"Científico de datos", la profesión del presente.

"Data scientist", today's job.

José Antonio Álvarez Jareño
Jose.A.Alvarez@uv.es
Universitat de València

Vicente Coll-Serrano
Vicente.Coll@uv.es
Universitat de València

Resumen

Debido a la explosión de datos que ha generado Internet y a la necesidad de transformar estos datos en información que aporte valor añadido, cada día es más demandado por las empresas y las instituciones el profesional conocido como "científico de datos". Las principales funciones de éste serán: entender los datos, comprender y resolver los problemas que se le planteen y conocer la tecnología disponible. Además, deberá disponer de dos cualidades personales que son: la curiosidad y la habilidad para comunicar. Dado que es imposible ser experto en todas las disciplinas que abarcaría el "científico de datos", es necesaria la especialización de los profesionales y la creación de equipos multidisciplinares.

Palabras clave

Big Data; Minería de datos; Machine Learning; Analítica Predictiva.

Abstract

Due to the data explosion that the Internet has generated and the need of transforming all these data into information that provides added value, the professional known as "Data Scientist" is increasingly demanded by firms and institutions. The main functions of a data scientist are (i) to understand the data, (ii) to understand the problem one needs to solve, and also (iii) to know and use the available technology. In addition, a data scientist must have curiosity and ability to communicate. Since it is impossible to be an expert in all disciplines in the scope of

a data scientist, professionals' specialization is required, as well as the creation of multidisciplinary teams.

Keywords

Big Data; Data Mining; Machine learning; Predictive analytics

Recibido: 16/02/2018
Aceptado:22/03/2018
DOI: http://dx.doi.org/10.5557/IIMEI9-N16-113129
Descripción propuesta: Álvarez Jareño, José Antonio; Coll-Serrano, Vicente, 2018. "Científico de datos", la profesión del presente. *Métodos de Información*, **9**(16), 113-129.

1. Introducción

En alguna ocasión todos nosotros hemos leído o escuchado eslóganes del estilo "quien tiene un cliente, tiene un tesoro", "el cliente es el rey" o "el cliente siempre tiene la razón", como parte de la estrategia de marketing o de calidad de las empresas. Sin embargo, estas afirmaciones no son una verdad absoluta. Selden y Colvin (2003) defienden la idea de que todos los clientes son valiosos es un mito de los negocios y que se debe tomar con la misma precaución que otras leyendas urbanas. Todos los clientes no son igual de rentables para una empresa. Si ordenásemos los clientes en función del beneficio que generan para la empresa, es fácil que nos encontráramos con la conocida como Ley de Pareto (Juran 2010). El 20% de los clientes más rentables aportan el 80% del beneficio de la empresa; al mismo tiempo, el quintil menos rentable no solo no aporta ningún beneficio, sino que incluso destruye valor. Identificar a los clientes que destruyen el valor de la empresa e invitarles amablemente a que se busquen otro proveedor es el trabajo de muchos gerentes, y supone una labor muy ardua.

Este problema que inicialmente era de difícil solución, puede ser resuelto, con la tecnología y el conocimiento disponible actualmente, por unos profesionales que han llegado para revolucionar la toma de decisiones en la

empresa y en otros muchos campos. Son los "científicos de datos", traducción del término anglosajón "Data Scientist", y que según la revista Harvard Business Review es "la profesión más sexy del siglo XXI" (Davenport y Patil 2012).

El 6 de abril de 2015 volvió a entrar en funcionamiento el Gran Colisionador de Hadrones de la Organización Europea para la Investigación Nuclear (CERN, por sus siglas en inglés). El director de Investigación y Ciencia Computacional del CERN, Sergio Bertolucci, en una entrevista concedida en 2015 a Lidia Montes, decía que "nuestra máquina produce 30 petabytes de datos cada año, y que esta cantidad de datos sería similar a una columna de 20 km de DVDs, sin carcasa". Para poder manejar y tratar este ingente volumen de datos será necesario disponer del hardware y del software más adecuado, porque, continúa diciendo Sergio Bertolucci en su entrevista, "ahora el juego está en ver cuál es el valor de estos datos".

Efectivamente, el volumen de datos disponible crece exponencialmente. Para captar la atención del público sobre esta afirmación, basta plantear la sencilla pregunta:

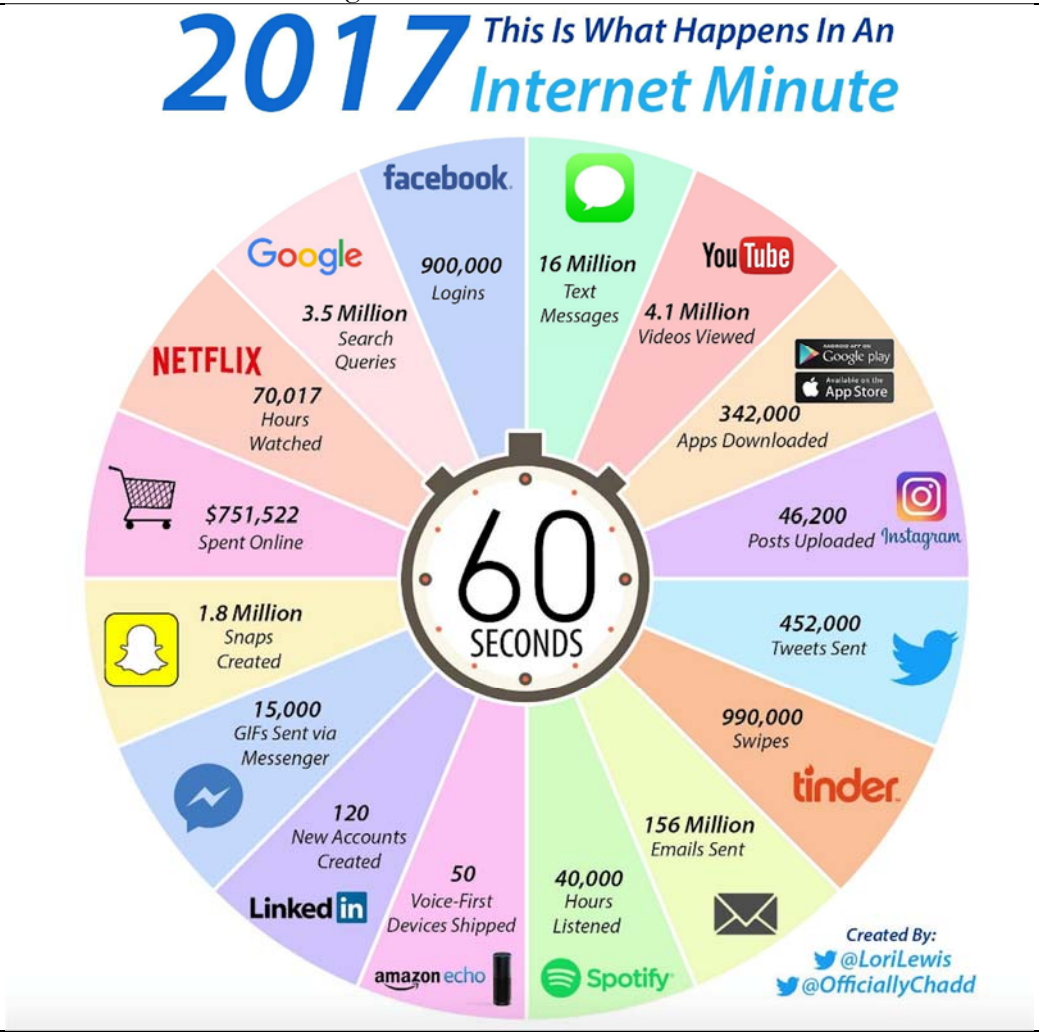
¿Qué pasa en un minuto en Internet?

Pensemos un instante en esta cuestión. ¡Exacto! La respuesta es: Muchísimas cosas pasan en Internet en un minuto.

Como se ilustra en la Figura 1, en un minuto se van a enviar más de 204 millones de correos electrónicos, se van a subir a YouTube 30 horas de video y se van a ver 1,3 millones de videos, se van a enviar más de 100.000 twits, etc. El volumen de datos generados crece de forma exponencial, al mismo ritmo que se expande la tecnología y sus aplicaciones por el mundo. La mayoría de esta información es creada por los usuarios de estas aplicaciones, y pueden aportar una información muy importante para muchas empresas e instituciones.

En la edición especial de febrero de 2010 de The Economist, Cukier (2010) exponía: "La cantidad de información digital se multiplica por 10 cada 5 años, mientras que la Ley de Moore (Moore, 1965) indica que la capacidad de procesamiento se duplica cada 18 meses". Los datos se incrementan mucho más deprisa que la capacidad para procesarlos.

Figura 1. Un minuto en Internet



Fuente: Lewis y Callahan (2017)

Es necesario que todo ese volumen de datos pase de ser información a ser conocimiento; de manera que los científicos y los profesionales puedan aprovecharlo para obtener mejoras sustanciales en los procesos a los que se dedican. La monitorización de los motores de avión realizada sobre los datos que facilitan los sensores con los que están equipados permite mejorar su eficiencia y prevenir posibles fallos de funcionamiento. De igual forma, los directivos de una empresa pueden tomar las decisiones más adecuadas en base a la información disponible, tanto interna, como externa.

2. ¿Cuál es el trabajo de un "científico de datos"?

Isabel Munera (2016), en su artículo "Profesiones que desaparecen y otras que son el futuro pero aún no existen" para el diario El Mundo, dice que "el

trabajo más cotizado este año será el de "growth hacker" y el más buscado el de especialista en Big Data". A continuación hace una descripción del trabajo que realizan estos últimos: "los especialistas en Big Data analizan los grandes datos que posee una empresa y los usan como indicadores fiables para proponer medidas correctoras y para ayudar en la toma de decisiones sobre el rumbo que tiene que seguir un determinado negocio". Finalmente, termina indicando que "ante la falta de formación en este ámbito, cada vez más universidades y escuelas de negocio están ofreciendo másters especializados en gestión, gobierno y arquitectura de datos".

El término "científico de datos" o "especialista en Big Data" abarca un amplio espectro de roles que van desde los centros de investigación -como el CERN-, las grandes empresas tecnológicas - como Google, Amazon o Microsoft-, a las administraciones públicas -como puede ser la Agencia Tributaria o la Seguridad Social. Sin embargo, todos ellos tendrán unas capacidades similares en cuanto a su formación y conocimientos.

No obstante, no será necesario disponer de un volumen de datos tan importante como el de estas empresas para explotar toda su potencialidad. Cualquier empresa, grande o pequeña, dispone de datos que pueden ser tratados para convertir la información en conocimiento. Hoy, gran cantidad de empresas disponen de una página web que les genera un tráfico continuo y constante de información que los "científicos de datos" pueden analizar.

Las tres principales capacidades que debe tener un "científico de datos" son, para Steve Hanks, científico de datos jefe en Whitepages.com, las que expone Bernard Marr (2016) en la revista Forbes: (1) Entender los datos, (2) Entender el problema a resolver y cómo pueden ayudar los datos, y (3) Comprender la tecnología disponible.

2.1. Entender los datos

Efectivamente, la principal virtud de un "científico de datos" será entender qué son y qué representa esos grandes conjuntos de datos. Para esta labor no se dispone de ningún procedimiento o técnica, y los profesionales se enfrentan a los datos con su intuición y su experiencia; esperando descubrir la punta del hilo del ovillo de Ariadna que les permita encontrar la salida del

laberinto. La mejor tecnología y los algoritmos más modernos no servirán de nada si no se comprenden los datos que se van a analizar, y la dificultad sería enorme, como buscar una aguja en un pajar.

Algunas veces, las soluciones serán sencillas y obvias, pero otras veces, la mayoría de ellas, serán complejas y difíciles. Los problemas a los que se enfrenta un "científico de datos" son, casi siempre, completamente nuevos. Nunca antes había habido tantos datos disponibles, ni en cantidad, ni en variedad, ni en calidad de los mismos. Por este motivo, Carlos Elías (2015) expone que entender los datos es un arte, "algo más sublime que una ciencia o una tecnología".

Los datos están empezando a ser un nuevo factor de producción, tal como indican Mayer-Schönberger y Cukier (2013), y el valor de los datos no disminuye con su uso. Es consecuentemente, un bien "no rival", diferentes personas pueden utilizar los mismos datos con diferentes finalidades, sin ser competencia entre ellos.

2.2. Comprender el problema a resolver

Solo una vez se comprende la materia prima con la que trabaja el "científico de datos", puede pasarse a la siguiente etapa: comprender el problema que se pretende resolver con los datos disponibles. ¿Qué conocimiento se puede extraer de los datos disponibles?

Para esta labor, el "científico de datos" dispone de una importante caja de herramientas (tool-kit) para analizar los datos y buscar correlaciones o patrones de comportamiento, que posteriormente permitirán realizar predicciones (Serrano-Cobos, 2014). El área de conocimiento que abarca estas técnicas se conoce en inglés como "Machine Learning" o "Statistical Learning", mientras que en castellano se han denominado "aprendizaje automático". La prestigiosa consultora Gartner publica cada año "The Hype Cycle for Emerging Technologies" (http://www.gartner.com/newsroom/id/3114217), que proporciona una perspectiva transversal de la industria de las tecnologías y muestra las tendencias que deberían considerar en el desarrollo de sus carteras de tecnología los Business Strategists, los líderes de innovación y desarrollo, los

emprendedores, los equipos de tecnologías emergentes, etc. En 2015, en el pico de las expectativas infladas del Hype Cycle de Gartner se situaba el "Internet of Things" y "Machine Learning", este último ocupando la posición que en 2014 era representada por "Big Data".

Los conjuntos de datos con los que se trabaja suelen ser enormemente grandes, siendo factible, en consecuencia, dividir este gran volumen de datos en dos subconjuntos: uno de entrenamiento (entre el 66% y el 80% de los datos originales), sobre que el que se buscarán los patrones y las correlaciones, y otro de test (los datos restantes), sobre el que se comprobará el modelo elaborado con los datos de entrenamiento. La calibración de los modelos es una de las fases más importantes en el proceso de extracción de conocimiento de los datos, tal como indica Silver (2014).

En una primera fase, se pretende obtener un modelo que describa los datos lo mejor posible sin que haya sobreajuste (overfitting). Fundamentalmente se dispone de una importante batería de métodos estadísticos que permiten extraer conocimiento de los datos (minería de datos), y que van desde la regresión logística, los árboles de decisión o las reglas de decisión (métodos clásicos), hasta los modernos sistemas de recomendación, las redes neuronales, los sistemas genéticos o el SVM (Support Vector Machine). Sin embargo, se han ido incluyendo otras teorías científicas que pueden aportar una mejor comprensión de los conjuntos de datos, como son, entre otros, la Teoría de Juegos o la Teoría de Decisión. La evolución es constante, y la potencia de cálculo de las computadoras actuales permite aplicar métodos de "bootstraping" como el "bagging", el "boosting" o los "random forest".

Con el modelo obtenido, se pasa a una segunda fase donde se pretende predecir el comportamiento del subconjunto de comprobación. No todos los modelos que describen bien el subconjunto de entrenamiento tienen porque predecir bien los resultados del subconjunto de test. Si el modelo es muy sensible a los datos se ajustará muy bien a los datos de entrenamiento, pero no será capaz de realizar buenas predicciones. Como indica Siegel (2014), los datos contienen una gran cantidad de ruido y no es posible modelizar correctamente la señal, se están suponiendo más cosas de las que muestran los datos.

Para solucionar estos problemas se han desarrollado técnicas basadas en el "bootstrap" y en el "ensamble de modelos". Seleccionar aleatoriamente

diferentes conjuntos de entrenamiento y, en lugar de ajustar un único modelo, obtener un modelo de cada conjunto diferente, haciendo que la solución predictiva para los datos del conjunto de test se obtenga por la mayoría de los modelos, como si de un proceso de votación se tratase.

Surowiecki (2005) expone que un individuo tiene pocas posibilidades de acertar el resultado exacto de una predicción, sin embargo, el colectivo a través de una votación la mejora considerablemente. La conclusión es que la media de las predicciones es más precisa que las predicciones individuales. El concepto se ha denominado "sabiduría de las masas".

El ensamblado de modelos vendría a reproducir esta idea de "sabiduría de las masas" desde un punto de vista matemático. Si se construyen diferentes modelos y se aplica un esquema de combinación, el resultado de la combinación debería tener un mayor rendimiento que el mejor modelo individual.

2.3. Conocer la tecnología disponible

Por último, la tercera capacidad del "científico de datos" es conocer la tecnología disponible y saber cómo utilizarla para la resolución de los problemas. De poco servirá resolver el problema si no se dispone de la infraestructura adecuada para obtener una solución práctica, precisa y correcta en tiempo y lugar. El "científico de datos" debe conocer la tecnología disponible para resolver el problema en tiempo y forma. Si la solución llega tarde, puede que el coste sea muy alto, y en algunos casos inasumible.

Actualmente está disponible, a coste cero, el software necesario para realizar el análisis estadístico de modelización y posterior calibración. Diferentes proyectos de "open source" -como R, Python, Weka y MOA- permitirán el entrenamiento de los modelos con las técnicas estadísticas más avanzadas. Sin embargo, debido a los grandes conjuntos de datos con los que lidian todos los días los "científicos de datos", estas aplicaciones pueden tardar horas o incluso días en alcanzar una solución para algunos de sus algoritmos y, evidentemente, las empresas o los centros de investigación no siempre disponen de tanto tiempo para tomar una decisión.

Muchas veces, no solo es necesario ajustar y calibrar un modelo adecuado, sino que además hay que hacerlo en un tiempo record. Si un algoritmo creado para la detección del uso fraudulento de tarjetas de crédito tarda más de unos minutos en detectar el uso inadecuado de una tarjeta, puede que cuando la empresa lo sepa la cuenta de su cliente este en "números rojos". Es necesario ser muy rápido en detectar el fraude para poder tomar las medidas oportunas y reducir las consecuencias perniciosas.

Por este motivo, y dado el tamaño de muchos conjuntos de datos, se han creado diferentes ecosistemas de programación que permiten manejar grandes bases de datos y realizar cálculos a gran velocidad. En estos momentos, el más conocido es Hadoop.

Hadoop permite que las tareas analíticas se dividan en fragmentos de trabajo y se distribuyan entre miles de ordenadores, ofreciendo un tiempo de análisis menor y un sistema de almacenamiento distribuido de grandes cantidades de datos. El almacenamiento lo proporciona HDFS (Hadoop Distributed File System) y el análisis MapReduce. Hadoop es, en esencia, un framework que permite procesar grandes cantidades de datos en paralelo a muy bajo coste.

HDFS es un sistema de ficheros distribuidos que fue creado a partir del Google File System (GFS). Al trabajar con un gran número de componentes de hardware (nodos), la probabilidad de que se produzca un fallo se incrementa rápidamente. HDFS replica los datos en diferentes clústeres y tiene una gran tolerancia a los fallos. Además está optimizado para trabajar con enormes flujos de información y realizar las tareas de lectura y escritura de grandes ficheros. Los datos se replican mediante un clúster de ordenadores en aras de la fiabilidad y disponibilidad.

Por su parte, realizar correctamente un análisis en un sistema distribuido es muy complejo, y MapReduce es capaz de simplificarlo mediante el procesamiento en paralelo. Básicamente, se dispone de dos funciones. Por un lado, *Map* funciona como extractor y asigna valores a determinadas claves para un documento o dato; por otro lado, *Reduce* realiza la función de agregación y combina las claves de múltiples documentos o datos para crear un valor reducido único por cada clave.

Dentro del ecosistema Hadoop se pueden integrar el software de análisis estadístico, pudiendo utilizar R o Phython, además de otras aplicaciones como Pig, Hive o Mahout. Que Hadoop sea el framework más conocido y

distribuido a través de diferentes compañías como Cloudera, Hortonworks o MapR, no quiere decir que sea el único, ni siquiera el que más futuro tiene. Hoy en día, su principal competidor es Spark, y algunos profesionales prefieren utilizar Scala o Akka por su versatilidad y escalabilidad.

Ser experto en las tres áreas de conocimiento –entender los datos, comprender el problema a resolver y conocer la tecnología disponible-– es literalmente imposible. Lo más lógico es crear equipos multidisciplinares, equipos de trabajo con diferentes expertos en las diferentes áreas. El "científico de datos jefe" debe comprender todas las piezas que integran el proceso, pero es imposible que sea experto en todas ellas. Así, un experto informático conocerá y sabrá aplicar la tecnología disponible en cada momento, pero será el experto en aprendizaje automático el que se encargue de la analítica predictiva y su posterior interpretación, ambos asesorados por alguien con la experiencia y el conocimiento para entender los datos que se están tratando.

Además, se debe añadir, que todas las piezas están en constante cambio y en un proceso de continua innovación, por lo que una tarea importantísima del "científico de datos" será la formación. Pero la formación que precisa no la encontrará en los canales que han sido habituales hasta ahora, y deberá adaptarse a las nuevas tecnologías de distribución del conocimiento y la información. Será necesario un proceso de reciclaje continuo para no quedar obsoleto en un período muy corto de tiempo.

3. Cualidades personales del "científico de datos"

Por si no fuera suficiente con lo enumerado hasta el momento, a los "científicos de datos" se les presumen dos cualidades personales: la curiosidad y la habilidad de comunicar.

Philip Ball (2013) nos recuerda que "hubo un tiempo en que la curiosidad era algo condenable: a fin de cuentas, por su culpa cometió Eva ese pecado original que al parecer aún estamos pagando". Sin embargo, gracias a esa curiosidad insaciable de algunos seres humanos la ciencia avanza y se dispone de un conocimiento que permite construir el LCH y "ver" el Big Bang con el que surgió nuestro universo.

El "científico de datos" deberá tener una gran curiosidad por los datos, lo que le llevará a estudiarlos y a analizarlos con la finalidad de descubrir y aprender cosas nuevas. Transformar la información en conocimiento.

En segundo lugar, aunque no por ello menos importante, estaría la capacidad de comunicar el conocimiento que se ha obtenido. El "científico de datos" debe ser un excelente comunicador, deberá entender los problemas a los que se enfrenta para, a continuación, poderlos exponer al equipo de trabajo de manera que sean comprensibles por todos ellos para, entre todos, componer la solución más adecuada. Los resultados que se obtienen en los procesos de aprendizaje automático no siempre son sencillos de interpretar. Mientras que los métodos clásicos -como los árboles de decisión- tienen una interpretación muy intuitiva, los modelos modernos -de redes neuronales o los random forest- muchas veces no ofrecen una interpretación inmediata y/o coherente del problema a resolver. Por este motivo, a algunos métodos se les conoce como "cajas negras", es decir, se conoce su funcionamiento pero no se puede explicar fácilmente la solución obtenida.

Cuando se trabaja con muchos datos y muchas variables, la complejidad de las relaciones y las correlaciones entre las mismas pueden no ser evidentes, y por tanto, no se dispone de una explicación sencilla sobre el fenómeno objeto de análisis. Kahneman (2011) expone que en la época de datos escasos, demostrar lo equivocadas que estaban las intuiciones causales era muy costoso. Sin embargo, con estas tecnologías será sencillo mostrar que hay muy poca o ninguna conexión estadística entre el efecto y su supuesta causa, y de esta forma, se mejorará la toma de decisiones, aunque no se comprenda la relación causal. En este sentido, la habilidad de comunicar del "científico de datos" será de gran utilidad.

4. Cualidades personales del "científico de datos"

Por si no fuera suficiente con lo enumerado hasta el momento, a los Dada la naturaleza del trabajo a desarrollar por parte de los "científicos de datos" y las cualidades personales requeridas, es evidente que no habrá un solo tipo de "científico de datos", de la misma forma que existirán necesidades diferentes por parte de las empresas y de los centros de investigación.

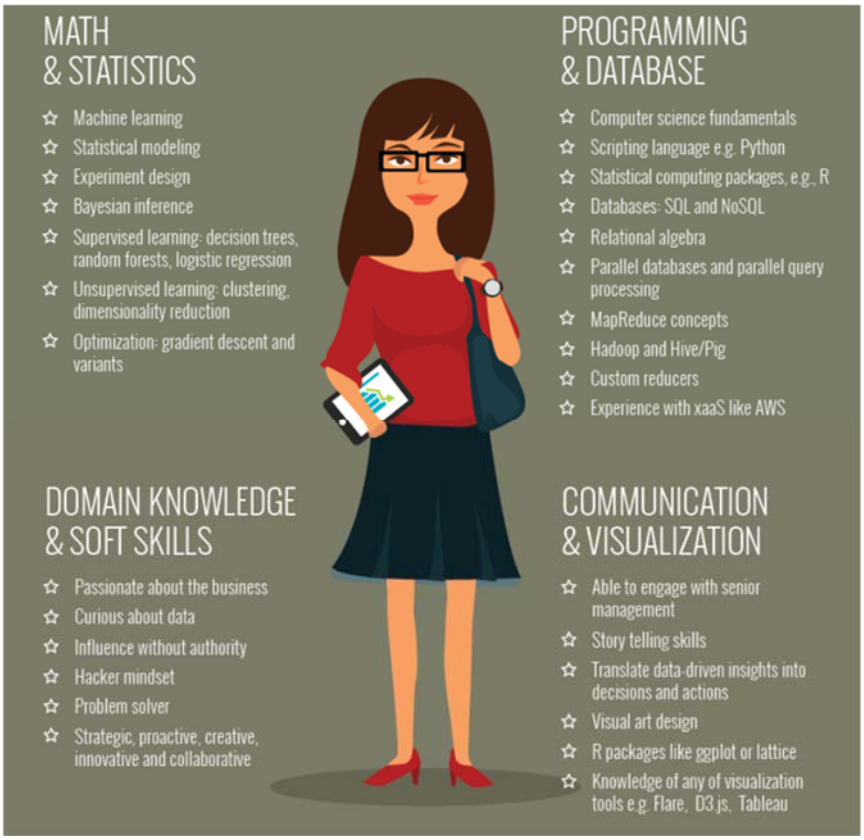


Figura 2. Habilidades de un "Científico de datos"
Fuente: Zawadzki (2014).

Según Zawadzki (2014), y como puede verse en la Figura 2, un "científico de datos" requiere tener un conjunto de habilidades multidisciplinares.

Harris, Murphy y Vaisman (2013), que son "científicos de datos", decidieron hacer lo que mejor se les da, y recopilaron una importante cantidad de información para determinar cuáles eran los tipos más importantes de "científico de datos" y sus principales conocimientos. Las personas que participaron en la encuesta se auto-identificaron en cuatro tipos de "científicos de datos": Desarrollador, Investigador, Creativo y Empresario. Además, a los encuestados se les pidió que indicasen qué habilidades consideraban que eran las más útiles para el desarrollo de su trabajo de entre un listado de 22 habilidades genéricas utilizadas habitualmente por los "científicos de datos". Estas habilidades se agruparon en cinco grandes grupos:

- Negocio: Desarrollo de producto y empresa.
- Machine Learning/Big Data: Datos estructurados y no estructurados, Machine Learning y almacenamiento de datos distribuidos.

- Matemáticas: Optimización, matemáticas, modelos gráficos, estadística bayesiana y modelos de Monte Carlo, algoritmos y simulación.
- Programación: Administración de sistemas y programación.
- Estadística: Visualización, series temporales, estadística espacial, estadística clásica, manejo de datos y supervivencia y marketing.

Relacionando auto-identificación del tipo de "científico de datos" y habilidades se obtuvo un perfil para cada uno de los cuatro tipos de "científico de datos":

- El Desarrollador de datos debe tener amplios conocimientos en programación y Machine Learning, en menor medida en matemáticas, y solo precisa de conocimientos genéricos en negocio y en estadística. Su función principal es aplicar la tecnología disponible para obtener la mejor solución del problema, es decir, seleccionar las infraestructuras y programar los algoritmos que darán solución al problema.
- El Investigador de datos deberá tener unos profundos conocimientos de estadística y matemáticas, y utilizará muy poco la programación, el Machine Learning y los negocios. Su principal característica será el planteamiento y solución del problema con los datos disponibles. La caja de herramientas del "científico de datos" es enorme, de tal suerte que su trabajo consistirá en determinar cuáles son las mejores metodologías para la resolución de un problema concreto.
- El Creativo de datos es el que precisa de unas habilidades más amplias, porque aunque no es necesario que sea experto en ninguna materia sí que tiene que tener amplios conocimientos de Machine Learning, programación y estadística. Imaginar soluciones creativas para un conjunto de datos no es una tarea fácil, y esa es la principal función de este tipo de "científico de datos".
- El Empresario de datos deberá ser experto en Negocios por delante de las otras cuatro habilidades. Debe tener conocimientos de Machine Learning y estadística, y solamente ciertas nociones de matemáticas o programación. Es el encargado de convertir en un producto comercializable el trabajo de sus otros compañeros. Si de los datos se puede obtener un "valor añadido" que puede suponer una ventaja frente a otros competidores, el empresario de datos será el más adecuado para llevar a la práctica la solución que se haya obtenido.

Es imposible ser experto en todas las disciplinas que abarcaría el "científico de datos", por lo que es necesaria la especialización de los profesionales y la creación de equipos multidisciplinares, donde diferentes puntos de vistas permitan afrontar los problemas desde distintas perspectivas y obtener soluciones más eficientes. Los equipos se deberán formar desde la discrepancia, tanto en formación como en valores, y con un objetivo común: alcanzar el consenso. El mestizaje es la principal fuente de innovación en un mundo que busca la uniformidad y la estandarización de los individuos.

Nate Silver (2014) nos previene en el uso de patrones de comportamiento con grandes volúmenes de datos: "*Encontrar patrones es muy fácil en un entorno con abundancia de datos y, de hecho, eso es justamente lo que hacen los apostantes mediocres. La clave está en saber decidir si esos patrones son ruido o señal*".

5. Conclusiones

Los científicos o investigadores siempre han sido "científicos de datos". Kepler fue un magnífico científico de datos, aunque nadie en su época lo hubiera llamado así, más bien astrónomo, o incluso astrólogo. La ciencia siempre ha avanzado en base a una metodología científica que ha estado basada en datos. La principal diferencia es que antes los datos eran escasos y caros, y ahora son abundantes y muy baratos. Se podría decir que los datos han pasado de ser un elemento de lujo a un bien de consumo de masas. Esta democratización de los datos implica que muchos empleos y profesiones deben cambiar su metodología de trabajo para pasar de experto en una materia, por ejemplo, marketing o recursos humanos, a científico de datos. Los expertos basan sus conocimientos en la intuición y la experiencia, mientras que los científicos de datos solo precisan de datos, muchos datos.

La mayoría de la metodología estadística que se utiliza en ciencia de datos se descubrió hace 50 años o más, pero muchas de ellas sólo se han podido poner en práctica cuando se ha dispuesto de grandes volúmenes de datos y de los ordenadores con la potencia de cálculo adecuada. Los ordenadores personales tienen hoy el suficiente poder de procesamiento para tratar volúmenes de datos considerables, aunque también se puede recurrir a servicios de cálculo en internet o al análisis de datos en la nube a un coste muy bajo.

En determinadas profesiones ya se utilizaba una metodología de trabajo muy similar a la de hoy en día para el desarrollo de sus actividades. Así, por ejemplo, los actuarios utilizan los datos disponibles en la compañía para analizar los riesgos y determinar el coste individual del seguro. La toma de decisiones se fundamenta en el análisis de los datos del negocio. Ahora, además de los datos internos de la empresa, se dispone de datos externos que son fáciles de obtener o muy baratos. Si a los datos internos, se añaden los datos externos, los resultados del análisis serán más consistentes, y consecuentemente, mejores.

No es exacto afirmar que el "científico de datos" es una nueva profesión, más bien es la democratización de la profesión de científico o investigador. La metodología es común a todas las ciencias empíricas, y la única novedad es la tecnología que permite disponer de más datos, analizarlos a muy bajo coste y aplicarlos a cualquier tipo de negocio o profesión. Los científicos de datos deberán pasar de ser una profesión, como se ve actualmente, a formar parte del acervo de conocimientos de cualquier empleo o profesión. Los abogados, los economistas o los médicos, entre otros, deberán tener amplios conocimientos en ciencia de datos, aunque no sea expertos, para poder desarrollar su profesión con garantías.

Los prejuicios, tabúes, limitaciones culturales y un largo etcétera impiden que las personas tomen las decisiones más adecuadas. Sin embargo, ahora se dispone de la metodología, los datos y las herramientas para poder realizar un análisis adecuado de la situación y tomar las decisiones con conocimiento de causa. Llegados a este punto, se debería tener en cuenta el uso que se hace del conocimiento que se obtiene, porque no siempre se va utilizar de forma ética y justa. Son varios los casos que se conocen, en los que los científicos de datos han utilizado el conocimiento extraído para llevar a cabo acciones ilícitas o amorales. Sería bueno que los científicos de datos dispusieran de un código deontológico que regulase su profesión, al igual que los médicos tienen su juramento hipocrático.

En definitiva, "científico de datos", un nuevo nombre para un trabajo muy antiguo. Seguramente no será el último, ahora mismo hay varios creativos trabajando en una nueva marca más potente. Renovarse o morir.

6. Bibliografía

BALL, P., 2013. *Curiosidad. Por qué todo nos interesa*. Madrid: Turner Publicaciones. ISBN 978-84-15832-74-4

CUKIER, K., 2010. Data, data everywhere [en línea]. *The Economist*. [Consulta: 13/02/2018]. Disponible en: https://www.emc.com/collateral/analyst-reports/ar-the-economist-data-data-everywhere.pdf

DAVENPORT, T.H. y PATIL, D.J., 2012. Data Scientist: The Sexiest Job of the 21st Century [en línea]. *Harvard Business Review*, October Issue. [Consulta: 13/02/2018]. Disponible en: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

ELÍAS, C., 2015. *El selfie de Galileo. Software social, político e intelectual del siglo XXI*. Barcelona: Ediciones Península. ISBN 978-84-9942-424-8

GARTNER, 2015. *Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor* [en línea]. [Consulta: 13/02/2018]. Disponible en: http://www.gartner.com/newsroom/id/3114217

HARRIS, H., MURPHY, S. y VAISMAN, M., 2013. *Analyzing the Analyzers. An Introspective Survey of Data Scientists and Their Work*. Sebastopol: O'Reilly Media. ISBN 781449368241

JURAN, J.M., 2010. *Juran's Quality Control Handbook*. 6th Edition. New York: McGraw-Hill. ISBN 0-07-034003-X

KAHNEMAN, D., 2011. *Pensar Rápido, Pensar Despacio*. Barcelona: Círculo de Lectores. ISBN 9788483068618

LEWIS, L. y CALLAHAN, C., 2017. *This is What Happens In An Internet Minute, 2017* [en línea]. [Consulta: 16/02/2018]. Disponible en: http://www.visualcapitalist.com/happens-internet-minute-2017/

MARR, B., 2016. Big Data Uncovered: What Does A Data Scientist Really Do? *Forbes*, enero [en línea]. [Consulta: 16/02/2018]. Disponible en: http://www.forbes.com/sites/bernardmarr/2016/01/07/big-data-uncovered-what-does-a-data-scientist-really-do/#535251066f7f

MAYER-SCHÖNBERGER, V. y CUKIER, K., 2013. *Big Data. La revolución de los datos masivos*. Madrid: Turner Publicaciones. ISBN 978-84-15427-81-0

MONTES, L., 2015. Imagina a un ordenador decir 'quizás', será el fin de la

casualidad. *El Mundo*, Madrid, 10 abril. [Consulta: 16/02/2018]. Disponible en: http://www.elmundo.es/economia/2015/04/10/5526badbe2704e104e8b4572.html

MOORE, G. E., 1965. Cramming more components onto integrated circuits. *Electronics,* **38**(8), 114-117. [Consulta: 16/02/2018]. Disponible en: https://www.cs.utexas.edu/~fussell/courses/cs352h/papers/moore.pdf

MUNERA, I., 2016. Profesiones que desaparecen y otras que son el futuro pero aún no existen. *El Mundo*, Madrid, 30 enero. [Consulta: 16/02/2018]. Disponible en: http://www.elmundo.es/economia/2016/01/30/56aba00222601d457c8b465f.html

SELDEN, L. y COLVIN, G., 2003. *Angel Customers and Demon customers: Discover Which is Which and Turbocharge Your Stock*. New York: Portfolio.

SERRANO-COBOS, J., 2014. Big data y analítica web. Estudiar las corrientes y pescar en un océano de datos. *El profesional de la información*, **23**(6), 561-565. DOI: http://dx.doi.org/10.3145/epi.2014.nov.01

SIEGEL, E., 2014. *Analítica Predictiva. Prediciendo el futuro utilizando Big Data*. Madrid: Anaya Multimedia. ISBN: 9788441534421

SILVER, N., 2014. *La señal y el ruido*. Barcelona: Ediciones Península, 2014. ISBN 9788499423234

SUROWIECKI, J., 2004. *Cien mejor que uno. La sabiduría de la multitud o por que la mayoría siempre es más inteligente que la minoría*. Barcelona: Ediciones Urano. ISBN 9788479535919

ZAWADZKI, K., 2014. Is Data Science a buzzword? Marketing Distilley, 29 Noviembre [en línea]. [Consulta: 16/02/2018]. Disponible en: https://mywebvault.wordpress.com/2017/05/18/is-data-science-a-buzzword-modern-data-scientist-defined-marketing-distillery