Xavier Duran

sense fronteres

L'imperi de les dades

El 'big data': oportunitats i amenaces

Premi Europeu de Divulgació Científica Estudi General

INTRODUCCIÓ

JO SÓC JO I LES MEVES DADES

Dades, dades, dades a dojo... Vivim en un món de dades, emmagatzemades en formes diverses: textos, nombres, imatges, gràfics... «Jo sóc jo, les meves circumstàncies... i les meves dades», diria avui el filòsof Ortega y Gasset.

N'hi ha tantes i tantes que ja no parlem de dades, sinó de *big data*, grans dades. El concepte ha fet fortuna i malgrat que sovint es deixa en anglès, també s'adapta a cada idioma. En català parlem de megadades o de dades massives. Farem servir preferentment *big data*, però també utilitzarem indistintament les dues traduccions. La idea sempre és que hi ha moltes dades.

Big data ens fa pensar en arxius digitals i en consultes per Internet. Pensem en Google i les muntanyes d'informació per on es deu moure aquest cercador per buscar el que li demanem. I potser en Facebook i en Instagram. Però, com anirem explicant al llarg del llibre, sembla que res no queda al marge d'enriquir el *big data*: ni missatges privats per WhatsApp, ni trucades telefòniques, ni compres amb targeta, ni, fins i tot, passejades amb el mòbil engegat. Però hi ha moltes més fonts de dades: les que envien els satèl·lits, les que proporcionen sensors escampats per les ciutats, pel camp o pels oceans, les imatges de càmeres de seguretat, les dades que proporcionen aparells mèdics o els anomenats

wearables –una mena de captadors de dades portables, que poden ser un braçalet o una peça de roba, com una samarreta.

Ja va escriure el filòsof anglès Francis Bacon, al final del segle XVI, que «coneixement és poder». Però dades i coneixement no és el mateix. De fet, fins i tot hi ha un pas intermedi, que és la informació. Confonem dades amb informació i són coses diferents. Un grup de músics tocant pel seu compte, per afinadament que ho facin i per virtuosos que siguin, són dades. Tots tocant en harmonia sota les ordres d'un director d'orquestra és informació.

Les dades són el combustible que permet resoldre problemes –de vegades, creats per les mateixes dades. Però un combustible sol no fa res. Les dades serveixen per fer funcionar la maquinària que busca les respostes als problemes. Per això, les dades són imprescindibles, però sense una estratègia per tractar-les i transformar-les no tindrem mai informació. I un cop reunida prou informació, encara ens quedarà analitzar-la, reflexionar. De la manera com ens en servim per produir coneixement dependrà la qualitat d'aquest.

Tot i així, no podem negar que, avui, les dades són poder. Hi ha qui en diu «el petroli del segle XXI». Tornem a la metàfora del combustible, però en aquest cas per alimentar màquines de fer diners –i de construir poder. Qui té moltes dades té molt de poder, si sap com utilitzar-les... o si les ven a algú que li interessi fer-ho.

Aquí explicarem d'on sorgeixen tantes dades, com circulen, com es guarden. Mostrarem com es processen –cosa que es pot fer bé o molt malament. I descriurem els beneficis que comporten i els riscos que representen. Molts possibles beneficis i molts possibles riscos. Alguns ja són

palpables –tant les derivacions positives com els perills– i altres estan a punt d'arribar, encara que semblin fantasies de pel·lícula de sèrie B.

En definitiva, proporcionarem al lector moltes dades, transformades en informació i amb l'esperança que generin coneixement. No sabem si la nostra aportació serà valuosa, però no tenim cap dubte que intentar-ho és necessari. Hi pot haver dades sense informació, però difícilment hi haurà informació sense dades. I encara menys, coneixement. Perquè l'imperi de les dades no ens engoleixi, cal estar mitjanament preparats. Només si els ciutadans tenen prou dades i les saben processar podran pressionar perquè la informació i el coneixement que se'n derivin siguin beneficiosos per a la societat.

1
VIATGE AL PAÍS DE LES DADES

> Hi havia 5 exabytes d'informació creats des de l'alba de la civilització fins al 2003, però aquesta informació ara es genera cada dos dies.
>
> Eric Schmidt (2010)

> El món ja no està dominat per les armes, ni per l'energia, ni pels diners. Està dominat per uns i zeros, per petits bits de dades. Tot està en els electrons.
>
> Cosmo, personatge de la pel·lícula *The sneakers* ('Els cibertafaners'), 1992

> Al llarg del segle xx han tingut gran repercussió tres conceptes científics profundament desestabilitzadors que l'han dividit en tres parts desiguals: l'àtom, el bit i el gen. [...] Cada un té el seu origen en una noció científica abstracta, però creix fins a acabar envaint un gran nombre de disciplines humanes i transformant la cultura, la societat, la política i el llenguatge.
>
> Siddhartha Mukherjee

Fremont Rider va aixecar la vista per contemplar els prestatges plens de llibres, va sospirar i immediatament va pensar en un futur més aviat negre o, si més no, molt complex. Rider era escriptor i bibliotecari de la Universitat Wesleyana a

Middleton (Connecticut, Estats Units) i l'any 1944 va llançar un crit d'alarma sobre el nombre de llibres que es publicaven anualment. Va calcular que les biblioteques nord-americanes duplicaven la seva mida cada setze anys. Segons Rider, a aquest ritme, la biblioteca de la Universitat de Yale, una de les principals del país, tindria, l'any 2040, «aproximadament 200.000.000 de volums, que ocuparien 9.656 quilòmetres de prestatges». El problema no seria només d'espai, sinó de gestió. Rider calculava que aquesta quantitat de llibres faria necessari un equip de més de sis mil persones per a catalogar-los.

Més de set dècades després de l'avís de Rider, el problema ja no són tant els llibres editats, sinó el conjunt de la informació. Internet ha provocat una explosió de dades. Només amb les que processa cada dia Google es podrien editar prou volums perquè, apilonats, s'estenguessin fins a meitat de camí entre la Terra i la Lluna. Potser Rider ni tan sols tindria esma de calcular quant de personal caldria per a catalogar-los –una senzilla regla de tres fent servir les dades del bibliotecari americà revela que serien més de 108.000 persones.

Per sort, aquestes dades no es troben en paper, sinó que més del 90% estan en suport digital. Per desgràcia, no hem de considerar només les cerques a Google, sinó tot allò que es genera en l'univers digital en diferents formats.

De tant en tant, algú fa càlculs semblants als de Rider, però ja no es poden limitar al paper. A més, solen quedar obsolets al cap de poc temps. El 1997, Michael Lesk, un informàtic i expert en sistemes d'informació, es va entretenir a calcular quanta informació hi havia en el món. Començava descrivint la Biblioteca del Congrés a Washington, amb els seus 20 milions de llibres, 13 milions de fotografies, 4 milions

de mapes, més de mig milió de pel·lícules i tres milions i mig d'enregistraments de so.

Però Lesk no es podia limitar a una biblioteca, per gran que fos, ni tan sols al material editat. Hi afegia que en un any es filmaven milers de pel·lícules, es prenien milers de milions de fotografies, s'emetien milions d'hores de televisió i de ràdio, s'editaven més de 400 milions de CD i més de 300 milions de cassets –molts duplicats, sens dubte, perquè d'alguns se'n fan milers de còpies–, hi havia bilions de minuts de converses telefòniques... Fent càlculs aproximats i basant-se en algunes altres fonts, assenyalava que potser al món hi havia 12.000 petabytes (PB) d'informació. Això significa 12 milions de gigues, per dir una unitat de mesura que a molta gent li resulta familiar.

Tot i aquestes xifres, concloïa que a la Terra hi hauria prou capacitat d'emmagatzematge per tot allò que la gent escrivís, digués, fotografiés o filmés en el futur.

De tot això fa vint anys i la quantitat d'informació ha augmentat de forma exponencial. I sembla que sí, que la tecnologia, almenys de moment, està solucionant el problema de guardar-la i fins i tot de fer-la accessible. Però quina utilitat pot tenir tanta informació? I com podem gestionar-la?

El naixement de les dades

Les dades neixen de la necessitat. Hi va haver dades abans que hi hagués mètodes per representar-les de forma comprensible per a tothom. Primer van ser les dades i, temps després, van aparèixer els nombres. Fa milers d'anys, un pastor veia que

del seu corral sortien moltes ovelles i que després de pasturar n'hi entraven moltes. Però, com podia saber si tornaven totes?

Per estar segur que no perdia cap ovella devia prendre una pedra o una branqueta per cada una que sortia de la cleda. I quan, després de pasturar, hi tornaven a entrar, devia enretirar una pedra o branqueta de la pila per cada una. Si no en quedava cap, totes havien tornat. Si quedaven pedres a la pila, alguna s'havia escapat. I si seguien arribant ovelles i ja havia acabat les pedres i les branquetes, o bé s'havia descomptat, o bé havia guanyat algun exemplar extra.

Després vindrien els sistemes per a simbolitzar les quantitats. Les societats evolucionaven, es feien complexes. Hi havia més producció agrícola i hi havia més bestiar. I es feien intercanvis comercials. Així van néixer els nombres. No els nombres actuals, sinó altres sistemes simbòlics per representar quantitats. Fa més de cinc mil anys ja hi havia fitxes d'argila amb símbols que corresponien a quantitats i fins i tot a càlculs.

Però la informació, les dades, no eren simplement numèriques. Hi havia textos, hi havia representacions simbòliques, hi havia gravats. Estalviem-nos uns quants mil·lennis i saltem al segle xv. Amb la impremta, la informació editada amb llibres i documents esclata i hi ha qui hi veu una allau difícil de gestionar. Els primers escèptics sobre la capacitat humana per a assimilar tants llibres no van poder veure com qualsevol de les seves previsions quedava curta en pocs decennis.

Tornem a fer un gran salt. A mitjan segle xx, la quantitat d'informació era immensa i a algú se li va acudir que hi havia d'haver alguna manera de quantificar-la. El 1948,

el nord-americà John W. Tukey, matemàtic i pioner de la informàtica, creà el bit, com a abreviatura de BInary digiT. A part de la contracció del concepte en tres lletres, devia jugar amb el significat de *bit* en anglès, peça petita. Ja teníem la unitat d'informació digital.

Al cap de pocs anys, el 1956, l'enginyer electrònic Werner Buchholz –nord-americà nascut a Alemanya, d'on va marxar escapant del nazisme– creà el byte. Als anys cinquanta, Buchholz treballava a la IBM i va formar part de l'equip que va dissenyar els primers ordinadors, com l'*IBM 701*. El bit era massa petit per mesurar la quantitat mínima d'informació, un sol caràcter, i per això va sorgir el byte. Al principi no hi havia una equivalència estàndard i un byte, segons el sistema o l'ordinador utilitzats, podia variar. Ara, un byte equival a 8 bits i per això també en diem octet.

Ja tenim el byte, però tot i la necessitat de definir la unitat que equival a un sol caràcter, una mesura tan petita té poca utilitat quan parlem de grans quantitats d'informació. Seria com mesurar distàncies astronòmiques en centímetres. Per això de seguida van sorgir els múltiples: quilobyte, megabyte... Però mega, un milió, es queda curt en molts casos i per això van aparèixer el giga (mil milions) i d'altres que progressivament multipliquen l'anterior per mil: tera, peta, exa, zetta, yotta... Amb aquest darrer arribem al quadrilió.

Explicàvem abans que Lesk havia situat en 12.000 petabytes la quantitat d'informació que hi havia en el món el 1997. Però amb aquestes xifres, a molta gent li passa com amb els pressupostos estatals o amb els beneficis de les grans empreses. Ens poden parlar de 17.000 milions d'euros, de

80.000 milions o de 250.000 milions. Comprenem que és moltíssim, però som incapaços de fer-nos-en una idea.

Per això, algunes comparacions seran útils. Un byte és un sol caràcter. Per tant, una sola lletra ocupa un byte. Si creem un document amb una sola lletra, *pesarà* 1 byte. A partir d'aquí, el primer pas no és difícil. Un quilobyte (KB) equival a mitja pàgina, uns mil caràcters. I un megabyte podria ser una novel·la curta.

Fem un breu incís. Sovint llegim que un KB són 1.024 bytes. Això es deu a l'origen del byte i al fet que els informàtics treballen en sistema binari i, per tant, amb potències de 2. Com que 2^{10} és 1.024, aquesta és l'equivalència que es fa servir sovint en l'àmbit dels ordinadors. Però per al sistema internacional de mesures, un KB són mil bytes.

Però la informació no està només en forma de text o de xifres. Podem tenir gràfics, dibuixos, fotografies... Fins i tot pel·lícules o sons. Cada afegit augmenta la quantitat d'informació. Una fotografia amb bona definició pot ocupar 2 megabytes. És a dir, com dues novel·les curtes.

Una filera de deu metres de llibres equival a 1 gigabyte (GB). I amb sis milions de llibres tindríem un terabyte (TB). Si reuníssim 7 milions d'hores de televisió d'alta definició tindríem un petabyte (PB). I Lesk deia que tota la informació que hi havia al món ocupava 12.000 petabytes! Avui, en només una hora ja es transmeten a tot el món 500 petabytes d'informació, equivalents a 6.600 anys de vídeo d'alta definició o a deu vegades totes les obres escrites per la humanitat des dels inicis de la història.

Totes aquestes comparacions són aproximades. La quantitat de bytes que té un text també depèn de les ordres que

hi incloem –format, estil i mida de lletra.... Una fotografia pot tenir molta qualitat o ben poca i el mateix passa amb una pel·lícula. D'altra banda, es fan comparacions amb coses ben difícils de mesurar amb exactitud. Així, s'ha dit que totes les paraules pronunciades per tota la humanitat al llarg de la història ocuparien 5 exabytes (EB). La idea també ha estat rebatuda i nous càlculs parlen de 42 zettabytes (ZB). Però és ben probable que ens faltin molts elements per poder valorar-la amb precisió.

Unitats d'informació i les seves equivalències

(Cada una multiplica por mil l'anterior)

1 byte	1 caràcter.
1 kilobyte (KB)	Mitja pàgina mecanoscrita.
1 megabyte (MB)	Una novel·la curta.
1 gigabyte (GB)	Una pel·lícula de dues hores.
1 terabyte (TB)	6 milions de llibres.
1 petabyte (PB)	2.000 anys seguits de música.
1 exabyte (EB)	100.000 vegades tot el material imprès –llibres, revistes, documents– de la Biblioteca del Congrés de Washington.
1 zettabyte (ZB)	152 milions d'anys de vídeo d'alta definició.
1 yottabyte (YB)	Tota la informació que pot contenir el centre de dades de l'NSA (National Security Agency) dels Estats Units a Utah, que té una superfície de 92.000 metres quadrats.