores decisiones y crear nuevos productos y servicios".

La estrategia analítica actualmente se lidera desde la Vicepresidencia de Innovación con el apoyo del área de tecnología. Reconocen que la tecnología es la "columna vertebral" de la estrategia, pero la estrategia es mucho más que un

4. El Grupo Bancolombia es el conglomerado de empresas financieras más grande de Colombia. http://contenido.grupobancolombia.com/webCorporativa/nosotros/contenido/historia2.asp

proceso de apropiación o incorporación de una tecnología. Con el fin de crear las capacidades para poder desarrollar la estrategia de analítica a nivel corporativo decidieron trabajar en los siguientes aspectos:

1. Definir la visión de la estrategia y tener claridad sobre lo que querían hacer y para qué les iba a servir.
2. Atracción y desarrollo de talento. Se integraron personas del banco que estaban trabajando en diferentes roles pero que tenían la formación necesaria para trabajar en analítica, se iniciaron procesos de formación con otros empleados que requerían nuevos conocimientos y habilidades, y trajeron otras personas de afuera. Se requerían diferentes perfiles: desde los más técnicos hasta los analíticos encargados de hacer los modelos, todos con la capacidad de entender el negocio.
3. Cultura. El tema de gestión del cambio en la cultura de la organización se realizó a través de proyectos conjuntos con las diferentes áreas. Según Pablo Arboleda, "el tema de cultura no es cuestión de una cartilla, se debe liderar en el día a día con los equipos; no es el discurso, sino el hacer".
4. Traer más conocimiento al banco. Este tema es diferente al del talento, va más allá de traer personas o formar empleados. El conocimiento se logra a través de la interacción, de las conversaciones con los diferentes actores, alrededor de los proyectos, y en este punto el gerente resalta la alianza CAOBA, la cual considera como un semillero de talento, como un nodo de una red que jalona un ecosistema y como un gran escenario para adquirir conocimiento.
5. Lograr esquemas de datos más robustos haciendo énfasis en la calidad de los mismos.
6. Arquitectura de información y tecnología que soporta el proceso.

La estrategia corporativa ha permitido ir alineando los equipos, que tienen diferentes grados de madurez y muy buenos niveles de conocimiento; saben cuánto cuesta desarrollar proyectos de analítica y ya se ven resultados que se pueden medir. Han desarrollado proyectos para solucionar temas de seguridad, riesgo, atención de clientes y otros para crear nuevos productos menos tradicionales. En síntesis, la estrategia analítica sirve para optimizar el negocio actual, crear nuevos negocios y también para encontrar otros espacios de competencia a través de proyectos experimentales aprovechando datos no estructurados y nuevas fuentes de información para resolver los problemas de formas no convencionales. Pablo dice que al poder combinar los datos de la organización con los datos de afuera "se pasa del mundo de la escasez al mundo de la abundancia".

Con CAOBA están desarrollando un proyecto que se llama "Comunidades", el cual parte de la hipótesis de que a partir de las transacciones de los clientes se pueden identificar las comunidades que hay entre ellos, entonces no se habla del riesgo de un cliente, sino de una comunidad.

Casa Editorial El Tiempo[5]

Hay industrias o compañías más predispuestas a aprovechar los datos, que es ahora la tendencia, como es el caso de El Tiempo. Desde hace 90 años, para poder entregar el periódico a cada suscriptor en su casa, debía tener los datos de cada uno de ellos completamente actualizados. En El Tiempo existe una cultura de manejo del dato y una cultura de manejo de las suscripciones que están en el ADN de la compañía. El *core* del negocio es su base de datos.

Hasta hace unos cuatro años los datos se usaban para lo que fueron creados: para entregar el periódico, para diseñar estrategias de fidelización de sus suscriptores y para hacer algunos ejercicios de analítica para evitar la deserción. El Tiempo contaba con una base consolidada de clientes y 28 millones de usuarios navegando en su red de portales. Habían implementado el proyecto de CRM para gestionar mejor la base de suscriptores y contaban con herramientas tecnológicas como *software* para minería de datos. Sin embargo, estas iniciativas estaban desarticuladas y se veía una gran oportunidad en integrar todas estas piezas que ya tenía la organización, juntar toda la data para tener una visual completa del cliente y aprovecharla para tareas diferentes a las tareas para las que fue creada.

Es así como se decide crear la Gerencia de Conocimiento y Gestión de Audiencias, liderada actualmente por Tito Neira, estadístico y matemático con especialización en mercadeo y MBA. Dentro de la gerencia se crearon tres áreas inspiradas en las características que debería tener un científico de datos:

- Área de gobierno de datos: cuyo alcance va más allá de cumplir la ley de protección de datos personales. Se han establecido reglas para garantizar la calidad, completitud, uso, integridad y ética para el manejo de la información al interior de la compañía. Gobierno de datos para lo digital y lo no digital.
- Área analítica: esta área toma la data que dispone el área de gobierno de datos y hace analítica de minería de datos, analítica de métricas digitales y analítica que se deriva de la investigación.
- Área de gestión de audiencias: esta área inicialmente tomó el CRM, que era la herramienta para gestionar la audiencia. Sin embargo, se amplía el universo

5. *El Tiempo* es el diario de más alta circulación en Colombia. http://www.eltiempo.com/

de los datos: por un lado, están los datos que tiene El Tiempo de las transacciones que hacen sus suscriptores con ellos, y por otro lado el universo digital de esos clientes. La política adoptada es tomar la data propia, es decir, la de los 28 millones de usuarios que navegan en sus portales y que ya autorizaron a El Tiempo a tomar la información, y asimismo, solo para los clientes que lo autoricen expresamente, tomar información de sus redes sociales, para enviarles ofertas o para generarle contenido personalizado.

Para implementar la estrategia han capacitado a personas de todas las áreas de la compañía. "Es muy difícil conseguir a alguien con el perfil de científico de datos. Es como volver a la época de Da Vinci", dice Tito Neira. Es así como nace una nueva línea de negocio de El Tiempo que consiste en vender conocimiento. Ya tienen dos productos basados en datos listos para ofrecer a sus clientes:

- Publicidad digital segmentada: ofrece a los anunciantes del mercado de la publicidad conocimiento para hacer acciones de marca. Son estudios que parten del análisis de los datos que tiene El Tiempo y que apoyan a las empresas en la toma de decisiones a la hora de escoger un segmento específico para sus planes de mercadeo.
- *Marketing* directo: este servicio que ofrece El Tiempo consiste en ejecutar campañas de mercadeo requeridas por sus anunciantes para hacer llegar a los usuarios/consumidores una información o un producto, a través de medios digitales o a través del canal físico, con el periódico, aprovechando este canal de distribución que llega a más de 500 municipios colombianos. Es un nuevo servicio, apalancado en información, aprovechando su capacidad instalada, donde se une lo digital con lo real.

Ya están avanzando en otros productos como, por ejemplo, el de ofrecer contenido personalizado a sus usuarios, de tal manera que cuando naveguen en la red de portales de El Tiempo, les aparezca el contenido que les interesa. Hay dos grados de personalización: orientada a segmentos o a nivel de usuario. Ya están pensando en usar técnicas de *machine learning* para aprender del comportamiento de los usuarios y que cada vez les aparezca contenido más pertinente. Y quieren ir más allá: ¡no solo que el contenido personalizado aparezca en la versión digital sino en el periódico impreso! O utilizar procesamiento de lenguaje natural para que podamos decirle "Tiempo: ¿cuáles son las noticias relevantes para mí el día de hoy?". En síntesis, quieren combinar las capacidades logísticas del diario impreso, con el valor de los datos y la computación cognitiva.

Rol de las empresas de tecnología

Las compañías multinacionales proveedoras de tecnología juegan un papel importantísimo en la masificación de *big data* y analítica en Colombia. Estas compañías hace muchos años vienen trabajando en análisis de información, desarrollando proyectos y soluciones para generar conocimiento y agregar valor en múltiples industrias del sector privado y del sector gobierno.

Javier Rengifo, Senior BD&Analytics Architect, comenta que IBM desde hace varios años ha desarrollado proyectos con el sector financiero enfocados en el análisis de clientes para entender mejor su comportamiento y poderle prestar un mejor servicio, facilitar el contacto, optimizar campañas de *marketing*, fidelización y retención de clientes, gestión del riesgo de crédito, de cartera, de incumplimiento de los pagos y cumplimiento de la reglamentación sobre lavado de activos y financiación del terrorismo. También proyectos con empresas del sector telecomunicaciones analizando los registros de llamadas para encontrar patrones y poder determinar problemas de congestiones en la red, incumplimiento de niveles de servicio, oportunidades comerciales, campañas de mercadeo, retención de clientes y mejoramiento del servicio al cliente. "Hoy en día IBM lidera a nivel mundial la computación cognitiva que va un paso más allá de *big data* y Analítica", dice Javier Rengifo. La herramienta desarrollada por IBM, llamada WATSON, permite aprovechar todas las fuentes de información ricas y variadas para utilizarlas y analizarlas con un sistema que aprende; inteligencia artificial aplicada al servicio de la humanidad; permite interactuar con las personas, que le hagan preguntas y contesta como un asesor externo que resuelve dudas y ayuda a tomar decisiones; permite aprender de la información y generar nuevo conocimiento.

Por su parte, Vivian Jones, *country manager* de SAS, dice que esta compañía lleva 40 años desarrollando e implementando soluciones en el campo del análisis de información. Ahora ofrece soluciones "end-to-end" como por ejemplo la plataforma de "Anti-Money Laundering" para detectar una transacción sospechosa, investigarla, alertar al sistema, reportar a la UIAF y cerrar el caso. O la plataforma de "Marketing Contextual" que permite desarrollar estrategias de *marketing* personalizadas y en tiempo real que tienen una efectividad de hasta tres veces las campañas tradicionales que se hacían solo con procesos de segmentación. Ahora se une la segmentación con la plataforma de *marketing*, se lanza la campaña, recibe la retroalimentación del cliente: la acepta o no la acepta. El modelo va aprendiendo y cada vez es más asertivo. Los volúmenes de información en sectores como la banca, telecomunicaciones y *retail* se han

multiplicado y aparece *big data* con las capacidades para almacenar, procesar y revisar **toda** la data, no una muestra, obteniendo un mayor nivel de asertividad. Sin embargo, es necesario monetizar los proyectos de *big data*: contar con toda la información no es suficiente; es necesario analizarla en tiempo real, explotarla, generar conocimiento, usarla para tomar decisiones. Las áreas de negocio necesitan tener una visión 360 grados del cliente. "La analítica es lo que de verdad le da valor al negocio", dice Vivian.

Las compañías de tecnología invierten grandes cantidades de dinero en investigación y desarrollo de innovaciones que puedan ayudar a sus clientes a solucionar sus problemáticas, por lo tanto, son un actor fundamental en la transferencia de conocimiento al país y en la formación de talento. Tanto SAS como IBM vieron en CAOBA una gran oportunidad para desarrollar proyectos con impacto real a nivel país, generar conocimiento con tecnologías de punta, contar con profesionales capacitados. "Con este esquema de colaboración, que es un esfuerzo país, único en Latinoamérica por integrar a los actores de la academia, el sector público, las empresas de tecnología y las empresas del sector real, Colombia se pone a la vanguardia en la implementación de proyectos de *big data* y Data Analytics", dice Javier Rengifo.

Aportes de la academia

Según Claudia Jiménez, profesora asociada del Departamento de Ingeniería de Sistemas y Computación de la Universidad de los Andes, hay tres tipos de ofertas académicas en el mundo alrededor de estos temas:

- La primera orientada hacia reforzar el área de negocios aprovechando la información como una oportunidad para mejorar los procesos.
- La segunda orientada hacia programas de tipo científico que buscan cómo hacer mejor ciencia o avanzar en procesos de innovación tecnológica a través de la información.
- La tercera opción gira alrededor de los retos de la tecnología, sobre cómo lograr diseñar e implementar infraestructuras que permitan aprovechar la información.

Teniendo en cuenta que en Colombia el mercado no es tan amplio como para ofrecer muchos programas con diferentes enfoques, la Universidad de los Andes diseñó un programa de Maestría en Ingeniería de Información, que busca que los estudiantes no solo tengan una visión de negocio, sino también que conozcan la tecnología necesaria para sacarle provecho a esa información y que además

puedan implementar proyectos y soluciones empresariales en los cuales la información es el centro del problema. Estamos haciendo la tecnología para las nuevas preguntas, tenemos que formar un montón de gente que sepa hacer las nuevas preguntas, que sepa aprovechar la información que circula por ahí y que sepa analizarla para tomar las nuevas decisiones. Eso es lo que hay detrás de la Maestría en Ingeniería de Información", afirma Claudia Jiménez.

La universidad ofrece también cursos de educación continuada y desde el Departamento de Ingeniería Industrial se ofrece la Maestría en Analítica, totalmente complementaria porque se enfoca en estudiar cómo la estadística, a través de diferentes modelos, ayuda a resolver los distintos tipos de problemas.

Por su parte, la Pontificia Universidad Javeriana inició este año una Maestría en Analítica para la Inteligencia de Negocios y también ofrece varios diplomados dentro de los programas de educación continuada. En relación con proyectos de investigación, la universidad, en alianza con el Hospital Universitario San Ignacio ha desarrollado proyectos de analítica y *big data* principalmente enmarcados en las necesidades del sector salud, dentro de los cuales vale la pena destacar: DISEARCH que permite buscar y priorizar historias clínicas electrónicas (HCE) usando en su búsqueda los datos estructurados y no estructurados proveniente de millones de historias clínicas electrónicas y EXEMED que es un *software* para el análisis de indicadores de guías de práctica clínica analizando también la información estructurada y no estructurada.

En cuanto a la participación en el Centro de Excelencia y Apropiación, Alexandra Pomares, profesora asociada del Departamento de Ingeniería de Sistemas, resalta como una de las fortalezas de CAOBA la complementariedad entre todos los actores y la agilidad que le imprime la presencia de las empresas del sector privado.

Para Darío Correal, profesor asociado de la Universidad de los Andes, CAOBA es una iniciativa muy importante en la que había que estar presente. Piensa que la universidad debe apoyar este tipo de iniciativas país y que es una oportunidad única para mostrar el tipo de proyectos que la universidad puede hacer. Estar cerca del Departamento Nacional de Planeación que tiene una visión macro del país, poder colaborar con otras universidades y con otras facultades de la misma universidad, formar estudiantes de doctorado y maestría, trabajar de la mano con el sector privado y resolver problemas reales parecía muy interesante y fueron algunas de las razones que los llevaron a tomar la decisión de formar parte de la alianza.

Reflexión final

El camino para la implementación de *big data* y Data Analytics es largo, pero ya se han dado importantes pasos. Si el país quiere ser referente no se puede desfallecer en el fortalecimiento y ampliación de las alianzas universidad-empresa-Estado. También se requerirá de la participación de los organismos multilaterales para, no solo replicar casos de éxito en otros países de la región, sino establecer alianzas entre los actores de los ecosistemas de las naciones.

Referencias bibliográficas

CAOBA (2015). *Creación y Operación de CAOBA - Colombian Center of Excellence and Appropriation on Big Data and Data Analytics.* Bogotá, Colombia.

DNP (2016). *Estrategia de Big Data para el Estado Colombiano*. Bogotá, Colombia.

MINTIC (2015). Informe de Caracterización de la Industria de Software y Tecnologías de la Información. http: //observatorioti.co/k_course/caracterizacion-del-sector-ti/.

MINTIC (2015). Estudio de Salarios del Sector de Software en Colombia 2015. http: // observatorioti.co/k_course/estudio-de-salarios-del-sector-ti/.

MINTIC (2015). Estudio de Caracterización de la Brecha de Talento Digital en Colombia — 2015. http: //observatorioti.co/k_course/brecha-digital/.

Schwab, K. (2016). The Fourth Industrial Revolution: What it means, How to respond https: //www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/

Capítulo 16

Oportunidades del *big data* en el sector público costarricense

Edwin Estrada Hernández*

¿Qué es *big data*?

Lo primero que hay que preguntarse al iniciar con este tema es ¿qué es *big data*? Se puede establecer que generalmente este concepto es utilizado para describir inmensas cantidades de datos, que podrían estar o no estructurados; y que en razón de esta cantidad y la diversidad de fuentes de las que provienen, no es posible manejarlos a partir de las bases de datos habituales desde el punto de vista informático, ya que se requeriría demasiado tiempo y recursos para examinarlas y estudiarlas de forma tradicional. Aunado a lo anterior, para que pueda determinar que existe big data, además del volumen de la información, se requiere necesariamente una variedad considerable de información, así como de alta velocidad de las redes.

La variedad se refiere a los diferentes tipos de datos que se puedan generar, como por ejemplo, audio, vídeo o ubicación, los cuales a su vez son recopilados por medio de los múltiples dispositivos que pone a disposición el mercado y los avances tecnológicos. Esta diversidad hace posible la gran cantidad de usos positivos que se le puede dar al *big data* en los procesos de toma de decisiones en la Administración Pública, lo cual conduce necesariamente a la velocidad, factor esencial del tema, con el objetivo de lograr que a partir de la cantidad y variedad de datos que se genera, exista la posiblidad de obtener nuevos datos que brinden información de manera rápida y oportuna para resolver de la mejor manera a los problemas que se planteen.

Este conjunto de las tres "V", volumen, variedad y velocidad, cuando confluyen es cuando se puede afirmar con certeza, que se está operando en big data.

Para ejemplarizar estas tres V se pueden ver los siguientes datos a nivel mundial[1]:

* Viceministro de Telecomunicaciones de Costa Rica.

1. Estudio: Big Data 2015, OBS Business School, http://landings.projectmanagement.obs-edu.com/informe-big-data-2015

- En volumen se tiene: 100 *terabytes* de datos se suben diariamente a Facebook; Akamai analiza 75 millones de eventos de un día para orientar los anuncios en línea; Walmart se ocupa de 1.000.000 transacciones de los clientes cada hora.
- En velocidad: en 1999, el almacén de datos de Wal Mart acumula 1.000 *terabytes* (1.000.000 *gigabytes*) de datos. En 2012, tuvo acceso a más de 2,5 *petabytes* (2.500.000 *gigabytes*) de datos. Cada minuto de cada día, que sube 100 horas de vídeo en Youtube, enviar más de 200 millones de correos electrónicos y enviar 300.000 tweets.
- En variedad: hoy el 90% de los datos generados es "no estructurados", que viene en todos los tamaños y formas de datos geoespaciales, desde los *tweets* que pueden ser analizadas por el contenido y sentimiento, a los datos visuales como fotos y vídeos.

Big data y el mundo

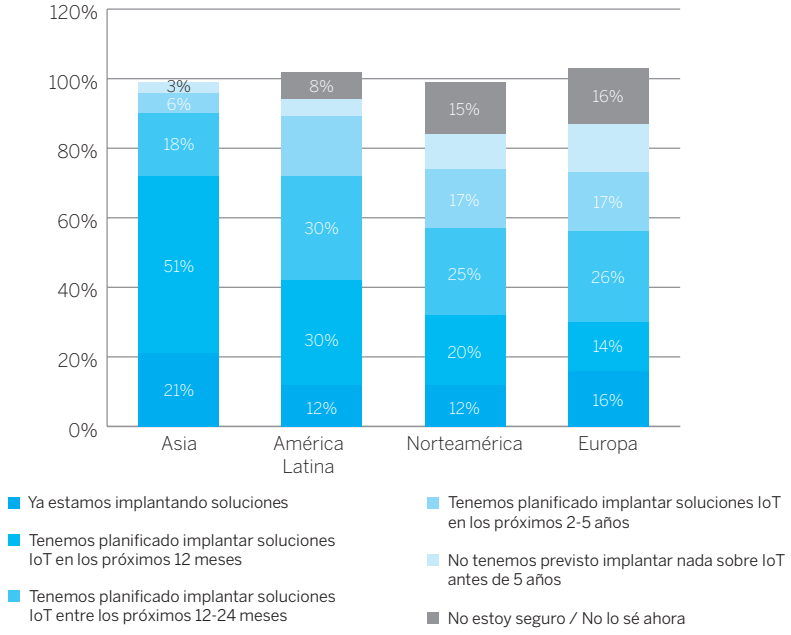
A nivel mundial, el OBS Business School presentó el estudio *big data* 2015[2], el cual refleja el crecimiento a nivel mundial del uso del *big data*. Al respecto indica que Norteamérica lidera la inversión y adopción de proyectos y herramientas *big data* con un 47%, además, señala que el resto de regiones registra también un aumento.

Asimismo dicho estudio indica que el 73% de las organizaciones están invirtiendo o tienen planificado invertir en *big data* en los próximos 24 meses, lo que confirma que a nivel mundial la tendencia de las organizaciones es crecer y aprovechar la gran generación de datos, como se demuestra en el gráfico de la página siguiente.

Como se puede observar, América Latina no es ajena a esa tendencia de crecimiento, y es por eso que los Estados de la región deben también aprovechar y utilizar para mejorar las funciones del Gobierno frente a los ciudadanos; más aún si se considera que para el año 2020 más de 30.000 millones de dispositivos estarán conectados a Internet[3], dentro de la evolución y crecimiento sostenido que tiene y tendrá el "Internet de las cosas" (IoT, sus siglas en inglés), esta enorme cantidad de información que se generará por medio de múltiples dispositivos, es necesario aprovecharlos y utilizarlos para mejorar la sociedad interconectada.

2. Estudio: Big Data 2015, OBS Business School, http://landings.projectmanagement.obs-edu.com/informe-big-data-2015
3. Estudio: Big Data 2015, OBS Business School, http://landings.projectmanagement.obs-edu.com/informe-big-data-2015, pág 29

Intención de implantar una solución de IoT en las organizaciones



Fuente: Estudio: Big Data 2015,OBS Business School, http://landings.projectmanagement.obs-edu.com/informe-big-data-2015, pág, 28.

› *Big data* en Costa Rica, ¿se ha hecho algo?

En Costa Rica, después de más de 60 años de operar en el mercado de las telecomunicaciones con un solo operador estatal en monopolio, en el año 2007 se tomó la decisión de abrir el mercado a la competencia, con el fin de promover el desarrollo y el uso de los servicios de telecomunicaciones dentro del marco de la sociedad de la información y el conocimiento, y como apoyo a sectores como salud, seguridad ciudadana, educación, cultura, comercio y gobierno electrónico, y así procurar que el país obtuviera los máximos beneficios del progreso tecnológico y de la convergencia. A partir de ahí, y de los avances tecnológicos, se ha dado en el país un aumento tanto en la cantidad de proveedores de servicios de telecomunicaciones, contabilizándose al día de hoy la suma de 231, en la variedad de servicios a disposición de los consumidores, así como en el tráfico de datos por medio del Internet en la red móvil, donde en el año 2015 se reportaron 74.933 *terabytes*, superando así a toda la región centroamericana en conjunto.

Por su parte, las diferentes actividades que realiza el ciudadano aprovechando los medios digitales es una realidad, las estadísticas del regulador del sector, la Superintendencia de Telecomunicaciones (SUTEL) en su último informe sobre Estadísticas del Sector Telecomunicaciones 2015, demuestran que las suscripciones en temas de transferencias de datos han venido en aumento, y por ende la dinámica de participación de la sociedad interconectada costarricense es aún mayor y, por lo tanto, la misma exige muchos más servicios aprovechando Internet.

Indicador	2011	2012	2013	2014	2015
	Transferencia de datos				
Suscripciones totales acceso a Internet	2.008.763	3.118.155	4.028.302	4.806.217	5.420.554
Suscripciones totales acceso de Internet fijo-inalámbrico	414.384	439.043	474.433	503.347	545.813
Suscripciones totales acceso a Internet fijo-inalámbrico	5.398	8.904	10.450	12.493	12.843
Suscripciones totales acceso a Internet móvil	1.588.981	2.670.208	3.543.419	4.290.377	4.861.898
Suscripciones totales acceso a Internet fijo/100 habitantes	9%	10%	10%	11%	12%
Suscripciones totales acceso a Internet fijo/ 100 viviendas	32%	34%	36%	37%	39%
Suscripciones totales acceso a Internet móvil/100 habitantes	35%	57%	75%	90%	101%
Suscripciones totales acceso a Internet móvil/ suscripciones totales telefonía móvil	38%	50%	50%	61%	65%
Cantidad total conexiones de líneas dedicadas	10.273	11.993	16.375	16.286	14.093

Este cambio viene acompañado de la forma en que las instituciones en el sector público han modernizado la manera de trabajar, ya que actualmente en su gran mayoría manejan una gran cantidad de información de manera digital, y que en mayor o menor medida ofrecen algún servicio a la ciudadanía en línea.

El Plan Nacional de Desarrollo de las Telecomunicaciones 2015-2021, el cual es el instrumento de política pública que establece el norte de las acciones a seguir en el sector de las telecomunicaciones y que fuera presentado por el Poder Ejecutivo en el mes de octubre de 2015, incluye metas dirigidas a la digitalización

de trámites y servicios brindados por el Gobierno mediante plataformas o terminales tecnológicas interoperables.

Pero ¿qué pasa con la gran cantidad de información que el Estado costarricense está recolectando?, ¿se está aprovechando realmente de manera eficiente y eficaz para lograr mejorar la productividad tanto del Estado como de cara a la ciudadanía?

Existen sectores que han avanzado en la digitalización de los procesos de trámites y servicios ciudadanos, un buen ejemplo de esta automatización es la banca, en la cual muchos de sus trámites se encuentran disponibles en línea aprovechando los datos de sus clientes. Pero el *big data* va más allá y se debe comenzar a plantear una ruta más eficiente del uso de los datos en el Estado, es por esta razón que se torna un gran reto, y sobre todo una gran oportunidad, para que comencemos un desarrollo estratégico del *big data* en el sector público.

Big data y el sector público costarricense: retos y oportunidades

El primer paso en el sector público costarricense para lograr aprovechar el potencial del *big data*, y además establecer una estrategia de cómo deben comunicarse entre sí las diferentes instituciones, ya no solo de manera presencial, sino de manera digital, es resolver la problemática de interoperabilidad. El Decreto nº 35776-PLAN-G-J Promoción del Modelo de Interoperabilidad en el Sector Público de 2010, emitido por el Poder Ejecutivo, establece en su artículo 1:

> *Artículo 1º.* ***Objeto****. El presentze decreto tiene por objeto promover, regular e implementar el modelo de interoperabilidad del Gobierno de la República, para la construcción de un Estado eficiente, transparente y participativo, prestando un mejor servicio a los ciudadanos mediante el aprovechamiento de las Tecnologías de la Información y la Comunicación (TIC).*
>
> *Entendiendo Interoperabilidad como un medio para la construcción de un Estado más eficiente, más transparente y participativo, y que preste mejores servicios a los ciudadanos, todo lo anterior, mediante el mejor aprovechamiento de las Tecnologías de la Información y las Comunicaciones. Entendiendo interoperabilidad como la habilidad de interactuar cooperar y transferir datos de manera uniforme y eficiente entre varias*

organizaciones y sistemas sin importar su origen o proveedor, fijando las normas, las políticas y los estándares necesarios para la consecución de esos objetivos[4].

En otras palabras, se debe lograr que las diferentes bases de datos del país logren comunicarse entre sí, de manera tal que se pueda utilizar, sin importar la estructura de la base que se implementó o su sistema, los datos para convertirlos en información útil. Indiscutiblemente se necesitan los diferentes departamentos de tecnologías de la información con que cuente cada institución para lograr llevar dicho objetivo, pero es claro que la calidad de informáticos que tienen el país, sin duda alguna podrán llevar a cabo este proyecto.

Teniendo comunicadas las bases de datos del sector público, existe otro gran reto: pasar la mayor cantidad, si no es toda, la información que procesa el sector público, de papel a digital, esto implica que se debe iniciar un proceso de digitalización de la documentación y de lograr automatizar y poner en línea los procesos que actualmente se realizan en las instituciones del Estado, el lograr este proceso depende de que se genere un efectivo gobierno electrónico, en la que unas de las metas propuestas es iniciar este proceso de automatización.

De igual forma, se debe pensar en generar una conexión adecuada a la población, tanto de la zona central del país, como del área rural, para aprovechar los beneficios que tendrían los servicios que podrían generarse a raíz del big data para todos los ciudadanos y al mismo tiempo disminuir brechas digitales.

Acompañado en todo este proceso, existe la necesidad de proteger estos datos, el país es consciente de la importancia para los ciudadanos de la protección de sus datos digitales, y sobre la protección especial que deben tener ciertos datos, como los datos de salud, por citar un ejemplo al ser estos datos sensibles, y, por lo tanto, la correcta manipulación de los mismos, así como el respeto a los derechos humanos de privacidad e intimidad, el derecho a la autodeterminación informativa, y la importancia que reviste para la innovación, el desarrollo económico y social y el uso de la información para el comercio electrónico transfronterizo. Es por esto que reviste importancia el marco legal que ya Costa Rica creó, con la Ley nº 8968 Ley de Protección de la Persona frente al Tratamiento de sus Datos Personales y su Reglamento, en el 2011, y por medio del Decreto Ejecutivo nº 37554-JP del 30 de octubre de 2012, el Reglamento a la Ley de Protección de la Persona frente al Tratamiento de sus Datos Personales, a su vez este marco

4. Decreto nº 35776-PLAN-G-J Promoción del Modelo de Interoperabilidad en el Sector Público del 2010.

normativo crea la Agencia de Protección de Datos de los Habitantes (PRODHAB), entidad encargada de hacer cumplir la normativa, tanto a nivel nacional como internacional, para garantizar la información privada de las personas e impulsar el intercambio científico, tecnológico en materia de protección de datos personales, y asegurar que se lleve a cabo una correcta transferencia de datos, respetando el marco normativo.

Es de vital importancia indicar que la normativa prevé el poder utilizar los datos con fines estadísticos, históricos o de investigación científica, cuando no exista riesgo de que las personas sean identificadas; de esta forma se protege a los ciudadanos y se puede utilizar estos datos para mejorar en investigación, salud pública y la administración de los sistemas de salud, por citar algunos ejemplos.

De todos estos retos señalados, existe uno que engloba a todos ellos, y es generar un marco de coordinación efectivo para que las diferentes instituciones del Estado trabajen para un fin común y un proyecto país, bajo este escenario debe existir la figura de un rector en materia de tecnologías que brinde un norte, y el viceministerio de Telecomunicaciones es el llamado a ejercer ese liderazgo.

Ahora bien, surge la pregunta: ¿en qué se puede aprovechar en el país el big data? Realmente las posibilidades son infinitas, pero se pueden trazar algunos sectores donde el impacto puede ser mayor, tanto para el Estado como para la ciudadanía.

En el sector salud, a partir de las metas planteadas por el Plan Nacional de Desarrollo de las Telecomunicaciones, se está desarrollando un enorme proyecto denominado Expediente Digital Único en Salud para los más de cuatro millones de asegurados que tiene la Caja Costarricense de Seguro Social, ya que en papel antes de empezar esta tarea podrían llegar a sumar alrededor de quince millones de historias clínicas. La tarea no es sencilla, ya que implica encontrar la mejor y más rápida forma de transferir dicha información a lo digital, lo cual tendría un costo en tiempo y dinero. Al finalizar dicho proyecto, se estima que se puedan analizar cerca de cinco mil variables como diagnósticos, medicamentos, imágenes médicas, internamientos y operaciones.

Extendiendo esta gran cantidad de información que se tiene de los usuarios para su uso en *big data* y su análisis por medio de la minería de datos, se pueden imaginar algunos usos que en otros países se han desarrollado, por ejemplo, el informe “Big Data in digital Health” de la Fundación Rock Health[5], el cual analizó el

5. Rock Report: Big Data & Healthcare https://rockhealth.com/rock-report-big-data-healthcare/

potencial del big data en el mundo de la salud. Según las conclusiones del informe, puede haber seis vías mediante las cuales big data puede cambiar la atención en materia de salud:

1. Apoyo a la investigación: genómica.
2. Transformar datos en Información.
3. Apoyo al autocuidado de las personas.
4. Apoyo a los proveedores de cuidados médicos.
5. Aumento del conocimiento y concienciación del estado de salud.
6. Agrupamiento de los datos para expandir el ecosistema.

Bajo este escenario, es posible que la información pueda ser utilizada para predecir, prevenir y personalizar enfermedades y con ello brindar a los pacientes atención más personalizada. Otros usos imaginables son que el sector salud podría utilizarlos para determinar con exactitud y en tiempo real dónde se está extendiendo un virus, logrando brindar una mejor respuesta tanto de manera profiláctica como correctiva.

Otro aspecto que el big data puede impactar en el país es en el uso fiscal. Para nadie es un secreto que la evasión fiscal es una constante lucha que tiene el gobierno, pero si a las actuales herramientas contra la lucha de la evasión fiscal se le suma el utilizar todas las bases de datos del sector público, analizarlas, aprovechando la minería de datos, y cruzar información para verificar compras, ingresos, egresos, etc., sería una forma de lograr disminuir este problema.

Este uso va de la mano con otro importante aspecto que se puede atacar con el big data, y es el tema de la transparencia y la lucha contra la corrupción, de manera tal que se puede aprovechar el análisis de volúmenes masivos de datos para la búsqueda de pruebas de forma más rápida, que permita prevenir, identificar e investigar las prácticas fraudulentas, para generar transparencia en todos los ámbitos. Y se puede pensar en un uso mucho mayor, aprovechando la combinación de modelos predictivos y la aplicación de algoritmos (por medio de minería de datos) permitirían analizar conductas e identificar patrones y tendencias de comportamiento que ayudarían a predecir casos de corrupción antes de que se produzcan.

En otro sector en el que se puede aprovechar las ventajas del big data es en el transporte. Con la problemática que se tiene en Costa Rica del aumento desmedido del parque automovilístico, generando las congestiones vehiculares,

principalmente en el área metropoilitana de la ciudad de San José, el big data puede permitir a conocer cambios en los patrones de movimiento en las diferentes vías de la capital, que pueden generar dar una respuesta inmediata a los problemas de transporte, como reducir los tiempos de viaje y rediseñar rutas y buscar su mejor aprovechamiento.

El papel del Viceministerio de Telecomunicaciones y el *big data* en el sector público

Mediante, la Ley n° 8660, Ley de Fortalecimiento y Modernización de las Entidades Públicas del Sector Telecomunicaciones y sus reformas, se crea la Rectoría del Sector Telecomunicaciones como parte del Ministerio de Ciencia, Tecnología y Telecomunicaciones. El artículo 39 de dicha ley indica las funciones que le corresponderá a la Rectoría, y dentro de estas existen funciones que se pueden relacionar y aprovechar para atender la temática de big data[6]:

Artículo 39.- Rectoría del Sector Telecomunicaciones*.

> *El rector del sector será el ministro o la ministra de Ciencia, Tecnología y Telecomunicaciones (Micitt), a quien le corresponderán las siguientes funciones:*
>
> *a) Formular las políticas para el uso y desarrollo de las telecomunicaciones.*
> *e) Dictar el Plan nacional de telecomunicaciones, así como los reglamentos ejecutivos que correspondan.*
> *h) Coordinar las políticas de desarrollo de las telecomunicaciones con otras políticas públicas destinadas a promover la sociedad de la información.*
>
> * Reformado mediante artículo 10 Ley n° 9046. Publicada en el Alcance n° 104 a la Gaceta n° 146 del 30 de julio de 2012.

Este escenario brinda al Viceministerio de Telecomunicaciones un papel fundamental en lograr desarrollar el big data en el país, lo cual es posible traducirlo en tres ejes principales en los que podría darse un aporte a mediano y largo plazo, en generar un desarrollo en la materia a nivel nacional.

6. Ley n° 8660, Ley de Fortalecimiento y Modernización de las Entidades Públicas del Sector Telecomunicaciones y sus reformas.

El primer aspecto que se puede desarrollar es la política pública referente al uso y aprovechamiento de los datos en el sector público costarricense. Esta política debe ser un aliado estratégico, primero para motivar y establecer las reglas en el sector público de su uso y aprovechamiento, y segundo que este uso genere el poder resolver los problemas que se detecten de una manera más eficiente y eficaz.

Dentro de este proceso es necesario pensar y establecer un eje de manejo de datos, un órgano que tenga la capacidad logística de manejar la información que se genere de las diferentes instituciones del país, podemos pensar en un centro de datos nacional, que logre la inmediatez necesaria para el análisis de los datos y permita acelerar la toma de decisiones.

No necesariamente debe ser una nueva institución, puede existir ya y asignársele la nueva función, pero deberá tener la capacidad real para afrontar el reto de ser eficientes y eficaces con los datos que los costarricenses generen, y que además esté siempre regida dentro del marco de derechos humanos, incluyendo dentro de todo su accionar los marcos de protección de datos.

A su vez, el Viceministerio podría establecer el estándar básico técnico que permita asegurar la interoperabilidad de la bases de datos con el Centro de Datos Nacional, con el fin de asegurar que no importe en qué desarrollaron la base datos, pero que se pueda utilizar la información, algo que actualmente es todo un reto a nivel nacional, de esta forma podremos establecer un norte común de uso de los datos en el estado.

› Conclusiones

Costa Rica, como se observa, está en el momento adecuado de lograr generar y aprovechar el big data. Las estadísticas demuestran que existe una gran cantidad de datos que día a día se generan, pero la pregunta fundamental es ¿y con este escenario qué va a hacer el Gobierno? La sociedad se ha convertido en una cibersociedad, donde resulta indispensable que se piense en un verdadero gobierno electrónico, que responda a una modernización del sector público, en que este nuevo esquema tecnosocial permita diseñar políticas públicas que mejoren la atención de problemas en áreas como movilidad, salud y seguridad ciudadana, por citar algunos; pero más aún, si se tiene claro lo que está ocurriendo con las personas, si se sabe lo que están opinando, y si hay certeza de dónde se están generando problemas en la sociedad, es posible que el Gobierno tome

decisiones mucho más acertadas, pero sin olvidar que el eje de mejorar la calidad de vida de las personas en su conjunto como sociedad debe ser el fin primordial que se debe buscar, es por esto que si los datos no se utilizan para mejorar las dificultades sociales, pierde sentido.

En definitiva, el alcance de aprovechamiento que se puede tener de esta gran herramienta implica e impacta a todos los sectores de la sociedad, y Costa Rica tiene todo el potencial para poder aprovecharlo, solamente se debe generar la voluntad política y social para encaminar este gran proyecto.

Referencias bibliográficas

Asamblea Legislativa de Costa Rica (2012). *Ley N° 8660, Ley de Fortalecimiento y Modernización de las Entidades Públicas del Sector Telecomunicaciones y sus reformas.* Obtenido de http: //www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_texto_completo.aspx?param1=NRTC&nValor1=1&nValor2=63786&nValor3=91177&strTipM=TC

Feldman, B., Martin, E. M., Skotnes, T. (2012). *Big Data in HealthcareHype and Hope.* Obtenido de https: //rockhealth.com/rock-report-big-data-healthcare/

OBS Business School (2015). *Big Data aquý y ahora.* Obtenido de http: //landings.projectmanagement.obs-edu.com/informe-big-data-2015

Poder Ejecutivo de Costa Rica (2010). *Decreto Ejecutivo Promoción del Modelo de Interoperabilidad en el Sector Público Nº 35776-PLAN-G-J.* Obtenido de http: //www.pgrweb.go.cr/scij/Busqueda/Normativa/Normas/nrm_norma.aspx?param1=NRM&nValor1=1&nValor2=67348&nValor3=79742&strTipM=FN

Superintendencia de Telecomunicaciones (SUTEL) (2016). *Estadísticas del sector de telecomunicaciones. Costa Rica. 2015.* Obtenido de https: //sutel.go.cr/informes-indicadores

Capítulo 17

La magia de trabajar con datos y gobiernos: mi experiencia en la ciudad de Los Ángeles

Juan Vasquez*

La tecnología no es magia. No hay sistema, base de datos o plataforma que solo o sola pueda cambiar una burocracia, gobierno o campaña. Igualmente, no hay ser humano que solo o sola puedan. La clave está en la combinación entre tecnología y humano. Carne y hueso más código y archivos. APIs y DNA. Cuando se mezcla el talento humano ideal con las herramientas correctas, ahí sí encontramos magia.

De eso se trata esta parte del manual y conversación colectiva que estamos teniendo, de los aspectos humanos necesarios al trabajar con *big data*, y las oportunidades y barreras que esa misma humanidad nos ofrece en el proceso.

Los vamos a hacer en dos partes:

1. Cómo trabajar con burócratas que no se sienten cómodos con *big data* ni nuevas tecnologías, y que pueden causar que tus proyectos fracasen.
2. El potencial e impacto que nos ofrece la colaboración entre sectores, principalmente cuando un gobierno local colabora con universidades y compañías tecnológicas.

Es importante tener esta conversación porque las dinámicas humanas y personales pueden impactar el éxito o fracaso asociado con un proyecto. Una persona, especialmente alguien que no sabe casi nada sobre la tecnología o que le tiene miedo a ella, puede ser más dañina para un proyecto que un ataque cibernético o cualquier virus.

Al terminar de cubrir estos tres temas espero que se sientan más cómodos manejando los lados humanos de *big data* y que puedan ser más eficaces en la implementación de nuevos sistemas e infraestructuras.

* Analista de datos en el Equipo de Innovación Operativa de la Alcaldía de Los Ángeles.

Pero antes de empezar, les cuento un poco sobre mi trabajo con el Equipo de Innovación Operativa en la Oficina del Alcalde de Los Ángeles, Eric Garcetti (@MayorOfLA). Nuestro apodo es el "O-Team," y tenemos tres iniciativas principales:

1. Modernizar los procesos de compra y contratación de la ciudad para impactar la economía local y crear acceso para negocios de minorías y mujeres.
2. Crear el portafolio de finca raíces de Los Ángeles y reformar la infraestructura de gestión.
3. Hacer más operativos y ágiles los procesos dedicados a la salud y el bienestar de los empleados de LA para reducir la cantidad de solicitudes de indemnización laboral

No somos como cualquier otro equipo gubernamental. El "O-Team" sale de una colaboración entre la Oficina del Alcalde, una organización sin fines de lucro llamada el Mayor's Fund, y una organización estilo *think tank* llamada Los Angeles Coalition for the Economy and Jobs. Esto es importante porque nos ofrece un nuevo modelo sobre cómo podemos reformar, optimizar y modernizar un gobierno. En EE. UU. estas colaboraciones están creciendo en popularidad, se llaman *public-private partnerships* y suelen a ser referidas como "p3's".

Mi equipo trabaja directamente dentro de la Alcaldía. Mi correo electrónico es de la Alcaldía, mi tarjeta tiene el sello del alcalde y dice que trabajo para el alcalde Eric Garcetti. Pero, mi salario es pagado por el Mayor's Fund, y el capital para pagar por mi equipo y nuestros proyectos fue donado por la Coalition. Esto en sí es una colaboración multisectorial.

El Mayor's Fund está directamente conectada a la Oficina del Alcalde pero tiene su propio cuerpo regulatorio y administrativo, y su meta es facilitar proyectos pilotos que tienen impacto cívico. Como el Fund es la entidad que me paga, tengo más libertad que si fuera un empleado de la ciudad. Por ende, puedo contratar los servicios y comprar las herramientas que necesito rápidamente, entre una semana y dos meses. Si fuera un empleado de la ciudad, esos mismos procesos nos llevarían entre siete meses y año y medio.

Llegando ya al final de nuestros proyectos somos un equipo de seis personas —una directora, una mánager de proyectos, un mánager de configuración dedicado exclusivamente a la configuración de sistemas, yo como estratega de datos

y comunicación, una analista y dos coordinadores enfocados en proyectos especiales y apoyo general—. Nuestro equipo se reporta al teniente de alcalde de la Oficina de Presupuesto e Innovación. Esto es importante porque crea una línea directa entre mi equipo y el director del Presupuesto de la ciudad, así podemos apoyar y empujar el financiamiento de proyectos específicos. Esta estructura —presupuesto junto a innovación— demuestra el nivel de compromiso que la ciudad y su gobierno local tienen con el uso de tecnología moderno y que ofrece impacto.

Como estratega de datos y comunicación, uso narración visual, análisis de información y una infraestructura de herramientas digital para reformar el gobierno de Los Ángeles, la segunda ciudad más grande en EE. UU.

Para reformar un gobierno, se necesita el apoyo de varios grupos. En mi caso, burócratas, corporaciones, pequeños negocios, organizaciones sin fines de lucro, organizaciones comunitarias y activistas. Sin mi experiencia como publicista y mánager de redes sociales, no podría ser exitoso en este trabajo.

Bueno, a lo que vinimos.

O me siguen, o me siguen: luchando con la burocracia

Hablemos sobre cómo trabajar con burócratas que no se sienten cómodos con *big data* ni nuevas tecnologías, y que pueden en ciertos casos, matar o paralizar tus proyectos. Pero, definamos el término "burócrata." Lo veo de esta forma:

1. En momento de tomar decisiones difíciles, suele a enfocarse en proceso y no personas. No me malentiendan, es importante considerar proceso y percepción externa. Por otro lado, si una organización quiere tener empleados que ofrezcan sus mejores ideas y que sean parte del crecimiento de la organización, importa demasiado crear un ambiente donde trabajadores se sientan apoyados y donde sean apreciados por sus habilidades y visiones.
2. Atrincherado en procesos muy específicos donde suelen crear barreras sin ofrecer valor. Son los que ante los cambios responden con un "aquí siempre se han hecho así". Debemos entender que las organizaciones realmente innovadoras están optimizando sus procesos y evolucionando constantemente.

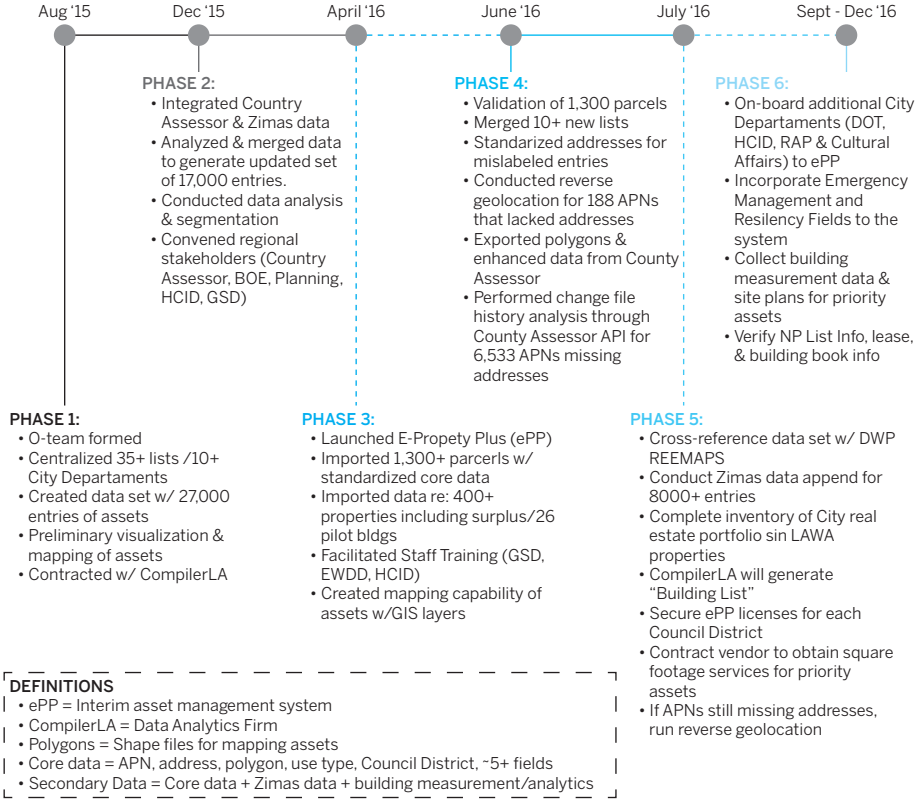
También es importante aclarar que no todas las personas que trabajan para el gobierno son burócratas. En al ayuntamiento de Los Ángeles tenemos rebeldes, innovadores, creadores, líderes, visionarios y de más. Igualmente, hay burócratas en todas las industrias —por ejemplo, los vemos en corporaciones, universidades, hospitales y cuerpos regulatorios—.

Trabajando con la Alcaldía me he estrellado directamente contra la pared burócrata varias veces. Les cuento las excusas o razones que han solido usar:

1. Pedir reportes y respuestas constantemente, específicamente sobre temas que realmente no entienden.
2. Detener un contrato antes de ser ejecutado, teniéndolo bajo "evaluación" por meses sin una justificación real.
3. Constantemente pedir que una decisión se aplace por su falta de entendimiento o confort con el tema.

Ciertamente existen muchas razones que dependen en gran medida de la capacidad de comunicar e informar a todos los involucrados en cada uno de los proyectos. Debemos entender que es parte de nuestra responsabilidad, mitigar o reducir los miedos de los burócratas ante los proyectos innovadores que podamos plantear, en definitiva, darles certidumbre donde ellos no la perciben.

Os voy a poner algunos ejemplos de cómo lo intentamos hacer en la ciudad de Los Ángeles. Por ejemplo, en este diagrama pueden observar el proceso que mi equipo manejó para crear la base de datos de fincas raíces de Los Ángeles. Representamos seis fases de trabajo usando pistas visuales como colores, figuras, y contenido presentado en una dirección progresiva.



La idea era simplificar un proceso bastante complicado. En vez de crear un reporte escrito en Times New Roman tamaño 12, color negro sobre blanco, de 10 páginas, decidí crear un diagrama que cabe en una página y es seriamente diferente y más eficaz a como un gobierno usualmente se comunica.

Empezamos en agosto 2015 y nos llevó hasta diciembre 2016. Unimos más de 50 diferentes conjuntos de datos de más de 15 fuentes, vimos problemas con identificadores únicos, y usamos APIs, GIS, *shapefiles* y portales de información abierta. Trabajé cercanamente con un *startup* llamada Compiler LA, la cual fue creada por dos programadores increíbles, Vyki Englert (@vyki_e) y Stephen Corwin (@Stephen_Corwin).

Vyki, Stephen y yo nos conocimos originalmente cuando todos trabajamos para NationBuilder, una compañía de tecnología dedicada a crear herramientas para organizar comunidades alrededor de campañas políticas, gobiernos y movimientos cívicos. Para ustedes que están armando esta clase de equipo en sus

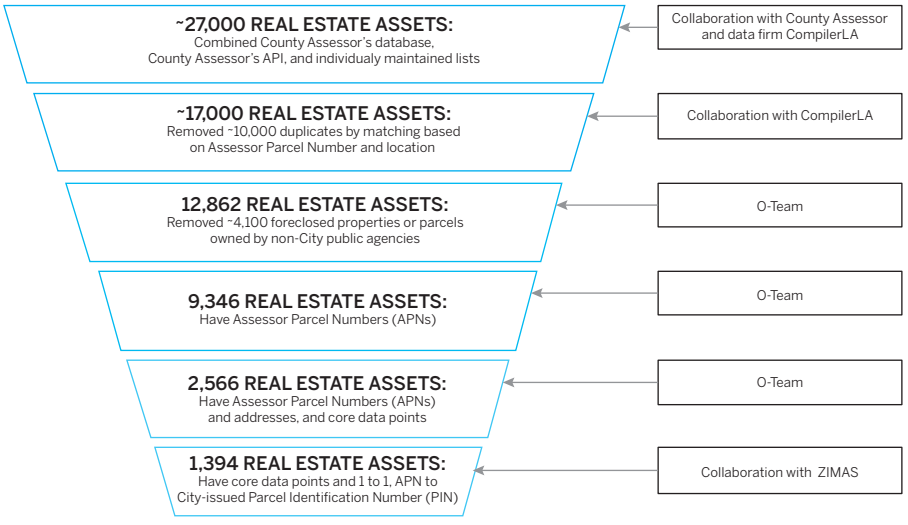
ayuntamientos, les recomiendo buscar candidatos que vienen con relaciones preexistentes con la comunidad de *hackers* cívicos locales. Los servicios de Compiler LA fueron claves para el éxito de nuestro proyecto en finca raíz.

Una vez terminamos la centralización de información, la configuramos en un sistema llamado ePropertyPlus, una herramienta que ayuda con la asistencia y el manejo de datos vinculados a fincas raíces.

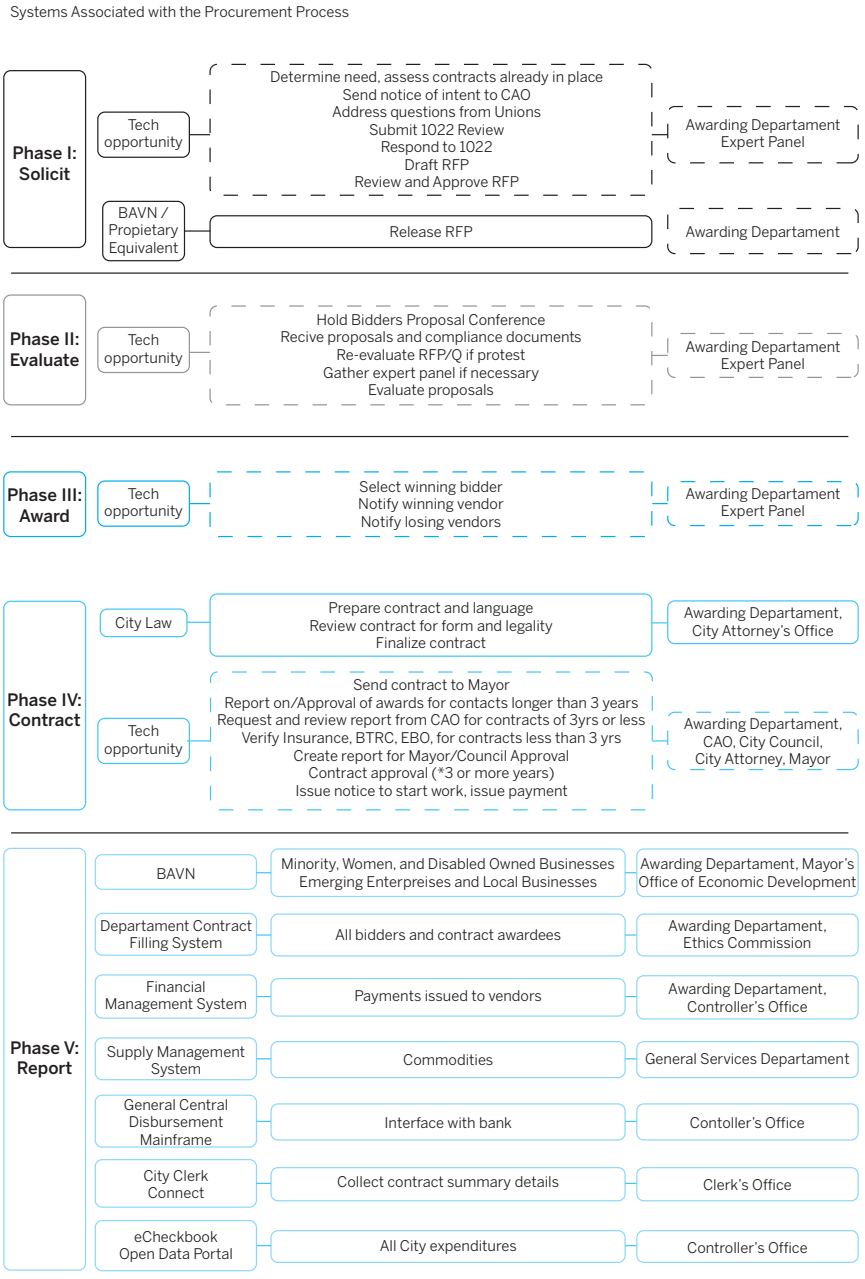
En este segundo ejemplo usamos la misma táctica para explicar cómo centralizamos los diferentes conjuntos de datos. Esta imagen la creé en diciembre de 2015, y en ese entonces teníamos mas de 27.000 récords individuales, los cuales reducimos a menos de 17.000 al eliminar duplicados, usando identificadores únicos y mapeando polígonos y puntos (direcciones). Si una dirección caía exactamente en el centro de un polígono, lo tratábamos como duplicado.

Después de una serie de análisis llegamos a un conjunto de datos piloto de 1.300 récords. Estos tenían la mayor cantidad de información, y la información más limpia y confirmada. Este grupo representaba más o menos 8% del portafolio total.

Le recomendaría a cualquier grupo haciendo una implementación de sistema que aprenda a usarlo y configurarlo con un subgrupo de información, y no con el conjunto en su totalidad.



El último ejemplo donde usamos narración visual para educar a burócratas y contextualizar temas complicados se enfoca en el proceso necesario para contratar con la ciudad de Los Ángeles.



En esta imagen vemos todos los pasos y sistemas necesarios para que la ciudad de Los Ángeles solicite un servicio, evalúe respuestas, escoja un vendedor, ejecute el contrato y reporte al final.

Las líneas continuas representan etapas que usan tecnología para ser ejecutadas. Las líneas discontinuas representan huecos donde el proceso está basado en papel. Como el proceso no es digital, la ciudad no puede recolectar datos. Sin datos, no hay métrica de rendimiento. Sin métrica de rendimiento, no hay una dirección clara.

Noten algo interesante, todas las líneas continuas están al final del proceso. Quiero decir, hay una falta de tecnología y métrica de rendimiento al principio del proceso. Esto es grave e impacta directamente en el conocimiento que puede ser extraído de estos procesos.

En este momento mi equipo está trabajando para alimentar esos huecos, específicamente en el primer paso del proceso —en la creación del documento que presenta la oportunidad, un RFP/RFQ o Request for Proposals/Request for Qualifications—.

Estas anécdotas sencillas que acabo de describir han servido y nos están sirviendo como ejemplos para demostrar cómo el uso de la narración visual a la hora de simplificar los procesos de comunicación, ayudan a mitigar la resistencia al cambio que suele existir en ciertos ámbitos de nuestras organizaciones. Cambiemos de tema.

En equipo se puede más: colaboración multisectorial

No cabe duda de que las colaboraciones entre sectores ofrecen oportunidades para crear impacto exponencial. En mi primer año y medio trabajando en la Alcaldía, mi equipo coordinó más de 15 colaboraciones con universidades, comunidades de *hacking* cívico, *startups*, organizaciones sin fines de lucro y corporaciones.

Para ejecutar una colaboración exitosa entre un gobierno y una universidad, recomiendo usar un modelo con X partes. Nosotros operamos así:

1. **Una meta específica**, con parámetros fijos y que se pueda completar en menos de 6 meses. *Ejemplos:*

- Creación de 3 perfiles representando las mesclas de variables que suelen a crear una reclamación de compensación de trabajadores costosa (Cornell University Applied Statistics Master's Program).
- Análisis de la cantidad de tiempo que un bombero va a estar indisponible basado en la parte del cuerpo que se lastimo, y la causa asociada con la reclamación de compensación de trabajadores (California State University, Los Ángeles —Clase de Big Data).
- Digitación y automatización del proceso de certificación de negocios con mujeres propietarias (California State University, Los Ángeles —Clase de Big Data).

2. **Gerente de proyecto** que mantiene el horario de implementación, facilita comunicación y coordinación, y se asegura de que todos estén cumpliendo con sus partes. En nuestras colaboraciones, mi equipo tiene esta responsabilidad.
3. **Cliente** donde se genera o se ve la meta especifica (1). *Ejemplos:*

   - El Departamento de Personal es el cliente en el proyecto con Cornell, y para el proyecto con el cuerpo de bomberos con California State University, Los Ángeles.
   - Para el proyecto de digitación y automatización, los clientes son la Agencia de Informática y la Oficina de Contratación.

4. **Equipo universitario** bajo un centro o programa de innovación o disciplina simularía o con un profesor con una clase en áreas de tecnología. *Ejemplos:*

   - Lloyd Greif Center for Entrepreneurial Studies, University of Southern California Marshall School of Business.
   - Global Center for Entrepreneurship and Innovation, California State, Los Angeles, College of Business and Economics.
   - Pepperdine Graziadio School of Business & Management, Pepperdine University.

5. **Entregable** como un archivo o formato especifico que pueda ser integrado fácilmente a la infraestructura de la ciudad. *Ejemplos:*

   - Tableau workbook.
   - Shapefiles.
   - Mapa en ArcGIS.

6. **Claridad** sobre el impacto del proyecto y el uso del entregable. *Ejemplos:*

   - Los 3 perfiles generados por Cornell para informar los procesos de la Unidad de Reclamación de Compensación de Trabajadores y la configuración del sistema de manejo de riesgo de la ciudad.
   - El análisis con el cuerpo de bomberos será usado por la Unidad de Metricas de Rendimiento y la Unidad de Manejo de Riesgo.
   - El proceso digital de certificación para negocios con mujeres propietarias ayuda a la ciudad a llegar a las métricas de rendimiento asociadas con una nueva póliza dedicada a la equidad de género.

7. **Herramienta** digital y accesible para manejo de proyectos. Mi preferencia es Trello, y suelo a dividir el proyecto en categorías como "Por Hacer," "Haciendo," "Hecho," "Datasets" y "Compañeros" y "Documentos."
8. **Alcance de proyecto** con todos los detalles y aspectos mencionados. Es muy importante tener este documento en la herramienta central.
9. **Acuerdo de no divulgación** si se está manejando información personal y privada. Recomiendo no usarlo si no es necesario. Igual que con el alcance, este documento debe vivir en la herramienta central.
10. **Reunión de lanzamiento de proyecto** para que todos los involucrados se conozcan y para responder cualquier pregunta que falte.
11. **Correos electrónicos** cada dos semanas o cuando haya preguntas. Usualmente entre el equipo universitario y el gerente de proyecto.
12. **Reunión de punto medio.**
13. **Presentación final.**
14. **Certificados de aprecio** de parte del ayuntamiento para la entidad académica, y si es posible para cada miembro del equipo universitario.

En mi experiencia, el tiempo requerido para coordinar todo antes de lanzar el proyecto depende mucho del cliente y de la presión que el gerente de proyecto puede aplicar cuando las cosas se estancan. Mi equipo ha planeado y lanzado colaboraciones en una semana para un proyecto de un mes.

La colaboración entre el cuerpo de bomberos y California State University, Los Ángeles, fue de seis meses para coordinar y finalizar todos los documentos. El proceso para alinear todo entre el Departamento de Personal y Cornell duró dos meses.

Los dejó con deseos de mucho éxito, un mundo de energía y un toque de perspectiva sobre la importancia de tener innovadores, rebeldes, tecnólogos y demás involucrados en sus gobiernos locales.

La labor de modernizar el gobierno y reformar sus procesos no es fácil. A veces duele. Puede ser frustrante. Requiere paciencia y diplomacia. Una muy buena cantidad de estrategia y flexibilidad.

Las victorias no son muchas y no vienen muy a menudo. Pero cuando llegan, son increíbles. Tienen impacto real. Un impacto visto en las calles donde familias exploran y aprenden. Espacios y estructuras donde individuos transforman ideas en metas, y metas en realidad.

Capítulo 18

Estimación de la pobreza utilizando datos de teléfonos celulares: evidencia de Guatemala

Marco Antonio Hernández Ore, Lingzi Hong, Vanessa Frías-Martínez, Andrew Whitby y Enrique Frías-Martínez

Extracto

La dramática expansión del uso de teléfonos móviles en países en desarrollo ha resultado en un incremento de fuentes ricas y mayormente intactas de información sobre las características de las comunidades y regiones. Los Registros de Detalles de Llamadas (CDR) obtenidos de los teléfonos celulares proveen una información altamente granular en tiempo real que puede ser usada para evaluar el comportamiento socioeconómico incluyendo consumo, movilidad y patrones sociales. Esta nota examina los resultados de un análisis CDR enfocado en cinco departamentos administrativos en la región suroeste de Guatemala, el cual usó datos de teléfonos celulares para predecir los índices de pobreza observados. Sus descubrimientos indican que los métodos de investigación basados en CDR tienen el potencial para replicar las estimaciones de pobreza obtenidos de las formas tradicionales de recolección de datos, como encuestas en hogares o censos, por una fracción del costo. En particular, los CDR fueron más útiles en predecir la pobreza urbana y total en Guatemala con más precisión que la pobreza rural. Además, mientras que los estimados de pobreza producidos por los análisis CDR no encajan perfectamente en aquellos generados por encuestas y censos, los resultados muestran que obtener más información exhaustiva podría mejorar enormemente su poder predictivo. El análisis CDR tiene especialmente aplicaciones prometedoras en Guatemala y otros países en desarrollo, lo cuales sufren altos índices de pobreza e inequidad, y donde los limitados recursos presupuestarios y fiscales complicarían la tarea de recolección de datos. Además, destacan la importancia de focalizar con precisión los gastos públicos para lograr su máximo impacto antipobreza.

Introducción

El explosivo crecimiento de las redes de telecomunicaciones en los países en desarrollo está arrojando una riqueza sin precedentes de datos altamente

granular en tiempo real, y los gobiernos solo han comenzado a aprovechar el enorme potencial de esta nueva fuente de información. Esta nota explora las metodologías analíticas para usar datos de teléfonos celulares agregados y cifrados para mapear la distribución de la pobreza en Guatemala. Para estimar los índices de pobreza, Guatemala, como muchos otros países en desarrollo, depende de la conducción y análisis de encuestas en hogares y censos de población que son tanto costosos como exigentes administrativamente. Por contraste, el análisis basado en datos de teléfonos celulares en combinación con el aprendizaje máquina tiene el potencial de generar información confiable y oportuna sobre la distribución espacial de la pobreza en hogares a un costo mucho menor que las tradicionales encuestas de hogares o censos de población.

› Pobreza en Guatemala

Los índices de pobreza en Guatemala son mayores que en otros países de ingresos medios comparables, y la distribución de la pobreza refleja una serie de dimensiones regionales, rural/urbana y étnicas superpuestas. Contrario a la tendencia observada en otros países de Latinoamérica, los índices de pobreza en Guatemala han crecido en años recientes. Los índices de pobreza[1] crecieron desde el 55% en 2001 al 60% en 2014, mientras que la cantidad de personas que viven por debajo de la línea de pobreza se incrementó en unos 2,8 millones. Los índices de pobreza varían dramáticamente por departamento administrativo. En 2014, el departamento más pobre de Guatemala, Alta Verapaz, tuvo un índice de pobreza del 83% y un índice de pobreza extrema[2] del 54%. Mientras tanto, los índices de pobreza y pobreza extrema en el departamento más rico, donde se encuentra localizada la Ciudad de Guatemala, fueron mucho más bajos ubicados en el 33% y 5%, respectivamente. Además, mientras que las áreas urbanas son ahora el hogar de una mayoría de la pobreza del país, los índices de pobreza se mantienen sustancialmente mayores en áreas rurales. En 2014, el 35% de la población rural estaba viviendo en pobreza extrema, comparado con el 11% de la población urbana. Los índices de pobreza son también significativamente mayores entre la población indígena de Guatemala. La gente indígena representa el 42% de la población total del país, pero en 2014 representaba el 52% de los pobres y el 66% de los extremadamente pobres en el país[3].

1. Los índices de pobreza moderada reflejan el consumo familiar per cápita equivalente a $4.00 al día en términos de poder adquisitivo.
2. Los índices de pobreza extrema reflejan el consumo familiar per cápita equivalente a $2.50 al día en términos de poder adquisitivo.
3. Sanchez, Scott y Lopez (2016).

Los altos índices de pobreza en Guatemala y la persistente inequidad en ingresos son reflejados en los débiles indicadores del desarrollo humano en el país. El acceso limitado y desigual a los servicios públicos como educación y el cuidado de la salud han restringido la formación del capital humano. Mientras que una falta de infraestructura básica ha incrementado los costos de producción y transporte y ha reducido las oportunidades de trabajo disponibles para los pobres. En conjunto, estos factores están contribuyendo al declive a largo plazo de la productividad económica. Mientras tanto, los bajos ingresos de impuestos en Guatemala limitan su capacidad para una política de redistribución fiscal y gastos en pro de los pobres. Los altos índices de pobreza y las fuertes restricciones fiscales de Guatemala destacan la crítica importancia de focalizar efectivamente el gasto público.

El rol de la información en la reducción de la pobreza

Las estrategias efectivas para reducir la pobreza requieren información detallada sobre la actual distribución geográfica de la pobreza y las características de los hogares que viven debajo de la línea de pobreza. Los censos y las encuestas de hogares pueden dar luz a un amplio rango de indicadores económicos y sociales. Sin embargo, estos métodos son costosos y consumen tiempo e implementarlos requiere de capacidad institucional. Además, las condiciones locales adversas como conflictos violentos, altos índices de crímenes o inestabilidad política puede hacer que las encuestas personales sean imposibles en ciertas áreas. Como resultado, los políticos responsables deben con frecuencia basar sus decisiones críticas en información incompleta u obsoleta.

Los científicos sociales están usando crecientemente el análisis de *big data* para suplementar más fuentes tradicionales de información. Las imágenes de satélites, registros de sensores (es decir, tráfico, clima), aplicaciones de teléfonos inteligentes e información de teléfonos celulares —el tema de esta nota— ya han arrojado perspectivas importantes en numerosos campos. A diferencia de las encuestas en hogares, las cuales están específicamente diseñadas para abordar ciertas preguntas de investigación, los grandes conjuntos de datos son usualmente recolectados en un contexto no investigativo, usualmente como el derivado de una actividad comercial o servicio público. Analizar *big data* requiere de nuevos métodos de investigación, muchos de los cuales están todavía en etapas iniciales de su desarrollo. Las metodologías de investigación emergentes basadas en Registros de Detalles de Llamadas (CDR) y técnicas avanzadas de aprendizaje máquina o *machine learning* tienen aplicaciones

especialmente prometedoras en países en desarrollo, ya que ellas pueden potencialmente generar datos confiables de pobreza a un costo mucho menor que las encuestas de hogares convencionales.

El análisis CDR puede jugar un rol vital al llenar los huecos espaciales y temporales dejados por los métodos tradicionales de investigación. Al hacer inferencias basadas en el uso de redes celulares, el análisis CDR puede proyectar confiablemente la evolución de las dinámicas de pobreza en un marco de tiempo específico. A diferencia de los censos y las encuestas de hogares, el análisis CDR es rápido y relativamente económico y puede ser realizado por un grupo pequeño de estadísticos usando registros que ya fueron recolectados por Operadoras de Redes Móviles (MNO).

Guatemala ofrece un claro ejemplo de los límites de la recolección tradicional de datos. El Censo de Hogares y Población más reciente tiene fecha de 2002 y todos los datos de pobreza nacional están derivados de solo 4 encuestas de hogares realizadas en los últimos 25 años. Por ejemplo, la más reciente encuesta de hogares de 2014 (Encuesta Nacional de Condiciones de Vida, ENCOVI) cubre alrededor de 11.500 hogares, se tardó dos años en completarla a un costo de alrededor de 2 millones de dólares. Por contraste, el análisis CDR realizado para este reporte tuvo un valor de alrededor de 100.000 dólares, y la mayoría de los gastos fue para al desarrollo del algoritmo de computadora, el cual es un costo fijo. Esto, si el análisis CDR es conducido nuevamente con nuevos datos, sería significativamente más económico.

Mientras que este fue el primer análisis de su tipo realizado en Guatemala y diseñado para probar principalmente la validez de varias metodologías, un ejercicio más minucioso requeriría solo de una pequeña fracción del tiempo y recursos humanos relativos a un censo o encuesta tradicional. Además, estos costos serían probablemente menor en iteraciones posteriores del análisis CDR, mientras la innovación y las pruebas tecnológicas son reemplazadas por la implementación rutinaria de técnicas establecidas. Mientras que el análisis CDR no puede reemplazar totalmente los métodos convencionales de investigación, pueden mejorar enormemente su valor al proveer actualizaciones de alta frecuencia e información complementaria. Además, si el análisis CDR puede demostrarse que puede proveer inferencias suficientemente precisas para permitir a los países extender ligeramente el tiempo entre las encuestas tradicionales, podría potencialmente generar un ahorro neto para el presupuesto nacional de investigación.

Antecedentes

Registro de detalles de llamada (CDR)

Las operadoras de redes móviles registran y almacenan datos sobre el uso de los teléfonos de sus clientes, primariamente para propósitos de cobro. Adicionalmente al registrar el consumo de datos celulares, las MNO recolectan información de cada llamada y mensaje. Los datos almacenados no reflejan generalmente el contenido de una llamada o mensaje. En vez de eso, se registran detalles circunstanciales, como hora y duración de la llamada, el tamaño del mensaje, las identidades de las partes involucradas y su información de la red. Se refieren a estos datos en la industria de telecomunicaciones como la CDR.

Figura 1. Ejemplo de registros de detalles de llamadas

Interacción	Dirección	ID Correspondiente	Hora y fecha	Duración de la llamada	ID de la antena
Llamada	Entrada	8f8ad28de134	2012-05-20 20:30:37	137	13084
Llamada	Salida	fe01d67aeccd	2012-05-20 20:31:42	542	13084
Texto	Entrada	c8f538f1ccb2	2012-05-20 21:10:31		13087

Fuente: http://bandicoot.mit.edu/docs/quickstart.html

Adicionalmente a las CDR, las MNO usualmente almacenan ciertos detalles personales sobre sus usuarios, incluyendo sus nombres y dirección de hogar, y en algunos casos su género, edad u otras características. Para clientes de prepago, los cuales son muy comunes en países de ingresos bajos y medios, las MNO típicamente guardan un registro de las recargas de créditos o "incremento de saldo".

Usar CDR en investigaciones sociales y económicas

Aunque las CDR pueden aparentar una serie de datos técnicos y estrechos debido a la dramática expansión del uso del teléfono móvil en las últimas décadas, estos registros pueden proveer una rica fuente de información sobre el comportamiento humano y las características de comunidades. Las CDR pueden ser usadas para inferir ciertos atributos personales sobre un usuario de celular, tales como la ubicación de su hogar. De allí, ellos pueden ser usados para analizar las redes sociales, ya que cada llamada puede ser vista como un vínculo entre los

clientes de un MNO. Este enfoque permite a los investigadores trazar las interacciones sociales, identificar puntos de nexo de las comunidades y examinar cómo se transmite la información a través de grupos y regiones[4].

Caja 1. La geografía virtual de los teléfonos celulares: determinar la ubicación de los usuarios desde los CDR

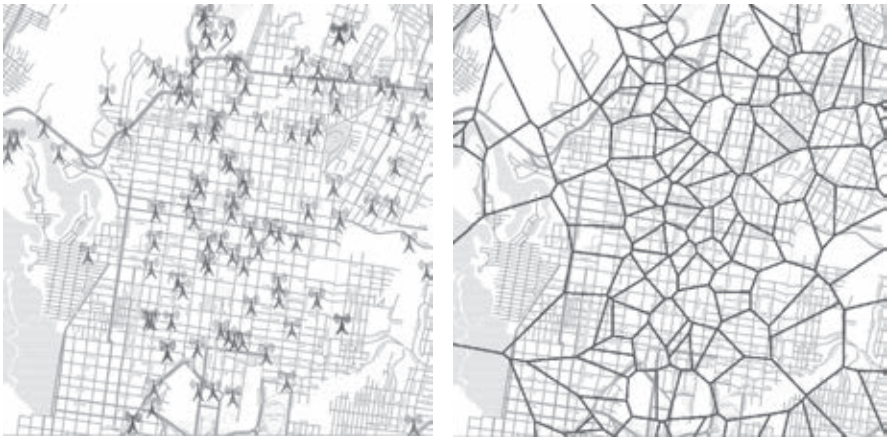
A diferencia de los celulares de satélite, los teléfonos celulares dependen de una red de torres conocidas como estaciones base, las cuales operan dentro de un rango limitado. Esto divide el área de cobertura de la red en "células" individuales. Mientras el cliente se mueve, su conexión a la red es transferida de una torre a la siguiente. Cuando un cliente realiza una llamada, envía un texto o inicia una sesión de datos, la identidad de la torre relevante es registrada en el CDR. Las MNO mantienen una lista de las coordenadas de cada torre, haciendo posible determinar la ubicación general de un teléfono cada vez que es usado.

Mientras que la ubicación precisa del usuario no puede ser identificada, áreas geográficas limitadas llamadas "polígonos de Voronoi" pueden ser construidas basadas en la asunción de que un móvil siempre se conecta a la torre celular más cercana. En realidad, factores como la diferente potencia de las antenas, restricciones de capacidad y terreno puede causar que un móvil se conecte a una antena más lejana. Sin embargo, los polígonos Voronoi continúan siendo una herramienta útil para aproximar la ubicación de un usuario de celular.

La estructura de la red determina el tamaño de cada polígono Voronoi. Ya que tanto el equipo físico de la antena y el espectro móvil tienen capacidad limitada, las MNO tienden a colocar más antenas en áreas de mayor uso para maximizar el rendimiento. Por ende, el tamaño de los polígonos tiende a correlacionarse inversamente con la densidad de población –es decir, los lugares densamente poblados tienden a tener más antenas y polígonos más pequeños–. Sin embargo, algunos lugares con una población residencial pequeña, pero con altos índices de actividad comercial, tales como distritos de negocios, centros comerciales y aeropuertos, pueden tener un número mayor de antenas. Los polígonos pueden variar en diámetro desde unos cientos metros a decenas de kilómetros, dependiendo de la red.

4. Estas aplicaciones son descritas con mayor detalle en Blondel *et al*. (2015).

Figura 2. Muestra de Las torres celulares alrededor de la Plaza Constitución de la Ciudad de Guatemala (izquierda) y los polígonos Voronoi que generan (derecha)



Nota: el Proyecto OpenCellID recolecta ubicaciones de redes celulares basado en los reportes de usuarios voluntarios quienes instalaron una aplicación de participación. Por lo tanto, estas ubicaciones de las torres son aproximadas. Ninguna información oficial de torres celulares ha sido usada para estos gráficos.
Fuente: opencellid.org

Estimar la pobreza en Guatemala usando datos de teléfonos celulares

Esta sección describe el resultado de un estudio reciente de métodos de investigación basados en los CDR en Guatemala que fueron diseñados para evaluar el valor potencial del análisis CDR como una herramienta de investigación socioeconómica. El objetivo del estudio era crear un modelo usando datos de CDR que pudiera predecir con precisión la incidencia observada de pobreza extrema. El estudio se enfocó en cinco municipios en los departamentos administrativos de Quetzaltenango, Suchitepéquez, Sololá, Totonicapán y San Marcos. Juntos, estos departamentos representan el 20% de la población guatemalteca.

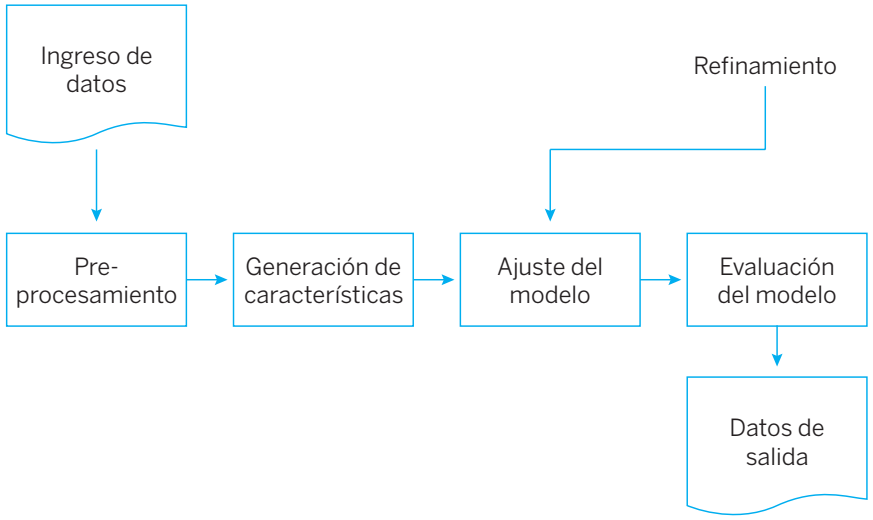
El estudio abordaba tres preguntas:

- ¿Pueden los datos de CDR ser usados para estimar confiablemente los índices de pobreza en Guatemala?

- ¿Son estos estimados más precisos en áreas urbanas, áreas rurales o a nivel nacional?
- ¿Pueden los datos de pobreza derivados de los CDR en el 2006 ser usados para predecir los índices de pobreza en el 2011?

Para responder a estas preguntas, el estudio empleó un enfoque de aprendizaje máquina, el cual es un método altamente general, iterativo para descubrir la relación entre datos de entrada y datos de salida, los cuales en este caso son los registros de teléfonos celulares y los índices de pobreza, respectivamente. La figura número 3 ilustra la metodología de aprendizaje de máquinas.

Figura 3. Pasos típicos en un análisis de aprendizaje de máquinas



Fuente: adaptado de Hong y Frías-Martínez [2015a].

Fuentes de datos

Para poder probar la validez del análisis CDR, sus descubrimientos fueron comparados con los estimados de pobreza del Banco Mundial, los cuales están basados en la Encuesta Nacional de Condiciones de Vida de Guatemala (ENCOVI) para 2006 y 2011 y el Censo de Hogares y Población de 2002. El ENCOVI es una encuesta de hogares tradicionales y su tamaño de muestra no provee estimados confiables a nivel municipal. Sin embargo, las técnicas de estimación de áreas pequeñas[5] que combinan los datos de ENCOVI con los datos del censo permiten

5. Ver Elbers, Lanjouw y Lanjouw (2003).

que la pobreza sea estimada a nivel municipal. Para los propósitos del estudio, estos estimados fueron tratados como "datos reales del terreno". En un análisis de aprendizaje de máquinas, los datos reales del terreno son obtenidos por observación directa, en vez de por un modelaje o inferencia. En este contexto, sin embargo, el término se refiere a los índices de pobreza determinados por métodos de estimación estadísticos estándar, los cuales proveen la única medida existente de la realidad del terreno. Sin embargo, debe tenerse en cuenta que todas las metodologías de estimación de índice de pobreza son predicados en asunciones, y en esta área ningún dato del terreno puede ofrecer una representación perfecta de la realidad.

Los modelos supervisados de aprendizaje máquina, como se describen aquí, requieren un serie de datos de entrenamiento, los cuales comprenden datos de referencia que representan la realidad del terreno. Ya que los CDR se relacionan directamente con las personas, pueden ser considerados datos de registro de unidad. Entonces es el detalle de los datos de la realidad del terreno los que determinan enormemente la resolución del modelo. Esta es usualmente la única opción disponible cuando los datos de la realidad del terreno están basados en estimaciones usando datos de encuestas de hogares, en los cuales ningún número de teléfono celular es recolectado durante la encuesta. En este caso, las características de nivel individual son extraídas de los CDR y luego combinadas para formar agregados estadísticos en el nivel geográfico elegido (es decir, medio, mediano, máximo o cuantiles por región). Un tamaño de muestra relativamente pequeño (por ejemplo, los 338 municipios de Guatemala) significa que la validación interna, como la validación en cruce o pruebas ocultas de datos, es difícil, por lo que pueden ser requeridas algunas validaciones externas.

Los índices de pobreza fueron calculados para cada municipio. Los índices agregados de pobreza rural, urbana y general estuvieron disponibles para 2006, pero solo los índices de pobreza rural estuvieron disponibles para 2011[6]. El estudio uso datos CDR agregados y cifrados para agosto de 2013, el cual se superpone con el periodo de encuesta del ENCOVI. En 2013 Guatemala tenía 140 cuentas celulares por cada 100 personas[7]. El modelo probó dos tipos de predicciones: (1) predicción de la misma encuesta, tal como predecir los índices de pobreza urbana de 2006 basado en un modelo de relación entre los datos de ENCOVI de 2006 y los CDR de 2013; y (2) predicción de encuestas diferentes, tal como

6. Estos datos provienen de un censo de las áreas rurales de 2011 diseñado para recolectar información para programas de protección social. Ningún censo nacional se realizó ese año.
7. Indicadores de Desarrollo Mundial 2016. Esto refleja múltiples cuentas por persona. Mientras que esto no indica que cada persona tiene un teléfono celular, sugiere que el uso de teléfono celular es alto.

predecir los índices de pobreza rural de 2011 basado en un modelo de la relación entre los datos de ENCOVI de 2006 y los CDR de 2013. La predicción de encuestas mismas es más similar a la aplicación en el mundo real del modelo de datos celulares conocidos como "relleno espacial" o "extrapolación espacial", mientras que la predicción de las encuestas diferentes es un ejemplo de "extrapolación de tiempo."

Preprocesamiento

Limpieza de datos y enriquecimiento

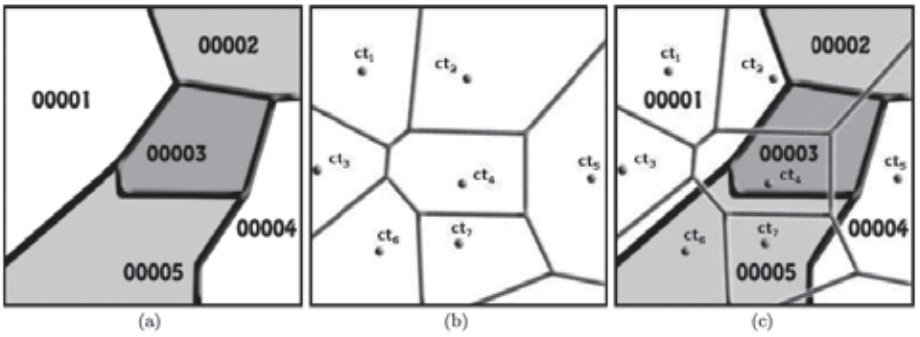
Los datos crudos de CDR son invariablemente ruidosos y requieren de un preprocesamiento antes de poder ser analizados. Las fuentes primarias de ruido en datos son: (1) huecos o inconsistencias causadas por decisiones de operación de las redes de las MNO, incluyendo cambios tecnológicos que afectan la comparabilidad de los CDR; y (2) la presencia de líneas de negocios, números para recargar textos u otras conexiones que no reflejan comunicaciones entre clientes individuales. El preprocesamiento comienza al identificar los CDR inconsistentes o irrelevantes y eliminarlos de la serie de datos. Cada cliente en la serie de datos de CDR es entonces asignado a una ubicación de hogar, la cual es generalmente inferida basada en la red celular en la cual ese cliente se encuentra más activo después de las 6pm. Sin embargo, si esa célula tiene 30% más actividad que la célula más activa siguiente durante el mismo periodo, se asume que el usuario es itinerante y no se asigna una ubicación de hogar.

Armonización espacial

Asegurar que todas las series de datos compartan una escala espacial común es un paso importante en la preparación de datos. Mientras que la red celular es la unidad espacial natural para los datos de CDR, los índices de pobreza son generalmente calculados usando límites administrativos, en este caso el *municipio*. Estas estructuras espaciales deben ser reconciliadas para poder analizar los datos.

En el modelo presentado a continuación, los datos de pobreza están trazados sobre las redes celulares. A cada célula le es asignado un promedio de área pesada de los índices de pobreza en los municipios que cubre. Las redes celulares ubicadas enteramente dentro de un municipio son asignadas al valor de dicho municipio. Este proceso, ilustrado en la figura 4, asegura que todas las series de datos usen la geografía común de la red celular, la cual entonces se convierte en la unidad de análisis. Los enfoques más complejos podrían ser adoptados, tales como ponderar la distribución por densidad de población.

Figura 4. Un ejemplo de armonización espacial



Nota: Unidades Geográficas Unidas (a) de los datos de encuestas (b) con la geografía de las redes celulares definidas como Diagramas Voronoi. (c) La incidencia imputada de pobreza para el polígono 0003 será "x" por ciento de aquel en ct2, "y" por ciento en ct4 y "z" por ciento en ct5. Donde x+y+z =100, representando el área completa del polígono 00003.
Fuente: reproducido de Frías-Martínez *et al.* (2012), figura 1.

Generación de características

Bajo el enfoque de aprendizaje de máquinas la "generación de características es el proceso de colapsar una rica serie de datos multidimensionales hacia un número pequeño de dimensiones cuidadosamente elegidas, las cuales son estadísticas de los datos subyacentes. Estas características entonces forman una representación intermedia de los datos y son la base para el modelo final basado en regresión o basado en clasificación. Por ejemplo, el reconocimiento de la letra escrita a mano es una tarea clásica de aprendizaje de máquinas en la cual una imagen de un dígito escrito a mano representa una serie de datos altamente multidimensional. En algunos casos, cada imagen puede ser 64 por 64 píxeles, para un total de 4.096 dimensiones. Ya que esto puede ser demasiado para que un modelo de clasificación funcione bien, un número más pequeño de características son extraídas usando algoritmos preexistentes bien conocidos. Estas características usualmente representan la presencia o ausencia de características geométricas de un nivel mayor como filos, curvas y esquinas. Este simplifica el modelo final, acelerando el proceso de entrenamiento[8].

8. El paradigma "característica" en el aprendizaje de máquinas está gradualmente siendo superado por enfoques de "aprendizaje profundo", los cuales trabajan directamente con los datos subyacentes altamente dimensionales. Este es especialmente cierto para tareas bien estudiadas como reconocimiento de imágenes. Para tareas más complejas como el modelaje de estructuras sociales basadas en los CDR, el aprendizaje profundo se mantiene como un área de investigación activa.

En el actual análisis, dos series de características son generadas para cada polígono de red:

- **Orientado a la casa del cliente.** Esta primera serie está basada en los registros de clientes quienes viven dentro del polígono, determinado por su ubicación inferida del hogar. Esto es similar a tomar una encuesta de personas en su residencia usual. Las características son calculadas primero por cliente, luego agregadas al polígono de red. Ellas incluyen medidas de consumo (es decir, el número de llamada hechas o recibidas) y medidas de movilidad (es decir, dónde, qué lejos y con qué frecuencia viaja un cliente). Los datos son agregados al tomar el medio sobre todos los clientes, si es apropiado o sino una variedad de valores umbrales —por ejemplo, el número de clientes viviendo en un polígono de red quienes viajan regularmente más allá de 80 kilómetros de su hogar—.
- **Orientado a la activad de polígonos.** La segunda serie está basada en la actividad que ocurre dentro del polígono, sin importar dónde se ubica el hogar del cliente involucrado. Esto es similar a tomar una encuesta de personas que pasan por un área. En este caso, las características basadas en actividad incluyen el número de clientes que entran al polígono de la red, pero viven fuera del mismo, la frecuencia con la que lo visitan y el volumen de llamadas entrantes y salientes que son procesadas. Estas características son calculadas directamente al nivel de la red de polígonos y no requieren futuras agregaciones.

Ajuste del modelo

Estimar los índices de pobreza desde las características de CDR es un ejemplo de un problema de aprendizaje de máquinas "supervisado", o uno donde los datos entrantes son usados para predecir resultados conocidos —en este caso, las características de CDR e índices de pobreza—. Una vez que el modelo es construido, puede ser aplicado a nuevos datos de ingreso para los cuales el resultado correspondiente es desconocido. En este caso, el resultado desconocido sería geografías diferentes o diferentes puntos en el tiempo.

Los problemas de aprendizaje de máquinas supervisados son basados en clasificación, en dicho caso la variable resultante es una de una serie discreta de clases (es decir, masculino/femenino, pobre/no pobre, etc.), o basado en regresión, en la cual la variable resultante es un número continuo real expresado como un decimal, coeficiente o porcentaje. Los índices de pobreza son mayormente modelados naturalmente como una variable resultante continua. Sin embargo, también es

posible agrupar los datos en una serie pequeña de clases reflejando índices de pobreza bajos, moderados o altos. Ambos enfoques fueron examinados en este ejemplo, el cual probó una variedad de escenarios al mezclar diferentes series de datos de entrenamiento y prueba, y empleando tantos métodos de regresión y clasificación. La figura número 5 ilustra las diferentes combinaciones de datos y metodología[9].

Figura 5. Combinaciones de datos y metodología probados

Datos del teléfono celular		Encuesta de entrenamiento	Encuesta de prueba		Método		
2013	X	2006 total	2006 total	X	Regresión		
		2006 rural	2006 rural		Clasificación:		
		2006 urbano	2006 urbano		Ancho igual Igual probable Medios-K Basado en características	X	Línea base SVM Bosques aleatorios Aumento del gradiente Estocástico Medios-K Mezcla gaussiana Modelos de temas supervisados

Fuente: adaptado de Hong y Frías-Martínez (2015).

Caja 2. Aplicando los modelos de pobreza basados en los CDR para rellenar huecos de datos

Mucha de la discusión alrededor del uso de los CDR para la investigación socioeconómica ignora las condiciones prácticas en las cuales serán aplicadas. Los datos de los CDR están casi siempre disponibles adicionalmente a los datos convencionales, tales como encuestas de hogares o censos. De poco sirve mostrar que los datos CDR pueden ser usados para predecir estos datos de encuestas existentes; en vez de eso, el modelo ajustado debe ser aplicado a datos nuevos que no se ven para responder preguntas nuevas. Existen al menos tres diferentes maneras en las cuales los datos de CDR pueden complementar las fuentes convencionales de datos:

9. Para más información, ver: Hong y Frías-Martínez (2015).

1. **Relleno espacial: generando estadísticas de áreas pequeñas.** Dada la relativamente alta resolución espacial de CDR, el relleno espacial posiblemente ofrece el mayor valor agregado como un método de investigación socioeconómica. Una encuesta de hogar particular con un limitado tamaño de muestra solamente puede apoyar los estimados en una resolución espacial relativamente áspera, tal como al nivel del departamento. Las señales de comportamiento de alta resolución en los CDR pueden mejorar la fortaleza estadística de los datos de encuestas, permitiendo estimados precisos y más detallados. Idealmente, el periodo de recolección de CDR debe coincidir con el periodo en el que la encuesta fuese conducida. Este enfoque implica el riesgo de ignorar el rol que el espacio y la geografía puedan jugar en la predicción de índices de pobreza.
2. **Interpolación/extrapolación de tiempo.** La alta frecuencia potencial de los CDR puede también complementar las fuentes convencionales de datos. Las encuestas de hogares son actualizadas generalmente en intervalos de 2 a 5 años. Los CDR pueden ser usados para actualizar estos estimados, proveyendo a los oficiales con información actual sobre temas específicos de políticas. Un modelo predictivo puede ser construido para un año de encuesta usando datos CDR contemporáneos. Este modelo puede entonces ser aplicados a datos CDR más recientes para los cuales los datos de encuestas correspondientes no están disponibles. Existe un riesgo, sin embargo, de que la relación entre señales de comportamiento CDR y el objetivo variable, en este caso el índice de pobreza, pudiera cambiar con el tiempo.
3. **Extrapolación espacial.** Esta es la más ambiciosa aplicación potencial de los datos de CDR. En países o regiones en los cuales no hay datos recientes de encuestas disponibles, tal como áreas afectadas por conflictos o países que han experimentado inestabilidad política severa, los estimados pueden ser generados al usar datos de encuestas y CDR para una ubicación similar y luego aplicar este modelo a los datos CDR de la ubicación objetivo. Este enfoque requiere asunciones significativas, pero podría ser útil en casos donde no existe una fuente fuerte de datos. Hasta la fecha, la investigación sobre las aplicaciones prácticas de la extrapolación espacial ha sido limitada.

Evaluando el modelo

Una gran desventaja de los modelos de aprendizaje de máquinas es un fenómeno conocido como sobreajuste (*overfitting*). Los datos de entrenamiento siempre reflejan "señales" significativas y un "ruido" aleatorio. El sobreajuste ocurre cuando el modelo que se acopla tiene muchos parámetros libres por lo que se acopla tanto a la señal como al ruido. En dicho caso, el modelo parecerá ser un acople excelente para los datos de entrenamiento, con alto R^2, pero tendrá un rendimiento pobre cuando se aplica a datos externos a la muestra.

En el caso de Guatemala, una técnica llamada "validación cruzada" fue usada para protegerse del sobreajuste. La validación cruzada divide los datos de entrenamiento en partes, entrenando al modelo en una subserie de los datos (75%) y luego probándolo en los datos restantes (25%). Esto permite que los valores diagnosticados de la serie de prueba provean una muestra más precisa de cómo el modelo rendiría en un escenario fuera de la muestra. Las subseries de validación cruzada pueden ser construidas en múltiples ocasiones en diferentes maneras, con resultados promediados, para proveer mejores resultados.

Para evaluar los resultados de regresión, la precisión del modelo de aprendizaje máquina es medida usando R^2, raíz cuadrada del error cuadrático medio y la correlación entre los valores reales y pronosticados. R^2 mide la extensión en la cual el modelo explica la variabilidad de los datos de respuesta sobre su media, mientras que la raíz cuadrada del error cuadrático medio indica la diferencia entre los valores reales y los valores pronosticados. La calidad de las técnicas de clasificación es analizada usando dos medidas; precisión y valor F1. La precisión refleja el porcentaje de las muestras probadas cuya clase pronosticada es la misma que sus muestras reales. El valor F1 es un parámetro que combina la precisión y la cobertura del método, es decir, el número de muestras que están correctamente clasificadas y el número de muestras para las cuales se proporciona una etiqueta. En general, las metodologías más fuertes tienen valores mayores tanto para precisión como para la sensibilidad.

Resultados

Todos los modelos basados en CDR exhibieron un grado significativo de valor predictivo. Sin embargo, las especificaciones de diferentes modelos influenciaron en cómo de bien estas predecían índices de pobreza. Los análisis arrojaron cuatro resultados generales:

Resultado #1. Los CDR pueden predecir los índices de pobreza en Guatemala

A través de todos los modelos, el análisis CDR consistentemente predijo los índices de pobreza a nivel municipal, aunque su valor predictivo estaba limitado bajo ciertos parámetros experimentales. Los mejores modelos predijeron niveles de pobreza total en el 2006 con un R^2 de 0,76, indicando que aproximadamente 76% de la variación en índices de pobreza a nivel *municipal* podría ser explicada por los datos de teléfonos celulares de 2013, y con valoraciones F1 de hasta 0,84 para los modelos de clasificación, indicando que 84% de los municipios estaban clasificados según la categoría correcta cuando tres categorías separadas (índices de pobreza bajo, medio y alto) eran considerados. Por otra parte, los índices de predicción para los datos urbanos en 2006 mostraron valores R^2 de 0,69 para técnicas de regresión y 0,73 para clasificación, indicando un valor predictivo más débil. Los índices más bajos de predicción fueron para los datos rurales en el 2011, con resultados R^2 de 0,46 y 0,59 para modelos de regresión y clasificación, respectivamente. Los resultados de clasificación fueron ligeramente mejores que los resultados de regresión a través de todos los modelos debido al hecho de que clasificar índices de pobreza en tres clases es un problema de predicción más simple que tratar de aproximar valores reales. Experimentalmente, mientras más clases eran incluidas, las precisiones predictivas para la clasificación disminuyeron y convergieron con aquellas para la regresión. Por lo tanto, el número de clases de pobreza seleccionadas implican una compensación entre la precisión y la granularidad de la predicción.

Resultado #2. En Guatemala los CDR predicen la pobreza urbana y total con más precisión que la pobreza rural

Tanto en modelos de regresión como de clasificación, los índices de pobreza rural fueron consistentemente más difíciles de predecir que los índices totales o urbanos. Los valores R2 cayeron hasta casi 0,25 para pobreza rural, insinuando que los datos CDR podrían explicar solo el 25% de la variación en índices de pobreza rural. La precisión de los modelos de clasificación disminuyó hasta entre 0,3 y 0,65 dependiendo de la especificación. Dos hipótesis podrían explicar este fenómeno. Primero, los índices de penetración celular en áreas urbanas tienden a ser mayores, y por ende los análisis basados en CDR proveen señales de modelado más robustas en áreas urbanas, donde representan el comportamiento de una porción más grande de la población. En áreas rurales, menos teléfonos y menos llamadas debilitan la señal que puede ser extraída de los datos CDR, y su respectivo poder predictivo disminuye. Segundo, las áreas urbanas tienden a tener más antenas celulares por kilómetro cuadrado, lo cual resulta en polígonos

más pequeños. Los polígonos de mayor tamaño en áreas rurales pueden tender a reducir la granularidad de los datos al agregar comportamientos, debilitando el poder predictivo del algoritmo. Probar estás hipótesis requerirá de más investigación.

Resultado #3. Más análisis será necesario para determinar la extensión en la cual los modelos basados en CDR sobre datos de pobreza pasada pueden ser utilizados para predecir futuras dinámicas de pobreza

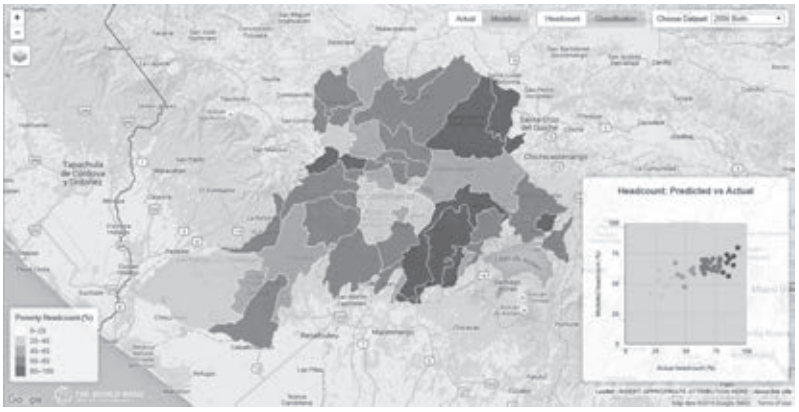
Los modelos basados en CDR podrían potencialmente ser usados para predicción temporal o extrapolación de tiempo (ver cajas 2 y 3), pero ninguna investigación ha podido todavía probar que esos modelos entrenados en valores de pobreza pasados pueden ser usados para predecir niveles de pobreza futuros. Dichos análisis requerirían una serie de datos extremadamente grandes y detallados. Por ejemplo, predecir los valores futuros de pobreza en Guatemala requeriría datos CDR del 2006 y 2011, además de los datos de encuestas de pobreza correspondientes para los mismos periodos de tiempo. Sin embargo, este análisis fue basado en datos de CDR de 2013 y niveles de pobreza rural para 2006 y 2011, y ningún dato de pobreza urbana o nacional fueron provistos. Como resultado, el modelo predictivo fue entrenado con los datos de CDR de 2013 y los datos de pobreza rural de 2006, mientras que los datos de CDR de 2013 fueron usados para predecir niveles de pobreza rural en el 2011. Mientras que este es el mejor enfoque metodológico dadas las restricciones de datos, los resultados preliminares mostraron valores bajos de R^2 de alrededor de 0,09 para regresión y valoraciones F1 máximas de 0,6.

Caja 3. Usando visualización de datos interactivos para comunicar resultados de modelos

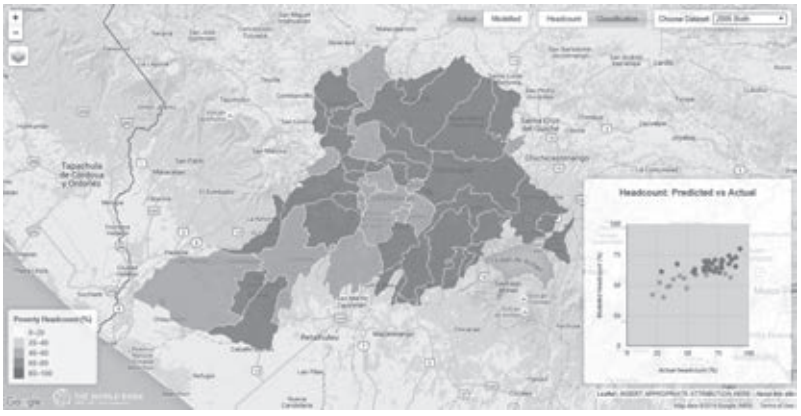
Los datos de Guatemala fueron mapeados en un sitio web interactivo. El mapa permitía a los usuarios cambiar entre modelos de regresión y clasificación, al igual que también diferentes periodos de datos. Los mapas muestran las estimaciones de pobreza mediante una escala de colores, mientras que los diagramas de dispersión y las matrices de confusión proporcionan evaluaciones más detalladas del rendimiento de cada modelo.

Figura 6. Índices actuales de pobreza (A) e índices de pobrezas modelados (B) para los municipios incluidos

(A)



(B)



Fuente: La visualización del autor basada en datos de Hong and Frías-Martínez (2015b).

Otras aplicaciones de datos de CDR para políticas de desarrollo

Mientras que los modelos de CDR no son aún lo suficientemente precisos para suplantar métodos tradicionales de investigación, las crecientes técnicas sofisticadas para combinar los CDR con otras fuentes de datos pudiera permitir que los datos celulares magnifiquen el valor de los censos y encuestas de hogares. Dos enfoques especialmente prometedores están descritos a continuación.

Predicciones a nivel de unidad usando encuestas móviles de datos reales del terreno

Un estudio reciente en Ruanda usó encuestas focalizadas de teléfono para recolectar datos sobre la riqueza personal en vez de la incidencia de pobreza[10]. La ventaja de este enfoque es que las respuestas de encuestas de teléfono pueden ser combinadas con los datos de CDR a nivel individual. Esto típicamente no es posible con encuestas de hogar oficiales, las cuales son usualmente anónimas y no contienen un número celular que pueda ser analizado con referencias cruzadas de los registros de CDR.

Al coordinar la recolección de respuestas de encuestas con análisis de CDR, el estudio pudo desarrollar una serie de datos individuales de gran calidad. Sin embargo, el uso de encuestas de teléfono para recolectar datos presenta sus propias limitaciones, ya que no es posible obtener la misma información de consumo detallado que las encuestas de hogar cara a cara generalmente revelan. En vez de eso, los valores aproximados de riqueza deben ser usados —en este caso, con preguntas sobre propiedad de activos—. Si los resultados no son cuidadosamente validados, esto puede llevar a errores en las variables resultantes. Adicionalmente, emparejar los datos con las encuestas de teléfonos requiere que los CDR no sean anónimos, lo cual genera preocupaciones sustanciales sobre privacidad.

El estudio en Ruanda ofreció una prueba motivadora de concepto para la aplicación de datos de CDR para la "extrapolación de tiempo". Encontró que un modelo de CDR para 2009 proveyó una evaluación de la pobreza más precisa que la encuesta de hogar obsoleta de 2007. Sin embargo, el modelo CDR no era únicamente apropiado para esta tarea, ya que los indicadores macroeconómicos como el crecimiento del PIB podría ser usado para crear una proyección similar precisa[11]. Además, el modelo CDR solo reproducía parcialmente los estimados de la riqueza de hogares promedio a nivel de distrito registrados por las encuestas de hogar. Aunque correlaciones de cerca de 0,9 fueron reportadas a nivel de distrito —mayor que en el análisis de Guatemala— parece que unos pocos distritos adinerados pueden haber generado los resultados obtenidos.

10. Blumenstock *et al*. (2015).
11. Ver Beegle (2016).

Combinando los datos de CDR con datos observables por el público

Durante los últimos años el proyecto WorldPop[12], con base en la Universidad de Southampton, ha producido mapas cuadriculados de la población para decenas de países, los cuales detallan la población estimada por cada 100 metros cuadrados[13]. WorldPop también ha producido estimaciones de pobreza cuadriculadas a una resolución de 1 kilómetro para un pequeño número de países usando un método similar basado en datos de encuestas de hogares. En colaboración con la Fundación Flowminder, WorldPop está ahora empezando a integrar características derivadas de CDR dentro del mismo marco. Mientras que relativamente poca información ha sido publicada sobre este enfoque, tiene la ventaja de poder incorporar todos los datos relevantes en un mismo formato. Los mapas resultantes deben no solo ser altamente precisos, sino también tan consistentes como sea posible con las series de datos subyacentes. Una advertencia sobre el enfoque WorldPop es que la inclusión de la dimensión de pobreza aún está en una etapa inicial, y no está claro cómo estos mapas son validados, ya que por diseño deberían correlacionar con las estimaciones de pobreza basadas en encuestas[14].

› Usar datos de CDR para trazar el nivel de la pobreza

Mientras que los enfoques técnicos del uso de los CDR para trazar el nivel de la pobreza continúan evolucionando, se requiere abordar un número de retos para poder poner en práctica esta metodología como una herramienta práctica para el análisis de políticas. Varios de estos retos están descritos abajo. Un rango de otros problemas éticos y legales está considerado en un informe oficial (Libro Blanco) realizado entre el Banco Mundial y Data-Pop Alliance[15].

Validación y transparencia

La validación es más complicada para modelos de CDR de lo que es para técnicas basadas en encuestas. Los instrumentos de encuestas son relativamente

12. Http://www.worldpop.org.uk
13. Estos mapas son construidos al combinar datos de censos con covariables físicas, como clima, elevación, inclinación y cuerpos de agua, y covariables humanas, tales como expansión humana, basada en imágenes de satélite, asentamientos conocidos, caminos y puntos de interés bajo un marco Bayesiano.
14. El Laboratorio de Innovación del Banco Mundial está patrocinando actualmente trabajos de análisis CDR adicionales con Flowminder/Worldpop en Haití, lo cual se espera que proveerá mayor introspectiva sobre sus métodos.
15. Letouzé y Vinck (2015).

transparentes. Estas pueden ser inspeccionadas, y el trabajo de campo cuantitativo y cualitativo realizado antes y después de la recolección de datos puede generar confianza. Las inconsistencias temporales y espaciales pueden ser revisadas y en raros casos las encuestas superpuestas pueden ser verificadas. Como resultados, los programas de encuestas como el Estudio de Medición de Estándares de Vida y Encuestas de Salud son usualmente tratados como lo más cercano a la realidad del terreno.

La validación es generalmente más difícil para investigaciones de grandes datos y para modelos de CDR en particular. Estos datos no son recolectados específicamente para análisis socioeconómicos, por lo que la inspección puede no ser útil y pruebas previas no son usualmente posibles. Además, los modelos de CDR son diseñados usualmente para interpolar entre modelos convencionales de encuesta bien sea en tiempo o espacio, por lo que la validación directa contra una encuesta no es posible.

Sin embargo, las estrategias de validación rigurosas deben ser desarrolladas. Como mínimo, las técnicas de validación dentro de muestras deben ser usadas[16]. El potencial para la validación fuera de muestras no se conoce, pero el acceso a las series de datos de CDR más largos, los cuales incluyen más de una encuesta de hogar, pudiera ayudar a lidiar con esta preocupación.

Privacidad

La privacidad de datos es un problema sensible y usualmente controvertido, por lo que ciertas precauciones deben ser tomadas antes de analizar datos de CDR. Los identificadores deben ser oscurecidos antes que los registros sean exportados desde los sistemas de las MNO, para que los números de teléfonos celulares o campos similares no permanezcan en los datos salientes. Este proceso es referido como "seudonimización". En general, la seudonimización debe asegurar que el mismo seudónimo sea aplicado a la historia completa de llamadas de un usuario dado.

La investigación sugiere que la seudonimización simple puede ser revertida por un determinado individuo armado con información auxiliar relativamente fácil de obtener, tal como las direcciones de trabajo y hogar de una persona y una o dos ubicaciones a las que se conoce que visitan en momentos particulares[17]. Por

16. Las técnicas de validación dentro de muestras incluyen series de pruebas de retención o validación en cruce de descartar-uno.
17. De Montjoye (2013).

lo tanto, las medidas técnicas e institucionales adicionales son necesarias para restringir el acceso a los datos. Esto generalmente implica acuerdos de no divulgación entre las MNO y los investigadores, y las investigaciones, desarrollos y comunidades MNO se están esforzando para racionalizar este proceso.

Adicionalmente, cualquier dato final como resultado del análisis CDR no debe comprometer la privacidad del individuo. Las precauciones similares a aquellas usadas cuando se liberan tabulaciones de los datos de encuestas tradicionales puede reforzar la privacidad de los datos de CDR. Estas medidas pueden incluir agrupar datos donde los tamaños de las células caen debajo de un número fijo de personas (es decir, 5 o 10), o variables de codificación altos y bajos en casos donde los valores extremos pudieran ser reveladores.

Otras aplicaciones prometedoras de la analítica CDR

Los CDR ofrecen una rica serie de datos para estudiar poblaciones, y aplicaciones numerosas están emergiendo en áreas más allá de la medición de la pobreza. Dos aplicaciones de relevancia particular al trabajo del Banco Mundial en Guatemala y la región de Latinoamérica están descritas a continuación[18].

Análisis de transporte

El análisis de transporte está entre las aplicaciones más prometedoras para los modelos de CDR. Las características económicas y sociales son solo registradas implícitamente en los datos de CDR, y por ende fuertes asunciones y modelajes complejos son requeridos para inferirlos. El comportamiento espacial, por otra parte, está explícitamente registrado por las ubicaciones de las antenas celulares. Por esta razón el análisis puramente espacial tiende a ser más simple y más robusto que otras formas de análisis de CDR[19].

Debido a la relativa simplicidad del análisis, algunas MNO y terceras partes están empezando a ofrecer productos de datos de transporte estándar a gobiernos locales y nacionales.

18. Una reseña más completa es provista en Blondel (2015).
19. Por ejemplo, Angelakis *et al.* (2013) usó datos CDR para examinar los patrones de transporte Cote d'Ivoire. Ellos descubrieron que varias matrices de viajes, incluyendo rutas y horas de viajes podrían ser calculados desde esos datos tanto a nivel nacional y a nivel de ciudad.

Preparación y respuesta ante desastres

Los grandes movimientos de la población usualmente ocurren a raíz de los desastres naturales, y estos movimientos pueden dejar tanto a los censos como las encuestas en lugares de predesastre en forma obsoleta. Para poder proveer asistencia humanitaria y restaurar servicios básicos en las áreas afectadas por el desastre, los gobiernos y las organizaciones requieren de datos actualizados de la población que puedan ser recolectados rápidamente y a un costo modesto. Este fue el caso de Haití después del terremoto de 2010, el cual inspiró a varios investigadores a examinar el potencial de datos de CDR para producir rápidamente información de rastreo de alta frecuencia sobre el desplazamiento de la población a corto plazo. Además, se encontró que esos datos históricos detallados sobre la movilidad de la población en áreas de predesastre podrían ser usados para predecir respuestas de residentes al terremoto, permitiéndole a las agencias prepararse mejor para futuros desastres[20]. Las técnicas similares han sido desde entonces aplicadas exitosamente durante y después de otros desastres naturales[21].

Conclusiones

La expansión dramática del uso de teléfonos móviles en países en desarrollo en años recientes ha producido una fuente de información rica y mayormente sin explotar sobre las características de las comunidades y regiones. Los métodos de investigación basados en CDR tienen el potencial para proveer estimados detallados y confiables de índices de pobreza en tiempo real y a un costo mucho menor que las encuestas tradicionales. Estos métodos tienen aplicaciones especialmente prometedoras en países en desarrollo, como en Guatemala, donde los altos índices de pobreza e inequidad y los limitados recursos fiscales y presupuestarios complican la tarea de recolectar datos y acentuar la importancia de focalizar con precisión el gasto público.

El análisis de CDR puede complementar métodos convencionales de investigación al mejorar la fortaleza estadística de datos de encuestas y al extrapolar estos datos a través del espacio y tiempo. El análisis presentado anteriormente estaba limitado a datos de CDR agregados y cifrados de solo cinco departamentos administrativos en el suroeste de Guatemala, y los resultados sugieren que expandir el tamaño de la muestra permitiría estimados de pobrezas más

20. Flowminder (2016a).
21. Ver para ejemplo Moumny *et al.* (2013) y Flowminder (2016b).

robustos y confiables. Los legisladores en Guatemala podrían obtener series de datos más exhaustivos al trabajar directamente con las MNO.

Mientras las metodologías analíticas son desarrolladas, las aplicaciones CDR podrían extenderse más allá del estudio de la pobreza. Los CDR podrían permitir a los legisladores rastrear patrones de crimen, inseguridad de comida, enfermedades epidémicas y otros problemas sociales y económicos en tiempo real. Los legisladores en Guatemala y otros países en desarrollo ahora tienen la capacidad de acceder a una fuente de riqueza de datos celulares. El establecer fuertes asociaciones con operadoras móviles será el primer paso para aprovechar el enorme potencial de los datos celulares para la investigación socioeconómica y análisis de la política.

Referencias bibliográficas

Angelakis, V., Gundlegård, D., Rajna, B., Rydergren, C., Vrotsou, K., Carlsson, R., Forgeat, J., Hu, TH., Liu EL., Moritz, S., Zhao, S., Zheng, Y. (2013). Mobility modeling for transport efficiency- analysis of travel characteristics based on mobile phone data. In: *Mobile phone data for development-analysis of mobile phone datasets for the development of Ivory Coast*. Orange D4D challenge, pp. 412-422.

Beegle, K., Christiaensen, L., Dabalen, A., Gaddis, I. (2016). *Poverty in a Rising Africa*. Washington, DC: World Bank. doi: 10.1596/978-1-4648-0723-7.

Blondel, Vicent D., Decuyper, A., Krings, G. (2015). "A survey of results on mobile phone datasets analysis". *EJP Data Science*. http: //link.springer.com/article/10.1140/ epjds/s13688-015-0046-0.

Blumenstock, J., Cadamuro, G., On, R. (2015). "Predicting poverty and wealth from mobile phone metadata", *Science,* 350: 1073-1076.

Elbers, Ch., Lanjouw, P., Lanjouw, J. (2003). "Micro-Level Estimation of Poverty and Inequality". *Econometrica*, vol. 71 (1): 355-364.

Flowminder (2016a). "Case Study: Haiti Earthquake 2010", http: //www.flowminder. org/case-studies/haiti-earthquake-2010

Flowminder (2016b). "Case Study: Nepal Earthquake 2015", http: //www.flowminder. org/case-studies/nepal-earthquake-2015

Frías-Martínez, V., Frías-Martínez, E., Oliver, N. (2010). "A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records". https: // www.aaai.org/ocs/index.php/SSS/SSS10/paper/viewFile/1094/1347

Frías-Martínez, V., Virseda, J. (2012). "On the relationship between economic factors and cell phone usage", International Conference on Technologies and Development, ICTD.

Hong, L., Frías-Martínez, E., Frías-Martínez, V. (2016). "Topic Models to Infer Socio-economic levels", Thirtieth International Conference on Artificial Intelligence, AAAI.

Hong, L., Frías-Martínez, V. (2015a). "Estimating Incidence Values Using Mobile Phone Data: Deliverable 2: Statistical Models". Unpublished manuscript, June 4.

Hong, L., Frías-Martínez, V. (2015b). "Prediction of Incidence Levels". Unpublished manuscript, June 4.

Letouzé, E., Vinck, P. (2015). "The Law, Politics and Ethics of Cell Phone Data Analytics". Data-Pop Alliance White Paper Series. Data-Pop Alliance, World Bank Group, Harvard Humanitarian Initiative, MIT Media Lab and Overseas Development Institute, April.

Montjoye, Y.-A. De, Hidalgo, C. A., Verleysen, M., Blondel, V. D. (2013). "Unique in the Crowd: The privacy bounds of human mobility". *Nature S.Rep.* 3.

Moumny, Y., Frías-Martínez, V., Frías-Martínez, E. (2013). "Characterizing Social Response to Urban Earthquakes using Cell-Phone Network Data: The 2012 Oaxaca Earthquake", Third Workshop on Pervasive Urban Applications @ Pervasive'13, Zurich, Switzerland.

Sánchez, S., Scott, K., López, H. (2016). *Guatemala: Closing Gaps to Generate More Inclusive Growth*. Washington, D.C.: The World Bank.

Simon, P. (2012). "IFC Mobile Money Scoping: Country Report: Guatemala", International Financial Corporation, World Bank Group, http: //www.ifc.org/wps/wcm/ connect/8b233f0043efb60d95b6bd869243d457/Guatemala+Public.pdf?MOD=AJPERES

Capítulo 19

Transformando datos en impacto sin gastar una fortuna: la experiencia experimental de Nesta

Juan Mateos García*

› Introducción

En años recientes, hemos visto una explosión en el volumen, variedad y velocidad de datos disponibles para el análisis (Manyika *et al*., 2011; Mayer-Schönberger and Cukier, 2013; OECD, 2014). Este fenómeno, consecuencia de la digitalización de la economía y la sociedad, viene acompañado de innovaciones en las tecnologías para el procesamiento de datos, y las técnicas para su análisis (esto es, para extraer información de los datos recogidos). Los resultados del análisis están transformando la toma de decisiones en empresas a lo largo y ancho de la economía, y siendo utilizados como base para nuevos productos, servicios y modelos de negocio (Bakhshi *et al*., 2014; Bakhshi and Mateos-Garcia, 2012).

¿Qué quiere decir todo esto para el sector público? Caben pocas dudas de que este necesita innovación para mejorar su efectividad y eficiencia, y desarrollar nuevos servicios para adaptarse a un mundo complejo, cambiante y lleno de retos. Es también innegable que una utilización efectiva de los datos puede contribuir a superar esos retos, por ejemplo, mejorando la diagnosis de los pacientes en los hospitales, personalizando la experiencia educativa en las escuelas, reduciendo el fraude en muchas partes del sector público, o mejorando la capacidad para medir la economía, y desarrollando políticas para apoyar el crecimiento. En 2011, McKinsey Global Institute estimaba un valor potencial de los datos masivos (*big data*) para el sector público en Europa de 250 billones de euros —más que el PIB de Grecia—.

El reto es transformar estas oportunidades en impactos tangibles. A trazo grueso, hay dos maneras de perseguir este objetivo: "de arriba abajo", llevando a cabo grandes inversiones en tecnologías "big data" y después buscando aplicaciones para ellas, o "de abajo arriba" realizando pilotos estratégicos para descubrir qué aplicaciones tienen más potencial e identificar las infraestructuras más adecuadas para desarrollarlas. La segunda estrategia tal vez sea menos vistosa, pero en

* Head of Innovation Mapping en Nesta.

mi opinión es más inteligente, reduciendo el riesgo de gastar recursos escasos en tecnologías irrelevantes.

En este capítulo, describo nuestra experiencia utilizando esta estrategia para crear capacidad de analítica de innovación en Nesta[1]. Aunque me centro en políticas de innovación, creo que muchas de las lecciones que hemos aprendido, y que describiré en las siguientes páginas, serán también relevantes para agencias en otras áreas interesadas en desarrollar su capacidad analítica, independientemente de su área de trabajo.

La estructura del capítulo es la siguiente: en la próxima sección describo los problemas que queríamos superar en el área de política de innovación, y las oportunidades ofrecidas por la analítica de innovación (que también definiré). Habiendo hecho esto, en la tercera sección hablo del proceso experimental que hemos seguido para crear esta capacidad, prestándole especial atención a la adquisición de conocimientos y la administración de proyectos. La sección cuarta concluye.

› Problemas y soluciones

La obsesión con los datos masivos y su potencial para revolucionar la economía y la sociedad ha dado lugar a que algunas organizaciones privadas y públicas lleven a cabo grandes inversiones en infraestructura para procesar petabits de datos sin hacerse antes las preguntas más importantes: ¿cuál es el problema que estos datos van a ayudar a resolver? ¿Qué pregunta van a ayudar a contestar? ¿Cómo van a ayudar a nuestra organización a cumplir sus objetivos?

En el caso de Nesta, la percepción del problema surgió tras muchos años de trabajo en el área de políticas de innovación, donde realizamos proyectos de investigación y desarrollo de políticas para incrementar los niveles de innovación y capacidad de crecimiento del Reino Unido. El problema con el que nos encontramos es que las fuentes de datos utilizadas tradicionalmente en

1. Nesta es la fundación de apoyo a la innovación del Reino Unido. Somos una agencia establecida en 1998 con la misión es apoyar la aplicación de nuevas ideas por el bien común, a través de programas prácticos (becas, educación y entrenamiento, premios y retos), inversiones e investigación. Utilizamos un enfoque experimental, trabajando con nuevos modelos y metodologías con el objetivo de reducir su incertidumbre y fomentar su adopción en otras partes del sector público, en aquellos casos en los que su potencial de impacto se confirme a través de evaluaciones rigurosas.

este campo sufren fuertes limitaciones que dificultan su utilización. Antes de explicar en qué consisten estas limitaciones, describo brevemente el propósito de las políticas de innovación, sus herramientas y etapas.

La importancia de la innovación y las políticas para apoyarla

La literatura económica ha reconocido en las últimas décadas que el principal determinante de la riqueza de las naciones, y su crecimiento a largo plazo, es su capacidad de generar y aplicar nuevas ideas —esto es, de la innovación—. Las políticas de innovación buscan fomentar ambas actividades, con el objetivo de generar crecimiento, sostenibilidad y bienestar (Corrado *et al.*, 2012; Gordon, 2016; Solow, 1957).

Ministerios de Ciencia y Tecnología, Ministerios de Industria y Agencias de desarrollo tecnológico buscan apoyar la innovación a través de financiación pública para la investigación básica y subsidios para la investigación y desarrollo (I+D) en el sector privado, un marco regulatorio que protege los derechos de propiedad intelectual, y políticas que fortalecen la colaboración entre distintas agentes en el "sistema de innovación" como universidades y empresas, y apoyo a la emprendeduría (entre otras cosas) (Edler *et al*., 2013).

Estas políticas siguen un "ciclo vital" similar a las de otras áreas: primero, una fase de diseño en la cual se identifica el problema u objetivo de la política, y se decide el mecanismo más apropiado para llevarlo a cabo; después, una fase de implementación para aplicar la política; y finalmente, una fase de supervisión y evaluación para determinar los impactos de la política, y decidir si hay que continuarla sin cambios, adaptarla o abandonarla[2].

Limitaciones en los datos

Cada una de las fases del ciclo vital de una política de innovación requiere evidencia para contestar a tres preguntas claves: ¿qué hacemos? ¿Cómo lo hacemos? ¿Qué tal lo hicimos? El análisis de distintas fuentes de datos genera esta evidencia.

Tradicionalmente, las principales fuentes de datos para las políticas de innovación han sido encuestas oficiales llevadas a cabo por agencias estadísticas y Ministerios de Industria e indicadores de Ciencia y Tecnología como el número de

2. Obviamente, esta descripción simplifica un proceso complejo en el cual hay importantes interacciones y retroalimentaciones entre las distintas fases.

publicaciones científicas, el número de patentes generadas por las empresas y el número de investigadores e ingenieros educados en las universidades (Fagerberg *et al*., 2006).

El problema es que estas fuentes de datos tienen limitaciones muy serias que restringen su utilidad para la política de innovación. Estas limitaciones son una consecuencia directa de cuatro características esenciales de los procesos de innovación que la hace difícil de medir con fuentes de datos estructuradas y estáticas (Nesta, 2016). Merece la pena señalar que muchas de esas características las encontramos también en otras áreas de política como la educación, la salud o la seguridad. Esto no es una coincidencia: rápidos cambios en tecnologías y demandas sociales generan turbulencia y complejidad en esas áreas, y dificultan su estudio utilizando fuentes de datos tradicionales.

En primer lugar, tenemos la novedad: la innovación ocurre en la intersección entre industrias y genera nuevas industrias. Una política de innovación efectiva necesita evidencia acerca de este cambio estructural en la economía, pero las fuentes de datos oficiales están basados en estándares sectoriales internacionales que dificultan su utilización para identificar y medir nuevas actividades[3]. Parafraseando al Premio Nobel de Economía Robert Solow, a las nuevas industrias las vemos en todos sitios, menos en las estadísticas oficiales.

En segundo lugar, la ubicuidad. La innovación no solo tiene lugar en sectores basados en la ciencia y la tecnología. Es también muy importante en sectores de servicios, en las industrias creativas y culturales, en las ONG (innovación social), y en el sector público (Harris and Halkett, 2007). Dado que estos sectores no suelen generar patentes o publicaciones, sus actividades innovadoras permanecen "escondidas" de las estadísticas oficiales, y son, por lo tanto, más difíciles de estudiar y apoyar.

Tercero, la complejidad: para comprender ecosistemas y dinámicas de innovación, hay que medir muchas variables distintas, y las relaciones entre ellas (por ejemplo, colaboraciones entre empresas y universidades en un *cluster i*ndustrial).

3. Algunas políticas de innovación se centran en sectores o áreas tecnológicas predeterminadas —en este sentido, tienen un aspecto de estrategia de industrial—. Esto se justifica en base a que algunos sectores tienen un potencial de crecimiento más elevado, desarrollan tecnologías que son más beneficiosas para el resto de la economía o tienen necesidades únicas que requieren programas específicos. El sector TIC (tecnologías de información y comunicación) es un ejemplo de un sector que crea "tecnologías de aplicación generalizada" que muchos países buscan fomentar. Otros economistas enfatizan la importancia de invertir en la economía "verde" para incrementar la sostenibilidad de la economía.

También son necesarias técnicas de análisis que capturen interdependencias sutiles entre múltiples variables. Desafortunadamente, pocas fuentes tradicionales capturan relaciones entre organizaciones, y los métodos de análisis estadísticos y econométricos convencionales buscan generalizar relaciones sencillas entre pocas variables.

En cuarto y final lugar está la multiplicidad: el sistema de innovación incluye muchos participantes distintos: políticos nacionales y locales, reguladores, universidades, inversores, empresas y trabajadores toman decisiones que, de forma conjunta, generan innovación y crecimiento —y todos ellos tiene preguntas por contestar—. El problema es que los formatos que utilizamos habitualmente para comunicar resultados —por ejemplo, informes en formato PDF– restringen la cantidad de información que es posible presentar, y deja muchas preguntas importantes sin contestar. De manera adicional, algunos de los participantes en el sistema de innovación están más interesados en información detallada ("¿cuál es la empresa que crece más rápido?") que en resúmenes estadísticos ("¿cuál es la media de crecimiento en un sector o región?"). Es difícil comunicar esta información utilizando informes "estáticos", o imposible incluso cuando el análisis está basado en fuentes oficiales anonimizadas y agregadas por industria o localidad.

La oportunidad analítica

La analítica de innovación ofrece importantes oportunidades para superar estas limitaciones, mejorando la calidad de la información disponible para el apoyo de la innovación. Este concepto incluye los siguientes componentes:

- *Nuevas fuentes de datos:* la explosión de los datos ha generado nuevas fuentes para comprender procesos de innovación complejos, estudiar la emergencia de nuevas industrias y medir la innovación en sectores que no patentan o publican. Dos ejemplos interesantes son las páginas web de las empresas, que se pueden analizar para identificar sus industrias o medir sus innovaciones, así como redes sociales y plataformas de colaboración digital como Twitter, que publican sus datos utilizando APIs (*Application Programming Interfaces*) abiertos. Estos datos pueden utilizarse para medir nuevos tipos de innovación y mapear las colaboración entre organizaciones e individuos.
- *Nuevas combinaciones de datos:* las nuevas fuentes de datos web pueden combinarse con fuentes de datos oficiales, públicas y abiertas. Estas fusiones nos permiten considerar más dimensiones en los sistemas y procesos de

innovación, y ofrecen una visión más completa y holística de fenómenos complejos, como *clusters* industriales que reúnen empresas, universidades, redes de colaboración y agencias de apoyo.

- *Nuevas técnicas de procesamiento y análisis:* las infraestructuras de datos masivos *(big data)* utilizan métodos de almacenamiento distribuido y procesamiento en paralelo para trabajar con grandes volúmenes de datos, y bases de datos "no-relacionales" *("NoSQL",* a diferencia de las fuentes de datos relacionales o *SQL*) para almacenar datos complejos con mayor eficiencia (por ejemplo, fuentes de datos desestructuradas donde una observación —digamos que una empresa— tiene un número indeterminado de variables —tal y como productos con diferentes características—).
  También ha habido importantes innovaciones en el desarrollo, usabilidad y efectividad de metodologías para la analítica de datos, como la minería de texto, que nos permite extraer información cuantitativa de fuentes de datos desestructuradas (por ejemplo, identificar el asunto al que se refiere un documento, o el "sentimiento" de un *tweet*); el análisis de redes sociales para medir la estructura de las redes de colaboración y el aprendizaje automático *(machine learning)* que utiliza algoritmos para generar modelos de las relaciones entre variables en una fuente de datos, y predecir o clasificar nuevas observaciones (por ejemplo, ¿podemos identificar las características de una empresa innovadora, y utilizar esta información para predecir que empresas innovarán en el futuro?).
- *Nuevos formatos para comunicar y diseminar los resultados:* la interactividad de la red posibilita nuevos formatos para publicar y presentar datos, como visualizaciones de datos interactivas, tableros de datos *(dashboards)* y publicación de datos abiertos (a través de archivos descargables o APIs) que los usuarios pueden explorar, reanalizar y fusionar con otros datos.

Experimentación

Los componentes de la analítica de innovación que acabo de describir permiten, al menos en teoría, superar muchas de las limitaciones de las fuentes de datos tradicionales en esta área. Pero los beneficios de su aplicación no están garantizados: estas nuevas fuentes de datos y metodologías de análisis tienen sus propias limitaciones (por ejemplo, no tienen el mismo control de calidad que las estadísticas oficiales, tienen otros sesgos, y pueden ser difíciles de presentar y comunicar). Además, adquirir las capacidades para trabajar con ellos tiene costes, especialmente dada la alta demanda laboral de expertos analistas (particularmente, científicos de datos o *data scientists,* a los que me referiré en breve).

En resumen, el reto organizacional para desarrollar estas capacidades es el mismo con el que se enfrenta cualquier innovador: hay que invertir recursos escasos en actividades con beneficios inciertos. En esta situación existe el riesgo de que triunfe la inercia: decidir que lo mejor es no hacer nada.

Como mencioné en la introducción, la mejor forma de hacer frente a este reto es a través de la experimentación, un modelo de trabajo que forma parte del DNA de Nesta. En vez de llevar a cabo inversiones a gran escala, se diseña un portafolio de pilotos con bajo coste para explorar distintas opciones. Los resultados de estos pilotos nos ayudan a determinar cuáles son las rutas más atractivas, identificar las capacidades para recorrerlas, y ganar la credibilidad interna para invertir en ellas.

En esta sección describo los componentes de la estrategia de experimentación que hemos utilizado en Nesta para llevar esto a cabo, comenzando con la adquisición de conocimiento y personal para realizar estos experimentos, y refiriéndome después a su administración. Termino con algunas estrategias complementarias —como hackatons, retos e inserción de personal— utilizadas por Nesta en otros campos[4].

Haciendo frente a la escasez de científicos de datos

La analítica de datos requiere conocimientos técnicos sofisticados, personificados en el rol del "científico de datos" (*data scientist)* que reúne conocimientos computacionales (para adquirir y procesar datos), analíticos (para implementar técnicas estadísticas y de aprendizaje automático) e industriales (comprende la realidad de un sector, de sus fuentes de datos y sabe cuáles son las preguntas más importantes) .

Esta fusión de conocimientos y habilidades es difícil de encontrar en un solo individuo, lo que ha resultado en un "déficit de capacidades" en el mercado laboral (Mateos-García *et al*., 2014b). En Nesta, hemos buscado superar este déficit a

4. La razón principal por la que decidimos desarrollar esta capacidad internamente en vez de trabajar con otras organizaciones externas es que encontramos un "hueco" en este mercado. Aunque hay agencias comerciales que utilizan metodologías de datos masivos para el análisis de empresas, con clientes en funciones de *marketing*, estrategia e inversión, su modelo de negocio, y, de forma particular, la opacidad de sus algoritmos, los hacen difíciles de utilizar en la toma de decisiones públicas en las que hay que ser capaces de justificar los resultados y las recomendaciones. Aunque hay muchos investigadores académicos trabajando en esta área, son o bien expertos en estudios de innovación que utilizan fuentes de datos tradicionales, o bien ingenieros informáticos con conocimiento limitado del campo de políticas de innovación.

través de una estrategia que combina el desarrollo profesional de personal dentro de la organización que ya sabía utilizar técnicas de análisis cuantitativo y conoce el campo de políticas de innovación, con el reclutamiento de personal externo. Hemos hecho esto aprovechando los siguientes recursos:

- Colaboración con otras organizaciones: en las fases iniciales de nuestra trayectoria, trabajamos con otras organizaciones en proyectos colaborativos que nos permitieron aumentar nuestra "capacidad de absorción" en diferentes áreas de la analítica de datos. Por ejemplo, ofrecimos becas a varias agencias comerciales e investigadores para que exploraran preguntas de interés para nosotros utilizando nuevos métodos, y colaboramos en proyectos con otras agencias en las cuales ellas recogieron los datos, y nosotros los analizamos (Collins *et al*., 2016). Esto nos permitió comprender mejor las opciones y herramientas disponibles, y a aprender nuevas habilidades y métodos de manera modular, colaborando con otras organizaciones con más experiencia que nosotros.
- Herramientas abiertas: muchas herramientas de analítica de datos son desarrolladas utilizando modelos de "código abierto" por investigadores académicos y empresas tecnológicas, y están disponibles de manera gratuita (menciono algunas de ellas en el apéndice). Con estas herramientas, hemos podido acceder a métodos avanzados con costes muy bajos. La contrapartida de esta poderosa funcionalidad es que estas herramientas son difíciles de utilizar: frecuentemente, requieren la escritura de *software* utilizando lenguajes de programación (una ventaja es que una vez que se ha escrito el código para un proyecto, es posible reutilizarlo o adaptarlo en el futuro, lo cual posibilita la automatización de tareas e incrementa la eficiencia).
- Infraestructuras en la nube: en algunos de nuestros experimentos (aunque no todos), hemos recopilado fuentes de datos demasiado grandes para almacenar en un solo computador, o tenido que paralelizar el análisis para hacerlo más rápido. Para satisfacer estas necesidades, hemos utilizado soluciones "en la nube" en vez de modificar nuestra infraestructura de información y comunicación. Esto nos ha permitido acceder a las capacidades tecnológicas necesarias de manera flexible y a un bajo coste.
- Programas de entrenamiento: hay una creciente oferta de cursos de corta y media duración centrados en la utilización de distintas herramientas y métodos analíticos. Estos programas de entrenamiento pueden servir de manera introductoria, o para afianzar conocimientos adquiridos a través del trabajo en proyectos aplicados.
- Plataformas digitales: también ha habido rápido crecimiento en la oferta de cursos *online* y MOOCs (*Massively Open Online Courses* o Cursos Web Masivos)

acerca de la ciencia de datos, muchos de ellos disponibles de manera gratuita, o a un precio bajo en comparación con las versiones presenciales. En el apéndice menciono algunos de los cursos que hemos utilizado para aprender nuevos métodos y herramientas analíticas.

El manejo de proyectos

Un experimento analítico tiene varios objetivos: testar el potencial de nuevas fuentes de datos y herramientas para hacer frente a retos organizacionales (en el caso de Nesta, mejorar la cantidad y calidad de la evidencia disponible para la política de innovación), generar resultados que ayudan a comunicar ese valor dentro de la organización y adquirir nuevas capacidades analíticas. Su análogo en el sector privado es la idea del "producto mínimo viable" en el paradigma del *lean start-up* (Eisenmann *et al*., 2012).

El reto es explorar estas nuevas posibilidades a la vez que se reduce el coste del fracaso. Los proyectos tienen dos componentes de riesgo: uno técnico (¿podemos adquirir los datos y analizarlos?) y otro de aplicación (¿son los resultados útiles o interesantes?). Un riesgo importante en estos experimentos es que el analista se "atasque" con una tarea de bajo valor y termine dedicándole demasiado tiempo.

Estas son las prácticas que hemos desarrollado para administrar nuestros experimentos analíticos:

- Preguntas claras: siempre buscamos definir de manera clara la pregunta de interés, dado que esto ayuda a concentrar la atención en aquellas actividades que crean más valor. Esta pregunta debe basarse en los intereses y necesidades de los usuarios finales —en nuestro caso, políticos de innovación—.
- Planificación realista: la planificación del experimento tiene en cuenta si este requiere la utilización de nuevas fuentes de datos, y el aprendizaje de nuevas técnicas, o, por el contrario, es más incremental (en cuyo caso los riesgos técnicos son más bajos, y hay más tiempo para el análisis y la presentación).
- Revisiones periódicas: intentamos identificar fases del experimento donde será posible presentar resultados intermedios a otros miembros del equipo, incluyendo analistas que pueden criticar métodos y ofrecer nuevas ideas, y audiencias con conocimiento del sector, que ayudan a interpretar los resultados e identificar posibles problemas con los datos. Este modelo, inspirado en la idea de la revisión por pares, ayuda a aprovechar la variedad

de conocimientos y habilidades en la organización, y ayuda a crear una sensación de comunidad analítica que beneficia la moral y retención de personal.

- Código y datos abiertos: En nuestros proyectos utilizamos Github, una plataforma de desarrollo de *software* colaborativa, para almacenar el código que creamos durante el proyecto y coordinar nuestro trabajo. Siempre que podemos, publicamos tanto el código como los datos que hemos utilizado para que otros usuarios los utilicen. Un beneficio indirecto de esta práctica es que crea incentivos para documentar adecuadamente todo nuestro análisis, lo cual facilita su futuro reuso.
- Publicación de resultados: Siempre publicamos nuestros resultados a través de informes, blogs y otros formatos interactivos. Esto nos ha ayudado a incrementar nuestra visibilidad y reputación, creando nuevas oportunidades de colaboración y atrayendo talento interesado en trabajar con nosotros.

Resultados provisionales

La estrategia experimental que he esbozado en esta sección nos ha permitido pasar de una situación inicial en la que solo utilizábamos fuentes de datos, métodos de análisis y formatos de comunicación presente, a una situación presente en la que tenemos un equipo con cinco científicos de datos que llevan a cabo ambiciosos proyectos de análisis utilizando fuentes de datos masivas, minería de datos y aprendizaje automático, y publicando resultados a través de visualizaciones interactivas y tableros de datos.

Entre los proyectos que hemos realizado, cabe destacar *A Map of the UK Video Games Industry*, un mapeo del sector de videojuegos británico utilizando fuentes de datos web que nos ha permitido demostrar la infraestimación del sector en los datos oficiales, y a identificar nuevos *clusters* de actividad (Mateos-Garcia *et al*., 2014a), *Tech Nation 2016,* un análisis de la economía digital británica combinando fuentes de datos oficiales, abiertos y digitales, y *The Geography of Creativity in the UK*, un análisis de la geografía de las industrias creativas en el Reino Unido acompañada de una visualización interactiva y una nueva fuente de datos abierta que está siendo utilizada para analizar y apoyar al sector en diferentes partes del Reino Unido (Mateos-Garcia and Bakhshi, 2016). Hemos publicado también una serie de visualizaciones interactivas que han recibido miles de visitantes y nos permiten comunicar resultados complejos —como por ejemplo, el rol de la BBC en el fortalecimiento de redes de innovación en el Reino Unido— a audiencias sin conocimientos técnicos.

En estos momentos, estamos trabajando es el desarrollo de un tablero de datos interactivo que fusionará y analizará información acerca de la situación y evolución de la innovación en Galés, para facilitar la toma de decisiones gubernamentales. Aunque este tablero no será lanzado hasta verano de 2017, hemos publicado en nuestro sitio web, y en GitHub, resultados preliminares y código analítico de 4 experimentos en los que hemos explorado distintas fuentes de datos, métodos analíticos y formatos de visualización —incluyendo análisis de redes de colaboración entre científicos, medición de la emergencia de nuevas tecnologías a través de las reuniones informales de sus comunidades, y un modelo predictivo de la futura especialización de economías locales basadas en su situación actual—.

Estrategias alternativas

Otros equipos en Nesta han utilizado estrategias distintas para llevar a cabo experimentos analíticos sin tener que adquirir una capacidad técnica interna, por ejemplo a través de "hackatones" en los cuales se reúnen a científicos de datos para responder preguntas de interés.

En una colaboración con el Open Data Institute, el equipo de Challenge Prizes de Nesta ha creado un proceso estructurado para organizar retos con datos abiertos (*Open Challenge Prizes*) que buscan incentivar el desarrollo de aplicaciones comerciales que utilizan fuentes de datos abiertas en áreas como la Salud, la Educación, El Desempleo o el Arte y la Cultura (Parkes and Philips, 2016). Este proceso incluye frases de exploración de preguntas, un "fin de semana de creación" en el cual se reúnen distintos equipos para producir prototipos y presentarlos a un panel de expertos, y una fase de incubación donde se desarrollan los mejores prototipos prestándole especial atención a sus modelos de negocio.

Otro modelo que hemos utilizado es la inserción de científicos de datos en organizaciones sin ánimo de lucro para ayudarles a resolver problemas internos importantes. En el proyecto *Data For Good* (Datos por el bien común), voluntarios de DataKind, una organización filantrópica que pone a científicos de datos a la disposición de ONG, analizaron fuentes de bases desestructuradas en el Citizen Advice Bureau (Oficina de Atención al Ciudadano) para identificar nuevos problemas sociales y su evolución (Baeck, 2015). En *Arts Data Impact* (Impacto de Datos en el Arte), un proyecto llevado a cabo en colaboración con Arts Council England y el Arts and Humanity Research Council, buscamos promover una mejor utilización de los datos en estrategias de *marketing* en diferentes organizaciones del arte y la cultura, insertando científicos de datos en organizaciones como la English National Opera o el National Theatre.

Conclusiones

Nuestra experimentación con la analítica de innovación nos ha ayudado a desarrollar una capacidad importante, ha demostrado el valor de sus métodos y ha incrementado nuestra visibilidad y reputación, ayudándonos a atraer colaboradores y talento.

Pienso que nuestro proceso contiene lecciones útiles para otras organizaciones planteándose la posibilidad de desarrollar su capacidad analítica:

- Realizar experimentos pequeños centrados en preguntas candentes.
- Invertir en personal interno que conoce cuáles son esas preguntas, y contratar a científicos de datos externos.
- Utilizar estrategias de innovación abierta, tanto en los *inputs* (herramientas de innovación e infraestructuras en la nube) como en los *outputs* (publicación de resultados y datos).

También hemos aprendido acerca de las limitaciones de estos nuevos métodos, que requieren la triangulación con fuentes de datos diversas, y la colaboración con expertos con conocimiento del campo siendo estudiada, para interpretarlos.

Es importante señalar que en este proceso la idea de "big data" ha sido para nosotros un destino más que un punto de comienzo: hay muchas oportunidades para crear valor en el sector público a través de los datos y no todas requieren datos masivos. Es importante evitar una obsesión con términos de moda, o proyectos faraónicos sin necesidades claras —esa es la razón por la cual preferimos el término "analítica de datos", que incluye datos masivos sin estar limitado a ellos—.

Otra pregunta importante se refiere a las transformaciones organizacionales que deben realizar las agencias de innovación si quieren crear valor con estas nuevas fuentes de datos y métodos: si bien podemos generar información acerca de la innovación en tiempo real, ¿es esta información necesaria? ¿Existen en las agencias de apoyo a la innovación los procesos y modelos de intervención necesarios para aprovecharlas?

Un asunto relacionado es el de las infraestructuras digitales para procesamiento y almacenamiento de datos. Aunque por el momento nuestro plan es seguir utilizando infraestructuras en la nube, esta opción no es viable en muchas otras partes del sector público, por razones de seguridad y privacidad. Una estrategia

experimental también será útil para ellas. Organizaciones como el Banco de Inglaterra o la Office for National Statistics han creado *data labs* que les permite experimentar con nuevas tecnologías, y el conocimiento generado les ayuda a decidir qué inversiones a realizar de manera mejor informada. Esta es la principal ventaja de una estrategia experimental sobre grandes proyectos *top down,* que corren el riesgo de terminar generando infraestructuras caras e inútiles, soluciones big data en búsqueda de un problema para resolver.

Apéndice: Herramientas y plataformas

Función	Herramientas/Plataformas/Soluciones abiertas
Almacenamiento y procesamiento de datos	Bases de datos estructuradas (MySQL, PostgreSQL) y sin estructurar (MongoDB). Infraestructuras de procesamiento de datos en la nube (Amazon Web Services)
Análisis de datos	Aplicaciones analíticas (R y Python)
Colaboración interna	Plataformas de desarrollo de *software* colaborativo (GitHub) y reproducibilidad de resultados (Jupyter Notebooks)
Visualización	Herramientas para visualización interactiva (D3.js), Mapeo geográfico (Leaflet.js) y mapeo de redes sociales (Gephi)
Aprendizaje	Cursos *online* (John Hopkins Bloomsberg School of Health Data Science Specialisation, Coursera, Codeacademy), plataformas de aprendizaje interactivo (Datacamp), Sitios web de Preguntas y Respuestas (Stack Overflow, Stack Exchange, Quora)

Referencias bibliográficas

Baeck, P. (2015). *Data for Good: How big and open data can be used for the common good*. Nesta, London.

Bakhshi, H., Bravo-Biosca, A., Mateos-García, J. (2014). *Inside the Datavores: Estimating The Effect Of Data And Online Analytics On Firm Performance*. Nesta, London.

Bakhshi, H., Mateos-García, J. (2012). *Rise of the Datavores*. Nesta, London.

Collins, L., Marston, L., Westlake, S. (2016). Big Data for better innovation policy | Nesta [WWW Document]. URL http: //www.nesta.org.uk/blog/big-data-better-innovation-policy (accessed 9.14.16).

Corrado, C. A., Haskel, J., Iommi, M., Jona-Lasinio, C. (2012). Intangible Capital and Growth in Advanced Economies: Measurement and Comparative Results (SSRN Scholarly Paper No. ID 2153512). Social Science Research Network, Rochester, NY.

Edler, J., Cunningham, P., Gök, A., Shapira, P. (2013). *Impacts of Innovation Policy: Synthesis and Conclusions.*

Eisenmann, T. R., Ries, E., Dillard, S. (2012). *Hypothesis-Driven Entrepreneurship: The Lean Startup* (SSRN Scholarly Paper No. ID 2037237). Social Science Research Network, Rochester, NY.

Fagerberg, J., Mowery, D. C., Nelson, R. R. (2006). *The Oxford handbook of innovation*. Oxford Handbooks Online.

Gordon, R. J. (2016). *The Rise and Fall of American Growth: The U.S. Standard of Living since the Civil War.* Princeton University Press.

Harris, D. M., Halkett, R. (2007). *Hidden Innovation: How innovation happens in six low innovation sectors*. NESTA.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A. H. (2011). *Big data: The next frontier for innovation, competition, and productivity*.

Mateos-García, J., Bakhshi, H. (2016). *Geography of Creativity 2016*. Nesta, London.

Mateos-García, J., Bakhshi, H., Lenel, M. (2014ª). *A Map of the UK Games Industry*. Nesta, London.

Mateos-García, J., Bakhshi, H., Windsor, G. (2014b). *Skills of the Datavores: Talent and the data revolution* | Nesta [WWW Document]. URL http: //www.nesta.org. uk/publications/skills-datavores-talent-and-data-revolution (accessed 7.22.16).

Mayer-Schönberger, V., Cukier, K. (2013). *Big Data: A Revolution that Will Transform how We Live, Work and Think.* Houghton Mifflin Harcourt.

Nesta (2016). *Innovation Analytics: A guide to new data and measurement in innovation policy* | Nesta. Nesta, London.

OECD (2014). *Data-driven Innovation for Growth and Well-being*. Interim Synthesis Report. OECD, Paris.

Parkes, E., Philips, B. (2016). *The Open Data Challenge Series Handbook* | Nesta. Nesta and ODI.

Solow, R. M., 1957. "Technical Change and the Aggregate Production Function". *Rev. Econ. Stat*., 39, 312-320. doi: 10.2307/1926047

Capítulo 20

Crear valor con *big data* privado en la Administración pública. Experiencias concretas

Richard Benjamins*, Elena Díaz Sánchez*, Tomas Trenor Escuin*, Pedro de Alarcón*, Javier Carro Calabor* y Florence Jane Broderick*

Introducción

La aplicación del *big data* a la Administración pública muchas veces se basa en datos públicos y datos abiertos. Pero también existen datos generados por el sector privado que pueden aportar mucho valor para los bienes públicos. Algunos sectores privados, por los servicios que ofrecen a sus clientes, generan muchos datos, simplemente como "efecto secundario" de su operación. Ejemplos son los bancos que tienen datos de uso de medios de pago y transferencias financieras, las empresas de telecomunicaciones que tienen datos de actividad de móviles y los supermercados que tienen datos sobre consumo y precios. En la época de *big data*, estos datos de operación se están convirtiendo en un activo muy valioso ya que revelan el comportamiento y el perfil de poblaciones en ciertos ámbitos. Por ejemplo, visualizamos la actividad de nuestra red móvil sobre un mapa justo antes, durante y después de un terremoto en Oaxaca, México (2012)[1]. Como se puede apreciar en las dos figuras abajo, se nota un incremento importante de actividad en la red (provocado por llamadas y SMS) durante el impacto.

* LUCA Data-Driven Decisions Telefónica.

1. Moumny, Y., Frías-Martínez, V., Frías-Martínez, E. (2013).

Esta visualización tiene un valor importante para la Administración pública y otras organizaciones humanitarias para comprender mejor los impactos de desastres naturales como terremotos, inundaciones, etc., y coordinar la respuesta temprana teniendo en cuenta los datos dinámicos de la actividad de población.

En este capítulo enseñamos algunos proyectos concretos que hemos ejecutado desde LUCA, la unidad de *big data* de Telefónica en el ámbito de la Administración pública. Los ámbitos de los proyectos son transporte, turismo, e impacto ambiental de tráfico. Para cada proyecto explicamos:

- El problema de la Administración Pública.
- Cuál era la aproximación antes.
- La propuesta de valor del *big data*.
- La solución.
- El resultado y los beneficios del proyecto.

Pero antes de entrar en los proyectos, vamos a explicar cómo tratamos los datos para que puedan generar estos *insights* relevantes, al mismo tiempo que respetar hasta el máximo la seguridad y la privacidad de los datos de nuestros clientes.

Del dato en crudo al dato de valor

Como casi siempre en las aplicaciones de *big data*, para relacionar los datos con eventos reales, usamos *proxies*" que son aproximaciones o indicadores cuantitativos que presentan un fuerte vínculo con la realidad. Por ejemplo, un "click" en un anuncio de publicad es un *proxy* para el interés del usuario; o una búsqueda

en Google para un coche es un *proxi* para una intención de compra de un coche. Igualmente, en nuestro caso, la actividad que generan los teléfonos móviles en nuestras antenas es un *proxy* para la actividad humana.

La disponibilidad de abundantes *proxies* que midan la actividad de los clientes es fundamental para convertir cualquier negocio en una actividad *data-driven* o lo que es lo mismo, orientada por los datos y el análisis de los mismos. En la era digital gran parte de la actividad humana se encuentra ya digitalizada (comunicaciones, compras, entretenimiento...) e incluso son las propias cosas las que están conectadas (IoT, "Internet of Things") y están generando datos de manera exponencial. Esto posibilita a las iniciativas *big data* analizar toda la información y retornarla a la sociedad en forma de valor económico, y por qué no, valor para las Administraciones públicas.

Nuestra plataforma de *big data* se alimenta de los eventos generados por las líneas móviles, por tanto, nos muestra el comportamiento real de los móviles, que usamos como un *proxy* de grupos de personas. Todos los eventos generados en la red móvil están georreferenciados, y tienen asociado un *timestamp*, lo que nos permite, analizar y estimar la actividad de grupos de personas, en diferentes localizaciones geográficas y a lo largo de diferentes ventanas temporales.

El siguiente gráfico nos muestra las etapas por las que pasan los eventos desde que se generan hasta el momento en el que están listos para la toma de decisiones:

1. **Extracción:** a través de sondas puestas en la red, diariamente se recogen los datos generados a lo largo de toda la geografía española.
2. **Anonimización:** una vez extraídos los datos, el siguiente paso es la anonimización de los mismos, y posterior almacenamiento en nuestra infraestructura *big data*. La anonimización se refiere al proceso de eliminación, codificación y alteración de datos personales por el cual la identidad de una

persona o entidad queda totalmente oculta. Es importante aclarar que la anonimización es un proceso básico en el tratamiento de datos que contengan información personal pero es necesario aplicar restricciones adicionales para asegurar aún más la identidad de las personas. Por ejemplo, no permitiendo agregaciones o filtros de los datos que impliquen a una cantidad menor de X individuos. Los algoritmos usados en el proceso de anonimización tienen dos características principales:

a) No nos permiten dar marcha atrás o hacer decodificación, por lo que una vez que los datos han llegado a nuestra infraestructura no podemos saber la identidad del usuario que los ha generado.
b) Los identificadores asociados a cada línea son persistentes a lo largo del tiempo.

3. Extrapolación: una vez anonimizados, los usuarios se extrapolan a la totalidad de la población de modos que los *insights* que generamos a partir de ellos harán referencia a la totalidad de la población y no solo a los usuarios que se conectan a nuestra red.
4. Agregación: finalmente antes de generar los ficheros o informes a los usuarios finales, los datos se agregan en base a diferentes variables sociodemográficas definidas según el caso de uso con el que trabajemos. En base a las restricciones impuestas por la Agencia Estatal de Protección de datos, en España nunca se podrá hacer referencia a comportamientos asociados a menos de 15 individuos.

El resultado de todo este procesamiento de los eventos de red que generan los móviles son *insights* (conocimiento profundo) relacionados con la movilidad de la población de un determinado país, ya que contamos con un número suficiente de clientes (muestra) para generar extrapolaciones a una determinada población. A partir de los datos crudos se aplican técnicas analíticas que nos permiten caracterizar asociar a las personas la permanencia o **estancia** (*dwell,* en inglés) en ciertas áreas geográficas, **puntos de interés** ("POI") y **desplazamientos** (*journey*). Si la localización de los móviles no cambia durante un periodo determinado, se considera que se trata de estancias. Si la localización cambia en menos de este periodo, y esto se repite varias veces, se considera que se trata de un desplazamiento. Estos dos tipos de *insights* tienen valor, y en los siguientes casos de usos explicamos cómo crean valor para distintas Administraciones públicas. El producto que forma la base de los análisis con *big data* es *smart steps* (https://luca-d3.com/business-insights/index.html), producto de LUCA (luca-d3.com), la unidad de *big data* B2B de Telefónica.

La Fiesta de las Flores

Introducción

La Fiesta de las Flores es una de las principales fiestas celebradas en el municipio de Girona. Constituye una de las fiestas primaverales más espectaculares del país, donde todos los rincones de Girona se llenan de composiciones florales, paralelamente se desarrollan concursos de fotografía y escaparates.

La Fiesta de las Flores comenzó a celebrarse en 1955 como una simple exposición de flores; a lo largo de los años la muestra ha evolucionado hasta convertirse en un evento que atrae cada año a un mayor número de visitantes tanto nacionales como internacionales

Cuál era la aproximación antes

Hasta 2014, el Ayuntamiento de Girona estimaba el número de visitantes usando técnicas tradicionales de investigación de mercados, que en este caso combinaba el uso de contadores con la realización de encuestas presenciales a los asistentes.

Debido a las dimensiones temporal y geográfica del evento, la realización de estos estudios es muy complicada y costosa. La fiesta tiene una duración de 9 días, y se desarrolla en más de 140 espacios ubicados en todo el municipio, por tanto, se tiene que realizar un muestreo de población y territorial, con el consiguiente aumento tanto del error asociado al estudio, como el coste de campo.

La solución y propuesta de valor del *big data*

Smart Steps Turismo permite usar técnicas de *big data* para estimar y segmentar el número de visitantes en diferentes espacios tanto temporales como geográficos, analizando el comportamiento real de los visitantes. Comparado con las técnicas tradicionales, nos permite

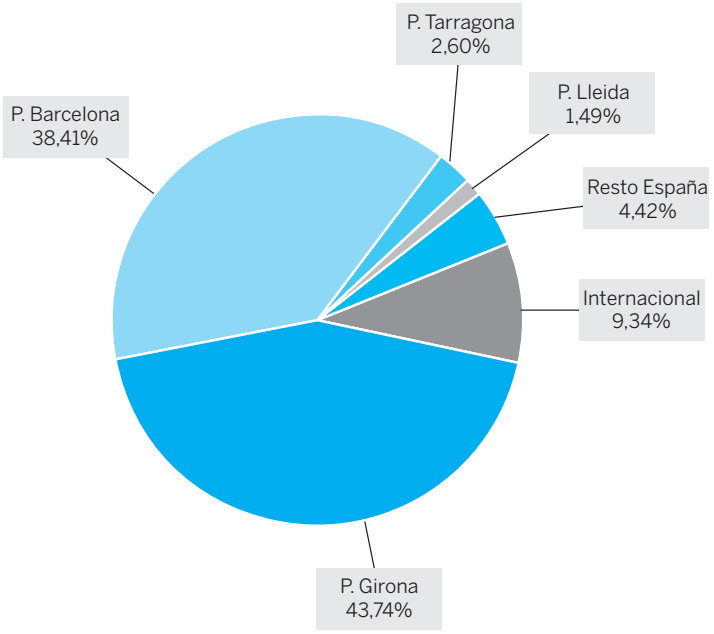
- Reducir el error: aumentamos de manera muy significativa el tamaño muestral, de modo que el error cometido tiende a cero.
- Eliminamos el sesgo asociado a:
  - Errores cometidos por los entrevistadores o fallos de los contadores.
  - Errores en las declaraciones de los entrevistados.

- Analizamos el movimiento de los visitantes a lo largo de todo el municipio, por lo que:
  - Eliminamos el muestreo territorial.
  - Podemos identificar las áreas que mayor atracción generan.

Los resultados y beneficios obtenidos

Los resultados del proyecto permiten al Ayuntamiento de Girona:

- Cuantificar el número de visitantes que han acudido al evento.

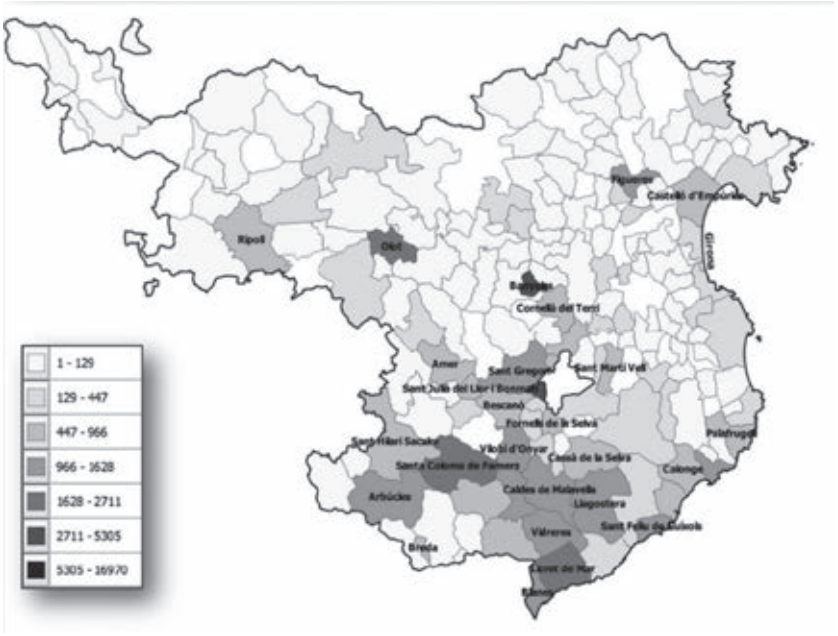


- Identificar la procedencia de los mismos con diferentes niveles de agregación:
  - País de origen para los internacionales.

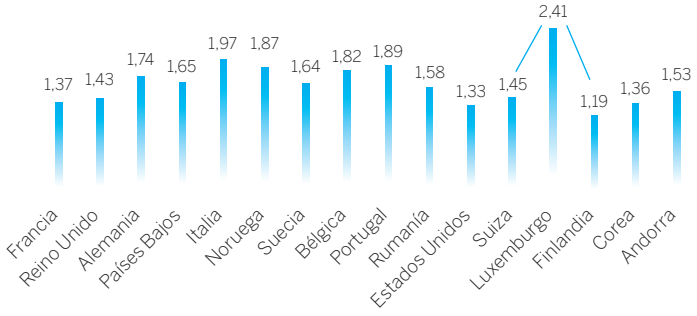
País	Visitantes		Estancia media
	Nº	%	
Francia	10.312	43,37%	1,12
Países Bajos	2.443	10,27%	1,30
Alemania	1.900	7,99%	1,32
Reino Unido	1.568	6,59%	1,57
Bélgica	1.022	4,30%	1,22
Italia	683	2,87%	1,37
Estados Unidos	670	2,82%	1,43
Polonia	550	2,31%	1,18
Rusia	431	1,81%	1,05
Andorra	306	1,29%	1,22
Suiza	299	1,26%	1,21
Resto nacionalidades	**3.595**	**15,12%**	**1,33**

– Provincia de origen para los visitantes nacionales.

– Municipio de origen para los visitantes locales.



- Estudiar la estancia media en días en el municipio.

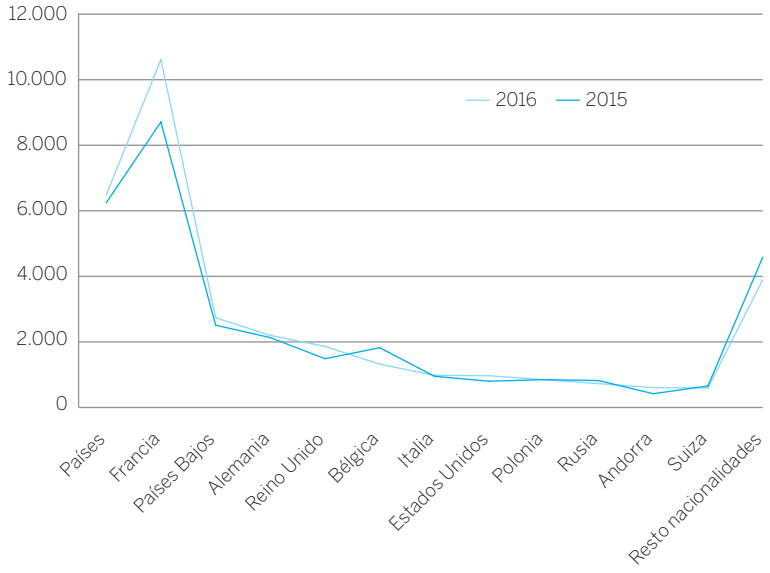


- Generar mapas de calor, a partir de los cuales se pueden identificar las diferentes áreas dentro del recorrido que mayor interés generan.

- Ver los lugares de más concentración por tipología de visitante, en el siguiente mapa mostramos las zonas en las que más se concentran los visitantes nacionales.

- Analizar el impacto de las campañas de *marketing* realizadas a través de la variación de visitantes por año y...

  - País de origen para los internacionales.
  - Provincia de origen para los nacionales.
  - Municipio de origen para los locales.



Países	2016	2015	Índice
Francia	43,37%	44,83%	97
Países Bajos	10,27%	9,30%	110
Alemania	7,99%	7,71%	104
Reino Unido	6,59%	5,00%	132
Bélgica	4,30%	6,41%	67
Italia	2,87%	2,75%	104
Estados Unidos	2,82%	2,11%	133
Polonia	2,31%	2,32%	99
Rusia	1,81%	2,20%	82
Andorra	1,29%	0,51%	250
Suiza	1,26%	1,51%	83
Resto nacionalidades	**15,12%**	**15,34%**	**99**

› Dirección General de Tráfico, España

Introducción

Entre otras competencias, la Dirección General de Tráfico (en adelante, DGT), dependiente del Ministerio de Interior del gobierno de España, se encarga de la

regulación, gestión y control del tráfico en vías interurbanas y travesías, así como de la planificación, dirección y coordinación de las instalaciones y tecnologías para el control, regulación, vigilancia y mejora de la seguridad vial.

El periodo 2011-2020 ha sido designado por Naciones Unidas como la "Década de la Acción para la Seguridad Vial". Esta "Década de la Acción" retaba a los países suscribientes a disminuir en un 50% el número de fallecidos en el mundo para 2020, un reto ya previamente lanzado por la Unión Europea en la década 2000-2010 y reiterado en 2011-2020.

A nivel del gobierno de España, el objetivo cero desarrollado es el relativo a la seguridad vial que se formuló a través del Plan Estratégico de Seguridad Vial 2011-2020.

Por lo que respecta a la recogida de datos sobre accidentes, sus consecuencias (víctimas mortales, heridos, seguimiento de los accidentes) y su publicación y proyección internacional, la DGT incluye todos los datos ocurridos en todo el territorio estatal, bien sean accidentes en vía urbana o interurbana, bien sean fallecidos hasta 30 días después del accidente.

Dentro de todas las acciones definidas, se recoge la de creación de un índice de riesgo de accidentalidad. Este índice permitirá asociar a una carretera, tramo o área determinada un nivel de riesgo de accidentalidad. Para la creación de dicho índice se requiere, entre otras variables, el número de KMs que recorren los vehículos en toda España y el origen y destino de los desplazamientos.

Para la DGT es un dato difícil de estimar debido principalmente a dos razones:

- En una red viaria del tamaño de la española es técnica y económicamente inabarcable conocer el volumen de tráfico en todos los puntos de la misma. Esto es, no es posible instalar y mantener estaciones de toma de datos (ETD) u otros dispositivos de conteo de vehículos en toda la red viaria y, por lo tanto, solo se instalan en los puntos más importantes de la misma.
- Aun conociendo el volumen de tráfico en los puntos más importantes de la red viaria, la DGT desconoce la distancia total de cada desplazamiento (desde su origen a su destino).

Cuál era la aproximación antes

Antes de la aparición de técnicas como el *big data*, este tipo de información se obtenía fusionando la información proveniente de ETDs con la realización de

encuestas o trabajos de campo. El principal problema de esta metodología es el elevado coste que implica realizar una encuesta, con suficiente base muestra, en todo el territorio español. Por este motivo, este tipo de trabajos se han realizado en las últimas décadas con muy poca frecuencia, lo que no permite evaluar las medidas que se han ido tomando al estar la información desactualizada.

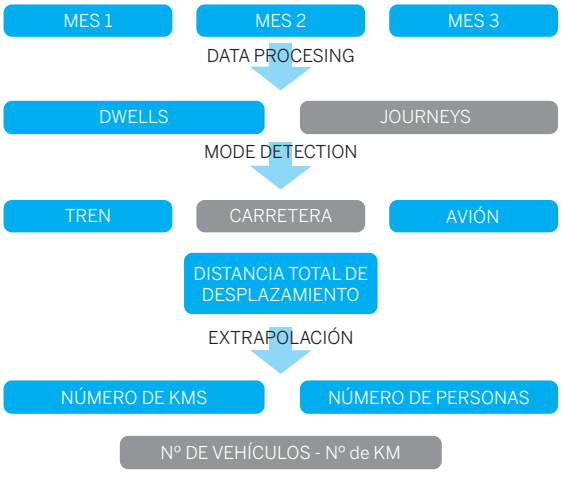
Se requiere, por lo tanto, una nueva metodología que permita obtener este tipo de información a un menor coste, con calidad asegurada, y que se puede aplicar de manera periódica tener la información siempre actualizada.

La solución y propuesta de valor del *big data*

Smart Steps, mediante tecnología y técnicas *big data*, recopila, anonimiza y agrega los eventos que generan los teléfonos móviles en la red de Telefónica, para, entre otras cosas, analizar cómo se mueven las personas en las ciudades o en el país (independientemente de la zona del país o de la carretera que utilizan). Recordar que se analizan las tendencias y los comportamientos de las multitudes, nunca de individuos.

Dado que Smart Steps recopila datos de todo el territorio español y tiene una muestra representativa de toda la población ha podido ayudar a la DGT a realizar una mejor estimación del número de KMs que recorren los usuarios en toda España y, también, de sus orígenes y destinos.

La metodología que se aplicó en este proyecto se resume, *grosso modo*, en la siguiente figura:



En el caso de la DGT se analizaron tres meses distintos para detectar la estacionalidad que existe en cuanto a la movilidad. Gracias a la existencia de este tipo de tecnologías, es ahora posible el análisis de esa cantidad información (30 billones de eventos aproximadamente por mes) y además aplicar técnicas *big data* que nos permitan, por ejemplo, inferir el modo de transporte de un desplazamiento, a través de los eventos que generan los teléfonos móviles.

Los resultados y beneficios obtenidos

Los resultados del proyecto han permitido a la DGT tener una **visión completa** (esto es, con independencia de la zona del país o de la carretera utilizada) de la movilidad en España y poder obtener una estimación **actualizada y precisa** del número de km que realizan los vehículos en diferentes periodos del año y, además, conocer sus orígenes y destino. A partir de este dato pueden obtener un mejor índice de riesgo de accidentalidad que les ayudará a evaluar y tomar medidas, en cuanto a la seguridad vial, basadas en los datos.

Entendiendo el tráfico de trabajo (*commuters*) para reducir la contaminación en grandes ciudades

El problema de las grandes ciudades

Las grandes ciudades y sus alrededores generan mucho tráfico debido al desplazamiento que hace la gente para ir al trabajo. Esto no solo genera atascos y la correspondiente ineficiencia en el aprovechamiento del tiempo, sino también puntualmente genera contaminación por encima de los niveles permitidos en puntos concretos. La calidad del aire es un gran desafío para la gestión de grandes áreas urbanas. La Organización Mundial de la Salud advierte del serio riesgo para la salud que constituye la polución del aire, dado que, en 2014, el 92% de la población mundial vivía en zonas donde los criterios de calidad del aire establecidos por la OMS no se cumplían.

La reducción del tráfico de vehículos a motor para mejorar la calidad del aire está siendo un auténtico problema para los gobiernos locales. La mayoría de ellos cuentan con sistemas avanzados para monitorizar un amplio rango de gases contaminantes de forma continua. Uno de ellos es el dióxido de nitrógeno (NO_2). La presencia de NO_2 está directamente asociada a la densidad

de tráfico de vehículos a motor, así como a las condiciones meteorológicas (la ausencia de lluvia y viento empeoran la situación).

En muchas de las grandes metrópolis existen medidas de emergencia (que impactan mucho en la población) para limitar el tráfico drásticamente cuando los niveles de contaminación se disparan, especialmente referidas a los niveles de NO_2. Conocer y entender cuáles son las horas punta, desde y hacia donde se mueve la gente en dichas horas es el primer paso hacia una posible optimización de las medidas a adoptar. Cruzar estos datos con datos de la calidad del aire permite dar todavía un paso más para generar conocimiento para la toma de decisiones.

Cuál era la aproximación antes

El gran caballo de batalla para la contabilización del tráfico es, como ocurre en muchas ocasiones, la adquisición fiable de los datos. La aproximación más tradicional en la adquisición de estos datos es la utilización de estaciones de medición de densidad de tráfico y la realización de encuestas. La utilización de sensores de medición de tráfico presentan la limitación en cuanto a su número (normalmente situados en vías principales o puntos conflictivos). En lo que se refiere a encuestas, tal y como se comentó anteriormente en el caso de uso de DGT, han tenido un valor bastante alto ya que los procedimientos para la preparación, realización y procesamiento estadístico posterior de las encuestas se han conseguido perfeccionar en gran medida a lo largo del tiempo. Sin embargo, la mayor carencia de las encuestas es su disponibilidad temporal, ya que habitualmente se necesita mucho tiempo desde que se elabora hasta que se obtienen los datos finales, con lo cual los resultados suelen llegar un tiempo después del momento en el que realmente tienen validez.

Se hace necesario entonces encontrar la forma de contabilizar el tráfico real y, a ser posible, en el mismo momento en que sucede. Los sensores que se colocan en las carreteras, las cámaras que muestran el tráfico en tiempo real, etc., ayudan a ello, pero también es cierto que el coste de estas instalaciones para las Administraciones públicas sería altísimo si se desea lograr un nivel de detalle suficiente sobre el tráfico.

La propuesta de valor del *big data*

Las operadoras de telecomunicaciones cuentan con una infraestructura con cobertura global con un potencial altísimo para casos como el que se está

describiendo: la red móvil. A través de los *insights* generados desde las plataformas de *big data* es posible conocer (de forma anonimizada) las antenas a las que se conectan los móviles en cada momento: cuando su dueño está en su trabajo (durante horas de oficina), en su hogar (durante la noche), haciendo desplazamientos habituales u ocasionales, etc. Incluso se puede inferir qué tipo de transporte (carretera, tren...) están utilizando en sus desplazamientos. Todo ello sin necesidad de que los ciudadanos hagan nada especial con sus teléfonos ya que no es necesaria la instalación de apps, uso de GPS, etc.

Obviamente, esto tiene una aplicación directa sobre la monitorización, incluso en tiempo real, de todo tipo de desplazamientos que cada día hacen los ciudadanos, en especial de los desplazamientos de trabajo. Como todo sistema de medida, tiene cierta limitación, ya que la localización que se realiza a través de las antenas actúa como *proxy* a la localización exacta. Sin embargo, los casos de uso llevados a cabo demuestran su gran fiabilidad cuando se aplican al análisis de desplazamientos de grupos de población.

La solución

Actualmente la plataforma de *big data* de Telefónica procesa sistemáticamente varios tipos de información que se pueden utilizar de forma recurrente, por ejemplo, lugares donde se hacen estancias largas, trayectos tanto habituales como ocasionales, etc. Y también otra información que se puede inferir, como el lugar estimado de trabajo, de residencia, etc. Tal y como se menciona más arriba, este tratamiento genera datos anonimizados y agregados convenientemente para asegurar y aumentar la protección de la privacidad.

A partir de este punto, la aplicación a casos de uso es prácticamente directa, de forma que se pueden obtener resultados útiles y concretos como los que se exponen a continuación.

Ejemplo 1: representación de la distribución de densidad de trabajadores de los ciudadanos de la Comunidad Autónoma de Madrid (dividida por códigos postales). Esto constituye una base para la realización de estudios de ordenación del transporte, del territorio, de infraestructuras empresariales, etc.

Ejemplo 2: representación de la densidad de distribución de los lugares de vivienda de ciudadanos que trabajan en un área específica (código postal). Es una buena herramienta para la valoración de áreas de influencia y, por tanto, para la ordenación urbanística y de infraestructuras.

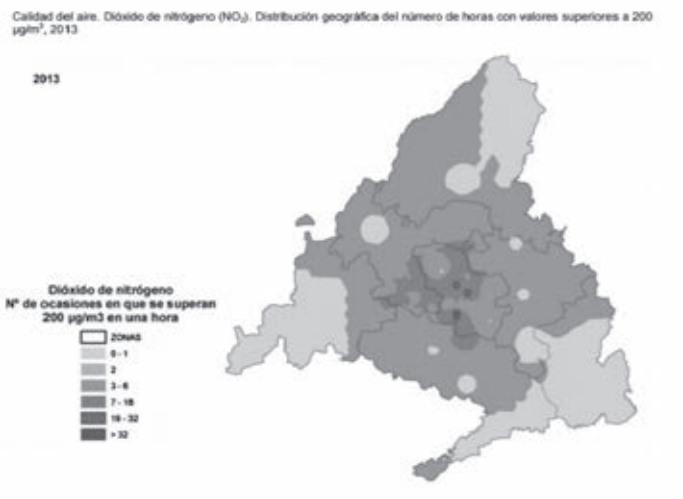
Ejemplo 3: uniendo los dos ejemplos anteriores se obtiene la base fundamental para una herramienta que permita hacer un estudio completo de movilidad, obteniendo por ejemplo áreas fuertemente conectadas debido a los ciudadanos que viajan cada día al trabajo, áreas de influencia más importantes, etc. Las áreas del mismo color corresponden a códigos postales muy conectados en

cuanto a movilidad de trabajadores. Dentro de cada zona, existen códigos postales que registran la mayor de actividad lo que significa que actúan como polos de atracción o zonas de gran afluencia de trabajadores.

A partir de aquí resulta interesante añadir una nueva capa de datos, utilizando datos abiertos de contaminación del aire publicados en el portal de datos abiertos del Ayuntamiento de Madrid. Estos datos están recogidos por 24 estaciones de medida distribuidas en la ciudad Madrid. En la siguiente figura mostramos cómo los niveles más altos de contaminación de NO_2 (representados en color rojo en los puntos, asociados a estaciones de medida) se corresponden visualmente con las zonas de mayor afluencia de trabajadores y, por tanto, mayor tráfico.

Obviamente no es posible disponer de centenas de sensores de medida de contaminación distribuidos por todo el territorio. Por ello, los datos de movilidad pueden ser un buen *proxy* para entender el impacto de la contaminación generada por vehículos a motor, abarcando una escala mucho mayor de la que es posible cubrir con sensores.

Como ejemplo, comparamos el mapa de densidad de trabajadores de toda la Comunidad de Madrid con el mapa de alertas por superación de NO_2 en dicha comunidad (fuente oficial) y puede apreciarse claramente el parecido.



El resultado y los beneficios del proyecto

Este proyecto es una prueba de concepto que nos permite demostrar cómo los datos de movilidad de vehículos (inferidos a partir de los *insights* generados con datos de Smart Steps) pueden apoyar los planes de sostenibilidad y calidad del aire diseñados por las ciudades. Dada la complejidad del problema creemos que la mejor solución pasa por planificar y actuar basándose en datos e *insights* generados desde un *big data* integrador de varias fuentes (movilidad, sensores de contaminación, datos meteorológicos, utilización de transporte público, datos censales...).

Conclusiones

En este capítulo hemos explicado cómo el *big data* de una empresa de telecomunicaciones, solo o en combinación con datos abiertos, puede aportar valor al bien público fomentando transparencia para los ciudadanos y optimizando los recursos públicos. Aplicando algoritmos de *machine learning* y estadística a los datos que pasan por la red móvil de una telco —anonimizados, agregados y extrapolados— es posible generar *insights* sobre movimientos y estancias de poblaciones que se usan para mejorar sectores como turismo, transporte o para tener un mejor control sobre la contaminación del aire en grandes ciudades. El uso del *big data* para resolver los problemas identificados supone un cambio importante en la forma de trabajar de las Administraciones públicas, pero los resultados son mejores y son más económicos.

Queda mucho camino para que todas las Administraciones públicas se hayan transformado para ser organizaciones *data-driven*, y por eso es importante que los decisores de Administración pública en varios países comparten conocimiento y *best practice*. Nuestra tecnología ya ha sido utilizada por las Administraciones públicas en una amplia gama de países, incluyendo España, Reino Unido, Alemania y Brasil, y ahora estamos formando alianzas estratégicas con otros telcos en otros países donde no tenemos red para que las organizaciones de otros países puedan aprovechar de nuestra herramienta de *big data* para tomar decisiones estratégicas.

En cada país, nos enfrentamos a retos diferentes y actitudes distintas hacia el uso de *big data*. Algunos países tienen iniciativas de datos abiertos más desarrollados, y hay muchas diferencias al nivel de legislación, pero en general, la

Administración pública sigue avanzando porque reconoce el valor que puede aportar al ciudadano.

Esperamos que el trabajo descrito aquí sirva como fuente de ejemplos tangibles sobre las implicaciones de ser una organización que toma sus decisiones basadas en los datos.

Agradecimientos

Queremos agradecer a la Dirección General de Tráfico de España y al Ayuntamiento de Girona por su colaboración en los proyectos descritos.

Referencias bibliográficas

Moumny, Y., Frías-Martínez, V., Frías-Martínez, E. (2013). "Characterizing Social Response to Urban Earthquakes using Cell-Phone Network Data: The 2012 Oaxaca Earthquake", Third Workshop on Pervasive Urban Applications @ Pervasive›13, Zurich, Switzerland.

http: //data-speaks.luca-d3.com/2016/11/commuter-traffic-can-big-data-solve.html

http: //data-speaks.luca-d3.com/2016/12/air-quality-how-can-open-data-and. html

Nota al cierre

La extinción de los dinosaurios

Chema Alonso*

No soy un experto en ciencias naturales ni mucho menos, y mis conocimientos sobre los dinosaurios tiene más que ver con los cromos que coleccionaba de niño, las películas de *Parque Jurásico* que vi de adolescente y joven adulto, y los capítulos del *Dinotren* que he visto con mi hija mayor, pero sí sé que se extinguieron porque su hábitat cambió y no pudieron adaptarse. No todos se extinguieron, pero sí los grandes reyes de aquel entonces por no poder adaptarse.

Con las empresas y las organizaciones, la metáfora de la extinción de los dinosaurios se ha utilizado muchas veces, y con la llegada de las empresas tecnológicas veloces, disruptivas y preparadas para el mundo digital, se ha visto cómo las grandes empresas de hace unas décadas han sido sobrepasadas.

El éxito de esas empresas es la adaptación a las nuevas tecnologías que en muchas ocasiones crean ellas mismas. Una empresa las crea, la otra las asimila y las transforma, para luego volver a crear una nueva tecnología que otra de ellas vuelve a asimilar.

El Cloud Computing, el *big data*, los servicios cognitivos y los nuevos paradigmas de desarrollo de *software* pasan de los investigadores a producción con un corto espacio de tiempo. De las pruebas de concepto a la generación de negocio por la vía rápida. Se adaptan rápidamente y prefieren crear ellos mismos las tecnologías para que les sea más fácil la adaptación.

En otras empresas y también en Administraciones públicas no es así. El día a día del negocio y el foco en aspectos menos tecnológicos hacen que la asimilación de las nuevas tecnologías sea lenta y queda cuasidescartado la creación de las mismas. Y esto es el caso del *big data* y su aplicación en las organizaciones.

* Chief Data Officer. Telefónica.

Mientras empresas como Google o Facebook hacen uso del *big data* desde su creación, lo transforman y lo evolucionan —los propios ingenieros de Facebook fueron responsables del nacimiento de Apache Cassandra— otras empresas siguen sin hacer uso y sacar partido de estas tecnologías ya "comoditizadas" para las grandes tecnológicas.

El poder tomar decisiones en base a datos —cuantos más mejor— hace mucho más eficaz el funcionamiento de una organización, y es por eso que las que no lo hacen, son como los dinosaurios. Yo suelo explicar esto a la gente cuando voy del trabajo a casa en coche.

Por supuesto que me conozco la ruta que me lleva de Telefónica a mi casa en la zona Sur de Madrid. Conozco la ruta, y la ruta más corta, pero esa no es la que me garantiza llegar antes. En base a experiencia puedo saber que a determinadas horas es conveniente evitar algunas rutas, debido a que sé que hay autobuses escolares, muchos semáforos o el tráfico es denso, y puede que sea mejor tomar otra ruta.

Sin embargo, nunca hago la ruta sin usar un sistema como Google Maps o Waze que cuenta con un *big data* que está recogiendo en tiempo real los datos del tráfico. Nunca mi conocimiento de la ruta más corta o mi experiencia van a ganar a una decisión tomada en base a datos en tiempo real.

Con las decisiones en las organizaciones sucede lo mismo. Por mucho que tengas conocimiento teórico, experiencia o intuición, nunca vas a poder competir con una organización que toma decisiones en base a datos —a grandes volúmenes de datos— en tiempo real. Y eso lo pueden hacer las empresas que han convertido el *big data* en una pieza más de la construcción de sus sistemas de información. Las otras, tal vez sean como los dinosaurios que aún no lo saben, pero están a punto de extinguirse.

Reseñas biográficas de los autores

Reseñas biográficas de los autores

Miguel Socías

Miguel Socías es licenciado en Economía por la Universidad Católica de Chile y Máster en Economía y Doctorado en Educación Internacional Comparada de Stanford University. Después de terminar sus estudios de doctorado formó parte del equipo del American Institutes for Research de Palo Alto como un Senior Research Scientist evaluando el impacto de políticas educativas fiscales y estatales en Estados Unidos. Miguel Socías pasó luego a ser Associate for Analytics de la Carnegie Foundation for the Advancement of Teaching en Stanford, estando a cargo del desarrollo de una plataforma para desarrollar un nuevo currículum de matemática y estadística en 17 universidades a lo largo de Estados Unidos. También creó su propio emprendimiento para facilitar el acceso a datos públicos de educación a familias en Estados Unidos. Después de su emprendimiento, Miguel Socías fue consultor para el Banco Mundial y el BID durante tres años, prestando asesoría en materia de manejo y análisis de grandes bases de datos, políticas fiscales y políticas educativas. Actualmente es un *data scientist en eShares,* una *startup* en Silicon Valley que hace uso del *big data* para crear productos financieros para inversionistas y emprendedores alrededor del mundo.

Eduard Martín Lineros

Eduard Martín Lineros, ingeniero superior en Informática e ingeniero técnico en Informática de Sistemas por la Universidad Oberta de Catalunya, Cambridge Diploma in Information Technology (UCLES); director de Estrategia para el Sector Público en España de Sopra Steria, decano del Ilustre Colegio de Ingeniería en Informática de Catalunya y exdirector de Innovación, Sociedad del Conocimiento y Arquitecturas TIC del Ayuntamiento de Barcelona.

Durante 27 años desarrolló su carrera profesional en el Ayuntamiento de Barcelona ocupando diversas responsabilidades técnicas hasta impulsar el proyecto de Innovación TIC en la ciudad durante los años 2011 a 2015, que fue decisivo para la nominación de la capital catalana como primera Capital Europea de la Innovación 2014-2016 por parte de la Comisión Europea.

En el Ayuntamiento de Barcelona impulsó el proyecto del Modelo de Gestión de la Innovación, siendo la primera entidad pública nacional que alcanzó la certificación UNE166002; participó y dirigió numerosos programas de actuación *smart* en el conjunto de la iniciativa Smart City Barcelona, destacando la ideación y definición del proyecto de Sistema Operativo de Ciudad (CityOs) y los programas de desarrollo educativo y social.

Actualmente como director de Estrategia para el Sector Público de Sopra Steria es el responsable directo de la expansión de la oferta comercial del grupo en el ámbito de las ciudades inteligentes, la "Internet of Things" y la innovación aplicada a los servicios públicos, así como de la mejora y consolidación de la oferta tradicional de la compañía.

Como decano del COEINF, es responsable directo de la creación de la nueva Asociación de Profesionales TIC de Catalunya y miembro de varios comités de trabajo para el impulso de la profesión a nivel catalán.

Óscar Corcho

Óscar Corcho es profesor titular de la Universidad Politécnica de Madrid (UPM) y pertenece al Grupo de Ingeniería de Ontologías (OEG). Como parte de su participación en el nodo Open Data Institute Madrid, Óscar lidera la red temática española en Datos Abiertos para ciudades inteligentes, donde las directrices conjuntas y vocabularios se proponen para la armonización de conjuntos de datos a través de portales de datos abiertos en España. También ha participado en la creación de la norma UNE178301: 2015 sobre datos abiertos para *smart cities*, que propone un modelo de madurez para evaluar y mejorar la calidad de las implementaciones de datos ciudades, y en la actual iniciativa "OjoalData100" para la identificación de los 100 *datasets* abiertos más relevantes para las ciudades. Ha asesorado en la implementación de la API de datos abiertos de Zaragoza.

Además, en 2013 Óscar cofundó Localidata, una empresa especializada en apoyar la implementación de estrategias abiertas de datos por las ciudades.

En 2016, recibió el premio López de Peñalver de la Real Academia de Ingeniería al investigador más prometedor de ingeniería de España.

Alberto González Yanes

Alberto González Yanes es licenciado en Matemáticas por la Universidad de La Laguna. Siendo Jefe de Servicio de Estadísticas Económicas del el Instituto Canario de Estadística (ISTAC) desde el año 2006, ha coordinado la actividad encomendada al mismo: estadísticas económicas, estadísticas ambientales, estadísticas laborales y estadísticas de ciencia y tecnología. Como tal ha contribuido en el desarrollo de proyectos de I+i asociados a las estimaciones en pequeñas áreas de Canarias, métodos avanzados de muestreo y metodologías para el aprovechamiento estadístico de registros administrativos.

Desde el año 2014 es Director de la Unidad Mixta de Metodología e Investigación en Estadística Pública entre el ISTAC y la Universidad de La Laguna; cuyas líneas estratégicas de investigación estarán dirigidas a mejorar la eficiencia productiva del Sistema Estadístico de Canarias, con especial atención a los métodos de regionalización de estadísticas.

A su vez desde 2016 es Director del Comité Técnico de la Infraestructura de Datos y Metadatos Estadísticos de Canarias (eDatos), en esa línea de actuación fue coordinador del proyecto europeo METAMAC, ejecutado por las Oficinas de Estadística de la Macaronesia Europea, de desarrollo de un sistema integrado de metadatos estadísticos e implantación del estándar SDMX (Statistical Data and Metadata Exchange). Complementariamente es Director del Comité Técnico del Sistema de Datos Integrados (iDatos) para la integración y geocodificación de fuentes administrativas con fines estadísticos.

Ejerce como representante suplente del Gobierno de Canarias en el Comité Interterritorial de Estadística (CITE) de España.

Noemí Brito

Noemí Brito es abogada digital y, en la actualidad, directora de Derecho Digital en LEGISTEL www.legistel.es y Sportic, así como IT GRC (IT Governance, Risk and Compliance) en COMTRUST www.comtrust.es. Forma parte de diversas organizaciones a nivel nacional e internacional, entre otras, es vocal de la Junta Directiva de la Asociación de Expertos Nacionales de Abogacía Digital (ENATIC), donde coordina además su Comisión de Corporate. Igualmente es corresponsable del Grupo de Derecho Digital del Capítulo español de la "European Law Institute" (Spanish ELI Hub), así como miembro de los Comités Operativos del Data Privacy Institute-ISMS Forum Spain y del Capítulo español de CSA-ES (Cloud Security Alliance). Asimismo, destaca su cargo como miembro del Comité de Dirección de la Sección TIC de AEADE (Asociación Europea de Arbitraje). Asimismo, es docente en diversas universidades e instituciones académicas y educativas de reconocido prestigio como son las universidades Carlos III de Madrid (UC3M), Universidad de Salamanca (USAL), Universidad Internacional de la Rioja (UNIR), la EOI (Escuela de Organización Industrial), la Escuela de Negocios Digital "The Valley", entre otras. Ha participado en diversas publicaciones sobre privacidad y *big data*, por ejemplo, recientemente como coautora en la nueva *Guía de Buenas Prácticas en Protección de Datos para Proyectos Big Data*, editada conjuntamente por la Agencia Española de Protección de Datos www.agpd.es e www.ismsforum.es que resulta accesible desde el siguiente enlace: https://www.agpd.es/portalwebAGPD/revista_prensa/revista_prensa/2017/notas_prensa/news/2017_05_11-ides-idphp.php

Enrique Ávila

Enrique Ávila Gómez es, en la actualidad, drector del Centro Nacional de Excelencia en Ciberseguridad, adscrito a la Universidad Autónoma de Madrid, además de jefe del Área de Seguridad de la Información de la Guardia Civil. Con anterioridad, ha sido subdirector de la Escuela de Inteligencia Económica, encuadrada dentro del Instituto de Ciencias Forenses y de la Seguridad perteneciente, a su vez, a la misma Universidad Autónoma de Madrid. Ha sido jefe de Seguridad de S.I. en el Tribunal Constitucional de España y responsable de Seguridad Perimetral en el Ministerio del Interior. Investigador del Instituto Universitario de Investigación sobre Seguridad Interior (IUISI) y del Instituto de Ciencias Forenses y de la Seguridad (ICFS), ha desarrollado un proyecto de investigación para el

primero en materia de "Inteligencia Económica aplicado a la Seguridad Interior". Profesor de varios títulos de Máster entre los que podemos citar los siguientes: Máster en Evidencias Digitales y Lucha contra el Cibercrimen (UAM), Máster en Ciberseguridad y Ciberdefensa, Escuela Superior de Guerra de Colombia, y Máster en Seguridad de los Sistemas de Información (UEM). Conferenciante habitual en foros de Inteligencia y Ciberseguridad.

Licenciado en Derecho por la Universidad Complutense de Madrid, Experto Universitario en Servicios de Inteligencia por el Instituto Universitario Gutiérrez Mellado y Máster Universitario en Evidencias Digitales y lucha contra el Cibercrimen.

Miguel Quintanilla

Ingeniero de Telecomunicaciones, Executive MBA, Advanced Management Program por el Instituto de Empresa y Project Management Profesional certificado por el PMI®. Los últimos cuatro años como director general de Nuevas Tecnologías y Telecomunicaciones en el Ayuntamiento de Las Palmas de Gran Canaria he sido el máximo responsable de la definición y ejecución del proceso de transformación digital de la región, liderando, entre otros, el plan director de Ciudad Inteligente de Las Palmas de Gran Canaria o el proceso de conversión de la ciudad en Destino Inteligente. Actualmente sigo apoyando este proceso de transformación digital como director ejecutivo de MHP Servicios de Control. En paralelo con mi actividad profesional he desarrollado otras actividades no remuneradas como la de miembro de la Junta Directiva y del Comité Ejecutivo del Club Financiero de Canarias, tesorero del Club Natación Metropole, vocal en el comité de normalización de ciudades inteligentes de AENOR, representante en la Red Española de Ciudades Inteligentes y ponente habitual en charlas y conferencias.

Diego May

Diego May es CEO de ixpantia (Ciencia de Datos), cofundador de Junar (Datos Abiertos) y emprendedor en la industria de DATA. Anteriormente, Diego trabajó en fondos de capital de riesgo en Boston y en Latinoamérica, así como en *startups* y en compañías multinacionales como Intel, Lucent, y Verizon, liderando diferentes cargos. Diego se formó como ingeniero en el Instituto Tecnológico de Buenos Aires, Argentina, y cuenta con un MBA del MIT Sloan School of Management.

Frans van Dunné

Frans van Dunné es CDO de ixpantia con más de 15 años de experiencia en datos. Frans ha colaborado con múltiples empresas en definición de estrategia y productos de datos. Como parte de estos procesos ha trabajado en modelado de procesos, minería de datos y gestión de datos con organizaciones alrededor del mundo. Frans se formó como biólogo, tiene un PhD de la Universidad de Amsterdam y ha enseñado estadísticas y modelado en universidades.

Ray Walshe

Ray Walshe es el jefe de la delegación de Big Data en la NSAI sobre la ISO JTC1 WG9. Investigador en el mayor centro de investigación de Europa en Análisis de Datos (Insight National Centre for Data Analytics), donde dirige los grupos de proyectos *big data* en Personal Sensing y Media Analytics. Ray es co-director de la Red STRTIC (Standars Research in ICT) y lidera múltiples iniciativas europeas H2020 referentes a estándares IT. Actualmente, se desempeña como director de Normas ISO en la ISO / IEC JTC1 / WG 9 20547 Big Data Reference Architecture, además de estar involucrado en otros grupos de trabajo de estándares en organizaciones como IEEE, NIST, ETSI y CENELEC. Ray tiene más de 30 años de experiencia en electrónica, *software* y telecomunicaciones trabajando para empresas como Electric Ireland, Ericsson, Siemens, Siemens Nixdorf y Software & Systems Engineering Ltd antes de unirse a Dublin City University. Trabaja con la Comisión Europea en el área de emprendimiento, representando a Irlanda en la Startup Europe University Network y coordina el StartUp Europe Week Dublin. Ray también sirve como Chief Architect de ReskiTV, Chief Architect de Performance Tracking Solutions y director de CLoudCore, Instituto de Investigación en Cloud Computing.

Jane Kernan

Jane Kernan tiene más de 20 años de experiencia como profesora en la Escuela de Informática en Dublin City University, Irlanda. Es coordinadora del 1er y 2° año del programa en Informática y exdirectora del programa máster en IT y de la diplomatura en Informática. Jane imparte clases de grado y posgrado en gestión de bases de datos y aplicaciones

orientadas a negocios y cuenta con una amplia experiencia supervisando proyectos de estudiantes. Jane realiza investigaciones en el Centro de Investigación CloudCORE y en la Asociación Europea de Investigación Universitaria de la Industria (EIURA).

Organizó el Foro EIURA Cloud, la 23º Conferencia sobre Inteligencia Artificial y Ciencia Cognitiva y ha participado más recientemente con la iniciativa Startup Europe de la Comisión Europea, incluyendo la Semana de Europa de la Startup, la Startup Europe University Network y la iniciativa SEC2U.

Juan Muñoz

Es doctor en Ciencias de la Computación. Se desempeña como director de Planeación y Normatividad Informática en el Instituto Nacional de Estadística y Geografía (INEGI) de México y como profesor investigador titular de la Universidad Autónoma de Aguascalientes (UAA). Colabora en distintos grupos internacionales coordinados por diferentes organismos como UNSD, UNECE, OECD, etc., en los que ha participado en distintos proyectos para desarrollar tecnologías y estándares para la modernización de la producción de información estadística, entre los que se incluye el uso de fuentes emergentes de datos conocidas comúnmente como *big data*.

Antonio Moneo

Antonio Moneo trabaja en el Banco Interamericano de Desarrollo promoviendo los datos abiertos como una herramienta de colaboración para los gobiernos, ciudadanos, emprendedores y universidades para resolver los retos del desarrollo. En el BID, lanzó el blog "Abierto al público" y organizó los primeros *hackathons* de innovación cívica. Es miembro de la Carta Internacional de Datos Abiertos y formador registrado en el Open Data Institute (ODI). Anteriormente, trabajó en la London School of Economics y la Universidad Nacional de Educación a Distancia, donde obtuvo un doctorado en Ciencia Política. También cuenta con una Maestría en Comercio Internacional. Actualmente es asociado sénior en Aprendizaje y Gestión del Conocimiento en el BID.

Lourdes Muñoz

Ingeniera técnica en Informática por la Universidad Politécnica de Cataluña (UPC), master en Sociedad de la Información y el Conocimiento por la Universidad Abierta de Cataluña (UOC). Fundadora y actual Co-Directora de Barcelona Iniciativa Open Data el Nodo Barcelona del Open Data Institute.

Ha sido diputada en el Congresp por la provincia de Barcelona del Partido de los Socialistas de Cataluña desde el 2002 a 2015. Además de su trabajo en temas de sociedad de la información y open data ha destacado su participación en la Comisión Mixta de los Derechos de la Mujer. En 2012 impulsó el proyecto "Partido abierto y transparencia" en el Partido Socialista de Cataluña. Es presidenta de la red catalana de mujeres "Dones en Xarxa"

Desde muy joven compatibilizó estudios, trabajo y compromiso social. Empezó a militar a los 15 años en las Juventudes Socialistas de Catalunya en 1984 y en el Partido de los Socialistas de Cataluña desde 1989, donde ha asumido diversos cargos de responsabilidad.

Trabajó como analista de gestión hasta 1999 que se incorpora al Ayuntamiento de Barcelona como consejera técnica del distrito de Les Corts. En 2001 como concejala de la Mujer del Ayuntamiento de Barcelona, es impulsora de diversas iniciativas para fomentar el uso de las tecnologías por parte de las mujeres.

Ha sido pionera en la blogosfera política española defendiendo las posibilidades de la web 2.0 para mejorar la conexión entre los políticos y la ciudadanía y avanzar en la transparencia. Inicia su blog en septiembre de 2005 para explicar su experiencia al quedar atrapada por el Huracán Katrina de Nueva Orleans. Es cofundadora de la Iniciativa Barcelona Open Data y coordinadora del Máster Open Data de Euncet Business School.

Su compromiso político y su compromiso feminista siempre han ido de la mano. Es cofundadora junto a la periodista Montserrat Boix y la escritora Gemma Lienas de Dones en Xarxa, red catalana por la igualdad, fundada en 2004 y que actualmente preside.

Nuria Oliver

Nuria Oliver es directora de Investigación en Ciencias de Datos en Vodafone, Chief Data Scientist en Data-Pop Alliance y Chief Scientific Advisor en Vodafone Institute. Es ingeniera superior de Telecomunicaciones por la UPM y doctora por el Massachusetts Institute of Technology (MIT) en inteligencia perceptual. Tiene más de 20 años de experiencia investigadora en MIT, Microsoft Research (Redmond, WA) y como primera directora científica (mujer) en Telefónica I+D (Barcelona). Su trabajo en el modelado computacional del comportamiento humano, la interacción persona-máquina, la informática móvil y el análisis de *big data* —especialmente para el bien social— es internacionalmente conocido con más de 150 publicaciones científicas, citadas más de 11.000 veces y con una decena de premios y nominaciones a mejor artículo científico. Es coinventora de 40 patentes y ponente invitada regularmente en congresos internacionales.

La Dra. Oliver es la primera investigadora mujer española nombrada Distinguished Scientist por el ACM (2015). Ha sido distinguida como Fellow del IEEE (2017, una investigadora mujer española en esta edición) y Fellow de la Asociación Europea de Inteligencia Artificial (2016, única española en esta edición).

Ha recibido numerosos premios por su trabajo, destacando el MIT TR100 Young Innovator Award (2004), el Gaudí Gresol Award a la Excelencia en Ciencia y Tecnología (2016), el Premio Nacional Ada Byron a la Mujer Tecnóloga (2016), el Premio Europeo Ada Byron a la Mujer Digital del Año (2016) y el Premio Nacional de Informatica Angela Ruiz Robles (2016). Ha sido seleccionada como "una de las más destacadas directoras en tecnología" (*El País*, 2012), una de los "100 líderes del futuro" (Capital, 2009), un Rising Talent por el Women's Forum for the Economy and Society (2009) y uno de los "40 jóvenes que marcarán el próximo milenio" (*El País*, 1999), entre otros.

Su pasión es mejorar la calidad de vida de las personas, tanto a nivel individual como colectivo, a través de la tecnología. También tiene un gran interés hacia la divulgación científica, por lo que es colaboradora frecuente con los medios de comunicación (prensa, radio, televisión) e imparte charlas de divulgación tecnológico-científica al público en general y especialmente a adolescentes, con particular énfasis en las chicas.

Pedro Huichalaf

Abogado de la Universidad de Valparaíso. Magíster (c) en Derecho Informático y de las Telecomunicaciones de la Universidad de Chile. Especializado en TIC (Tecnologías de la Información y Comunicaciones) y en Telecomunicaciones, teniendo experiencia en regulación nacional e internacional en temas de propiedad intelectual, datos personales, delitos informáticos, comercio electrónico, regulación sector telecomunicaciones, derecho administrativo. En sus inicios, promotor y participante activo de movimientos ciudadanos relacionados con tecnología, donde ha interactuado con diversos actores relevantes de la industria, gobierno y parlamento. Posteriormente fue asesor experto tanto de sector público como sector legislativo en temas relacionados con TIC. Desde marzo de 2014 a octubre de 2016 se desempeñó como viceministro de Telecomunicaciones de Chile, donde tuvo a cargo la conducción de las políticas públicas de Chile en materia de Telecomunicaciones. Actualmente realiza diversas consultorías tanto a nivel nacional como internacional en materia de políticas públicas de telecomunicaciones y tecnología.

Sebastián Vargas

Estudio CC. Empresariales en la Universidad Complutense de Madrid y es Máster en Comercio Internacional por la Escuela Europea de Dirección y Empresa. Traductor inglés/español (C) acreditado por la Universidad Europea de Madrid.

Tras sus inicios en Nueva York en la iniciativa sobre Responsabilidad Social Empresarial de las Naciones Unidas "Global Compact" apoyando al equipo de comunicaciones (2007), Sebastián se incorporó al equipo del Gabinete del viceministro chileno de Telecomunicaciones Pedro Huichalaf, donde asesoró en materias relativas al ámbito internacional, tales como: coordinador de respuestas a la OCDE en materia de telecomunicaciones, coordinador del cuarto informe de Chile sobre objetivos del milenio PNUD en materia TIC, delegado junto al viceministro de Telecomunicaciones y delegación chilena representando a Chile en Corea del Sur en la Conferencia de Plenipotenciarios de la UIT, responsable también de la elaboración del MOU firmado entre Chile y Corea del Sur en materia tecnológica, entre otras labores (2014).

Posteriormente colaboró activamente en el programa "Exporta Digital" de la Dirección General de Relaciones Económicas Internacionales perteneciente al

Ministerio de Relaciones Exteriores de Chile, enfocado en instalar a empresas chilenas en plataformas digitales para incorporarlas en la dinámica de exportaciones vía Internet cuya finalidad es la de diversificar la matriz productiva nacional.

Actualmente se encuentra finalizando su formación como traductor profesional inglés/español en España.

María Isabel Mejía

Ingeniera de Sistemas y Computación de la Universidad de los Andes, donde también hizo una especialización en Gerencia Estratégica de Informática. Tiene más de 30 años de experiencia en el sector de las tecnologías de la información y las comunicaciones, trabajando tanto en entidades públicas, como empresas privadas y la academia.

María Isabel Mejía fue la coordinadora del Proyecto Año 2000 (Y2K) de Colombia. Fue la responsable de liderar las acciones requeridas por el país para llevar tecnológicamente a Colombia al cambio del milenio.

Entre el 2000 y el 2006 asumió la Dirección Ejecutiva de Computadores para Educar, diseñando las estrategias de promoción, gestión de donaciones, reacondicionamiento de computadores y acompañamiento educativo que hoy por hoy benefician a millones de niños de escasos recursos.

En el 2006, se encargó del diseño de la Estrategia Nacional de Gobierno en línea, programa del cual fue directora hasta el 2010. En ese momento, de acuerdo con el E-Government Survey de Naciones Unidas, ella logró posicionar a Colombia como líder de la región de América Latina y el Caribe en Gobierno Electrónico y llevó al país a ocupar el noveno puesto a nivel mundial en desarrollo de servicios en línea.

De 2010 a 2012 se desempeñó como gerente general de la compañía Marinas de Colombia S.A.S., responsable de la construcción y ejecución del proyecto Marina Puerto Velero, la marina más completa y moderna del Caribe colombiano.

Entre 2012 y 2016 trabajó en el Ministerio de Tecnologías de la Información y las Comunicaciones como viceministra de Tecnologías y Sistemas de la Información. Desde esta posición María Isabel desempeñó el rol de CIO de Colombia, liderando el fortalecimiento de la industria de Tecnologías de la Información, así como el fortalecimiento de la gestión de Tecnologías de la Información en el Estado.

Actualmente es la directora ejecutiva de Info Projects, empresa dedicada al desarrollo de proyectos y soluciones innovadoras para la transformación digital de las organizaciones, aprovechando tecnologías punta como *big data*, "Internet de las cosas", computación cognitiva y Blockchain.

Edwin Estrada

Es licenciado en Derecho por la Universidad de San José. Tiene una Especialidad en Derecho Agrario y Ambiental y un Técnico en Administración y Regulación de Telecomunicaciones por la Universidad de Costa Rica. Es egresado de la Maestría en Derecho Público de la Universidad Autónoma de Centroamérica. Experto en Regulación de Telecomunicaciones" emitido por la Universidad Autónoma de Centroamérica. Está incorporado al Colegio de Abogados desde 1996. Se ha desempeñado como abogado en instituciones del sector público y como asesor parlamentario, especialmente del Plenario Legislativo y en las comisiones relacionadas con las telecomunicaciones. Ha sido profesor universitario por más de 10 años. Labora en la institución desde el 15 de julio de 2009 y ha desempeñado los cargos de gerente de Normas y Procedimientos y director de Concesiones y Normas. A partir del 17 de mayo del año 2016 funge como viceministro de Telecomunicaciones.

Juan Vasquez

Juan Sebastián Vasquez une analisis de datos, narración visual, implementación de sistemas y reforma estructural en su trabajo con el Equipo de Innovación Operativa de la Alcaldía de Los Ángeles. Previo a trabajar con el gobierno, Juan manejó las redes sociales y ventas en Latinoamérica y España para NationBuilder, una compañía de tecnología enfocada en campañas políticas, gobiernos y organizaciones sin fines de lucro. Juan empezó su carrera como publicista en la Florida y es originalmente de Bogota, Colombia.

En su tiempo libre Juan es conferencista y facilita entrenamientos en crecimiento de comunidades digitales e innovación en el gobierno.

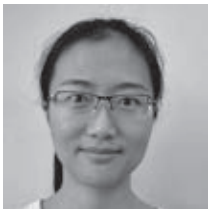
Marco Hernández

Marco Antonio Hernández Oré es líder de Programa para el Sudeste de Europa en el Banco Mundial, centrándose en el crecimiento, finanzas e instituciones. Se unió al Banco Mundial en 2008 y ha trabajado como economista de países de África Subsahariana, Europa y América Latina. En esta capacidad, Marco ha apoyado a los gobiernos en la búsqueda de soluciones para gestionar mejor las políticas que promueven el crecimiento económico y reducen la pobreza. Antes de unirse al Banco Mundial, Marco trabajó como economista en Charles River Associates International, en Boston, centrándose en la regulación económica y el análisis de las políticas públicas. También trabajó como consultor en el Ministerio de Economía y Finanzas del Perú. Marco Tiene una licenciatura en Economía por el MIT y un Ph.D. por la Universidad de Oxford.

Andrew Whitby

Andrew Whitby es científico de datos en el grupo de desarrollo de datos del Banco Mundial. Trabajó previamente en grandes proyectos en los laboratorios de innovación del Banco Mundial. Su trabajo se centra en econometría y ciencia de datos; también en la economía de la tecnología, innovación y crecimiento. Anteriormente, trabajaba como investigador en Nesta, un grupo de expertos en innovación de Londres, y como consultor económico en The Allen Consulting Group (actualmente, ACIL Allen Consulting), con sede en Brisbane, Australia, focalizado en proyectos en política social (salud, educación, discapacidad, pobreza) y regulación económica (gas, electricidad, agua, aviación / aeropuertos). Recibió su M. Phil. Y Ph.D. en Economía por la Universidad de Oxford.

Lingzi Hong



Lingzi Hong es estudiante de doctorado de tercer año en la Universidad de Maryland en College Park. Es asistente de investigación en el Laboratorio de Informática Urbana y miembro del Laboratorio de Linguística Computacional y Procesamiento de Información. Sus intereses de investigación incluyen aplicaciones del *machine-learning*, minería de datos espacio-temporales y procesamiento natural del lenguaje. Concretamente, utiliza datos de telefonía móvil para desvelar la relación entre la

situación socioeconómica regional y los patrones de movilidad humana, incluyendo el análisis de medios sociales para comprender la comunicación entre el gobierno y los ciudadanos durante los eventos sociales y bajo inclemencias meteorológicas. Posee una licenciatura en Sistemas de Información y Maestría en Ciencias de la Información por la Universidad de Beijing, Beijing, China.

Vanessa Frías-Martínez

Vanessa Frías-Martínez es profesora asistente en el iSchool y profesora asistente de Afiliados del Departamento en Ciencias de la Computación de la Universidad de Maryland, College Park. Directora del Laboratorio de Informática Urbana y miembro de Laboratorio de Lingüística Computacional y Procesamiento de Información, del Centro de Ciencias de la Información Geoespacial y del Centro de Investigación de Población. El enfoque de la investigación de Vanessa es aplicar minería de datos, *machine learning* y técnicas de procesamiento natural del lenguaje para analizar el comportamiento humano en el mundo físico. Su objetivo es desarrollar modelos conductuales que tengan un impacto social en áreas como la planificación urbana, el desarrollo económico o la epidemiología. Vanessa recibió su M.Sc. Y Ph.D. en Informática por la Universidad de Columbia. De 2009 a 2013, fue investigadora en el gupo de minería de datos y modelado de usuarios de Telefónica Research en Madrid, España.

Enrique Frías-Martínez

Enrique Frías-Martínez es investigador en Telefónica Research, Madrid, España, la empresa de investigación y desarrollo del Grupo Telefónica. El foco de investigación de Enrique se encuentra en computación urbana, movilidad humana y *big data* para el bien social. Antes de unirse a Telefónica, trabajó para el Departamento de Ingeniería Biomédica de la Universidad de California, Los Ángeles (UCLA), para el Departamento de Sistemas de Información y Computación de la Universidad de Brunel en Londres, y para el Instituto Courant de Ciencias Matemáticas, Universidad de Nueva York. Enrique recibió un Ph.D. Ingeniería Informática por la Universidad Politécnica de Madrid, Madrid, España, en el año 2000, y un Ph.D. en Sistemas de Información de la Universidad de Brunel, Londres, U.K., en 2007. Su tesis fue primer premio de la Escuela de Informática de la Universidad Politécnica de Madrid en el año 2001.

Juan Mateos García

Juan Mateos García es jefe de Mapeo de la Innovación en Nesta Policy and Research. En su trabajo, Juan utiliza nuevas fuentes de datos y métodos analíticos para mejorar las políticas y prácticas de innovación. Recientemente, Juan ha trabajado en Arloesiadur (una plataforma de datos de innovación para el gobierno Galés), TechNation (un mapa de la economía digital del Reino Unido) y The Geography of Creativity (un mapa de la economía creativa en el Reino Unido).

Juan está interesado en las dinámicas de emergencia de nuevas tecnologías e industrias, en la difusión de ideas a través de redes y en la gestión de este proceso de cambio continuo para beneficio de todos. Técnicamente, Juan está interesado en el potencial del aprendizaje automático (*machine learning*) y la ciencia de redes (*network science*) para entender una economía compleja, y en la reproducibilidad como una herramienta para generar confianza en nuevas fuentes de datos y métodos, haciéndolos más adecuados para su aplicación pública.

Juan es licenciado en Económicas por la Universidad de Salamanca y tiene un Master en Science and Technology Policy por SPRU, University of Sussex.

Richard Benjamins

Richard Benjamins es director External Positioning & Big Data for Social Good en LUCA, la unidad de Big Data B2B de Telefónica. Antes ha sido director de Big Data para optimizar y mejorar el negocio de Telefónica. Richard ha sido profesor e investigador en varias universidades internacionales y ha sido cofundador de un *start-up* de Internet. Richard obtuvo su doctorado en Ciencia Cognitiva / Inteligencia Artificial en 1993 en la Universidad de Amsterdam.

Elena Díaz Sánchez

Elena Díaz Sánchez es consultor analítico en Go to Market en LUCA, la unidad de Big Data B2B de Telefónica. Elena antes ha trabajado como consultor analítico en diferentes Consultoras de Marketing Analítico, así como en empresas proveedoras de datos a terceros. Elena obtuvo su licenciatura en Ciencias Matemáticas en 2005 en la Universidad Complutense de Madrid, en 2008 un Máster en Aprendizaje Estadístico y Data Mining en la UNED, y en 2014 un Máster en Marketing Digital Analítica y UX en el IEBS.

Tomás Trenor

Tomás Trenor es Business Analyst en LUCA, la unidad de Big Data B2B de Telefónica. Tomás cuenta con más de diez años de experiencia en desarrollo de *software* y análisis de datos en diferentes sectores. En concreto, ha ocupado diferentes posiciones en empresas de I+D, Consultoría, Banca y ahora Telecomunicaciones. Tomás es ingeniero superior en Informática, ingeniero de Telecomunicaciones y posee un Máster de Negocios en Internet del prestigioso Instituto Superior de Desarrollo de Internet.

Pedro de Alarcón

Pedro de Alarcón es ingeniero de computadores por la UGR y actualmente es Senior Data Scientist en LUCA Data-Driven Decisions (Telefónica), enfocado en proyectos de Big Data for Social Good. Su carrera profesional empezó en el mundo científico con un doctorado en bioinformática por el Centro Nacional de Biotecnología (CSIC) y el San Diego Supercomputer Center (USA). En 2003 fue socio fundador y CTO de Integromics. En 2007 se unió al equipo de Telefónica donde ha coordinado el desarrollo de productos de eHealth y el equipo global de analítica avanzada de datos de televisión.

Javier Carro

Javier Carro Calabor es ingeniero de Telecomunicación y PMP Certified. Actualmente trabaja como Data Scientist en el área External Positioning & Big Data for Social Good de LUCA. Anteriormente trabajó para BI de productos internos de Telefónica, especialmente para el área de vídeo. También ha sido Desarrollador y Program Manager en proyectos de tecnologías móviles y exposición de APIs.

Florence Broderick

Florence Broderick trabaja actualmente en LUCA, la unidad de Big Data de Telefónica, como Strategic Marketing Manager. Viene de Reino Unido y ha trabajado en varias áreas de Telefónica incluyendo el departamento de Coches Conectados, tecnología VOIP e *insights* de datos móviles —trabajando en roles de desarrollo de negocio y *marketing* en los sectores de transporte, turismo, retail y publicad exterior. Es embajadora de One Young World y ha asistido a la cumbre de Dublín (2014) y Ottawa (2016), en la que ha hablado de la oportunidad de usar *big data* para el bien social. Además es fundadora de la red "Millennial Network" de Telefónica.

Chema Alonso

Chema Alonso es actualmente CDO de Telefónica. Previamente fue el fundador y CEO de Eleven Paths empresa filial de Telefónica Digital centrada en la innovación en productos de seguridad y el director General de Global Security Business en la unidad B2B de Telefónica Business Solutions. Anteriormente trabajó y dirigió Informática 64 durante 14 años, empresa centrada en Seguridad Informática y formación. Es Dr. en Seguridad Informática por la Universidad Rey Juan Carlos de Madrid, ingeniero informático por la URJC e ingeniero informático de Sistemas por la Universidad Politécnica de Madrid, que además le nombró Embajador Honorífico de la Escuela Universitaria de Informática en el año 2012.

Tamara Dull

Tamara Dull es directora de Tecnologías Emergentes para SAS Best Practices, un equipo de liderazgo en SAS Institute. A través de compromisos clave con la industria, artículos provocativos y publicaciones, ofrece una perspectiva pragmática sobre *big data*, "Internet de las cosas", código abierto, privacidad y ciberseguridad. Tamara comenzó su aventura en el mundo de la alta tecnología mucho antes de que Internet naciera, y ha ocupado posiciones técnicas y administrativas para múltiples proveedores de tecnología, consultorías y una organización sin fines de lucro. Ella está en la lista de las "25 mujeres más influyentes en el IoT" del Instituto IoT y en la lista Onalytica Big Data Top 100 durante los últimos tres años. Tamara es también miembro del consejo asesor de la Comunidad del Internet de las Cosas.

Manuel Acevedo

Manuel Acevedo Ruiz es consultor independiente en el área de las TIC y redes de desarrollo, así como en evaluación y planificación estratégica con organizaciones de desarrollo como IDRC (Canadá), Hivos (Holanda), agencias de Naciones Unidas, EuropeAid y gobiernos nacionales. Coordinó la Red de Telecentros de Latinoamérica y el Caribe (Red LAC) (2012-2013). Entre 1994 y el 2003 trabajó en el Programa de Naciones Unidas para el Desarrollo (PNUD) y en UN Voluntarios (UNV), estableciendo, entre otros, el Servicio de Voluntariado en Línea de NN. UU. (www.onlinevolunteering.org) y el programa UNITeS (UN Information Technology Service) de voluntariado en el área de TIC y Desarrollo. Ha participado como docente en varios programas universitarios de posgrado sobre TIC, redes y desarrollo humano, y es un doctorando en la Universidad Politécnica de Madrid (sobre modelos en red de Cooperación al Desarrollo), donde es miembro del Centro de Innovación en Tecnología para el Desarrollo (itdUPM) y del Grupo de Investigación sobre Organizaciones Sostenibles.

Pablo Díaz

Ingeniero industrial, Msc in Project Management por La Salle Business School y Msc in International Business por la Universidad Pompeu Fabra. Fundador y socio-director del Grupo EVM, donde se encarga de liderar la unidad de consultoría estratégica y operaciones.

Cuenta con amplia experiencia en el diseño, planificación, desarrollo y puesta en marcha de proyectos de base tecnológica en organizaciones públicas, siempre intentando anteponer el valor social y la utilidad pública de la tecnología para mejorar la manera en la que los ciudadanos se relacionan con sus gobiernos y gobernantes.

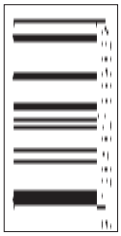
Es coautor de los libros *Open Government - Gobierno Abierto* (Algón Editores, 2010) y *Guía práctica para abrir Gobiernos* (Goberna América Latina, 2015).

También ayuda a otros emprendedores a convertir en realidad sus ideas, centrándose en proyectos de tecnología cívica (Civic Tech).

Algunos ejemplos son Civiclick.es, desarrollos tecnológicos que buscan mejorar la gestión pública a través de la aplicación del paradigma del gobierno abierto, o Buong, una plataforma *online* para la difusión de documentales centrados en mostrar la realidad social del mundo.

Si existe un fenómeno tecnológico que ha inundado de manera masiva los medios generalistas (y también los especializados) en los últimos años, ese es el *big data*. Sin embargo, como ocurre con muchas tendencias tecnológicas, hay cierta confusión en su definición que genera incertidumbre y dudas cuando se trata de entender cómo esta tecnología puede ser usada desde el sector público para mejorar la forma en la que se toman decisiones o se prestan bienes y servicios a la ciudadanía.

A través de la lectura de los veinte capítulos de este manual, el lector podrá descubrir y descifrar los diferentes aspectos que son necesarios analizar antes de formular y desarrollar políticas o proyectos públicos en los que el uso de herramientas de ciencias de datos o vinculadas al *big data* sean la clave del éxito.



Con la colaboración de:

Telefónica