Fundamentos de *data science*

Julià Minguillón

PID_00235534

UOC
Universitat Oberta de Catalunya

Ninguna parte de esta publicación, incluido el diseño general y la cubierta, puede ser copiada, reproducida, almacenada o transmitida de ninguna forma, ni por ningún medio, sea éste eléctrico, químico, mecánico, óptico, grabación, fotocopia, o cualquier otro, sin la previa autorización escrita de los titulares del copyright.

Índice

1. La sociedad de la información 5

2. Datos, información, conocimiento, ¿sabiduría? 10

3. ¿Qué es un dato? 12
3.1. Datos simples 12
3.2. Datos compuestos o estructurados 12
3.3. Datos semiestructurados o no estructurados 16

4. Ciclo de vida de los datos 17
4.1. Captura 17
4.2. Almacenamiento 21
4.3. Preprocesado 25
4.4. Análisis 27
4.5. Visualización 30
4.6. Publicación 31

Resumen 33

Bibliografía 35

1. La sociedad de la información

La rápida aceleración de la tecnología propia de la sociedad industrial ha dado paso a la llamada sociedad de la información, basada en el uso intensivo de las tecnologías de la información y la comunicación (TIC). Actualmente es más importante (tanto en valor económico como estratégicamente) la información que se genera, gestiona y distribuye que el hardware que soporta dichos procesos. La aparición de internet y, posteriormente, su popularización mediante la World Wide Web a partir de 1994, han permitido a un porcentaje importante de la población mundial cambiar su forma de comunicarse, estudiar, trabajar y relacionarse con otras personas. Según datos del 2014, el porcentaje de población conectada a internet es del 40%, cifra que llega hasta el 78% en el caso de los países desarrollados. No obstante, la sociedad de la información no es exclusiva del uso de internet. La progresiva digitalización de recursos y servicios del mundo real ha permitido su expansión, globalizando aspectos comunes como la educación, el trabajo o el consumo, entre otros, mediante diferentes dispositivos, incluyendo ordenadores y teléfonos móviles pero también tarjetas de crédito, por ejemplo.

Los usuarios de esta sociedad de la información producen una gran cantidad de datos cuando interaccionan con este mundo digital. Todas las acciones realizadas por un usuario son susceptibles de ser analizadas, a partir del rastro que ha generado dicha acción, con el doble objetivo de mejorar la experiencia del usuario y avanzar en el conocimiento que se tiene de él. Cada día millones de usuarios pagan con sus tarjetas de crédito, utilizan sus teléfonos móviles, se conectan a internet mediante sus ordenadores, estacionan sus vehículos en aparcamientos inteligentes, toman fotografías o vídeos, acceden a edificios mediante una tarjeta de identificación, utilizan la red de transporte público, etc. Todos los ámbitos de la vida humana están empapados de tecnología que transforma todas nuestras acciones en datos factibles de ser manipulados, almacenados, analizados y visualizados, con fines bien distintos. Por otra parte, muchos otros procesos también generan datos que son factibles de ser o bien capturados y procesados, mediante sensores o indicadores que permiten conocer su estado, o bien almacenados directamente una vez que han sido generados.

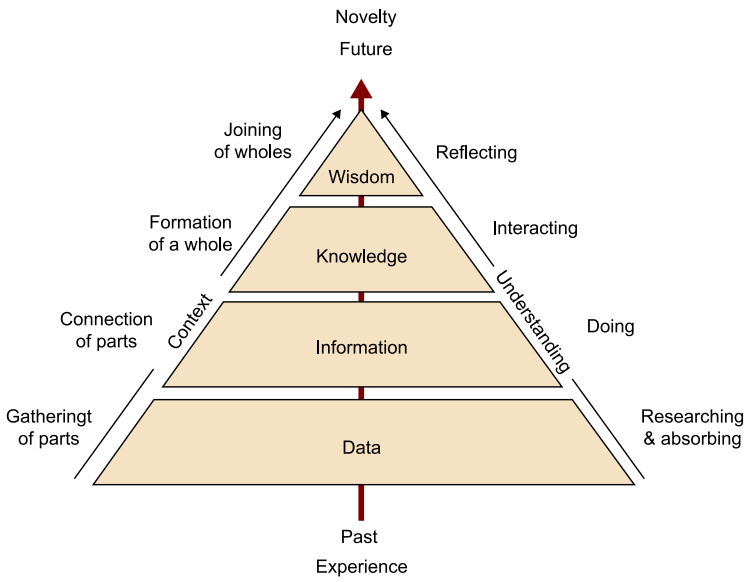
Así pues, esta sociedad de la información, o también del conocimiento, como otros prefieren llamarla, está basada en la generación e intercambio de datos entre usuarios, servicios y recursos, mediada por la tecnología digital. Aunque información y conocimiento son conceptos diferentes, ambos se basan en un elemento primordial: los datos.

Es lo que se conoce como la pirámide D-I-K-W (*data*, *information*, *knowledge* y *wisdom*), de modo que la información se define a partir de los datos disponibles, el conocimiento se extrae de dicha información y la sabiduría es entendida como la habilidad para aplicar dicho conocimiento en beneficio propio o común[1].

(1)Gu Jifa; Zhang Lingling (2014). «Data, DIKW, Big Data and *data science*». *Procedia Computer Science* (vol. 31, págs. 814-821). ISSN 1877-0509. http://0-dx.doi.org.cataleg.uoc.edu/10.1016/j.procs.2014.05.332

Este proceso combina métodos y técnicas de diferentes ámbitos, como la psicología, la estadística, la inteligencia artificial, la minería de datos y la visualización, entre otros. Se trata, pues, de un área **multidisciplinar** que incorpora perfiles muy diferentes, dada su complejidad y amplitud.

Pirámide D-I-K-W



Fuente: http://vishalkumarg325.blogspot.com.es/2013/03/dikw-pyramid-theory.html

Actualmente existe una gran confusión en torno al nombre más adecuado para este nuevo ámbito del conocimiento, dependiendo de la aproximación tomada. Es habitual hablar de *data science*, pero también es posible usar *data engineering* o, sencillamente, *data mining*. Una posible solución es plantearlo en términos de los objetivos que se tienen en función de cada perfil:

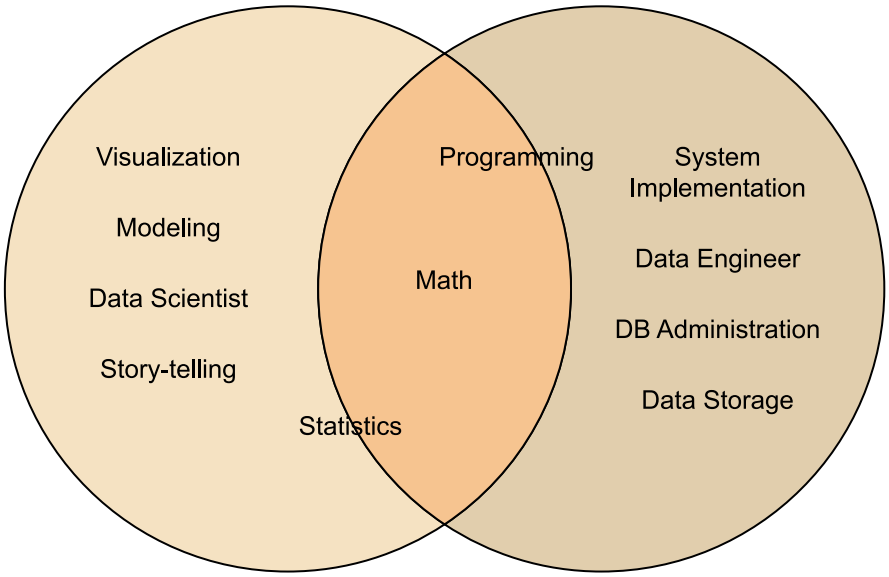
- Un *data scientist* es una persona que es capaz de plantearse las preguntas adecuadas a partir de un conjunto de datos relativo a un dominio, y establecer qué métodos y técnicas son los más adecuados para extraer el conocimiento necesario para responder a dichas preguntas, para posteriormente realizar dicha tarea. Este perfil está orientado principalmente a resolver el «¿qué?».

- Un *data engineer* es una persona capaz de preparar un conjunto de datos de forma que tenga la estructura e información adecuadas para su posterior análisis, así como de llevar a la práctica una solución o prueba de concepto y convertirla en una implementación que pueda usarse en un entorno productivo real. Este perfil está más orientado a resolver el «¿cómo?».

- Un *data miner* es una persona que, a partir de un conjunto de datos con un objetivo bien definido, es capaz de utilizar diferentes métodos y herra-

mientas para maximizar el resultado esperado. Es, por lo tanto, una especialización del primer perfil orientada específicamente al análisis de los datos.

Funciones del *data scientist* y del *data engineer*



Estos perfiles se complementan y, de hecho, comparten algunos objetivos y habilidades necesarias para llevarlos a cabo. La respuesta a una pregunta inicial relacionada con datos nunca es un proceso completamente lineal, sino que normalmente es necesario realizar diferentes iteraciones en dicho proceso, de manera que sea posible ir mejorando los diferentes elementos que lo constituyen, entre otros:

- La pregunta que se desea responder.
- Los datos de los que se dispone.
- Los métodos y técnicas más adecuados.
- La evaluación e interpretación de los resultados obtenidos.

El orden de estos elementos no siempre es el mismo. Lo más habitual es que a partir de un hecho observado en cierto entorno o contexto, los responsables de dicho entorno se planteen preguntas que necesiten respuestas, idealmente a partir de los datos disponibles o que puedan obtenerse. Esta es la aproximación tradicional de la estadística, intentar demostrar o refutar una hipótesis inicial, a partir de los datos disponibles.

Por ejemplo, ante un descenso de las ventas, una compañía podría preguntarse por las razones que han causado dicho descenso. Usando instrumentos convencionales como encuestas y entrevistas con sus consumidores habituales, se podrían determinar los factores que han causado el cambio de hábitos de consumo, utilizando para el análisis técnicas estadísticas, o establecer si se trata de un problema relacionado con la aparición de nueva competencia en el sector, por ejemplo.

En otras ocasiones, sin embargo, es posible que observando los datos disponibles se detecten patrones o irregularidades que permitan plantearse nuevas preguntas, nunca anteriormente imaginadas. El nuevo paradigma big data es un claro ejemplo de las oportunidades que ofrece analizar datos a gran escala. Es posible plantearse preguntas a las que nunca anteriormente se había podido responder, no solo por limitaciones relacionadas con la infraestructura tecnológica, sino también por la ausencia de los datos necesarios.

Por ejemplo, una cadena de supermercados podría analizar todos los patrones de compra de sus clientes durante un periodo de tiempo y detectar qué productos son comprados conjuntamente y con qué frecuencia, con el objetivo de proponer ofertas y/o establecer nuevas estrategias que promuevan el consumo de otros productos parecidos o en promoción.

No obstante...

...tal y como indica Jeff Leek, la palabra clave en *data science* es *science*, no *data*.

Finalmente, otro concepto muy de moda hoy día es el de **periodismo de datos**.

Básicamente, se trata de convertir datos en historias, utilizando medios digitales para narrarlas, especialmente visualizaciones interactivas, que permitan al lector participar en ellas y seleccionar aquellos fragmentos de la historia que le parezcan más interesantes.

Pero el periodista de datos también es capaz de ir a buscar los datos donde estos se encuentran (especialmente datos en abierto), utilizando las herramientas adecuadas, o de encontrar las conexiones entre diferentes fuentes de datos que permiten hallar la noticia enterrada en datos oficiales[2].

(2) M. Charski (2015). «Data Journalism: how to create compelling content from data». *EContent* (vol. 38(5), págs. 10-14).

Ejemplo

Por ejemplo, es posible cruzar los datos relativos al precio del barril de petróleo, que publican diariamente diferentes organismos oficiales, con los precios de los carburantes en la red de estaciones de servicio españolas, publicados por el Ministerio de Industria, Energía y Turismo. El resultado de este análisis permite probar la tan temida sensación de que cuando el precio del petróleo sube, los carburantes lo siguen a la par inmediatamente, mientras que cuando el precio del petróleo baja, los carburantes lo hacen más despacio, aprovechando las distribuidoras el margen disponible y repercutiéndolo en los usuarios finales.

Así pues, un buen experto en datos es aquella persona que sabe plantearse las preguntas adecuadas en un ámbito de conocimiento concreto, conoce los mecanismos para obtener los datos necesarios y prepararlos para su posterior análisis, mediante las técnicas estadísticas y de minería de datos más adecuadas, interpretar los resultados obtenidos y ponerlos en contexto, relacionándolos con las preguntas planteadas y el ámbito de conocimiento, contrastando con resultados anteriores, así como visualizar el conocimiento extraído durante este proceso para identificar los patrones, variables y otros elementos relevantes

que son clave en dicho proceso y que permiten avanzar en su comprensión. Se trata, por lo tanto, de combinar un arsenal de herramientas y técnicas científicas, que incluyen, entre otras, matemáticas, informática, de visualización, analíticas, estadísticas, de diseño experimental, de definición de problemas, de construcción de modelos y de validación. Todo ello con el objetivo de convertir los datos en conocimiento.

2. Datos, información, conocimiento, ¿sabiduría?

¿Qué es un dato, entonces? ¿Cuándo un dato se convierte en información o, mejor aún, en un conocimiento que se puede aprovechar? La pirámide D-I-K-W es el marco conceptual que permite entender y dar significado a dichos términos. Por ejemplo:

«42»

Es un dato, en este caso un número entero. Se puede dar por entendido que está escrito (representado) en base decimal, aunque esto no sería tan obvio en según qué contextos donde otros sistemas de numeración como el binario o el hexadecimal son más habituales. También se podría haber representado mediante la frase *cuarenta y dos*, aunque entonces solo los castellanohablantes reconocerían dicho dato, mientras que los habitantes de la costa este de África lo preferirían escrito como *arobaini* na *miwili*, seguramente. Por lo tanto, todo dato necesita una **representación** adecuada para su correcto uso.

Un **dato** es, en principio, una cantidad o cualidad que describe un atributo de una entidad, dentro de un rango de valores posibles. Es un valor «dado» al respecto de algo observado, de acuerdo con la raíz latina que da origen al término (*datum*).

De todos modos, es factible preguntarse ¿42 qué?, ¿a qué se refiere dicho dato? Es en este momento cuando un dato se convierte en información, cuando es capaz de responder a una pregunta concreta y adquiere significado, por ejemplo:

«¿Cuál es la temperatura del paciente?»

En este caso, 42 responde a dicha pregunta, pero ¿lo hace de forma adecuada? Es muy fácil darse cuenta de que no:

- Si se trata de 42 grados centígrados, el paciente tiene una fuerte fiebre.
- Si se trata de 42 grados Fahrenheit, el paciente es un cadáver frío.
- Si se trata de 42 grados Kelvin, el paciente es un cadáver frío flotando por el espacio exterior, seguramente.

Por lo tanto, un dato no es una información veraz si no va acompañado de una **precisión** y unas **unidades** que lo definan adecuadamente. Si para la misma pregunta anterior la respuesta fuera «cuarenta y algo», la vida del paciente podría estar en peligro por falta de precisión.

Este ejemplo puede parecer anecdótico o incluso trivial, pero el 23 de septiembre de 1999 la NASA perdió el contacto con la *Mars Climate Orbiter*, un satélite diseñado para estudiar la superficie, atmósfera y clima del planeta Marte. La razón fue que el satélite entró en órbita a una altitud insuficiente, lo que causó su destrucción. El motivo de dicho error fue que una parte del software utilizado para el cálculo de las trayectorias orbitales usaba el sistema métrico

decimal, mientras que otros módulos del software usaban el sistema basado en unidades inglesas (pies, pulgadas, etc.). Este error costó a la NASA (de hecho, a los ciudadanos estadounidenses) un total de 327,6 millones de dólares, el total de construir el satélite, lanzarlo al espacio y controlarlo hasta su puesta en órbita, sin tener en cuenta los problemas de imagen y el retraso en la misión original.

En este ejemplo, la información disponible puede aportar conocimiento sobre el estado del paciente. La experiencia acumulada en el tratamiento de casos similares dice que un paciente con una fiebre o temperatura corporal de 42 grados centígrados puede padecer lesiones cerebrales irreversibles. Este conocimiento ha sido extraído de múltiples ejemplos previos, donde la mayoría de los pacientes en la misma situación desarrollaban dicha patología. Es este conocimiento el que permite transformar una información basada en datos en una serie de acciones que permitan cambiar el estado del entorno, actuando sobre aquellos elementos que tienen incidencia sobre el proceso que ha sido «medido», en este caso la temperatura corporal del paciente. Así pues, este conocimiento podría expresarse mediante una regla, por ejemplo:

Ejemplo

«Si la temperatura del paciente alcanza los 42 grados centígrados, se pueden producir lesiones cerebrales irreversibles».

Finalmente, es la aplicación de este conocimiento (y su actualización constante) la que permite alcanzar la sabiduría tal y como esta se define en la pirámide D-I-K-W. No obstante, muchos autores prefieren no definir *sabiduría* (ni incluirla en este marco) dada su ambigüedad, y centrarse en la transformación de datos en información y la extracción de conocimiento de dicha información, aunque sí aplicarlo, por cierto. En este ejemplo, este conocimiento acerca de la temperatura del paciente y sus implicaciones debería ser aplicado con el objetivo de reducir su temperatura corporal de la manera más rápida posible.

3. ¿Qué es un dato?

Los datos pueden presentarse de muchas formas diferentes. En el ejemplo de «42», se trataba de un número decimal entero, pero los datos son de naturaleza muy diversa y se pueden clasificar de acuerdo con diferentes criterios, entre otros según su estructura; así, tenemos:

- **Simples**: datos atómicos indivisibles, con un significado propio, de acuerdo con la definición (un valor de un atributo).

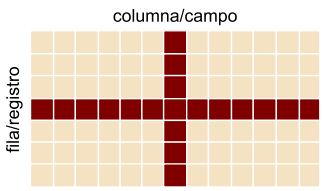
- **Compuestos** o **estructurados**: datos que son una combinación de otros datos simples y/o compuestos, de acuerdo con una estructura fija y conocida a priori.

- **Semiestructurados** o **no estructurados**: datos estructurados que siguen una estructura parcial o que puede cambiar según el contexto, o que no siguen ninguna estructura.

3.1. Datos simples

Ejemplos de datos simples son los números enteros, los reales (pero no los complejos, que serían un ejemplo de datos estructurados), los caracteres y, excepcionalmente también, las cadenas, aun cuando estas son en realidad secuencias de caracteres, por lo que son un tipo de dato compuesto (no obstante, su popularidad y utilización provocan que sea más sencillo pensar en las cadenas como de un tipo simple). Otro ejemplo parecido es la tupla [latitud, longitud], que identifica una posición en un mapa, ya que se trata de dos números reales pero que adquieren significado cuando se procesan conjuntamente.

3.2. Datos compuestos o estructurados

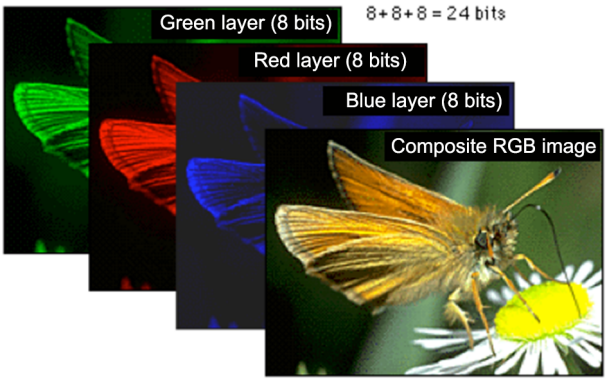


A su vez, los datos se agrupan en otras estructuras más complejas de acuerdo con su dimensionalidad. Un dato concreto o puntual no tiene dimensionalidad (o se podría decir que tiene cero dimensiones). Una secuencia de datos (por ejemplo, provenientes de un sensor de temperatura) constituye un *array* 1D o vector de una dimensión. Una estructura que describe, por ejemplo, los productos de un comercio (con su código, su descripción, su fotografía, su precio de coste y su precio de venta al público) es lo más habitual, hablándose en este caso de array 2D, tabla o matriz. Cada uno de los productos ocupa una **fila** de la tabla, mientras que cada uno de los atributos que lo describen ocupa una **columna**. Cada fila de la tabla también se suele llamar **registro**, mientras que cada atributo que describe un registro se suele llamar **campo**, siguiendo la nomenclatura propia de las bases de datos relacionales. El concepto de **di-**

mensionalidad puede ser extendido a más de dos dimensiones, obviamente. Por ejemplo, la misma tabla 2D puede verse como un «corte» de una tabla 3D donde en la tercera dimensión se dispone de la tabla 2D en diferentes momentos, por ej. el seguimiento de una cohorte de estudiantes semestralmente.

Por otra parte, actualmente existen otros tipos de datos que se manipulan como si fueran simples pero en realidad se trata de datos compuestos con una estructura bien conocida. Por ejemplo:

- Imágenes: se trata de una matriz o tabla de dos dimensiones (altura por anchura), donde cada elemento (o píxel) se define a su vez por una tupla de valores enteros, uno para cada canal (tres canales, RGB en el caso más habitual de imágenes a color).



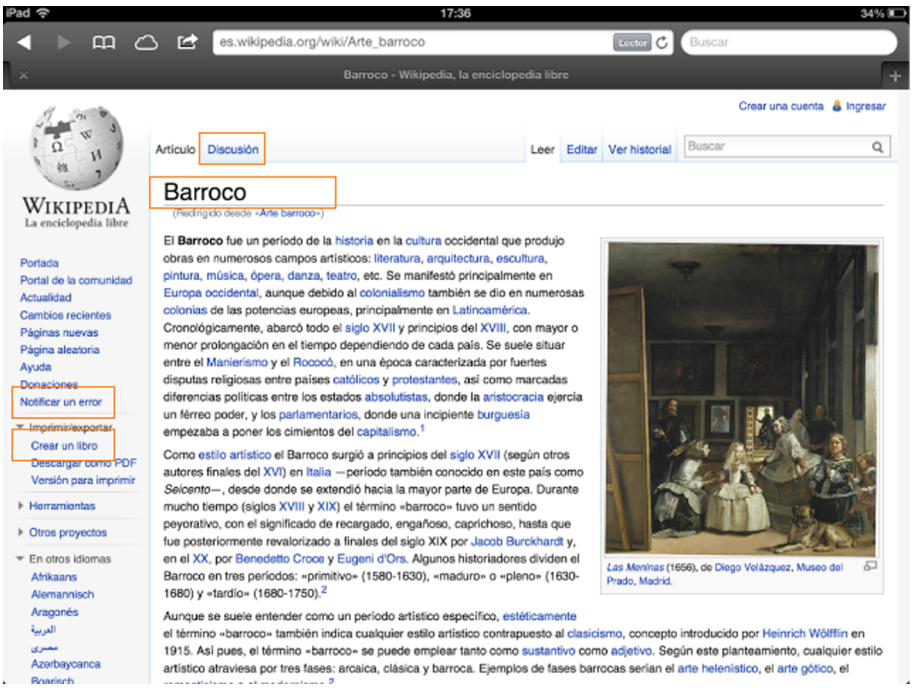
Fuente: http://vesta.astro.amu.edu.pl/Library/WWW/Tutorial1/graphics/display_primer.html

- Tuits: se trata de una estructura compleja que «rodea» los 140 caracteres, que son los que percibe el usuario que lo lee o escribe. Cada tuit ocupa de hecho unos pocos KB, ya que incluye información del emisor del tuit, la fecha y hora de creación, datos de geolocalización, referencias a otros tuits si se trata de un RT (retuit) o MT (mention), etc. Esta estructura está prefijada de antemano, aunque no todos los campos o atributos que describen un tuit pueden estar presentes en cada tuit.

- Publicaciones: se trata de un texto o documento, que puede estar en formatos muy diferentes, y que incluye unos datos típicos que lo definen, como un título, uno o más autores con sus respectivas afiliaciones, una fecha de publicación y dónde ha sido publicado. Aunque los datos más interesantes se encuentran en el propio documento (su contenido), también puede resultar interesante analizar el resto de los datos, ya que se pueden establecer relaciones entre documentos a partir, por ejemplo, de analizar la red de coautores que publican conjuntamente.

- Páginas de Wikipedia: se trata de una página web con un formato especial, que enlaza y/o es enlazada a/desde otras páginas web, todas ellas con la misma estructura y con el mismo tipo de contenidos (texto, enlaces, tablas, referencias, imágenes, etc.) y formando parte de una misma colección de páginas (por idioma). Wikipedia puede verse como una enorme red social

donde usuarios (anónimos y registrados) editan páginas que forman una red de contenidos enlazados entre sí. Cada página tiene un título, que la identifica unívocamente (es, de hecho, el enlace para acceder a ella) y un conjunto de datos adicionales que la complementan (historial de ediciones, página de discusión, etc.).



Fuente: Wikipedia

Un caso aparte son los **metadatos**. A veces los datos necesitan ser descritos mediante otros datos, es lo que se conoce como metadatos.

Ejemplo

Por ejemplo, una fotografía tomada con una cámara digital en formato JPEG tiene una resolución en píxeles y una profundidad de color (número de canales y resolución de cada canal), pero también puede ir acompañada de datos como la geolocalización del lugar donde se tomó la fotografía, la fecha y hora, con qué cámara se tomó, si se usó *flash* o no, etc. Este conjunto de metadatos (datos sobre datos) en el caso de las imágenes se conoce como Exif (*exchangeable image file format*). En el caso de las publicaciones, existen diferentes estándares para describirlas, como MARC 21 o Dublin Core, entre otros, ampliamente usados en bibliotecas y repositorios digitales.

En otras ocasiones lo que se desea representar no son características o atributos de una entidad, sino las **relaciones** existentes entre diferentes entidades. En este caso, los datos son 3-tuplas o tripletas de la forma [E1, R, E2], de modo que se puede leer «la entidad E1 está relacionada mediante R con la entidad E2».

Ejemplo

Por ejemplo, si «a» y «b» son usuarios de Twitter y R es la relación «el usuario es seguidor del usuario», el hecho de que el usuario «a» siga al usuario «b» se puede representar como:

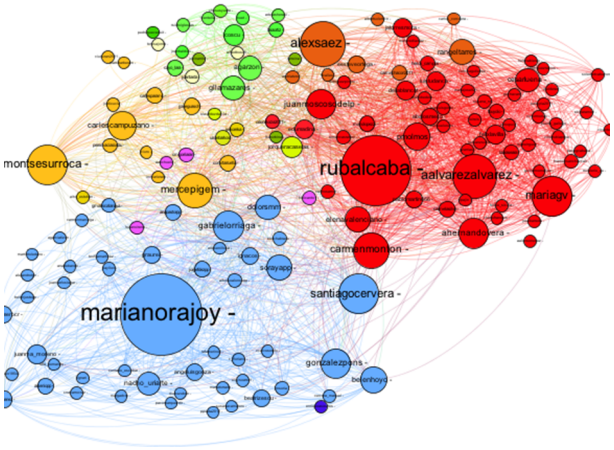
[a, «es seguidor de», b]

o más sencillamente como:

a → b

Obviamente, cuando se agregan dichos datos (en este ejemplo, para más de un usuario de Twitter), se puede construir un grafo con todas las relaciones existentes (es decir, quién es seguidor de quién en Twitter). Este grafo puede ser analizado mediante las técnicas adecuadas para, por ejemplo, descomponerlo en comunidades y subcomunidades y estudiar su estructura, detectar los elementos de mayor importancia, etc. El análisis de estas estructuras creadas a partir de las relaciones que se dan en las redes sociales es un ámbito de investigación muy importante hoy día.

Ejemplo de grafo en Twitter



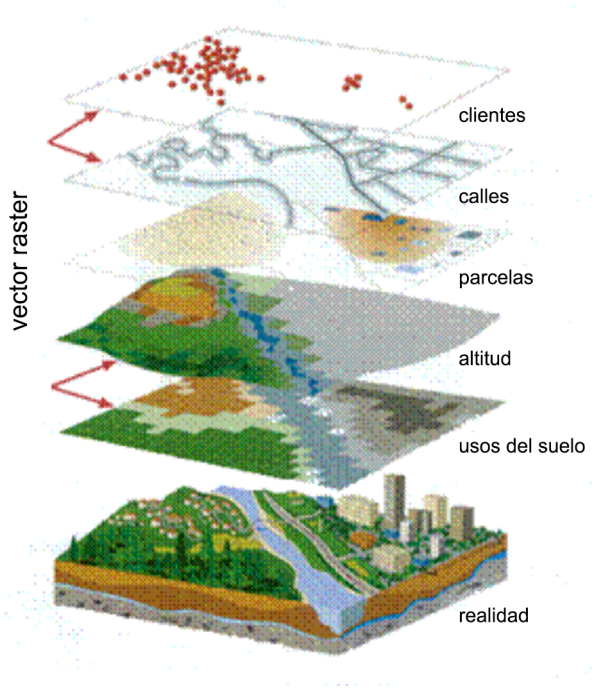
Fuente: http://www.k-government.com/2011/12/29/el-congreso-de-los-diputados-en-twitter/

Ejemplo

Otro ejemplo interesante sobre datos que describen relaciones entre datos lo proporciona la reciente iniciativa Wikidata, que pretende describir el contenido de Wikipedia de acuerdo con las relaciones que se establecen entre los diferentes elementos, incluyendo sus propios atributos. Así, si en Wikipedia existe una página para el concepto «Oxígeno», en Wikidata existe una entrada equivalente (aunque independiente del idioma) que lo enlaza, por ejemplo, con el concepto «Carl_Wilhelm_Scheele», mediante la relación «discoverer or inventor». A su vez, este está enlazado con el concepto «Stralsund», mediante la relación «place of birth» (lugar de nacimiento). Así, Wikidata configura una red que permite enlazar todo el conocimiento de Wikipedia, de manera que sea posible realizar consultas con una semántica más avanzada, por ej., «inventores nacidos en una cierta ciudad».

Otros datos que por su naturaleza tienen unas particularidades propias son, por ejemplo, los mapas o diagramas. En este caso existe una relación de proximidad (métrica) entre los elementos que aparecen en el mapa (es decir, la distancia entre ellos), que es la que determina su posición, siendo esta muy importante, a diferencia de la representación usada para los grafos, donde lo más importante son las relaciones entre elementos, no su posición. Normalmente, sobre los mapas se superponen capas con información que se desea relacionar con su situación para mostrar diferencias entre zonas, por ejemplo. En general, todos los datos procedentes del mundo real que se desean representar en un mapa se gestionan mediante lo que se conoce como sistema de información geográfica, teniendo en cuenta su referencia espacial, de modo

que es posible realizar búsquedas por proximidad o por criterios relativos a la naturaleza de dichos datos, por ej., superponiendo una capa relativa a la vegetación del terreno con otra relativa a su explotación agrícola.



Fuente: http://sabia.tic.udc.es/gc/Contenidos%20adicionales/trabajos/3D/Internet%20GIS/sig.html

3.3. Datos semiestructurados o no estructurados

Finalmente, por lo que respecta a su estructura, existen otros tipos de datos que se consideran o bien semiestructurados, básicamente debido a la complejidad de la estructura subyacente, o bien no estructurados.

Un ejemplo es este mismo documento de texto: está formado por caracteres o cadenas de texto, pero estos se agrupan en líneas, que a su vez se agrupan en párrafos, secciones, capítulos, etc. No existe una estructura prefijada para los documentos de texto, sino que depende de su tipología (libro, artículo, etc.), por lo que su estructura está parcialmente definida. Otro ejemplo serían las páginas web escritas en lenguaje HTML, un lenguaje de marcas que describe la posición y el significado de cada contenido (textos, tablas, imágenes, enlaces, etc.) en una página. Aunque a priori HTML es un lenguaje bien estructurado, su complejidad y la posibilidad de crear páginas que no respeten exactamente la sintaxis pero sigan siendo válidas hacen que su análisis sea complicado.

Los datos no estructurados suelen ser generados por humanos, mientras que los estructurados en muchos casos son generados por máquinas.

4. Ciclo de vida de los datos

En la literatura tradicional de la gestión de datos, los datos son generados (creados) o capturados (extraídos o encontrados), almacenados, preprocesados, analizados, visualizados y publicados, de manera que se cierra el círculo y se permite su reutilización. Cada una de estas fases tiene un objetivo, generando valor a partir de los datos en cada una de ellas. No todas las fases son estrictamente necesarias y, además, pueden solaparse o realizarse simultáneamente en algunos casos. De esta manera, se entiende que las fases típicas del ciclo de vida de los datos son las siguientes:

- Captura
- Almacenamiento
- Preprocesado
- Análisis
- Visualización
- Publicación

4.1. Captura

La fase de captura tiene como objetivo recopilar todos los datos que se generan durante un proceso, mediante dos mecanismos básicos complementarios:

- Creación: se trata de integrar en el propio proceso de generación de los datos un mecanismo que almacene los que se consideren relevantes cada vez que se generen.

Ejemplo

Por ejemplo, cada vez que un usuario paga con su tarjeta de crédito (o débito) en un establecimiento, se genera un nuevo registro que define perfectamente qué cantidad de dinero se ha gastado en qué establecimiento, en qué fecha y hora y con qué tarjeta.

- Extracción: se utiliza cuando no es posible intervenir en el proceso de generación de los datos, sino que es necesario irlos capturando conforme se van encontrando, idealmente inmediatamente después de su generación.

Ejemplo

Por ejemplo, es posible capturar los tuits que contienen una cierta palabra clave o *hashtag* tal y como van apareciendo en el flujo (*timeline*) de tuits de un usuario o el flujo público, ya que Twitter es un servicio que genera datos en abierto.

Es decir, o se tiene acceso a los datos *a priori*, en el mismo momento de su creación, o *a posteriori*, una vez estos han sido generados por otro proceso sobre el cual no se tiene posibilidad de intervención.

Un ejemplo puede ser el seguimiento de la navegación de los usuarios de un entorno virtual, para analizar qué servicios son los más requeridos por sus usuarios, cuáles son los caminos o itinerarios seguidos, etc. La opción que parece más sencilla es analizar el rastro que dejan los usuarios en los *logs* de los servidores web que alojan el entorno virtual, aunque realmente dichos *logs* contienen muchos otros datos relativos a la carga de los elementos que conforman cada página web, los cuales no aportan ninguna información relevante para analizar el comportamiento de los usuarios que navegan por él. La extracción de la navegación de los usuarios de un entorno virtual web es, en general, muy compleja y costosa. Por el contrario, si se tiene acceso al entorno web, es posible introducir marcas (código que genera los datos deseados) en aquellos servicios de los cuales realmente se quiere obtener información, recogiendo (generando) solamente aquellos datos que formarán parte del análisis posterior.

Desafortunadamente, no siempre se tiene acceso al nivel necesario para poder recoger los datos en el momento exacto de su creación, por lo que las estrategias de captura de datos basadas en la extracción de datos ya existentes o su generación mediante mecanismos alternativos son las más habituales, entre otras:

a) Acceso a los datos mediante un repositorio: como resultado de su publicación en abierto, los datos pueden estar disponibles en un repositorio digital, o simplemente en una web que permita su acceso. Normalmente los datos se encuentran ya en ficheros de acuerdo con una clasificación preestablecida, en función de su naturaleza o del dominio al cual pertenecen. Estos datos pueden verse como una fotografía de un proceso o escenario en un momento dado, por lo que pueden no haber sido actualizados recientemente. Por ejemplo, el Instituto Nacional de Estadística (INE) publica datos demográficos, como la población de cada municipio, con una periodicidad anual. Se trata, por lo tanto, de datos **estáticos**, a los cuales se accede como un conjunto entero, sin poder especificar qué datos se desean descargar. No obstante, cada vez es más común que los proveedores de datos incorporen mecanismos para facilitar el acceso a datos específicos, aunque no es una práctica habitual. La mayoría de los repositorios de datos en abierto solo ofrecen la opción de descargar ficheros con los datos y no incluyen ninguna opción para seleccionarlos a priori.

Ejemplos de repositorios con datos en abierto

Catálogo de datos en datos.gob.es, reutiliza la información pública: http://datos.gob.es/catalogo

European Union Open Data Portal: https://open-data.europa.eu/en/data/

U.S. Government's open data: https://www.data.gov/

Registry of Research Data Repositories: http://www.re3data.org/

b) Acceso mediante una API (*application programming interface*): en algunos casos, sí que es posible utilizar un mecanismo que permite realizar consultas específicas contra un conjunto de datos, obteniendo solamente aquellos que han sido requeridos de acuerdo con los parámetros de la consulta. Se puede hablar, entonces, de datos **dinámicos**, ya que estos se generan de acuerdo con la consulta realizada. Aunque el concepto de API fue inicialmente pensado para acceder a servicios y librerías de software, hoy día está también extendido su uso para el acceso e intercambio de datos, no solamente para la ejecución de servicios. Actualmente, la mayoría de los servicios como Twitter, Flickr y otras redes sociales permiten acceder a estas mediante una API, por lo que resulta posible desarrollar herramientas y aplicaciones que extraigan datos de dichas redes sociales de forma automatizada, sin apenas intervención humana excepto para iniciar la consulta. En algunos casos, las API están integradas en lo que se conoce como *sandbox*, un entorno controlado que permite a los desarrolladores efectuar pruebas y observar el funcionamiento de la API. Un aspecto importante que se debe tener en cuenta es la existencia de mecanismos de *throttling*, que limitan el número y la frecuencia de las peticiones que se realizan contra una API para evitar su colapso si es accedida de forma desenfrenada, o para limitar su uso hasta un cierto número de peticiones diarias (o mensuales) en un servicio gratuito, mientras que es posible superar dicho límite si se paga por el servicio de acceso a la API. Es lo que se conoce como modelo *freemium*, un concepto típico de servicios en línea como Spotify, por ejemplo.

Ejemplos de API

Los ejemplos de API disponibles son:

Twitter: https://dev.twitter.com/rest/public

Flickr: https://www.flickr.com/services/api/

Scopus: http://dev.elsevier.com/sc_apis.html

Listado de API: http://www.programmableweb.com/apis/directory

c) Manipulación de los parámetros de búsqueda: la mayoría de las veces, un repositorio digital o una web no disponen de una API para la captura de los datos deseados, sino que es necesario realizar operaciones de búsqueda mediante la interfaz de usuario disponible para acceder a estos, seleccionando aquellos conjuntos de datos que resultan de interés. En algunos casos, la propia navegación indica la arquitectura de la web, mediante los parámetros que aparecen en la URL y que van cambiando de acuerdo con la navegación del usuario. Entonces es posible manipular dichos parámetros y programar unos *scripts* sencillos que automaticen dicho proceso, accediendo a la misma URL una y otra vez pero modificando los parámetros de búsqueda, con lo que se obtienen diferentes resultados en cada llamada.

Ejemplo de manipulación de los parámetros de búsqueda y contenido

http://www.idescat.cat/nadons/?sexe=<SEXO>&res=<LUGAR>&t=<AÑO>&lang=es

Se pueden manipular el parámetro <SEXO> ("1" niños, "2" niñas, "0" todos, sin las comillas), el parámetro <LUGAR> (p.e., "a" para toda Cataluña, "d13" para la comarca del Barcelonès) y el parámetro <AÑO> (p.e., "2014"), obteniendo los nombres puestos a los nacidos de ese sexo en ese año en esa región (Datos IDESCAT). Por ejemplo:

http://www.idescat.cat/nadons/?sexe=0&res=a&t=2014&lang=es

d) Captura de datos mediante *scraping*: una situación aún peor es cuando ni tan solo es posible manipular manualmente las URL para acceder a los datos de manera semiautomática. En este caso la única opción posible es utilizar herramientas (también llamadas *bots*) que simulen la navegación de un usuario por un conjunto de páginas web y que extraigan el contenido de las páginas visitadas, analizando su estructura. Esto es lo que se conoce como *web scraping*. Actualmente las páginas web son creadas de forma automatizada, por lo que su estructura interna es bastante estable y coherente. Utilizando herramientas básicas (como, por ejemplo, la posibilidad de inspeccionar la página web desde el propio navegador), es posible determinar dicha estructura interna del documento HTML y programar un bot para acceder solamente a la información deseada.

Ejemplos de herramientas para hacer *web scraping*

Algunos ejemplos de herramientas para hacer web scraping son:

Scrapy: http://scrapy.org/

PhantomJS: http://phantomjs.org/

Otras herramientas: http://www.garethjames.net/a-guide-to-web-scrapping-tools/

e) Extracción de datos de documentos de texto: en ocasiones se publican datos en formatos no pensados para su reutilización, en PDF por ejemplo, en forma de tablas, listas, etc. En este caso es necesario extraer dicha información usando opciones del tipo corta y pega, pero esto no siempre es posible y se suele perder información relativa al formato de tabla. Afortunadamente existen otras herramientas que permiten extraer tablas directamente, generando datos en formato tabular listos para ser reusados. Cuando los datos se publican como imágenes escaneadas, es necesario recurrir a herramientas para el reconocimiento óptico de caracteres (en inglés, OCR), aunque este proceso no está exento de errores en función de la calidad de la imagen.

Ejemplos de herramientas para la extracción de texto

Herramientas para la extracción de texto:

Extracción de datos en tablas de PDF con Tabula: http://tabula.technology/

Reconocimiento de caracteres con Tesseract: https://github.com/tesseract-ocr

f) Formularios: a veces lo más sencillo y eficaz es preguntar directamente a los usuarios de un servicio, recurso o sistema, con el objetivo de recabar datos, tanto del servicio en cuestión como de los propios usuarios. Aunque se siguen haciendo muchas encuestas a pie de calle, se trata de un proceso lento y costoso, así como limitado en el tiempo y en el espacio. Actualmente existen muchas herramientas gratuitas para la creación y ejecución de formularios en línea, de modo que es posible acceder a una gran cantidad de usuarios a partir de sus direcciones de correo o mediante una red social.

Ejemplos para la creación de formularios en línea

Algunos ejemplos para la creación de formularios en línea:

Google Forms: https://www.google.com/forms/about/

LimeSurvey: https://www.limesurvey.org

g) *Crowdsourcing*: se trata de aprovechar la «sabiduría de las masas» (*wisdom of crowds*), que son capaces de resolver problemas que resultan muy complicados incluso para el estado del arte en temas de inteligencia artificial y minería de datos, aprovechando el intelecto humano.

Ejemplo

Por ejemplo, la figura 1 muestra un sistema de control anti- *bots* basado en *captchas*, que evita que una máquina se registre automáticamente en un servicio en línea, por ejemplo. El sistema obliga al usuario a introducir dos palabras, una de ellas distorsionada y que solamente un humano es capaz de leer, en este caso «admipa». La otra palabra («exposure») se trata en realidad de un caso de *crowdsourcing*, es una palabra proveniente de un sistema óptico de reconocimiento de caracteres que no ha podido ser procesada correctamente, por su baja resolución, por ejemplo. Cuando unos cuantos usuarios acceden al servicio en línea e introducen «exposure», el sistema aprende que ese es el texto asociado a esa imagen que no se había podido reconocer. Aunque algunos usuarios se equivoquen o mientan con respecto a dicha palabra, la mayoría de ellos seguramente la introducirá

correctamente, por lo que se puede dar por válida cuando un cierto número de usuarios coincide.

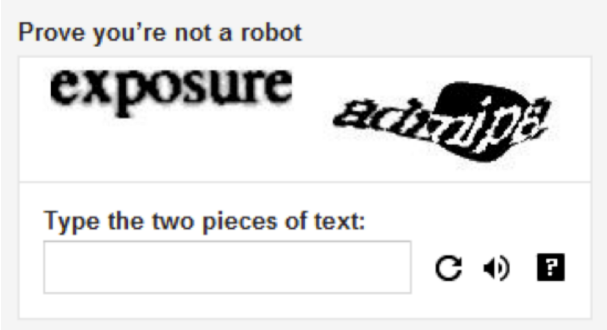


Figura 1. Ejemplo de *crowdsourcing* usando un *captcha*

h) Datos cualitativos: a veces es necesario obtener datos directamente de los usuarios porque interesa conocer de primera mano algunos aspectos que no pueden recogerse mediante una encuesta, especialmente el «¿por qué?» y el «¿cómo?», y aspectos difícilmente cuantificables, como emociones o sentimientos. Normalmente se recurre a entrevista semiestructuradas, donde mediante una serie de preguntas se obtiene información directa de los participantes. Estas entrevistas suelen ser grabadas y luego transcritas y codificadas, de acuerdo con unos criterios establecidos con anterioridad, de manera que se identifican aquellos aspectos que se querían conocer mediante la entrevista.

Ejemplo de herramientas para la captura de datos cualitativos

Herramientas para la captura de datos cualitativos:

RQDA: http://rqda.r-forge.r-project.org/

4.2. Almacenamiento

Los datos capturados son almacenados en un formato que permita su posterior manipulación, conforme a la representación más adecuada, teniendo en cuenta tanto su tipología como el uso que se va a querer efectuar de ellos. Sin entrar en detalles técnicos relativos a la infraestructura tecnológica subyacente, los datos se almacenan o bien en **ficheros simples** (o colecciones de ficheros), o bien en **bases de datos**, en cuyo caso se intenta reproducir de cierta manera la estructura de relaciones entre ellos. En función de su objetivo y complejidad se puede hablar de:

1) Ficheros simples: los datos son almacenados en ficheros (o colecciones de ficheros) de acuerdo con uno o más criterios, como por ejemplo el origen de los datos y/o la fecha de creación o captura.

Un ejemplo de esto pueden ser los ficheros de log generados por los servidores web, que contienen todas las peticiones que se realizan cuando los usuarios navegan por las páginas web de un servicio en línea. Normalmente se trata de ficheros en formatos planos que permiten una manipulación relativamente sencilla, aunque también puede darse el caso contrario, como una colección de documentos de texto o de imágenes, de acuerdo con algún formato concreto.

2) Bases de datos: se trata de una estructura más o menos compleja que permite representar los datos de acuerdo con su naturaleza, teniendo también en cuenta las relaciones entre todos los elementos que los componen. Se suele hablar de bases de datos relacionales, pero recientemente se han populariza-

do las llamadas no relacionales, que pretenden dar solución a problemas relacionados con la escalabilidad de las relacionales, especialmente para grandes volúmenes de datos.

En este segundo caso, es posible construir sistemas que optimizan ciertas operaciones:

- Almacenes de datos: del inglés *data warehouse*, son un tipo de bases de datos que está orientado a almacenar datos con el objetivo de optimizar las consultas y generar informes, agregando y resumiendo datos de diferentes fuentes. Suelen alimentarse de las diferentes bases de datos usadas en una organización y permiten un acceso centralizado con el objetivo de poder realizar consultas más complejas y eficientes.

- *Datamarts*: en este caso se trata de un subconjunto de un almacén de datos, con el propósito de dar soporte a un área específica, con unos objetivos concretos. Se puede ver como una capa de acceso sobre un almacén de datos que filtra y selecciona aquellos datos que desean ser analizados desde una perspectiva concreta, simplificando la toma de decisiones.

Un aspecto muy importante para facilitar la manipulación de los datos almacenados es que estos sigan algún formato de fichero abierto, de modo que no sea necesario ningún software específico para su manipulación, ni disponer de una licencia para ello, asegurando también su preservación a largo plazo, especialmente en el caso de ficheros, no tanto en el caso de bases de datos donde grandes nombres como IBM u Oracle se disputan el segmento, aunque también existen soluciones basadas en código abierto, como MySQL o PostgreSQL.

Por otra parte, es importante que los datos sean realmente accesibles y que el formato de fichero no sea solamente abierto por lo que respecta a la disponibilidad de software libre y a la ausencia de licencias, sino también porque sea posible acceder y manipular los datos almacenados en él. Los mismos datos pueden ser almacenados de muchas formas, y no todas ellas promueven su reutilización posterior. Es lo que se conoce como esquema de 5 estrellas, creado por Tim Berners-Lee y que se muestra en la figura 2.

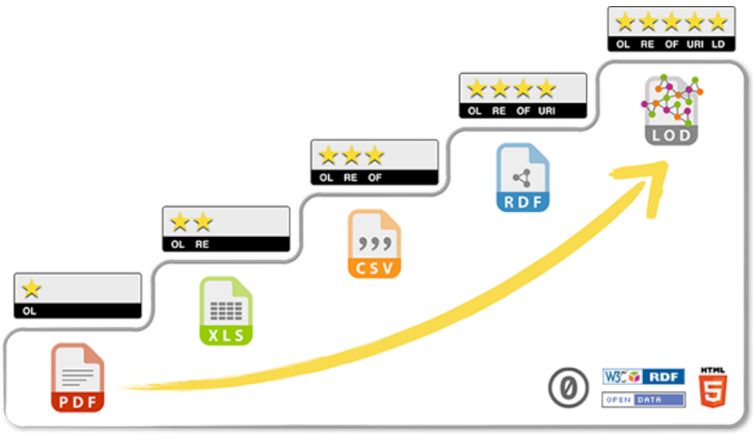


Figura 2. Esquema de 5 estrellas para la publicación de datos.
Fuente: http://5stardata.info/

El nivel más bajo (una estrella) corresponde a datos publicados en formatos que no permiten su extracción directa (excepto mediante herramientas como la mencionada Tabula). El segundo nivel sí que permite una manipulación del dato pero mediante software que requiere una licencia, como por ejemplo Microsoft Excel. El tercer nivel, el más usado en la actualidad, utiliza formatos no propietarios que también permiten la manipulación del dato. El cuarto y el quinto niveles tienen como objetivo alcanzar lo que se conoce como *linked data*, es decir, datos que están descritos de manera que se puede trazar su origen (los datos están descritos mediante URL, por lo que es posible apuntar hacia ellos) y además están enlazados con otros datos de terceros, creando entonces una red o grafo complejo que permite realizar consultas más elaboradas.

En el tercer nivel, cuando se trabaja con **datos no jerárquicos**, es decir, cuando todos los campos de un registro se encuentran al mismo nivel, se suele utilizar un formato plano como el CSV, o valores separados por comas, un fichero de texto donde cada registro ocupa una línea y los campos de cada registro van separados por comas «,» (o a veces, punto y coma «;»). Este formato se ha popularizado por su simplicidad y por la existencia de muchas herramientas que lo soportan, y es también posible manipularlo con un simple editor de textos y/o pequeños *scripts* o comandos del sistema operativo.

Por el contrario, cuando se trabaja con **datos jerárquicos**, se suele optar por formatos que permiten representar dicha jerarquía, así como la presencia o no de elementos opcionales, incluyendo entonces la posibilidad de alcanzar los niveles 4 y 5 del esquema de 5 estrellas. En este caso, se suelen utilizar los diferentes formatos:

JSON: acrónimo de *JavaScript object notation*, se trata de un formato ligero para el intercambio de datos, basado en el uso de parejas atributo-valor (con el formato "<atributo>": <valor>); es independiente del lenguaje de programación, y existen multitud de librerías para la manipulación de datos en dicho formato.

Por ejemplo, un sensor capturando datos de temperatura podría producir datos en formato JSON de la siguiente manera, indicando también su posición en ese momento (usando GeoJSON):

```
{ "measureID": "TS123456_1454070512",
  "sensorID": 123456,
  "date": "2016-01-29 12:28:32 GMT",
  "tempCelsius": 22.7,
  "position": { "type": "point",
    "coordinates": [2.1940713, 41.4056256, 16.96]
  }
}
```

XML: acrónimo de *extensible markup language*, se trata de un metalenguaje extensible mediante etiquetas, de forma que es posible definir datos mediante la especificación de un conjunto de etiquetas que los describen y la estructura jerárquica que componen. Está pensado también para el intercambio de datos de manera independiente respecto al lenguaje de programación usado, y existe una gran colección de adaptaciones para tipos de datos o ámbitos específicos, como por ejemplo MathML, un lenguaje para la descripción de contenido matemático, o XSIL, orientado al intercambio de datos científicos.

El ejemplo anterior en XML sería:

```
<sensorMeasure>
   <measureID>"TS123456_1454070512"</measureID>
   <sensorID>123456</sensorID>
   <date>"2016-01-29 12:28:32 GMT"</date>
   <tempCelsius>22.7</tempCelsius>
   <position>
     <type>"point"</type>
     <coordinates>
       <longitude>2.1940713</longitude>
       <latitude>41.4056256</latitude>
       <altitude>16.96</altitude>
     </coordinates>
   </position>
</sensorMeasure>
```

RDF: acrónimo de resource description framework, se trata también de un metalenguaje que permite describir recursos de acuerdo con un marco preestablecido, mediante el uso de tripletas sujeto-predicado-objeto.

Una posible representación parcial del ejemplo anterior en RDF (sin especificar los espacios de nombres) podría ser la siguiente:

"TS123456_1454070512" taken_by 123456

De hecho, la representación real en RDF es muy parecida a la de XML, pero incluyendo información relativa al origen de los datos, y describiéndolos mediante metadatos, de manera que sea más sencillo procesarlos de modo automático, no manual.

4.3. Preprocesado

El objetivo de esta etapa es preparar los datos para su análisis posterior, de forma que puedan ser usados directamente por cualquier investigador o data scientist, sin tener que preocuparse por aspectos relacionados con su calidad, procedencia, etc.

Entre otras operaciones típicas de esta etapa, se pueden destacar las siguientes:

1) Fusión: a veces los datos se obtienen de diferentes fuentes, por lo que es necesario combinarlos en una única estructura, normalmente una tabla.

Por ejemplo, puede que de un conjunto de usuarios de un servicio o red social se conozcan sus hábitos de navegación, por una parte, y sus hábitos de compra, por otra. Si cada usuario tiene un identificador que (valga la redundancia) lo identifica de forma unívoca en ambos conjuntos de datos, es posible usar dicho identificador como clave para realizar la fusión de los dos conjuntos, combinando registros de dos ficheros o tablas diferentes en uno solo para cada elemento.

2) Selección/filtrado: pensando en los datos como si fueran un hipercubo (que combina datos de diferentes fuentes ya fusionados), este proceso consiste en realizar cortes sobre dicho hipercubo, seleccionando aquellas regiones que contienen los datos de interés, de acuerdo con uno o más criterios de búsqueda. El caso más sencillo corresponde a la selección de datos de una tabla de acuerdo con un criterio.

Por ejemplo seleccionar solamente los datos de ciudadanos varones menores de 18 años de un municipio en concreto.

3) Conversión: a veces los datos están en formatos que dificultan su análisis posterior, por lo que es necesario convertirlos a un formato plano más cercano a la idea de tabla.

Por ejemplo, los datos de Twitter que son capturados en formato JSON contienen una gran cantidad de información relativa a la estructura de los tuits que puede ser innecesaria para el análisis, así que se pueden seleccionar solamente aquellos atributos que se desean (por ej. el texto del tuit) y volcarlos en un formato plano, por ej. CSV.

4) Limpieza: también conocida como *data cleansing*, consiste en eliminar todas las inconsistencias en los datos que puedan ser detectadas, ya sea borrándolos o marcándolos para una inspección manual posterior, más detallada, que persiga la «validez» de los datos, de modo que sean realmente útiles.

Por ejemplo, si se trata de usuarios de una red social y la edad indicada en el perfil del usuario es mayor de 100 años, se puede pensar que se ha cometido un error al introducirla, marcando dicho registro como posible candidato a ser borrado o corregido manualmente. O si se especifica el código postal, comprobar que este es válido y coherente con la población especificada, por ejemplo. Este proceso también incluye la resolución de incoherencias como las que pueden darse cuando se fusionan datos de diferentes fuentes, detectando por ejemplo si un mismo cliente aparece con datos diferentes, con lo que es necesario determinar cuál es la información correcta.

5) Agregación: en algunos casos, puede resultar interesante resumir un subconjunto de datos en uno solo, de forma que se simplifique el conjunto de datos original y se genere una nueva variable que tenga un mayor poder predictivo.

Por ejemplo, si para cada usuario se dispone de todas sus conexiones diarias a lo largo de un periodo de tiempo, se podrían agregar los datos semanalmente, con lo que se reduce el volumen de datos drásticamente.

6) Creación de nuevas variables/indicadores: a veces es necesario realizar cálculos sobre las variables o campos disponibles para, por ejemplo, realizar una conversión entre unidades o calcular el ratio entre dos variables, generando nuevas variables o indicadores. Este proceso está relacionado con la extracción de características y la reducción de la dimensionalidad, que suelen verse ya como operaciones que forman parte de la etapa de análisis.

Estas operaciones pueden realizarse de muchas formas diferentes, en función de la naturaleza de los datos y de las herramientas disponibles, normalmente combinando más de una herramienta. Las opciones más habituales son estas:

- Mediante una base de datos: los datos de diferentes fuentes se cargan en una base de datos (generalmente relacional) que respeta la estructura de estos, facilitando su manipulación posterior mediante sentencias SQL, que incluyen entre otras, SELECT y JOIN, por ejemplo.

- Mediante herramientas específicas: para la manipulación de datos con formatos diversos existen herramientas como OpenRefine, que permite realizar las operaciones mencionadas anteriormente y exportar el resultado a formatos más aptos para su análisis posterior.

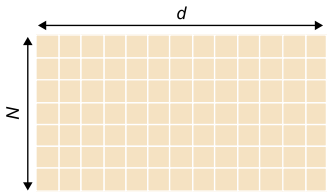
- Mediante lenguajes de programación: finalmente, utilizando lenguajes como Perl, y especialmente hoy día Python, es posible manipular datos en múltiples formatos de entrada (JSON, XML, etc.) y realizar operaciones sobre ellos de manera sencilla; existen incluso librerías y entornos que simplifican dichas operaciones.

Es importante destacar que esta etapa resulta muy importante y posibilita las que vienen a continuación, ya que la calidad de los resultados obtenidos (y de los modelos construidos) dependerá, directamente, de la calidad de los datos sobre los cuales hayan sido fundamentados. Es lo que se conoce como *garbage in*, *garbage out*, es decir, si se usan datos «basura» para crear modelos o visualizaciones, estas serán, con toda seguridad, también «basura». Además, asegura la obtención de un fichero con datos de calidad que pueden ser preservados a lo largo del tiempo, para futuras reutilizaciones.

4.4. Análisis

Una vez que los datos ya se consideran válidos, se puede proceder a su análisis.

> El objetivo de esta etapa es crear uno o más modelos que expliquen cómo son los datos y sus características principales, y poder responder a las preguntas planteadas en el marco del proyecto donde se estén utilizando dichos datos.



En este punto ya se piensa en los datos como en una tabla 2D que contiene una fila por cada elemento (también llamado observación) descrito del cual se tienen datos, en forma de atributos (también llamados variables), uno en cada columna. En general, se suele hablar de *N* datos de *d* dimensiones, aun cuando las dimensiones puedan ser a su vez estructuras de datos complejas. Estos dos parámetros (*N* y *d*) condicionan, en parte, el análisis que se puede realizar, junto con la naturaleza y tipología de los datos.

Si *N* es muy pequeño (unas pocas decenas de datos), resultará imposible construir modelos robustos, ya que resultará imposible generalizar los resultados obtenidos a nuevos datos de los cuales se pueda disponer en un futuro. En particular, con un valor de *N* pequeño seguramente será imposible observar el fenómeno deseado, y si se observa también será imposible asegurar que sea una característica del conjunto de datos, es decir, para valores de *N* pequeños es prácticamente imposible inferir nada, excepto describir exactamente el conjunto de datos del cual se dispone. El tamaño de *N* determina aspectos como los intervalos o niveles de confianza que se pueden estimar a partir de un conjunto de datos o muestra extraída de una población.

Por el contrario, si *N* es muy grande, será necesario realizar técnicas de muestreo para reducir el número de datos que se utilizan para construir un modelo, especialmente si este tiene una complejidad cuadrática o superior, tanto por lo que respecta a la complejidad espacial como la temporal. Un algoritmo con complejidad temporal cuadrática que ve incrementada la cantidad de datos en un orden de magnitud (es decir, *N* x *10*) necesitará 100 veces más tiempo que en el caso del conjunto de datos original con *N* elementos.

Por otra parte, el valor de *d* va asociado a un problema conocido, llamado maldición de la dimensionalidad, que aparece cuando *d* es realmente grande, del orden de cientos o miles. Si se piensa en los datos como en un hipercubo de *d* dimensiones, cuando *d* crece también lo hace el volumen de dicho hipercubo, de modo que, para un valor de *N* fijo, si d crece entonces los datos están más dispersos dentro de dicho hipercubo.

Por ejemplo, un conjunto de datos descrito mediante 10 variables binarias puede tomar 2^10 ≈ 10^3 valores diferentes, por lo que si *N* no es del orden de 1.000 elementos, se podría dar el caso de que hubiera un solo caso en cada «esquina» del hipercubo, con lo que resultaría muy difícil la extracción de características comunes o la agrupación por similitud.

Finalmente, la relación entre *N* y *d* también es importante. En función del tipo de análisis que se desea realizar, se suele recomendar que *N* sea, al menos, de un orden de magnitud mayor que *d*, es decir, que $N > 10 \times d$, pero se suele recomendar que *N* sea, al menos, mayor que $30 \times d$. Esta restricción suele causar problemas, por ejemplo, en el análisis de datos provenientes de una encuesta, donde el número de respondientes (*N*) puede ser insuficiente si el número de preguntas o ítems de la encuesta (*d*) es muy elevado. De ahí la necesidad de desarrollar técnicas para la reducción de la dimensionalidad de los datos, tratando de capturar toda su riqueza con un número menor de dimensiones *d'*, y mejorar así el ratio *N* / *d'*.

Así pues, una vez que se dispone de un conjunto de datos se puede proceder con el análisis; y en función de su naturaleza y de los objetivos que se persigan, este puede ser:

a) Análisis estadístico descriptivo: se trata de describir los datos mediante un conjunto reducido de valores que permitan modelarlos, de acuerdo con alguna distribución conocida, con el fin de describir apropiadamente las características esenciales de dicho conjunto, sin tener en cuenta si forman o no parte de un conjunto de datos mayor.

Por ejemplo, es habitual que en el análisis de los resultados de una encuesta se incluya un apartado que describa el sexo de los participantes, su edad, estado civil y situación laboral, entre otros, indicando el porcentaje de cada valor posible para cada variable.

b) Análisis estadístico inferencial: en este caso, se trata de modelar los datos de acuerdo con una distribución desconocida, teniendo en cuenta que, seguramente, solo se dispone de una fracción de la totalidad de los datos que conforman toda la población (en el sentido estadístico del término). El objetivo es inferir cómo es la totalidad de la población, asumiendo un grado de error en las estimaciones causado por la no disponibilidad de todos los datos, sino solamente de una fracción.

Así, por ejemplo, en el caso de una encuesta se suele especificar el número de participantes y, a partir de este, calcular los intervalos de confianza para los resultados obtenidos.

c) Extracción de características: a partir de todos los atributos disponibles, el objetivo es crear nuevos (llamados características o features) que capturen la naturaleza de los datos originales pero representando mejor al mismo tiempo aquellos aspectos que se desean analizar.

Por ejemplo, de un grupo de personas se podría disponer de diferentes medidas, como el peso (w) y la altura (h), entre otros. Si el objetivo es decidir si una persona está excesivamente delgada u obesa, en lugar de usar su peso y altura se podría utilizar el índice de masa corporal, que está más ligado al concepto de peso razonable para una altura.

Índice de masa corporal

$$\mathrm{IMC} = \frac{w}{h^2}$$

w (en kg)

h^2 (en m)

Esto es lo que se conoce como la extracción de características dependientes del ámbito de aplicación. Por otra parte, existen técnicas generales que permiten extraer las características a partir del análisis, generalmente, de la varianza de los datos disponibles. Entre las más conocidas, se pueden destacar el análisis de componentes principales (PCA) o el reciente (al menos por lo que respecta al término usado) *deep learning*, basado en la combinación de diferentes redes neurales para la extracción de conocimiento de los datos.

d) Reducción de la dimensionalidad: de manera análoga a la extracción de características, el objetivo es reducir el número de variables (d) para que el ratio N / d sea mejor, y permitir una representación de los datos usando recursos gráficos 2D o 3D, con el objetivo de detectar agrupaciones o disimilitudes. Además del ya mencionado PCA, existen otros algoritmos como el escalado multidimensional (en inglés MDS), por ejemplo, que permiten visualizar en 2D o 3D datos de dimensionalidad elevada.

e) Modelos supervisados: en este caso, se dispone de datos que incluyen una o más variables que se consideran el objetivo del problema que hemos de resolver. Por lo tanto, se construyen modelos que intentan predecir dicha variable (también llamada dependiente) en función de las otras (llamadas independientes).

Por ejemplo, utilizando la experiencia acumulada durante años, las compañías de seguros ajustan las pólizas (la cantidad a pagar cada mes) en función de diferentes parámetros, como la edad del conductor, su experiencia, el modelo de coche, su registro previo de partes e incidencias, etc., de manera que para un nuevo cliente potencial o para la renovación de uno existente sea posible determinar el valor óptimo (seguramente, desde la óptica de la compañía de seguros) de su póliza, ajustándola al máximo para que resulte atractiva para el cliente pero no arriesgada para la compañía.

Existe una larga colección de métodos y algoritmos para el aprendizaje supervisado; entre otros, se pueden destacar los árboles de decisión, las redes neuronales, el clasificador bayesiano ingenuo o los modelos lineales generalizados.

f) Modelos no supervisados: cuando no se dispone de una variable objetivo, el análisis de los datos se enfoca a compararlos entre sí y encontrar similitudes y diferencias, con el objetivo de detectar cualquier estructura interna que permita crear agrupaciones en función de algún criterio. Entran en esta categoría de aprendizaje no supervisado algoritmos como los de *clustering*, cierto tipo de redes neuronales, los mapas autoorganizativos y el algoritmo EM (*expectation-maximization*), entre otros. La creación de modelos no supervisados puede usarse también para la extracción de nuevas características que pueden ser usadas por otros algoritmos, especialmente supervisados.

g) Visualización: aunque esta metodología merece un apartado por sí misma, visualizar los datos disponibles en una primera opción es una estrategia de análisis muy recomendable, dado que para los humanos es muy fácil detectar patrones e irregularidades en los datos o en sus descriptores estadísticos básicos.

Por ejemplo, es posible realizar un análisis descriptivo utilizando visualizaciones basadas en el histograma de los datos o detectar relaciones entre variables mediante el uso de diagramas de dispersión.

El análisis no se limita a la construcción de modelos, sino que debe también explicar el resultado obtenido, mediante una interpretación del modelo y su puesta en contexto con respecto al problema original. Esto incluye también la evaluación del propio modelo, identificando qué variables o características son las más relevantes, la capacidad de generalización ante datos nunca utilizados previamente en la creación del modelo o su capacidad de adaptación a los cambios en los datos. Es en este punto donde un buen data scientist será capaz de establecer un equilibrio entre precisión, complejidad y actualización, eligiendo el modelo o la combinación de modelos más adecuados en función de los objetivos perseguidos y la naturaleza de los datos disponibles.

4.5. Visualización

Los humanos disponen de un sistema visual muy complejo y avanzado, que incluye desde el ojo con todos sus elementos (pupila, córnea, retina, etc.) hasta el córtex visual, el encargado de procesar toda la información recogida por el primero. Este sistema es capaz de capturar y procesar hasta 10 Mb/s, y de hecho, es conocido que el 93% de la información procesada por el cerebro es considerada no verbal. Los humanos son, principalmente, máquinas de procesado visual, con diversos subsistemas que se encargan de procesar eficientemente diferentes aspectos de dicha tipología de información de manera separada: forma, movimiento, color, etc.

Por lo tanto, cualquier aproximación basada en la transmisión de información visual es muy eficiente.

Por ejemplo, se considera que el procesado de información escrita alcanza un máximo de unas 1.000 palabras por minuto para personas muy capacitadas, y el promedio es de unas 200-250 palabras por minuto. El proceso de convertir imágenes en texto e interpretarlo es mucho más lento que el proceso de la imagen equivalente. El dicho «una imagen vale más que mil palabras» no deja de ser razonablemente cierto en este sentido. Las imágenes tienen una capacidad de captar la atención muy importante, y puede ser explotada para comunicar hechos (basados en datos) de forma más efectiva que mostrando los datos originales, por ejemplo.

Otro aspecto interesante de la visualización de datos es que puede convertirse en la propia interfaz de navegación de los propios datos, permitiendo ciertas operaciones básicas (selección, agregación, etc.), de manera que sea posible añadir cierta interactividad a la visualización. Siguiendo el mantra de Ben Shneiderman[3], es necesario presentar los datos resumidos, ofrecer las opciones

(3)Shneiderman, B. (1996, September). «The eyes have it: A task by data type taxonomy for information visualizations». *Visual Languages*, 1996. Proceedings., IEEE Symposium on (pp. 336-343). IEEE.

de zoom y filtro (selección) y, finalmente, añadir los detalles necesarios si estos son solicitados por el usuario. Así es posible combinar una gran cantidad de información en una misma visualización sin sobrecargar al usuario, dejándole la opción de determinar la cantidad de datos que desea visualizar. Concretamente (usando los términos originales descritos por Shneiderman):

- *Overview*: observar patrones en los datos.
- *Zoom*: seleccionar un subconjunto de los datos.
- *Filter*: seleccionar de acuerdo con un criterio o valor.
- *Detail on demand*: obtener valores para los datos seleccionados.
- *Relate*: comparar valores.
- *History*: mantener un registro de las acciones realizadas.
- *Extract*: marcar y extraer (capturar) los datos seleccionados.

Con este esquema, es posible basar el análisis de los datos a partir de su visualización, de forma que se combinen la capacidad visual humana por lo que respecta a la detección de patrones, tendencias, etc., con la potencia de un sistema informático que permita seleccionar, filtrar o comparar datos. Todo buen data scientist siempre empieza visualizando los datos para comprender mejor su naturaleza y para detectar posibles relaciones, dependencias y características que los definen, individual y grupalmente.

4.6. Publicación

Como etapa final que cierra el círculo, el resultado de las tres fases anteriores puede ser publicado en forma de nuevos datos, de manera que sea posible para terceros poder reutilizarlos con otros propósitos, especialmente los datos ya preprocesados listos para su análisis, en forma de una o más tablas.

Aunque hoy en día existen diversos mecanismos para compartir datos, es recomendable utilizar espacios que estén optimizados para ello, más allá de su simple publicación en una web. Concretamente, resulta muy ventajoso compartir los datos a través de un repositorio digital, de forma que se aseguren dos objetivos complementarios: su preservación y su diseminación. Por **preservación** se entiende el proceso que asegura la disponibilidad de los datos en un largo periodo de tiempo, teniendo en cuenta (e incluso interviniendo en) aspectos como el formato de los datos, su estructura interna, etc., de manera que sea posible acceder a ellos años después, aunque ya no existan las herramientas originales usadas para crearlos y almacenarlos. Por otra parte, la **diseminación** tiene como objetivo asegurar la reutilización de dichos datos, de forma que sea posible encontrarlos, acceder a ellos y entender su estructura

y formato. Para ello, resulta necesario disponer de los metadatos adecuados que describan los datos, además de indicar de claramente qué condiciones de uso delimitan lo que se puede hacer con ellos. De no ser así, resulta imposible en la práctica reutilizar dichos datos si se desconoce su origen y naturaleza.

Idealmente, el repositorio debería incorporar una API para sea factible desarrollar servicios y aplicaciones que puedan acceder a los datos deseados mediante una consulta bien especificada, sin intervención humana. También es interesante que el repositorio permita elegir en qué formato se devolverán los datos deseados, si en una tabla plana (por ej. usando el formato CSV) o mediante una estructura jerárquica (en cuyo caso los formatos JSON y XML son los más habituales).

Actualmente existen diversas iniciativas que promueven la publicación de datos en repositorios, desde perspectivas diferentes. Por un lado, existe un movimiento impulsado por administraciones y organizaciones públicas que pretenden ofrecer en abierto la mayoría de los datos que se generan en su seno, la mayor parte de ellos con origen en los propios ciudadanos o contribuyentes, intentado cumplir con criterios de transparencia y participación ciudadana. Por otro lado, también cada vez más empresas deciden compartir sus datos intentando hacer emerger nuevas aplicaciones que los exploten y permitan extraer valor añadido de ellos. En algunos casos, este desarrollo se promueve y soporta mediante una *hackathon* (combinación de *hacking* y *marathon*), un evento donde programadores e interesados en el tema pueden poner en práctica sus ideas. Cuando estos eventos están orientados a la manipulación y explotación de datos, suelen recibir el nombre de *datathon*.

Finalmente, un repositorio diferente que merece una atención especial es el UC Irvine Machine Learning Repository (o sencillamente, UCI ML), que tiene como objetivo la compartición de conjuntos de datos usados para la comparación y afinación de algoritmos de clasificación, regresión y predicción. Contiene más de 300 *datasets* de diferentes ámbitos de conocimiento, incluyendo los más característicos y conocidos de la literatura en minería de datos, como el *dataset* Iris de R.A. Fisher, por ejemplo, presentado en un trabajo de 1936 para mostrar las posibilidades del análisis discriminante lineal. Cualquier avance en el área de *machine learning* (un nuevo algoritmo o la afinación de los parámetros de un algoritmo conocido) pasa por la evaluación usando conjuntos disponibles en este repositorio.

Resumen

En este módulo se ha introducido el concepto de *data science*, usando el ciclo de vida de los datos para mostrar todas las competencias necesarias que incluye y requiere dicho ámbito de conocimiento. Se trata de un área multidisciplinar que combina conocimientos científicos, principalmente matemáticos y estadísticos, con otros provenientes de un contexto propio de la ingeniería, donde mediante el uso y la combinación de diferentes herramientas se resuelven problemas, como los relacionados con la captura y el preprocesado de los datos. Cuando esto se combina con el conocimiento del área de aplicación, se dispone del potencial adecuado para extraer conocimiento de los datos disponibles, obteniendo un valor añadido que puede ser usado para avanzar a la competencia y/o mejorar procesos y/o servicios, así como para entender mejor la naturaleza del área de aplicación en cuestión. La figura 3 resume esta triple combinación, situando *data science* en el centro de la intersección de los tres ámbitos mencionados.

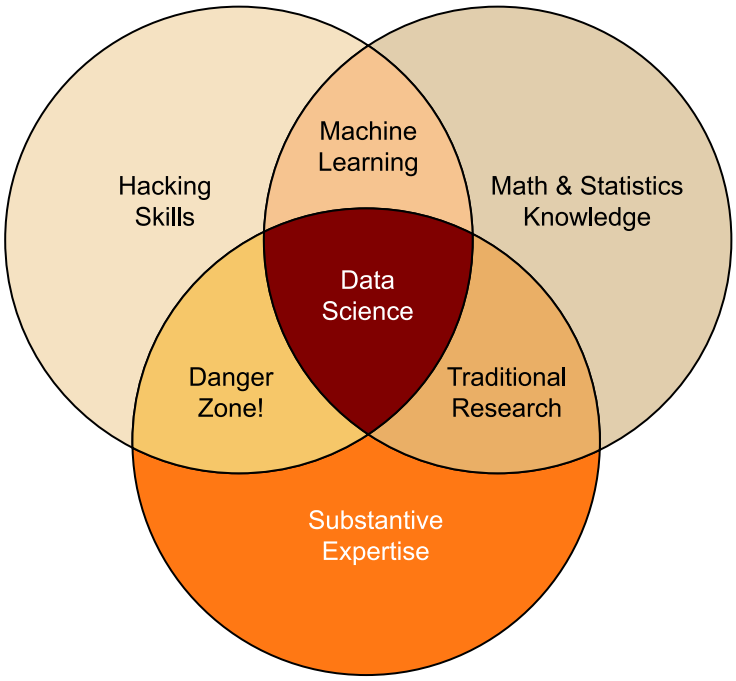


Figura 3. *Data science* como disciplina emergente.
Fuente: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Bibliografía

Cohen, Y.; Cohen, J. Y. (2008). *Statistics and Data with R: An applied approach through examples*. John Wiley & Sons.

Downey, A. (2012). *Think Python*. O'Reilly Media, Inc.

Fry, B.; Reas, C. (2007). *Processing*.

Leek, J. (2015). *The Elements of Data Analytic Style*. Leanpub.com.

Murray, S. (2013). *Interactive data visualization for the Web*. O'Reilly Media, Inc.

Scott, J. (2012). *Social network analysis*. Sage.

Spector, P. (2008). *Data manipulation with R*. Springer Science & Business Media.

Stanton, J. M. (2013). *Introduction to data science*.

Verborgh, R.; De Wilde, M. (2013). Using OpenRefine. Packt Publishing Ltd.

Witten, I. H.; Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.