TEMA 1

Introducció a la mineria de dades

Índex

- Què és la mineria de dades?
- Per què la mineria de dades: motivació i beneficis?
- Quin tipus de dades minar?
- Quan minar les dades?
- Com organitzar el procés de la mineria?
- Quins són els desafiaments de la mineria de dades?

Perquè mineria de dades?

Inundació de dades:

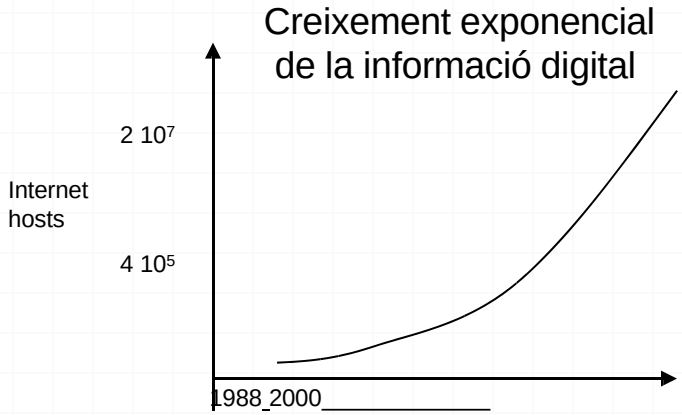
- Bancs, telecomunicacions, altres transaccions comercials...
- Dades científiques: astronomia, biologia, etc.
- Web, text, i el comerç electrònic

Perquè ara?

Fonts de sobrecàrrega de dades:

- Fonts de dades distribuïdes
- La teledetecció/sensors
- Internet
- Dades multimèdia

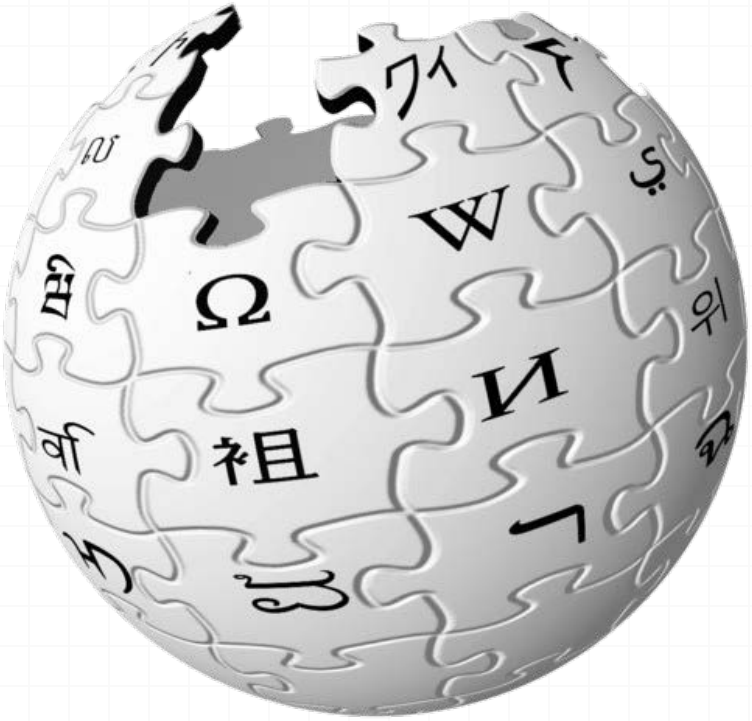
....



Existeix un forat entre la capacitat tecnològica de recol·lecció i organització de dades, i les habilitats per analitzar grans conjunts de dades i extraure'n coneixement útil per a la presa de decisions.

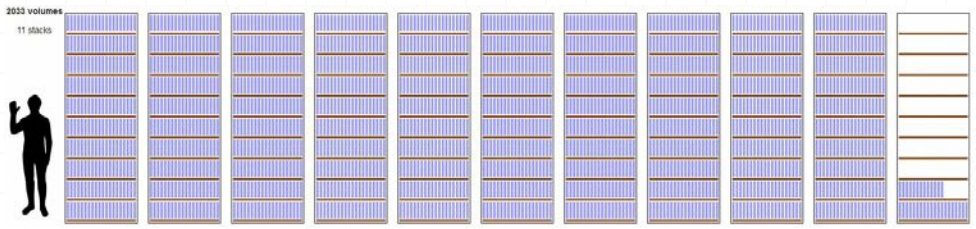
"We are Drowning in Data..."

The following slides are taken from Aidan Hogan's course on "Massive Data Processing"

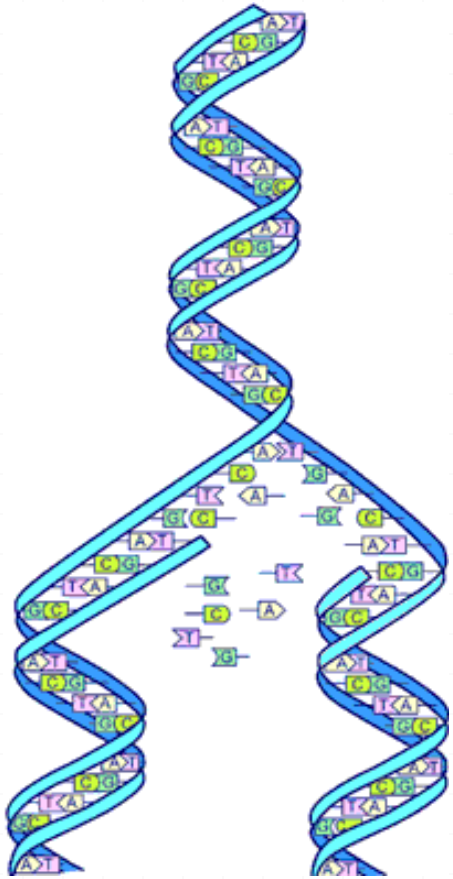


Wikipedia
≈ 10 TB of data
(May *2016 Dump*)

1 Wiki = 1 Wikipedia



“We are Drowning in Data...”

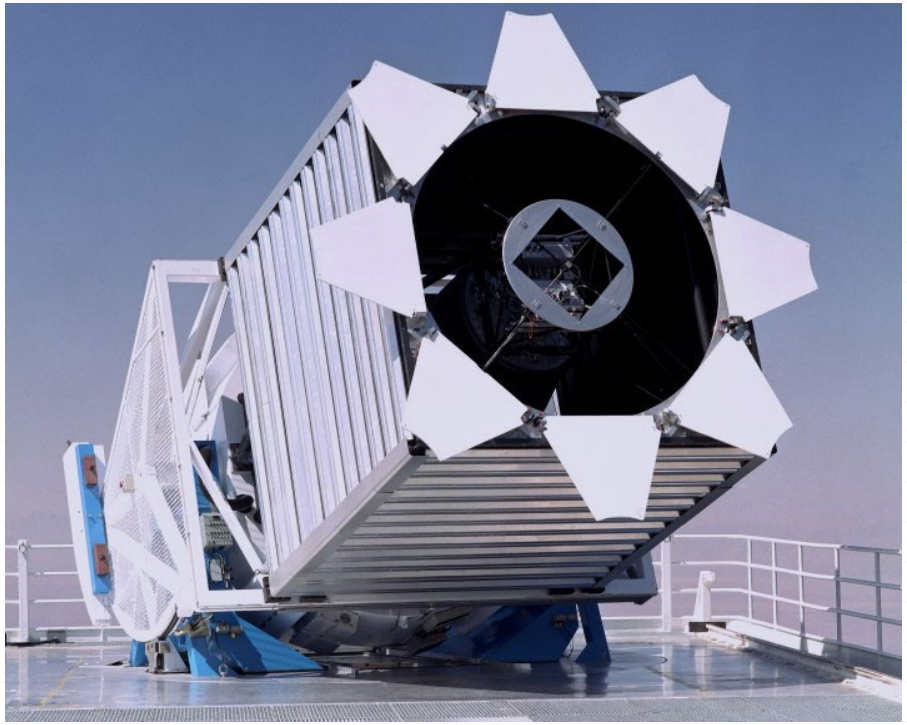


Human Genome
≈ 4 GB/person
≈ 0.0004 Wiki/person
≈ 2.4M Wiki/humankind

"We are Drowning in Data..."

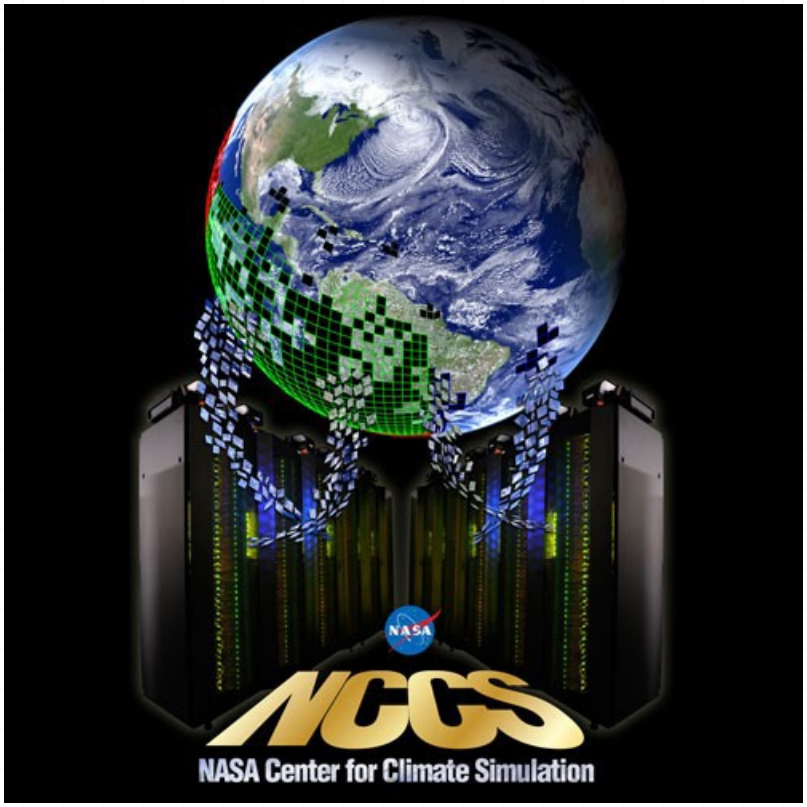
US Library of Congress
≈235 TB archived
≈23.5 Wiki

“We are Drowning in Data...”



Sloan Digital Sky Survey
≈200 GB/day
≈73 TB/year
≈7.3 Wiki/year

“We are Drowning in Data...”

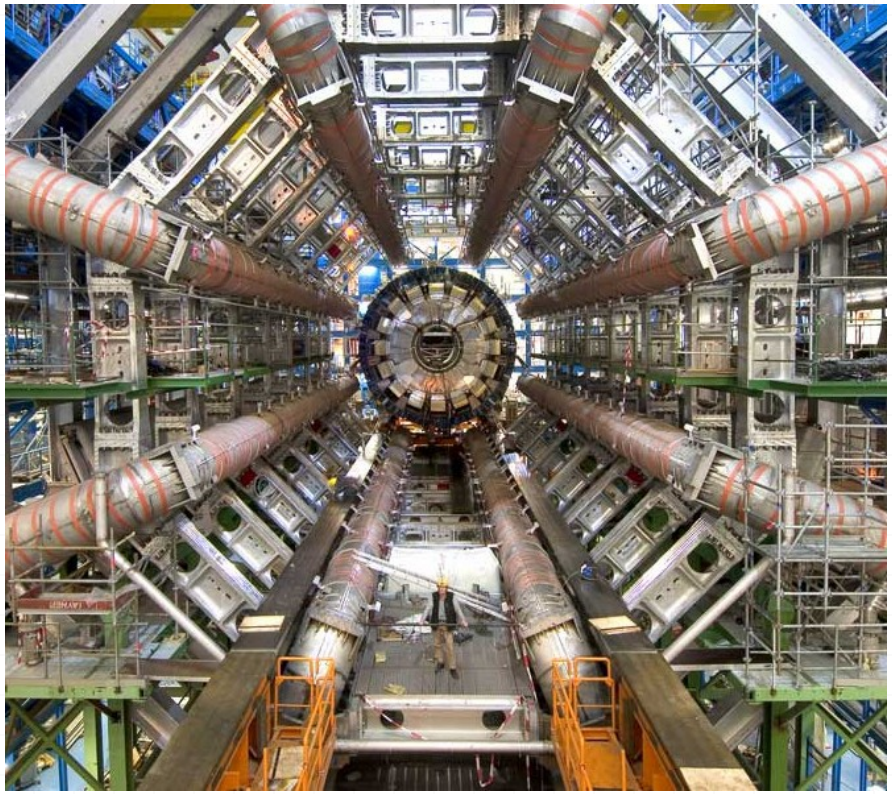


NASA Center for Climate Simulation
≈ 32 PB archived
≈ 3,200 Wiki

"We are Drowning in Data..."

Facebook
≈12 TB/day added
≈1.2 Wiki/day
≈438 Wiki/year
(*as of Mar. 2010*)

“We are Drowning in Data...”



Large Hadron Collider
≈15 PB/year
≈1,500 Wiki/year

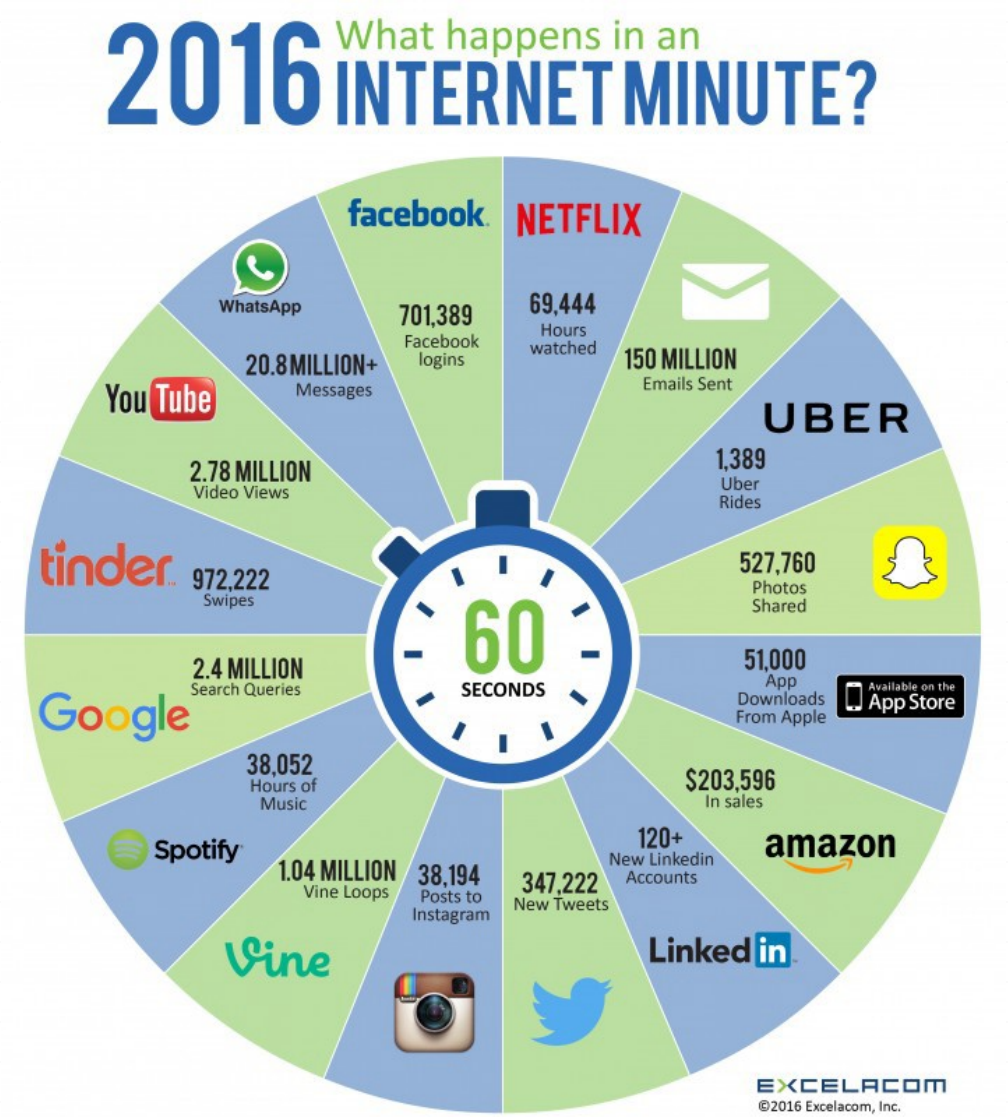
"We are Drowning in Data..."

Google
≈20 PB/day processed
≈2,000 Wiki/day
≈730,000 Wiki/year
(Jan. 2010)

"We are Drowning in Data..."

Internet (2016)
≈1.3 ZB/year
≈130,000,000 Wiki/year
(2016 IP traffic; Cisco est.)

"We are Drowning in Data..."



De quantes dades parlem?

MEDLINE: Base de dades d'articles en medicina	• 12 milions d'articles publicats
Google	• 4.2 mil milions de pàgines Web indexades • 80 milions de visitants per dia
CALTRANS dades del sensor en bucle permanent	• Cada 30 segons, milers de sensors, 2Gbytes per segon
Satèl·lit MODIS de la NASA	• Cobertura en la resolució 250, 37 bandes, terra sencera, tots els dies
Dades de transaccions a *Walmart*:	• Ordre de 100 milions de transaccions per dia

La mineria de dades s'estén

FINANCIAL INSTITUTIONS

RETAIL INDUSTRY

TELECOMMUNICATION INDUSTRY

HEALTH INDUSTRY

SCIENCE & ENGINEERING

GOVERNMENT

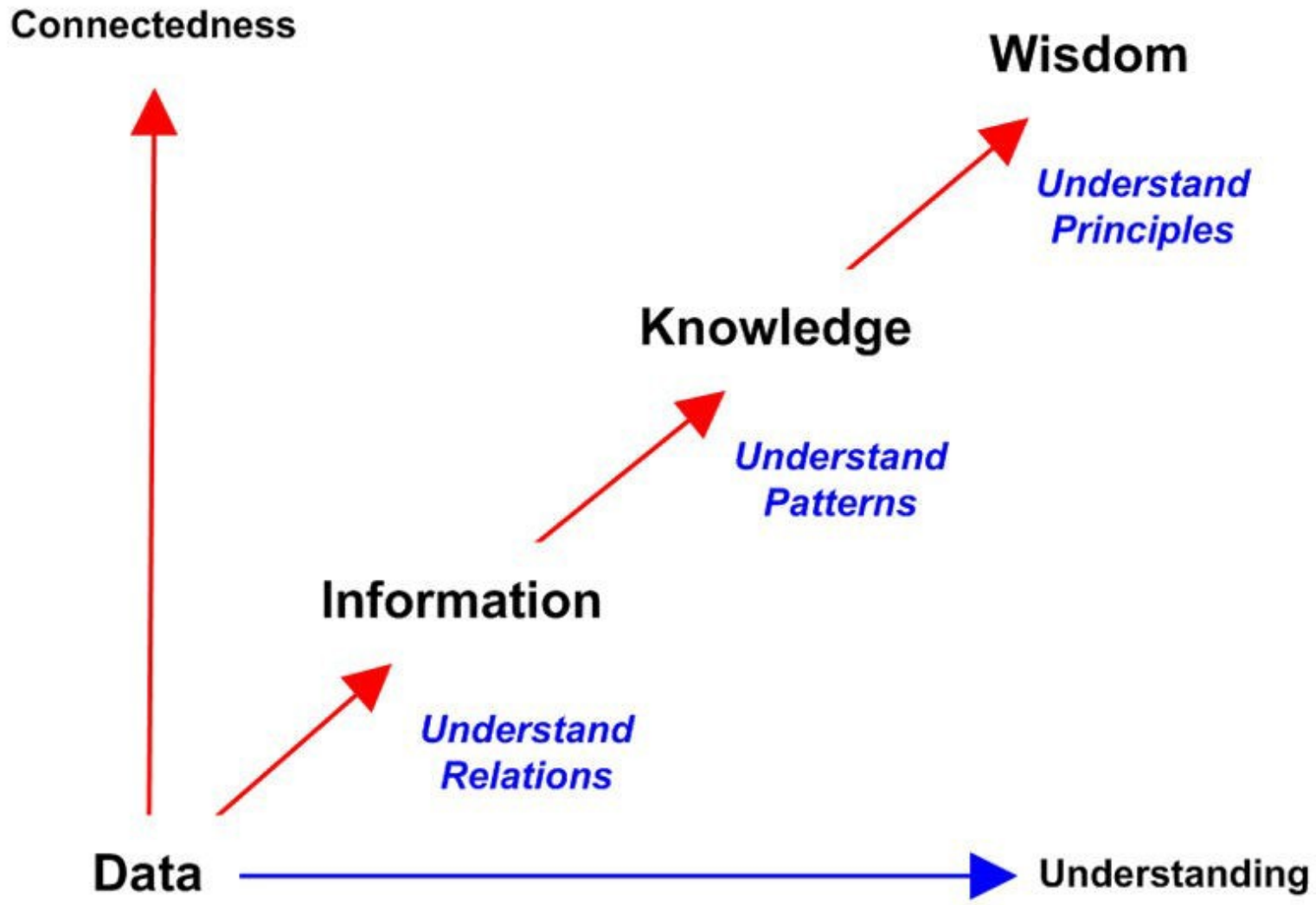
E-COMMERCE

De quantes dades parlem?

- **Explosió de dades fa que les dades es desaprofitin:**
  - Només una petita porció (5% - 10%) de les dades recopilades s'analitzen.
  - Les dades que mai podran ser analitzades es segueixen recopilant malgrat la despesa que suposa

Ens estem ofegant en dades, però no en traiem CONEIXEMENT

Data, Information, Knowledge, and Wisdom



Gene Bellinger, Durval Castro and Anthony Mills. "Transforming Data to Wisdom."

De dades a coneixement: exemple

Medical Data by Dr. Tsumoto, Tokyo Med. & Dent. Univ., 38 atributs

...
10, M, 0, 10, 10, 0, 0, 0, SUBACUTE, 37, 2, 1, 0,15,-,-, 6000, 2, 0, abnormal, abnormal,-, 2852, 2148, 712, 97, 49, F,-,multiple,,2137, negative, n, n, ABSCESS,VIRUS

12, M, 0, 5, 5, 0, 0, 0, ACUTE, 38.5, 2, 1, 0,15, -,-, 10700,4,0,normal, abnormal, +, 1080, 680, 400, 71, 59, F,-,ABPC+CZX,, 70, negative, n, n, n, BACTERIA, BACTERIA

15, M, 0, 3, 2, 3, 0, 0, ACUTE, 39.3, 3, 1, 0,15, -, -, 6000, 0,0, normal, abnormal, +, 1124, 622, 502, 47, 63, F, -,FMOX+AMK, , 48, negative, n, n, n, BACTE(E), BACTERIA

16, M, 0, 32, 32, 0, 0, 0, SUBACUTE, 38, 2, 0, 0, 15, -, +, 12600, 4, 0,abnormal, abnormal, +, 41, 39, 2, 44, 57, F, -, ABPC+CZX, ?, ? ,negative, ?, n, n, ABSCESS, VIRUS
...

Attribut Numéric — Attribut categoric — valors perduts — Etiqueta de classe

IF a9=ACUTE AND a33=negative AND a2=m AND a10<30
THEN BACTERIA [87,5%]

Precisió en la predicció

Què és el Data Mining?

La Mineria de Dades (*Data Mining*) és un procés per la extracció automàtica de coneixement ocult no obvi a partir de conjunts de dades de gran volum.

En la pràctica DM es refereix a:

- Trobar patrons/models en grans conjunts de dades
- Descobrir informació desconeguda oculta en les dades

Canvi de paradigma

En molts dominis hi ha un canvi de perspectiva que passa del modelatge i anàlisi clàssic basat en primers principis, al desenvolupament de models i anàlisis corresponents directament de les dades.

Arrels de la mineria de dades

- Estadística
  - Centrat en la construcció de models
- Base de dades
  - Centrat en la gestió de grans quantitat de dades
- Aprenentatge Automàtic
  - Centrat en els algorismes
- Visualització de Dades

Tipus de tasques de Data Mining

Acme Investors Incorporated

Customer ID	Account Type	Margin Account	Transaction Method	Trades/ Month	Sex	Age	Favorite Recreation	Annual Income
1005	Joint	No	Online	12.5	F	30–39	Tennis	40–59K
1013	Custodial	No	Broker	0.5	F	50–59	Skiing	80–99K
1245	Joint	No	Online	3.6	M	20–29	Golf	20–39K
2110	Individual	Yes	Broker	22.3	M	30–39	Fishing	40–59K
1001	Individual	Yes	Online	5.0	M	40–49	Golf	60–79K

Clustering:

- Puc desenvolupar una caracterització general / Perfil de diferents tipus d'inversors?

Classificació

- Quines característiques distingeixen entre els inversors i el Broker Online?

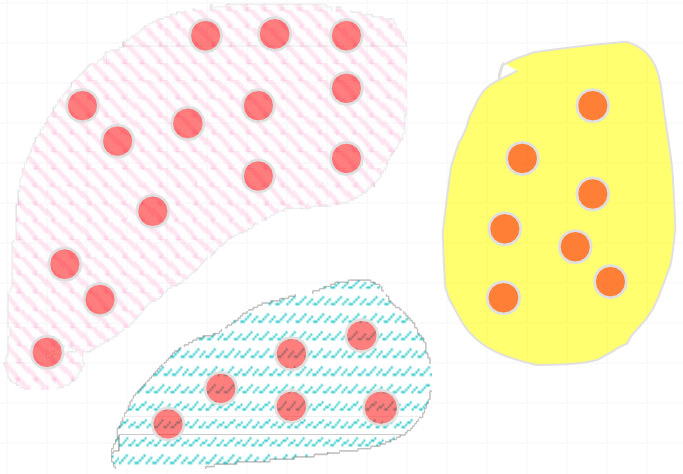
Descobriment de patrons

- Puc descobrir quan l'ocurrencia d'uns events estan relacionats amb d'altres?

Clustering

- Given a set of data points, and a similarity measure among them, find clusters such that
  - Data points in one cluster are similar to one another
  - Data points in separate clusters are different from each other

- Result:
  - a descriptive grouping of data points

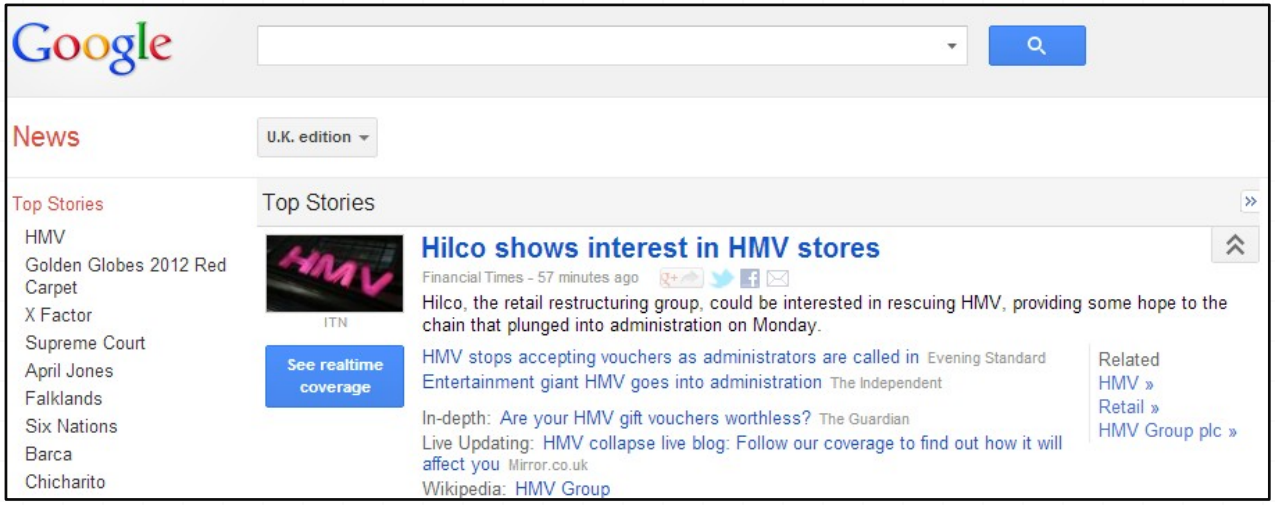


Clustering: Applications

- Application area: Market segmentation
- Goal: Subdivide a market into distinct subsets of customers
  - where any subset may be conceived as a marketing target to be reached with a distinct marketing mix
- Approach:
  - Collect information about customers
  - Find clusters of similar customers
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters

Clustering: Applications

- Application area: Document Clustering
- Goal: Find groups of documents that are similar to each other based on the important terms appearing in them
- Approach
  - Identify frequently occurring terms in each document
  - Define a similarity measure based on the frequencies of different terms
- Application Example: Grouping of stories in Google News



Tipus de tasques de Data Mining

Acme Investors Incorporated

Customer ID	Account Type	Margin Account	Transaction Method	Trades/ Month	Sex	Age	Favorite Recreation	Annual Income
1005	Joint	No	Online	12.5	F	30–39	Tennis	40–59K
1013	Custodial	No	Broker	0.5	F	50–59	Skiing	80–99K
1245	Joint	No	Online	3.6	M	20–29	Golf	20–39K
2110	Individual	Yes	Broker	22.3	M	30–39	Fishing	40–59K
1001	Individual	Yes	Online	5.0	M	40–49	Golf	60–79K

Clustering:

- Puc desenvolupar una caracterització general / Perfil de diferents tipus d'inversors?

Classificació

- Quines característiques distingeixen entre els inversors i el Broker Online?

Descobriment de patrons

- Puc descobrir quan l'ocurrencia d'uns events estan relacionats amb d'altres?

Classification

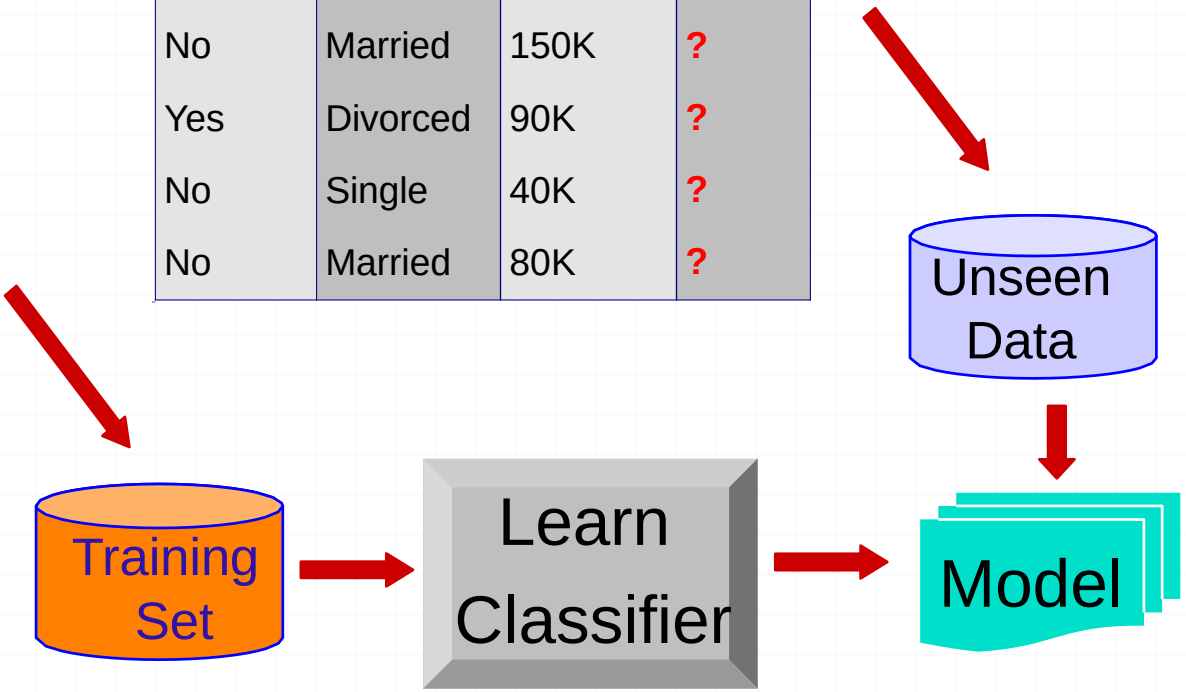
- Given a collection of records (training set)
  - each record contains a set of attributes
  - one of the attributes is the class (label) that should be predicted
- Find a *model* for class attribute as a function of the values of other attributes
- Goal: previously unseen records should be assigned a class as accurately as possible
  - A test set is used to validate the accuracy of the model
  - Training set may be split into training and validation data

Classification Example

Class / Label Attribute

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Applications

- Application area: Direct Marketing
- Goal: Reduce cost of mailing by targeting
  a set of consumers
  which are likely to buy a new cell phone
- Approach:
  - Use the data for a similar product introduced before
  - We know which customers decided to buy and which did not
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model

Classification: Applications

- Application area: Fraud Detection
- Goal: Recognize fraudulent cases in credit card transactions
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes
    - When and where does a customer buy? What does he buy?
    - How often he pays on time? etc.
  - Label past transactions as *fraud* or *fair* transactions
    This forms the *class attribute*
  - Learn a model for the class of the transaction
  - Use this model to detect fraud by observing credit card transactions on an account

Tipus de tasques de Data Mining

Acme Investors Incorporated

Customer ID	Account Type	Margin Account	Transaction Method	Trades/ Month	Sex	Age	Favorite Recreation	Annual Income
1005	Joint	No	Online	12.5	F	30–39	Tennis	40–59K
1013	Custodial	No	Broker	0.5	F	50–59	Skiing	80–99K
1245	Joint	No	Online	3.6	M	20–29	Golf	20–39K
2110	Individual	Yes	Broker	22.3	M	30–39	Fishing	40–59K
1001	Individual	Yes	Online	5.0	M	40–49	Golf	60–79K

Clustering:

- Puc desenvolupar una caracterització general / Perfil de diferents tipus d'inversors?

Classificació

- Quines característiques distingeixen entre els inversors i el Broker Online?

Descobriment de patrons

- Puc descobrir quan l'ocurrencia d'uns events estan relacionats amb d'altres?

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
- produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered
{Diaper, Milk} → {Beer}
{Milk} → {Coke}

Association Rule Discovery: Applications

- Application area: Marketing and Sales Promotion
- Example rule discovered:

  {Bagels, Coke} --> {Potato Chips}
- Insights:
  - promote bagels to boost potato chips sales
  - if selling bagels is discontinued, this will affect potato chips sales
  - coke should be sold together with bagels to boost potato chips sales

Association Rule Discovery: Applications

- Customers who bought this product also bought…
  - ...do terrorists order bomb building parts on Amazon?

http://thenewdaily.com.au/news/world/2017/09/21/amazon-bomb-explosives-ingredients-algorithm-frequently-bought-together/

Association Rule Discovery: Applications

- Real example:
  - Target (American grocery store)
  - Analyzes customer buying behavior
  - Sends personalized advertisement

- Famous case in the USA:
  - Teenage girl gets advertisement for baby products
  - ...and her father is mad

http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/

Association Rule Discovery: Applications

- Bottom line of the Target teenage girl story:
  - Janet Vertesi, Princeton university
  - Tried to hide her pregnancy from computers

- Measures taken:
  - using Tor for online surfing
  - no social media posts about her pregnancy
  - paying all pregnancy/baby related products in cash
  - a fresh Amazon account delivering to a local locker
    - paying with cash-payed gift cards

read the full story at
http://mashable.com/2014/04/26/big-data-pregnancy/

- Outcome:
  - massive buying of gift cards in a convenience store was reported to tax authorities

Tipus de tasques de Data Mining

Classificació

- trobar la descripció de diverses classes predefinides i classificar un element de dades en un d'ells.

Regressió

- assigna una dada a una variable de predicció de valor real

Clustering

- la identificació d'un conjunt finit de categories o grups per descriure les dades (p.e. la personalització)

Sumarització

- la recerca d'una descripció compacta per a un subconjunt de les dades

Modelització de dependències

- la recerca d'un model que descriu les dependències significatives entre les variables.

Detecció de canvis i desviacions

- descobrir canvis significatius en les dades

Data Mining: Altres noms

Information Harvesting

Knowledge Mining

Data Mining

Intelligent Data Analysis

Knowledge Discovery in Databases

Data Dredging

Data Pattern Processing

Data Archaeology

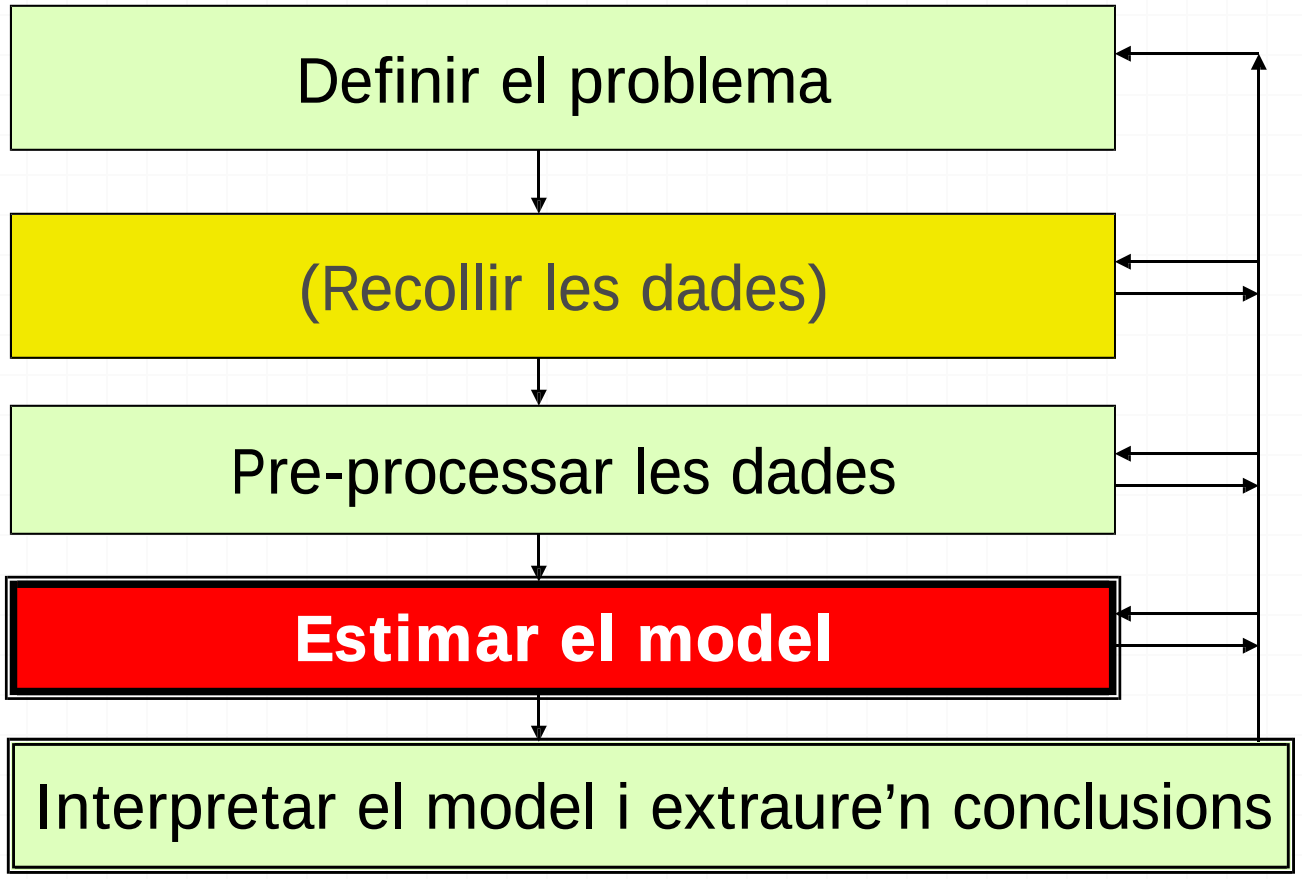
Database Mining

Siftware

Data Fishing

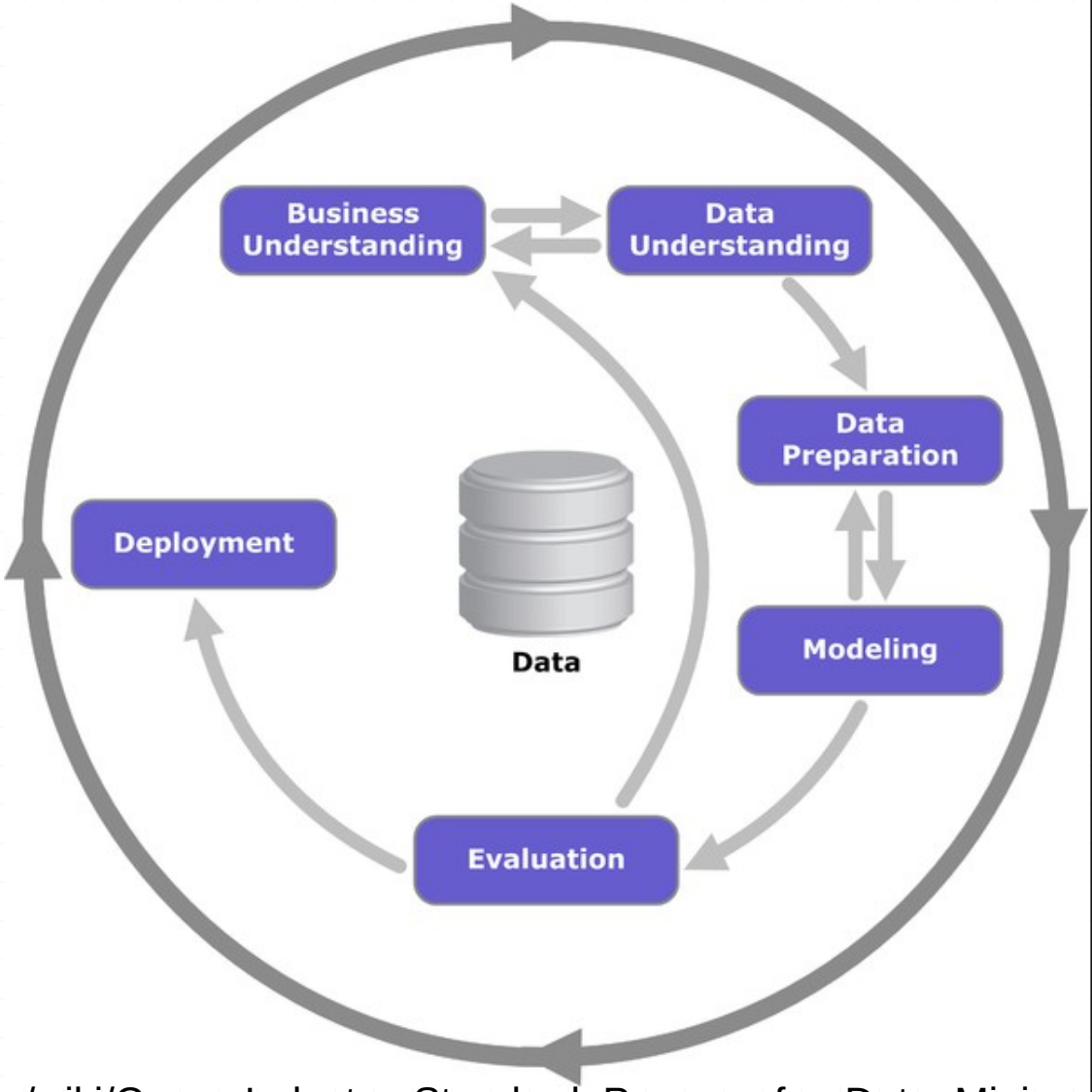
Knowledge Extraction

Data Mining és un procés



CRISP-DM: Reference Model

- **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- De facto standard for conducting data mining projects.
- Defines tasks and outputs.
- Now developed by IBM as the Analytics Solutions Unified Method for Data Mining/Predictive Analytics (**ASUM-DM**).
- SAS has **SEMMA** and most consulting companies use their own process.



https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

Tasks in the CRISP-DM Model

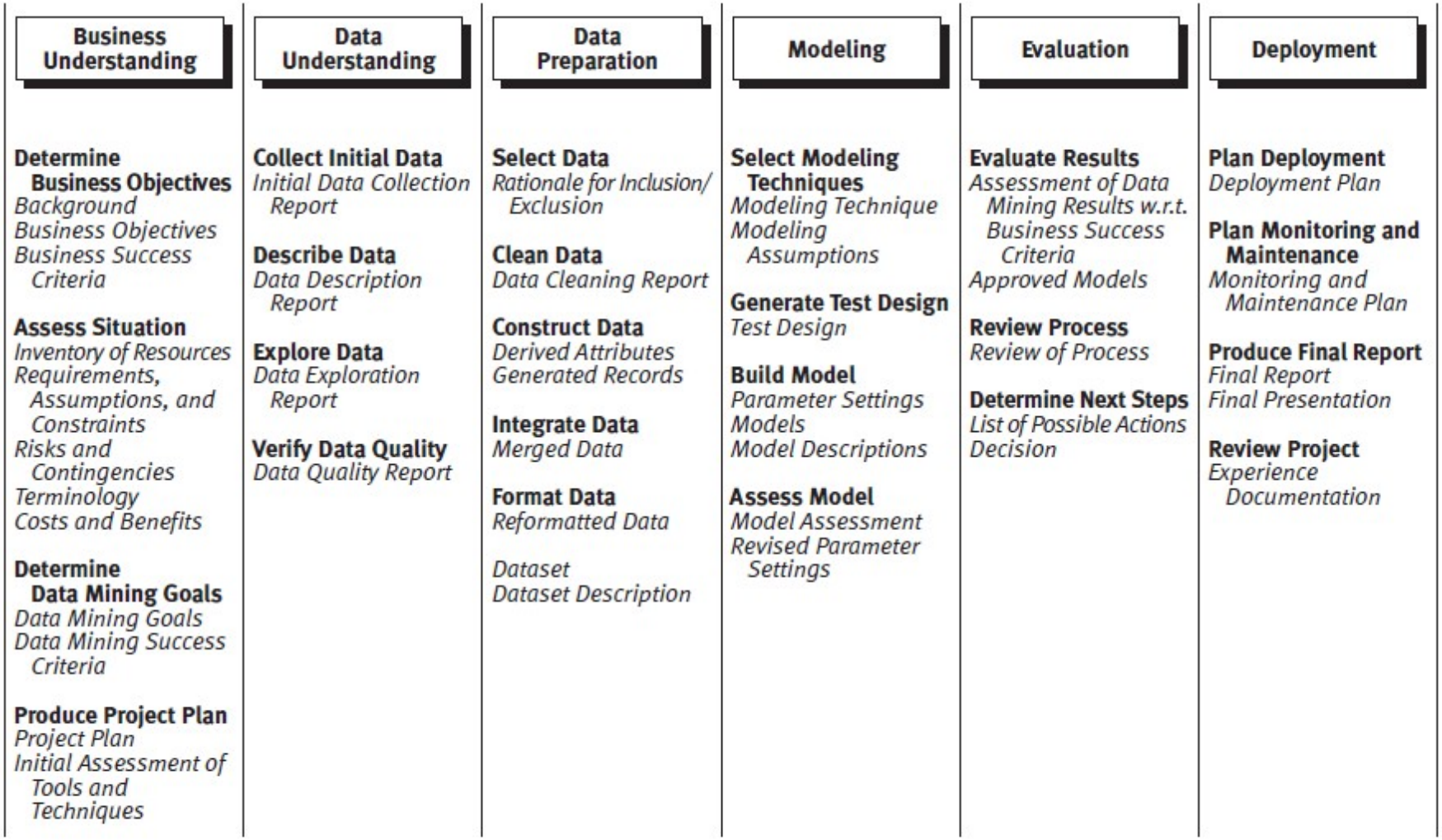


Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Data Mining és un procés... iteratiu e interactiu

- La mineria de dades és un procés iteratiu i interactiu.
- Estigueu preparats per a generar una gran quantitat de "escombraries" fins arribar a alguna cosa que sigui susceptible de recurs i de ser significativa i útil.

Data Mining és un procés... amb pre-condicions

- ✓ Hi ha d'haver un problema ben definit
- ✓ Les dades han d'estar disponibles
- ✓ Les dades han de ser pertinents, adequades i netes
- ✓ El problema no ha de poder ser resolt per mitjà de consulta ordinàries, OLAP (*On-Line Analytical Processing*) o altres eines de bases de dades

- *Els resultats han de ser validables!*

Element fonamental

Es tracta de deixar parlar les dades, perquè tenen molt a dir. Però:

1- En son moltes

2- les dades poden ser sorolloses, incompletes, heterogènies, amb dades irrellevants, etc.

Característiques de les dades originals (*raw data*)

- Dades desconegudes,
- Dades perdudes,
- Dades errònies
- Dades heterogènies,
- Amb diferents estructures i formats
- Redundants
- Irrellevants
- Amb components implícits temporals i espacials (canvis de població estadística)

Pre-processament de les dades

Les dades en el món real són "brutes".
Alguns exemples:

- Incompletes / perdudes (*missing*) : Atribut no té valors e.g. occupació=""
- Sorolloses (*noisy*): contenen errors, alguns detectables
  - e.g. Salari="-10"
- inconsistents: Tenen discrepancies en valors o en els codis
  - e.g. Edat="42" Data Naixement="03/07/1997"
  - e.g. Es valorava "1,2,3", i ara com "A, B, C"
  - e.g. Discrepancies en registres repetits

Pre-pocessament de les dades

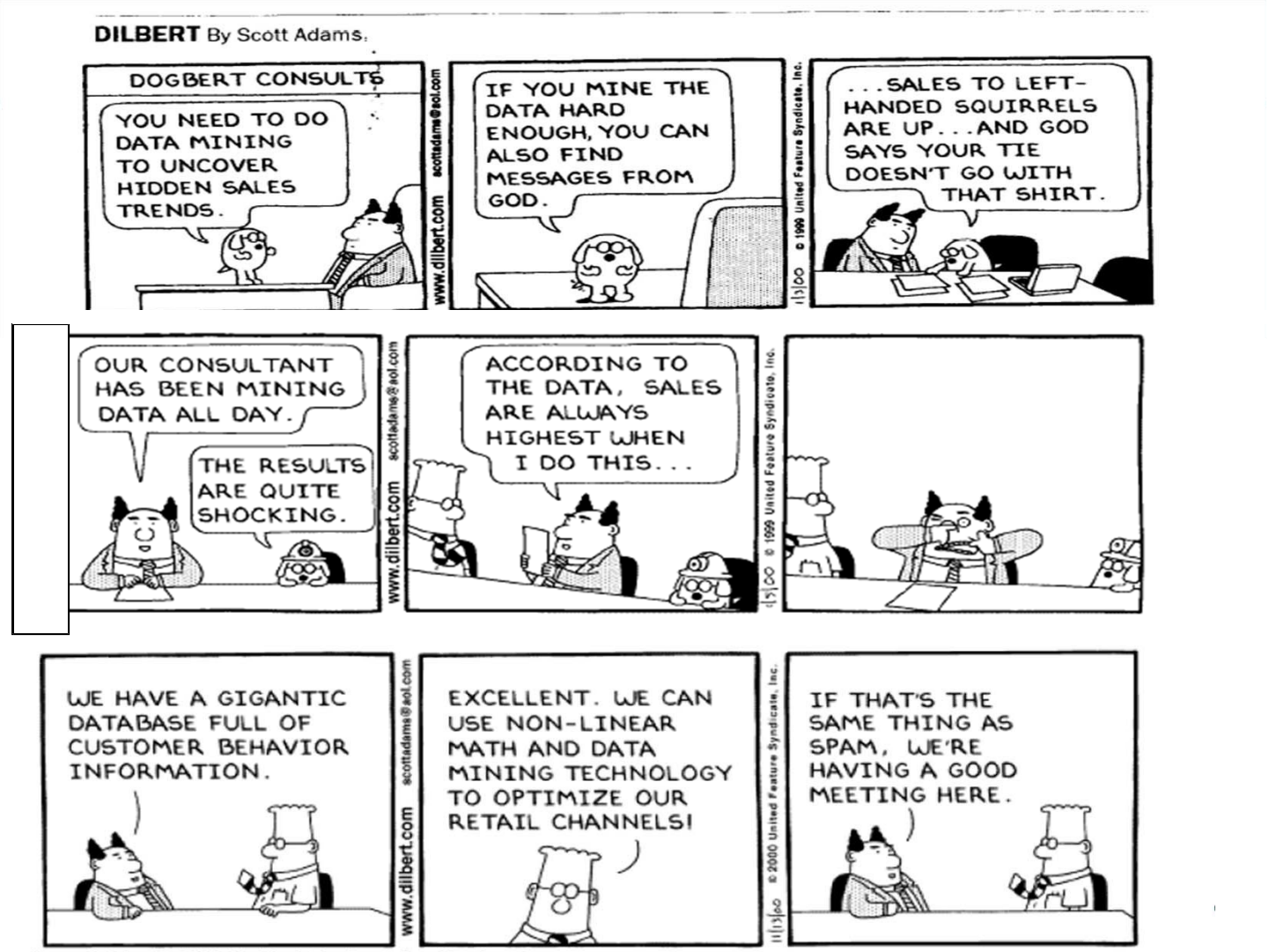
- És necessari un pre-processament de les dades
- Per cada problema descrit és necessari prendre un criteri de processament
- Algunes solucions
  - Eliminar dades inconsistents, sorolloses o perdudes
  - Intentar omplir correctament les dades inconsistents o perdudes
  - ...

Pre-pocessament de les dades

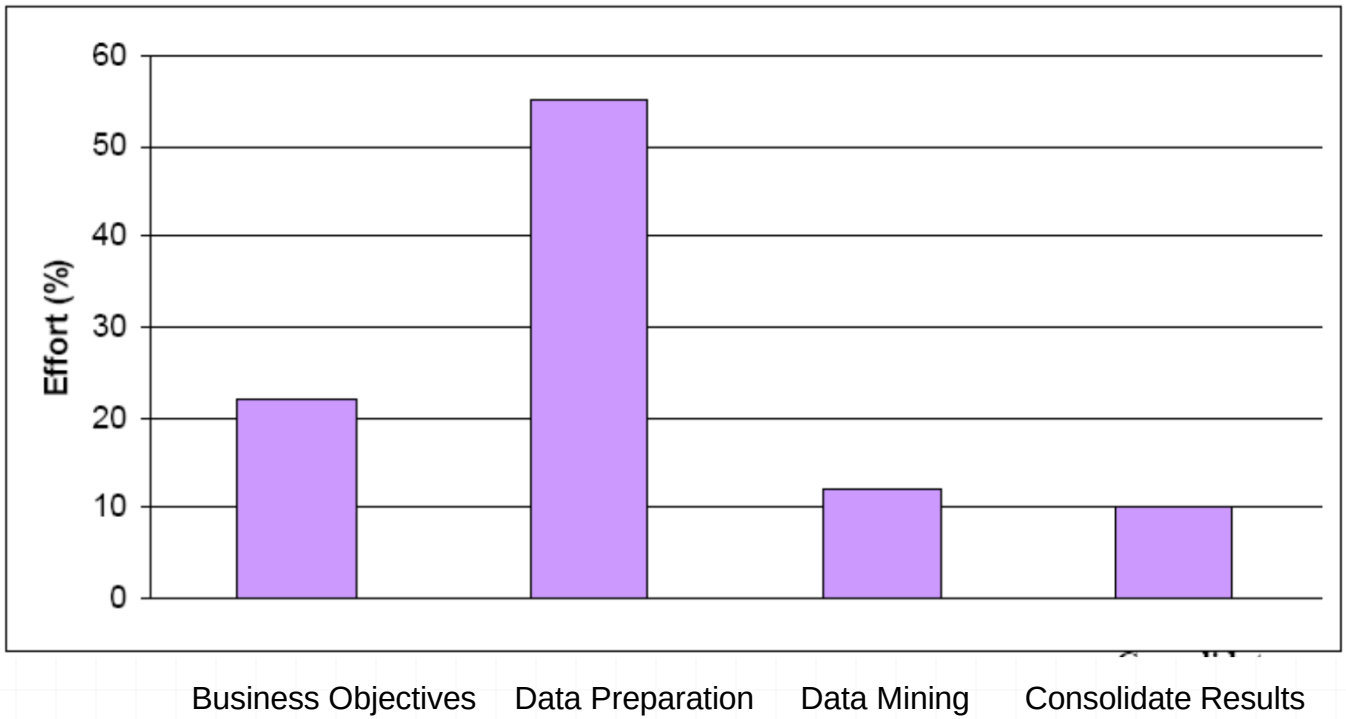
Altres ALGORISMES de **PREPROCESSAMENT:**

- Escalat i Normalització
- Codificació
- Detecció i eliminació de valors atípics (*Outlier Detection & Removal*)
- Selecció i Composició de Característiques (*Feature Selection & Composition*)
- Neteja de dades
- Eliminació de dades errònies o inconsistents
- Suavitzat de les dades
- Mostreig (Sampling)

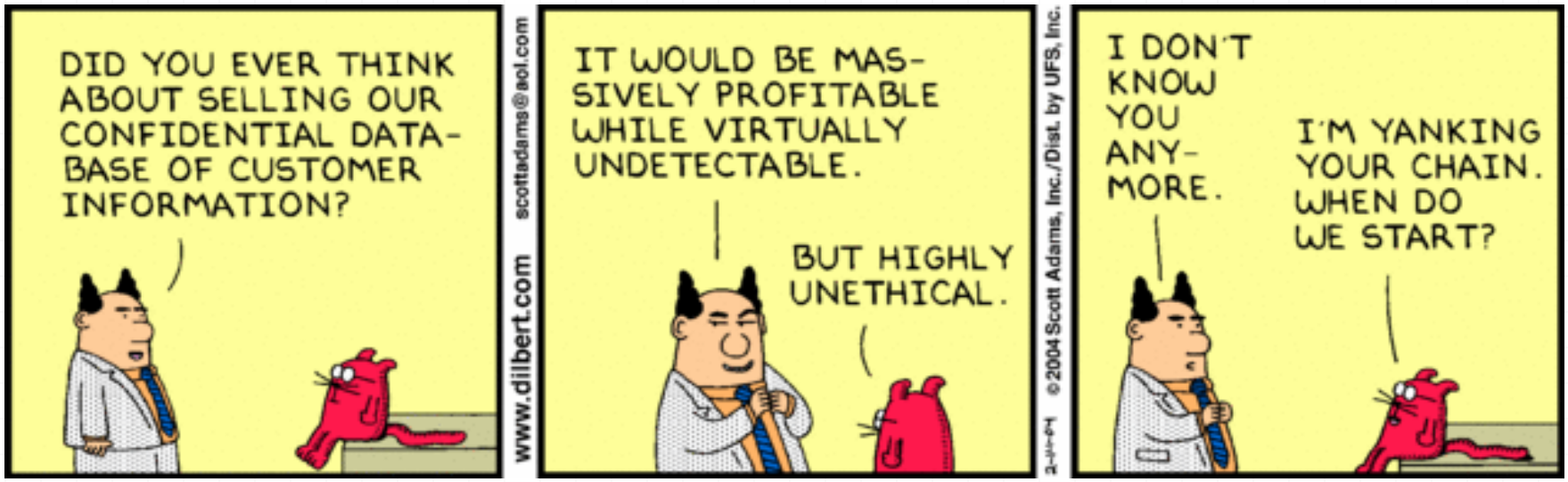
Verificable



Distribució d'esforços en MD



Legal, Privacy and Security Issues



?

Legal, Privacy and Security Issues

- Are we allowed to **collect** the data?
- Are we allowed to **use** the data?
- Is **privacy** preserved in the process?
- Is it **ethical** to use and act on the data?

- Problem: Internet is global but legislation is local!

Legal, Privacy and Security Issues

The New York Times

Data-Gathering via Apps Presents a Gray Legal Area

By KEVIN J.O'BRIEN

Published: October 28, 2012

BERLIN — Angry Birds, the top-selling paid mobile app for the iPhone in the United States and Europe, has been downloaded more than a billion times by devoted game players around the world, who often spend hours slinging squawking fowl at groups of egg-stealing pigs.

When Jason Hong, an associate professor at the Human-Computer Interaction Institute at Carnegie Mellon University, surveyed 40 users, all but two were *unaware that the game was storing their locations so that they could later be the targets of ads*....


Pokémon
GO

What the small print says...

USA Today Network Josh Hafner, 2:38 p.m. EDT July 13, 2016

Pokémon Go's constant location tracking and camera access required for gameplay, paired with its skyrocketing popularity, could provide data like no app before it.

"Their privacy policy is vague," *Hong said. "I'd say deliberately vague, because of the lack of clarity on the business model."*

...

The agreement says Pokémon Go collects data about its users as a ***"business asset."*** *This includes data used to personally identify players such as email addresses and other information pulled from Google and Facebook accounts players use to sign up for the game.*

If Niantic is ever sold, the agreement states, all that data can go to another company.