Introducció a les dades

PID_00272091

Blas Torregrosa García

Temps mínim de dedicació recomanat: 1 hora

Universitat Oberta de Catalunya

Blas Torregrosa García

Enginyer en Informàtica i màster universitari en Seguretat de les Tecnologies de la Informació i de les Comunicacions (MISTIC) per la Universitat Oberta de Catalunya (UOC). Especialitzat en ciberseguretat. Professor col·laborador del màster de Ciència de Dades de la UOC i professor associat a la Universitat de Valladolid (UVA).

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Ferran Prados Carrasco (2020)

Primera edició: febrer 2020

Av. Tibidabo, 39-43, 08035 Barcelona
Realització editorial: FUOC

Índex

Introducció

L'objectiu d'aquest mòdul és definir què són les dades i la relació existent entre les dades, la informació i el coneixement. També s'introdueixen els diferents tipus de dades.

En aquest mòdul s'utilitzarà el cicle de vida de les dades per descriure les diferents fases per les quals poden passar aquestes abans de convertir-se en informació i després en coneixement.

1. Concepte de dades

En primer lloc parlarem una mica del que significa el concepte **dada**.

> «Del llatí *datum* ('el que es dona').
>
> Informació sobre alguna cosa concreta que permet el seu coneixement exacte o serveix per deduir les conseqüències derivades d'un fet.
>
> Document, testimoniatge, fonament.
>
> Informació disposada de manera adequada per al seu tractament per una computadora.»
>
> Diccionari RAE

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» respecte a alguna cosa observada, d'acord amb l'arrel llatina que dona origen al terme (*datum*).

1.1. Què és una dada?

Suposem que algú ens transmet la dada següent:

42

Immediatament ens apareix la pregunta «42 què?». Amb aquest senzill exemple pretenem mostrar, de moment, dues coses:

1) Una dada, sense el seu **context**, manca de significat.
2) El **format** de representació de la dada és important.

Suposem que ara ens diuen «la temperatura del pacient és de 42 graus». Hem dotat de significat al 42, quan la dada (42) és la resposta a una pregunta («Quina és la temperatura del pacient?»). Hem avançat un nivell i ja podem parlar d'**informació**. Però encara no som prou precisos. Què vol dir que «la temperatura del pacient és de 42 graus»? Doncs coses ben diferents:

- Si són graus Celsius, el pacient té febre (42 °C).
- Si són graus Fahrenheit, el pacient és un cadàver fred (5 °C).
- Si són graus Kelvin, el pacient és un cadàver congelat a –231 °C.

Així doncs, per poder parlar d'informació amb propietat necessitem un tercer element:

3) Les dades tenen **unitats** i un rang associat.

El 23 de setembre de 1999, la NASA va perdre el contacte amb la sonda espacial Mars Climate Orbiter, un satèl·lit dissenyat per 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 va causar la seva destrucció. El motiu d'aquest error va ser que una part del programari utilitzat per al càlcul de les trajectòries orbitals usava el sistema mètric decimal, mentre que altres mòduls del programari usaven el sistema anglosaxó d'unitats (peus, polzades, etc.). Aquest error va costar a la NASA un total de 327,6 milions de dòlars, l'import 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.

1.2. Dades, informació i coneixement

Dades i *informació* són dues paraules usades indistintament. De fet, una participa en la definició de l'altra, encara que no són sinònims.

Les **dades** són els fets o detalls dels quals es deriva la **informació**.

Informació

Del llatí, *informatio*: 'concepte' o 'explicació d'una paraula'. Acció i efecte d'informar (donar forma o descriure).

Com hem vist, les dades individuals en comptades vegades són útils per si mateixes: necessiten tenir un context perquè es puguin convertir en informació.

El **coneixement** és la capacitat de saber, la capacitat d'actuar i la capacitat d'entendre que resideix en el cervell.

Coneixement

Acció i efecte de conèixer (del llatí, *cognoscĕre*: 'esbrinar per l'exercici de les facultats intel·lectuals la naturalesa, les qualitats i les relacions de les coses').

Taula 1. Comparació dades-informació-coneixement

	Definició	Respostes
Dades	Símbols que representen propietats d'objectes i esdeveniments. Propietats bàsiques sense refinar ni filtrar que s'han de processar.	
Informació	Quan les dades s'han processat, organitzat, estructurat i posat en context, resultant llavors útils.	Dona resposta a les preguntes «qui», «què», «on» i «quan».
Coneixement	Resideix en les persones i resulta de l'aplicació de les dades i la informació.	Dona resposta a la pregunta «com».
Saviesa	Resideix en les persones i és l'aplicació del coneixement.	Dona resposta a la pregunta «per què».

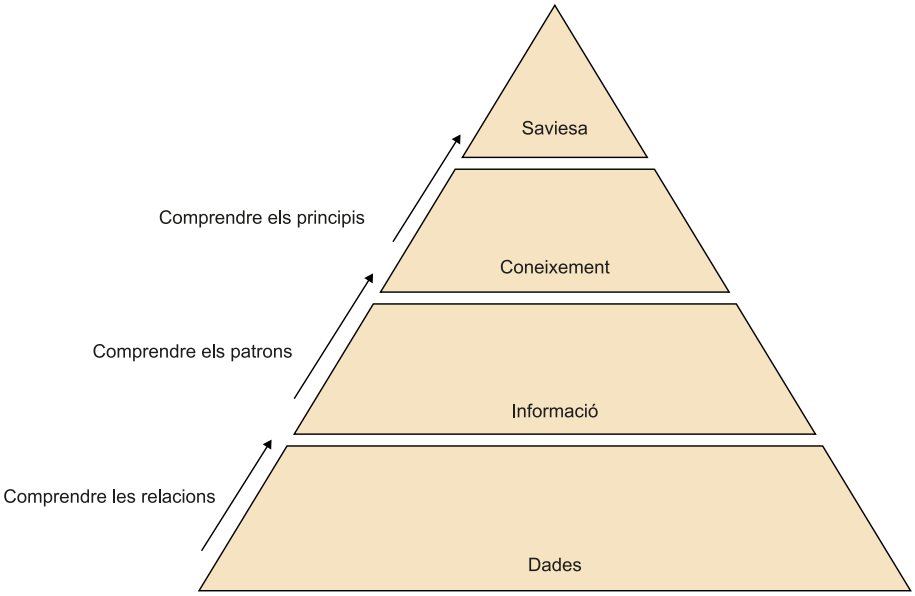
La informació i el coneixement són dos conceptes molt diferents, encara que tots dos es basen en un element primordial: les dades. És el que es coneix com la **piràmide DIKW** (*Data*, *Information*, *Knowledge* i *Wisdom*).

En la piràmide DIKW les **dades** són el nivell més bàsic, la **informació** afegeix el context, el **coneixement** afegeix com usar la informació i la **saviesa** incorpora el perquè (aplicar aquest coneixement en benefici propi o comú).

Saviesa

De sabedor: grau més alt del coneixement adquirit per mitjà de l'estudi o de l'experiència.

Figura 1. Piràmide o jerarquia DIKW



1.3. Tipus de dades

Un detall fonamental quan es tracta amb dades és que el seu tractament depèn de la seva naturalesa o tipus de dada. Les dades poden ser quantitatives o qualitatives.

1.3.1. Dades quantitatives

Les **dades quantitatives** (també denominades **dades numèriques**) són aquelles que es poden mesurar o quantificar (que poden ser explicades).

Poden ser de dues classes:

1) Dades quantitatives contínues. Hi ha un continu de valors possibles de la dada, que no es restringeix a valors discrets. Els valors es mesuren en comptes d'explicar-se. Les operacions disponibles inclouen igual (=), diferent (≠), menor (<), major (>), suma (+), resta (–), multiplicació (×), divisió (÷), etc.

Exemples de dades quantitatives contínues

Altura: 181,3 cm, 194,1 cm, 167,44 cm, etc.

Diners: 23,53 €, 11,18 $, 1.053,99 ¥, etc.

2) Dades quantitatives discretes. No admeten valors intermedis. S'enumeren (expliquen) més que es mesuren. Solen prendre solament valors sencers. Les operacions disponibles inclouen igual (=), diferent (≠), menor (<), major (>), suma (+), resta (–), multiplicació (×), divisió (÷), etc.

Exemples de dades quantitatives discretes

Nombre de queixes dels clients: 1, 2, 5, etc.

Nombre de persones en el curs: 10, 15, 25, etc.

1.3.2. Dades qualitatives

Les **dades qualitatives** (també anomenades **dades categòriques**) són aquelles que no es poden expressar numèricament i representen una qualitat o atribut que classifica o descriu cada subjecte en una d'entre diverses categories. Hi ha un nombre acotat de possibles categories.

1) Nominals. Representen categories sense un ordre intrínsec. Les úniques operacions disponibles són igual (=) i diferent (≠).

Exemples de dades qualitatives nominals

Color del cabell: ros, bru, castany, pèl-roig, etc.

Gènere: home, dona, etc.

Estat civil: casat/a, solter/a, vidu/a, divorciat/a, etc.

2) Ordinals. Representen categories en què hi ha un ordre lògic, precedència o jerarquia (sigui natural o assignada segons alguna preferència). La distància entre les categories no és, en general, coneguda. Les operacions disponibles són igual (=), diferent (≠), menor (<) i major (>). Les dades ordinals poden ser:

a) Seqüencials: hi ha un valor inicial (o zero) i tots els valors parteixen d'aquest.

b) Divergents: és possible identificar un punt central (o zero) i les dades estan per damunt i per sota d'aquest.

c) Cícliques: els valors es repeteixen formant cicles.

Exemples de dades qualitatives ordinals

Nivell: baix, mitjà, alt

Educació: bàsica, secundària, universitària

Mesos de l'any: gener, febrer, març, etc.

1.4. Estructures de dades

Les dades es poden presentar de moltes formes diferents. 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 d'altres segons la seva estructura. Així, tenim:

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

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

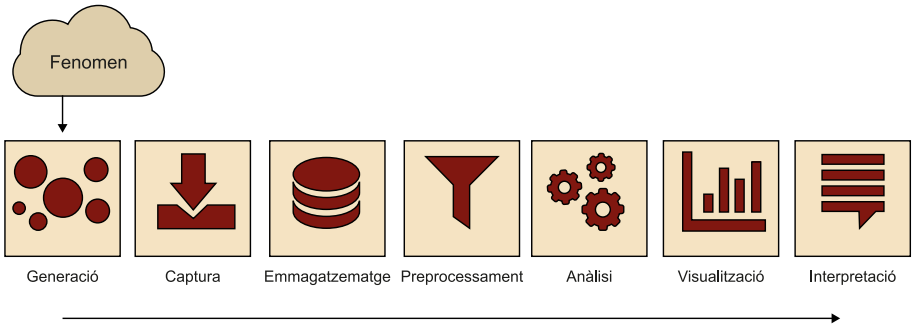
> Un **conjunt de dades** (*Dataset*) és una col·lecció de dades relacionades entre si a les quals es pot accedir individualment o en combinació i que es gestionen conjuntament.

Els conjunts de dades han estat capturats i organitzats per a un propòsit concret. Els conjunts de dades poden estar segmentats en diverses parts formant subconjunts de dades separades.

2. Cicle de vida de les dades

Hi ha una gestió de les dades, des que es generen les dades fins que s'interpreten, denominada **cicle de vida de les dades**, que consta d'una sèrie de fases, cadascuna de les quals té com a objectiu generar valor a partir de les dades. No totes les fases són estrictament necessàries i, a més, poden superposar-se o realitzar-se simultàniament en alguns casos (figura 2).

Figura 2. Cicle de vida de les dades



2.1. Generació

El cicle comença amb la **generació** de les dades. Suposem que hi ha un fenomen (físic, social, cultural, de comportament, etc.) que volem conèixer amb més detall. Aquest fenomen serà l'origen de les dades.

Al món actual es generen contínuament dades: cada cerca en un motor de cerca, cada clic a la web, cada vídeo que es reprodueix, missatge que s'envia o lloc que es visita, contribueix a la massiva empremta digital que diàriament es genera. A més, hi ha dades de sensors que monitoritzen infraestructures, càmeres de vigilància per al control del tràfic i seguretat, i també la internet de les coses (IoT), amb més sensors que generen més dades.

IoT

Internet of Things és un concepte d'interconnexió d'objectes a internet.

2.2. Captura

Una vegada generades les dades, la fase següent és la **captura**. Aquesta fase té com a objectiu recopilar les dades generades. Per motius de rellevància, qüestions de consideració d'importància o de capacitat de processament, no sempre es recopilen necessàriament totes les dades generades. La recopilació es realitza 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. Per exemple, cada vegada que un usuari paga amb targeta de crèdit en un establiment es genera un nou registre que defineix

perfectament quina quantitat de diners s'ha gastat, a quin establiment, a 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 capturant-les a mesura que es van trobant, idealment immediatament després de la seva generació. Per exemple, és possible capturar les piulades que contenen una certa paraula clau o *hashtag* tal com van apareixent en el flux (*Timeline*) de piulades d'un usuari o en el flux públic, ja que Twitter és un servei que genera dades en obert.

Per desgràcia, 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 són més freqüents les estratègies de captura de dades basades en l'extracció de dades ja existents o en la seva generació mitjançant mecanismes alternatius, entre els quals:

1) 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 permeti el seu accés.

2) Accés mitjançant una API (*Application Programming Interface*): en alguns casos sí que és possible utilitzar un mecanisme que permet realitzar consultes específiques sobre un conjunt de dades, obtenint solament aquelles que han estat sol·licitades d'acord amb els paràmetres de la consulta.

3) Captura de dades mitjançant *scraping*: quan no es disposa d'una API per accedir a les dades, de vegades és possible utilitzar eines (també anomenades *Bots*) que simulin la navegació d'un usuari per pàgines web i que extreguin el contingut de les pàgines visitades, analitzant la seva estructura i dades.

4) Extracció de dades de documents de text: en certs casos es publiquen dades en diversos formats, no sempre pensats per a la seva reutilització, que contenen taules, llistes, etc.

5) 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. Altres vegades cal obtenir dades directament proporcionades pels usuaris perquè interessa conèixer de primera mà alguns aspectes que no es poden recollir mitjançant una enquesta, especialment el «per què?» i el «com?», i aspectes difícilment quantificables, com ara emocions o sentiments.

2.3. Emmagatzematge

Les dades capturades s'han d'emmagatzemar en un format que permeti la seva posterior manipulació, d'acord amb la representació més adequada, tenint en compte tant la seva tipologia com l'ús que es vulgui efectuar amb aquestes. En funció del seu objectiu, de la freqüència d'accés i de la complexitat es pot parlar de:

- **Arxius simples**: les dades s'emmagatzemen en arxius (o col·leccions d'arxius) segons uns certs criteris. Per exemple, l'origen de les dades o la data de la captura.

Exemple d'arxius simples

Un exemple d'arxiu simple poden ser els fitxers del registre d'activitat dels servidors (*log*) que contenen totes les peticions que es realitzen quan els usuaris naveguen per les pàgines web d'un servei en línia.

- **Bases de dades**: es tracta d'una distribució més o menys complexa que permet representar les dades d'acord amb la seva estructura, també tenint en compte les relacions entre tots els elements que les componen. Se sol parlar de bases de dades relacionals o SQL, i també de les denominades no tradicionals o NoSQL, que pretenen donar solució a problemes que tenen a veure amb l'escalabilitat de les relacionals, especialment per a grans volums de dades o per a bases de dades de grafs.

2.4. Preprocessament

L'objectiu d'aquesta etapa és preparar les dades per a la seva anàlisi posterior, de manera que es puguin usar en processos de ciència de dades o d'aprenentatge automàtic, sense haver de preocupar-se per aspectes relacionats amb la seva qualitat, procedència, etc. Entre altres operacions típiques d'aquesta etapa es poden destacar les següents:

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

2) Selecció/filtratge: consisteix a obtenir algunes dades que són d'interès, segons uns certs criteris de cerca.

3) Conversió: freqüentment les dades estan en formats que dificulten la seva anàlisi posterior, per la qual cosa cal convertir-les en un format més manejable.

4) Neteja: també coneguda com a *Data Cleaning*, consisteix a eliminar totes les inconsistències en les dades que puguin ser detectades, amb el propòsit de garantir la validesa de les dades.

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

6) Creació de noves variables/indicadors (variables derivades): de vegades cal realitzar càlculs amb les variables per, per exemple, calcular la relació entre dues variables, generant noves variables o indicadors.

És important destacar la importància d'aquesta fase, ja que la qualitat dels resultats obtinguts dependrà, directament, de la qualitat de les dades a partir de les quals s'hagin obtingut els resultats.

És el que es coneix com a *Garbage in*, *Garbage out*, és a dir, si s'usen dades brossa per crear models, aquests també seran, probablement, brossa. A més, resulta convenient emmagatzemar les dades netes per a futures reutilitzacions.

2.5. Anàlisi

Una vegada les dades ja es consideren vàlides, es pot procedir a la seva **anàlisi**. Aquí s'inclouen totes les tècniques computacionals i estadístiques d'anàlisi de dades amb el propòsit de: obtenir coneixement, construir classificadors, construir predictors o inferir causalitat mitjançant la utilització d'algorismes i mètodes d'intel·ligència artificial, de la mineria de dades, de l'aprenentatge automàtic i de la teoria de la inferència estadística.

L'objectiu d'aquesta etapa és obtenir **models** que expliquin com són les dades i les seves característiques principals, i poder respondre les preguntes plantejades sobre el fenomen que ha originat les dades.

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 fenomen original. Això també inclou l'avaluació del propi model, identificant quines variables o característiques són les més rellevants, la capacitat de generalització amb dades fins ara mai utilitzades en la creació del model, o la seva capacitat d'adaptació als canvis en les dades.

2.6. Visualització

Els humans disposem d'un sistema visual molt complex i avançat, que inclou des de l'ull fins al còrtex visual, l'encarregat de processar tota la informació recollida pel primer. Els humans som, principalment, 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.

La **visualització** de dades ajuda a presentar els resultats de l'anàlisi d'una forma clara i senzilla que un ésser humà pugui comprendre visualitzant-la. En aquesta fase, per transmetre millor els resultats de l'anàlisi, cal considerar, a més de la funcionalitat, l'estètica i la capacitat de percepció visual humana.

Un aspecte interessant de la visualització de dades és que es pot convertir en una interfície de navegació de les pròpies dades, permetent certes operacions bàsiques (selecció, agregació, etc.), de manera que sigui possible afegir interactivitat a la visualització.

D'aquesta forma, és possible basar l'anàlisi de les dades a partir de la seva visualització, de manera que es combinin la capacitat visual humana per a la detecció de patrons, tendències, etc., amb la potència d'un sistema informàtic que permeti seleccionar, filtrar o comparar dades.

2.7. Interpretació

Mitjançant la **interpretació** proporcionem una explicació del que significa la visualització, i generem un relat que expliqui el context, les implicacions i les possibles opcions que s'extreuen de la visualització.

Aquesta interpretació pot ser publicada en forma de noves dades, de manera que sigui possible que tercers puguin reutilitzar-les amb altres propòsits, especialment les dades ja processades llestes per a la seva anàlisi, en forma d'una o més taules.

En tot el cicle de vida de les dades s'han obviat les possibles realimentacions existents. Inevitablement, després de presentar algunes observacions i descobriments sobre les dades, es poden generar nous interrogants i qüestions que podrien exigir recopilar més dades o realitzar altres tipus d'anàlisis.

Bibliografia

Ackoff, R. (1989). «From data to wisdom». *Journal of Applied Systems Analysis* (vol. 16, núm. 1, pàg. 3-9).

Chen, M. i altres (2009). «Data, Information, and Knowledge in Visualization». *IEEE Computer Graphics and Applications* (vol. 29, núm. 1, pàg. 12-19).

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

Stanton, J. M. (2013). *Introduction to Data Science*. Nova York: Syracuse University.

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