Els rols, àmbits i noms de la ciència de dades

PID_00261826

Marçal Mora Cantallops

Universitat Oberta de Catalunya

Marçal Mora Cantallops

Enginyer industrial i enginyer informàtic per la UPC, màster en Data Science per la UAH i doctorand en Comunicació, Informació i Tecnologia de la Societat en Xarxa per la mateixa universitat. Investigador en l'àmbit dels *game studies*, la ciència de dades i, en particular, l'anàlisi de xarxes socials; està interessat en l'ús d'aquestes tècniques per a l'extracció de coneixement i informació. Ha treballat en la creació i optimització de models estadístics per a logística i planificació de la demanda i actualment participa en diversos projectes relacionats amb l'estadística i la ciència de dades.

L'encàrrec i la creació d'aquest recurs d'aprenentatge UOC han estat coordinats pel professor: Josep Maria Marco (2019)

Índex

Introducció

Benvinguts a Ciència de dades! En aquest mòdul s'introduiran molts conceptes i terminologia de la ciència de dades, la majoria dels quals no us seran familiars.

Això és del tot normal perquè la ciència de dades és una disciplina relativament moderna, que es basa en molts elements i principis que provenen de disciplines molt diferents, des de la matemàtica i l'estadística fins a les ciències de la computació, passant per les activitats empresarials. És molt difícil trobar persones que dominin aquests tres elements i encara ho és més que ho facin en més d'un àmbit. Així que no us amoïneu, apreneu tants conceptes com pugueu, que us sonin, i més endavant, a mesura que avanceu en els estudis, ja els anireu entenent i consolidant. En aquest mòdul es tracta de proporcionar una perspectiva general i global de la ciència de dades i això té dues conseqüències: que apareixeran molts conceptes nous i que no es podran desenvolupar amb la profunditat que mereixen. Però res que no es pugui solucionar!

L'estructura del mòdul és la següent:

1) En primer lloc, es presentarà el context històric en què sorgeix i neix la ciència de dades, la qual cosa ajudarà a entendre la seva emergència i les preguntes a les quals busca donar resposta.

2) En segon lloc, s'elaborarà la definició del científic de dades: què és, què fa, quines habilitats ha de cultivar i de quines eines disposa.

3) En tercer lloc es proposarà una visió molt general d'alguns àmbits en els quals ja s'està aplicant la ciència de dades i es parlarà del potencial que presenta en alguns altres.

4) Finalment, s'introduirà un petit glossari amb explicacions i exemples sobre els termes i conceptes més comuns de la ciència de dades (i, per tant, termes que apareixeran de manera habitual en el futur).

En resum, aquesta unitat és un mapa que té l'objectiu d'ajudar l'estudiant a situar-se i orientar-se en el món de la ciència de dades. És només el primer pas del camí fins a l'objectiu final. Ànims i endavant!

1. Origen i evolució de la ciència de dades

Tant l'anàlisi de dades com una de les seves branques més esteses i presents, l'anomenada intel·ligència de negoci (*business intelligence*, BI), han guanyat un protagonisme especial en les últimes dècades; les oportunitats derivades de l'ús de la informació disponible i la seva anàlisi en qualsevol organització han generat un creixement notable de l'interès en aquestes disciplines. Es pot entendre el BI com les tècniques, sistemes, tecnologies, pràctiques, aplicacions i metodologies que serveixen per extreure valor de les dades que, al seu torn, aconsegueixin que el negoci (o l'organització) prengui decisions més informades i que, per tant, tinguin un retorn positiu. La ciència de dades, entesa en el seu sentit més ampli, no és ni de bon tros nova, no obstant els seus orígens es poden seguir fins pràcticament a mitjan segle passat, amb els primers intents de dotar d'intel·ligència les primeres (i mastodòntiques) computadores. El que sí que és relativament nou és la seva popularitat i projecció, esperonada per unes capacitats tècniques que per primera vegada aconsegueixen acostar-se a les seves promeses. Un article famós de la prestigiosa *Harvard Business Review* la presentava, el 2012, com «la feina més sexy del segle XXI»; tendència accentuada en els últims anys, especialment als Estats Units.

Però, deixant de banda les xifres i les perspectives econòmiques, és necessari entendre com s'ha arribat fins aquí. Tant les dades massives (*big data*) com la ciència de dades (*data science*) són dues de les paraules més de moda de la dècada, però no són idees noves. De fet, no són ni idees d'un sol àmbit i porten més de cinquanta anys fent-se lloc en la indústria i les disciplines d'anàlisi. És moment, doncs, de mirar mig segle enrere.

1.1. Models estadístics i mineria de dades

Per trobar la primera referència al canvi aportat per la computació és necessari remuntar-se a l'any 1962, quan John Tukey va adonar-se del potencial de la intersecció entre l'estadística i la computació en una cita que és, a dia d'avui, cèlebre:

> «[...] a mesura que he vist evolucionar l'estadística, he tingut motius per reflexionar i dubtar [...] crec que he descobert que el meu interès principal és l'anàlisi de dades...» (Tukey, 1962).

Tukey està referint-se, d'alguna manera, a l'amor a primera vista que va sentir quan, mitjançant ordinadors, els resultats estadístics podien ser obtinguts en hores, molt més ràpidament que els dies o setmanes que es tardava amb els mètodes manuals.

Un altre nom important és el de Peter Naur que, avançat al seu temps, va publicar el *Concise Survey of Computer Methods* (Naur, 1974), un compendi de mètodes de processament de dades en múltiples aplicacions. El més curiós d'aquest cas és que ja citava diverses vegades el terme *ciència de dades*, que definia de la manera següent:

> «la ciència de treballar amb dades, una vegada establertes, mentre la relació de les dades amb el que representen es deixa a altres camps i ciències».

Qui havia de dir que el terme *ciència de dades* té més de quaranta anys, oi?

Tot i que la definició és poc clara i que les seves idees tardessin a ser enteses per la comunitat científica i empresarial, es pot considerar que aquest és un dels primers intents de recollir tota la feina realitzada en aquest nou camp.

La dècada dels setanta és, de fet, una de les més importants en el desenvolupament de nous models estadístics que aprofiten el nou paradigma computacional. Moltes de les tècniques utilitzades avui dia segueixen basant-se en els avanços teòrics d'aquesta dècada prodigiosa en la qual es fundà l'Associació Internacional per a l'Estadística Computacional (IASC, per les seves sigles en anglès) l'any 1977. El seu origen es troba en la voluntat d'enllaçar, precisament, les metodologies estadístiques tradicionals i la tecnologia moderna que aportaven els ordinadors. Però anava més enllà, també buscava integrar els experts i especialistes de cada domini per tal de convertir aquestes dades en informació i coneixement. Aquest segon pas és allò que defineix la ciència de dades.

Aquell mateix any, Tukey (1977), que seguia investigant els primers passos del nou camp científic, va publicar *Exploratory Data Analysis*, en el qual torna a destacar la importància d'aprofitar les dades per seleccionar les hipòtesis en qualsevol experiment i fa, també, una crida a combinar els enfocaments exploratoris i confirmatoris en l'anàlisi de dades per obtenir millors resultats. Tot i la promesa teòrica dels avanços registrats en la dècada dels setanta, la limitada capacitat computacional real i la dificultat d'accés als centres de càlcul i ordinadors van alentir el procés d'integració. Pràcticament en paral·lel es començava a desenvolupar una altra tècnica relacionada: el *data mining* o mineria de dades. Considerada una pràctica repudiable durant els seixanta i setanta, les males llengües l'anomenaven «pesca de dades» o «dragat de dades», un títol despectiu que es referia a l'anàlisi de dades sense hipòtesi prèvia. A l'inici dels vuitanta, no obstant, alguns analistes de bases de dades van començar a canviar la connotació del terme cap a la més positiva «experimentació», parlant de *database mining*. El terme, no obstant, va acabar tornant a *data mining* de la manera més americana possible, ja que *database mining* era una frase que ja estava registrada per una companyia.

Més tard, els esforços d'un conegut científic de dades, Gregory Piatetsky-Shapiro, per avançar en l'extracció d'informació de les bases de dades van desembocar en una línia d'investigació que va batejar amb el nom de *Knowledge Discovery in Databases* (KDD), la qual cosa va portar a l'organització l'any 1989 de la primera conferència sobre la cerca de coneixement en bases de dades, avui coneguda com *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* i que se celebra anualment.

1.2. Intel·ligència de negoci

En els anys noranta es produeix una transició a les empreses que continuarà fins al canvi de mil·lenni. Basant-se en els mètodes estadístics desenvolupats durant els setanta i les tècniques de mineria de dades dels vuitanta, les tecnologies i aplicacions més habituals de la industria (encara avui) es van començar a popularitzar. Aquesta primera aproximació de presa de decisions basada en dades (la intel·ligència de negoci) parteix dels fonaments de la gestió de les bases de dades i, per tant, es basa en informació estructurada, recollida per les mateixes companyies (habitualment amb sistemes i servidors de programari antiquat, com els ordinadors centrals que controlen processos de producció) i que s'emmagatzemen en sistemes comercials de gestió de bases de dades relacionals (RDBMS). D'aquesta necessitat de gestió i emmagatzematge de dades neix el BI. Per poder prendre decisions informades, apareix la necessitat de transformar aquestes dades massives i de baix nivell en informació més intel·ligible per als equips de direcció i de presa de decisions.

En aquest entorn, el disseny de *data marts* i eines per a l'extracció, la transformació i la càrrega de dades (anomenats ETL) són essencials per convertir i integrar les dades específiques de cada negoci. De la mateixa manera, els analistes de dades necessiten eines per explorar les dades; eines que obtenen de llenguatges de consulta de bases de dades i dels cubs OLAP (*Online Analytical Processing*), pensats per facilitar l'accés a les dades de treball, i que són, també, complementats per eines gràfiques rudimentàries que permeten explorar algunes característiques de les dades. De cara a la direcció, cal destacar especialment el desenvolupament del *Business Performance Management* (BPM), amb el suport dels registres de resultats però, sobretot, dels quadres de comandament, inspirats en la proposta de quadre de comandament integral de Kaplan i Norton (1996), que ajuden a analitzar i visualitzar una sèrie de mètriques de rendiment. A més de totes aquestes eines orientades a generar informes, també cal afegir l'expansió de les tècniques mencionades anteriorment, tant estadístiques com de mineria de dades, per associar, segmentar, agrupar i classificar dades, preparar models de regressió, de detecció d'anomalies i, fins i tot, fer prediccions cada vegada en més aplicacions de negoci.

ETL

Extracció, Transformació i Càrrega (*Load*) és el nom que es dona al procés que permet moure dades de múltiples fonts, transformar-les i netejar-les per posteriorment carregar-les de nou a un procés de negoci.

OLAP

L'objectiu dels cubs OLAP és agilitzar la consulta de grans quantitats de dades. Es pot imaginar com una reordenació del contingut de les bases de dades relacionals, una vista, per tal de fer més eficients les operacions de consulta.

BusinessWeek va publicar, el setembre de 1994, un article de portada sobre allò que va anomenar «màrqueting de base de dades». El més interessant és un passatge que transpira indecisió i que serveix per il·lustrar la dècada:

> «Les empreses estan recollint muntanyes d'informació sobre tu, processant-la per preveure la probabilitat que compris un producte, i utilitzant aquest coneixement per construir un missatge de màrqueting calibrat per tal que ho facis [...] Una empenta d'entusiasme causada per l'expansió dels escàners de compra als vuitanta va acabar en decepció general: moltes empreses estaven prou sobrepassades per l'elevada quantitat de dades com per fer res útil amb la informació [...] Sigui com sigui, moltes companyies pensen que no tenen alternativa a desafiar la barrera del màrqueting de base de dades».

El terme *ciència de dades* comença a utilitzar-se gradualment aproximadament al final de la dècada. En primer lloc, el 1996, la conferència bianual de la Federació Internacional de Societats de Classificació (IFCS) és titulada «Data science, classification, and related methods». El mateix any, Fayyad i altres. (1996) parlen de la diferència entre la simple mineria de dades (o aplicació d'algoritmes) i l'obtenció d'informació a partir de bases de dades, que implica passos addicionals com la «preparació de les dades, selecció, neteja, incorporació d'informació d'altres fonts i interpretació dels resultats de la mineria de dades», essencials per a obtenir coneixement útil a partir de les dades.

Jeff Wu, en el seu discurs d'inauguració del curs 1997 a la Universitat de Michigan, va proposar canviar el nom de l'estadística pel de *ciència de dades* i el d'estadista per *científic de dades*. Però la frase que millor il·lustra el pas al nou mil·lenni prové de Jacob Zahavi (desembre de 1999), que afegeix dos elements importants, els inicis de les dades massives i la influència creixent d'internet:

> «L'escalabilitat és un problema gegant per a la mineria de dades [...] els mètodes estadístics convencionals funcionen bé en conjunts de dades petits. Però les bases de dades d'avui dia poden estar formades per milions de files i una pila de columnes de dades [...] Un altre repte tècnic és desenvolupar models que puguin fer una millor tasca analitzant dades, detectant relacions no lineals i interaccions entre elements [...] És possible que s'hagin de desenvolupar eines especials de mineria de dades per a la presa de decisions a les pàgines web».

1.3. Internet i la web 2.0

Tot i que internet sigui bastant més antic, no és fins a principi del 2000 quan les oportunitats de recollida de dades i d'analítica comencen a aparèixer. Primer, en forma de l'anomenada web 1.0, caracteritzada pels cercadors (com Google i Yahoo), però també per l'emergent comerç electrònic (eBay i Amazon), que obtenen dades de primera mà dels seus usuaris. Ja no es tracta tan sols de treballar amb les dades tradicionals de productes i de negoci de les bases de dades relacionals, sinó que el contingut en línia proporciona detalls per IP de cada usuari sobre les seves cerques i la seva interacció amb les pàgines, dades que són recollides per mitjà de registres detallats (*logs*) i galetes (*cookies*) i que representen una oportunitat sense precedent d'identificar les necessitats de cada client i noves oportunitats de negoci. Moltes d'aquestes empreses han convertit l'explotació de les dades en el nucli dels serveis que ofereixen i en el seu avantatge competitiu.

La quantitat d'informació que es pot extreure de la gran xarxa és inacabable. L'anàlisi de clics de clients, per exemple, origina eines d'anàlisi web com Google Analytics, orientades a descobrir patrons de comportament i de compra. D'aquesta anàlisi en deriven posteriorment el disseny de pàgines web,

l'optimització del màrqueting i del posicionament, l'anàlisi de mercat i les recomanacions. Així es fa visible com la ciència de dades no tan sols aprofita l'extracció da dades per prendre decisions informades en l'àmbit de negoci, sinó que acaba transformant tot el que l'envolta, des de la manera de comprar fins a la influència que té sobre les pàgines web que els usuaris visiten.

El problema, per dir-ho d'alguna manera, és que l'increment s'excedeix amb l'explosió del contingut generat per usuaris. Els fòrums, grups en línia, blogs, plataformes socials i fins i tot entorns virtuals omplen la web de final de la dècada dels 2000 d'inabastable contingut, difícilment tractable amb la capacitat de processament del moment. En l'àmbit de màrqueting es parla de l'oportunitat que aquest tipus de contingut suposen per als negocis d'observar i ser partícips de la conversa entre proveïdors i usuaris, canviant el paradigma tradicional d'una sola direcció.

Si la dècada va començar amb els plans de Cleveland (2001) sobre com els científics de dades haurien de preparar-se per als requeriments del futur, va acabar amb la popularització (i moda) del terme, que se sol atribuir a DJ Patil y Jeff Hammerbacher, de LinkedIn i Facebook, respectivament. De l'any 2009 també cal destacar el retorn (per quedar-se) de les bases de dades NoSQL (o no relacionals).

1.4. Ciència de dades

El naixement de la disciplina de la ciència de dades tal i com la coneixem avui és a principi de la dècada de 2010. La clau és, en realitat, una tempesta perfecta d'esdeveniments.

Hadoop

El projecte Hadoop neix l'any 2006 i parteix de Nutch, un intent d'indexar la totalitat de la web.

En primer lloc, l'existència de dades massives, provinents ja no tan sols dels portals d'internet sinó de múltiples sensors de qualsevol aparell (allò que es coneix com a internet de les coses o *Internet of things*, IoT). És l'anomenat *big data*, que no és res més que el nom que rep la quantitat massiva de dades de les quals es disposa avui en pràcticament qualsevol aplicació. Segons IBM, el 90% de les dades de les quals es disposa l'any 2018 han estat generades únicament en els dos anys anteriors. Sí, això significa bàsicament que entre 2016 i 2017 es van generar més dades que des del principi de la humanitat fins a l'any 2015.

En segon lloc, arquitectures de processament distribuït com Hadoop i HDFS (i l'ecosistema *open source* que el complementa), de les quals parlarem més endavant, que permeten processar grans quantitats d'informació en clústers de computadores convencionals[1].

(1)El 2018, Google processa 40.000 cerques per segon.

Però, sobretot, allò que més destaca d'aquests últims anys és l'anomenat atac dels exponencials (vegeu la figura 1):

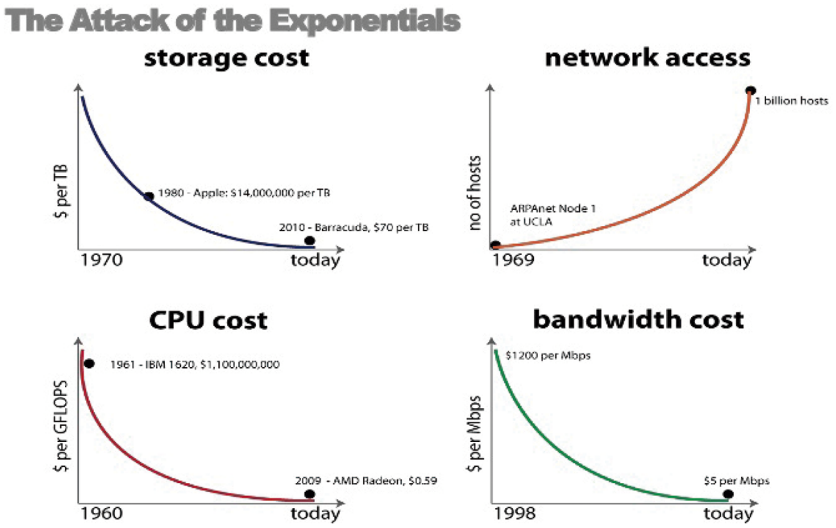
- El cost de l'emmagatzematge ha baixat de manera dràstica.

- L'accés a internet s'ha multiplicat (aproximadament mig món, pràcticament 4 mil milions de persones accedeixen habitualment a la xarxa).

- El cost de les CPU s'ha reduït, també, exponencialment.

- El cost de l'amplada de banda ha deixat de ser rellevant.

Figura 1. Gràfics que mostren l'evolució exponencial d'alguns paràmetres crítics (2011)



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

En resum, uns costos econòmics cada vegada menys rellevants fan que molts problemes que abans no era rendible estudiar passin a ser-ho. Les xarxes de sensors i la generació de dades en línia fan més fàcil l'accés a les dades. I, per si fos poc, a la baixada dels costos s'hi uneix el *cloud computing* o, el que és el mateix, la possibilitat de «llogar» capacitat de computació per fer front a les necessitats puntuals de qualsevol empresa, sense haver d'invertir en grans equips o servidors. El *software as a service* (SaaS) fa la resta.

Es pot parlar, doncs, de tres grans elements que han canviat en l'anàlisi de dades orientada a oferir serveis:

1) Ja no són només dades, sinó que són dades obtingudes de manera massiva, ràpida i amb un processament que pot arribar a ser similar al temps real en alguns casos. L'abaratiment de la memòria RAM també hi ha col·laborat.

2) L'anàlisi és ràpida i a gran escala. Ja no és només que llenguatges com R o Python s'hagin expandit i formin part del nucli de les eines d'anàlisi més habituals, sinó que també han aparegut noves arquitectures distribuïdes com Spark amb potencials elevadíssims i que permeten treballar amb petabytes[2] de dades.

[(2)]Un petabyte equival a un milió de gigabytes, és a dir, 1 PB = 10^6 GB = 10^{15} bytes.

3) Els serveis oferts poden ser més específics i més centrats en cada aplicació concreta.

SaaS

El SaaS és un model de distribució de programari en el qual aquest està allotjat directament en el servidor del proveïdor (i el client hi accedeix mitjançant internet).

Durant els últims anys, la ciència de dades ha crescut i ha inclòs tant el món acadèmic com els negocis i les organitzacions del món sencer. Ja avui dia és una realitat que utilitzen els governs, els enginyers, astrònoms i metges, entre molts d'altres, arreu. El pas al *big data* no representa només un canvi d'escala, sinó que també ha portat noves maners d'entendre i processar les dades, canviant la manera d'estudiar-les i analitzar-les.

Apache Spark

Apache Spark és un motor unificat d'analítica orientat al processament de dades massives, amb mòduls destinats a la reproducció en continu (*streaming*), SQL, aprenentatge automàtic i processament de grafs.

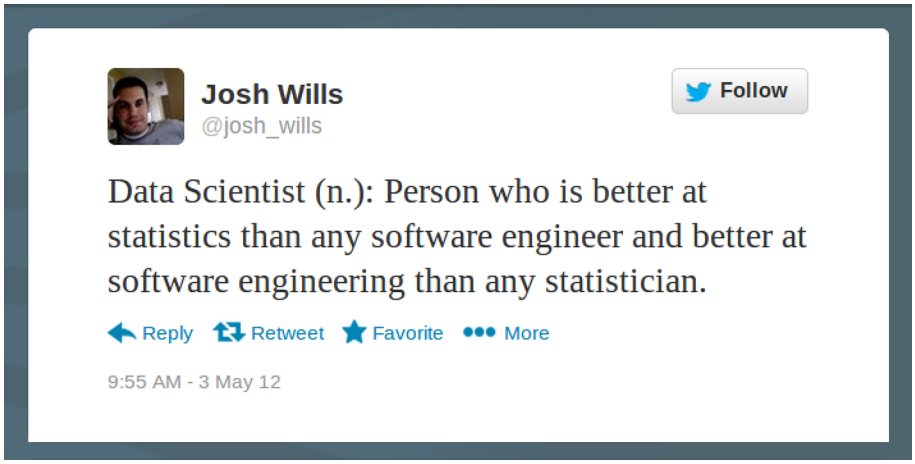
Així doncs, la ciència de dades s'ha convertit en part important de la recerca tant empresarial com acadèmica. Els àmbits són amplis, però inclouen la traducció automàtica, la robòtica, el reconeixement de veu, l'economia digital i els motors de cerca, entre d'altres. Les disciplines també són transversals: des de la biologia, la medicina i la salut, fins a les humanitats i les ciències socials. L'anàlisi que proporciona la ciència de dades influencia, en el dia a dia, l'economia, la política i les finances. Els darrers anys han proporcionat, a més a més, el desenvolupament i la millora de tècniques que fins no fa tant eren costoses en l'àmbit de producció. L'aprenentatge automàtic (*machine learning*) ha liderat les tècniques utilitzades (i és on probablement es troben algunes de les més madures), però també és notable el creixement que han experimentat noves àrees com l'aprenentatge profund (*deep learning*), les xarxes neuronals o l'anàlisi de xarxes socials. Encara hi ha, no obstant, molt camí per recórrer. Molts problemes són constants i complexos de resoldre i, igualment, no en deixen d'aparèixer de nous!

2. El rol del científic de dades

2.1. Què és un científic de dades?

Josh Wills, de Slack, va definir el científic de dades en una famosa piulada de la manera següent:

Figura 2. Tweet de Josh Wills sobre el científic de dades



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

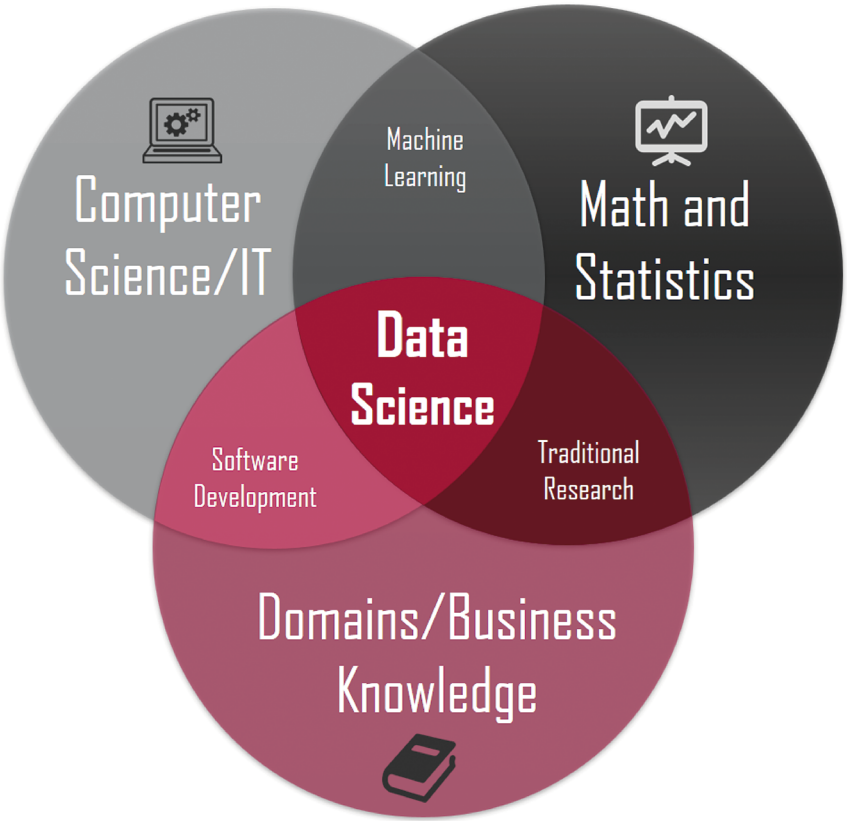
La seva definició (en clau d'humor) no se situa massa lluny de la realitat, com veurem a continuació, encara que és cert que el rol del científic de dades és una mica més complex que el que s'extreu d'aquests cent quaranta caràcters.

De la disponibilitat de grans volums de dades apareix la necessitat, especialment en l'àmbit de negoci, d'utilitzar-les per guanyar un avantatge competitiu. Queda clar que les empreses i organitzacions que siguin capaces d'utilitzar de manera efectiva aquest tipus d'informació seran també propenses a prendre millors decisions i posar-se davant de la resta de competidores.

Per tal d'inferir informació raonable i útil de tal quantitat de dades apareix la necessitat de comptar amb professionals amb un conjunt d'habilitats i aptituds que no existien. Aquests perfils, que s'anomenen científics de dades, combinen bàsicament tres disciplines bàsiques i les dominen en profunditat:

- matemàtiques i estadística
- ciències de la computació i programació
- coneixement de l'àrea de negoci

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



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

Tal com es pot observar en la figura 3, la intersecció entre cada parella de les disciplines requerides proporciona un perfil que ajudarà a entendre com es relacionen aquestes habilitats. Així:

- Un professional que tingui habilitats de programació i de matemàtiques i estadística és un perfil d'aprenentatge automàtic. En realitat, es tracta d'un perfil que a vegades es considera perillós o que té poc sentit en una empresa, ja que extreure conclusions sobre un domini desconegut per a l'investigador pot ser contraproduent.

- El domini de la informàtica i el coneixement del sector (o domini) en el qual es treballa porta a ser desenvolupador de programari d'aquell àmbit, normalment de manera específica.

- La utilització de les habilitats matemàtiques i estadístiques per a la investigació en un domini concret és el que ha fet des del principi la recerca anomenada tradicional, contrastant hipòtesis, per exemple.

- En el punt de trobada entre les tres disciplines es troba el científic de dades, habitualment considerat una *rara avis*.

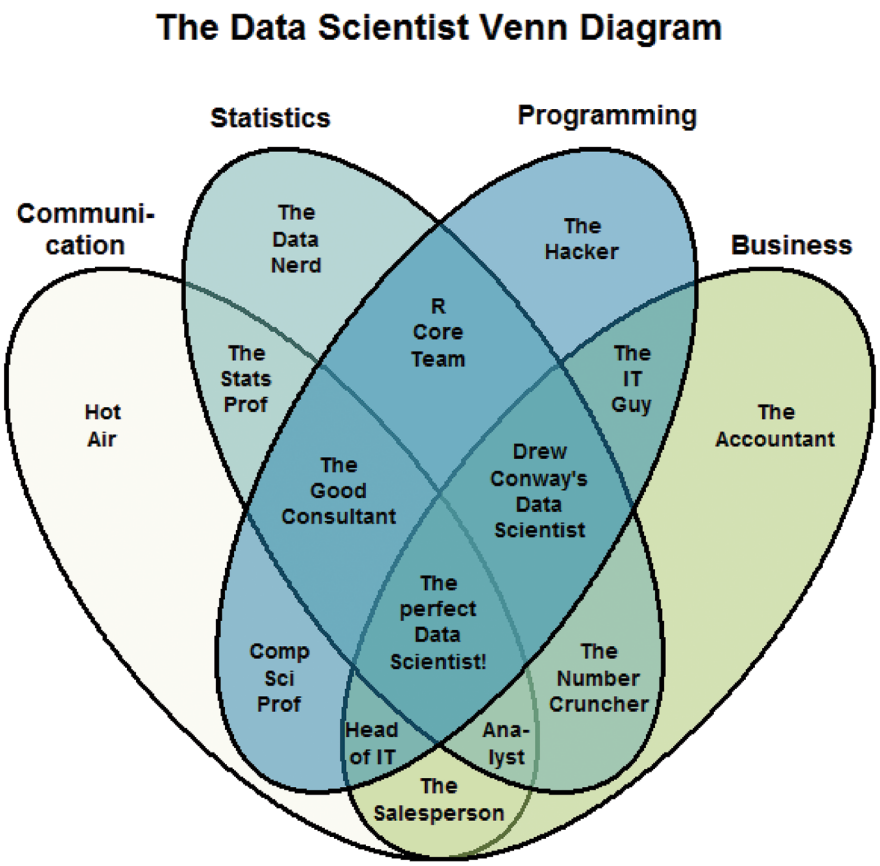
El problema d'aquest perfil professional és, com es pot albirar, que tradicionalment les dues disciplines base (la branca computacional i la matemàtica) s'han tractat per separat i és estrany trobar informàtics amb una base matemà-

tica profunda (és una mica menys estrany trobar matemàtics amb bona competència informàtica, no obstant). A més a més, tant uns com els altres han ocupat, habitualment, posicions molt concretes a les empreses, no sempre a prop de l'activitat empresarial. Per aquest motiu, els professionals que reuneixen habilitats avançades en totes tres disciplines reben, en alguns sectors, el nom d'«unicorns», qüestionant de manera jocosa la seva existència. El camí habitual del científic de dades és, doncs, el d'aconseguir dominar dues de les disciplines, primer, i, després, ja sigui mitjançant formació addicional o amb l'entrada al món laboral, desenvolupar la tercera. Fins aquí la definició tradicional, proposada per Drew Conway a principi de la dècada.

No obstant, alguns han suggerit que, a falta de les habilitats requerides, els científics de dades necessiten, i cada vegada més, destacar en una quarta disciplina: la comunicació. El raonament és senzill: de poc serveix dominar l'estadística, la programació i l'activitat empresarial si no s'és capaç d'explicar els resultats obtinguts i, encara més important, modular aquest missatge per a cada un dels departaments o actors implicats. Presentar les conclusions a la direcció, per exemple, pot requerir eines de visualització clares i intel·ligibles; explicar el model per a la seva implementació als desenvolupadors de programari és una història completament diferent, tot i que el treball hagi estat el mateix.

A la figura 4 es mostra com quedaria el diagrama de Venn transformat amb aquesta nova variable.

Figura 4. Diagrama de Venn expandit del científic de dades



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

És interessant fixar-se en com la comunicació modifica lleugerament la idea dels perfils del primer diagrama de Venn; així, un professional que combini l'alta capacitat de comunicació amb:

- Estadística, seria un professor d'estadística.
- Programació, seria un professor de ciències de la computació.
- Activitat empresarial, seria el perfil del venedor.
- Programació i estadística, seria un bon consultor.
- Estadística i negoci, seria el que s'anomena un analista de dades.
- Programació i negoci, seria director de sistemes dins de l'empresa.
- I la unió de les quatre habilitats és el que s'espera cada vegada més del científic de dades.

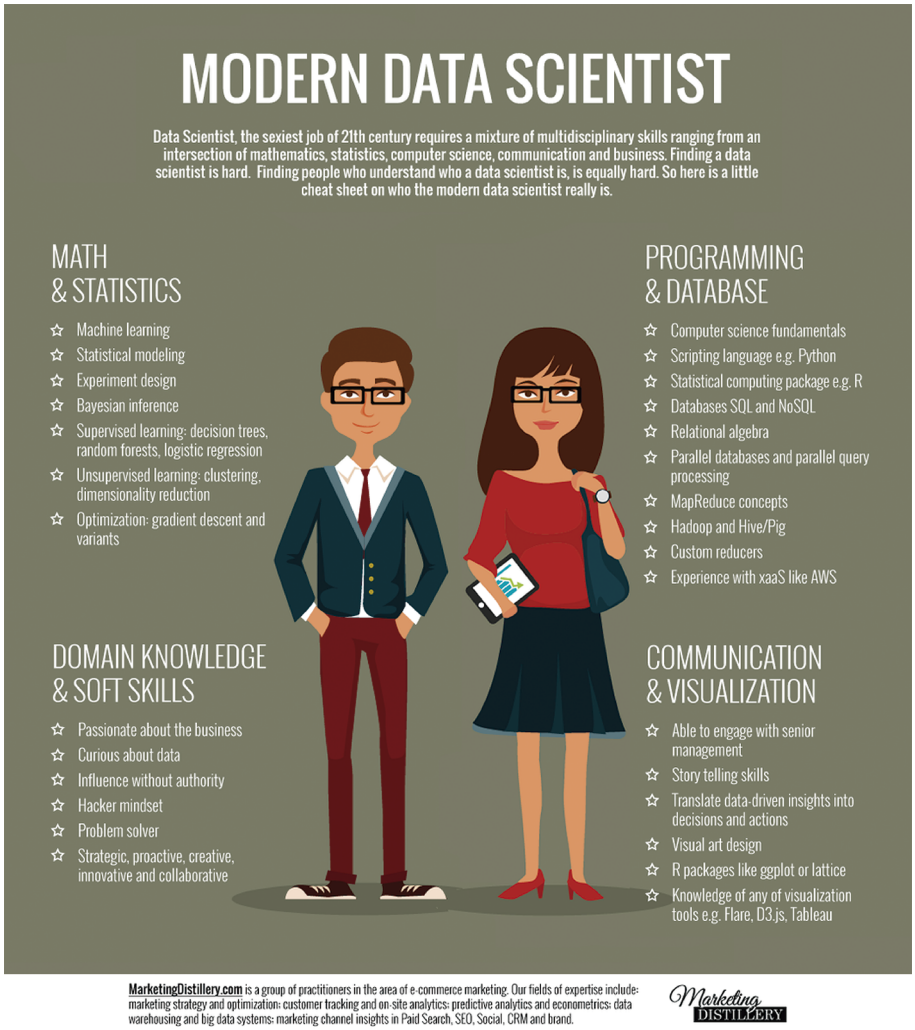
Quina és doncs la diferència entre un científic de dades i un analista de dades? En realitat els perfils tenen àrees en les quals s'assemblen, però també diferei-xen en algunes altres:

- Un analista de dades té com a objectiu la interpretació de les dades per tal d'obtenir coneixement útil per al negoci o organització. El seu punt fort ha de ser l'estadística (i la matemàtica, per extensió), però també han de tenir capacitats moderades en coneixement del negoci i programació (per, com

a mínim, ser capaços de transformar les dades). En resum, un analista de dades recull, processa i aplica algoritmes estadístics a dades estructurades per tal de respondre a una sèrie de preguntes predeterminades pel mateix negoci.

- La missió del científic de dades és similar a la de l'analista: obtenir el mateix tipus de coneixement. No obstant, el científic de dades s'ha d'enfrontar també a complicacions tècniques (volums de dades més grans i velocitats elevades de creació de les mateixes) i d'arquitectura (dades sense estructura). De tot aquest conjunt ha de ser capaç d'identificar primer les preguntes o hipòtesis que s'ha de fer (que en el cas de l'analista ja solen estar determinades) i complementar les dades de les quals disposa amb altres fonts. A més a més, neteja i transforma les dades per preparar-les per al processament i crea nous algoritmes i cerques. Per si fos poc, com s'ha vist anteriorment, també necessita habilitats comunicatives, narratives i de visualització per ser capaç de compartir els resultats a qualsevol nivell dins de l'organització.

Figura 5. Detall de les habilitats del científic de dades



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

Es podria resumir de la manera següent: mentre l'analista de dades busca conclusions i alertes sobre mètriques que la companyia considera crítiques, el científic de dades construeix nous models i cerca coneixement sobre indicadors que la companyia encara no sap que són importants.

Així que, simplificant una altra vegada, també es podria dir que la feina de l'analista de dades és crítica en el dia a dia de la companyia (supervivència diària), mentre que la del científic de dades està més orientada al mitjà-llarg termini (avantatge competitiu).

2.2. Què fa un científic de dades?

Una vegada hem vist el rol del científic de dades és més fàcil entendre les seves funcions. El científic de dades és un perfil que gaudeix de certa llibertat en la seva feina diària; ha de tenir una mentalitat oberta i curiosa, però també un cert escepticisme respecte del coneixement «establert». Aplica el mètode científic, així que és important que conegui la metodologia adequada per a la generació i comparació d'experiments. Ha de ser capaç de programar i de modificar codi, d'interactuar tant amb la gent de sistemes com de direcció i de crear històries a partir de dades.

La seva feina diària es resumeix en deu punts:

1) Fer-se (bones) preguntes: Què no se sap? Què es voldria saber? Què seria útil saber? Com es podria saber?

2) Definir i posar a prova hipòtesis, mitjançant experiments que segueixin el mètode científic.

3) Extreure, obtenir, fer *scraping*, mostrejar, etc., dades rellevants per al negoci.

Scraping

Consisteix a extreure dades de formats llegibles per humans (com una pàgina web) per al seu posterior processament.

4) Adaptar les dades a les seves necessitats de forma, distribució i format.

5) Descobrir noves dades a partir de l'exploració i de mètriques desconegudes.

6) Modelar tant les dades com els algoritmes.

7) Entendre relacions entre dades.

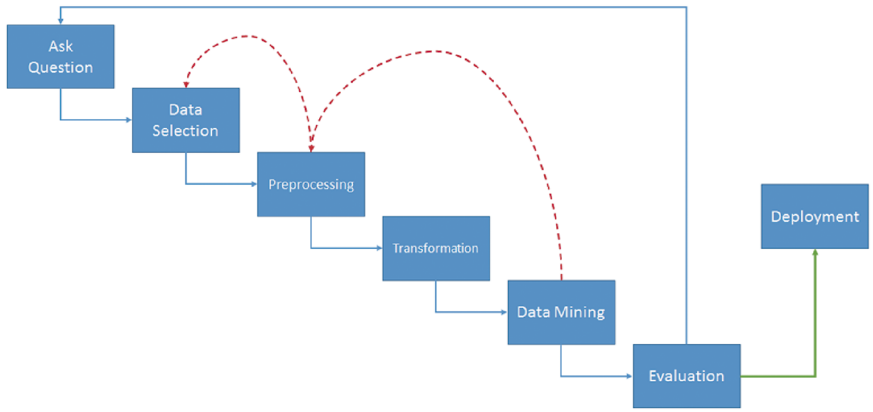
8) Aplicar aprenentatge automàtic de manera controlada i informada.

9) Crear programes i productes que proporcionin coneixement a l'empresa.

10) Explicar històries a partir de les dades, amb una narrativa fàcil de comprendre.

Aquest procés es pot representar com en la figura 6.

Figura 6. El procés de la ciència de dades



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

2.2.1. Fer-se (bones) preguntes

La feina del científic de dades comença definint el problema que es vol resoldre o el plantejament que es vol comprovar. Un bon científic de dades intenta eliminar qualsevol biaix personal que pugui tenir; no es tracta de comprovar allò que un pensa, sinó de respondre una pregunta que es planteja. És important tenir també en compte els objectius empresarials. Tenir dades i més dades no implica que siguin útils, així que entre les funcions del científic de dades s'inclou la de convertir-les en informació accionable. És també el moment d'escriure un pla, un mapa, que portarà des de les dades a un estat final desitjat.

2.2.2. Selecció de dades

Una vegada realitzat el plantejament inicial i dissenyat l'experiment, és el moment de seleccionar les dades que s'utilitzaran; aquí és important pensar no tan sols en les dades internes de les quals disposa l'organització, sinó també en les internes de les quals encara no disposa (i que es poden intentar obtenir) i en les externes que es poden incorporar al procés, algunes a cost zero (dades obertes) i altres de pagament (dades de mercat, per exemple).

2.2.3. Preprocessament

Un pas important és observar les dades. I no es tracta de començar a tocar-les i transformar-les, sinó d'avaluar-les, comprovar si són completes, quin tipus de distribució segueixen, quins problemes tenen... Obviar aquest pas pot ser un desastre per al científic de dades ja que podria deixar passar dades de dubtosa qualitat al sistema i elaborar un model amb més problemes dels desitjats. El

preprocessament pot derivar en canvis en la selecció de dades (descartant, per exemple, dades que siguin manifestament errònies o que tinguin excessius valors no informats).

2.2.4. Transformació

Arribat aquest pas, el científic de dades ja té clara la direcció en què va i les dades que té per treballar-hi. Així, aquest és el moment de transformar les dades en brut i adaptar-les per tal de poder-les utilitzar en els algoritmes desitjats. Aquest pas és també crític, perquè les dades solen estar en un estat poc òptim. Dades de diferents formats, incompletes, amb problemes de distribució i d'escala... idealment, al final del procés de transformació, el científic de dades tindrà un conjunt de dades a punt per passar al pas següent.

2.2.5. Descobriment de coneixement (o mineria de dades)

Sota aquest epígraf es troba, en realitat, la cara més visible de la ciència de dades; curiosament, també es podria dir que és una de les menys complexes. Aquí s'apliquen els algoritmes desitjats. S'ha de decidir si es vol fer servir un algoritme supervisat o no supervisat, si es vol classificar o preveure, etc. Tant els objectius inicials com les dades disponibles tenen influència sobre els algoritmes aplicables. Per què es considera una de les menys complexes si el fons matemàtic de la majoria d'algoritmes és realment complex? Perquè la major part dels algoritmes estan força definits en l'actualitat, així que rarament es baixa al pla matemàtic. Els algoritmes, a més, ja estan disponibles i només cal aplicar-los sobre el conjunt de les dades que s'han d'analitzar (no cal tornar-los a escriure i verificar, es poden reaprofitar i adaptar amb facilitat, etc.). Algunes veus opinen que aquesta dinàmica porta a un cert comportament «conservador», és a dir, fa més difícil l'aparició de canvis bruscos en les tècniques utilitzades. En tot cas, allò que queda clar és que la resta de fases requereixen molta més experiència i són, per tant, més difícils de dominar.

2.2.6. Avaluació

El model resultant és simplement això, un model. Al científic de dades li correspon comprovar i entendre si el model respon a les qüestions plantejades inicialment. És per això que el coneixement del domini o del negoci és especialment important per ser capaç d'interpretar els resultats obtinguts. L'avaluació es realitza al final de cada iteració i també és moment de considerar si fa falta tornar a començar, sigui des de la fase que sigui, per millorar el resultat obtingut o si, en cas contrari, el model és acceptable per al seu pas a producció.

2.2.7. Pas a producció

És important tenir sempre en ment que el motiu principal de tot el procés és que el coneixement obtingut sigui útil i que, per tant, pugui passar a producció. La major part dels que utilitzaran el nou model no són hàbils en la

ciència de dades i, per tant, s'ha de poder presentar el resultat d'una manera intel·ligible. Pot ser mitjançant visualitzacions personalitzades, quadres de comandament o, fins i tot, com a entrada per a altres sistemes de la companyia. Aquest pas és el que diferencia la ciència de dades d'altres disciplines exploratòries, l'objectiu de les quals no és necessàriament un producte final accionable per una companyia.

2.2.8. Tornar a començar

El procés (o la feina) del científic de dades és una constant iteració que no acaba mai. El pas a producció es pot produir en qualsevol moment en què es consideri que el balanç entre el cost d'una nova iteració i la utilitat del model ja sigui adequat per al negoci. Però cal entendre que un model no acaba mai, ja que sempre es poden afegir noves dades, sempre apareixen noves dades, sempre es generen nous algoritmes. Així, és gairebé més important generar una metodologia, un procés, que sigui repetible, reproduïble i robust, que no un model que funcioni de meravella en un primer moment però que sigui fràgil davant de qualsevol interferència. Perquè les interferències són el dia a dia de l'activitat empresarial i les coses canvien constantment, així que cal construir els models amb aquesta idea en ment.

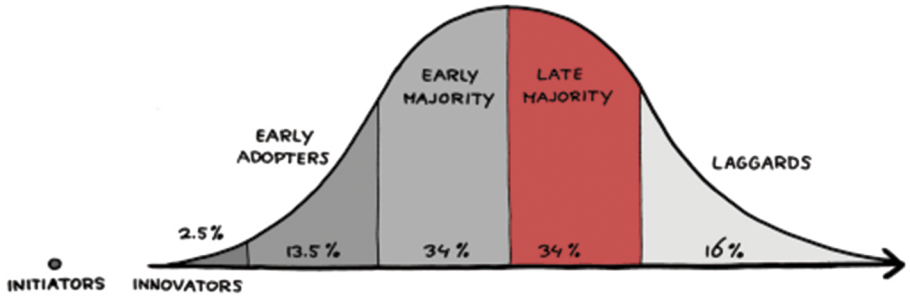
2.3. La caixa d'eines del científic de dades

L'evolució i l'expansió del nombre d'eines i d'empreses que ofereixen serveis relacionats amb la ciència de dades i les dades massives és només comparable a l'explosió de la mateixa disciplina. En el ja tradicional «Big Data Landscape» de 2018 (figura 8) cal destacar, també, la inclusió del terme *AI* ('intel·ligència artificial'), que comença a estar gairebé més «de moda» que el terme *dades massives* (i això que en la majoria de casos la gent els utilitza per descriure la mateixa realitat).

«Big Data Landscape»

Per veure la imatge de manera més detallada es pot visitar la pàgina personal del seu creador: http://mattturck.com/bigdata2018/.

Figura 7. El cicle tradicional de canvi tecnològic



Font: http://mattturck.com/bigdata2018/

Les tecnologies que treballen amb dades (aprenentatge automàtic, ciència de dades, intel·ligència artificial) segueixen creixent, cada vegada més eficients i amb més presència en negocis d'arreu del món. És també l'època de l'anomenada «transformació digital», que pot semblar estranya, ja que els ordinadors són indispensables a la major part d'empreses des de fa més de trenta anys, però que mostra com la majoria d'indústries tradicionals estan ara compromeses a convertir-se en empreses que es basen en la informació de les dades. A la figura 7, que reflecteix el cicle tradicional de canvi de tecnologia, es podria dir que l'època actual se situa just al final de l'*early majority* o, dit d'una altra manera, d'aquella primera meitat d'empreses que ja estan o han canviat de paradigma i que convencen o forcen la resta a seguir el mateix camí. Cal destacar, també, l'increment de la importància del núvol amb un creixement imparable dels proveïdors principals (AWS d'Amazon, Azure de Microsoft, Google Cloud Platform i IBM) i la creixent integració de les eines d'aprenentatge automàtic i *data engineering* en les mateixes plataformes que ofereixen.

Agenda Digital

Tal és la importància d'aquest tema que la transformació digital té fins i tot el seu propi ministeri, el d'Agenda Digital, sota el paraigua d'Indústria.

Data engineering

Tot i que no se n'hagi parlat específicament, el *data engineering* és la disciplina encarregada de donar forma a les dades.

L'ecosistema de la ciència de dades és immens (figura 8); tan immens, de fet, que allò habitual és que cada científic de dades s'especialitzi en una petita part i que esculli, d'entre totes les possibilitats, les que millor encaixen amb la seva caixa d'eines. A continuació es resumeixen en grans grups (explicar-ho amb detall és poc útil en aquest punt i els noms canvien any rere any):

- Les eines d'infraestructura, que són les que proporcionen l'entorn de treball i l'estructura funcional. Aquí es troben els proveïdors del núvol, les bases de dades tradicionals, NoSQL i NewSQL, les de graf, eines d'integració i transformació de les dades, d'emmagatzematge, de monitorització...

- Les eines d'analítica, entre les que es troben les plataformes d'anàlisi i de ciència de dades, de *business intelligence*, de visualització, d'aprenentatge automàtic, de tractament del llenguatge, d'anàlisis social i de comerç electrònic, entre d'altres.

- Cal destacar aquesta nova tendència d'integrar-les ambdues, les eines d'infraestructura més les d'analítica, en algunes de les eines existents (AWS, Google Cloud, Azure, SAS, etc.).

- També hi ha un conjunt d'aplicacions més orientades a cada:
  - àmbit: aplicacions per a vendes, màrqueting, servei al client, recursos humans, legals, finances, seguretat, etc.
  - indústria: aplicacions dissenyades per a l'educació, la publicitat, els governs, les assegurances, les finances (tant d'inversió com convencionals), la salut, el transport o l'agricultura

- El bloc més interessant i més popular en l'àmbit global és l'anomenat *open source* o eines de codi obert. Aquí se situa tot l'ecosistema Hadoop i Spark, Hive, bases de dades NoSQL, eines de reproducció en continu, estadística

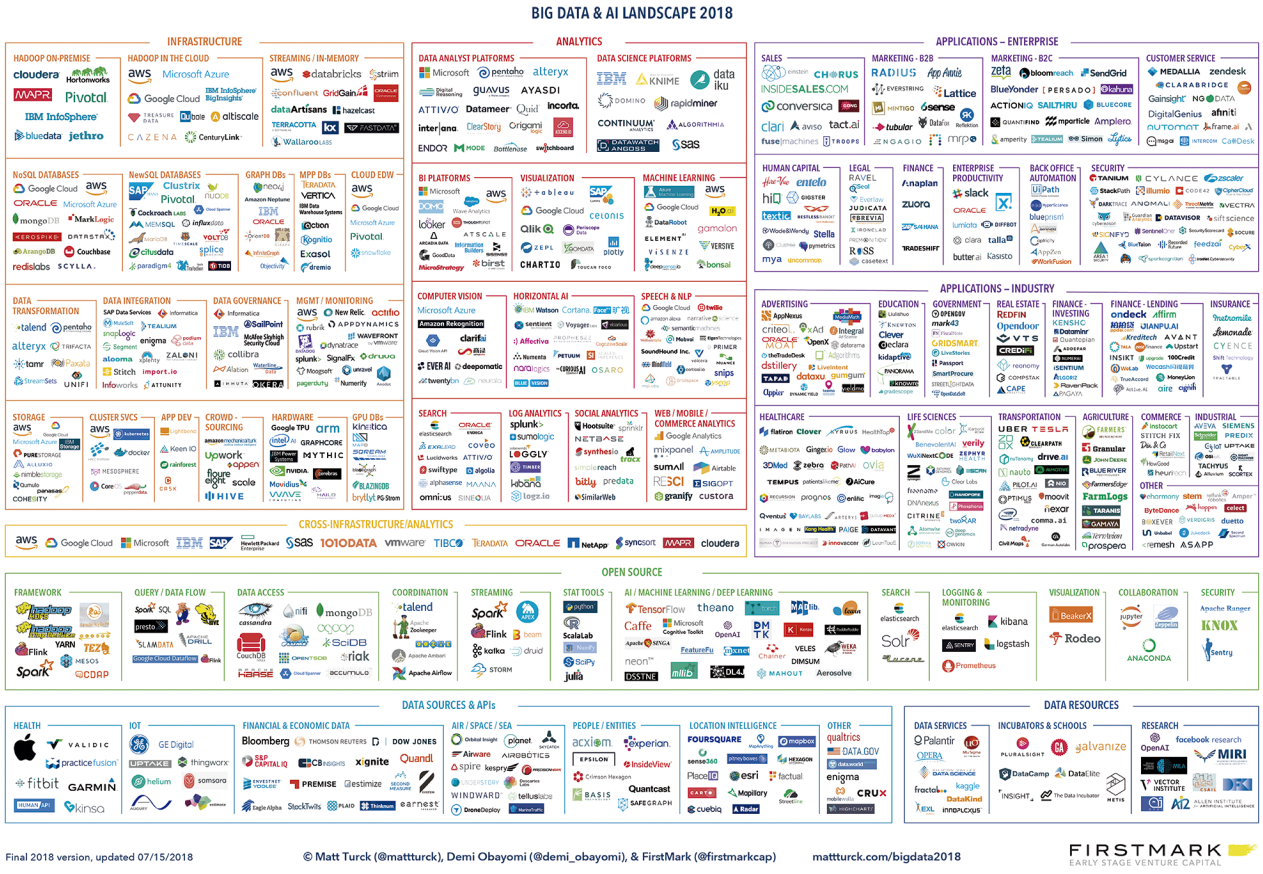
API

Correspon a *application programming interface* i no són més que un conjunt de funcions que ens permeten consultar una aplicació o un altre servei i obtenir respostes de manera directa.

(R i Python) i aprenentatge automàtic. És el que el científic de dades es trobarà de manera més habitual.

- Finalment, algunes eines destinades a l'obtenció de dades, ja siguin APIs o repositoris.

Figura 8. «Big Data Landscape» (2018)



Font: http://mattturck.com/bigdata2018/

3. Àmbits de la ciència de dades

Entesa la importància de la ciència de dades i el rol del científic de dades, és el moment de repassar breument alguns dels àmbits que s'han vist més transformats en els últims anys per l'emergència d'aquesta disciplina, així com les seves perspectives de futur.

3.1. Màrqueting

El màrqueting és un sector que necessita adaptar-se constantment al consumidor; quan els hàbits de consum canvien o evolucionen, toca seguir el mateix camí per no perdre el tren. Els últims anys han estat temps d'una marcada digitalització i aquest camp és cada dia més tècnic. Decisions que antigament es podien basar en la intuïció ara es poden, com a mínim, recolzar en dades. La ciència de dades també possibilita que activitats comercials, publicitat i campanyes siguin analitzables, mesurables i comparables. Tal és el canvi que moltes empreses ja han incorporat perfils més tècnics en els entorns de màrqueting i han afegit fermament la tecnologia de la ciència de dades com a responsabilitat dels anomenats CMO (*chief marketing officer*).

Alguns àmbits d'aplicació són:

- **Optimització de pressupostos**. Una de les tasques que més temps consumeix als responsables de màrqueting de les marques sol ser la gestió del pressupost. Canals, activitats, campanyes, descomptes, mitjans... cadascun amb un retorn estimat de la inversió, un valor esperat o la intenció de guanyar certa quota de mercat. La incorporació del científic de dades pot ajudar a modelitzar el retorn en funció de les assignacions pressupostàries, optimitzant l'ús dels preuats (i habitualment escassos) recursos dels quals es disposa, així com analitzar les suposicions i casos anteriors per establir criteris més pròxims a la realitat.

- **Segmentació**. Habitualment, les campanyes de màrqueting es dirigien al públic general, i esperen captar l'atenció del públic objectiu a partir de l'exposició global. No obstant, aprofitar l'estudi de la informació pot ajudar a segmentar els demogràfics i els espais dels quals s'obté el retorn més òptim. És possible, per exemple, comparar entre localitzacions d'un anunci, veure si una campanya funciona millor en una zona de la ciutat que en una altra, decidir emetre un anunci en una franja horària concreta d'un canal de televisió determinat o, fins i tot, fer publicitat únicament en canals digitals si es detecta que l'audiència a la qual es dirigeix l'acció rarament veu la televisió. Aquesta part requereix creuar dades internes (de re-

sultats) i externes (de quota de pantalla i audiències, pas de persones o clics, per exemple).

- **Retenció**. La ciència de dades obre la porta a conèixer (o perfilar) els clients d'una companyia pel seu comportament i no només pels seus atributs. D'aquesta manera, és possible intentar identificar aquells clients que estan a punt de deixar la companyia (aquesta és una estratègia que s'utilitza molt, per exemple, en les empreses de telecomunicacions). Així, les despeses de màrqueting per oferir descomptes als clients que es volen retenir abans que marxin són molt més eficients en la seva acció i, a la vegada, també es pot aconseguir que, amb menys contactes o trucades, els clients fidels contractin millors serveis i, per tant, siguin més rendibles per a la companyia. Parlant de contactes, la ciència de dades també s'utilitza per predir en quin moment i per quin mitjà és millor contactar amb un client. Hi pot haver, per exemple, qui vegi amb millors ulls un correu electrònic que una trucada, que pot percebre com més agressiva, mentre altres prefereixen el contacte directe. La ciència de dades pot ajudar a eliminar la prova i error que caracteritzava aquest tipus de contactes.

- **Prioritzar**. En general, per a qualsevol dels casos anteriors, la ciència de dades possibilita que els responsables de màrqueting puguin comparar i construir models temporals per veure si, per exemple, és millor contactar amb els clients que estan a punt de marxar a l'estiu i amb els que potencialment poden millorar el seus serveis a la tardor. La construcció d'aquest tipus de models pot ajudar a prioritzar, a organitzar les activitats i, en definitiva, a optimitzar el temps de tot tipus de recursos.

- **Xarxes socials**. El màrqueting actual comença a tenir més presència a les xarxes socials que en els entorns considerats tradicionals (televisió i premsa escrita). No es tracta tan sols de centrar les activitats en els entorns digitals, sinó d'entendre que cada tuit, cada missatge de Facebook i cada foto d'Instagram és una mina de dades. Utilitzant anàlisis textuals, de sentiment o, fins i tot, anàlisi de xarxes socials, és possible no tan sols saber de primera mà l'opinió dels clients, sinó anticipar futures necessitats.

3.2. Finances

El sector de les finances és un sector que tradicionalment es recolza molt en les dades, però que mai no havia accedit a una capacitat de processament com l'actual. De fet, gràcies a les grans possibilitats d'innovació i ús de la tecnologia del sector, ha aparegut fins i tot una branca paral·lela a les finances tradicionals: les anomenades *fintech*, que no són més que empreses financeres absolutament basades en la tecnologia i la ciència de dades i que aspiren a (i, en molts camps, ho aconsegueixen) competir amb els mètodes tradicionals que segueixen utilitzant les grans entitats bancàries. A més, és important destacar

l'elevada quantitat de dades de les quals disposen les entitats financeres, que les han convertit en el sector que disposa de departaments més avançats dedicats exclusivament a la ciència de dades.

En què apliquen aquestes companyies la ciència de dades?

- **Riscos**. La gestió de riscos és, probablement, l'àrea crítica de qualsevol institució financera. Convé a l'entitat deixar una quantitat determinada de diners a un individu o empresa? Té sentit aquesta inversió? Quin preu és raonable per a una assegurança amb certes característiques? Aquestes són tasques ideals per als processos d'aprenentatge automàtic, que permeten identificar, prioritzar i monitoritzar els riscos associats a les operacions habituals d'un banc. És un camp amb un potencial elevadíssim que tot just s'està començant a treballar i que requereix no tan sols la integració d'aquests processos en el nucli de la companyia, sinó també una millora de les capacitats de ciència de dades de la major part dels treballadors.

- **Gestió de dades**. S'ha destacat que el sector financer és, probablement, el que més dades genera i guarda sobre els seus usuaris. La gestió d'aquestes dades és un altre àmbit en què la ciència de dades pot ajudar, primer, per seleccionar-les i, segon, per extreure'n informació útil. Per exemple, com afecten certes notícies al comportament dels usuaris? És possible preveure si contractaran més cert producte quan hi ha un canvi polític, per exemple?

- **Predicció**. Seguint amb el punt anterior, la capacitat d'extreure informació útil de les dades històriques obre la porta utilitzar-les per preveure esdeveniments futurs. Així, és possible intentar preveure, per exemple, els moviments de la borsa (o, com a mínim, decidir en quin moment és més raonable intervenir).

- **Detecció de frau**. Potser és l'exemple més habitual. La detecció de frau no és tan sols una obligació en l'àmbit de responsabilitat de cara als usuaris, sinó que en molts casos es converteix també en una responsabilitat legal. És possible saber si una operació realitzada amb targeta de crèdit és legítima o un frau? La resposta és que sí, i que cada vegada ho és més. Així, és possible detectar tant usuaris que intenten manipular operacions com robatoris de targeta i, per tant, prevenir aquestes operacions. Per als clients és ja cada vegada més habitual observar bloquejos preventius de targetes quan apareix una operació a l'estranger o en un patró poc comú (per exemple, una gran operació d'un client que sol fer moltes petites transaccions).

- **Anàlisi de clients**. Les entitats financeres també poden barrejar tècniques de màrqueting amb productes bancaris per oferir els productes d'una manera més dirigida.

- **Inversió algorítmica**: encara que aquesta no sigui una tendència pel ciutadà mitjà, és potser la tendència que més impacte ha tingut en l'economia mundial (i de manera relativament silenciosa). Certs algoritmes (d'intel·ligència artificial) prenen decisions de compra i venda a les borses internacionals de manera constant. Sí, la immensa majoria de transaccions borsàries provenen d'un ordinador. Els *traders* i *brokers* han deixat el seu lloc a processadors que gestionen no tan sols els preus de les accions, sinó que també incorporen informació externa a l'hora de prendre decisions.

3.3. Salut

Juntament amb el màrqueting i les finances, el sector de la salut és probablement el que més ús de la ciència de dades està fent avui dia, amb la projecció de seguir avançant molt més en aquest camp considerant les possibilitats que ofereix. L'objectiu final (obviant per un moment els interessos farmacèutics i empresarials) és clar: conèixer millor el cos humà i ser capaços de salvar més vides.

Alguns exemples que estan canviant la manera d'entendre i enfocar la salut són:

- ***Wearables***. La irrupció dels dispositius portables, com rellotges o roba que incorpora sensors, permet recollir terabytes de dades sobre el funcionament diari del cos humà. Així, no es difícil imaginar el seu potencial per detectar ritmes cardíacs anòmals, per controlar riscos cardíacs o respiratoris, diabetis i, en general, per preveure possibles atacs o aturades.

- **Millora de diagnòstics**. Tot i la gran quantitat de dades i l'experiència acumulada, els diagnòstics erronis són encara freqüents (es calculen entorn del 5%, que pot semblar poc però cal veure els milions de persones que representa) i la detecció precoç és menor que la desitjada. L'ús de la ciència de dades per detectar patrons en les dades de pacients, per identificar possibles indicadors de certes malalties i per processar resultats d'anàlisis, radiografies i altres, poden millorar els diagnòstics i proporcionar informació addicional tant als metges com als seus pacients.

- **Tractaments personalitzats**. És ben sabut que no tots els pacients responen igual als mateixos tractaments o principis actius. L'ús de la tecnologia pot possibilitar l'agrupació de pacients per perfils i generar l'anomenada *medicina de precisió*, que va més enllà de l'ampli espectre i se centra en l'efectivitat.

- **Recerca farmacèutica**. Les cures, vacunes o tractaments de malalties que encara escapen de les possibilitats de la medicina (com el càncer, l'ebola o la malaltia d'Alzheimer) poden rebre ajuda del potencial de la ciència de

dades, tant per analitzar milions de casos com per proporcionar informació addicional sobre tractaments experimentals o avançaments en la cura. En casos de malalties infeccioses, a més, pot proporcionar oportunitats de facilitar el control i la neutralització de l'expansió.

- **Control de prescripcions**. Tot i que sembli estrany, encara hi ha milers de casos de prescripcions errònies, que a vegades acaben en desenllaços fatals. La prescripció també pot rebre recomanacions i alertes; si al 99% dels pacients se'ls recepta el mateix, quan s'intenti prescriure un medicament diferent pot saltar una alerta per assegurar que és una opció correcta. Succeeix el mateix si un medicament prescrit té una interacció amb un altre fàrmac que ja pren el pacient (i que pot estar prescrit per un altre metge o hospital) o si hi ha possibilitat d'al·lèrgia.

- **Reducció de costos**. De nou, com en la resta de sectors, l'ús adequat de la informació pot permetre la reducció de costos. Quants dies necessita algú estar ingressat? Quina és la dosi exacta de medicina requerida? Quin és el tractament més efectiu (o amb menys efectes secundaris que poden significar costos addicionals)? Són només algunes de les preguntes que es fa la ciència de dades aplicada a la salut.

3.4. Educació

Un altre camp en què la ciència de dades es va obrint pas és el de l'educació, batejat en aquest cas amb el nom de *learning analytics*. El pas cada vegada més comú a l'educació en línia ha permès, per una banda, l'aparició d'una gran quantitat de dades que poden ajudar a millorar l'experiència de l'estudiant i, per una altra, ha iniciat la necessitat de disposar de certes capacitats d'automatització per oferir un millor servei a les grans quantitats d'alumnes que opten per aquestes modalitats.

La major part d'aquestes iniciatives van dirigides a conèixer el comportament dels alumnes per tal d'oferir-los una experiència adaptada al seu perfil. Què poden aportar?

- **Predicció de rendiment**. Un dels beneficis que pot aportar la ciència de dades a l'aprenentatge és informació sobre el rendiment de l'estudiant; no sols en el moment actual, sinó en el futur durant el curs. Les possibilitats d'aquestes dades no són per aprovar o suspendre abans d'hora, és clar. La idea és que, davant la previsió d'un futur suspens d'un alumne, és possible proveir suport addicional a temps per evitar-ho. De la mateixa manera, és possible veure si certes activitats o materials contribueixen positivament a millorar el rendiment acadèmic i, també, l'aprenentatge.

- **Experiència personalitzada**. Mitjançant *learning analytics* es poden proporcionar i generar experiències d'aprenentatge personalitzades a cada

perfil i/o alumne. Una persona pot requerir més material per comprendre un mòdul en què ha invertit molt més temps del que és normal per a la resta o de la seva pròpia mitjana. Algú preferirà materials per a llegir, altres en vídeo. La idea és que no hi ha dues persones que aprenguin exactament de la mateixa manera, així que un sistema apropiat pot assegurar que l'experiència sigui la més òptima possible per a cada un d'ells.

- **Motivació**. Una conseqüència de l'aplicació de la predicció de rendiment i l'experiència personalitzada és un augment de la motivació (o, com a mínim, un abandonament menys elevat). Si algú no està gaudint d'una bona experiència o intueix que no podrà superar el curs, és més fàcil que decideixi deixar-ho. L'aplicació de les estratègies anteriors pot disminuir l'abandonament.

- **Iteració**. No es tracta només que l'ús de la ciència de dades en l'educació pugui beneficiar els estudiants actuals, sinó que també pot fer-ho amb els futurs. Es poden detectar materials o unitats problemàtiques (que podran ser revisats pel curs següent), metodologies que funcionen millor que d'altres, maneres i freqüències de contacte... en definitiva, una millora en eficiència i utilitat curs rere curs.

- **Reducció del cost**. Com en totes les perspectives de negoci, aquí també es parla del cost. Però en educació aquest cost no és tan sols econòmic, sinó que també hi ha un important component temporal. Tots els materials són utilitzats pels alumnes? Hi ha algun mòdul que passen per alt? És possible millorar-lo o eliminar-lo per assignar aquells recursos a un altre element amb un retorn d'aprenentatge més gran? Són preguntes que fins fa ben poc eren difícils de respondre i que, gràcies a la ciència de dades, cada dia som més a prop de millorar.

3.5. IoT

Encara que no entrarem en detall en les implicacions de l'IoT (*Internet of things*), la idea bàsica és que està relacionat amb les dades que proporcionen els sensors, que a dia d'avui són barats i estan incorporats a pràcticament qualsevol dispositiu. El seu potencial radica en la capacitat que tenen per obtenir dades del seu entorn, que a la vegada poden ser analitzades i creuades amb altres dades per tal de detectar patrons.

Els exemples relacionats amb IoT són inacabables, però a continuació se'n citen alguns:

- **Anàlisi de vídeo**. Tot i que pugui tenir una part controvertida, la idea és que és possible monitorar en vídeo (d'una càmera de vigilància, per exemple) per detectar anomalies i generar avisos de seguretat o identificar per-

sones. De la mateixa manera es pot controlar el trànsit o llegir les emocions de les persones que hi apareixen.

- **Mòbils**. Aquests dispositius intel·ligents que pràcticament tothom porta sempre a la butxaca són una font d'informació inesgotable. Les dades de geolocalització, per exemple, que poden preveure aglomeracions, comptar capacitat o identificar els patrons de moviment i circulació de les persones en una botiga o un centre comercial.

- **Ús de productes**. Semblava ciència ficció, però ja és habitual tenir neveres, rentadores i cafeteres connectades a la xarxa. Aquests dispositius poden proporcionar informació valuosa sobre hàbits d'ús (quants cafés i de quin tipus es fa un usuari cada dia), sobre consum (utilitza el programa de la rentadora adequat?) o nivells d'inventari (no hi ha llet a la nevera). El potencial d'aquesta informació creuada amb altres dades, com les de màrqueting, és il·limitat.

- **Dades de xarxes socials**. I si pensem en Twitter o Facebook com una gran font d'informació que pot ajudar, per exemple, en cas d'un desastre natural? Si un grup d'usuaris proporciona informació sobre un accident, un incendi o un altre esdeveniment a les xarxes socials, és possible creuar-ho amb altres dades provinents de sensors pròxims per tenir una anàlisis completa de la situació i saber, per exemple, quants efectius cal mobilitzar des del primer moment.

3.6. Seguretat

Un altre àmbit que és cada dia més important és el de la seguretat pública, amenaçada per, per exemple, atacs terroristes. La ciència de dades també s'utilitza per entendre les maneres que tenen aquests grups de comunicar-se, per identificar atacs potencials o grups radicalitzats i per detenir-los. Analitzant atacs anteriors i comunicacions passades és possible detectar patrons i executar accions preventives.

Per altra banda, com en totes les perspectives, però en aquest cas especialment, s'han de tenir en compte les implicacions legals i la protecció de dades.

Per exemple, és possible imaginar la predicció de crims. Un algoritme pot aprendre de múltiples perfils i determinar que un individu té el potencial de cometre un robatori. Aquesta informació, no obstant, no pot portar a la detenció d'una persona per un crim «hipotètic» a l'estil de *The Minority Report*. No obstant, cal pensar en altres opcions, com la de determinar zones, cases i hores en què és més probable que hi hagi un robatori: això permetria determinar els recorreguts de les patrulles o el nombre d'efectius requerits en cada cas i cada moment.

The Minority Report

Aquesta història curta de Philip K. Dick (de 1956) es basava en la premissa de poder detectar crims abans que es produïssin. L'adaptació al cine és *Minority Report* (2002), dirigida per Steven Spielberg i protagonitzada per Tom Cruise. Superarà la realitat a la ficció?

3.7. Altres

Podríem seguir enumerant àmbits i aplicacions meravelloses de la ciència de dades però la llista no acabaria mai. Tot i això, és convenient citar altres àmbits específics en els quals la ciència de dades ja està canviant les vides de les persones en el seu dia a dia:

- **Cerques i cercadors a Internet**. El fet que s'utilitzi Google[3] (o Bing, Ask, Duckduckgo, etc.) com si fos la cosa més natural del món no és excusa per entendre que funcionen utilitzant ciència de dades sobre unes quantitats exagerades d'informació. Sense ciència de dades valdria més tenir una agenda telefònica (de direccions web) ben àmplia.

[(3)]Google processa desenes de petabytes de dades cada dia.

- **Anuncis**. S'ha buscat informació sobre el Louvre recentment? A qui li estranya rebre una oferta d'un viatge a París, un vol o un hotel? Els anuncis que es mostren en una pàgina o al correu són diferents per a cada usuari i apareixen en funció del seu historial i de les tendències d'usuaris anteriors; els anuncis generals en línia són una espècie en extinció.

- **Sistemes de recomanació**. Has comprat un llapis? Et recomanem també que compris aquesta goma i aquesta maquineta. Que has vist una sèrie d'acció? Tenim també totes aquestes que potser que t'agradin. La recomanació ja no és un tema només de compra (compra X perquè has comprat Y abans), sinó que s'ha convertit en un tema relacionat amb l'experiència d'usuari. Si un proveïdor té milers de pel·lícules, per a l'usuari és molt beneficiós veure un filtre que mostri les pel·lícules que són més rellevants segons els seus gustos. En cas contrari, se'n podria avorrir.

- **Reconeixement d'imatge**. En aquest àmbit, les possibilitats també són infinites: reconeixement de persones a fotos pujades a les xarxes socials, obtenció d'informació sobre un quadre al fer-hi una foto, informació sobre la localització retratant el carrer on l'usuari s'ha perdut...

- **Reconeixement de veu**. Tot i que els usuaris informàtics s'hagin acostumat a l'ús del text i del teclat per interactuar amb les màquines, el mitjà que és més natural per a l'ésser humà és el llenguatge oral. El reconeixement del llenguatge natural està experimentant un creixement sense precedents, tot i que encara queda molt per davant. L'aparició i expansió dels assistents domèstics n'és tot un símptoma.

- **Videojocs**. Aquí podríem aplicar-ho a molts aspectes. Per exemple, és possible modificar l'experiència del jugador en funció dels seus hàbits. En jocs competitius en línia, es pot determinar si el jugador és més social i habitualment juga amb amics o si, en cas contrari, és un jugador solitari, i oferir serveis adequats i personalitzats. En altres jocs de llarg recorregut (*World of Warcraft*, *League of Legends*) s'utilitza l'anàlisi de dades per veure què fan els jugadors dins la partida i determinar els canvis o adaptacions necessaris

per mantenir el balanç adequat o prevenir que els subscriptors deixin el joc. I no oblidem la part competitiva, els *eSports*. Els equips professionals tenen persones dedicades únicament a analitzar les mètriques de rendiment per entendre què funciona i què no.

- **Comparació de preus**. Un usuari té un producte a la seva cistella, el compra i, minuts més tard, baixa de preu dràsticament. Aquesta situació, impensable fins ara en el comerç tradicional, és habitual entre els minoristes en línia. Darrere hi ha algoritmes que comparen el preu amb els preus dels competidors i que aprenen del comportament dels usuaris per oferir els preus més atractius en el moment adequat. Ja hi ha, fins i tot, botigues físiques que incorporen marcadors de preus variables amb aquestes estratègies.

- **Rutes aèries**. La proliferació de vols, companyies aèries i la saturació de les rutes fa que les línies aèries confiïn cada vegada més en la ciència de dades per, per exemple, decidir quants avions han de comprar, quines rutes són més eficients i preveure retards.

- **Logística**. Un dels sectors en què cada cèntim compta més és el de la logística, que busca guanyar cada segon possible i millorar l'eficiència, i en què la competència és dura. Utilitzant la ciència de dades es planifiquen rutes de repartiment, es poden preveure horaris òptims per a cada client, triar el mitjà de transport i, fins i tot, maneres d'agrupar els paquets al magatzem.

4. Conceptes de ciència de dades

Els apartats anteriors han mostrat que hi ha una gran quantitat de conceptes i noms (i que canvien i n'apareixen de nous cada dia) entorn de la ciència de dades, així que en aquest apartat intentarem aportar algunes definicions i un mapa conceptual de termes relacionats amb la disciplina. Primer, una sèrie de termes fonamentals, bàsics en la ciència de dades, seguits dels camps d'interès principals i un breu glossari sobre conceptes estadístics (que, com s'ha dit, forma la base sobre la qual se sustenta la ciència de dades). Més endavant desenvoluparem algunes notes sobre les parts del procés d'anàlisi, detalls de les tècniques d'aprenentatge automàtic i un breu resum de noms relacionats amb els programes i arquitectures que s'utilitzen més freqüentment.

4.1. Termes fonamentals

A continuació veurem els termes fonamentals de ciència de dades:

- **Algoritme**. Un algoritme és un conjunt d'instruccions que es donen a un ordinador per tal que les executi. Pot correspondre als passos per resoldre una fórmula matemàtica, per exemple.

- **Base de dades**. És el sistema d'emmagatzematge de dades, així de simple. Tradicionalment relacionals o SQL (és a dir, on la informació es guardava de manera molt concreta, organitzada i formalitzada i on primava la relació coneguda entre les dades). També n'existeixen (i cada vegada s'utilitzen més) les bases de dades NoSQL. Els programes de ciència de dades solen interactuar amb les bases de dades.

- ***Big data***. Aquest és un terme que ha perdut una mica el sentit actualment perquè s'utilitza massa freqüentment, però la idea bàsica és pensar en una quantitat de dades prou gran com perquè no sigui trivial processar-la. El *big data* es caracteritza per les quatre V problemàtiques: alt volum de dades, varietat de tipus, necessitat de comprovar la veracitat i tot a alta velocitat (tant d'arribada com de processament).

- ***Data warehouse***. Un *data warehouse* és un sistema pensat per fer anàlisis ràpides de dades de diferents fonts en un entorn empresarial. Bàsicament és una manera de fer més «fàcil» l'anàlisi als analistes de dades i que, així, no hagin de conèixer la tecnologia que suporta l'aplicació.

- **Entrenament i test**. En el procès de construcció d'un model d'aprenentatge automàtic, normalment se separen les dades disponibles

en una part d'entrenament (per construir-lo) i una de prova o test (per comprovar com funciona i que no hi hagi sobreentrenament).

- ***Front/Back end***. Així s'anomena la part visible (*front*) d'un programa (és a dir, el que veu l'usuari o client) i la part com està programat (el que hi ha al darrere, *back*).

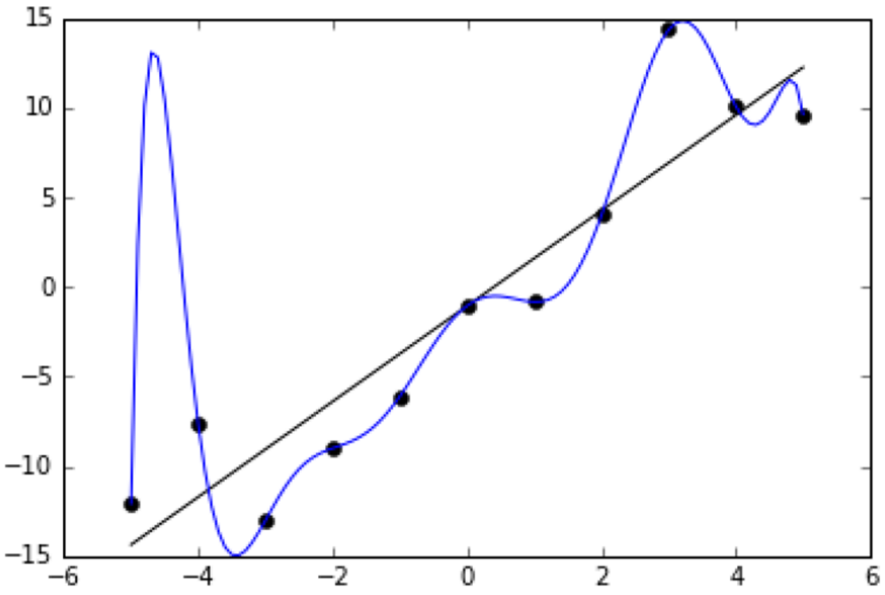
- **Lògica difusa**. És una abstracció de la lògica booleana (la dels 0 i 1) que assigna valors intermedis i que, per tant, permet que una afirmació no tan sols sigui certa o falsa, sinó que pugui ser una mica certa o pràcticament falsa, per exemple.

- ***Machine learning* o aprenentatge automàtic**. S'anomena algoritme d'aprenentatge automàtic aquell que és capaç d'obtenir informació a partir d'un conjunt de dades i fer prediccions en funció d'aquesta informació. Hi ha múltiples tècniques d'aprenentatge automàtic.

- **Regressió**. La regressió és un problema d'aprenentatge automàtic supervisat que se centra en explicar com canvia una variable numèrica en funció de la resta.

- **Sobreentrenament**. En anglès *overfitting*, és el que passa quan es proporciona informació excessiva al model, que memoritza i no aprèn. Memoritzar implica que el model obtindrà resultats excel·lents amb les dades d'entrenament, però quan s'utilitzin les dades de prova (o una predicció real) s'obtindran resultats indesitjats (figura 9). El contrari és l'*underfitting*, que passa quan el model té molt poques dades.

Figura 9. Tot i que la línia polinòmica s'ajusti perfectament als punts, la línia negra (lineal) és més generalitzable



Font: https://en.wikipedia.org/wiki/Overfitting

4.2. Camps d'interès

Els camps d'interès principals de la ciència de dades són:

- **Anàlisis de dades**. L'anàlisi de dades és com una versió reduïda de la ciència de dades, centrada a donar resposta a certes preguntes predeterminades, utilitzant l'estadística més bàsica i molta menys programació.

- ***Business intelligence***. El BI es podria resumir en com utilitzar programari divers per generar informes i trobar informació important per al negoci entre les dades. És essencialment descriptiu i se centra en les mètriques de negoci.

- ***Data engineering***. L'enginyeria de dades és la disciplina que s'encarrega de facilitar la feina del científic de dades assegurant que les dades de treball estan en el format més adequat. En equips petits, el científic de dades també s'encarrega de l'enginyeria de dades.

- ***Data journalism***. El periodisme de dades és la versió narrativa de la ciència de dades, és a dir, s'encarrega d'explicar històries de rellevància informativa mitjançant dades i basades en dades. La immensa majoria de les vegades es complementa amb una visualització de dades impactant i clara.

- ***Data science***. La ciència de dades és la disciplina que utilitza dades i estadística avançada per fer prediccions, entendre la informació i generar coneixement útil.

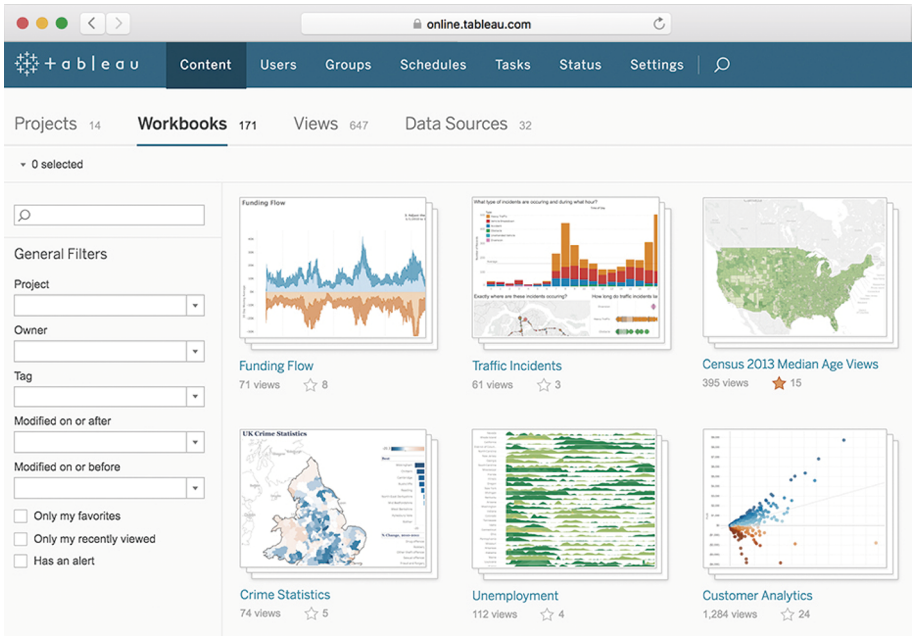
Vegeu també

Per consultar més informació sobre el *data science*, vegeu l'apartat 1.

- **Intel·ligència artificial**. La IA és la disciplina que se centra en la investigació i el desenvolupament de màquines que tenen consciència del seu entorn, que són capaces de, per exemple, resoldre una tasca concreta. Cotxes autònoms, robots mèdics o els mateixos videojocs en són alguns exemples.

- **Visualització de dades**. La visualització de dades s'ha erigit com una disciplina amb molta projecció, considerant la complexitat de comunicar de manera clara la informació obtinguda de grans volums de dades. Utilitza infogràfics, gràfics tradicionals o programari específic (Tableau, Qlik,...). (Vegeu la figura 10.)

Figura 10. Tableau permet visualitzacions molt elaborades



Font: https://www.tableau.com/

4.3. Conceptes estadístics

Els conceptes estadístics principals de la ciència de dades són:

- **Correlació**. La correlació és una mesura que indica com de relacionats estan dos conjunts de valors. Pot ser positiva (si un augmenta, l'altre també), negativa (si un augmenta, l'altre disminueix) o nul·la (quan no hi ha cap tendència).

- **Desviació estàndard**. Si la mitjana mostra el valor esperat, la desviació indica com de dispersos són els valors. Si l'elevem al quadrat obtenim la variància (σ^2).

- **Error residual**. La diferència entre el valor real i el valor calculat basat en el model obtingut és l'error residual. Si un model calcula que una persona de 170 cm hauria de pesar 70 kg però en realitat en pesa 65, l'error és de 5.

- **Estadísticament significant**. Un resultat és estadísticament significant si no es pot assegurar que és causat per un efecte aleatori.

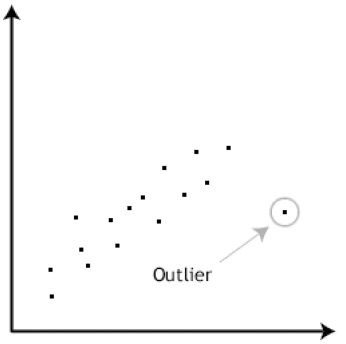
- **Mediana**. Si es posen totes les dades ordenades, la mediana és la dada que queda al mig de totes elles. Combinada amb la mitjana serveix per veure si hi ha dades anormalment grans o petites.

- **Mitjana**. La mitjana mostra el valor típic que s'espera trobar en un conjunt de dades. S'ha d'anar amb compte perquè la mitjana, per si sola, no serveix de gaire.

- **Mostra**. És el conjunt de dades a les quals es té accés i que es pretén utilitzar per extreure conclusions sobre la població (que seria «el món real»).

- **Normalitzar**. El procés de normalització és el que es duu a terme per equiparar totes les dades en un mateix rang. Molts algoritmes d'aprenentatge automàtic són sensibles al valor absolut de les dades i, per tant, sol ser necessari normalitzar l'entrada.

- ***Outlier* o data atípica**. Un *outlier* és una observació, una dada, que està exageradament lluny de la resta i que pot ser deguda a un error (una dada mal escrita) o a un punt excepcional (els sous dels futbolistes d'elit). S'han de considerar i tractar, si és necessari, en la fase inicial.



Representació gràfica d'un outlier

4.4. Processos

A continuació exposarem els processos principals de la ciència de dades:

- **ETL**. Un procés ETL és el que serveix per extreure, transformar i carregar un conjunt de dades a un *data warehouse*.

- **Exploració de dades**. És el procés inicial, en el que el científic de dades intenta entendre el context de les dades i es fa preguntes bàsiques per posar-se en situació i preparar una anàlisis completa posterior.

- **Mineria de dades o *data mining***. És un terme general que s'atribueix al procés d'extreure informació d'un conjunt de dades i utilitzar-lo. Així, inclou des de la neteja fins a l'aplicació d'algoritmes.

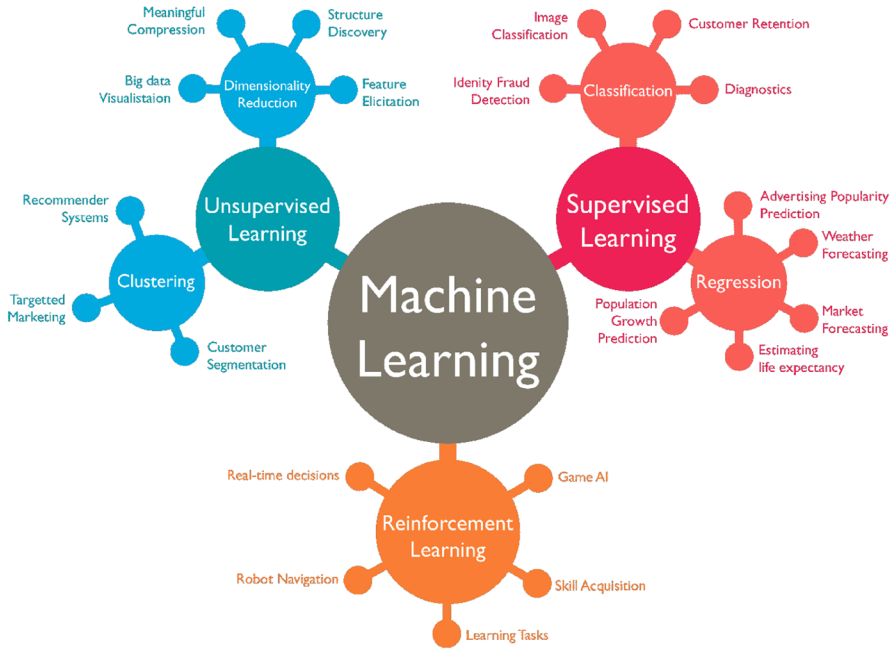
- ***Pipeline***. Un *pipeline* és un conjunt de funcions o algoritmes que s'apliquen en sèrie (un darrere l'altre, en ordre). Així, el resultat del primer serveix d'entrada al segon.

- ***Scraping web***. És la tècnica utilitzada per extreure dades del codi font d'una pàgina web i normalment requereix que el programador identifiqui les etiquetes necessàries per a l'extracció.

4.5. Tècniques d'aprenentatge automàtic

Entorn de l'aprenentatge automàtic hi ha moltes tècniques, orientades a problemes diversos. A la figura 11 se'n representen alguns.

Figura 11. Classificació dels algoritmes d'aprenentatge automàtic



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

En general, es parla de les tècniques següents: tècniques supervisades, tècniques no supervisades i tècniques de reforç.

4.5.1. Tècniques supervisades

Són aquelles en les quals el científic de dades sap quina és la variable resultat. Així, es miren de construir models que expliquin la variable final per, així, entendre millor el problema.

> Per exemple, si es vol saber quines característiques indiquen l'aparició de la miopia es tracta d'un problema supervisat: es tenen les dades dels pacients i si són o no miops. En general es poden dividir en algoritmes de regressió i de classificació.

Tècniques de regressió

Són aquelles tècniques en les quals la variable resultat és una variable «real». La més habitual és la regressió lineal.

Regressió lineal

La regressió lineal és una tècnica que busca modelar la relació entre una variable resposta (o dependent) i una sèrie de variables explicatives (o independents).

> Per exemple, si intentem calcular l'alçada d'una persona a partir del seu pes, obtindrem una relació entre ambdues que serà una recta amb una certa pendent.

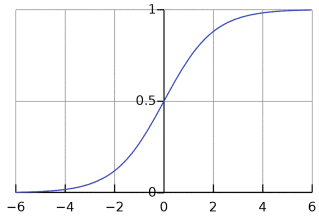
La regressió lineal simple, d'una sola variable, té la forma $y = ax + b$ i els coeficients s'obtenen de minimitzar la suma de residus al quadrat.

Tècniques de classificació

Són els mètodes supervisats en els quals la variable resposta és una categoria (per exemple, home/dona, sí/no). N'hi ha molts (i alguns de molt complexos), però els dos exemples més comuns i senzills són la regressió logística i els arbres de decisió.

Regressió logística

Els algoritmes de regressió logística intenten predir la probabilitat que passi un esdeveniment en funció de la funció logística. Així, el resultat és un nombre entre 0 i 1, que indica la probabilitat que passi l'esdeveniment concret. Per exemple, un frau amb la targeta de crèdit.

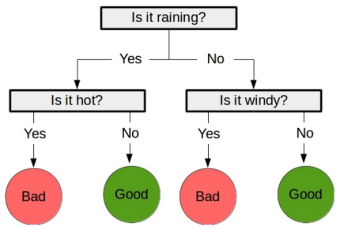


La funció logística

Arbres de decisió

Els arbres de decisió fan servir una sèrie de camins que es recorren en funció de la resposta a la pregunta de cada node i que, a les fulles, acaba amb la classificació.

Per exemple, si es vol predir la nota final d'un alumne es pot fer amb un arbre de decisió: la primera pregunta podria ser si la nota del primer parcial és aprovat o suspès. Els aprovats anirien per una branca i els suspesos per una altra.



Un arbre de decisió

4.5.2. Tècniques no supervisades

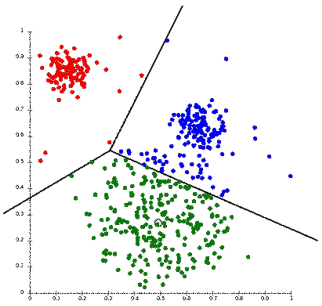
En aquest cas, la interpretació es deixa en mans de l'ordinador. Si es té un conjunt de clients i es volen agrupar en perfils, no se sap quants perfils diferents existeixen. Així, l'algoritme d'aprenentatge intentarà buscar una classificació raonable (que pot ser o no pròxima a la realitat). Les tècniques no supervisades principals són la clusterització i la reducció de la dimensionalitat.

Clusterització

Les tècniques de clusterització intenten agrupar les dades en grups que són similars o, com a mínim, propers els uns als altres. Depenen de la manera de mesurar la «distància» entre punts i la complexitat incrementa a mesura que ho fan les dimensions i el nombre de dades. Un exemple és el K-Means.

K-Means

L'algoritme de clusterització K-Means intenta dividir les observacions en K clústers, de tal manera que cada observació pertanyi al clúster més proper a la seva mitjana.



Exemple de K-Means

Reducció de la dimensionalitat

Els conjunts de dades reals solen tenir grans quantitats de variables; així, moltes vegades és necessari (o adequat) reduir-ne el nombre. Normalment es fa o bé escollint les variables més relevants (selecció de variables) o bé obtenint una descomposició en variables principals (anàlisi de components principals).

Selecció de variables

Aquest procés consisteix a mesurar la rellevància de cada una de les variables en la predicció del resultat final i seleccionar posteriorment les més adequades per obtenir un model de rendiment el més òptim possible.

Anàlisi de components principals

L'anàlisi de components principals és una mica més complex d'entendre, però bàsicament busca transformar les variables en una descomposició ortogonal i sense correlació entre elles. Tot i la seva utilitat exploratòria, l'explicació posterior del model es converteix en molt més abstracta.

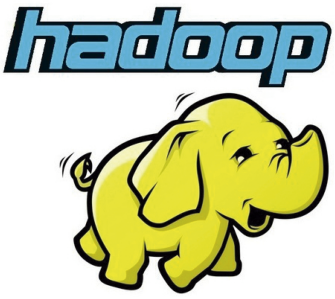
4.5.3. Tècniques de reforç

Aquesta àrea s'encarrega d'intentar determinar quines accions ha d'escollir un agent en un entorn donat per tal de maximitzar una recompensa o un premi. Són les tècniques que s'utilitzen, per exemple, per entrenar els sistemes que juguen a escacs o a Go.

4.6. Programari

En aquest subapartat presentem una llista de noms propis perquè sonin a l'estudiant per a futures matèries:

1) Hadoop. Hadoop és un marc *open source* de processament distribuït que s'utilitza per treballar amb grans quantitats de dades. Marca un abans i un després en la ciència de dades avançada. Permet utilitzar processament en paral·lel entre diverses màquines (anomenats clústers).

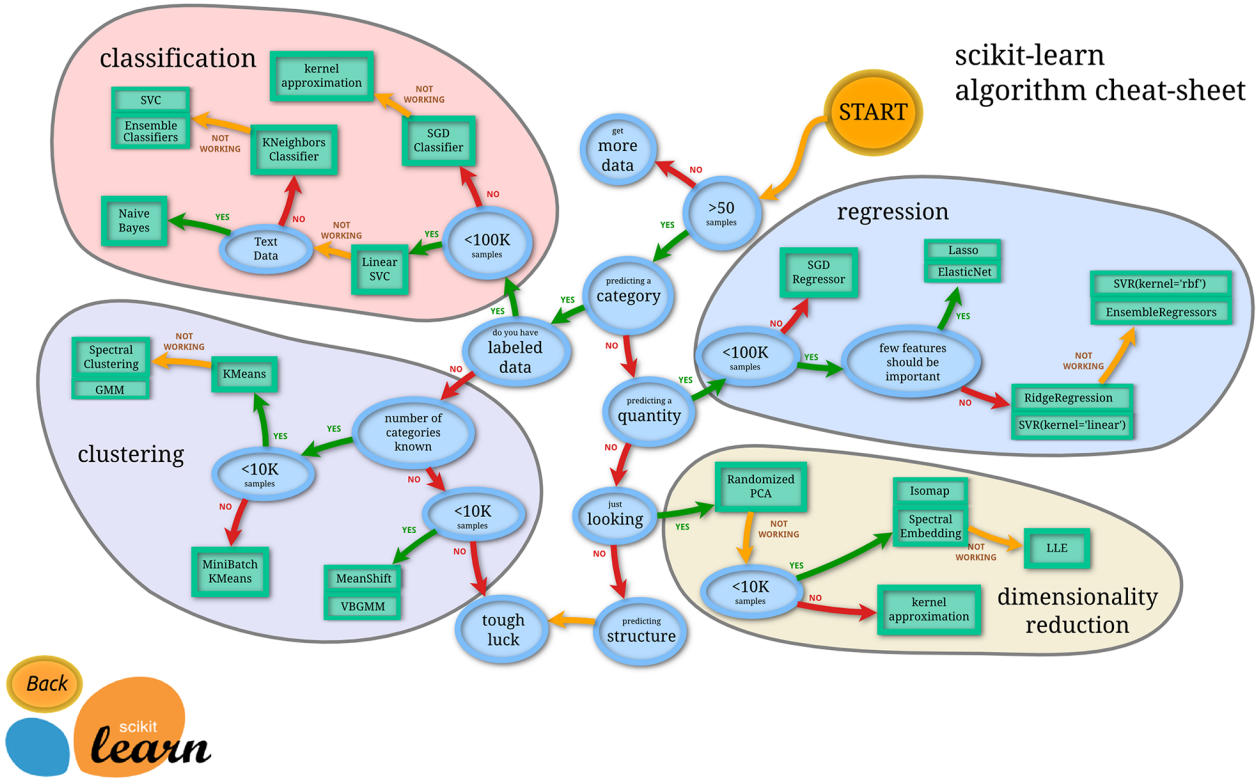


L'elefant de Hadoop

2) Python. Python és un llenguatge de programació *open source*, utilitzat en moltes aplicacions, com la programació en general, la ciència de dades i l'aprenentatge automàtic. Es considera fàcil d'aprendre, d'alt nivell i té una comunitat molt activa. Algunes de les llibreries més populars de ciència de

dades són per a Python, com Scikit-learn (aprenentatge automàtic, vegeu la figura 12), NTLK (processament del llenguatge natural) o NetworkX (anàlisis de xarxes socials).

Figura 12. Arbre de decisió per decidir quin algoritme utilitzar a Scikit-learn



Font: http://scikit-learn.org

3) R. R és tant un llenguatge com un entorn *open source* orientat a la computació estadística. És molt extensible (la comunitat és molt activa) i té implementades la immensa majoria de tècniques existents. R es gratuït i en l'àmbit acadèmic i de desenvolupament se sol preferir a les alternatives de pagament (SPSS, SAS), precisament perquè permet veure, controlar i modificar o adaptar els algoritmes que incorpora.

4) Spark. Apache Spark és un altre marc *open source* de processament distribuït que s'utilitza per treballar amb grans quantitats de dades, però que dona molta més flexibilitat que Hadoop. Permet utilitzar Java, Python, Scala i R i suporta SQL, reproducció en continu i algoritmes d'aprenentatge automàtic.

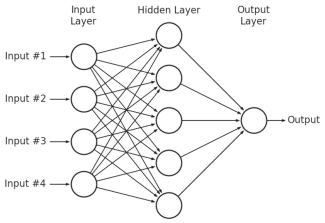
Apache Spark

4.7. Altres conceptes

A continuació explicarem els conceptes d'aprenentatge profund i xarxes neuronal, *open data*, *open source* i sistemes de recomanació.

4.7.1. Aprenentatge profund i xarxes neuronals

Les xarxes neuronals pertanyen a l'aprenentatge automàtic, però les seves característiques especials fan que destaquin. Es basen (lliurement) en com funcionen les connexions de les neurones en el cervell i bàsicament són un conjunt de nodes organitzats per capes que s'entrenen per fer prediccions. L'anomenat aprenentatge profund (*deep learning*) no és més que l'extensió a xarxes neuronals molt grans, com les que s'utilitzen per identificar cares o imatges.



Una xarxa neuronal

4.7.2. *Open data*

Les dades obertes són aquelles que són lliures i qualsevol persona les pot extreure i utilitzar com vulgui, sense drets d'autor, patents o mecanismes de control. Alguns ajuntaments, com els de Barcelona o Madrid, proporcionen dades obertes que permeten que qualsevol ciutadà consulti dades de transport o de qualitat de l'aire, per exemple.

4.7.3. *Open source*

Les eines de codi obert són aquelles que permeten accedir i editar el seu codi font i que, per tant, els usuaris poden modificar. No s'ha de confondre amb gratuït, tot i que la majoria de vegades ho siguin. De fet, moltes eines de codi obert presenten, per exemple, dificultats en la configuració (com Hadoop) i certes empreses ofereixen paquets, de pagament, configurats i als quals donen suport tècnic.

4.7.4. Sistemes de recomanació

Són sistemes d'aprenentatge automàtic que se situen entre la regressió i la classificació i que utilitzen la informació per recomanar elements que poden ser d'interès a l'usuari. Per exemple, productes relacionats amb compres anteriors o pel·lícules basades en els gustos i les puntuacions d'usuaris que tenen un perfil similar.

Bibliografia

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

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

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

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

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