Fonaments de *data science*

PID_00247382

Julià Minguillón

Temps mínim de dedicació recomanat: 3 hores

Universitat Oberta de Catalunya

Índex

1. La societat de la informació 5

2. Dades, informació, coneixement, saviesa? 10

3. Què és una dada? 12
3.1. Dades simples 12
3.2. Dades compostes o estructurades 12
3.3. Dades semiestructurades o no estructurades 16

4. Cicle de vida de les dades 17
4.1. Captura 17
4.2. Emmagatzematge 21
4.3. Preprocessat 24
4.4. Anàlisi 26
4.5. Visualització 29
4.6. Publicació 30

Resum 33

Bibliografia 35

1. La societat de la informació

La ràpida acceleració de la tecnologia pròpia de la societat industrial ha donat pas a l'anomenada societat de la informació, basada en l'ús intensiu de les tecnologies de la informació i la comunicació (TIC). Actualment és més important (tant en valor econòmic com estratègicament) la informació que es genera, gestiona i distribueix que no pas el maquinari que suporta aquests processos. L'aparició d'internet i, posteriorment, la seva popularització mitjançant la World Wide Web a partir de 1994 han permès a un percentatge important de la població mundial canviar la seva manera de comunicar-se, estudiar, treballar i relacionar-se amb altres persones. Segons dades del 2014, el percentatge de població connectada a internet és del 40%, xifra que arriba fins al 78% en el cas dels països desenvolupats. No obstant això, la societat de la informació no és exclusiva de l'ús d'internet. La progressiva digitalització de recursos i serveis del món real n'ha permès l'expansió i han globalitzat aspectes comuns com l'educació, la feina o el consum, entre d'altres, mitjançant diferents dispositius, incloent-hi ordinadors i telèfons mòbils però també targetes de crèdit, per exemple.

Els usuaris d'aquesta societat de la informació produeixen una gran quantitat de dades quan interaccionen amb aquest món digital. Totes les accions dutes a terme per un usuari són susceptibles de ser analitzades a partir del rastre que generen, amb el doble objectiu de millorar l'experiència de l'usuari i d'avançar en el coneixement que se'n té. Cada dia milions d'usuaris paguen amb les seves targetes de crèdit, utilitzen els seus telèfons mòbils, es connecten a internet mitjançant els seus ordinadors, estacionen els vehicles en aparcaments intel·ligents, fan fotografies o vídeos, accedeixen a edificis mitjançant una targeta d'identificació, utilitzen la xarxa de transport públic, etc. Tots els àmbits de la vida humana estan amarats de tecnologia que transforma totes les nostres accions en dades factibles de ser manipulades, emmagatzemades, analitzades i visualitzades, amb finalitats ben diferents. D'altra banda, molts altres processos també generen dades que són factibles de ser o bé capturades i processades mitjançant sensors o indicadors que permeten conèixer el seu estat, o bé emmagatzemades directament un cop generades.

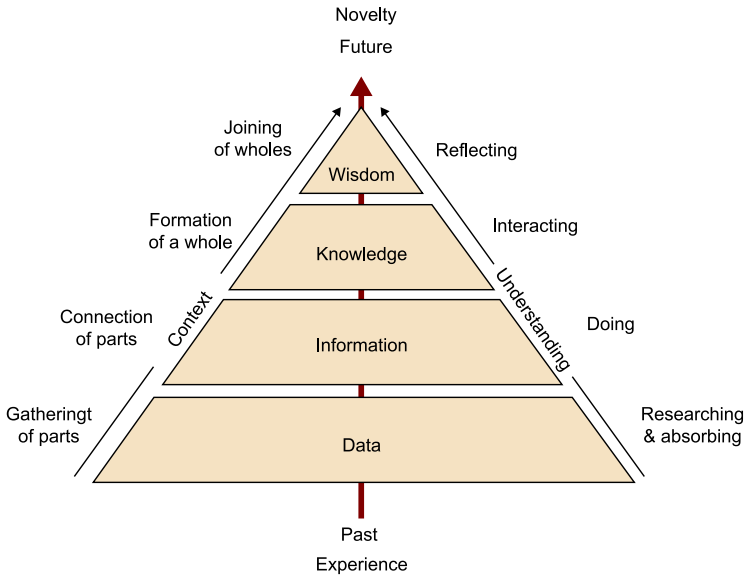
Així doncs, aquesta societat de la informació, o també del coneixement, com prefereixen anomenar-la altres autors, està basada en la generació i intercanvi de dades entre usuaris, serveis i recursos i mediada per la tecnologia digital. Encara que informació i coneixement són conceptes diferents, tots dos es basen en un element primordial: les dades.

> És el que es coneix com la piràmide D-I-K-W (*data*, *information*, *knowledge* i *wisdom*), de manera que la informació es defineix a partir de les dades disponibles, el coneixement s'extreu d'aquesta informació i la saviesa és entesa com l'habilitat per a aplicar aquest coneixement en benefici propi o comú[1].

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

Aquest procés combina mètodes i tècniques de diferents àmbits com la psicologia, l'estadística, la intel·ligència artificial, la mineria de dades i la visualització, entre d'altres. Es tracta, doncs, d'una àrea **multidisciplinària** que incorpora perfils molt diferents, atesa la seva complexitat i amplitud.

Piràmide D-I-K-W



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

Actualment hi ha una gran confusió entorn del nom més adequat per a aquest nou àmbit del coneixement, depenent de l'aproximació que s'assumeixi. És habitual parlar de *data science*, però també és possible utilitzar *data engineering* o, senzillament, *data mining*. Una possible solució és plantejar-ho en termes dels objectius que es tenen en funció de cada perfil:

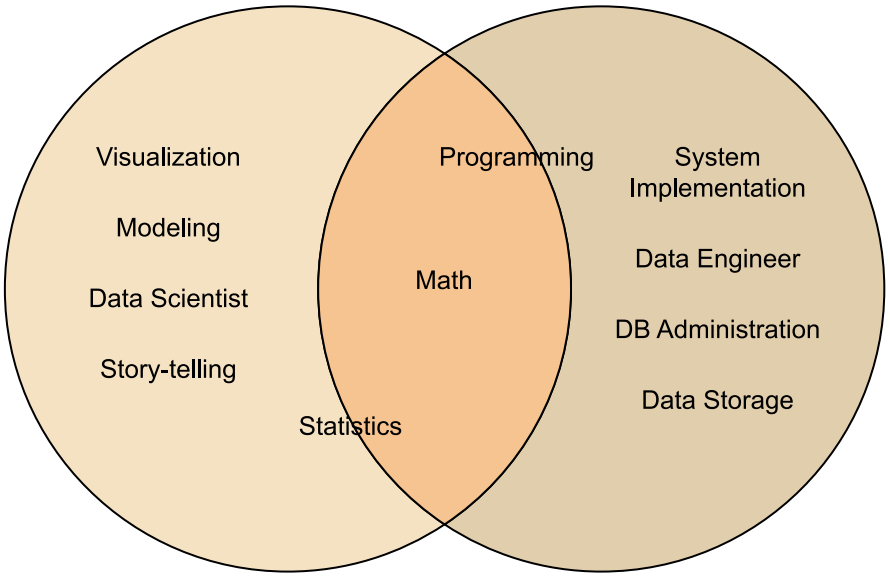
- Un *data scientist* és una persona que és capaç de plantejar-se les preguntes adequades a partir d'un conjunt de dades relatiu a un domini, i establir quins mètodes i tècniques són els més adequats per extreure el coneixement necessari per a respondre aquestes preguntes, per posteriorment dur a terme aquesta tasca. Aquest perfil està orientat principalment a resoldre el «què?».

- Un *data engineer* és una persona capaç de preparar un conjunt de dades de manera que tingui l'estructura i la informació adequades per a una posterior anàlisi, així com de portar a la pràctica una solució o prova de concepte i convertir-la en una implementació que pugui usar-se en un entorn productiu real. Aquest perfil està més orientat a resoldre el «com?».

- Un *data miner* és una persona que, a partir d'un conjunt de dades amb un objectiu ben definit, és capaç d'utilitzar diferents mètodes i eines per

maximitzar el resultat esperat. És, per tant, una especialització del primer perfil orientada específicament a l'anàlisi de les dades.

Funcions del *data scientist* i del *data engineer*



Aquests perfils es complementen i, de fet, comparteixen alguns objectius i habilitats necessàries per dur-los a terme. La resposta a una pregunta inicial relacionada amb dades mai és un procés completament lineal, sinó que normalment cal dur a terme diferents iteracions en aquest procés, de manera que sigui possible anar millorant els diferents elements que el constitueixen, entre d'altres:

- La pregunta que es desitja respondre.
- Les dades de les quals es disposa.
- Els mètodes i tècniques més adequats.
- L'avaluació i interpretació dels resultats obtinguts.

L'ordre d'aquests elements no sempre és el mateix. El més habitual és que a partir d'un fet observat en determinat entorn o context, els responsables d'aquest entorn es plantegin preguntes que necessitin respostes, idealment a partir de les dades disponibles o que puguin obtenir-se. Aquesta és l'aproximació tradicional de l'estadística: intentar demostrar o refutar una hipòtesi inicial a partir de les dades disponibles.

Per exemple, davant un descens de les vendes, una companyia podria preguntar-se per les raons que han causat aquest descens. Utilitzant instruments convencionals com enquestes i entrevistes amb els seus consumidors habituals i fent servir tècniques estadístiques per a l'anàlisi, es podrien determinar els factors que han causat el canvi d'hàbits de consum o concloure que es tracta d'un problema relacionat amb l'aparició de nova competència en el sector, per exemple.

En altres ocasions, tanmateix, és possible que, observant les dades ja disponibles, es detectin patrons o irregularitats que permetin plantejar-se noves preguntes, mai abans imaginades. El nou paradigma big data és un clar exemple de les oportunitats que ofereix analitzar dades a gran escala. És possible plantejar-se preguntes a les quals mai abans s'havia pogut respondre, no solament per limitacions relacionades amb la infraestructura tecnològica, sinó també per l'absència de les dades necessàries.

Per exemple, una cadena de supermercats podria analitzar tots els patrons de compra dels seus clients durant un període de temps i detectar quins productes es compren de manera conjunta i amb quina freqüència, amb l'objectiu de proposar ofertes i/o establir noves estratègies que promoguin el consum d'altres productes semblants o en promoció.

No obstant això...

...tal com indica Jeff Leek, la paraula clau en *data science* és *science*, i no pas *data*.

Finalment, un altre concepte molt de moda avui dia és el de **periodisme de dades**.

Bàsicament, es tracta de convertir dades en històries utilitzant mitjans digitals per a narrar-les, especialment visualitzacions interactives, que permetin al lector participar-hi i seleccionar aquells fragments de la història que li semblin més interessants.

Però el periodista de dades també és capaç d'anar a buscar les dades allà on es troben (especialment dades en obert) utilitzant les eines adequades o de trobar les connexions entre diferents fonts de dades que permeten localitzar la notícia enterrada en dades oficials[2].

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

Exemple

Per exemple, és possible creuar les dades relatives al preu del barril de petroli, que publiquen diàriament diferents organismes oficials, amb els preus dels carburants a la xarxa d'estacions de servei espanyoles, publicats pel Ministeri d'Indústria, Energia i Turisme. El resultat d'aquesta anàlisi permet provar la tan temuda sensació que, quan el preu del petroli puja, el preu dels carburants també puja immediatament, mentre que quan el preu del petroli baixa, els carburants ho fan més a poc a poc i les distribuïdores aprofiten el marge disponible i el repercuteixen en els usuaris finals.

Així doncs, un bon expert en dades és aquella persona que sap plantejar-se les preguntes adequades en un àmbit de coneixement concret, i que coneix els mecanismes per a obtenir les dades necessàries i preparar-les per a la seva posterior anàlisi mitjançant les tècniques estadístiques i de mineria de dades més adequades, per a interpretar els resultats obtinguts i posar-los en context relacionant-los amb les preguntes plantejades i l'àmbit de coneixement contrastant amb resultats anteriors, així com per a visualitzar el coneixement extret durant aquest procés per a identificar els patrons, les variables i altres elements rellevants que són clau en aquest procés i que permeten avançar en la seva comprensió. Es tracta, per tant, de combinar un arsenal d'eines i tècniques científiques, que inclouen, entre d'altres, eines matemàtiques, informà-

tiques, de visualització, analítiques, estadístiques, de disseny experimental, de definició de problemes, de construcció de models i de validació. Tot això amb l'objectiu de convertir les dades en coneixement.

2. Dades, informació, coneixement, saviesa?

Què és una dada, aleshores? ¿Quan una dada es converteix en informació o, millor encara, en un coneixement que es pot aprofitar? La piràmide D-I-K-W és el marc conceptual que permet entendre i donar significat a aquests termes. Per exemple:

«42»

És una dada, en aquest cas un nombre enter. Es pot donar per entès que està escrit (representat) en base decimal, encara que això no seria tan obvi en segons quins contextos on altres sistemes de numeració com el binari o l'hexadecimal són més habituals. També es podria haver representat mitjançant la frase *quaranta-dos*, encara que aleshores només els catalanoparlants reconeixerien aquesta dada, mentre que els habitants de la costa est d'Àfrica probablement preferirien la frase *arobaini* na *miwili*. Per tant, tota dada necessita una **representació** adequada perquè se'n pugui fer un ús correcte.

Una **dada** és, en principi, una quantitat o qualitat que descriu un atribut d'una entitat, dins d'un rang de valors possibles. És un valor «donat» referent a alguna cosa observada, d'acord amb l'arrel llatina que dona origen al terme (*datum*).

De totes maneres, és factible preguntar-se: 42 què?, a què es refereix aquesta dada? És en aquest moment quan una dada es converteix en informació, quan és capaç de respondre una pregunta concreta i adquireix significat, per exemple:

«Quina és la temperatura del pacient?»

En aquest cas, 42 respon a aquesta pregunta. Però ho fa de manera adequada? És molt fàcil adonar-se que no:

- Si es tracta de 42 graus centígrads, el pacient té una febre molt forta.
- Si es tracta de 42 graus Fahrenheit, el pacient és un cadàver fred.
- Si es tracta de 42 graus Kelvin, el pacient és un cadàver fred surant per l'espai exterior, segurament.

Per tant, una dada no és una informació veraç si no va acompanyada d'una **precisió** i unes **unitats** que la defineixin adequadament. Si per a la mateixa pregunta anterior la resposta fos «quaranta i alguna cosa», la vida del pacient podria estar en perill per manca de precisió.

Aquest exemple pot semblar anecdòtic o fins i tot trivial, però el 23 de setembre de 1999 la NASA va perdre el contacte amb la *Mars Climate Orbiter*, un satèl·lit dissenyat per a estudiar la superfície, atmosfera i clima del planeta Mart. La raó va ser que el satèl·lit va entrar en òrbita a una altitud insuficient, la qual cosa en va causar la destrucció. El motiu d'aquest error va ser que una part del programari utilitzat per al càlcul de les trajectòries orbitals utilitzava el sistema mètric decimal, mentre que altres mòduls del programari usaven el

sistema basat en unitats angleses (peus, polzades, etc.). Aquest error va costar a la NASA (de fet, als ciutadans nord-americans) un total de 327,6 milions de dòlars, el total de construir el satèl·lit, llançar-lo a l'espai i controlar-lo fins a la seva posada en òrbita, sense tenir en compte els problemes d'imatge i el retard en la missió original.

En aquest exemple, la informació disponible pot aportar coneixement sobre l'estat del pacient. L'experiència acumulada en el tractament de casos similars diu que un pacient amb una febre o temperatura corporal de 42 graus centígrads pot patir lesions cerebrals irreversibles. Aquest coneixement ha estat extret de múltiples exemples previs, on la majoria dels pacients en la mateixa situació desenvolupaven aquesta patologia. És aquest coneixement el que permet transformar una informació basada en dades en una sèrie d'accions que permetin canviar l'estat de l'entorn, actuant sobre aquells elements que tenen incidència sobre el procés que ha estat «mesurat», en aquest cas la temperatura corporal del pacient. Així doncs, aquest coneixement podria expressar-se mitjançant una regla, per exemple:

Exemple

«Si la temperatura del pacient assoleix els 42 graus centígrads, es poden produir lesions cerebrals irreversibles».

Finalment, és l'aplicació d'aquest coneixement (i la seva actualització constant) el que permet aconseguir la saviesa tal com aquesta es defineix en la piràmide D-I-K-W. No obstant això, molts autors prefereixen no definir *saviesa* (ni incloure-la en aquest marc) atesa la seva ambigüitat i centrar-se en la transformació de dades en informació i l'extracció de coneixement d'aquesta informació, i, això sí, aplicar-lo. En aquest exemple, aquest coneixement sobre la temperatura del pacient i les seves implicacions hauria de ser aplicat amb l'objectiu de reduir-ne la temperatura corporal de la manera més ràpida possible.

3. Què és una dada?

Les dades poden presentar-se de moltes maneres. En l'exemple de «42», es tractava d'un nombre decimal sencer, però les dades són de naturalesa molt diversa i es poden classificar d'acord amb diferents criteris, entre altres, segons la seva estructura; així, tenim dades:

- **Simples**: dades atòmiques indivisibles, amb un significat propi, d'acord amb la definició (un valor d'un atribut).

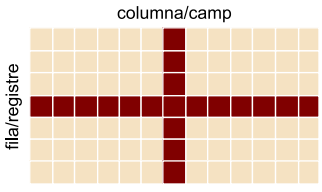
- **Compostes** o **estructurades**: dades que són una combinació d'altres dades simples i/o compostes, d'acord amb una estructura fixa i coneguda a priori.

- **Semiestructurades** o **no estructurades**: dades estructurades que segueixen una estructura parcial, que poden canviar segons el context o que no segueixen cap estructura.

3.1. Dades simples

Exemples de dades simples són els nombres enters, els reals (però no els complexos, que serien un exemple de dades estructurades), els caràcters i, excepcionalment també, les cadenes, tot i que aquestes són en realitat seqüències de caràcters, per la qual cosa són un tipus de dada composta (no obstant això, la seva popularitat i utilització provoquen que sigui més senzill pensar en les cadenes com un tipus de dada simple). Un altre exemple semblant és la tupla [latitud, longitud], que identifica una posició en un mapa, ja que es tracta de dos nombres reals que adquireixen significat quan es processen conjuntament.

3.2. Dades compostes o estructurades

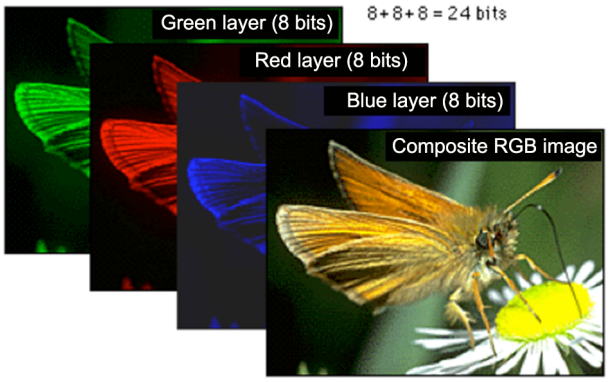


Al seu torn, les dades s'agrupen en altres estructures més complexes d'acord amb la seva dimensionalitat. Una dada concreta o puntual no té dimensionalitat (o es podria dir que té zero dimensions). Una seqüència de dades (per exemple, provinents d'un sensor de temperatura) constitueix un *array* 1D o vector d'una dimensió. Una estructura que descriu, per exemple, els productes d'un comerç (amb el seu codi, la seva descripció, la seva fotografia, el seu preu de cost i el seu preu de venda al públic) és el més habitual, i en aquest cas es parla d'array 2D, taula o matriu. Cadascun dels productes ocupa una **fila** de la taula, mentre que cadascun dels atributs que el descriuen ocupa una **columna**. Cada fila de la taula també se sol anomenar **registre**, mentre que cada atribut que descriu un registre se sol anomenar **camp**, seguint la nomenclatura pròpia de les bases de dades relacionals. El concepte de **dimensionalitat** pot ser estès a més de dues dimensions, òbviament. Per exemple, la mateixa taula 2D pot

veure's com un «tall» d'una taula 3D on, en la tercera dimensió, es disposa de la taula 2D en diferents moments, per exemple, el seguiment d'una cohort d'estudiants semestralment.

D'altra banda, actualment hi ha altres tipus de dades que es manipulen com si fossin simples però en realitat es tracta de dades compostes amb una estructura ben coneguda. Per exemple:

- Imatges: es tracta d'una matriu o taula de dues dimensions (altura per amplària), on cada element (o píxel) es defineix al seu torn per una tupla de valors sencers, un per a cada canal (tres canals, RGB en el cas més habitual d'imatges a color).



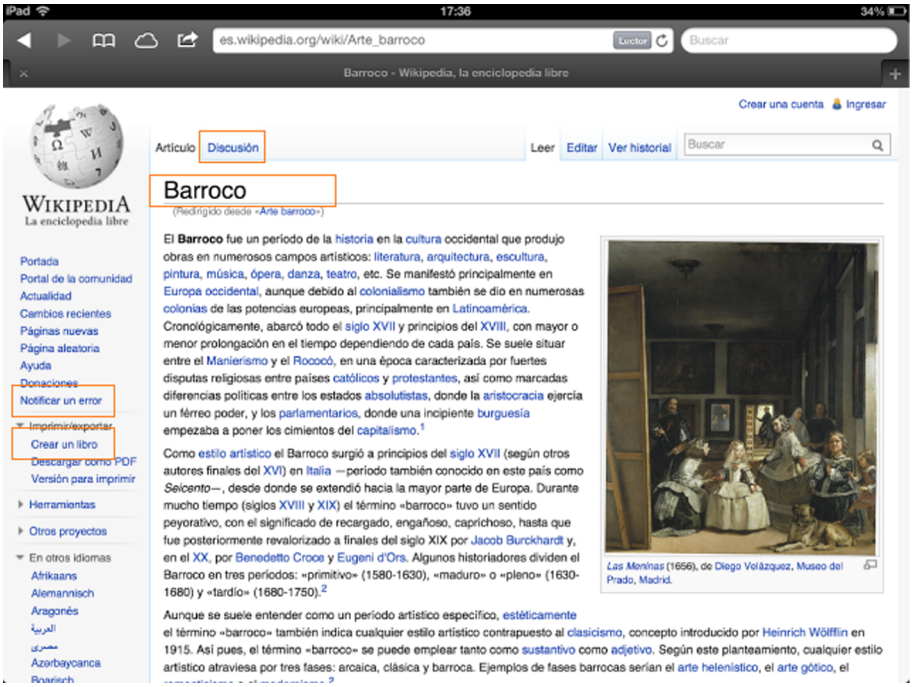
Font: http://vesta.astro.amu.edu.pl/library/www/tutorial1/graphics/display_primer.html

- Tuits: es tracta d'una estructura complexa entorn dels 140 caràcters, que són els que percep l'usuari que el llegeix o escriu. Cada tuit ocupa de fet uns pocs KB, ja que inclou informació de l'emissor del tuit, la data i hora de creació, dades de geolocalització, referències a altres tuits si es tracta d'un RT (retuit) o MT (mention), etc. Aquesta estructura està prefixada per endavant, encara que no tots els camps o atributs que descriuen un tuit poden estar presents en cada tuit.

- Publicacions: es tracta d'un text o document, que pot estar en formats molt diferents, i que inclou unes dades típiques que el defineixen. Aquestes dades podrien ser un títol, un o més autors amb les seves respectives afiliacions, una data de publicació i on ha estat publicat. Tot i que les dades més interessants es troben en el propi document (el seu contingut), també pot resultar interessant analitzar la resta de les dades, ja que es poden establir relacions entre documents a partir, per exemple, d'analitzar la xarxa de coautors que publiquen conjuntament.

- Pàgines de Wikipedia: es tracta d'una pàgina web amb un format especial, que enllaça i/o és enllaçada a/des d'altres pàgines web, totes amb la mateixa estructura i amb el mateix tipus de continguts (text, enllaços, taules, referències, imatges, etc.) i que formen part d'una mateixa col·lecció de pàgines (per idioma). Wikipedia pot veure's com una enorme xarxa social on usuaris (anònims i registrats) editen pàgines que formen una xarxa de

continguts enllaçats entre si. Cada pàgina té un títol que la identifica unívocament (és, de fet, l'enllaç per accedir-hi) i un conjunt de dades addicionals que la complementen (historial d'edicions, pàgina de discussió, etc.).



Font: Wikipedia

Un cas apart són les **metadades**. De vegades les dades han de ser descrites mitjançant altres dades. Aquestes dades que descriuen altres dades es coneixen com a metadades.

Exemple

Per exemple, una fotografia feta amb una càmera digital en format JPEG té una resolució en píxels i una profunditat de color (nombre de canals i resolució de cada canal), però també pot anar acompanyada de dades com la geolocalització del lloc on es va fer la fotografia, la data i hora, amb quina càmera es va fer, si es va fer servir *flaix* o no, etc. Aquest conjunt de metadades (dades sobre dades) en el cas de les imatges es coneix com a Exif (*exchangeable image file format*). En el cas de les publicacions, hi ha diferents estàndards per a descriure-les, com MARC 21 o Dublin Core, entre altres, àmpliament utilitzats en biblioteques i repositoris digitals.

En altres ocasions el que es desitja representar no són característiques o atributs d'una entitat, sinó les **relacions** que hi ha entre diferents entitats. En aquest cas, les dades són 3-tuples o tripletes de la forma [E1, R, E2], de manera que es pot llegir «l'entitat E1 està relacionada mitjançant R amb l'entitat E2».

Exemple

Per exemple, si «a» i «b» són usuaris de Twitter i R és la relació «l'usuari és seguidor de l'usuari», el fet que l'usuari «a» segueixi a l'usuari «b» es pot representar com:

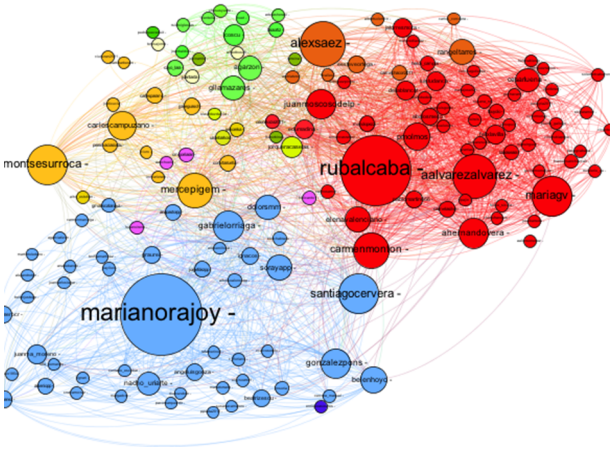
[a, «és seguidor de», b]

o més senzillament com:

a → b

Òbviament, quan s'agreguen aquestes dades (en aquest exemple, per a més d'un usuari de Twitter), es pot construir un graf amb totes les relacions existents (és a dir, qui és seguidor de qui a Twitter). Aquest graf pot ser analitzat mitjançant les tècniques adequades per a, per exemple, descompondre'l en comunitats i subcomunitats i estudiar-ne l'estructura, detectar els elements de més importància, etc. L'anàlisi d'aquestes estructures creades a partir de les relacions que es donen a les xarxes socials és un àmbit de recerca molt important avui dia.

Exemple de graf a Twitter

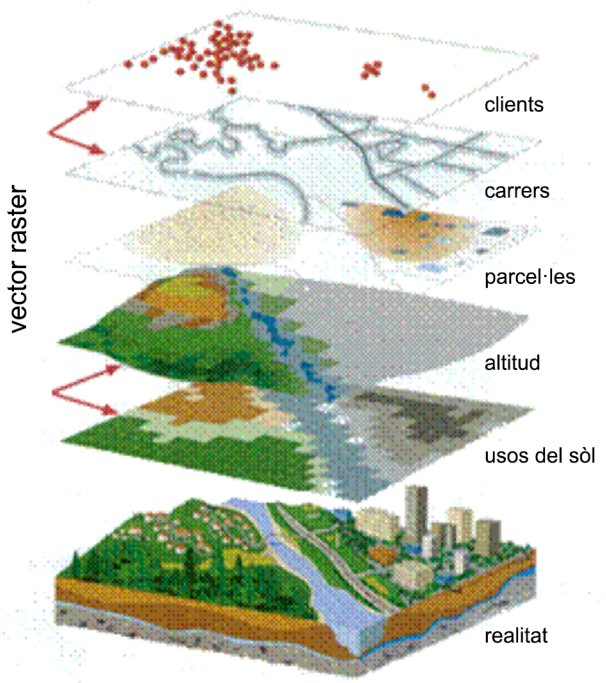


Font: http://www.k-government.com/2011/12/29/el-congreso-de-los-diputados-en-twitter/

Exemple

Un altre exemple interessant sobre dades que descriuen relacions entre dades el proporciona la recent iniciativa Wikidata, que pretén descriure el contingut de Wikipedia d'acord amb les relacions que s'estableixen entre els diferents elements, incloent-hi els seus propis atributs. Així, si a Wikipedia hi ha una pàgina per al concepte «Oxigen», a Wikidata hi ha una entrada equivalent (encara que independent de l'idioma) que l'enllaça, per exemple, amb el concepte «Carl_Wilhelm_Scheele», mitjançant la relació «discoverer or inventor». Al seu torn, aquest està enllaçat amb el concepte «Stralsund», mitjançant la relació «place of birth» (lloc de naixement). Així, Wikidata configura una xarxa que permet enllaçar tot el coneixement de Wikipedia, de manera que sigui possible realitzar consultes amb una semàntica més avançada, per exemple, «inventors nascuts en una determinada ciutat».

Altres dades que per la seva naturalesa tenen unes particularitats pròpies són, per exemple, els mapes o diagrames. En aquest cas hi ha una relació de proximitat (mètrica) entre els elements que apareixen en el mapa (és a dir, la distància entre ells), que és la que determina la seva posició, i aquesta és molt important, a diferència de la representació utilitzada per als grafs, on el més important són les relacions entre elements, i no pas la seva posició. Normalment, en els mapes se superposen capes amb informació que es desitja relacionar amb la seva situació per a mostrar diferències entre zones, per exemple. En general, totes les dades procedents del món real que es desitgen representar en un mapa es gestionen mitjançant el que es coneix com a sistema d'informació geogràfica, tenint en compte la seva referència espacial, de manera que és possible fer cerques per proximitat o per criteris relatius a la naturalesa d'aquestes dades, per exemple, superposant una capa relativa a la vegetació del terreny amb una altra relativa a la seva explotació agrícola.



Font: http://sabia.tic.udc.es/gc/contenidos%20addicionals/treballs/3D/Internet%20GIS/sig.html

3.3. Dades semiestructurades o no estructurades

Finalment, pel que fa a la seva estructura, hi ha altres tipus de dades que es consideren o bé semiestructurades, bàsicament a causa de la complexitat de l'estructura subjacent, o bé no estructurades.

> Un exemple és aquest mateix document de text: està format per caràcters o cadenes de text, però aquests s'agrupen en línies, que al seu torn s'agrupen en paràgrafs, seccions, capítols, etc. No hi ha una estructura prefixada per als documents de text, sinó que depèn de la seva tipologia (llibre, article, etc.), per la qual cosa la seva estructura està parcialment definida. Un altre exemple serien les pàgines web escrites en llenguatge HTML, un llenguatge de marques que descriu la posició i el significat de cada contingut (textos, taules, imatges, enllaços, etc.) en una pàgina. Encara que a priori HTML és un llenguatge ben estructurat, la seva complexitat i la possibilitat de crear pàgines que no respectin exactament la sintaxi però segueixin sent vàlides fan que la seva anàlisi sigui complicada.

Les dades no estructurades solen ser generades per humans, mentre que les estructurades en molts casos són generades per màquines.

4. Cicle de vida de les dades

En la literatura tradicional de la gestió de dades, les dades són generades (creades) o capturades (extretes o oposades), emmagatzemades, preprocessades, analitzades, visualitzades i publicades, de manera que es tanca el cercle i se'n permet la reutilització. Cadascuna d'aquestes fases té un objectiu, i es genera valor a partir de les seves dades. No totes les fases són estrictament necessàries i, a més, poden solapar-se o realitzar-se simultàniament en alguns casos. D'aquesta manera, s'entén que les fases típiques del cicle de vida de les dades són les següents:

- Captura
- Emmagatzematge
- Preprocessat
- Anàlisi
- Visualització
- Publicació

4.1. Captura

La fase de captura té com a objectiu recopilar totes les dades que es generen durant un procés, mitjançant dos mecanismes bàsics complementaris:

- Creació: es tracta d'integrar en el propi procés de generació de les dades un mecanisme que emmagatzemi les que es considerin rellevants cada vegada que es generin.

  Exemple

  Per exemple, cada vegada que un usuari paga amb la seva targeta de crèdit (o dèbit) en un establiment, es genera un nou registre que defineix perfectament quina quantitat de diners s'ha gastat en quin establiment, en quina data i hora i amb quina targeta.

- Extracció: s'utilitza quan no és possible intervenir en el procés de generació de les dades, sinó que cal anar-les capturant segons es van trobant, de manera ideal immediatament després de la seva generació.

Exemple

Per exemple, és possible capturar els tuits que contenen una determinada paraula clau o *hashtag* tal com van apareixent en el flux (*timeline*) de tuits d'un usuari o el flux públic, ja que Twitter és un servei que genera dades en obert.

És a dir, o es té accés a les dades *a priori*, en el mateix moment de la seva creació, o *a posteriori*, una vegada han estat generades per un altre procés sobre el qual no es té possibilitat d'intervenció.

Un exemple pot ser el seguiment de la navegació dels usuaris d'un entorn virtual, per a analitzar quins serveis són els més requerits pels seus usuaris, quins són els camins o itineraris seguits, etc. L'opció que sembla més senzilla és analitzar el rastre que deixen els usuaris als *logs* dels servidors web que allotgen l'entorn virtual, encara que realment aquests *logs* contenen moltes altres dades relatives a la càrrega dels elements que conformen cada pàgina web, els quals no aporten cap informació rellevant per a analitzar el comportament dels usuaris que hi naveguen. L'extracció de la navegació dels usuaris d'un entorn virtual web és, en general, molt complexa i costosa. Per contra, si es té accés a l'entorn web, és possible introduir marques (codi que genera les dades desitjades) en aquells serveis dels quals realment es vol obtenir informació, recollint (generant) solament aquelles dades que formaran part de l'anàlisi posterior.

Desafortunadament, no sempre es té accés al nivell necessari per poder recollir les dades en el moment exacte de la seva creació, per la qual cosa les estratègies de captura de dades basades en l'extracció de dades ja existents o la seva generació mitjançant mecanismes alternatius són les més habituals, entre altres:

a) Accés a les dades mitjançant un repositori: com a resultat de la seva publicació en obert, les dades poden estar disponibles en un repositori digital, o simplement en una web que en permeti l'accés. Normalment les dades es troben ja en fitxers d'acord amb una classificació preestablerta, en funció de la seva naturalesa o del domini al qual pertanyen. Aquestes dades poden veure's com una fotografia d'un procés o escenari en un moment donat, per la qual cosa poden no haver estat actualitzades recentment. Per exemple, l'Institut Nacional d'Estadística (INE) publica dades demogràfiques, com la població de cada municipi, amb una periodicitat anual. Es tracta, per tant, de dades **estàtiques**, a les quals s'accedeix com un conjunt sencer, sense poder especificar quines dades es desitgen descarregar. No obstant això, cada vegada és més comú que els proveïdors de dades incorporin mecanismes per a facilitar l'accés a dades específiques, encara que no és una pràctica habitual. La majoria dels repositoris de dades en obert només ofereixen l'opció de descarregar fitxers amb les dades i no inclouen cap opció per seleccionar-les a priori.

Exemples de repositoris amb dades en obert

Catàleg de dades a datos.gob.es, reutilitza la informació pública: http://datos.gob.es/catalogo

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

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

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

b) Accés mitjançant una API (*application programming interface*): en alguns casos, sí que és possible utilitzar un mecanisme que permet realitzar consultes específiques contra un conjunt de dades, obtenint només aquelles que han estat requerides d'acord amb els paràmetres de la consulta. Es pot parlar, aleshores, de dades **dinàmiques**, ja que aquestes es generen d'acord amb la consulta realitzada. Encara que el concepte d'API va ser inicialment pensat per a accedir a serveis i llibreries de programari, avui dia també n'està estès l'ús per a l'accés i l'intercanvi de dades, no només per a l'execució de serveis. Actualment, la majoria dels serveis com Twitter, Flickr i altres xarxes socials hi permeten accedir mitjançant una API, per la qual cosa resulta possible desenvolupar eines i aplicacions que extreguin dades d'aquestes xarxes socials de manera automatitzada, gairebé sense intervenció humana excepte per a iniciar la consulta. En alguns casos, les API estan integrades en el que es coneix com a *sandbox*, un entorn controlat que permet als desenvolupadors efectuar proves i observar el funcionament de l'API. Un aspecte important que cal tenir en compte és l'existència de mecanismes de *throttling*, que limiten el nombre i la freqüència de les peticions que es realitzen a una API per evitar que es col·lapsi si s'hi accedeix de manera desenfrenada, o per limitar-ne l'ús fins a un cert nombre de peticions diàries (o mensuals) en un servei gratuït, mentre que és possible superar aquest límit si es paga pel servei d'accés a l'API. És el que es coneix com a model *freemium*, un concepte típic de serveis en línia com Spotify, per exemple.

Exemples d'API

Els exemples d'API disponibles són:

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

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

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

Llistat d'API: http://www.programmableweb.com/apis/directory

c) Manipulació dels paràmetres de cerca: la majoria de les vegades, un repositori digital o una web no disposen d'una API per a la captura de les dades desitjades, sinó que cal dur a terme operacions de cerca mitjançant la interfície d'usuari disponible per a accedir-hi, seleccionant aquells conjunts de dades que resulten d'interès. En alguns casos, la pròpia navegació indica l'arquitectura de la web, mitjançant els paràmetres que apareixen a la URL i que van canviant d'acord amb la navegació de l'usuari. Llavors és possible manipular aquests paràmetres i programar uns *scripts* senzills que automatitzin aquest procés, accedint a la mateixa URL una vegada i una altra però modificant els paràmetres de cerca, amb la qual cosa s'obtenen diferents resultats a cada crida.

Exemple de manipulació dels paràmetres de cerca i contingut

http://www.idescat.cat/nadons/?sexe=<SEXE>&res=<LLOC>&t=<ANY>&lang=es

Es poden manipular el paràmetre <SEXE> ("1" nens, "2" nenes, "0" tots, sense les cometes), el paràmetre <LLOC> (per exemple, "a" per a tot Catalunya, "d13" per a la comarca del Barcelonès) i el paràmetre <ANY> (per exemple, "2014"), per obtenir els noms posats als nadons d'aquest sexe nascuts en aquest any en aquesta regió (Dades IDESCAT). Per exemple:

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

d) Captura de dades mitjançant *scraping*: una situació encara pitjor és quan ni tan sols és possible manipular manualment les URL per a accedir a les dades de manera semiautomàtica. En aquest cas l'única opció possible és utilitzar eines (també anomenades *bots*) que simulin la navegació d'un usuari per un conjunt de pàgines web i que extreguin el contingut de les pàgines visitades, analitzant-ne l'estructura. Això és el que es coneix com a *web scraping*. Actualment les pàgines web són creades de manera automatitzada, per la qual cosa la seva estructura interna és bastant estable i coherent. Utilitzant eines bàsiques (com, per exemple, la possibilitat d'inspeccionar la pàgina web des del propi navegador), és possible determinar aquesta estructura interna del document HTML i programar un bot per accedir solament a la informació desitjada.

Exemples d'eines per a fer *web scraping*

Alguns exemples d'eines per fer web scraping són:

Scrapy: http://scrapy.org/

PhantomJS: http://phantomjs.org/

Altres eines: http://www.garethjames.net/a-guide-to-web-scrapping-tools/

e) Extracció de dades de documents de text: en ocasions es publiquen dades en formats no pensats per a la seva reutilització, en PDF per exemple, en forma de taules, llistes, etc. En aquest cas cal extreure aquesta informació usant opcions del tipus tallar i enganxar, però això no sempre és possible i se sol perdre informació relativa al format de taula. Afortunadament hi ha altres eines que permeten extreure taules directament, generant dades en format tabular llestes per ser reutilitzades. Quan les dades es publiquen com a imatges escanejades, cal recórrer a eines per al reconeixement òptic de caràcters (en anglès, OCR), encara que aquest procés no està exempt d'errors en funció de la qualitat de la imatge.

Exemples d'eines per a l'extracció de text

Eines per a l'extracció de text:

Extracció de dades en taules de PDF amb Tabula: http://tabula.technology/

Reconeixement de caràcters amb Tesseract: https://github.com/tesseract-ocr

f) Formularis: de vegades el més senzill i eficaç és preguntar directament als usuaris d'un servei, recurs o sistema, amb l'objectiu de recaptar dades, tant del servei en qüestió com dels propis usuaris. Encara que se segueixen fent moltes enquestes a peu de carrer, es tracta d'un procés lent i costós, així com limitat en el temps i en l'espai. Actualment hi ha moltes eines gratuïtes per a la creació i execució de formularis en línia, de manera que és possible accedir a una gran quantitat d'usuaris a partir de les seves adreces de correu o mitjançant una xarxa social.

Exemples per a la creació de formularis en línia

Alguns exemples per a la creació de formularis en línia:

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

LimeSurvey: https://www.limesurvey.org

g) *Crowdsourcing*: es tracta d'aprofitar la «saviesa de les masses» (*wisdom of crowds*), que són capaces de resoldre problemes que resulten molt complicats fins i tot per a l'estat de l'art en temes d'intel·ligència artificial i mineria de dades, i aprofitar l'intel·lecte humà.

Exemple

Per exemple, la figura 1 mostra un sistema de control anti- *bots* basat en *captcha*, que evita que una màquina es registri automàticament en un servei en línia, per exemple. El sistema obliga l'usuari a introduir dues paraules, una d'elles distorsionada i que solament un humà és capaç de llegir, en aquest cas «admipa». L'altra paraula («exposure») es tracta en realitat d'un cas de *crowdsourcing*, és una paraula provinent d'un sistema òptic de reconeixement de caràcters que no ha pogut ser processada correctament, per la seva baixa resolució, per exemple. Quan uns quants usuaris accedeixen al servei en línia i introdueixen «exposure», el sistema aprèn que aquest és el text associat a aquesta imatge que no s'havia pogut reconèixer. Encara que alguns usuaris s'equivoquin o menteixin pel que fa a aquesta paraula, la majoria d'ells segurament la introduirà correctament, per la qual cosa es pot donar per vàlida quan un cert nombre d'usuaris coincideix.

Figura 1. Exemple de *crowdsourcing* fent servir un *captcha*

h) Dades qualitatives: de vegades cal obtenir dades directament dels usuaris perquè interessa conèixer de primera mà alguns aspectes que no poden recollir-se mitjançant una enquesta, especialment el «per què?» i el «com?», i aspectes difícilment quantificables, com emocions o sentiments. Normalment es recorre a entrevistes semiestructurades, on mitjançant una sèrie de preguntes s'obté informació directa dels participants. Aquestes entrevistes solen ser gravades i després transcrites i codificades, d'acord amb uns criteris establerts amb anterioritat, de manera que s'identifiquen aquells aspectes que es volien conèixer mitjançant l'entrevista.

Exemple d'eines per a la captura de dades qualitatives

Eines per a la captura de dades qualitatives:
RQDA: http://rqda.r-forge.r-project.org/

4.2. Emmagatzematge

Les dades capturades són emmagatzemades en un format que en permeti la posterior manipulació, conforme a la representació més adequada, tenint en compte tant la seva tipologia com l'ús que se'n voldrà efectuar. Sense entrar en els detalls tècnics relatius a la infraestructura tecnològica subjacent, les dades s'emmagatzemen o bé en **fitxers simples** (o col·leccions de fitxers) o bé en **bases de dades**. En aquest cas, s'intenta reproduir de certa manera l'estructura de relacions entre ells. En funció del seu objectiu i complexitat es pot parlar de:

1) Fitxers simples: les dades són emmagatzemades en fitxers (o col·leccions de fitxers) d'acord amb un o més criteris, com per exemple l'origen de les dades i/o la data de creació o captura.

> Un exemple d'això poden ser els fitxers de log generats pels servidors web, que contenen totes les peticions que es fan quan els usuaris naveguen per les pàgines web d'un servei en línia. Normalment es tracta de fitxers en formats plans que permeten una manipulació relativament senzilla, encara que també pot donar-se el cas contrari, com una col·lecció de documents de text o d'imatges, d'acord amb algun format concret.

2) Bases de dades: es tracta d'una estructura més o menys complexa que permet representar les dades d'acord amb la seva naturalesa, tenint també en compte les relacions entre tots els elements que les componen. Se sol parlar de bases de dades relacionals, però recentment s'han popularitzat les anomenades dades no relacionals, que pretenen donar solució a problemes relacionats amb l'escalabilitat de les relacionals, especialment per a grans volums de dades.

En aquest segon cas, és possible construir sistemes que optimitzen certes operacions:

- Magatzems de dades: de l'anglès *data warehouse*, són un tipus de bases de dades que està orientat a emmagatzemar dades amb l'objectiu d'optimitzar les consultes i generar informes agregant i resumint dades de diferents fonts. Solen alimentar-se de les diferents bases de dades utilitzades en una organització i permeten un accés centralitzat amb l'objectiu de poder fer consultes més complexes i eficients.

- *Datamarts*: en aquest cas es tracta d'un subconjunt d'un magatzem de dades que té el propòsit de donar suport a una àrea específica, amb uns objectius concrets. Es pot veure com una capa d'accés sobre un magatzem de dades que filtra i selecciona aquelles dades que es desitja analitzar des d'una perspectiva concreta, la qual cosa simplifica la presa de decisions.

Un aspecte molt important per a facilitar la manipulació de les dades emmagatzemades és que aquestes segueixin algun format de fitxer obert, de manera que no calgui cap programari específic per a la seva manipulació ni cap llicència i que es pugui assegurar la seva preservació a llarg termini, especialment en el cas de fitxers i no tant en el cas de bases de dades on grans noms com IBM o Oracle es disputen el segment. També hi ha solucions basades en codi obert com MySQL o PostgreSQL.

D'altra banda, és important que les dades siguin realment accessibles i que el format de fitxer no només sigui obert pel que fa a la disponibilitat de programari lliure i a l'absència de llicències, sinó també perquè és possible accedir i manipular les dades que hi té emmagatzemades. Les mateixes dades poden emmagatzemar-se de moltes maneres, i no totes en promouen la reutilització posterior. És el que es coneix com a esquema de 5 estrelles, creat per Tim Berners-Lee i que es mostra a la figura 2.

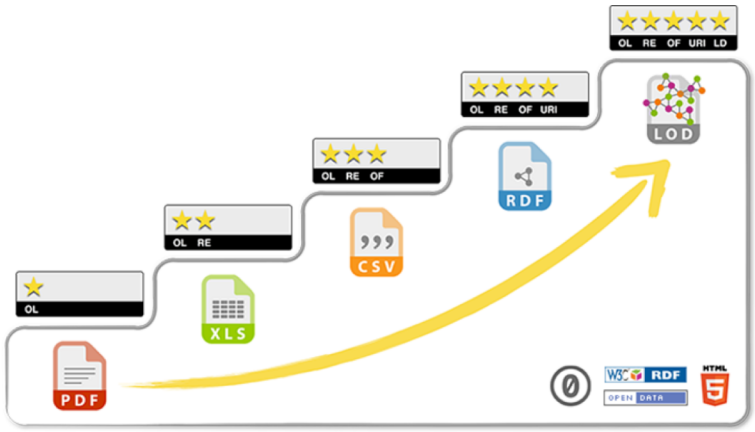


Figura 2. Esquema de 5 estrelles per a la publicació de dades.
Font: http://5stardata.info/

El nivell més baix (una estrella) correspon a dades publicades en formats que no en permeten l'extracció directa (excepte mitjançant eines com l'esmentada Tabula). El segon nivell sí que permet una manipulació de la dada però mitjançant programari que requereix una llicència, com per exemple Microsoft Excel. El tercer nivell, el més utilitzat en l'actualitat, fa servir formats no propietaris que també permeten la manipulació de la dada. Els nivells quart i cinquè tenen com a objectiu aconseguir el que es coneix com a *linked data*, és a dir, dades que estan descrites de manera que se'n pot traçar l'origen (les dades estan descrites mitjançant URL, per la qual cosa és possible apuntar-hi). A més estan enllaçades amb altres dades de tercers, de manera que aleshores es crea una xarxa o graf complex que permet fer consultes més elaborades.

En el tercer nivell, quan es treballa amb **dades no jeràrquiques**, és a dir, quan tots els camps d'un registre es troben al mateix nivell, se sol utilitzar un format pla com el CSV, o valors separats per comes, un fitxer de text on cada registre ocupa una línia i els camps de cada registre van separats per comes «,» (o de vegades, punt i coma «;»). Aquest format s'ha popularitzat per la seva simplicitat i per l'existència de moltes eines que el suporten. A més, és també possible manipular-lo amb un simple editor de textos i/o petits *scripts* o comandaments del sistema operatiu.

Per contra, quan es treballa amb **dades jeràrquiques**, se sol optar per formats que permeten representar aquesta jerarquia, així com la presència o no d'elements opcionals, incloent-hi aleshores la possibilitat d'aconseguir els nivells 4 i 5 de l'esquema de 5 estrelles. En aquest cas, se solen utilitzar els diferents formats:

JSON: acrònim de *Javascript object notation*, es tracta d'un format lleuger per a l'intercanvi de dades, basat en l'ús de parelles atribut-valor (amb el format "<atribut>": <valor>); és independent del llenguatge de programació, i existeixen multitud de llibreries per a la manipulació de dades en aquest format.

Per exemple, un sensor capturant dades de temperatura podria produir dades en format JSON de la següent manera, indicant també la seva posició en aquest moment (usant 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: acrònim d'*extensible markup language*, es tracta d'un metallenguatge extensible mitjançant etiquetes, de manera que és possible definir dades mitjançant l'especificació d'un conjunt d'etiquetes que les descriuen i l'estructura jeràrquica que componen. Està pensat també per a l'intercanvi de dades de manera independent del llenguatge de programació utilitzat, i hi ha una gran

col·lecció d'adaptacions per a tipus de dades o àmbits específics, com per exemple MathML, un llenguatge per a la descripció de contingut matemàtic, o XSIL, orientat a l'intercanvi de dades científiques.

L'exemple anterior en XML seria:

```
<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: acrònim de resource description framework, es tracta també d'un metallenguatge que permet descriure recursos d'acord amb un marc preestablert, mitjançant l'ús de tripletes subjecte-predicat-objecte.

Una possible representació parcial de l'exemple anterior en RDF (sense especificar els espais de noms) podria ser la següent:

"TS123456_1454070512" taken_by 123456

De fet, la representació real en RDF és molt semblant a la d'XML, però inclou informació relativa a l'origen de les dades i les descriu mitjançant metadades, de manera que sigui més senzill processar-les de manera automàtica, no manual.

4.3. Preprocessat

L'objectiu d'aquesta etapa és preparar les dades per a la seva anàlisi posterior, de manera que puguin ser usades directament per qualsevol investigador o data scientist, sense que s'hagi de preocupar per aspectes relacionats amb la seva qualitat, procedència, etc.

Entre altres operacions típiques d'aquesta etapa, es poden destacar les següents:

1) Fusió: de vegades les dades s'obtenen de diferents fonts, per la qual cosa cal combinar-les en una única estructura, normalment una taula.

Per exemple, pot ser que, d'un conjunt d'usuaris d'un servei o xarxa social, se'n coneguin, d'una banda, els hàbits de navegació, i d'una altra, els hàbits de compra. Si cada usuari té un identificador que (valgui la redundància) l'identifica de manera unívoca en tots dos conjunts de dades, és possible utilitzar aquest identificador com a clau per a la fusió dels dos conjunts, combinant registres de dos fitxers o taules diferents en un de sol per a cada element.

2) Selecció/filtrat: pensant en les dades com si fossin un hipercub (que combina dades de diferents fonts ja fusionades), aquest procés consisteix a realitzar talls sobre aquest hipercub, tot seleccionant aquelles regions que contenen les dades d'interès, d'acord amb un o més criteris de cerca. El cas més senzill correspon a la selecció de dades d'una taula d'acord amb un criteri.

> Per exemple, seleccionar solament les dades de ciutadans homes menors de 18 anys d'un municipi en concret.

3) Conversió: de vegades les dades estan en formats que en dificulten l'anàlisi posterior, per la qual cosa cal convertir-les a un format pla més proper a la idea de taula.

> Per exemple, les dades de Twitter capturades en format JSON contenen una gran quantitat d'informació relativa a l'estructura dels tuits que pot ser innecessària per a l'anàlisi, així que es poden seleccionar només els atributs que es desitgin (per exemple, el text del tuit) i bolcar-los en un format pla, per exemple CSV.

4) Neteja: també coneguda com a *data cleansing*, consisteix a eliminar totes les inconsistències en les dades que puguin ser detectades, ja sigui esborrant-les o marcant-les per a una inspecció manual posterior, més detallada, que persegueixi la «validesa» de les dades, de manera que siguin realment útils.

> Per exemple, si es tracta d'usuaris d'una xarxa social i l'edat indicada en el perfil de l'usuari és de més de 100 anys, podem pensar que s'ha comès un error en introduir-la i marcar aquest registre com a possible candidat a ser esborrat o corregit manualment. O si s'especifica el codi postal, podem comprovar que aquest és vàlid i coherent amb la població especificada, per exemple. Aquest procés també inclou la resolució d'incoherències com les que poden donar-se quan es fusionen dades de diferents fonts, tot detectant per exemple si un mateix client apareix amb dades diferents, amb la qual cosa cal determinar quina és la informació correcta.

5) Agregació: en alguns casos, pot resultar interessant resumir un subconjunt de dades en un de sol, de manera que se simplifiqui el conjunt de dades original i es generi una nova variable que tingui un major poder predictiu.

> Per exemple, si per a cada usuari es disposa de totes les seves connexions diàries al llarg d'un període de temps, es podrien agregar les dades setmanalment, amb la qual cosa es redueix el volum de dades dràsticament.

6) Creació de noves variables/indicadors: de vegades cal fer càlculs sobre les variables o camps disponibles per a, per exemple, convertir unitats o calcular la ràtio entre dues variables, tot generant noves variables o indicadors. Aquest procés està relacionat amb l'extracció de característiques i la reducció de la dimensionalitat, que solen veure's ja com a operacions que formen part de l'etapa d'anàlisi.

Aquestes operacions poden realitzar-se de moltes maneres diferents, en funció de la naturalesa de les dades i de les eines disponibles, normalment combinant més d'una eina. Les opcions més habituals són aquestes:

- Mitjançant una base de dades: les dades de diferents fonts es carreguen en una base de dades (generalment relacional) que en respecta l'estructura i

en facilita la manipulació posterior mitjançant sentències SQL, que inclouen, entre altres, SELECT i JOIN, per exemple.

- Mitjançant eines específiques: per a la manipulació de dades amb formats diversos hi ha eines com OpenRefine, que permet realitzar les operacions esmentades anteriorment i exportar el resultat a formats més aptes per a la seva anàlisi posterior.

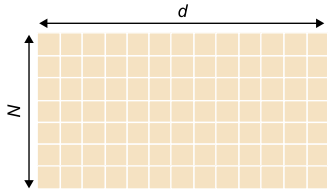
- Mitjançant llenguatges de programació: finalment, utilitzant llenguatges com Perl, i especialment avui dia Python, és possible manipular dades en múltiples formats d'entrada (JSON, XML, etc.) i realitzar-hi operacions de manera senzilla; hi ha fins i tot llibreries i entorns que simplifiquen aquestes operacions.

És important destacar que aquesta etapa resulta molt important i possibilita les que vénen a continuació, ja que la qualitat dels resultats obtinguts (i dels models construïts) dependrà, directament, de la qualitat de les dades sobre les quals hagin estat fonamentats. És el que es coneix com a *garbage in*, *garbage out*, és a dir, si s'usen dades «escombraries» per a crear models o visualitzacions, aquestes també seran, amb tota seguretat, «escombraries». A més, assegura l'obtenció d'un fitxer amb dades de qualitat que poden preservar-se al llarg del temps, per a futures reutilitzacions.

4.4. Anàlisi

Una vegada les dades ja es consideren vàlides, es pot procedir a la seva anàlisi.

> L'objectiu d'aquesta etapa és crear un o més models que expliquin com són les dades i les seves característiques principals i poder respondre a les preguntes plantejades en el marc del projecte on s'estiguin utilitzant aquestes dades.



En aquest punt ja es pensa en les dades com en una taula 2D que conté una fila per a cada element descrit (també anomenat observació) del qual es tenen dades, en forma d'atributs (també anomenats variables), un a cada columna. En general, se sol parlar de N dades de d dimensions, tot i que les dimensions puguin ser al seu torn estructures de dades complexes. Aquests dos paràmetres (N i d) condicionen, en part, l'anàlisi que es pot fer, juntament amb la naturalesa i tipologia de les dades.

Si N és molt petit (unes poques desenes de dades), resultarà impossible construir models robusts, ja que no es podran generalitzar els resultats obtinguts a noves dades de què es pugui disposar en un futur. En particular, amb un valor de N petit, segurament serà impossible observar el fenomen desitjat, i si s'observa també serà impossible assegurar que sigui una característica del

conjunt de dades, és a dir, per a valors de *N* petits és pràcticament impossible inferir-hi res, excepte descriure exactament el conjunt de dades de què es disposa. La mida de *N* determina aspectes com els intervals o nivells de confiança que es poden estimar a partir d'un conjunt de dades o mostra extreta d'una població.

Per contra, si *N* és molt gran, caldrà dur a terme tècniques de mostreig per a reduir el nombre de dades que s'utilitzen per a construir un model, especialment si aquest té una complexitat quadràtica o superior, tant pel que fa a la complexitat espacial com la temporal. Un algorisme amb complexitat temporal quadràtica que veu incrementada la quantitat de dades en un ordre de magnitud (és a dir, *N* x *10*) necessitarà 100 vegades més temps que en el cas del conjunt de dades original amb *N* elements.

D'altra banda, el valor de *d* va associat a un problema conegut, anomenat maledicció de la dimensionalitat, que apareix quan *d* és realment gran, de l'ordre de centenars o milers. Si es pensa en les dades com en un hipercub de *d* dimensions, quan *d* creix també ho fa el volum d'aquest hipercub, de manera que, per a un valor de *N* fix, si d creix, aleshores les dades estan més disperses dins d'aquest hipercub.

> Per exemple, un conjunt de dades descrit mitjançant 10 variables binàries pot prendre 2^10 ≈ 10^3 valors diferents, per la qual cosa si *N* no és de l'ordre de 1.000 elements, es podria donar el cas que hi hagués un sol cas a cada «cantonada» de l'hipercub, de manera que resultaria molt difícil l'extracció de característiques comunes o l'agrupació per similitud.

Finalment, la relació entre *N* i *d* també és important. En funció del tipus d'anàlisi que es vulgui fer, se sol recomanar que *N* sigui, almenys, d'un ordre de magnitud major que *d*, és a dir, que $N > 10 \times d$, però se sol recomanar que *N* sigui, almenys, major que $30 \times d$. Aquesta restricció acostuma a causar problemes, per exemple, en l'anàlisi de dades provinents d'una enquesta, on el nombre de respondents (*N*) pot ser insuficient si el nombre de preguntes o ítems de l'enquesta (*d*) és molt elevat. D'això la necessitat de desenvolupar tècniques per a la reducció de la dimensionalitat de les dades, tractant de capturar tota la seva riquesa amb un nombre menor de dimensions *d'*, i millorar així la ràtio N / d'.

Així doncs, una vegada que es disposa d'un conjunt de dades, es pot procedir amb l'anàlisi, i en funció de la seva naturalesa i dels objectius que es persegueixin, aquesta pot ser:

a) Anàlisi estadística descriptiva: es tracta de descriure les dades mitjançant un conjunt reduït de valors que permetin modelar-les, d'acord amb alguna distribució coneguda, amb la finalitat de descriure apropiadament les característiques essencials d'aquest conjunt, sense tenir en compte si formen o no part d'un conjunt de dades major.

Per exemple, és habitual que en l'anàlisi dels resultats d'una enquesta s'inclogui un apartat que descrigui el sexe dels participants, la seva edat, estat civil i situació laboral, entre altres, indicant el percentatge de cada valor possible per a cada variable.

b) Anàlisi estadística inferencial: en aquest cas, es tracta de modelar les dades d'acord amb una distribució desconeguda, tenint en compte que, segurament, només es disposa d'una fracció de la totalitat de les dades que conformen tota la població (en el sentit estadístic del terme). L'objectiu és inferir com és la totalitat de la població, assumint un grau d'error en les estimacions causat per la no disponibilitat de totes les dades, sinó només d'una fracció.

Així, per exemple, en el cas d'una enquesta se sol especificar el nombre de participants i, a partir d'aquest, calcular els intervals de confiança per als resultats obtinguts.

c) Extracció de característiques: a partir de tots els atributs disponibles, l'objectiu és crear-ne de nous (anomenats característiques o *features*) que capturin la naturalesa de les dades originals però alhora representant millor aquells aspectes que es desitgen analitzar.

Per exemple, d'un grup de persones es podria disposar de diferents mesures, com el pes (w) i l'altura (h), entre altres. Si l'objectiu és decidir si una persona està excessivament prima o obesa, en lloc d'utilitzar el seu pes i altura es podria fer servir l'índex de massa corporal, que està més lligat al concepte de pes raonable per a una altura.

Índex de massa corporal

$$\mathrm{IMC} = \frac{w}{h^2} \qquad 1.1$$

w (en kg)

h^2 (en m)

Això és el que es coneix com l'extracció de característiques dependents de l'àmbit d'aplicació. D'altra banda, hi ha tècniques generals que permeten extreure les característiques a partir de l'anàlisi, generalment, de la variància de les dades disponibles. Entre les més conegudes, es poden destacar l'anàlisi de components principals (PCA) o el recent (almenys pel que fa al terme utilitzat) *deep learning*, basat en la combinació de diferents xarxes neurals per a l'extracció de coneixement de les dades.

d) Reducció de la dimensionalitat: de manera anàloga a l'extracció de característiques, l'objectiu és reduir el nombre de variables (d) perquè la ràtio N / d sigui millor, i permetre una representació de les dades usant recursos gràfics 2D o 3D, amb l'objectiu de detectar agrupacions o dissimilituds. A més del ja esmentat PCA, hi ha altres algorismes com l'escalat multidimensional (en anglès MDS), per exemple, que permeten visualitzar en 2D o 3D dades de dimensionalitat elevada.

i) Models supervisats: en aquest cas, es disposa de dades que inclouen una o més variables que es consideren l'objectiu del problema que hem de resoldre. Per tant, es construeixen models que intenten predir aquesta variable (també anomenada depenent) en funció de les altres (anomenades independents).

Per exemple, utilitzant l'experiència acumulada durant anys, les companyies d'assegurances ajusten les pòlisses (la quantitat a pagar cada mes) en funció de diferents paràmetres, com l'edat del conductor, la seva experiència, el model de cotxe, el seu registre previ de parts i incidències, etc., de manera que per a un nou client potencial o per a la renovació d'un d'existent sigui possible determinar el valor òptim (segurament, des de l'òptica de la companyia d'assegurances) de la seva pòlissa, ajustant-la al màxim perquè resulti atractiva per al client però no arriscada per a la companyia.

Hi ha una llarga col·lecció de mètodes i algorismes per a l'aprenentatge supervisat; entre altres, es poden destacar els arbres de decisió, les xarxes neuronals, el classificador bayesià ingenu o els models lineals generalitzats.

f) Models no supervisats: quan no es disposa d'una variable objectiu, l'anàlisi de les dades s'enfoca a comparar-les entre si i a trobar-hi similituds i diferències, amb l'objectiu de detectar qualsevol estructura interna que permeti crear agrupacions en funció d'algun criteri. Entren en aquesta categoria d'aprenentatge no supervisat algorismes com els de *clustering*, cert tipus de xarxes neuronals, els mapes autoorganitzatius i l'algorisme EM (*expectation-maximization*), entre altres. La creació de models no supervisats pot utilitzar-se també per a l'extracció de noves característiques que poden ser usades per altres algorismes, especialment supervisats.

g) Visualització: encara que aquesta metodologia mereix un apartat per si sola, visualitzar les dades disponibles en una primera opció és una estratègia d'anàlisi molt recomanable, atès que per als humans és molt fàcil detectar patrons i irregularitats en les dades o en els seus descriptors estadístics bàsics.

> Per exemple, és possible fer una anàlisi descriptiva utilitzant visualitzacions basades en l'histograma de les dades o detectar relacions entre variables mitjançant l'ús de diagrames de dispersió.

L'anàlisi no es limita a la construcció de models, sinó que també ha d'explicar el resultat obtingut mitjançant una interpretació del model i la seva posada en context pel que fa al problema original. Això inclou també l'avaluació del propi model, identificant quines variables o característiques són les més rellevants, la capacitat de generalització davant dades mai utilitzades prèviament en la creació del model o la seva capacitat d'adaptació als canvis en les dades. És en aquest punt on un bon data scientist serà capaç d'establir un equilibri entre precisió, complexitat i actualització, triant el model o la combinació de models més adequats en funció dels objectius perseguits i la naturalesa de les dades disponibles.

4.5. Visualització

Els humans disposen d'un sistema visual molt complex i avançat que inclou des de l'ull amb tots els seus elements (pupil·la, còrnia, retina, etc.) fins al còrtex visual, encarregat de processar tota la informació recollida per l'ull. Aquest sistema és capaç de capturar i processar fins a 10 Mb/s, i de fet, se sap que el 93% de la informació processada pel cervell és considerada no verbal. Els humans són sobretot màquines de processament visual, amb diversos subsistemes que s'encarreguen de processar eficientment diferents aspectes d'aquesta tipologia d'informació de manera separada: forma, moviment, color, etc.

Per tant, qualsevol aproximació basada en la transmissió d'informació visual és molt eficient.

Per exemple, es considera que el processament d'informació escrita aconsegueix un màxim d'unes 1.000 paraules per minut per a persones molt capacitades, i la mitjana és d'unes 200-250 paraules per minut. El procés de convertir imatges en text i interpretar-lo és molt més lent que el procés de la imatge equivalent. La dita «una imatge val més que mil paraules» no deixa de ser raonablement certa en aquest sentit. Les imatges tenen una capacitat de captar l'atenció molt important, i pot ser explotada per comunicar fets (basats en dades) de manera més efectiva que mostrant les dades originals, per exemple.

Un altre aspecte interessant de la visualització de dades és que pot convertir-se en la pròpia interfície de navegació de les pròpies dades, permetent certes operacions bàsiques (selecció, agregació, etc.), de manera que sigui possible afegir certa interactivitat a la visualització. Seguint el mantra de Ben Shneiderman[3], cal presentar les dades resumides, oferir les opcions de zoom i filtre (selecció) i, finalment, afegir els detalls necessaris si l'usuari els sol·licita. Així és possible combinar una gran quantitat d'informació en una mateixa visualització sense sobrecarregar l'usuari, deixant-li l'opció de determinar la quantitat de dades que vol visualitzar. Concretament (fent servir els termes originals descrits per Shneiderman):

[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 (pàg. 336-343). IEEE.

- *Overview*: observar patrons en les dades.
- *Zoom*: seleccionar un subconjunt de les dades.
- *Filter*: seleccionar d'acord amb un criteri o valor.
- *Detail on demand*: obtenir valors per a les dades seleccionades.
- *Relate*: comparar valors.
- *History*: mantenir un registre de les accions realitzades.
- *Extract*: marcar i extreure (capturar) les dades seleccionades.

Amb aquest esquema, és possible basar l'anàlisi de les dades a partir de la seva visualització, de manera que es combinin la capacitat visual humana pel que fa a la detecció de patrons, tendències, etc., amb la potència d'un sistema informàtic que permeti seleccionar, filtrar o comparar dades. Un bon data scientist sempre comença visualitzant les dades per a comprendre més bé la seva naturalesa i per detectar possibles relacions, dependències i característiques que les defineixen, individualment i grupalment.

4.6. Publicació

Com a etapa final que tanca el cercle, el resultat de les tres fases anteriors pot ser publicat en forma de noves dades, de manera que sigui possible que tercers les reutilitzin amb altres propòsits, especialment les dades ja preprocessades i preparades per a analitzar, en forma d'una o més taules.

Encara que avui dia hi ha diversos mecanismes per a compartir dades, és recomanable utilitzar espais optimitzats per a això, més enllà de la seva simple publicació en una web. Concretament, resulta molt avantatjós compartir les dades a través d'un repositori digital, de manera que s'assegurin dos objectius complementaris: la seva preservació i la seva disseminació. Per **preservació** s'entén el procés que assegura la disponibilitat de les dades en un llarg període de temps, tenint en compte (i fins i tot intervenint en) aspectes com el format de les dades, la seva estructura interna, etc., de manera que sigui possible accedir-hi anys després, encara que ja no existeixin les eines que es van utilitzar originalment per crear-les i emmagatzemar-les. D'altra banda, la **disseminació** té com a objectiu assegurar la reutilització d'aquestes dades, de manera que sigui possible trobar-les, accedir-hi i entendre'n l'estructura i el format. Per a això, cal disposar de les metadades adequades que descriguin les dades, a més d'indicar de clarament quines condicions d'ús delimiten el que se'n pot fer. Si no és així, en la pràctica resulta impossible reutilitzar aquestes dades si se'n desconeix l'origen i naturalesa.

Idealment, el repositori hauria d'incorporar una API perquè sigui factible desenvolupar serveis i aplicacions que puguin accedir a les dades que es volen obtenir mitjançant una consulta ben especificada, sense intervenció humana. També és interessant que el repositori permeti triar en quin format es retornaran les dades, si en una taula plana (per exemple, usant el format CSV) o mitjançant una estructura jeràrquica (en aquest cas els formats JSON i XML són els més habituals).

Actualment hi ha diverses iniciatives que promouen la publicació de dades en repositoris, des de perspectives diferents. D'una banda, hi ha un moviment impulsat per administracions i organitzacions públiques que pretenen oferir en obert la majoria de les dades que es generen en el seu si, la major part d'elles amb origen en els propis ciutadans o contribuents, intentant complir amb criteris de transparència i participació ciutadana. D'altra banda, també cada vegada més empreses decideixen compartir les seves dades intentant fer emergir noves aplicacions que les explotin i permetin extreure'n valor afegit. En alguns casos, això es promou i es desenvolupa mitjançant una *hackathon* (combinació de *hacking* i *marathon*), un esdeveniment on programadors i interessats en el tema poden posar en pràctica les seves idees. Quan aquests esdeveniments estan orientats a la manipulació i explotació de dades, acostumen a rebre el nom de *datathon*.

Finalment, un repositori diferent que mereix una atenció especial és l'UC Irvine Machine Learning Repository (o senzillament, UCI ML), que té com a objectiu la compartició de conjunts de dades usades per a la comparació i afinació d'algorismes de classificació, regressió i predicció. Conté més de 300 *datasets* de diferents àmbits de coneixement, incloent-hi els més característics i coneguts de la literatura en mineria de dades, com el *dataset* Iris de R.A. Fisher, per exemple, presentat en un treball de 1936 per a mostrar les possibilitats de

l'anàlisi discriminant lineal. Qualsevol avenç en l'àrea de *machine learning* (un nou algorisme o l'afinació dels paràmetres d'un algorisme conegut) passa per l'avaluació mitjançant conjunts disponibles en aquest repositori.

Resum

En aquest mòdul s'ha introduït el concepte de *data science*, utilitzant el cicle de vida de les dades per a mostrar totes les competències necessàries que inclou i requereix aquest àmbit de coneixement. Es tracta d'una àrea multidisciplinària que combina coneixements científics, principalment matemàtics i estadístics, amb altres de provinents d'un context propi de l'enginyeria, on mitjançant l'ús i la combinació de diferents eines es resolen problemes com els que tenen a veure amb la captura i el preprocessat de les dades. Quan això es combina amb el coneixement de l'àrea d'aplicació, es disposa del potencial adequat per a extreure coneixement de les dades disponibles, obtenint un valor afegit que pot ser usat per avançar a la competència i/o millorar processos i/o serveis, així com per entendre més bé la naturalesa de l'àrea d'aplicació en qüestió. La figura 3 resumeix aquesta triple combinació, situant *data science* al centre de la intersecció dels tres àmbits esmentats.

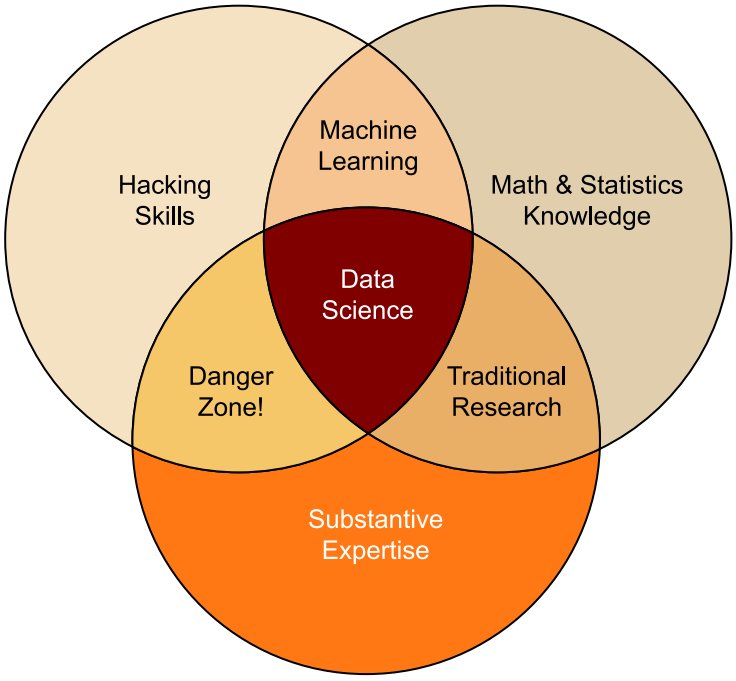


Figura 3. *Data science* com a disciplina emergent.
Font: http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Bibliografia

Cohen, I.; Cohen, J. I. (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, I. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.