UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

SCDM

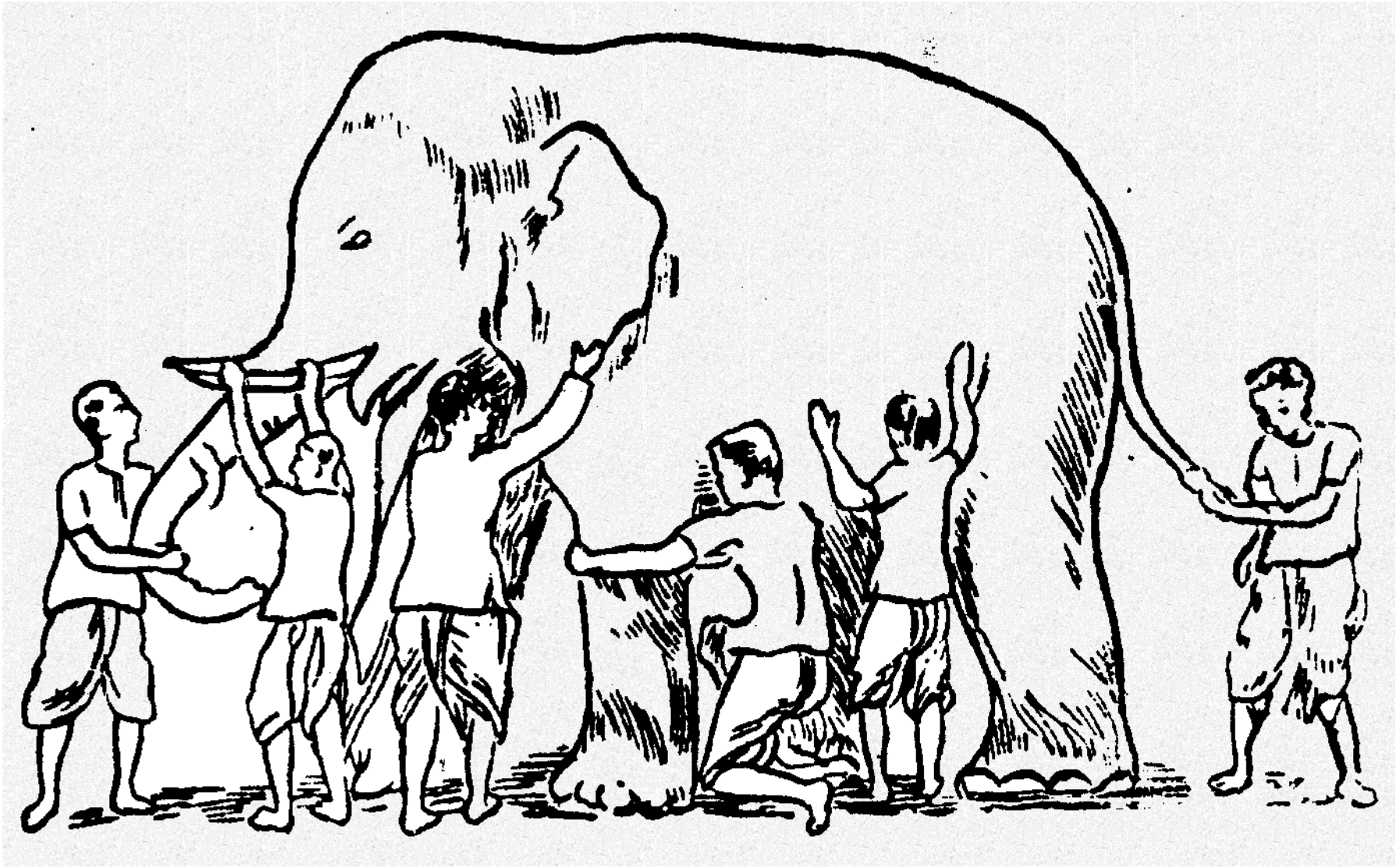
Dels significats de BIG DATA

Ricard Gavaldà
Universitat Politècnica de Catalunya
17 de març de 2015

Sessió “Noves eines per a la gestió de la informació clínica”,
Societat Catalana de Documentació Mèdica

D’on ha vingut el Big Data?

- Disc barat
- Cloud
- Web
- Open data
- Sensorització
- Data Analytics



1. Big Data com a infraestructura

... i com a problema per als informàtics

Definició habitual de Big Data

... però inútil per a la gran majoria

Tenim un problema de Big Data quan el volum o la complexitat de les nostres dades fa impossible tractar-les amb les tecnologies informàtiques (software i hardware) habituals

Molt grans o molt complexes?

Tres V's:

- Volum
- Velocitat
- Varietat

En la meva experiència,

Quan tenim dades però no sabem

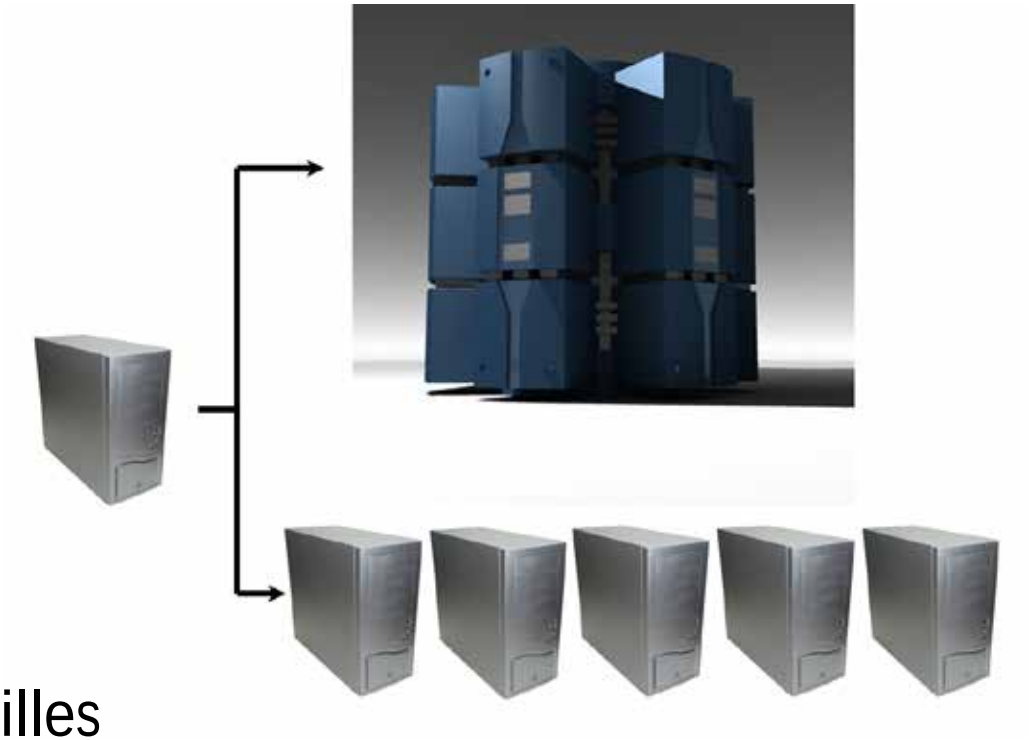
- per on començar a mirar-les,
- quines preguntes formular,
- quin profit podríem treure'n

Les eines habituals?

- BD relacionals - SQL
- Servidors potents

vs.

- BD NoSQL
- Sistemes distribuïts
- Més màquines però senzilles



Altres eines?

- BD no relacionals: Cassandra, Hbase, MongoDB, CouchDB
  - sense gaire esquema predefinit
  - Escalen a desenes...milers de màquines

- Processadors de dades distribuits
  - Hadoop (*però és taaan 2010...)*
  - Spark

2. Big Data com a “ho tenim guardat tot”

Font: http://images.wisegeek.com/elephant.jpg

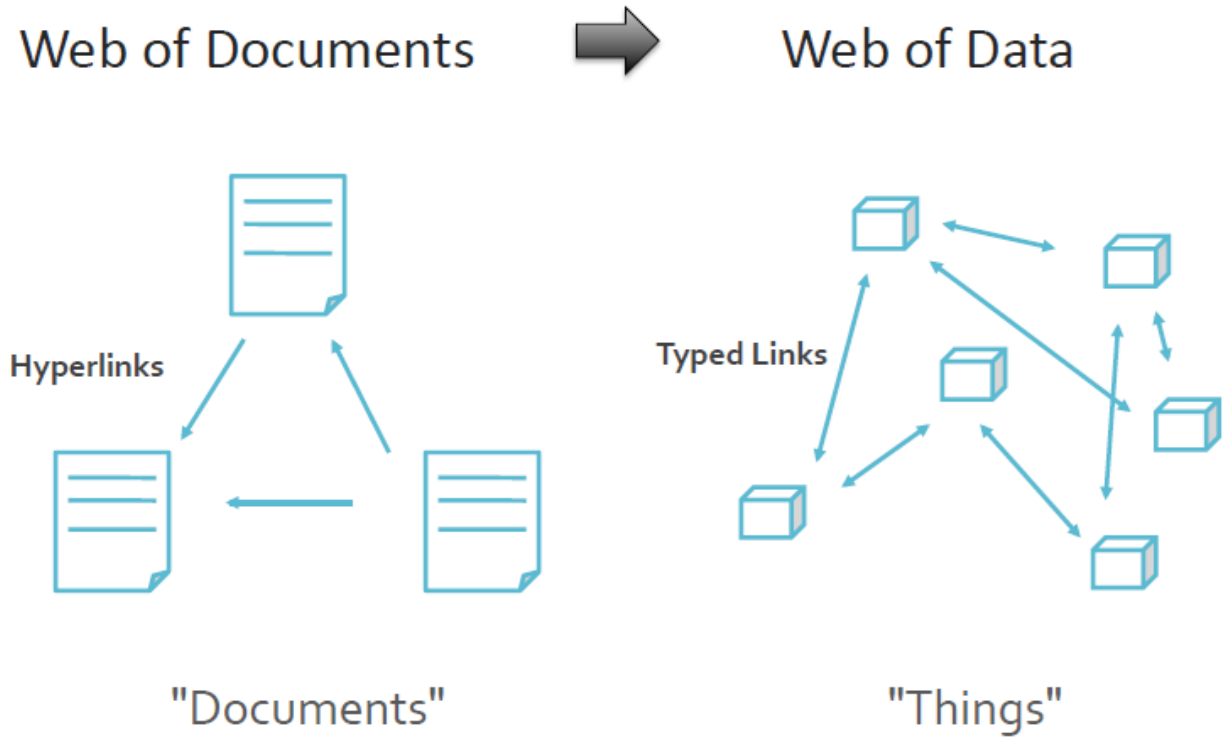
Les meves dades

són

les meves dades

i tot el que hi ha a la Web

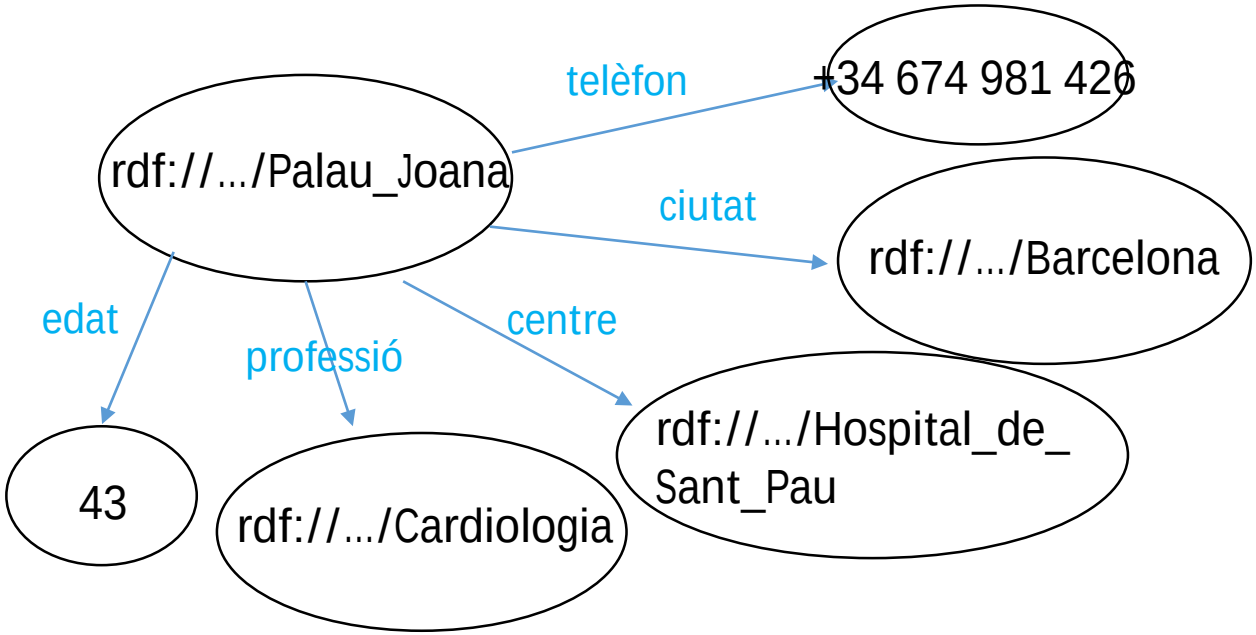
Cap a la web de les dades



Source: Maria-Esther Vidal: A tutorial on Semantic Data Management Techniques for Linked Open Data

Cap a la web de les dades

Nom	Edat	Professió	Centre	Ciutat	Telèfon
Palau, Joana	43	Cardiòloga	Hosp. de Sant Pau	Barcelona	+34 674 981 426



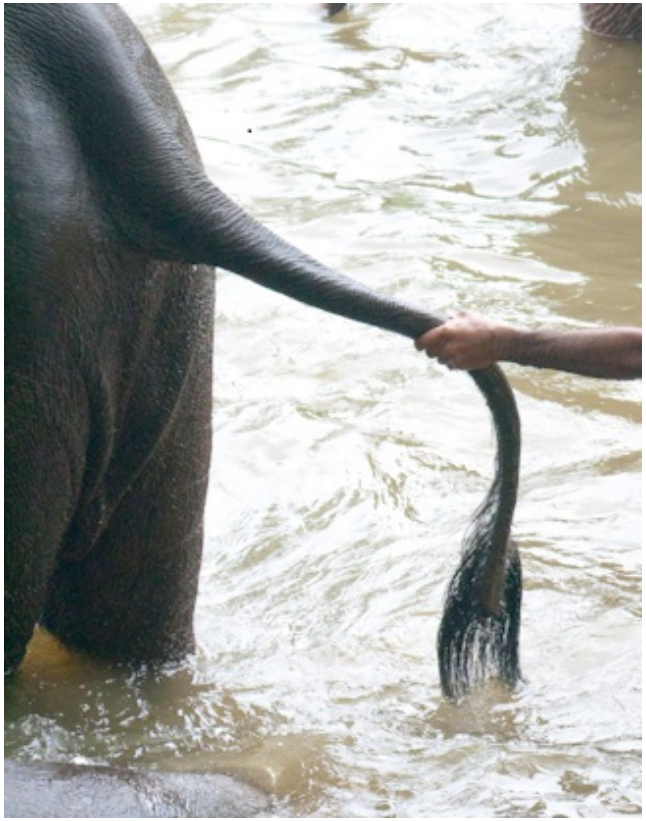
RDF = Resource Description Framework

Les cinc estrelles del Open Data

	Posa la teva informació a la web (qualsevol format)
	Posa-la en un format llegible (Excel i no un escanejat o pdf)
	Usa formats no propietaris (Csv i no Excel)
	Usa URLs per donar noms a les coses
	Enllaça les teves dades a les d'altres per donar-hi context

T. Berners-Lee. Font: http://5stardata.info/

3. Big Data, Big Leak, Big Brother



Font: https://www.flickr.com/photos/photosofsrilanka/3406040079/

Anonimització:

- Sabem que no és fàcil
- *Privacy Preserving Data Mining*

L'efecte porter

4. Big Data = font de coneixement nou i útil

- Dades ≠ Coneixement ≠ Saviesa
- El valor no és en les dades, és en el coneixement
- I sobretot el coneixement que permet triar accions

Tecnologies

- Estadística, és clar
- Machine Learning, Data Mining
- Anàlisi de dades en xarxa
- Sistemes recomanadors
- Comprensió de la parla
- Processament del Llenguatge Natural
- Visualització d'informació complexa
- Anàlisi en temps real

Tecnologies

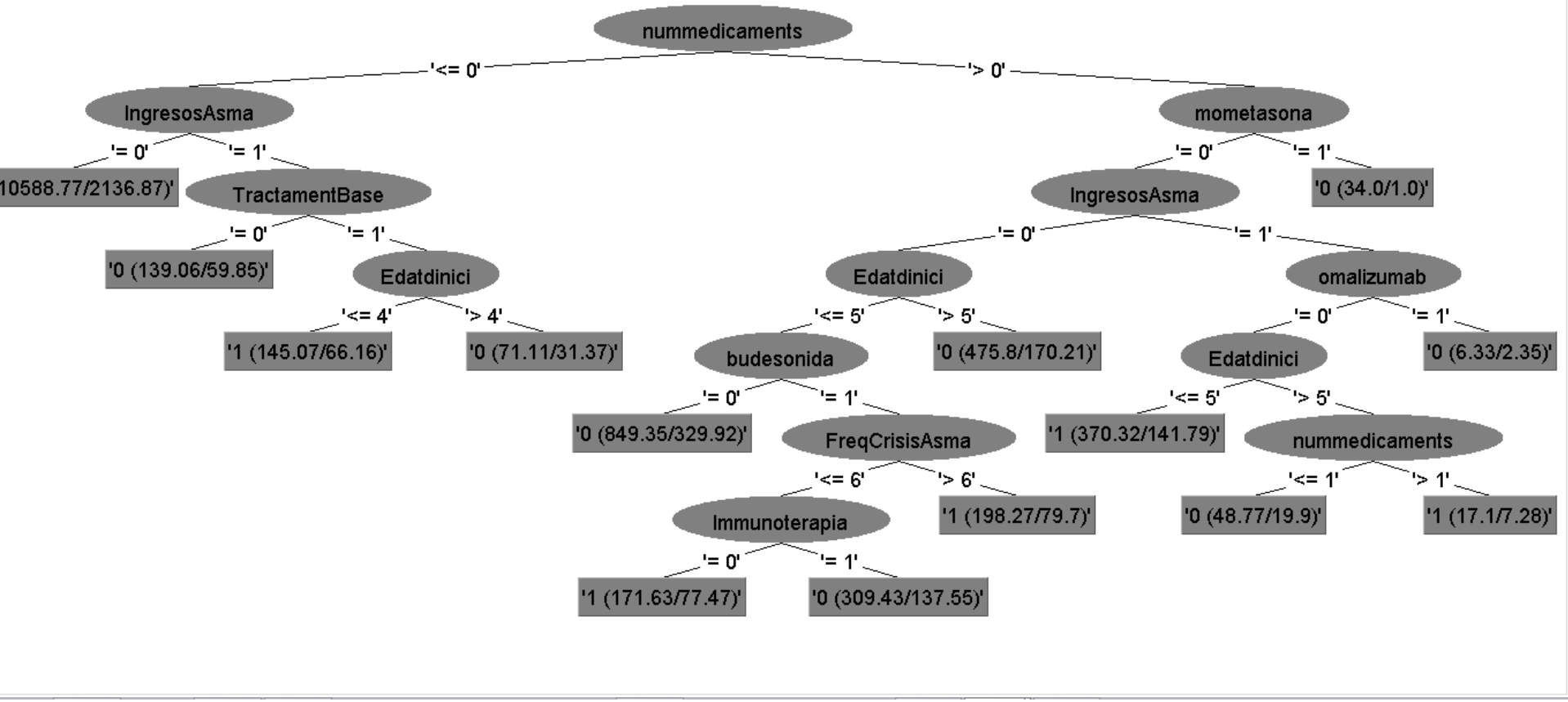
- Sí: Trobar relacions subtils, no evidents
- Sí: Trobar respostes a preguntes no formulades
- No: Trobar variables correlacionades, trobar la tendència, calcular una mitjana

Predicció

- Trobar un model que prediu una variable de sortida en funció de les d'entrada
- Entrada: variables categòriques, numèriques, text, ...
- Algorismes per trobar el millor model d'un tipus desitjat
- Molts tipus de models possibles: Regressió lineal i quadràtica, arbres de decisió, xarxes baiesianes, xarxes neuronals, "support vector machines", "deep learning"....

Exemple: arbre de decisió

- Molt no lineal
- Fàcil de construir = de trobar el millor arbre
- Fàcilment interpretable
- Classifica els individus en grups, implícitament


nummedicaments
'<= 0'
'> 0'
IngresosAsma
mometasona
'= 0'
'= 1'
'= 0'
'= 1'
10588.77/2136.87)'
TractamentBase
IngresosAsma
'0 (34.0/1.0)'
'= 0'
'= 1'
'= 0'
'= 1'
'0 (139.06/59.85)'
Edatdinici
Edatdinici
omalizumab
'<= 4'
'> 4'
'<= 5'
'> 5'
'= 0'
'= 1'
'1 (145.07/66.16)'
'0 (71.11/31.37)'
budesonida
'0 (475.8/170.21)'
Edatdinici
'0 (6.33/2.35)'
'= 0'
'= 1'
'<= 5'
'> 5'
'0 (849.35/329.92)'
FreqCrisisAsma
'1 (370.32/141.79)'
nummedicaments
'<= 6'
'> 6'
'<= 1'
'> 1'
Immunoterapia
'1 (198.27/79.7)'
'0 (48.77/19.9)'
'1 (17.1/7.28)'
'= 0'
'= 1'
'1 (171.63/77.47)'
'0 (309.43/137.55)'

Cerca de patrons freqüents, anomalies i regles

- No intenten expliquen la majoria de les dades
- petits nínxols, casos minoritaris però interessants

- Malaltia A + Malaltia B à Malaltia C (1%, 55%)
- Edat > 70 + Antecedents familiars à Mal pronòstic (2%, 30%)

Exemple: Una variable resposta és certa quan 5 variables d'entrada d'entre 100 són certes

- 100x100x100x100x100 = 10.000 milions de combinacions

I què obtenim?

- Llista de patrons, models, regularitats, anomalies, ... que l'expert pot
  - Interpretar
  - Verificar
  - Descartar
  - Incorporar al sistema d'informació

Alternativa: sistema autònom en temps real

I en l'àmbit de la salut? Infinit

- Històries clíniques + analítiques + imatge + òmiques + dades de gestió +...
- Internet de les coses (sensorització) + xarxes socials + dades interconnectades + anàlisis en temps real
- Visió global del sistema + visió local (pacient)

Moltes gràcies!

- gavalda@cs.upc.edu
- http://www.cs.upc.edu/~gavalda