# Big Data i estadística oficial: oportunitat o amenaça?

# (Yier Wu, Yang Chen, David Serfaty i David Graupere)

## Introducció

El tema que hem tractat en aquest estudi és el famós concepte/fenomen del <Big Data>. Hem escollit aquest tema perquè ens sembla de gran interès i útil per a la carrera que estem estudiant, ja que, molts diuen que el <Big data> és el futur de l’estadística.

Comencem l’estudi amb una introducció i presentació de que és el BIG data, intentem explicar i detallar algunes característiques més importants i significatives. I veurem les aplicacions que té per a les empreses. Posteriorment ho hem relacionat amb l’estadística pública, l’ètica professional i la presentació d’un projecte de l’IDESCAT que malda per fer l’empelt del <big data> en l’estadística pública per al ‘bé comú’, concepte que ja s’explicarà.

## Generació de Big Data

Big Data és un concepte que es va popularitzar el 2011, que va donar nom a un conjunt de dades que superen la capacitat del software habitual per ser capturats, gestionats i processats en un temps raonable.

El Big Data és compost per 5 aspectes bàsics que són:

* Volum: el valor i la informació que pot tenir les dades depèn del volum.
* Varietat: gran varietat de tipus de dades.
* Velocitat: velocitat d’obtenció de dades.
* Veracitat: qualitat de les dades obtingudes.
* Valor: l’ús adequat del Big Data comporta una reducció dels costos.

I hi ha d’altres aspectes que podem comentar: <*variability>*, les possibles variacions que poden tenir algunes dades poden influir en el procediment de tractar-les. Un segon aspecte, <*Complexity>*, un gran volum de dades que provenen de diferents fonts.

## *Data mining*: D’on vénen les dades?

*<Data mining>* o mineria de dades és un procés no trivial d'identificació d’informació vàlida, nova, potencialment útil i entenedora, de patrons comprensibles que es troben ocults a dins de les dades. Engloba tot un conjunt de tècniques encaminades a l'extracció de coneixement processable, implícit en les [bases de dades](https://ca.wikipedia.org/wiki/Bases_de_dades). És íntimament lligat a la supervisió de processos industrials, perquè resulta molt útil per aprofitar les dades emmagatzemades a les bases de dades. Les bases de la mineria de dades es troben a **la** [**intel·ligència artificial**](https://ca.wikipedia.org/wiki/Intel%C2%B7lig%C3%A8ncia_artificial) **i en l'**[**anàlisi estadística**](https://ca.wikipedia.org/wiki/Estad%C3%ADstica).

Les eines de la mineria de dades ajuden a extreure la informació interessant en arxius corporatius o en registres públics, arxivats; es combinen fàcilment i poden analitzar-se i processar-se ràpidament. La mineria de dades produeix cinc tipus d'informació:

* Associacions.
* Seqüències.
* Classificacions.
* Agrupaments.
* Pronòstics.

Procés de funcionament de la mineria de dades:

Les tres tecnologies més importants de *Data Mining* són: la recol·lecció massiva de dades, les computadores potents amb multiprocessadors i l’Algorismes de *Data Mining.*

## *Data analytic:* Com analitzar les dades?

Anàlisi de dades és el procés d’examinar un conjunt de dades per descobrir patrons ocults, correlacions desconegudes, tendències del mercat, les preferències dels clients i... més informació útil.

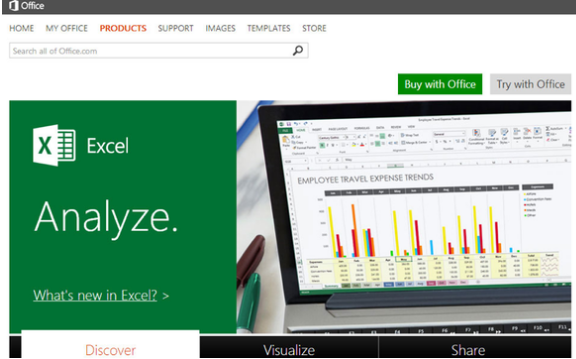
Les fonts de dades no estructurades no encaixen en els magatzems de les dades tradicionals. A més a més, els magatzems de les dades tradicionals poden no ser capaços de manejar les demandes del procediment de big data. Com a resultat, sorgeix una nova classe de tecnologia de big data utilitzada en anàlisi de big data. Les tecnologies relacionades amb el anàlisi de dades inclouen bases de dades grans NoSQL, Hadoop o MapReduce. Aquestes tecnologies formen el nucli de marc de software que suporta el procediment d’anàlisi de big data.

5 aspectes basics per comentar en anàlisi de big data:

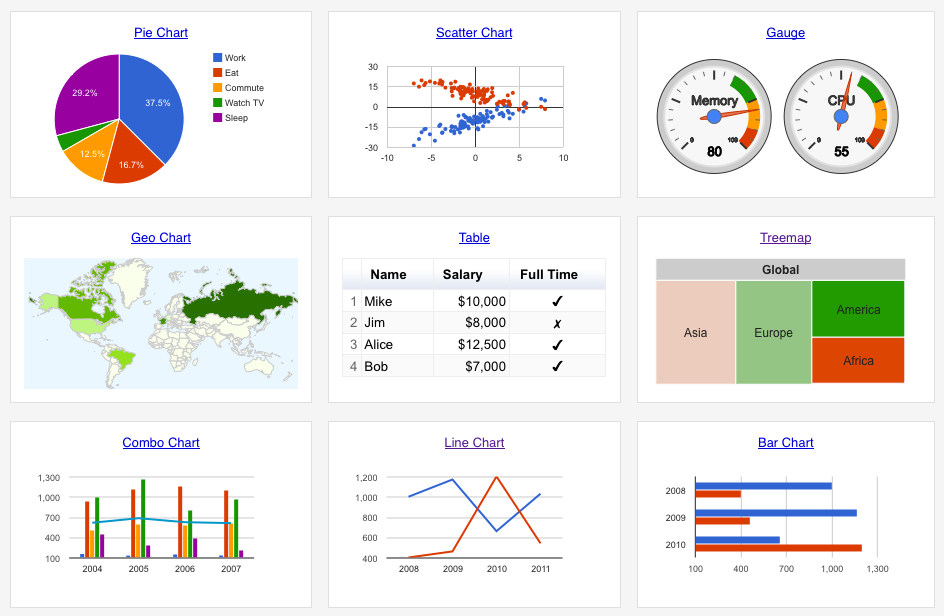
1. Visualitzacions Analítiques

L'objectiu principal de la visualització de dades és comunicar informació clarament i de forma eficient a través de gràfiques estadístiques i gràfiques d'informació.

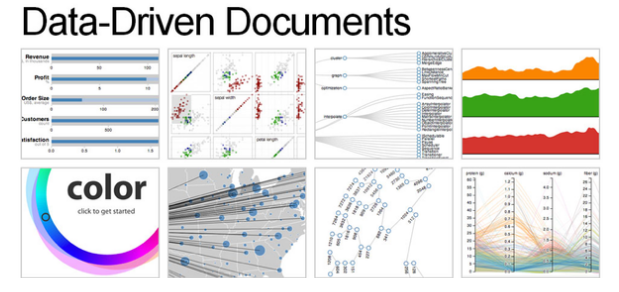
Ex: EXCEL



google Chart API



D3 ( Data Driven Documents)

Chartng Fonts



Carto DB ( relaciona les dades amb la mapa)



etc.

1. Algorismes de Mineria de Dades

Un algoritme de <data mining> és un conjunt d’algorismes i càlculs que permeten crear un model a partir de dades. Per a crear un model, l’algorisme analitza primer les dades proporcionades, en cerca de tipus específics de relacions o tendències. L’algorisme usa els resultats d’anàlisis en moltes iteracions per a determinar els paràmetres òptims per a crear el model. Després aquest model s’aplica en tot el conjunt de dades.

Tipus d’algorismes : algorismes de classificació , algorismes de regressió , algorismes de segmentació , algorismes d’associació , algorismes d’anàlisi secundari

1. Capacitats Analítiques Predictives

L’ anàlisi predictiu agrupa una varietat de tècniques estadístiques de modelització, aprenentatges automàtics i mineria de dades que analitza les dades actuals i històriques reals per fer predicció del futur o esdeveniments desconeguts.

1. Motors Semàntics

Es una procés per extreure informacions dels documents de manera intel·ligent.

1. La qualitat i la gestió de les dades

Les dades han de ser d’alta qualitat i ben gestionades abans de ser analitzades. Amb l’ajuda de les eines i els processos estandarditzats, es garanteix que la qualitat de les dades es mantingui alta.

## *Data application:* Us del Big Data

El <Big data> és un concepte relativament nou, però ja famós entre molts professionals. Això és per causa del seu gran potencial en molts sectors, com per exemple, la industria, el comerç, els serveis, etc. Per això, el <Big data> també té una aplicació molt àmplia, citarem les 5 més importants a l’actualitat:

* L’exploració de gran volum de dades.

El tractament i l’accessibilitat a volum massives de dades és molt interessant per a les grans empreses, perquè els ajuda a prendre decisions, millorar les operacions i reduir el risc.

* Visió general sobre el client

Les empreses tenen com a objectiu, comprendre el comportament del client i predir les seves accions, per tant, els cal obtenir informació de fonts internes i externes. Denominem fonts internes, tota la informació que el client pot deixar a l’empresa (les cerques a la pàgina web, el tractament personal amb un empleat, etc). I denominem fonts externes les que provenen de xarxes socials, correus electrònics, etc.

* Extensió de la seguretat intel·ligent

L’interès en localitzar, solucionar i prevenir possibles anomalies i atacs en el sistema o pàgina web de l’empresa. Aquests mecanismes permeten trobar entre el volum massiu de dades possibles relacions ocultes o patrons de conducta que amenacin la seva seguretat.

* Anàlisi d’operacions

Obtenir una visibilitat en temps real de les operacions, experiència, transaccions i comportament del client; permet que l’empresa incrementi l’eficiència de les operacions. I veure’n d’aquesta manera el rendiment.

* Augmenta l’emmagatzematge de dades i <*Data Warehouse>*

Ampliar l’estructura d’emmagatzematge de dades ja existent aplicant els avantatges del <Big Data> per incrementar el seu valor. Augmentant el <*Data Warehouse>* de l’empresa, s’ aprofitaran els diferents tipus de dades per guanyar noves perspectives de l’empresa en temps real, optimitzar l’estructura d’emmagatzematge de dades per facilitar el treball tot reduint els costos.

## La tendència del desenvolupament del Big Data

Les tendències en futur del big data seran :

* Les dades són els recursos principals: el <Big data> serà el punt focus de totes les empreses i les societats.
* Big data i computació al núvol seran inseparables: la computació en el núvol és la plataforma bàsica del big data.
* Trencarà la teoria científica: amb l’augment de la mineria de dades, l’aprenentatge automàtic i les tecnologies d’intel·ligència artificial podrien canviar les teories i els algorismes tradicionals en el món de les dades.
* S’establiran més d’estudis relacionats amb la ciència de les dades: en el futur, la ciència de les dades serà una disciplina específica. Hi haurà moltes universitats que crearan una carrera relacionada amb les dades.
* Un efecte de fuita de dades: la taxa de creixement de fuga de dades serà del 100%, és a dir, necessiten proposar més mitjans per protegir tota aquesta massa de dades.
* La competència principal entre les empreses serà la gestió de dades
* La qualitat de les dades serà el punt clau de BI ( intel·ligència empresarial)
* Enfocament del sistema del big data: el món de <big data> no és un únic concepte, sinó que és compost per un sistema de diversos elements. Actualment, el sistema s’acaba de formar. En un futur, es desenvoluparan components interns que enfortiran més el sistema.

## Els efectes del Big Data

Els beneficis que porta la utilització màxima del <Big Data> són:

* Avantatges competitius en la gran distribució: permet actualitzar, optimitzar i afinar inventaris en funció de la demanda en temps real. L’anàlisi de les dades extretes del <*data mining>* ens facilita el comportament dels clients, i per tant poder oferir millor productes per complir les expectatives del client.
* Millora de l’eficiència i els costos: l’anàlisi del <Big Data> pot accelerar la velocitat en què es desenvolupa un producte, en alguns sectors pot arribar fins i tot a reduir un 50% el temps de desenvolupament d’un producte.
* Millora en la gestió empresarial: ajuda a prendre decisions importants basant-se en dades.
* Emmagatzematge en el núvol: el gran volum de dades no comporta dificultats d’emmagatzematge.

Però, també hem de pensar els possibles inconvenients. Per exemple, ens podem assegurar que la gran quantitat de mostres triades és representativa? Si no ho és, ens portaria a uns resultats esbiaixats.

En conclusió: podem dir que <Big Data> aporta a les empreses una millora de les seves accions en temps real.

## El <Big Data> com a bé social?

El big data sembla una gran oportunitat de negoci per a les empreses. El big data ha nascut per iniciativa de l’empresa privada tot i que va ser afavorit per l’aparició d’internet que va ser una iniciativa pública. És bo que la iniciativa privada desenvolupi les aplicacions de les dades que poden generar benefici per la mateixa empresa i revertir en serveis desitjats pels usuaris. Cap problema. O gairebé cap problema. Deixant per més endavant el tema de la privacitat i de l’ús de la informació personal, el tema que més ens hauria de capficar en el desenvolupament del <big data> com a negoci és la tendència dels negocis en aquesta àrea a concentrar els ingressos i els guanys en no gaires mans, deixant xavalla per a la majoria dels participants. Hom pot argumentar que aquesta és una tendència general de l’economia del segle XXI, però és una tendència particularment agressiva entre les empreses el negoci de les quals és la informació, les dades. La desproporció entre el volum i la rapidesa d’acumulació de guanys d’empreses com <Google>, <Facebook> o <Airbnb> comparat amb els petits beneficis que obtenen els seus clients és de magnituds mai no vistes fins a les darreres dècades. Calia fer-ne esment.



Admès que és bo i desitjable que les empreses treguin profit econòmic de l’explotació de les dades massives, voldríem concentrar aquesta aportació en el tema que realment ens interessa: quin és el profit, i com obtenir-lo, que es pot extreure de les dades massives per al bé comú, per al bé social, per a la millora de la societat i de les condicions de vida dels ciutadans, per a l’anàlisi, avaluació i millora de les polítiques públiques. Diguem-ne de tot això <el bé comú> per ara, ja afinarem més endavant.

Hi han dos aspectes a tractar prèviament per enfocar l’ús de les dades massives per al bé comú: **la propietat de les dades en qüestió** i **els límits que el respecte de la privacitat imposa al seu ús i difusió**.

La propietat privada de la gran majoria de les dades del <big data> és un fet. Però que siguin privades no pot ser un obstacle definitiu per al seu ús en favor del bé públic. Alguns països ja han desenvolupat eines legals per donar dret a l’accés a dades propietat d’empreses privades per part d’organismes públics amb finalitats exclusivament estadístiques i això permet utilitzar-les sobretot conjuntament amb informació d’origen administratiu, una autèntica oportunitat per a l’estadística pública. Però aleshores emergeix un altre obstacle: molta de la informació <big data> no és només de propietat privada sinó que pertany a empreses estrangeres o multinacionals que són fora de la legislació estatal. Idealment això s’hauria d’enfocar a una escala com a mínim de la Unió Europea tot obligant a aquestes empreses a facilitar la reutilització de les seves dades per a finalitats estadístiques, possiblement amb reconeixement del dret a rebre compensacions econòmiques per les despeses que això els pugui produir. Però també cal cercar altres vies d’accés a la informació, com de fet ja existeixen en algun cas. Pot esser d’interès de les mateixes empreses facilitar part de la informació que tenen sobre els seus productes, com fa per exemple <twitter> quan facilita informació sobre la localització geogràfica de les piulades fetes amb dispositius mòbils. També pot ser interessant la pressió que poden fer les administracions públiques tot condicionant la contractació de serveis a la cessió de dades o a la seva publicació com a dades obertes.

L’altre gran tema que afecta a l’ús del big data per a la millora del bé comú és la privacitat. Per exemple, si volem utilitzar dades de mobilitat de telèfons mòbils i fer-les públiques, hi ha el perill que hom pugui re-identificar un aparell individual i per tant poder seguir els moviments d’una persona. En aquest cas, difonent només dades agregades es pot obviar el problema, a costa de perdre part de la informació. Curiosament ens trobem de vegades que més que la preservació de la privacitat a les companyies telefòniques els pesa més la por... o pànic que els seus clients percebin que es cedeixen les seves dades. Seria per tant important algun codi de bones pràctiques que facilités l’auto-regulació de les empreses propietàries alhora que permetés la utilització pública d’una bona part de les dades. En l’ús de dades, i més si es tracta de dades massives, no és possible eliminar del tot els riscos per a la privacitat. I no oblidem que sempre hi haurà dues menes de riscos: els causats a una voluntat d’emprar les dades per a benefici sense consentiment i els derivats de manca de seguretat. Per tant, caldrà sempre ser curosos en el seu ús, emmagatzematge i transmissió, però també ser molt actius tant en l’autoregulació com en la cerca i el disseny de mètodes que permetin l’explotació segura de les dades. Així, ja en el disseny de les bases de dades cal separar les dades que permeten identificació directa i regular estrictament el seu tractament. D’altra banda, l’anàlisi rigorós dels riscos de seguretat i l’aplicació de les mesures adequades per a la prevenció ha d’acompanyar una política de transparència i de facilitar l’accés a la informació al actors afectats.

Però fixem-nos-hi bé, sigui <big data>, <open data> o <admindata>, el que és important és com ho hem de fer per tal que l’ús de les dades reverteixi en bé del conjunt de la societat. Les administracions han de poder convenir amb les empreses propietàries de les dades per reutilitzar-les, possiblement prèvia combinació amb les dades que en diem <admindata>, dades d’origen administratiu. Els organismes responsables de l’estadística pública han de ser capaços d’utilitzar les noves fonts d’informació per posar al dia els mètodes de producció. Hi ha alguns problemes no gens fàcils de superar per aconseguir-ho: pel seu origen, el <big data> té manca d’universalitat, estandardització i comparabilitat, tres propietats que són importants per a l’estadística pública.

La universalitat és important, si volem reflectir com és la societat o el país, no podem fer-ho només usant dades de la gent o les empreses que són a les xarxes socials, per dir-ho de qualque manera. L’estandardització és important per poder fer interpretació correcta de les dades, i el <big data> evoluciona tan ràpid que és difícil d’estar segur què i com estem mesurant al curs del temps, per exemple, tenim milions de dades sobre transaccions de cartes de crèdit. La comparabilitat, en el temps i en l’espai és fonamental per a un bon ús de les dades. Serà important poder comparar el que mesurem avui amb el que vam mesurar l’any passat o el què farem d’aquí cinc anys, tant com serà important poder comparar la nostra situació amb la de països veïns. Aquests problemes es poden resoldre mitjançant la integració de les dades massives amb la dades provinent de fonts com <l’admindata> o les enquestes, que són més lentes, menys detallades i potser més cares, però que tenen la universalitat, l’estandardització i la comparabilitat que manca al <big data>.

Podem trobar exemples d’un bon ús de les dades per al bé social. Dades de mobilitat telefònica s’estan fent servir per millorar la planificació del transport públic. La gran capacitat de recol·lecció i integració de dades del sistema de salut català pot proporcionar, i s’estan posant les bases per fer-ho, un gran cabdal de coneixement al servei dels recercadors per a la millora de la salut de la població. I combinant dades de salut amb dades de medi ambient, de condicions de vida, d’educació, justícia o criminalitat es pot arribar a escatir quins són els factors que permetrien millorar la salut i les condicions de vida dels ciutadans, on val la pena invertir capital i esforços socials de manera eficient per la millora de les polítiques públiques.

## Què passarà amb l’ètica professional quan el <big data> sigui d’utilitat? Repassem-ne els principis!

## 

I no només podem trobar exemples d’un bon ús de les dades per al bé social, sinó que cal enfortir i desenvolupar el codi deontològic de la mateixa societat estadística perquè l’impacte del <big data> serà descomunal: la implicació de l’Institut Nacional d’Estadística (International Statistical Institute), en l’establiment d’una declaració d’ètica professional, es va perllongar durant 7 anys en la seva gestació. El comitè de l’institut en resposta de les sol·licituds dels seus membres a la proposta del comitè de direccions de futur de l’institut, va establir un comitè per al codi d’ètica dels estadístics el 1979, durant 42ª sessió de la IIE a Manila. El comitè va treballar per preparar una reunió general per a la següent sessió de Buenos Aires, el 1981, durant la qual es va acordar, per consens, dissenyar un codi: <el codi> havia de ser enllestit perquè l’Institut l’aprovés durant la celebració del centenari 1985.

La finalitat d’aquesta declaració és la de permetre a l’estadístic de fonamentar els seus judicis ètics i la seves decisions personals sobre valors i experiències compartides, més que noi pas sobre regles dictades per la professió. La declaració cerca en conseqüència extreure uns principis d’investigació estadística que siguin àmpliament reconeguts i identificar els obstacles per posar-los en pràctica. S’ha establert tot tenint present que, a vegades, l’aplicació d’un principi en contradirà un altre, que els estadístics – com altres grups professionals – tenen obligacions concurrents que no poden esser complides ensems. Implícitament o explícitament, a vegades, s’haurà de triar entre aquests principis. La present declaració no pretén evitar aquests tries, ni privilegiar un principi respecte un altre. En lloc d’això, ofereix un marc en el qual l’estadístic conscient hauria de poder treballar amb comoditat en allò essencial. Quan l’estadístic es plantegi de separar-se d’aquest marc, hauria d’esser més un acte reflexiu que no un acte de la ignorància.

El text és dividit en quatre parts. A dins de cadascuna, l’enunciat de principis o conjunt de principis aquí no és seguit sobre les dificultats i els conflictes inherents a la seva posada en pràctica, per una qüestió d’espai i perquè no és l’objectiu primari d’aquest treball. Els diversos principis són lligats entre si i, en conseqüència, han d’esser considerats en conjunt: l’ordre de presentació no ha d’esser interpretat com un ordre de prioritat. Fem-ne un repàs sintètic i exhaustiu!

1. Obligacions envers la societat:

1.1 Prendre en consideració els interessos contradictoris:

L’estadístic ha d’estar atent a les possibles conseqüències dels seus treballs. El dret que la societat té de conèixer els trets col·lectius i el dret de la persona a protegir la seva intimitat entren a vegades en conflicte.

1.2 Eixamplar l’abast de l’estadística: Els estadístics han d’emprar les facultats que hom els ofereixen per eixamplar la recerca estadística i difondre les seves conclusions per al profit per a la més àmplia comunitat possible. Concretament, els estadístics tenen més de possibilitats d’evitar les restriccions imposades al seu treball quan són capaços d’estipular amb anterioritat conceptes i tècniques (per aplegar, analitzar o interpretar la informació) sobre els quals exerciran un control. Altrament dit: les més grans dificultats semblen produir-se quan aquestes qüestions es deixen en suspens fins a la sortida dels resultats.

1.3 Cercar l’objectivitat: tot i treballar a dins del sistema de valors de les societats on viuen, els estadístics han de maldar preservar la seva integritat professional, sense pors (prejudicis...) o favors.

La ciència no pot esser del tot objectiva i l’estadística, per tant, no n’és una excepció. Els estadístics tenen en concret l’obligació professional d’oposar-se a maneres de recollida, d’anàlisis, d’interpretació o de publicació de la informació que puguin (de manera explícita o implícita) informar malament o induir a error en lloc de contribuir a l’avenç del coneixement.

2. Obligacions envers els <sponsors> i els empresaris

2.1 explicitar les obligacions i els papers: els estadístics han d’establir clarament amb anterioritat les obligacions respectives del seu empresari o de qui finança l’enquesta i també establir les seves obligacions pròpies.

2.2 avaluar amb imparcialitat les diverses vies: els estadístics han d’examinar els diversos mètodes i procediments a la seva disposició per a la investigació prevista i han d’oferir al seu spònsor o empresari una apreciació imparcial dels mèrits i inconvenients

2.3 compromisos segons els resultats: els estadístics no han d’acceptar cap clàusula, l’aplicació de la qual depengui del resultat obtingut en la recerca estadística programada.

2.4 mantenir la confidencialitat sobre la informació privada: els estadístics reben sovint del seu empresari o spònsor informacions de les quals aquest pot exigir legítimament que es mantinguin confidencials. Els mètodes i procediments estadístics emprats per produir

3. Obligacions envers els col·legues

3.1 preservar la confiança en l’estadística: els estadístics tenen necessitat de la confiança en el públic. En els seus treballs, han d’intentar de promoure i de preservar aquesta confiança, sense exagerar l’exactitud ni el poder explicatiu dels seus resultats.

3.2 descriure i verificar els mètodes i resultats: tot respectant els límits de la confidencialitat, els estadístics han de donar als seus col·legues les informacions adients perquè els seus mètodes, procediments, tècniques i resultats puguin esser jutjats. Aquests judicis s’han de dirigir als mètodes mateixos, i no pas a les persones que els han triats o emprats.

3.3 fer conèixer els principis ètics: per tal de dur a terme certes recerques, els estadístics han de col·laborar amb especialistes d’altres disciplines i amb els enquestadors, el personal administratiu, els estudiants, etc. En aquests casos han d’exposar clarament els seus principis ètics i tenir en compte els dels seus col·laboradors.

4. Obligacions envers els individus

4.1 Evitar les intrusions injustificades: els estadístics han d’esser conscients del potencial d’intrusió que comporten alguns dels seus treballs. Cap prerrogativa concreta els autoritza a estudiar tots els fenòmens. Ni l’avenç del coneixement ni la recerca d’informació justifiquen per ells mateixos que es transgredeixin altres valors socials o culturals.

4.2 Obtenir el ple consentiment: les enquestes estadístiques que comporten la participació activa d’individus humans han d’esser, sempre que sigui factible, fetes amb ple consentiment, és a dir, donat lliurement i amb coneixement de causa. Fins i tot quan la participació dels individus és legalment obligatòria, la seva acceptació ha de reposar sobre una informació com més completa possible millor. Quan la participació és voluntària, no s’ha de fer creure els individus que aquesta és obligatòria; han d’esser informats del seu dret a refusar, en qualsevol moment i per la raó que sigui, i de retirar les informacions que acaben de subministrar. Cap informació susceptible d’afectar la voluntat de participació de l’individu no pot esser deliberadament silenciada.

4.3 Modificacions al ple consentiment: en alguns casos raons tècniques o pràctiques impedeixen l’obtenció prèvia del ple consentiment dels individus. Els interessos d’aquests han de romandre protegits per d’altres vies. Així:

a. Respecte dels drets en els estudis de camp: l’espai privat d’un individu o d’un grup varia d’una cultura a una altra.

b. Passada per un tercer: quan hi ha indicis que fan pensar que l’individu s’oposaria que fossin desvelades certes informacions, aquestes dades no s’han d’anar a cercar per mitjà d’una tercera persona.

c. Emprament secundari de les dades: no exonera l’estadístic del deure de prendre en consideració els interessos de les persones implicades, inclòs el dret a l’anonimat.

d. Desorientació dels individus: dissimular una informació essencial, o equivocar els individus, constitueix una mentida, per acció o omissió, temporal o definitiva, que comporta una legítima condemna si no es pot justificar.

4.4 Protegir els interessos dels individus: ni el consentiment dels individus ni l’obligació legal de respondre exoneren l’estadístic de protegir tant com sigui possible els individus dels efectes potser perjudicials de la seva participació. L’estadístic ha d’esforçar-se a molestar el mínim els individus i les seves relacions amb l’entorn.

4.5 Assegurar la confidencialitat: l’estadística no s’ocupa de les identitats individuals. Les dades són recollides per respondre a demandes o qüestions del tipus <quants?> o <quina proporció?> i no pas de la mena <qui?>. La identitat dels individus que responen i els que no ho fan, i les informacions recollides gràcies a ells, han de restar en conseqüència confidencials, hagi estat o no la confidencialitat promesa de forma explícita.

4.6 Impedir que les identitats siguin revelades: els estadístics han de prendre totes les mesures convenients per impedir que els resultats siguin publicats o difosos sota una forma que permeti que la identitat de cap individu sigui revelada o deduïda.

## L’empelt del <big data> en les dades públiques de l’IDESCAT. Projectes que hi són relacionats.

Tornant a la faceta més tècnica i deixant ara l’aspecte ètic i professional que comporta l’empelt del <big data> en l’estadística pública, veiem que segons els estàndards internacionals, la modernització de la producció estadística oficial consisteix, entre altres aspectes, en la integració de les diverses fonts d'informació estadística, en contrast amb el mètode tradicional, anomenat sovint stovepipe approach (aproximació per operacions aïllades).

Actualment, les fonts d'informació estadística són les d'origen administratiu, les provinents d'enquestes pròpies o d'altri i, eventualment, les dades generades per empreses públiques o privades.

L'objectiu del Sistema integrat d'informació estadística (SIIE) és possibilitar que dues o més fonts d'orígens similars o diversos puguin ser utilitzades conjuntament quan ho requereixi una operació estadística oficial. El SIIE va ser creat per la Llei 23/1998, d'estadística de Catalunya, i ara n'estem desenvolupant la versió del segle XXI.

Per aconseguir aquest objectiu, l'Idescat ha engegat en els darrers anys dues línies o camins de treball en paral·lel: [A] la línia tecnològica (amb la construcció de la Plataforma Cerdà) i [B] la línia de processos estadístics (amb el Projecte Qualitas).

[A]

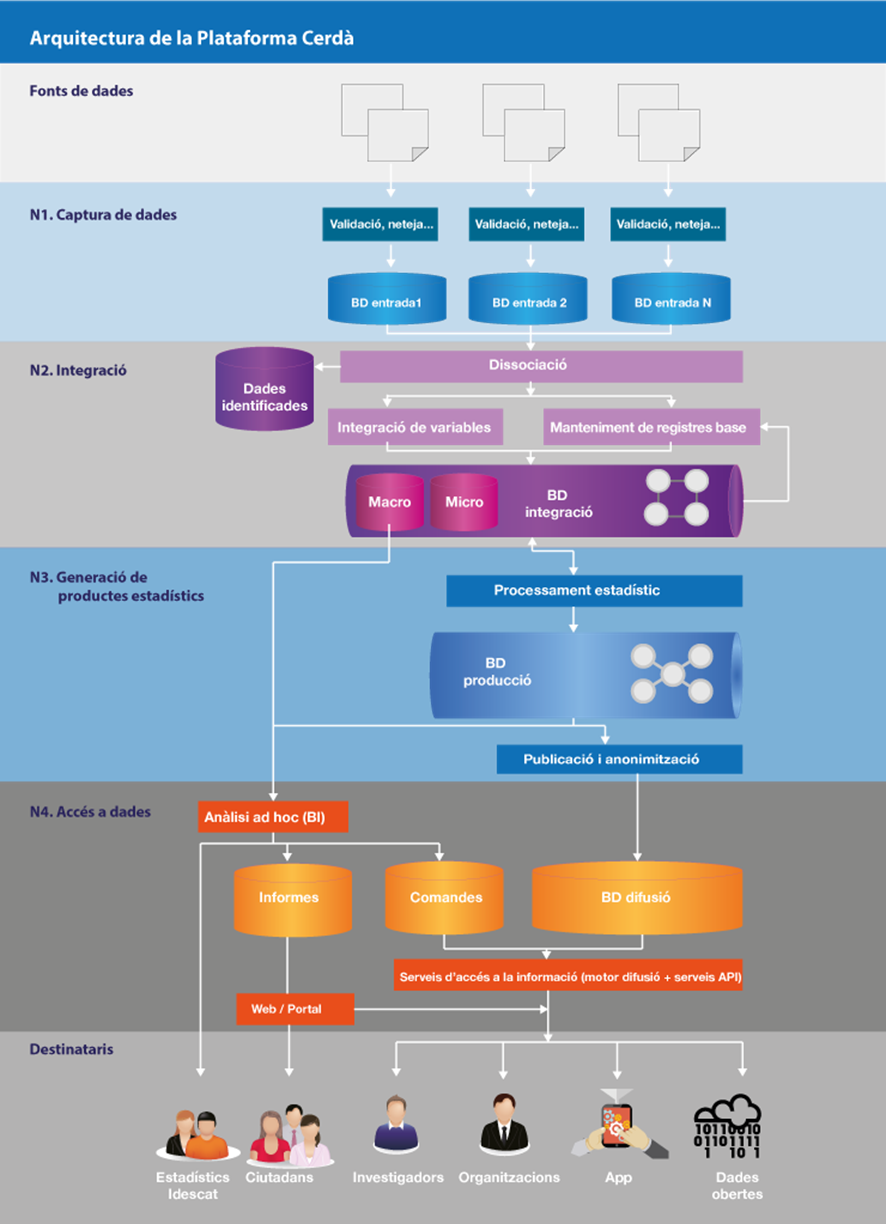
Per aconseguir la integració de la informació, la Plataforma Cerdà estructura conceptualment el procés de la informació estadística en quatre fases:

I. Captura de dades: assegura que tota la informació d'interès estadístic que entra a l'Idescat resti registrada i emmagatzemada ordenadament, després de passar pels processos de validació necessaris.

II. Integració de la informació: codificació de tota la informació segons els registres estadístics bàsics, les classificacions oficials i els instruments disponibles. Per preservar el secret estadístic, la informació identificadora dels subjectes estadístics (persones, empreses, entitats o elements territorials) és separada de la resta de la informació i mantinguda a banda en una zona sotmesa a mesures especials de seguretat i control d'accés, d'acord amb la Llei 23/1998, d'estadística de Catalunya.

III. Generació de productes estadístics: en la mesura necessària per al desplegament, l'elaboració i la difusió de les estadístiques oficials, s'encreua informació provinent de diferents fonts o altres operacions per produir els resultats desitjats. En aquesta fase també s'apliquen els protocols necessaris per preservar la confidencialitat i el secret estadístic.

IV. Accés a les dades: adaptació de la informació, amb finalitats estadístiques, als diferents destinataris a què s'adreça. Cada tipus de destinatari requereix un diferent tractament de la informació, un nivell de desagregació, un detall territorial, uns nivells de seguretat... Des de la feina dels tècnics de l'Idescat (que poden tenir accés complet a les dades) fins a la publicació de les dades al web (que són totalment obertes) es defineixen diferents protocols de difusió de la informació.



[B]

El Projecte Qualitas consisteix en la descripció dels processos estadístics adaptats a un entorn d'informació integrada. Per aconseguir la màxima eficiència en la producció estadística, cal que tota la informació disponible treballi conjuntament en la descripció d'una realitat social i econòmica complexa i canviant.

En una primera fase, el Projecte Qualitas **ha efectuat una descripció completa i acurada dels processos en què està treballant l'Idescat**, prenent com a guia el model GSBPM, un estàndard internacional desenvolupat i adoptat per l'OCDE, l'Eurostat i la UNECE. A partir d'una setantena de projectes ja documentats, s'està treballant en la seva transformació i integració en la Plataforma Cerdà.

L'objectiu del Projecte Qualitas és doble: d'una banda, utilitzar la documentació detallada dels processos per garantir la qualitat (rellevància, puntualitat, eficiència, exactitud i fiabilitat, coherència i comparabilitat, accessibilitat i claredat); d'altra banda, proporcionar una guia i un control, a diferents nivells, dels cinc aspectes següents: l'automatització de les fases del procés que ho permetin, les persones involucrades en el procés, la millora i l'aprofitament de sinergies entre diferents processos, els experts interessats en els detalls de l'elaboració de les estadístiques, el públic potencialment interessat en la informació produïda.



## Cloenda

Per concloure, un comentari sobre el problema de la privacitat versus l’ús públic de les dades, siguin big o small. Per poder fer anàlisis interessants dels efectes de, posem per cas, les polítiques públiques sobre les persones, com més detallades siguin les dades que disposem més riques seran les conclusions possibles. Però si hem de treballar amb dades personals no ens podem enganyar, l’anonimització no és possible! Si tenim prou informació, sempre serà possible re-identificar els individus amb què treballem. I si no és possible... com s’ha d’adaptar el codi ètic de l’estadístic?

Llavors ens resten dues vies de treball: d’una banda podem difondre informació agregada que pugui ser pública sense risc de re-identificació, i d’això en diem <open data>, o podem restringir l’ús de la informació més detallada i rica a les persones o institucions que poden donar garanties que no faran cap procés de re-identificació ni difusió de cap característica individual.

# Bibliografia:

*“Big Data: A Revolution That Will Transform How We Live, Work, and Think.” Viktor Mayer-Schönberger. 2013*

[*https://en.wikipedia.org/wiki/Big\_data*](https://en.wikipedia.org/wiki/Big_data)

[*https://en.wikipedia.org/wiki/Data\_mining*](https://en.wikipedia.org/wiki/Data_mining)

<http://www.springer.com/cda/content/document/cda_downloaddocument/9781852337032-c1.pdf?SGWID=0-0-45-128448-p27228572>.

[*https://ca.wikipedia.org/wiki/Visualitzaci%C3%B3\_de\_dades*](https://ca.wikipedia.org/wiki/Visualitzaci%C3%B3_de_dades)

[*http://www.36dsj.com/archives/58886*](http://www.36dsj.com/archives/58886)

[*http://ismailvtl.blogspot.com.es/2014/07/display-live-data-on-your-site-use.html*](http://ismailvtl.blogspot.com.es/2014/07/display-live-data-on-your-site-use.html)

[*https://zhuanlan.zhihu.com/p/25780330*](https://zhuanlan.zhihu.com/p/25780330)

[*https://msdn.microsoft.com/es-es/library/ms175595.aspx*](https://msdn.microsoft.com/es-es/library/ms175595.aspx)

[*https://blog.es.logicalis.com/analytics/las-cinco-principales-aplicaciones-de-big-data*](https://blog.es.logicalis.com/analytics/las-cinco-principales-aplicaciones-de-big-data)

[*http://www.viu.es/las-aplicaciones-del-big-data-mas-importantes/*](http://www.viu.es/las-aplicaciones-del-big-data-mas-importantes/)

[*https://blog.es.logicalis.com/analytics/big-data-ventajas-de-la-revolucion-de-los-datos-masivos?\_\_hstc=61804339.004b6f0727df1907a88059568351a029.1494931998065.1494931998065.1494934896983.2&\_\_hssc=61804339.1.1494934896983&\_\_hsfp=23383515*](https://blog.es.logicalis.com/analytics/big-data-ventajas-de-la-revolucion-de-los-datos-masivos?__hstc=61804339.004b6f0727df1907a88059568351a029.1494931998065.1494931998065.1494934896983.2&__hssc=61804339.1.1494934896983&__hsfp=23383515)

[*http://news.sciencenet.cn/htmlnews/2016/12/363189.shtm*](http://news.sciencenet.cn/htmlnews/2016/12/363189.shtm)

[*http://www.cbdio.com/BigData/2017-01/23/content\_5437741.htm*](http://www.cbdio.com/BigData/2017-01/23/content_5437741.htm)

<https://www.bigdatabcn.com/cal-utilitzar-dades-massives-millora-social/><http://bigdataweek.com/blog/2014/03/31/big-data-i-estadistica-oficial-oportunitat-o-amenaca/><http://www1.unece.org/stat/platform/display/GSBPM/GSBPM+v5.0><http://www.idescat.cat/metodes/siie/>