Los roles, ámbitos y nombres de la ciencia de datos

PID_00261827

Marçal Mora Cantallops

Universitat Oberta de Catalunya

Marçal Mora Cantallops

Ingeniero industrial e ingeniero informático por la UPC, máster en Data Science por la UAH y doctorando en Comunicación, Información y Tecnología de la Sociedad en Red por la misma universidad. Investigador en el ámbito de los *game studies*, la ciencia de datos y, en particular, el análisis de redes sociales; está interesado en el uso de estas técnicas para la extracción de conocimiento e información. Ha trabajado en la creación y optimización de modelos estadísticos para logística y planificación de la demanda y actualmente participa en varios proyectos relacionados con la estadística y la ciencia de datos.

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Josep Maria Marco (2019)

Índice

Introducción

¡Bienvenidos a la Ciencia de datos! En este módulo se introducirán muchos conceptos y terminología de la ciencia de datos, la mayoría de los cuales no os serán familiares.

Esto es del todo normal porque la ciencia de datos es una disciplina relativamente moderna, que se basa en muchos elementos y principios que provienen de disciplinas muy diferentes, desde la matemática y la estadística hasta las ciencias de la computación, pasando por las actividades empresariales. Es muy difícil encontrar personas que dominen estos tres elementos y todavía lo es más que lo hagan en más de un ámbito. Así que no os preocupéis, aprended tantos conceptos como podáis, que os suenen, y más adelante, a medida que avancéis en los estudios, ya los iréis entendiendo y consolidando. En este módulo se trata de proporcionar una perspectiva general y global de la ciencia de datos y esto tiene dos consecuencias: que aparecerán muchos conceptos nuevos y que no se podrán desarrollar con la profundidad que merecen. ¡Pero nada que no se pueda solucionar!

La estructura del módulo es la siguiente:

1) En primer lugar, se presentará el contexto histórico en el que surge y nace la ciencia de datos, que ayudará a entender su emergencia y las preguntas a las que busca dar respuesta.

2) En segundo lugar, se elaborará la definición del científico de datos: qué es, qué hace, qué habilidades debe cultivar y de qué herramientas dispone.

3) En tercer lugar se propondrá una visión muy general de algunos ámbitos en los que ya se está aplicando la ciencia de datos y se hablará del potencial que presenta en algunos otros.

4) Finalmente, se introducirá un pequeño glosario con explicaciones y ejemplos sobre los términos y conceptos más comunes de la ciencia de datos (y, por lo tanto, términos que aparecerán de forma habitual en el futuro).

En resumen, esta unidad es un mapa que tiene el objetivo de ayudar al estudiante a situarse y orientarse en el mundo de la ciencia de datos. Es solo el primer paso del camino hasta el objetivo final. ¡Ánimo y adelante!

1. Origen y evolución de la ciencia de datos

Tanto el análisis de datos como una de sus ramas más extendidas y presentes, la llamada inteligencia de negocio (*business intelligence*, BI), han ganado un protagonismo especial en las últimas décadas; las oportunidades derivadas del uso de la información disponible y su análisis en cualquier organización han generado un crecimiento notable del interés en estas disciplinas. Se puede entender el BI como las técnicas, sistemas, tecnologías, prácticas, aplicaciones y metodologías que sirven para extraer valor de los datos que, a su vez, consigan que el negocio (o la organización) tome decisiones más informadas y que, por lo tanto, tengan un retorno positivo. La ciencia de datos, entendida en su sentido más amplio, no es ni mucho menos nueva, y sus orígenes se pueden seguir hasta prácticamente a mediados de siglo pasado, con los primeros intentos de dotar de inteligencia a las primeras (y mastodónticas) computadoras. Lo que sí es relativamente nuevo es su popularidad y proyección, espoleada por unas capacidades técnicas que por primera vez consiguen acercarse a sus promesas. Un artículo famoso de la prestigiosa *Harvard Business Review* la presentaba, en 2012, como «el trabajo más sexy del siglo XXI»; tendencia acentuada en los últimos años, especialmente en los Estados Unidos.

Sin embargo, dejando de lado las cifras y las perspectivas económicas, es necesario entender cómo se ha llegado hasta aquí. Tanto los datos masivos (*big data*) como la ciencia de datos (*data science*) son dos de las palabras más de moda de la década, pero no son ideas nuevas. De hecho, no son ni ideas de un solo ámbito y llevan más de cincuenta años haciéndose lugar en la industria y las disciplinas de análisis. Es momento, pues, de mirar medio siglo atrás.

1.1. Modelos estadísticos y minería de datos

Para encontrar la primera referencia en el cambio aportado por la computación es necesario remontarse al año 1962, cuando John Tukey se dio cuenta del potencial de la intersección entre la estadística y la computación en una cita que es, a día de hoy, célebre:

> «[...] a medida que he visto evolucionar la estadística, he tenido motivos para reflexionar y dudar [...] creo que he descubierto que mi interés principal es el análisis de datos...» (Tukey, 1962).

Tukey está refiriéndose, de alguna manera, al amor a primera vista que sintió cuando, mediante ordenadores, los resultados estadísticos podían ser obtenidos en horas, mucho más rápidamente que los días o semanas que se tardaba con los métodos manuales.

Otro nombre importante es el de Peter Naur que, avanzado a su tiempo, publicó el *Concise Survey of Computer Methods* (Naur, 1974), un compendio de métodos de procesamiento de datos en múltiples aplicaciones. Lo más curioso de este caso es que ya citaba varias veces el término *ciencia de datos*, que definía de la forma siguiente:

> «la ciencia de trabajar con datos, una vez establecidos, mientras la relación de los datos con lo que representan se deja a otros campos y ciencias».

¿Quién diría que el término *ciencia de datos* tiene más de cuarenta años, verdad?

A pesar de que la definición es poco clara y que sus ideas tardaran en ser entendidas por la comunidad científica y empresarial, se puede considerar que este es uno de los primeros intentos de recoger todo el trabajo realizado en este nuevo campo.

La década de los setenta es, de hecho, una de las más importantes en el desarrollo de nuevos modelos estadísticos que aprovechan el nuevo paradigma computacional. Muchas de las técnicas utilizadas hoy en día siguen basándose en los avances teóricos de esta década prodigiosa en la cual se fundó la Asociación Internacional para la Estadística Computacional (IASC, por sus siglas en inglés) en 1977. Su origen se encuentra en la voluntad de enlazar, precisamente, las metodologías estadísticas tradicionales y la tecnología moderna que aportaban los ordenadores. Pero iba más allá, también buscaba integrar a los expertos y especialistas de cada dominio para convertir estos datos en información y conocimiento. Este segundo paso es lo que define la ciencia de datos.

Aquel mismo año, Tukey (1977), que seguía investigando los primeros pasos del nuevo campo científico, publicó *Exploratory Data Analysis*, en el que vuelve a destacar la importancia de aprovechar los datos para seleccionar las hipótesis en cualquier experimento y hace también un llamamiento a combinar los enfoques exploratorios y confirmatorios en el análisis de datos para obtener mejores resultados. A pesar de la promesa teórica de los avances registrados en la década de los setenta, la limitada capacidad computacional real y la dificultad de acceso a los centros de cálculo y ordenadores retardaron el proceso de integración. Prácticamente en paralelo se empezaba a desarrollar otra técnica relacionada: el *data mining* o minería de datos. Considerada una práctica repudiable durante los sesenta y setenta, las malas lenguas lo denominaban «pesca de datos» o «dragado de datos», un título despectivo que se refería al análisis de datos sin hipótesis previa. Al inicio de los ochenta, no obstante, algunos analistas de bases de datos empezaron a cambiar la connotación del término hacia la más positiva «experimentación», hablando de *database mi-*

ning. El término, no obstante, acabó volviendo a *data mining* de la manera más americana posible, puesto que *database mining* era una frase que ya estaba registrada por una compañía.

Más tarde, los esfuerzos de un conocido científico de datos, Gregory Piatetsky-Shapiro, para avanzar en la extracción de información de las bases de datos desembocaron en una línea de investigación que bautizó con el nombre de *Knowledge Discovery in Databases* (KDD), y eso llevó a la organización en 1989 de la primera conferencia sobre la busca de conocimiento en bases de datos, hoy conocida como *ACM SIGKDD Conference of Knowledge Discovery and Data Mining (KDD)* y que se celebra anualmente.

1.2. Inteligencia de negocio

En los años noventa se produce una transición en las empresas que continuará hasta el cambio de milenio. Basándose en los métodos estadísticos desarrollados durante los setenta y las técnicas de minería de datos de los ochenta, las tecnologías y aplicaciones más habituales de la industria (todavía hoy) se empezaron a popularizar. Esta primera aproximación de toma de decisiones basada en datos (la inteligencia de negocio) parte de los fundamentos de la gestión de las bases de datos y, por lo tanto, se basa en información estructurada, recogida por las mismas compañías (habitualmente con sistemas y servidores de software anticuado, como los ordenadores centrales que controlan procesos de producción) y que se almacenan en sistemas comerciales de gestión de bases de datos relacionales (RDBMS). De esta necesidad de gestión y almacenamiento de datos nace el BI. Para poder tomar decisiones informadas, aparece la necesidad de transformar estos datos masivos y de bajo nivel en información más inteligible para los equipos de dirección y de toma de decisiones.

En este entorno, el diseño de *data marts* y herramientas para la extracción, la transformación y la carga de datos (denominados ETL) son esenciales para convertir e integrar los datos específicos de cada negocio. Del mismo modo, los analistas de datos necesitan herramientas para explorar los datos; herramientas que obtienen de lenguajes de consulta de bases de datos y de los cubos OLAP (*Online Analytical Processing*), pensados para facilitar el acceso a los datos de trabajo, y que son, también, complementados por herramientas gráficas rudimentarias que permiten explorar algunas características de los datos. De cara a la dirección, cabe destacar especialmente el desarrollo del *Business Performance Management* (BPM), con el apoyo de los registros de resultados, pero, sobre todo, de los cuadros de mando, inspirados en la propuesta de cuadro de mando integral de Kaplan y Norton (1996), que ayudan a analizar y visualizar una serie de métricas de rendimiento. Además de todas estas herramientas orientadas a generar informes, también hay que añadir la expansión de las técnicas mencionadas anteriormente, tanto estadísticas como de minería de

ETL

Extracción, Transformación y Carga (*Load*) es el nombre que se da al proceso que permite mover datos de múltiples fuentes, transformarlos y limpiarlos para posteriormente cargarlos de nuevo en un proceso de negocio.

OLAP

El objetivo de los cubos OLAP es agilizar la consulta de grandes cantidades de datos. Se puede imaginar como una reordenación del contenido de las bases de datos relacionales, una vista, para hacer más eficientes las operaciones de consulta.

datos, para asociar, segmentar, agrupar y clasificar datos, preparar modelos de regresión, de detección de anomalías e, incluso, hacer predicciones cada vez en más aplicaciones de negocio.

BusinessWeek publicó, en septiembre de 1994, un artículo de portada sobre lo que denominó «marketing de base de datos». Lo más interesante es un pasaje que transpira indecisión y que sirve para ilustrar la década:

> «Las empresas están recogiendo montañas de información sobre ti, procesándola para prever la probabilidad de que compres un producto, y utilizando este conocimiento para construir un mensaje de marketing calibrado para que lo hagas [...] Un empujón de entusiasmo causado por la expansión de los escáneres de compra en los ochenta acabó en decepción general: muchas empresas estaban bastante sobrepasadas por la elevada cantidad de datos como para hacer nada útil con la información [...] Sea como fuere, muchas compañías piensan que no tienen alternativa a desafiar la barrera del marketing de base de datos».

El término *ciencia de datos* empieza a utilizarse de forma gradual aproximadamente al final de la década. En primer lugar, en 1996, la conferencia bianual de la Federación Internacional de Sociedades de Clasificación (IFCS) es titulada «Data science, classification, and related methods». El mismo año, Fayyad y otros (1996) hablan de la diferencia entre la simple minería de datos (o aplicación de algoritmos) y la obtención de información a partir de bases de datos, que implica pasos adicionales como la «preparación de los datos, selección, limpieza, incorporación de información otras fuentes e interpretación de los resultados de la minería de datos», esenciales para obtener conocimiento útil a partir de los datos.

Jeff Wu, en su discurso de inauguración del curso 1997 en la Universidad de Michigan, propuso cambiar el nombre de la estadística por el de *ciencia de datos* y el de estadista por *científico de datos*. Pero la frase que mejor ilustra el paso al nuevo milenio proviene de Jacob Zahavi (diciembre de 1999), que añade dos elementos importantes, los inicios de los datos masivos y la influencia creciente de internet:

> «La escalabilidad es un problema gigante para la minería de datos [...] los métodos estadísticos convencionales funcionan bien en conjuntos de datos pequeños. Pero las bases de datos de hoy en día pueden estar formadas por millones de filas y una pila de columnas de datos [...] Otro reto técnico es desarrollar modelos que puedan hacer una mejor tarea analizando datos, detectando relaciones no lineales e interacciones entre elementos [...] Es posible que se tengan que desarrollar herramientas especiales de minería de datos para la toma de decisiones en las páginas web».

1.3. Internet y la web 2.0

A pesar de que internet sea bastante más antiguo, no es hasta principios del 2000 cuando las oportunidades de recogida de datos y de analítica empiezan a aparecer. Primero, en forma de la llamada web 1.0, caracterizada por los buscadores (como Google y Yahoo), pero también por el emergente comercio electrónico (eBay y Amazon), que obtienen datos de primera mano de sus usuarios. Ya no se trata tan solo de trabajar con los datos tradicionales de productos y de negocio de las bases de datos relacionales, sino que el contenido en línea

proporciona detalles por IP de cada usuario sobre sus búsquedas y su interacción con las páginas, datos que son recogidos por medio de registros detallados (*logs*) y galletas (*cookies*) y que representan una oportunidad sin precedente de identificar las necesidades de cada cliente y nuevas oportunidades de negocio. Muchas de estas empresas han convertido la explotación de los datos en el núcleo de los servicios que ofrecen y en su ventaja competitiva.

La cantidad de información que se puede extraer de la gran red es inacabable. El análisis de clics de clientes, por ejemplo, origina herramientas de análisis web como Google Analytics, orientadas a descubrir patrones de comportamiento y de compra. De este análisis derivan posteriormente el diseño de páginas web, la optimización del marketing y del posicionamiento, el análisis de mercado y las recomendaciones. Así se hace visible como la ciencia de datos no solo aprovecha la extracción de datos para tomar decisiones informadas en el ámbito de negocio, sino que acaba transformando todo lo que le rodea, desde la forma de comprar hasta la influencia que tiene sobre las páginas web que los usuarios visitan.

El problema, para decirlo de alguna manera, es que el incremento se excede con la explosión del contenido generado por usuarios. Los foros, grupos en línea, blogs, plataformas sociales e incluso entornos virtuales llenan la web de finales de la década de los 2000 de inabarcable contenido, difícilmente tratable con la capacidad de procesamiento del momento. En el ámbito del marketing se habla de la oportunidad que este tipo de contenido supone para los negocios de observar y ser partícipes de la conversación entre proveedores y usuarios, cambiando el paradigma tradicional de una sola dirección.

Si la década empezó con los planes de Cleveland (2001) sobre como los científicos de datos tendrían que prepararse para los requerimientos del futuro, acabó con la popularización (y moda) del término, que se suele atribuir a DJ Patil y Jeff Hammerbacher, de LinkedIn y Facebook, respectivamente. Del año 2009 también cabe destacar el regreso (para quedarse) de las bases de datos NoSQL (o no relacionales).

1.4. Ciencia de datos

El nacimiento de la disciplina de la ciencia de datos tal y como la conocemos hoy es a principios de la década de 2010. La clave es, en realidad, una tormenta perfecta de acontecimientos.

Hadoop

El proyecto Hadoop nace en 2006 y parte de Nutch, un intento de indexar la totalidad de la web.

En primer lugar, la existencia de datos masivos, provenientes ya no solo de los portales de internet sino de múltiples sensores de cualquier aparato (lo que se conoce como internet de las cosas o *Internet of things*, IoT). Es el llamado *big data*, que no es nada más que el nombre que recibe la cantidad masiva de datos de las que se dispone hoy en prácticamente cualquier aplicación. Según IBM, el 90% de los datos de los que se dispone en 2018 han sido generados

únicamente en los dos años anteriores. Sí, esto significa básicamente que entre 2016 y 2017 se generaron más datos que desde el principio de la humanidad hasta el año 2015.

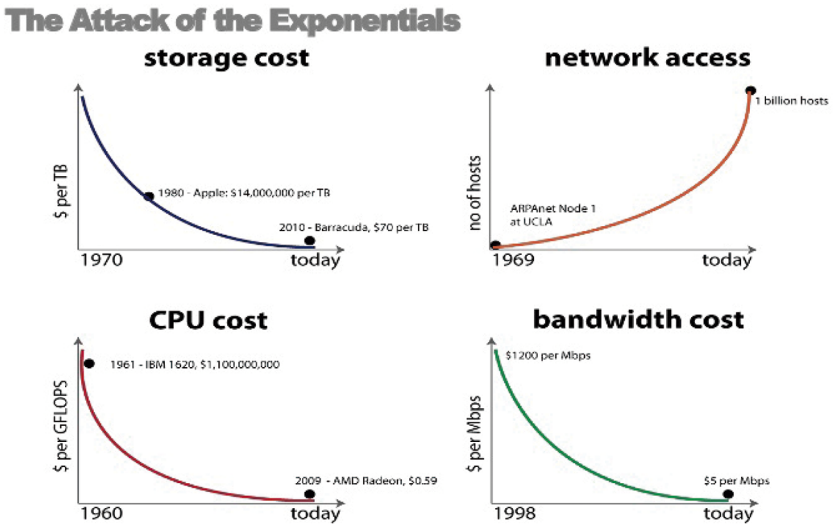
En segundo lugar, arquitecturas de procesamiento distribuido como Hadoop y HDFS (y el ecosistema *open source* que lo complementa), de las que hablaremos más adelante, que permiten procesar grandes cantidades de información en clústeres de computadoras convencionales[1].

(1)En 2018, Google procesa 40.000 búsquedas por segundo.

Pero, sobre todo, lo que más destaca de estos últimos años es el llamado ataque de los exponenciales (ver la figura 1):

- El coste del almacenamiento ha bajado de forma drástica.
- El acceso a internet se ha multiplicado (aproximadamente medio mundo, prácticamente 4 mil millones de personas acceden habitualmente a la red).
- El coste de las CPU se ha reducido, también, exponencialmente.
- El coste de la anchura de banda ha dejado de ser relevante.

Figura 1. Gráficos que muestran la evolución exponencial de algunos parámetros críticos (2011)



Fuente: Building Data Start-Ups: Fast, Big, and Focused (2011). *slideshare.net*

En resumen, unos costes económicos cada vez menos relevantes hacen que muchos problemas que antes no era rentable estudiar pasen a serlo. Las redes de sensores y la generación de datos en línea hacen más fácil el acceso a los datos. Y, por si fuera poco, a la bajada de los costes se le une el *cloud computing* o, lo que es lo mismo, la posibilidad de «alquilar» capacidad de computación

para hacer frente a las necesidades puntuales de cualquier empresa, sin tener que invertir en grandes equipos o servidores. El *software as a service* (SaaS) hace el resto.

Se puede hablar, pues, de tres grandes elementos que han cambiado en el análisis de datos orientado a ofrecer servicios:

1) Ya no son solo datos, sino que son datos obtenidos de forma masiva, rápida y con un procesamiento que puede llegar a ser similar al tiempo real en algunos casos. El abaratamiento de la memoria RAM también ha colaborado.

2) El análisis es rápido y a gran escala. Ya no es solo que lenguajes como R o Python se hayan expandido y formen parte del núcleo de las herramientas de análisis más habituales, sino que también han aparecido nuevas arquitecturas distribuidas como Spark con potenciales elevadísimos y que permiten trabajar con petabytes[2] de datos.

[(2)]Un petabyte equivale a un millón de gigabytes, es decir, 1 PB = 10^6 GB = 10^{15} bytes.

3) Los servicios ofrecidos pueden ser más específicos y más centrados en cada aplicación concreta.

SaaS

El SaaS es un modelo de distribución de software en el que este está alojado directamente en el servidor del proveedor (y el cliente accede a él mediante internet).

Durante los últimos años, la ciencia de datos ha crecido y ha incluido tanto el mundo académico como los negocios y las organizaciones del mundo entero. Ya hoy en día es una realidad que utilizan los gobiernos, los ingenieros, astrónomos y médicos, entre muchos otros, por todas partes. El paso al *big data* no representa solo un cambio de escala, sino que también ha traído nuevas maneras de entender y procesar los datos, cambiando la forma de estudiarlos y analizarlos.

Apache Spark

Apache Spark es un motor unificado de analítica orientado al procesamiento de datos masivos, con módulos destinados a la reproducción en continuo (*streaming*), SQL, aprendizaje automático y procesamiento de grafos.

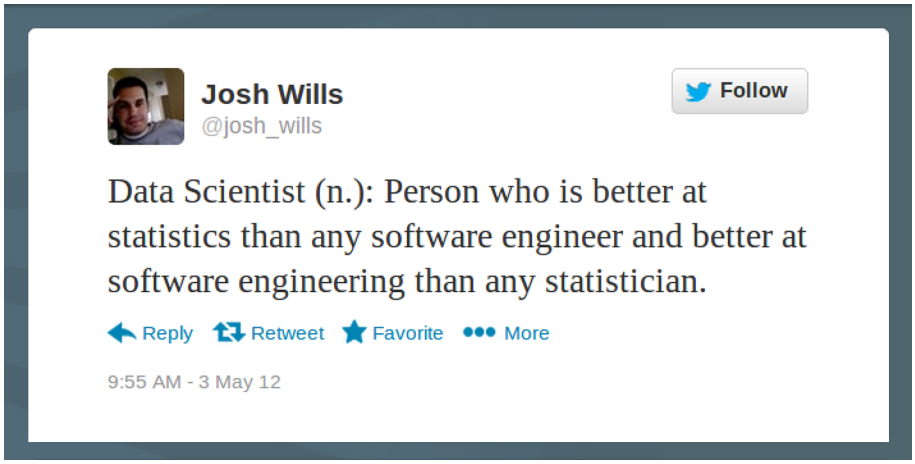
Así pues, la ciencia de datos se ha convertido en parte importante de la investigación tanto empresarial como académica. Los ámbitos son amplios, pero incluyen la traducción automática, la robótica, el reconocimiento de voz, la economía digital y los motores de busca, entre otros. Las disciplinas también son transversales: desde la biología, la medicina y la salud, hasta las humanidades y las ciencias sociales. El análisis que proporciona la ciencia de datos influye, en el día a día, en la economía, la política y las finanzas. Los últimos años han proporcionado, además, el desarrollo y la mejora de técnicas que hasta no hace tanto eran costosas en el ámbito de producción. El aprendizaje automático (*machine learning*) ha liderado las técnicas utilizadas (y es donde probablemente se encuentran algunas de las más maduras), pero también es notable el crecimiento que han experimentado nuevas áreas como el aprendizaje profundo (*deep learning*), las redes neuronales o el análisis de redes sociales. Todavía hay, no obstante, mucho camino para recorrer. Muchos problemas son constantes y complejos de resolver e, igualmente, ¡no dejan de aparecer nuevos problemas!

2. El rol del científico de datos

2.1. ¿Qué es un científico de datos?

Josh Wills, de Slack, definió el científico de datos en un famoso tuit de la forma siguiente:

Figura 2. Tuit de Josh Wills sobre el científico de datos



Fuente: https://twitter.com/josh_wills/status/198093512149958656

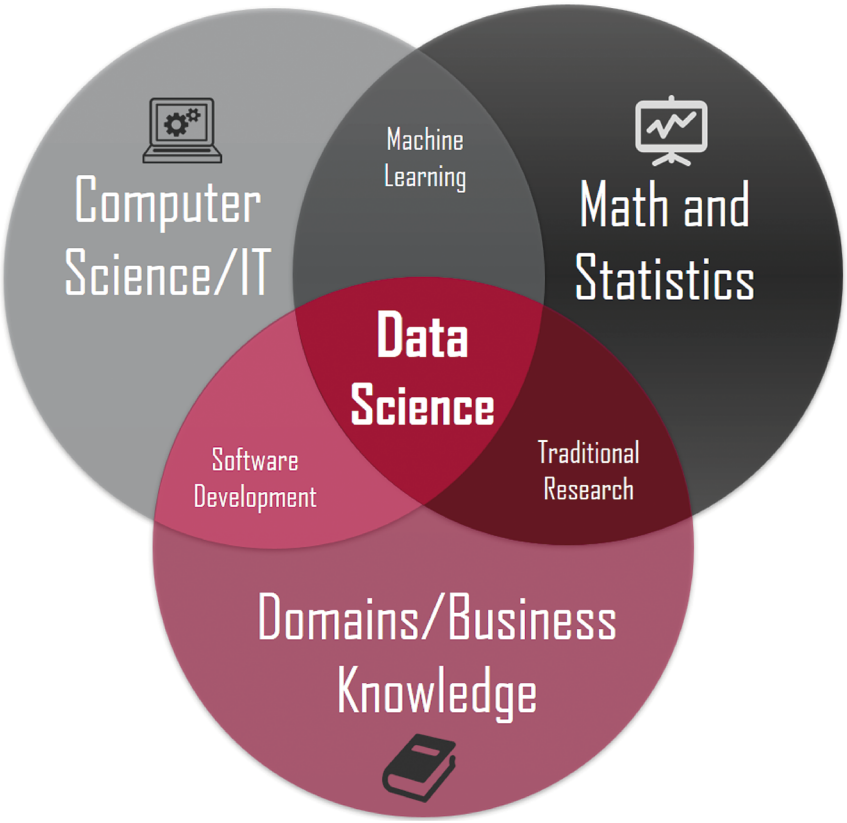
Su definición (en clave de humor) no se sitúa demasiado lejos de la realidad, como veremos a continuación, aunque es cierto que el rol del científico de datos es algo más complejo que el que se extrae de estos ciento cuarenta caracteres.

De la disponibilidad de grandes volúmenes de datos aparece la necesidad, especialmente en el ámbito de negocio, de utilizarlas para ganar una ventaja competitiva. Queda claro que las empresas y organizaciones que sean capaces de utilizar de manera efectiva este tipo de información serán también propensas a tomar mejores decisiones y ponerse por delante del resto de competidoras.

Para inferir información razonable y útil de tal cantidad de datos aparece la necesidad de contar con profesionales con un conjunto de habilidades y aptitudes que no existían. Estos perfiles, que se denominan científicos de datos, combinan básicamente tres disciplinas básicas y las dominan en profundidad:

- matemáticas y estadística
- ciencias de la computación y programación
- conocimiento del área de negocio

Figura 3. Diagrama de Venn del *data scientist* (Drew Conway)



Fuente: https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

Tal como se puede observar en la figura 3, la intersección entre cada pareja de las disciplinas requeridas proporciona un perfil que ayudará a entender cómo se relacionan estas habilidades. Así:

- Un profesional que tenga habilidades de programación y de matemáticas y estadística es un perfil de aprendizaje automático. En realidad, se trata de un perfil que a veces se considera peligroso o que tiene poco sentido en una empresa, puesto que extraer conclusiones sobre un dominio desconocido para el investigador puede ser contraproducente.

- El dominio de la informática y el conocimiento del sector (o dominio) en el que se trabaja lleva a ser desarrollador de software de aquel ámbito, normalmente de manera específica.

- La utilización de las habilidades matemáticas y estadísticas para la investigación en un dominio concreto es lo que ha hecho desde el principio la investigación llamada tradicional, contrastando hipótesis, por ejemplo.

- En el punto de encuentro entre las tres disciplinas se encuentra el científico de datos, habitualmente considerado una *rara avis*.

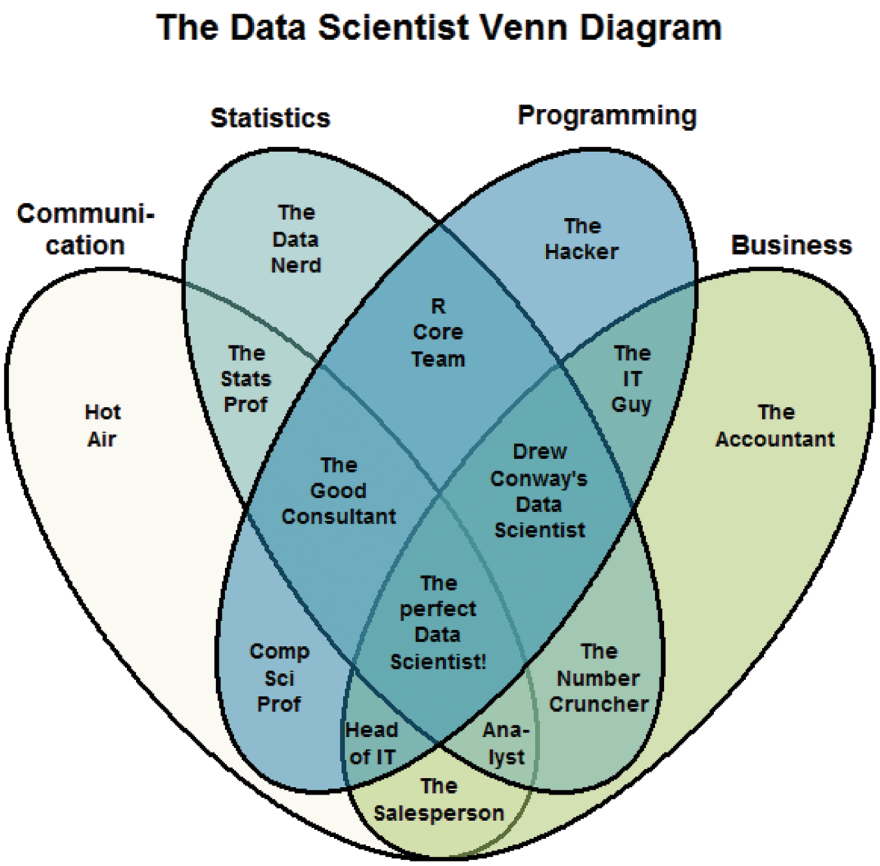
El problema de este perfil profesional es, como se puede divisar, que tradicionalmente las dos disciplinas base (la rama computacional y la matemática) se han tratado por separado y es extraño encontrar informáticos con una base

matemática profunda (es algo menos extraño encontrar matemáticos con buena competencia informática, no obstante). Además, tanto unos como otros han ocupado, habitualmente, posiciones muy concretas en las empresas, no siempre cerca de la actividad empresarial. Por este motivo, los profesionales que reúnen habilidades avanzadas en las tres disciplinas reciben, en algunos sectores, el nombre de «unicornios», cuestionando de forma jocosa su existencia. El camino habitual del científico de datos es, pues, el de conseguir dominar dos de las disciplinas, primero, y, después, ya sea mediante formación adicional o con la entrada al mundo laboral, desarrollar la tercera. Hasta aquí la definición tradicional, propuesta por Drew Conway a principios de la década.

No obstante, algunos han sugerido que, a falta de las habilidades requeridas, los científicos de datos necesitan, y cada vez más, destacar en una cuarta disciplina: la comunicación. El razonamiento es sencillo: de poco sirve dominar la estadística, la programación y la actividad empresarial si no se es capaz de explicar los resultados obtenidos y, todavía más importante, modular este mensaje para cada uno de los departamentos o actores implicados. Presentar las conclusiones a la dirección, por ejemplo, puede requerir herramientas de visualización claras e inteligibles; explicar el modelo para su implementación a los desarrolladores de software es una historia completamente diferente, a pesar de que el trabajo haya sido el mismo.

En la figura 4 se muestra cómo quedaría el diagrama de Venn transformado con esta nueva variable.

Figura 4. Diagrama de Venn expandido del científico de datos



Fuente: https://datascience.stackexchange.com/questions/2403/data-science-without-knowledge-of-a-specific-topic-is-it-worth-pursuing-as-a-ca

Es interesante fijarse en cómo la comunicación modifica ligeramente la idea de los perfiles del primer diagrama de Venn; así, un profesional que combine la alta capacidad de comunicación con:

- Estadística, sería un profesor de estadística.
- Programación, sería un profesor de ciencias de la computación.
- Actividad empresarial, sería el perfil del vendedor.
- Programación y estadística, sería un buen consultor.
- Estadística y negocio, sería lo que se denomina un analista de datos.
- Programación y negocio, sería director de sistemas dentro de la empresa.
- Y la unión de las cuatro habilidades es lo que se espera cada vez más del científico de datos.

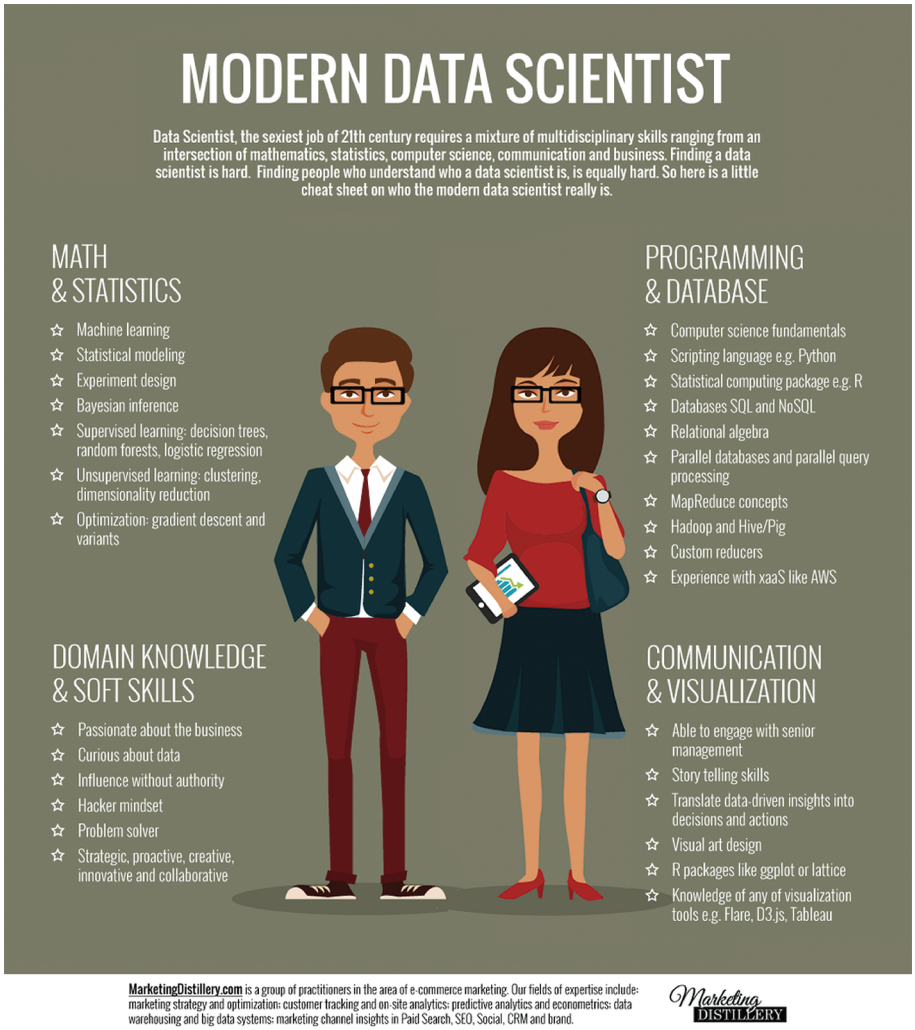
¿Cuál es pues la diferencia entre un científico de datos y un analista de datos? En realidad los perfiles tienen áreas en las que se asemejan, pero también difieren en algunas otras:

- Un analista de datos tiene como objetivo la interpretación de los datos para obtener conocimiento útil para el negocio u organización. Su punto fuerte tiene que ser la estadística (y la matemática, por extensión), pero también debe tener capacidades moderadas en conocimiento del negocio

y programación (para, como mínimo, ser capaces de transformar los datos). En resumen, un analista de datos recoge, procesa y aplica algoritmos estadísticos a datos estructurados para responder a una serie de preguntas predeterminadas por el mismo negocio.

- La misión del científico de datos es similar a la del analista: obtener el mismo tipo de conocimiento. No obstante, el científico de datos debe enfrentarse también a complicaciones técnicas (volúmenes de datos más grandes y velocidades elevadas de creación de los mismos) y de arquitectura (datos sin estructura). De todo este conjunto tiene que ser capaz de identificar primero las preguntas o hipótesis que se deben hacer (que en el caso del analista ya suelen estar determinadas) y complementar los datos de los que dispone con otras fuentes. Además, limpia y transforma los datos para prepararlos para el procesamiento y crea nuevos algoritmos y búsquedas. Por si fuera poco, como se ha visto anteriormente, también necesita habilidades comunicativas, narrativas y de visualización para ser capaz de compartir los resultados a cualquier nivel dentro de la organización.

Figura 5. Detalle de las habilidades del científico de datos



Fuente: https://towardsdatascience.com/my-journey-from-physics-into-data-science-5d578d0f9aa6

Se podría resumir de la forma siguiente: mientras el analista de datos busca conclusiones y alertas sobre métricas que la compañía considera críticas, el científico de datos construye nuevos modelos y busca conocimiento sobre indicadores que la compañía todavía no sabe que son importantes.

Así que, simplificando otra vez, también se podría decir que el trabajo del analista de datos se critica en el día a día de la compañía (supervivencia diaria), mientras que el del científico de datos está más orientado al medio-largo plazo (ventaja competitiva).

2.2. ¿Qué hace un científico de datos?

Una vez hemos visto el rol del científico de datos es más fácil entender sus funciones. El científico de datos es un perfil que disfruta de cierta libertad en su trabajo diario; debe tener una mentalidad abierta y curiosa, pero también un cierto escepticismo respecto del conocimiento «establecido». Aplica el método científico, así que es importante que conozca la metodología adecuada para la generación y comparación de experimentos. Tiene que ser capaz de programar y de modificar código, de interactuar tanto con el personal de sistemas como de dirección y de crear historias a partir de datos.

Su trabajo diario se resume en diez puntos:

1) Hacerse (buenas) preguntas: ¿Qué no se sabe? ¿Qué se querría saber? ¿Qué sería útil saber? ¿Cómo se podría saber?

2) Definir y poner a prueba hipótesis, mediante experimentos que sigan el método científico.

3) Extraer, obtener, hacer *scraping*, muestrear, etc., datos relevantes para el negocio.

Scraping

Consiste en extraer datos de formatos legibles por humanos (como una página web) para su posterior procesamiento.

4) Adaptar los datos a sus necesidades de forma, distribución y formato.

5) Descubrir nuevos datos a partir de la exploración y de métricas desconocidas.

6) Modelar tanto los datos como los algoritmos.

7) Entender relaciones entre datos.

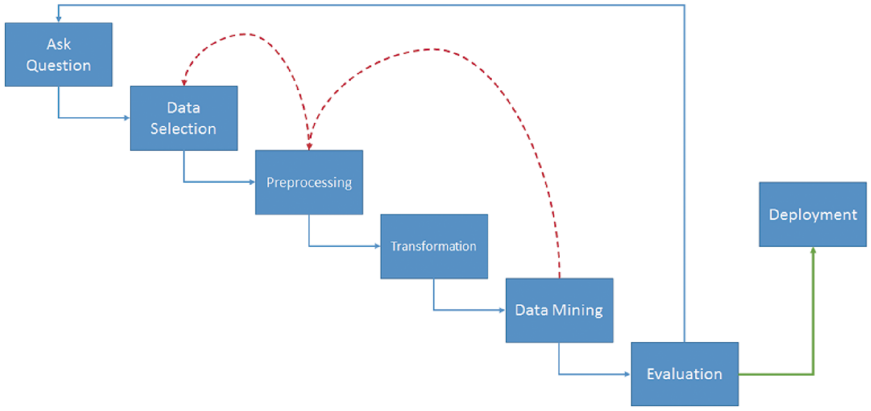
8) Aplicar aprendizaje automático de manera controlada e informada.

9) Crear programas y productos que proporcionen conocimiento a la empresa.

10) Explicar historias a partir de los datos, con una narrativa fácil de comprender.

Este proceso se puede representar como en la figura 6.

Figura 6. El proceso de la ciencia de datos



Fuente: https://datascienceex.files.wordpress.com/2015/12/datascienceprocess.png

2.2.1. Hacerse (buenas) preguntas

El trabajo del científico de datos empieza definiendo el problema que se quiere resolver o el planteamiento que se quiere comprobar. Un buen científico de datos intenta eliminar cualquier sesgo personal que pueda tener; no se trata de comprobar aquello que uno piensa, sino de responder una pregunta que se plantea. Es importante tener también en cuenta los objetivos empresariales. Tener datos y más datos no implica que sean útiles, así que entre las funciones del científico de datos se incluye la de convertirlos en información accionable. Es también el momento de escribir un plan, un mapa, que llevará desde los datos hasta un estado final deseado.

2.2.2. Selección de datos

Una vez realizado el planteamiento inicial y diseñado el experimento, es el momento de seleccionar los datos que se utilizarán; aquí es importante pensar no solo en los datos internos de los que dispone la organización, sino también en los internos de los que todavía no dispone (y que se pueden intentar obtener) y en los externos que se pueden incorporar en el proceso, algunos a coste cero (datos abiertos) y otros de pago (datos de mercado, por ejemplo).

2.2.3. Preprocesamiento

Un paso importante es observar los datos. Y no se trata de empezar a tocarlos y transformarlos, sino de evaluarlos, comprobar si están completos, qué tipo de distribución siguen, qué problemas tienen... Obviar este paso puede ser un

desastre para el científico de datos, puesto que podría dejar pasar datos de dudosa calidad al sistema y elaborar un modelo con más problemas de los deseados. El preprocesamiento puede derivar en cambios en la selección de datos (descartando, por ejemplo, datos que sean manifiestamente erróneos o que tengan excesivos valores no informados).

2.2.4. Transformación

Llegado este paso, el científico de datos ya tiene clara la dirección por la que va y los datos que tiene para trabajar. Así, este es el momento de transformar los datos en bruto y adaptarlos para poderlos utilizar en los algoritmos deseados. Este paso es también crítico, porque los datos suelen estar en un estado poco óptimo. Datos de diferentes formatos, incompletos, con problemas de distribución y de escala... Idealmente, al final del proceso de transformación, el científico de datos tendrá un conjunto de datos a punto para pasar al paso siguiente.

2.2.5. Descubrimiento de conocimiento (o minería de datos)

Bajo este epígrafe se encuentra, en realidad, la cara más visible de la ciencia de datos; curiosamente, también se podría decir que es una de las menos complejas. Aquí se aplican los algoritmos deseados. Se tiene que decidir si se quiere usar un algoritmo supervisado o no supervisado, si se quiere clasificar o prever, etc. Tanto los objetivos iniciales como los datos disponibles tienen influencia sobre los algoritmos aplicables. ¿Por qué se considera una de las menos complejas si el fondo matemático de la mayoría de algoritmos es realmente complejo? Porque la mayor parte de los algoritmos están bastante definidos en la actualidad, así que raramente se baja al plan matemático. Los algoritmos, además, ya están disponibles y solo hay que aplicarlos sobre el conjunto de los datos que se deben analizar (no hay que volverlos a escribir y verificar, se pueden reaprovechar y adaptar con facilidad, etc.). Algunas voces opinan que esta dinámica lleva a un cierto comportamiento «conservador», es decir, hace más difícil la aparición de cambios bruscos en las técnicas utilizadas. En todo caso, aquello que queda claro es que el resto de fases requieren mucha más experiencia y son, por lo tanto, más difíciles de dominar.

2.2.6. Evaluación

El modelo resultante es simplemente esto, un modelo. Al científico de datos le corresponde comprobar y entender si el modelo responde a las cuestiones planteadas inicialmente. Es por eso que el conocimiento del dominio o del negocio es especialmente importante para ser capaz de interpretar los resultados obtenidos. La evaluación se realiza al final de cada iteración y también es momento de considerar si es necesario volver a empezar, sea desde la fase que sea, para mejorar el resultado obtenido o si, en caso contrario, el modelo es aceptable para su paso a producción.

2.2.7. Paso a producción

Es importante tener siempre en mente que el motivo principal de todo el proceso es que el conocimiento obtenido sea útil y que, por lo tanto, pueda pasar a producción. La mayoría de los que utilizarán el nuevo modelo no son habilidosos en la ciencia de datos y, por lo tanto, se tiene que poder presentar el resultado de una manera inteligible. Puede ser mediante visualizaciones personalizadas, cuadros de mando o, incluso, como entrada para otros sistemas de la compañía. Este paso es lo que diferencia la ciencia de datos de otras disciplinas exploratorias, cuyo objetivo no es necesariamente un producto final accionable por una compañía.

2.2.8. Volver a empezar

El proceso (o el trabajo) del científico de datos es una constante iteración que no acaba nunca. El paso a producción se puede producir en cualquier momento en el que se considere que el balance entre el coste de una nueva iteración y la utilidad del modelo ya sea adecuado para el negocio. Pero hay que entender que un modelo no acaba nunca, puesto que siempre se pueden añadir nuevos datos, siempre aparecen nuevos datos, siempre se generan nuevos algoritmos. Así, es casi más importante generar una metodología, un proceso, que sea repetible, reproducible y robusto, que no un modelo que funcione de fábula en un primer momento pero que sea frágil ante cualquier interferencia. Porque las interferencias son el día a día de la actividad empresarial y las cosas cambian constantemente, así que hay que construir los modelos con esta idea en mente.

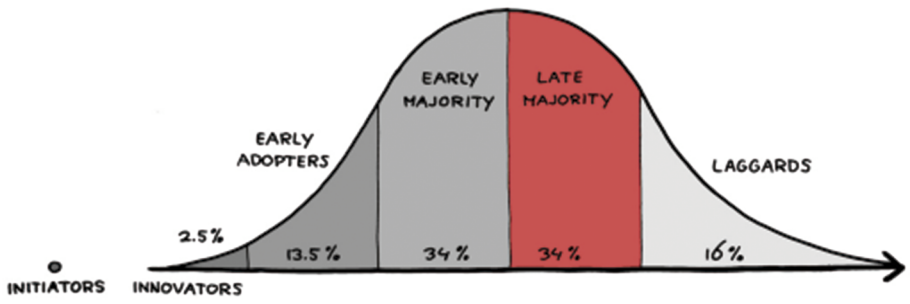
2.3. La caja de herramientas del científico de datos

La evolución y la expansión del número de herramientas y de empresas que ofrecen servicios relacionados con la ciencia de datos y los datos masivos es solo comparable a la explosión de la misma disciplina. En el ya tradicional «Big Data Landscape» de 2018 (figura 8) hay que destacar, también, la inclusión del término *AI* ('inteligencia artificial'), que empieza a estar casi más «de moda» que el término *datos masivos* (y eso que en la mayoría de casos la gente los utiliza para describir la misma realidad).

«Big Data Landscape»

Para ver la imagen de forma más detallada se puede visitar la página personal de su creador: http://mattturck.com/bigdata2018/.

Figura 7. El ciclo tradicional de cambio tecnológico



Fuente: http://mattturck.com/bigdata2018/

Las tecnologías que trabajan con datos (aprendizaje automático, ciencia de datos, inteligencia artificial) siguen creciendo, cada vez más eficientes y con más presencia en negocios de todo el mundo. Es también la época de la llamada «transformación digital», que puede parecer extraña, puesto que los ordenadores son indispensables en la mayor parte de empresas desde hace más de treinta años, pero que muestra como la mayoría de industrias tradicionales están ahora comprometidas a convertirse en empresas que se basan en la información de los datos. En la figura 7, que refleja el ciclo tradicional de cambio de tecnología, se podría decir que la época actual se sitúa justo al final de la *early majority* o, dicho de otro modo, de aquella primera mitad de empresas que ya están o han cambiado de paradigma y que convencen o fuerzan al resto a seguir el mismo camino. Cabe destacar, también, el incremento de la importancia de la nube con un crecimiento imparable de los proveedores principales (AWS de Amazon, Azure de Microsoft, Google Cloud Platform e IBM) y la creciente integración de las herramientas de aprendizaje automático y *data engineering* en las mismas plataformas que ofrecen.

Agenda Digital

Tal es la importancia de este tema que la transformación digital tiene incluso su propio ministerio, el de Agenda Digital, bajo el paraguas de Industria.

Data engineering

A pesar de que no se haya hablado de él específicamente, el *data engineering* es la disciplina encargada de dar forma a los datos.

El ecosistema de la ciencia de datos es inmenso (figura 8); tan inmenso, de hecho, que lo habitual es que cada científico de datos se especialice en una pequeña parte y que escoja, de entre todas las posibilidades, las que mejor encajan con su caja de herramientas. A continuación se resumen en grandes grupos (explicarlo con detalle es poco útil en este punto y los nombres cambian año tras año):

- Las herramientas de infraestructura, que son las que proporcionan el entorno de trabajo y la estructura funcional. Aquí se encuentran los proveedores de la nube, las bases de datos tradicionales, NoSQL y NewSQL, las de grafo, herramientas de integración y transformación de los datos, de almacenamiento, de monitorización...

- Las herramientas de analítica, entre las que se encuentran las plataformas de análisis y de ciencia de datos, de *business intelligence*, de visualización, de aprendizaje automático, de tratamiento del lenguaje, de análisis social y de comercio electrónico, entre otros.

- Cabe destacar esta nueva tendencia de integrarlas ambas, las herramientas de infraestructura más las de analítica, en algunas de las herramientas existentes (AWS, Google Cloud, Azure, SAS, etc.).

- También hay un conjunto de aplicaciones más orientadas a cada:
  - ámbito: aplicaciones para ventas, marketing, servicio al cliente, recursos humanos, legales, finanzas, seguridad, etc.

  - industria: aplicaciones diseñadas para la educación, la publicidad, los gobiernos, los seguros, las finanzas (tanto de inversión como convencionales), la salud, el transporte o la agricultura

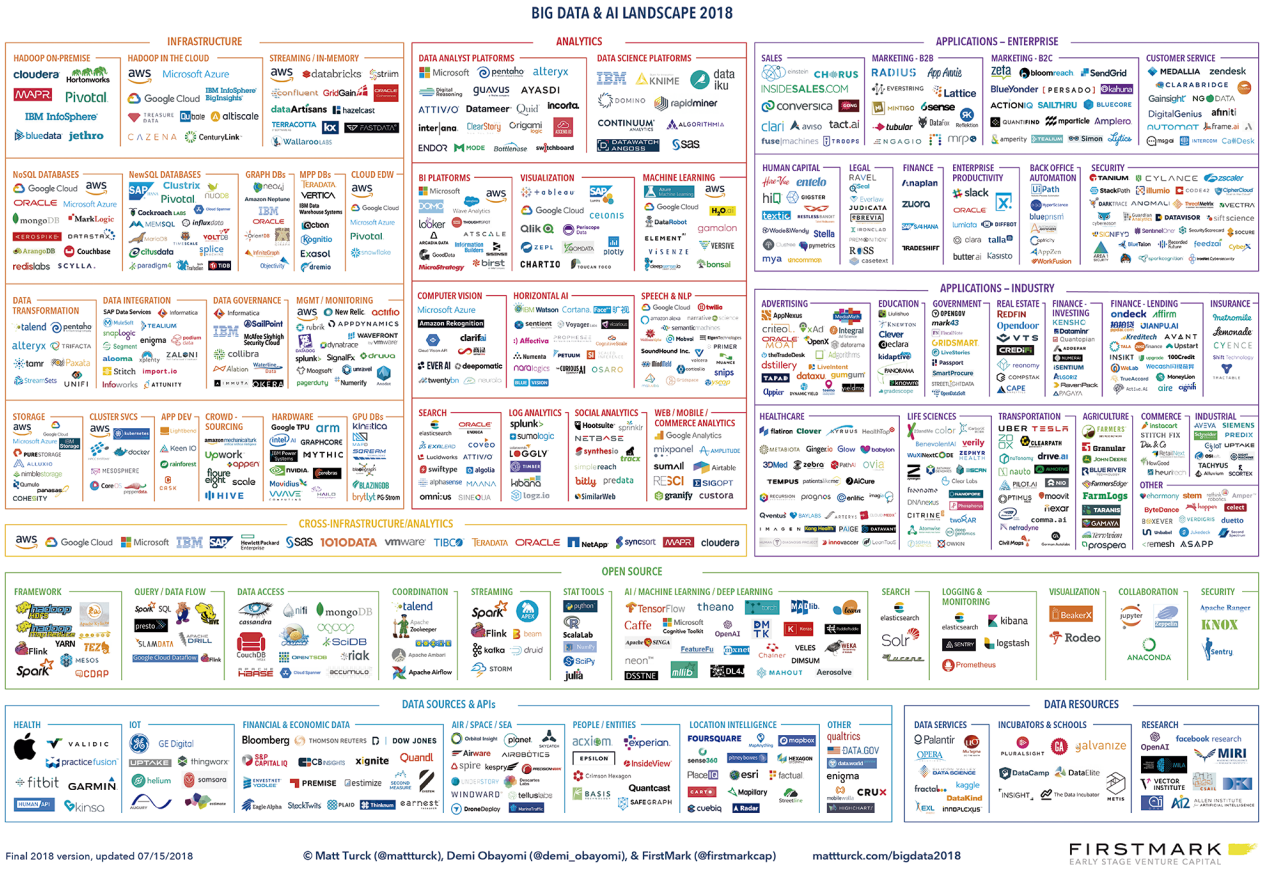
API

Corresponde a *application programming interface* y no son más que un conjunto de funciones que nos permiten consultar una aplicación u otro servicio y obtener respuestas de manera directa.

- El bloque más interesante y más popular en el ámbito global es el llamado *open source* o herramientas de código abierto. Aquí se sitúa todo el ecosistema Hadoop y Spark, Hive, bases de datos NoSQL, herramientas de reproducción en continuo, estadística (R y Python) y aprendizaje automático. Es lo que el científico de datos se encontrará de forma más habitual.

- Finalmente, algunas herramientas destinadas a la obtención de datos, ya sean APIs o repositorios.

Figura 8. «Big Data Landscape» (2018)



Fuente: http://mattturck.com/bigdata2018/

3. Ámbitos de la ciencia de datos

Entendida la importancia de la ciencia de datos y el rol del científico de datos, es el momento de repasar brevemente algunos de los ámbitos que se han visto más transformados en los últimos años por la emergencia de esta disciplina, así como sus perspectivas de futuro.

3.1. Marketing

El marketing es un sector que necesita adaptarse constantemente al consumidor; cuando los hábitos de consumo cambian o evolucionan, toca seguir el mismo camino para no perder el tren. Los últimos años han sido tiempos de una marcada digitalización y este campo es cada día más técnico. Decisiones que antiguamente se podían basar en la intuición ahora se pueden, como mínimo, apoyar en datos. La ciencia de datos también posibilita que actividades comerciales, publicidad y campañas sean analizables, medibles y comparables. Tal es el cambio que muchas empresas ya han incorporado perfiles más técnicos en los entornos de marketing y han añadido firmemente la tecnología de la ciencia de datos como responsabilidad de los llamados CMO (*chief marketing officer*).

Algunos ámbitos de aplicación son:

- **Optimización de presupuestos**. Una de las tareas que más tiempo consume a los responsables de marketing de las marcas suele ser la gestión del presupuesto. Canales, actividades, campañas, descuentos, medios... cada uno con un retorno estimado de la inversión, un valor esperado o la intención de ganar cierta cuota de mercado. La incorporación del científico de datos puede ayudar a modelizar el retorno en función de las asignaciones presupuestarias, optimizando el uso de los preciados (y habitualmente escasos) recursos de los que se dispone, así como analizar las suposiciones y casos anteriores para establecer criterios más próximos en la realidad.

- **Segmentación**. Habitualmente, las campañas de marketing se dirigían al público general, y esperan captar la atención del público objetivo a partir de la exposición global. No obstante, aprovechar el estudio de la información puede ayudar a segmentar los demográficos y los espacios de los que se obtiene el retorno más óptimo. Es posible, por ejemplo, comparar entre localizaciones de un anuncio, ver si una campaña funciona mejor en una zona de la ciudad que en otra, decidir emitir un anuncio en una franja horaria concreta de un canal de televisión determinado o, incluso, hacer publicidad únicamente en canales digitales si se detecta que la audiencia a la que se dirige la acción raramente ve la televisión. Esta parte requiere

cruzar datos internos (de resultados) y externos (de cuota de pantalla y audiencias, paso de personas o clics, por ejemplo).

- **Retención**. La ciencia de datos abre la puerta a conocer (o perfilar) a los clientes de una compañía por su comportamiento y no solo por sus atributos. De este modo, es posible intentar identificar aquellos clientes que están a punto de dejar la compañía (esta es una estrategia que se utiliza mucho, por ejemplo, en las empresas de telecomunicaciones). Así, los gastos de marketing para ofrecer descuentos a los clientes que se quieren retener antes de que se marchen son mucho más eficientes en su acción y, a la vez, también se puede conseguir que, con menos contactos o llamadas, los clientes fieles contraten mejores servicios y, por lo tanto, sean más rentables para la compañía. Hablando de contactos, la ciencia de datos también se utiliza para predecir en qué momento y por qué medio es mejor contactar con un cliente. Puede haber, por ejemplo, quien vea con mejores ojos un correo electrónico que una llamada, que puede percibir como más agresiva, mientras otros prefieren el contacto directo. La ciencia de datos puede ayudar a eliminar la prueba y error que caracterizaba a este tipo de contactos.

- **Priorizar**. En general, para cualquiera de los casos anteriores, la ciencia de datos posibilita que los responsables de marketing puedan comparar y construir modelos temporales para ver si, por ejemplo, es mejor contactar con los clientes que están a punto de marcharse en verano y con los que potencialmente pueden mejorar sus servicios en otoño. La construcción de este tipo de modelos puede ayudar a priorizar, a organizar las actividades y, en definitiva, a optimizar el tiempo de todo tipo de recursos.

- **Redes sociales**. El marketing actual empieza a tener más presencia en las redes sociales que en los entornos considerados tradicionales (televisión y prensa escrita). No se trata solo de centrar las actividades en los entornos digitales, sino de entender que cada tuit, cada mensaje de Facebook y cada foto de Instagram es una mina de datos. Utilizando análisis textuales, de sentimiento o, incluso, análisis de redes sociales, es posible no solo saber de primera mano la opinión de los clientes, sino anticipar futuras necesidades.

3.2. Finanzas

El sector de las finanzas es un sector que tradicionalmente se apoya mucho en los datos, pero que nunca había accedido a una capacidad de procesamiento como el actual. De hecho, gracias a las grandes posibilidades de innovación y uso de la tecnología del sector, ha aparecido incluso una rama paralela a las finanzas tradicionales: las llamadas *fintech*, que no son más que empresas financieras absolutamente basadas en la tecnología y la ciencia de datos y que aspiran (y, en muchos campos, lo consiguen) a competir con los métodos tra-

dicionales que siguen utilizando las grandes entidades bancarias. Además, es importante destacar la elevada cantidad de datos de los que disponen las entidades financieras, que los ha convertido en el sector que dispone de departamentos más avanzados dedicados exclusivamente a la ciencia de datos.

¿En qué aplican estas compañías la ciencia de datos?

- **Riesgos**. La gestión de riesgos es, probablemente, el área crítica de cualquier institución financiera. ¿Conviene a la entidad dejar una cantidad determinada de dinero a un individuo o empresa? ¿Tiene sentido esta inversión? ¿Qué precio es razonable para un seguro con ciertas características? Estas son tareas ideales para los procesos de aprendizaje automático, que permiten identificar, priorizar y monitorizar los riesgos asociados a las operaciones habituales de un banco. Es un campo con un potencial elevadísimo que apenas se está empezando a trabajar y que requiere no tan solo la integración de estos procesos en el núcleo de la compañía, sino también una mejora de las capacidades de ciencia de datos de la mayor parte de los trabajadores.

- **Gestión de datos**. Se ha destacado que el sector financiero es, probablemente, el que más datos genera y guarda sobre sus usuarios. La gestión de estos datos es otro ámbito en el que la ciencia de datos puede ayudar, primero, para seleccionarlos y, segundo, para extraer información útil. Por ejemplo, ¿cómo afectan ciertas noticias en el comportamiento de los usuarios? ¿Es posible prever si contratarán más cierto producto cuando hay un cambio político, por ejemplo?

- **Predicción**. Siguiendo con el punto anterior, la capacidad de extraer información útil de los datos históricos abre la puerta a utilizarlos para prever acontecimientos futuros. Así, es posible intentar prever, por ejemplo, los movimientos de la bolsa (o, como mínimo, decidir en qué momento es más razonable intervenir).

- **Detección de fraude**. Quizás es el ejemplo más habitual. La detección de fraude no es solo una obligación en el ámbito de responsabilidad de cara a los usuarios, sino que en muchos casos se convierte también en una responsabilidad legal. ¿Es posible saber si una operación realizada con tarjeta de crédito es legítima o un fraude? La respuesta es sí, y que cada vez lo es más. Así, es posible detectar tanto a usuarios que intentan manipular operaciones como robos de tarjeta y, por lo tanto, prevenir estas operaciones. Para los clientes es ya cada vez más habitual observar bloqueos preventivos de tarjetas cuando aparece una operación en el extranjero o en un patrón poco común (por ejemplo, una gran operación de un cliente que suele hacer muchas pequeñas transacciones).

- **Análisis de clientes**. Las entidades financieras también pueden mezclar técnicas de marketing con productos bancarios para ofrecer los productos de una forma más dirigida.

- **Inversión algorítmica**: aunque esta no sea una tendencia para el ciudadano medio, es quizás la tendencia que más impacto ha tenido en la economía mundial (y de manera relativamente silenciosa). Ciertos algoritmos (de inteligencia artificial) toman decisiones de compra y venta en las bolsas internacionales de forma constante. Sí, la inmensa mayoría de transacciones bursátiles provienen de un ordenador. Los *traders* y *brokers* han dejado su lugar a procesadores que gestionan no solo los precios de las acciones, sino que también incorporan información externa a la hora de tomar decisiones.

3.3. Salud

Junto con el marketing y las finanzas, el sector de la salud es probablemente el que más uso de la ciencia de datos está haciendo hoy en día, con la proyección de seguir avanzando mucho más en este campo considerando las posibilidades que ofrece. El objetivo final (obviando por un momento los intereses farmacéuticos y empresariales) está claro: conocer mejor el cuerpo humano y ser capaces de salvar más vidas.

Algunos ejemplos que están cambiando la manera de entender y enfocar la salud son:

- ***Wearables***. La irrupción de los dispositivos portables, como relojes o ropa que incorpora sensores, permite recoger terabytes de datos sobre el funcionamiento diario del cuerpo humano. Así, no es difícil imaginar su potencial para detectar ritmos cardíacos anómalos, para controlar riesgos cardíacos o respiratorios, diabetes y, en general, para prever posibles ataques o paradas.

- **Mejora de diagnósticos**. A pesar de la gran cantidad de datos y la experiencia acumulada, los diagnósticos erróneos son todavía frecuentes (se calculan en torno al 5%, que puede parecer poco pero hay que ver los millones de personas que representa) y la detección precoz es menor que la deseada. El uso de la ciencia de datos para detectar patrones en los datos de pacientes, para identificar posibles indicadores de ciertas enfermedades y para procesar resultados de análisis, radiografías y otros, pueden mejorar los diagnósticos y proporcionar información adicional tanto a los médicos como a sus pacientes.

- **Tratamientos personalizados**. Es muy sabido que no todos los pacientes responden igual a los mismos tratamientos o principios activos. El uso de la tecnología puede posibilitar la agrupación de pacientes por perfiles y

generar la llamada *medicina de precisión*, que va más allá del amplio espectro y se centra en la efectividad.

- **Investigación farmacéutica**. Las curas, las vacunas o los tratamientos de enfermedades que todavía escapan de las posibilidades de la medicina (como el cáncer, el ebola o la enfermedad de Alzheimer) pueden recibir ayuda del potencial de la ciencia de datos, tanto para analizar millones de casos como para proporcionar información adicional sobre tratamientos experimentales o adelantos en la cura. En casos de enfermedades infecciosas, además, puede proporcionar oportunidades de facilitar el control y la neutralización de la expansión.

- **Control de prescripciones**. A pesar de que parezca extraño, todavía hay miles de casos de prescripciones erróneas, que a veces acaban en desenlaces fatales. La prescripción también puede recibir recomendaciones y alertas; si al 99% de los pacientes se les receta lo mismo, cuando se intente prescribir un medicamento diferente puede saltar una alerta para asegurar que es una opción correcta. Sucede lo mismo si un medicamento prescrito tiene una interacción con otro fármaco que ya toma el paciente (y que puede estar prescrito por otro médico u hospital) o si hay posibilidad de alergia.

- **Reducción de costes**. De nuevo, como en el resto de sectores, el uso adecuado de la información puede permitir la reducción de costes. ¿Cuántos días necesita alguien estar ingresado? ¿Cuál es la dosis exacta de medicina requerida? ¿Cuál es el tratamiento más efectivo (o con menos efectos secundarios que pueden significar costes adicionales)? Son solo algunas de las preguntas que se hace la ciencia de datos aplicada a la salud.

3.4. Educación

Otro campo en el que la ciencia de datos se va abriendo paso es el de la educación, bautizado en este caso con el nombre de *learning analytics*. El paso cada vez más común en la educación en línea ha permitido, por un lado, la aparición de una gran cantidad de datos que pueden ayudar a mejorar la experiencia del estudiante y, por otro, ha iniciado la necesidad de disponer de ciertas capacidades de automatización para ofrecer un mejor servicio a las grandes cantidades de alumnos que optan por estas modalidades.

La mayor parte de estas iniciativas van dirigidas a conocer el comportamiento de los alumnos para ofrecerles una experiencia adaptada a su perfil. ¿Qué pueden aportar?

- **Predicción de rendimiento**. Uno de los beneficios que puede aportar la ciencia de datos al aprendizaje es información sobre el rendimiento del estudiante; no solo en el momento actual, sino en el futuro durante el curso. Las posibilidades de estos datos no son para aprobar o suspender

antes de tiempo, está claro. La idea es que, ante la previsión de un futuro suspenso de un alumno, es posible proveer apoyo adicional a tiempo para evitarlo. Del mismo modo, es posible ver si ciertas actividades o materiales contribuyen positivamente a mejorar el rendimiento académico y, también, el aprendizaje.

- **Experiencia personalizada**. Mediante *learning analytics* se pueden proporcionar y generar experiencias de aprendizaje personalizadas a cada perfil y/o alumno. Una persona puede requerir más material para comprender un módulo en el que ha invertido mucho más tiempo del que es normal para el resto o de su propia media. Alguien preferirá materiales para leer, otros en vídeo. La idea es que no hay dos personas que aprendan exactamente del mismo modo, así que un sistema apropiado puede asegurar que la experiencia sea la más óptima posible para cada uno de ellos.

- **Motivación**. Una consecuencia de la aplicación de la predicción de rendimiento y la experiencia personalizada es un aumento de la motivación (o, como mínimo, un abandono menos elevado). Si alguien no está disfrutando de una buena experiencia o intuye que no podrá superar el curso, es más fácil que decida dejarlo. La aplicación de las estrategias anteriores puede disminuir el abandono.

- **Iteración**. No se trata solo de que el uso de la ciencia de datos en la educación pueda beneficiar a los estudiantes actuales, sino que también puede hacerlo con los futuros. Se pueden detectar materiales o unidades problemáticas (que podrán ser revisados el curso siguiente), metodologías que funcionan mejor que otras, maneras y frecuencias de contacto... en definitiva, una mejora en eficiencia y utilidad curso tras curso.

- **Reducción del coste**. Como en todas las perspectivas de negocio, aquí también se habla del coste. Pero en educación este coste no es tan solo económico, sino que también hay un importante componente temporal. ¿Todos los materiales son utilizados por los alumnos? ¿Hay algún módulo que pasan por alto? ¿Es posible mejorarlo o eliminarlo para asignar aquellos recursos a otro elemento con un retorno de aprendizaje más grande? Son preguntas que hasta hace muy poco eran difíciles de responder y que, gracias a la ciencia de datos, cada día estamos más cerca de mejorar.

3.5. IoT

Aunque no entraremos en detalle en las implicaciones del IoT (*Internet of things*), la idea básica es que está relacionado con los datos que proporcionan los sensores, que a día de hoy son baratos y están incorporados a prácticamen-

te cualquier dispositivo. Su potencial radica en la capacidad que tienen para obtener datos de su entorno, que a la misma vez pueden ser analizados y cruzados con otros datos para detectar patrones.

Los ejemplos relacionados con IoT son inacabables, pero a continuación se citan algunos:

- **Análisis de vídeo**. A pesar de que pueda tener una parte controvertida, la idea es que es posible monitorizar en vídeo (de una cámara de vigilancia, por ejemplo) para detectar anomalías y generar avisos de seguridad o identificar personas. Del mismo modo se puede controlar el tráfico o leer las emociones de las personas que aparecen en el vídeo.

- **Móviles**. Estos dispositivos inteligentes que prácticamente todo el mundo lleva siempre en el bolsillo son una fuente de información inagotable. Los datos de geolocalización, por ejemplo, pueden prever aglomeraciones, contar capacidad o identificar los patrones de movimiento y circulación de las personas en una tienda o un centro comercial.

- **Uso de productos**. Parecía ciencia ficción, pero ya es habitual tener neveras, lavadoras y cafeteras conectadas a la red. Estos dispositivos pueden proporcionar información valiosa sobre hábitos de uso (cuántos cafés y de qué tipo se hace un usuario cada día), sobre consumo (¿utiliza el programa de la lavadora adecuado?) o niveles de inventario (no hay leche en la nevera). El potencial de esta información cruzada con otros datos, como las de marketing, es ilimitado.

- **Datos de redes sociales**. ¿Y si pensamos en Twitter o Facebook como una gran fuente de información que puede ayudar, por ejemplo, en caso de un desastre natural? Si un grupo de usuarios proporciona información sobre un accidente, un incendio u otro acontecimiento en las redes sociales, es posible cruzarlo con otros datos provenientes de sensores próximos para tener un análisis completo de la situación y saber, por ejemplo, cuántos efectivos hay que movilizar desde el primer momento.

3.6. Seguridad

Otro ámbito que es cada día más importante es el de la seguridad pública, amenazada por, por ejemplo, ataques terroristas. La ciencia de datos también se utiliza para entender las formas que tienen estos grupos de comunicarse, para identificar ataques potenciales o grupos radicalizados y para detenerlos. Analizando ataques anteriores y comunicaciones pasadas es posible detectar patrones y ejecutar acciones preventivas.

Por otro lado, como en todas las perspectivas, pero en este caso especialmente, se deben tener en cuenta las implicaciones legales y la protección de datos.

Por ejemplo, es posible imaginar la predicción de crímenes. Un algoritmo puede aprender de múltiples perfiles y determinar que un individuo tiene el potencial de cometer un robo. Esta información, no obstante, no puede llevar a la detención de una persona por un crimen «hipotético» al estilo de *The Minority Report*. No obstante, hay que pensar en otras opciones, como la de determinar zonas, casas y horas en que es más probable que haya un robo: esto permitiría determinar los recorridos de las patrullas o el número de efectivos requeridos en cada caso y cada momento.

The Minority Report

Esta historia corta de Philip K. Dick (de 1956) se basaba en la premisa de poder detectar crímenes antes de que se produjeran. La adaptación al cine es *Minority Report* (2002), dirigida por Steven Spielberg y protagonizada por Tom Cruise. ¿Superará la realidad a la ficción?

3.7. Otros

Podríamos seguir enumerando ámbitos y aplicaciones maravillosas de la ciencia de datos pero la lista no acabaría nunca. Aun así, es conveniente citar otros ámbitos específicos en los que la ciencia de datos ya está cambiando las vidas de las personas en su día a día:

- **Búsquedas y buscadores en Internet**. El hecho de que se utilice Google[3] (o Bing, Ask, Duckduckgo, etc.) como si fuera la cosa más natural del mundo, no es excusa para entender que funcionan utilizando ciencia de datos sobre unas cantidades exageradas de información. Sin ciencia de datos valdría más tener una agenda telefónica (de direcciones web) muy amplia.

[(3)]Google procesa decenas de petabytes de datos cada día.

- **Anuncios**. ¿Se ha buscado información sobre el Louvre recientemente? ¿A quién le extraña recibir una oferta de un viaje a París, un vuelo o un hotel? Los anuncios que se muestran en una página o en el correo son diferentes para cada usuario y aparecen en función de su historial y de las tendencias de usuarios anteriores; los anuncios generales en línea son una especie en extinción.

- **Sistemas de recomendación**. ¿Has comprado un lápiz? Te recomendamos también que compres esta goma y este sacapuntas. ¿Que has visto una serie de acción? Tenemos también todas estas que quizás que te gusten. La recomendación ya no es un tema solo de compra (compra X porque has comprado Y antes), sino que se ha convertido en un tema relacionado con la experiencia de usuario. Si un proveedor tiene miles de películas, para el usuario es muy beneficioso ver un filtro que muestre las películas que son más relevantes según sus gustos. En caso contrario, se podría aburrir.

- **Reconocimiento de imagen**. En este ámbito, las posibilidades también son infinitas: reconocimiento de personas en fotos subidas a las redes sociales, obtención de información sobre un cuadro al hacer una foto, información sobre la localización retratando la calle donde el usuario se ha perdido...

- **Reconocimiento de voz**. A pesar de que los usuarios informáticos se hayan acostumbrado al uso del texto y del teclado para interactuar con las

máquinas, el medio que es más natural para el ser humano es el lenguaje oral. El reconocimiento del lenguaje natural está experimentando un crecimiento sin precedentes, a pesar de que todavía queda mucho por delante. La aparición y expansión de los asistentes domésticos es todo un síntoma de ello.

- **Videojuegos**. Aquí podríamos aplicarlo en muchos aspectos. Por ejemplo, es posible modificar la experiencia del jugador en función de sus hábitos. En juegos competitivos en línea, se puede determinar si el jugador es más social y habitualmente juega con amigos o si, en caso contrario, es un jugador solitario, y ofrecer servicios adecuados y personalizados. En otros juegos de largo recorrido (*World of Warcraft*, *League of Legends*) se utiliza el análisis de datos para ver qué hacen los jugadores dentro de la partida y determinar los cambios o adaptaciones necesarios para mantener el balance adecuado o prevenir que los subscriptores dejen el juego. Y no olvidemos la parte competitiva, los *deportes*. Los equipos profesionales tienen personas dedicadas únicamente a analizar las métricas de rendimiento para entender qué funciona y qué no.

- **Comparación de precios**. Un usuario tiene un producto en su cesta, lo compra y, minutos más tarde, baja de precio drásticamente. Esta situación, impensable hasta ahora en el comercio tradicional, es habitual entre los minoristas en línea. Detrás hay algoritmos que comparan el precio con los precios de los competidores y que aprenden del comportamiento de los usuarios para ofrecer los precios más atractivos en el momento adecuado. Ya hay, incluso, tiendas físicas que incorporan marcadores de precios variables con estas estrategias.

- **Rutas aéreas**. La proliferación de vuelos, compañías aéreas y la saturación de las rutas hace que las líneas aéreas confíen cada vez más en la ciencia de datos para, por ejemplo, decidir cuántos aviones tienen que comprar, qué rutas son más eficientes y prever retrasos.

- **Logística**. Uno de los sectores en el que cada céntimo cuenta más es el de la logística, que busca ganar cada segundo posible y mejorar la eficiencia, y en el que la competencia es dura. Utilizando la ciencia de datos se planifican rutas de reparto, se pueden prever horarios óptimos para cada cliente, elegir el medio de transporte e, incluso, maneras de agrupar los paquetes en el almacén.

4. Conceptos de ciencia de datos

Los apartados anteriores han mostrado que hay una gran cantidad de conceptos y nombres (y que cambian y aparecen unos nuevos cada día) en torno a la ciencia de datos, así que en este apartado intentaremos aportar algunas definiciones y un mapa conceptual de términos relacionados con la disciplina. Primero, una serie de términos fundamentales, básicos en la ciencia de datos, seguidos de los campos de interés principales y un breve glosario sobre conceptos estadísticos (que, como se ha dicho, forma la base sobre la que se sustenta la ciencia de datos). Más adelante desarrollaremos algunas notas sobre las partes del proceso de análisis, detalles de las técnicas de aprendizaje automático y un breve resumen de nombres relacionados con los programas y las arquitecturas que se utilizan más frecuentemente.

4.1. Términos fundamentales

A continuación veremos los términos fundamentales de ciencia de datos:

- **Algoritmo**. Un algoritmo es un conjunto de instrucciones que se dan a un ordenador para que las ejecute. Puede corresponder a los pasos para resolver una fórmula matemática, por ejemplo.

- **Base de datos**. Es el sistema de almacenamiento de datos, así de simple. Tradicionalmente relacionales o SQL (es decir, donde la información se guardaba de manera muy concreta, organizada y formalizada y donde primaba la relación conocida entre los datos). También existen (y cada vez se utilizan más) las bases de datos NoSQL. Los programas de ciencia de datos suelen interactuar con las bases de datos.

- ***Big data***. Este es un término que ha perdido un poco el sentido actualmente porque se utiliza con demasiada frecuencia, pero la idea básica es pensar en una cantidad de datos bastante grande como para que no sea trivial procesarla. El *big data* se caracteriza por las cuatro V problemáticas: alto volumen de datos, variedad de tipo, necesidad de comprobar la veracidad y todo a alta velocidad (tanto de llegada como de procesamiento).

- ***Data warehouse***. Un *data warehouse* es un sistema pensado para hacer análisis rápidos de datos de diferentes fuentes en un entorno empresarial. Básicamente es una forma de hacer más «fácil» el análisis a los analistas de datos y que, así, no tengan que conocer la tecnología que soporta la aplicación.

- **Entrenamiento y test**. En el proceso de construcción de un modelo de aprendizaje automático, normalmente se separan los datos disponibles en

una parte de entrenamiento (para construirlo) y una de prueba o test (para comprobar cómo funciona y que no haya sobreentrenamiento).

- ***Front/Back end***. Así se denomina a la parte visible (*front*) de un programa (es decir, lo que ve el usuario o cliente) y la parte como está programado (lo que hay detrás, *back*).

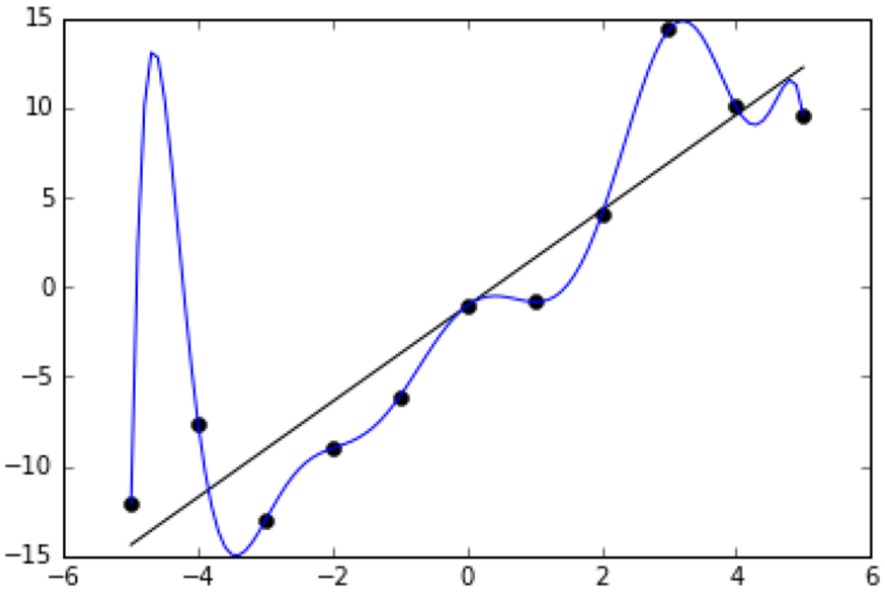
- **Lógica difusa**. Es una abstracción de la lógica booleana (la de los 0 y 1) que asigna valores intermedios y que, por lo tanto, permite que una afirmación no tan solo sea cierta o falsa, sino que pueda ser un poco cierta o prácticamente falsa, por ejemplo.

- ***Machine learning* o aprendizaje automático**. Se denomina algoritmo de aprendizaje automático aquel que es capaz de obtener información a partir de un conjunto de datos y hacer predicciones en función de esta información. Hay múltiples técnicas de aprendizaje automático.

- **Regresión**. La regresión es un problema de aprendizaje automático supervisado que se centra en explicar cómo cambia una variable numérica en función del resto.

- **Sobreentrenamiento**. En inglés *overfitting*, es lo que pasa cuando se proporciona información excesiva al modelo, que memoriza y no aprende. Memorizar implica que el modelo obtendrá resultados excelentes con los datos de entrenamiento, pero cuando se utilicen los datos de prueba (o una predicción real) se obtendrán resultados indeseados (figura 9). Lo contrario es el *underfitting*, que pasa cuando el modelo tiene muy pocos datos.

Figura 9. A pesar de que la línea polinómica se ajuste perfectamente a los puntos, la línea negra (lineal) es más generalizable



Fuente: https://en.wikipedia.org/wiki/overfitting

4.2. Campos de interés

Los campos de interés principales de la ciencia de datos son:

- **Análisis de datos**. El análisis de datos es como una versión reducida de la ciencia de datos, centrada en dar respuesta a ciertas preguntas predeterminadas, utilizando la estadística más básica y mucha menos programación.

- ***Business intelligence***. El BI se podría resumir en cómo utilizar software diverso para generar informes y encontrar información importante para el negocio entre los datos. Es esencialmente descriptivo y se centra en las métricas de negocio.

- ***Data engineering***. La ingeniería de datos es la disciplina que se encarga de facilitar el trabajo del científico de datos asegurando que los datos de trabajo están en el formato más adecuado. En equipos pequeños, el científico de datos también se encarga de la ingeniería de datos.

- ***Data journalism***. El periodismo de datos es la versión narrativa de la ciencia de datos, es decir, se encarga de explicar historias de relevancia informativa mediante datos y basadas en datos. La inmensa mayoría de las veces se complementa con una visualización de datos impactante y clara.

- ***Data science***. La ciencia de datos es la disciplina que utiliza datos y estadística avanzada para hacer predicciones, entender la información y generar conocimiento útil.

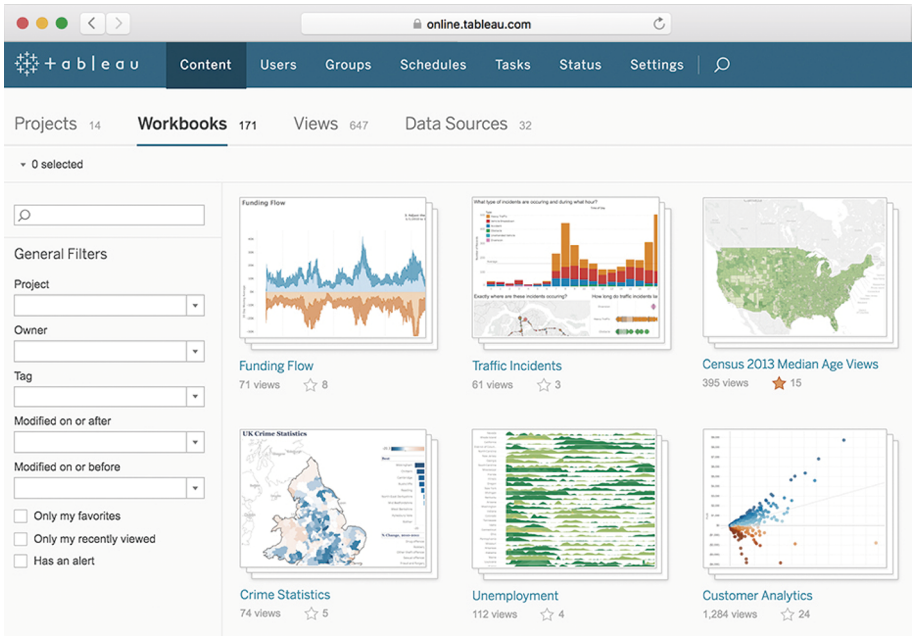
Ved también

Para consultar más información sobre el *data science*, consultad el apartado 1.

- **Inteligencia artificial**. La IA es la disciplina que se centra en la investigación y el desarrollo de máquinas que tienen conciencia de su entorno, que son capaces de, por ejemplo, resolver una tarea concreta. Coches autónomos, robots médicos o los videojuegos son algunos ejemplos.

- **Visualización de datos**. La visualización de datos se ha erigido como una disciplina con mucha proyección, considerando la complejidad de comunicar de manera clara la información obtenida de grandes volúmenes de datos. Utiliza infográficos, gráficos tradicionales o software específico (Tableau, Qlik,...). (Ver la figura 10.)

Figura 10. Tableau permite visualizaciones muy elaboradas



Fuente: https://www.tableau.com/

4.3. Conceptos estadísticos

Los conceptos estadísticos principales de la ciencia de datos son:

- **Correlación**. La correlación es una medida que indica lo relacionados que están dos conjuntos de valores. Puede ser positiva (si uno aumenta, el otro también), negativa (si uno aumenta, el otro disminuye) o nula (cuando no hay ninguna tendencia).

- **Desviación estándar**. Si la media muestra el valor esperado, la desviación indica lo dispersos que son los valores. Si lo elevamos al cuadrado obtenemos la varianza (σ^2).

- **Error residual**. La diferencia entre el valor real y el valor calculado basado en el modelo obtenido es el error residual. Si un modelo calcula que una persona de 170 cm tendría que pesar 70 kg pero en realidad pesa 65, el error es de 5.

- **Estadísticamente significante**. Un resultado es estadísticamente significante si no se puede asegurar que es causado por un efecto aleatorio.

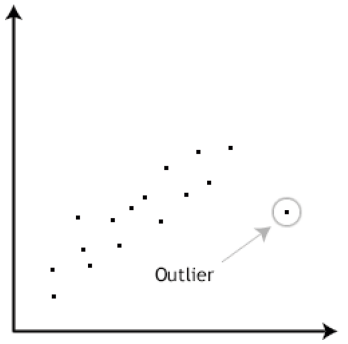
- **Mediana**. Si se ponen todos los datos ordenados, la mediana es el dato que queda en medio de todas ellas. Combinada con la media sirve para ver si hay datos anormalmente grandes o pequeños.

- **Media**. La media muestra el valor típico que se espera encontrar en un conjunto de datos. Se debe tener cuidado porque la media, por sí sola, no sirve de mucho.

- **Muestra**. Es el conjunto de datos a los que se tiene acceso y que se pretende utilizar para extraer conclusiones sobre la población (que sería «el mundo real»).

- **Normalizar**. El proceso de normalización es el que se lleva a cabo para equiparar todos los datos en un mismo rango. Muchos algoritmos de aprendizaje automático son sensibles al valor absoluto de los datos y, por lo tanto, suele ser necesario normalizar la entrada.

- ***Outlier* o valor atípico**. Un *outlier* es una observación, un dato, que está exageradamente lejos del resto y que puede ser debido a un error (un dato mal escrito) o a un punto excepcional (los sueldos de los futbolistas de élite). Deben considerarse y tratarse, si es necesario, en la fase inicial.



Representación gráfica de un outlier

4.4. Procesos

A continuación expondremos los procesos principales de la ciencia de datos:

- **ETL**. Un proceso ETL es el que sirve para extraer, transformar y cargar un conjunto de datos a un *data warehouse*.

- **Exploración de datos**. Es el proceso inicial, en el que el científico de datos intenta entender el contexto de los datos y se hace preguntas básicas para ponerse en situación y preparar un análisis completo posterior.

- **Minería de datos o *data mining***. Es un término general que se atribuye al proceso de extraer información de un conjunto de datos y utilizarlo. Así, incluye desde la limpieza hasta la aplicación de algoritmos.

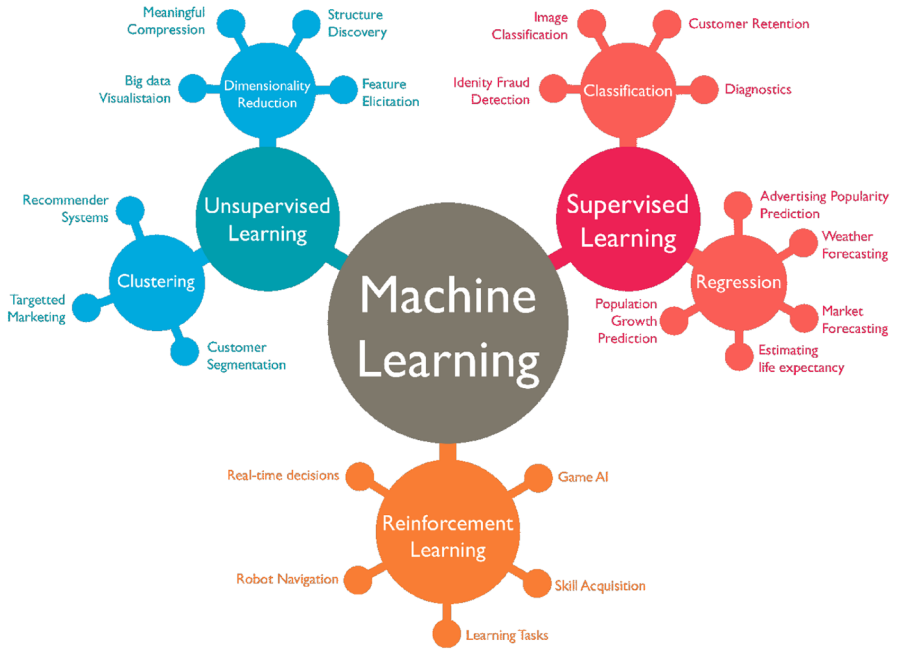
- ***Pipeline***. Un *pipeline* es un conjunto de funciones o algoritmos que se aplican en serie (uno tras otro, en orden). Así, el resultado del primero sirve de entrada al segundo.

- ***Scraping web***. Es la técnica utilizada para extraer datos del código fuente de una página web y normalmente requiere que el programador identifique las etiquetas necesarias para la extracción.

4.5. Técnicas de aprendizaje automático

En torno al aprendizaje automático hay muchas técnicas, orientadas a problemas diversos. En la figura 11 se representan algunos de ellos.

Figura 11. Clasificación de los algoritmos de aprendizaje automático



Fuente: https://www.slideshare.net/awahid/big-data-and-machine-learning-for-businesses

En general, se habla de las técnicas siguientes: técnicas supervisadas, técnicas no supervisadas y técnicas de refuerzo.

4.5.1. Técnicas supervisadas

Son aquellas en las que el científico de datos sabe cuál es la variable resultado. Así, se intentan construir modelos que expliquen la variable final para, así, entender mejor el problema.

> Por ejemplo, si se quiere saber qué características indican la aparición de la miopía se trata de un problema supervisado: se tienen los datos de los pacientes y si son o no miopes. En general se pueden dividir en algoritmos de regresión y de clasificación.

Técnicas de regresión

Son aquellas técnicas en las cuales la variable resultado es una variable «real». La más habitual es la regresión lineal.

Regresión lineal

La regresión lineal es una técnica que busca modelar la relación entre una variable respuesta (o dependiente) y una serie de variables explicativas (o independientes).

Por ejemplo, si intentamos calcular la altura de una persona a partir de su peso, obtendremos una relación entre ambas que será una recta con una cierta pendiente.

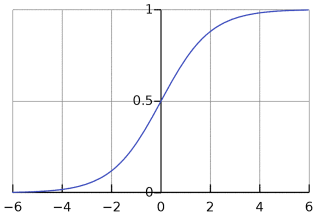
La regresión lineal simple, de una sola variable, tiene la forma $y = ax + b$ y los coeficientes se obtienen de minimizar la suma de residuos al cuadrado.

Técnicas de clasificación

Son los métodos supervisados en los que la variable respondida es una categoría (por ejemplo, hombre/mujer, sí/no). Hay muchos (y algunos muy complejos), pero los dos ejemplos más comunes y sencillos son la regresión logística y los árboles de decisión.

Regresión logística

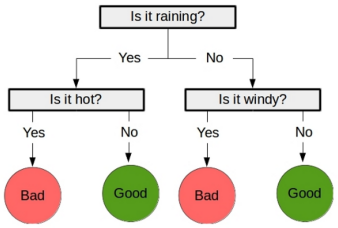
Los algoritmos de regresión logística intentan predecir la probabilidad de que pase un acontecimiento en función de la función logística. Así, el resultado es un número entre 0 y 1, que indica la probabilidad de que pase el acontecimiento concreto. Por ejemplo, un fraude con la tarjeta de crédito.



La función logística

Árboles de decisión

Los árboles de decisión usan una serie de caminos que se recorren en función de la respuesta a la pregunta de cada nodo y que, en las hojas, acaba con la clasificación.



Un árbol de decisión

Por ejemplo, si se quiere predecir la nota final de un alumno se puede hacer con un árbol de decisión: la primera pregunta podría ser si la nota del primer parcial es aprobado o suspenso. Los aprobados irían por una rama y los suspensos por otra.

4.5.2. Técnicas no supervisadas

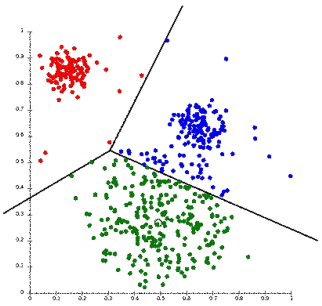
En este caso, la interpretación se deja en manos del ordenador. Si se tiene un conjunto de clientes y se quieren agrupar en perfiles, no se sabe cuántos perfiles diferentes existen. Así, el algoritmo de aprendizaje intentará buscar una clasificación razonable (que puede ser o no próxima a la realidad). Las técnicas no supervisadas principales son la clusterización y la reducción de la dimensionalidad.

Clusterización

Las técnicas de clusterización intentan agrupar los datos en grupos que son similares o, como mínimo, cercanos los unos a los otros. Dependen de la manera de medir la «distancia» entre puntos y la complejidad se incrementa a medida que lo hacen las dimensiones y el número de datos. Un ejemplo es el K-Means.

K-Means

El algoritmo de clusterización K-Means intenta dividir las observaciones en K clústeres, de tal manera que cada observación pertenezca al clúster más cercano a su media.



Ejemplo de K-Means

Reducción de la dimensionalidad

Los conjuntos de datos reales suelen tener grandes cantidades de variables; así, muchas veces es necesario (o adecuado) reducir el número. Normalmente se hace o bien escogiendo las variables más relevantes (selección de variables) o bien obteniendo una descomposición en variables principales (análisis de componentes principales).

Selección de variables

Este proceso consiste en medir la relevancia de cada una de las variables en la predicción del resultado final y seleccionar posteriormente las más adecuadas para obtener un modelo de rendimiento lo más óptimo posible.

Análisis de componentes principales

El análisis de componentes principales es algo más complejo de entender, pero básicamente busca transformar las variables en una descomposición ortogonal y sin correlación entre ellas. A pesar de su utilidad exploratoria, la explicación posterior del modelo se convierte en algo mucho más abstracto.

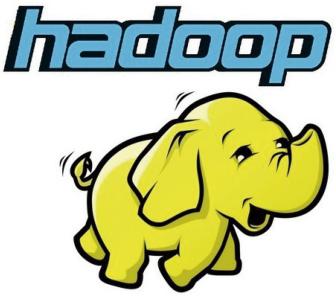
4.5.3. Técnicas de refuerzo

Esta área se encarga de intentar determinar qué acciones tiene que escoger un agente en un entorno dado para maximizar una recompensa o un premio. Son las técnicas que se utilizan, por ejemplo, para entrenar a los sistemas que juegan a ajedrez o a Go.

4.6. Software

En este subapartado presentamos una lista de nombres propios para que le suenen al estudiante para futuras materias:

1) Hadoop. Hadoop es un marco *open source* de procesamiento distribuido que se utiliza para trabajar con grandes cantidades de datos. Marca un antes y un después en la ciencia de datos avanzada. Permite utilizar procesamiento en paralelo entre varias máquinas (llamados clústeres).

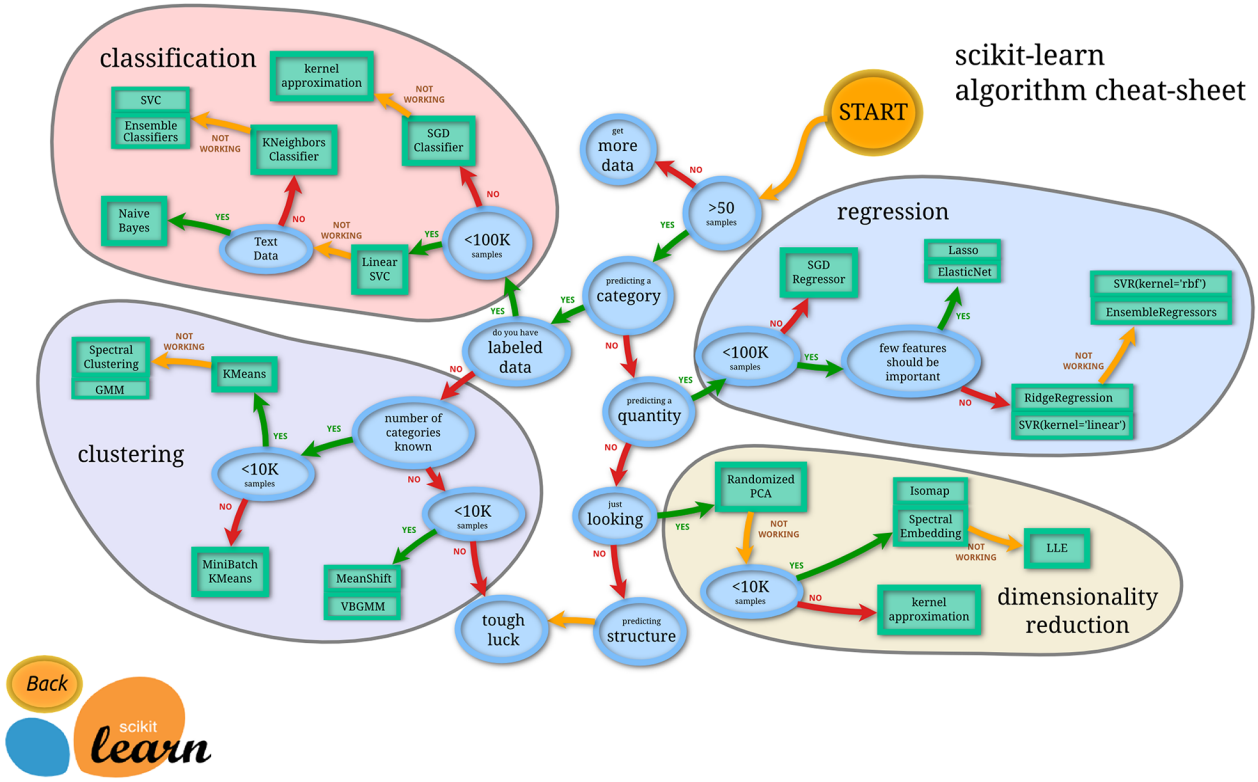


El elefante de Hadoop

2) Python. Python es un lenguaje de programación *open source*, utilizado en muchas aplicaciones, como la programación en general, la ciencia de datos y el aprendizaje automático. Se considera fácil de aprender, de alto nivel y tiene una comunidad muy activa. Algunas de las librerías más populares de ciencia

de datos son para Python, como Scikit-learn (aprendizaje automático, ver la figura 12), NTLK (procesamiento del lenguaje natural) o NetworkX (análisis de redes sociales).

Figura 12. Árbol de decisión para decidir qué algoritmo utilizar en Scikit-learn



Fuente: http://scikit-learn.org

3) R. R es tanto un lenguaje como un entorno *open source* orientado a la computación estadística. Es muy extensible (la comunidad es muy activa) y tiene implementadas la inmensa mayoría de técnicas existentes. R es gratuito y en el ámbito académico y de desarrollo se suele preferir a las alternativas de pago (SPSS, SAS), precisamente porque permite ver, controlar y modificar o adaptar los algoritmos que incorpora.

4) Spark. Apache Spark es otro marco *open source* de procesamiento distribuido que se utiliza para trabajar con grandes cantidades de datos, pero que da mucha más flexibilidad que Hadoop. Permite utilizar Java, Python, Scala y R y soporta SQL, reproducción en continuo y algoritmos de aprendizaje automático.

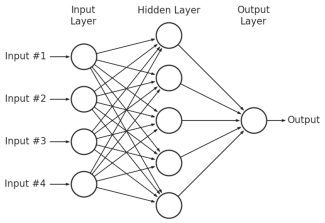
Apache Spark

4.7. Otros conceptos

A continuación explicaremos los conceptos de aprendizaje profundo y redes neuronales, *open data*, *open source* y sistemas de recomendación.

4.7.1. Aprendizaje profundo y redes neuronales

Las redes neuronales pertenecen al aprendizaje automático, pero sus características especiales hacen que destaquen. Se basan (libremente) en cómo funcionan las conexiones de las neuronas en el cerebro y básicamente son un conjunto de nodos organizados por capas que se entrenan para hacer predicciones. El llamado aprendizaje profundo (*deep learning*) no es más que la extensión en redes neuronales muy grandes, como las que se utilizan para identificar caras o imágenes.



Una red neuronal

4.7.2. *Open data*

Los datos abiertos son aquellos que son libres y cualquier persona los puede extraer y utilizar como quiera, sin derechos de autor, patentes o mecanismos de control. Algunos ayuntamientos, como los de Barcelona o Madrid, proporcionan datos abiertos que permiten que cualquier ciudadano consulte datos de transporte o de calidad del aire, por ejemplo.

4.7.3. *Open source*

Las herramientas de código abierto son aquellas que permiten acceder y editar su código fuente y que, por lo tanto, los usuarios pueden modificar. No debe confundirse con gratuito, a pesar de que la mayoría a veces lo sean. De hecho, muchas herramientas de código abierto presentan, por ejemplo, dificultades en la configuración (como Hadoop) y ciertas empresas ofrecen paquetes, de pago, configurados y a los que dan soporte técnico.

4.7.4. Sistemas de recomendación

Son sistemas de aprendizaje automático que se sitúan entre la regresión y la clasificación y que utilizan la información para recomendar elementos que pueden ser de interés para el usuario. Por ejemplo, productos relacionados con compras anteriores o películas basadas en los gustos y las puntuaciones de usuarios que tienen un perfil similar.

Bibliografía

Cleveland, W. S. (2001). «*Data science: an action plan for expanding the technical areas of the field of statistics*». *International statistical review* (vol. 69, n.º 1, pág. 21-26).

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996). «From data mining to knowledge discovery in databases». *AI magazine* (vol. 17, n.º 3, pág. 37).

Naur, P. (1974). *Concise survey of computer methods*. Nueva York: Petrocelli Books.

Tukey, J. W. (1962). «The Future of Data Analysis». *Ann. Math. Statist.* (vol. 33, n.º 1, pág. 1–67). doi:10.1214/aoms/1177704711.

Tukey, J. W. (1977). *Exploratory data analysis* (vol. 2). Londres: Pearson.