UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

SOCIETAT CATALANA
D'ESTADÍSTICA

Analítica en temps real: el model de "Data Streams"

Ricard Gavaldà
Dept. de Ciències de la Computació
Universitat Politècnica de Catalunya

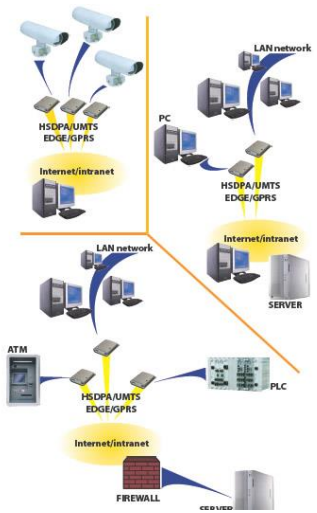
L'Estadística de les Coses
Jornada de primavera de la Societat Catalana d'Estadística
5 de juny de 2019

Índex

- De què parlem quan parlem de...
- El model de “Data Stream”
- Exemple: models predictius
- Conclusió: Un canvi en el procés

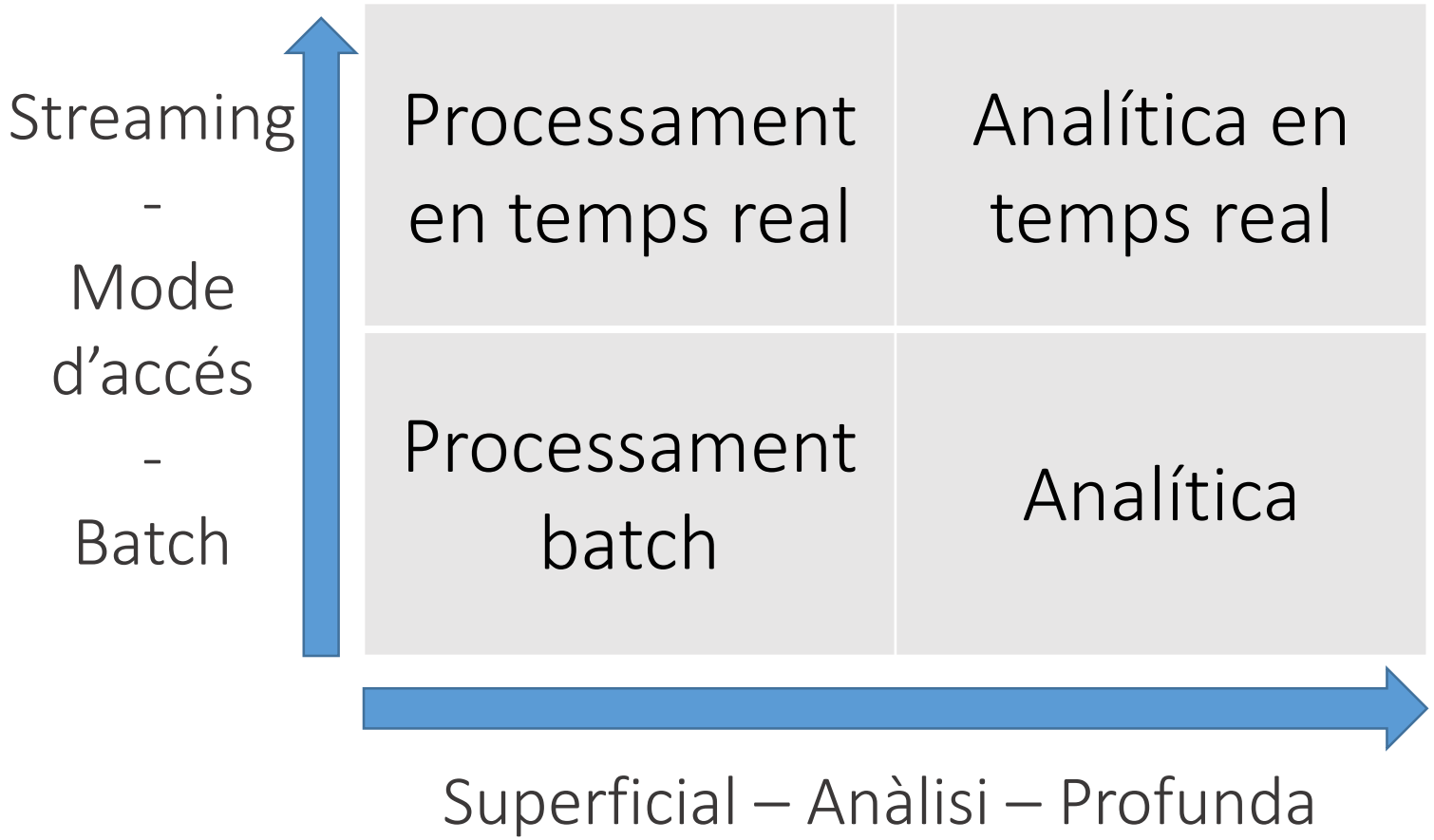
Analítica en temps real

Per què?

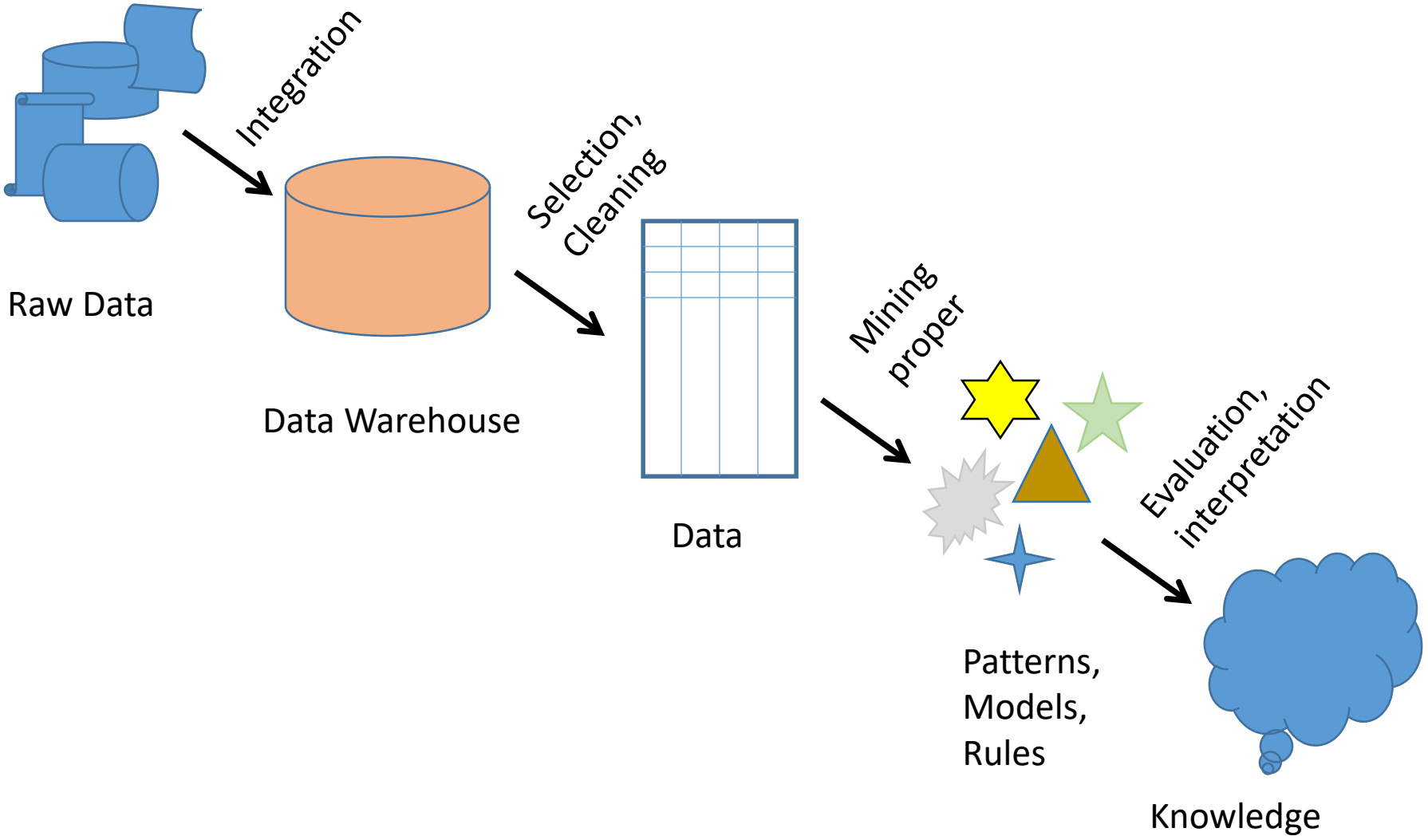


- Telcos – trucades telefòniques
- Satèl·lits, radar,
- Xarxes de sensors; urbanes i domòtiques
- “Wearables”, implants…
- Monitoratge de xarxes i “datacenters”
- “Logs” de cerca i accessos
- Comerç electrònic
- Activitat en xarxes socials
- …

Quatre quadrants



El procés clàssic en Data Science



La realitat canvia

Exemples recurrents

- Activitat d'una persona gran a casa
- Activitat urbana – vianants i vehicles
- Fuga de clients de comerç electrònic
- Propagació de notícies en xarxes socials

El model de “Data Streams”

Data Stream Analysis

"Ginger Rogers did everything Fred Astaire did, but she did it backwards and in high heels"

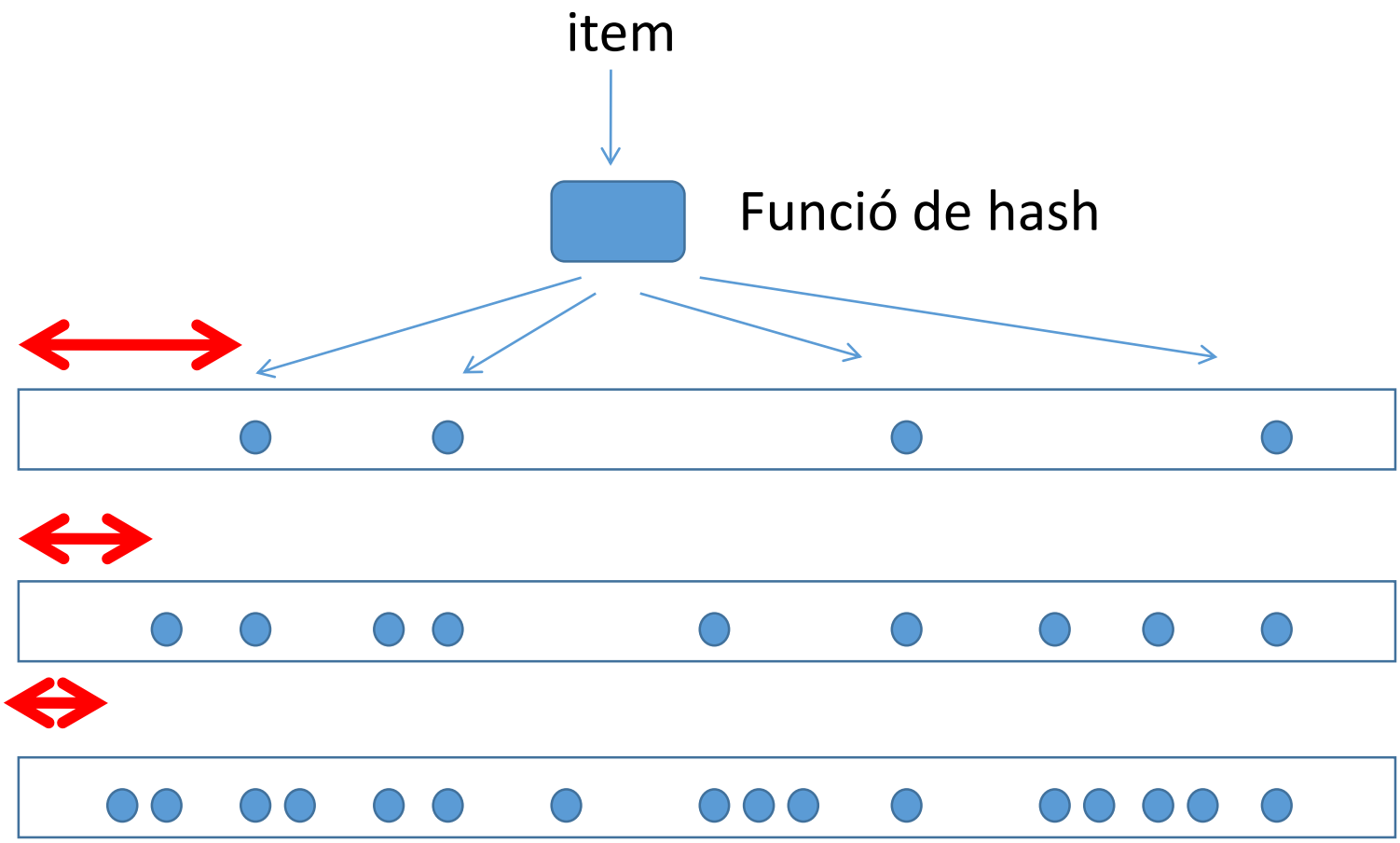
5 axiomes dels “Data Streams”

1. Dades seqüencials, una sola passada
2. Temps constant per element llegit
3. Memòria independent de la llargària
4. Respostes en qualsevol moment
5. S’adapta als canvis de la realitat

Limitacions computacionals

- Mostreig
- “Load shedding”
- “Sketches”: Petits algorismes que guarden estadístics suficients amb poc temps i memoria
- Paradigma del “10% error el 99% dels cops”

Sketching: Elements diferents



Sketching: Elements diferents

Algorithm:

For each item x: If f(x) < min then min = f(x)

Claim: with probability >3/4,

$$0.2\ D \leq \min \leq 4\ D$$

Trick: Run $\frac{4}{\varepsilon^2}\ln\frac{1}{\delta}$ copies in parallel, $f_1 \ldots f_k$

Then take the median

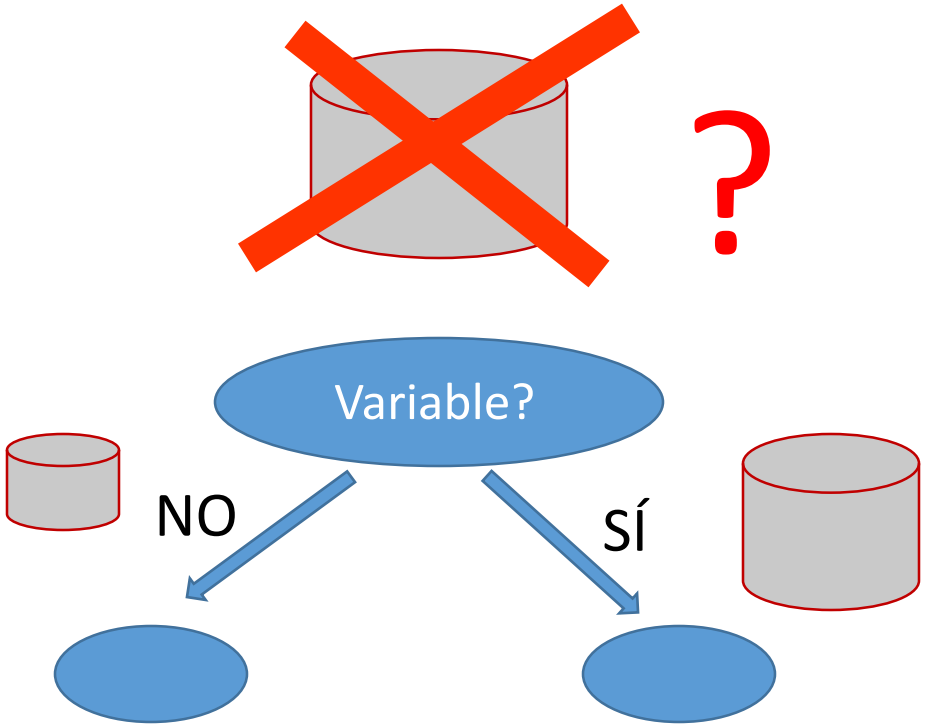
This gives an approximation within **ε** a fraction **1-δ** of times

Gestió del canvi

- Tests d'hipòtesi (CUSUM)
- Finestres de mida fixa
- Finestres de mida variable
- Pesos decreixents (EWMA)

Exemple: Models predictius

Exemple: Arbres de decisió



1. Tria variable
2. Divideix dades
3. Construeix fills recursivament

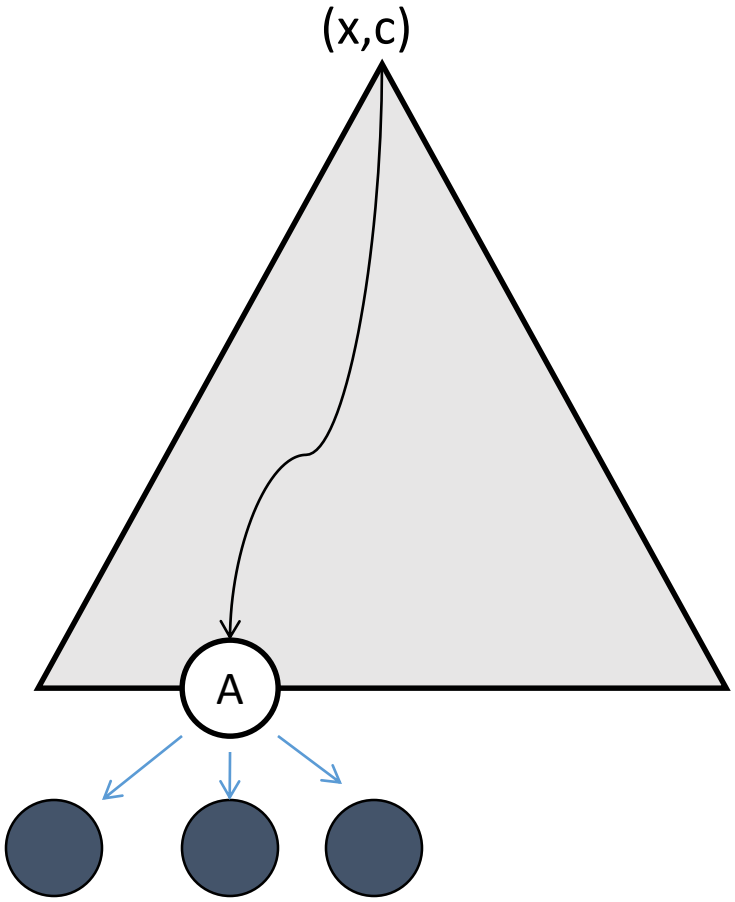
[Domingos-Hulten 2000]

VFDT

Acumulem exemples (i evidència) a les fulles

Només quan hi ha prou evidència per triar la millor, expandim la fulla

(Test d'hipòtesi "és A millor que B?")

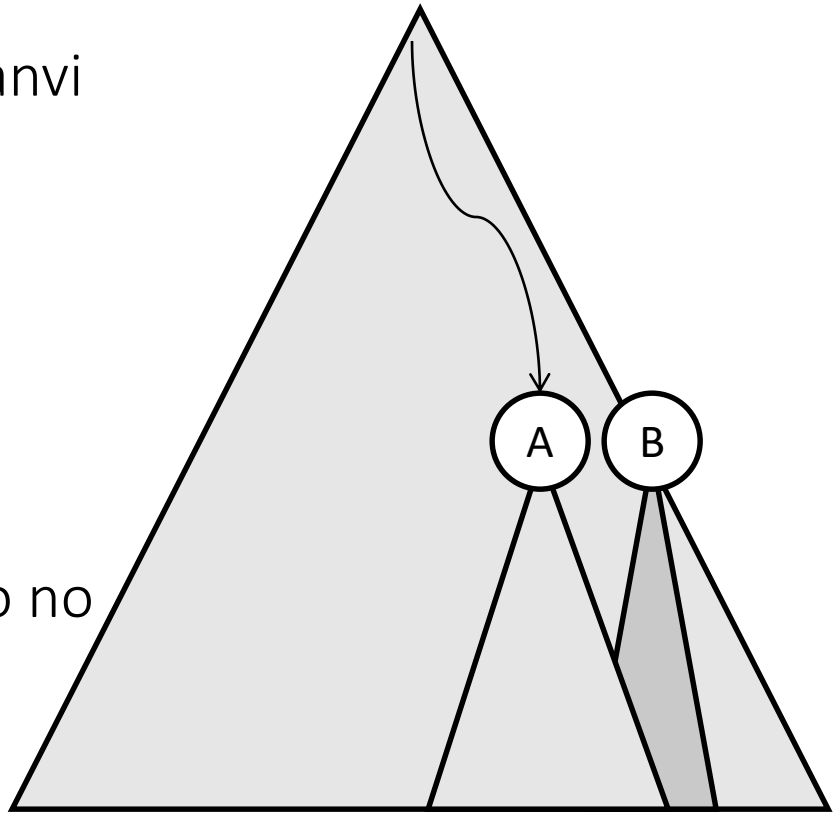


CVFDT – Revisem?

Monitorizem si sembla que hi ha canvi en un node

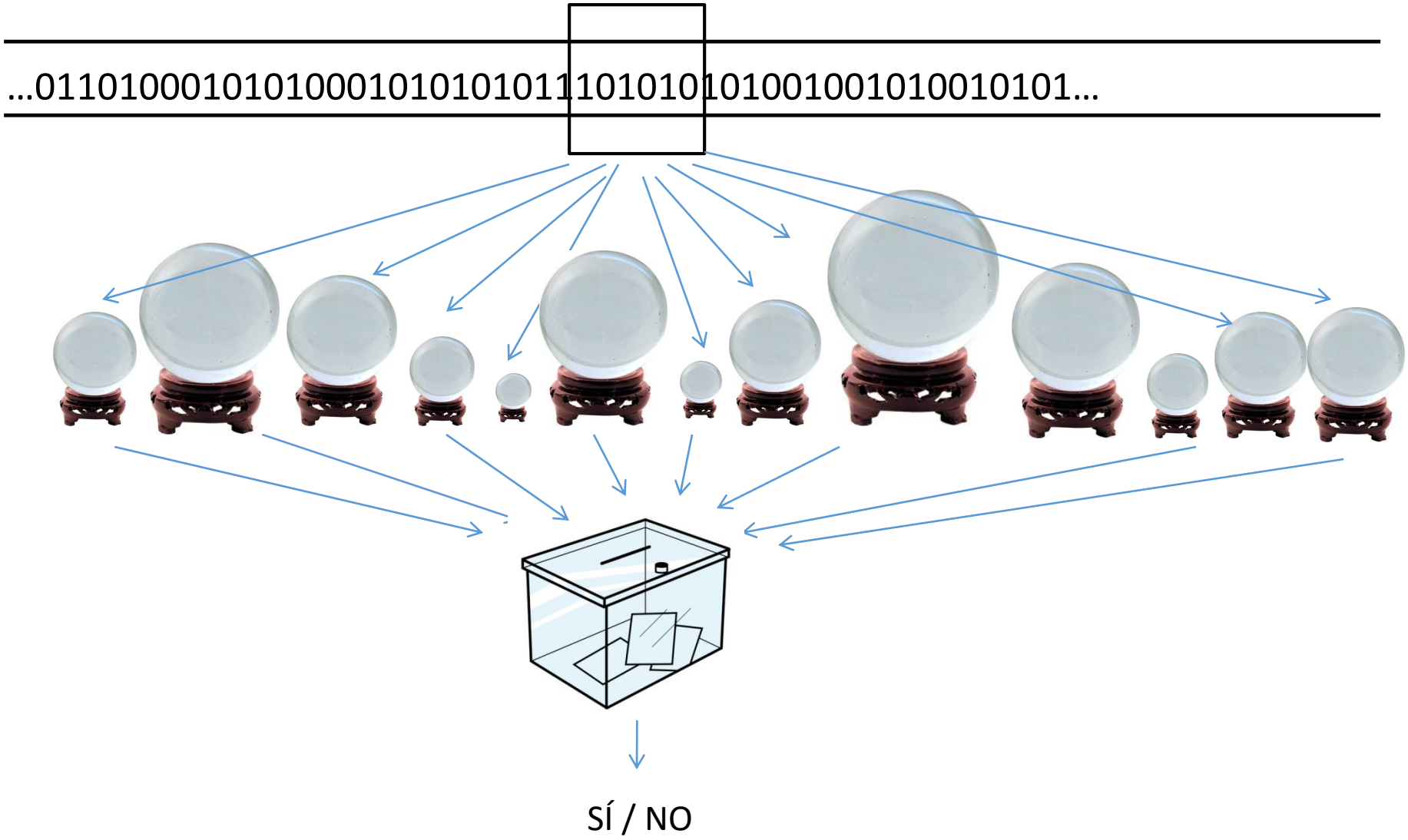
Si sí, hi comencem a fer crèixer un subarbre alternatiu

Més endavant decidim si canviem o no

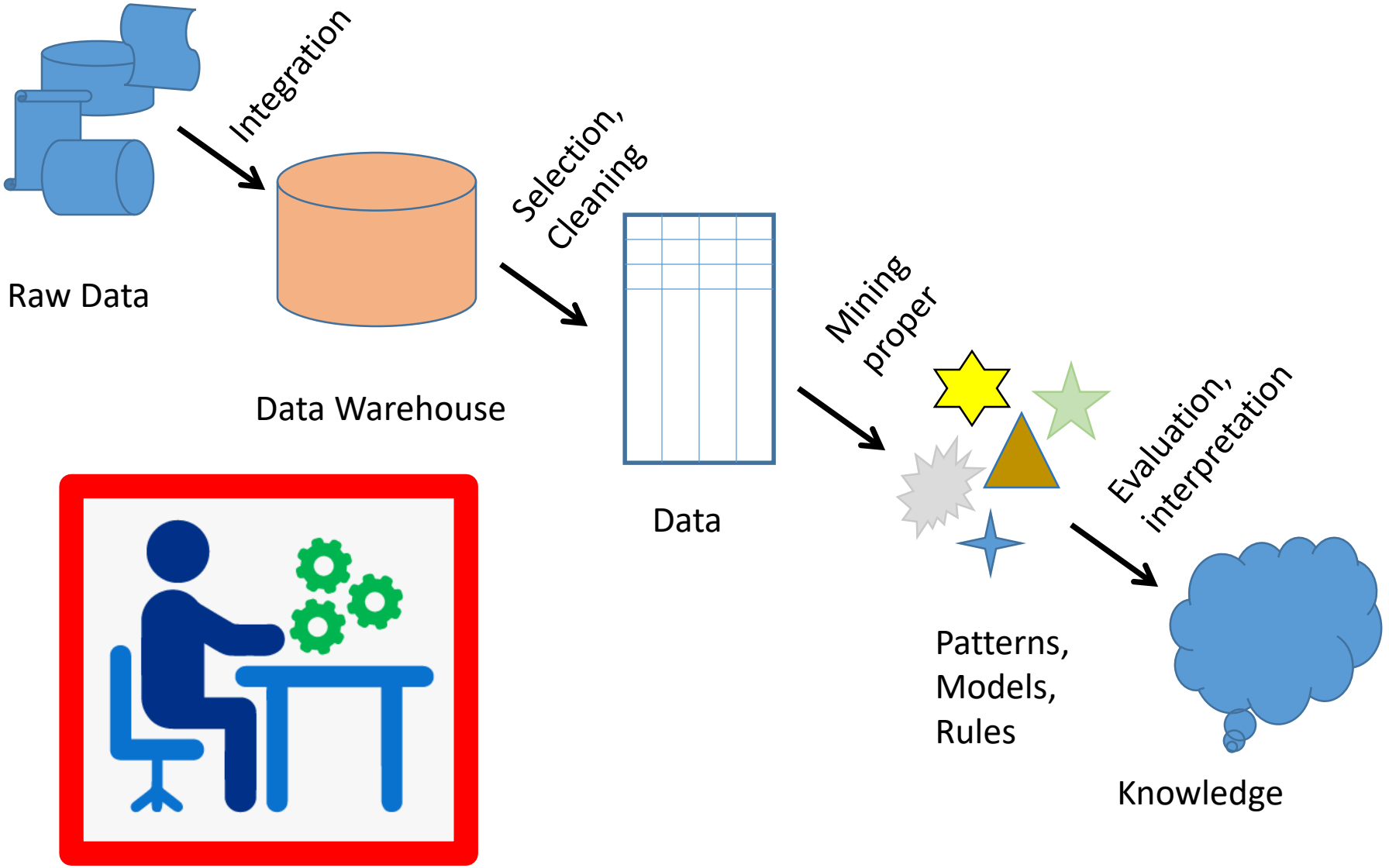


[Hulten-Spencer-Domingos 2001]

Ensemble – voting methods



El procés clàssic en Data Science



El nou procés

1. Arriba un element x
2. Fem una predicció de F(x)

Al cap d'una mica de temps… o *molt* temps…. o *mai*…

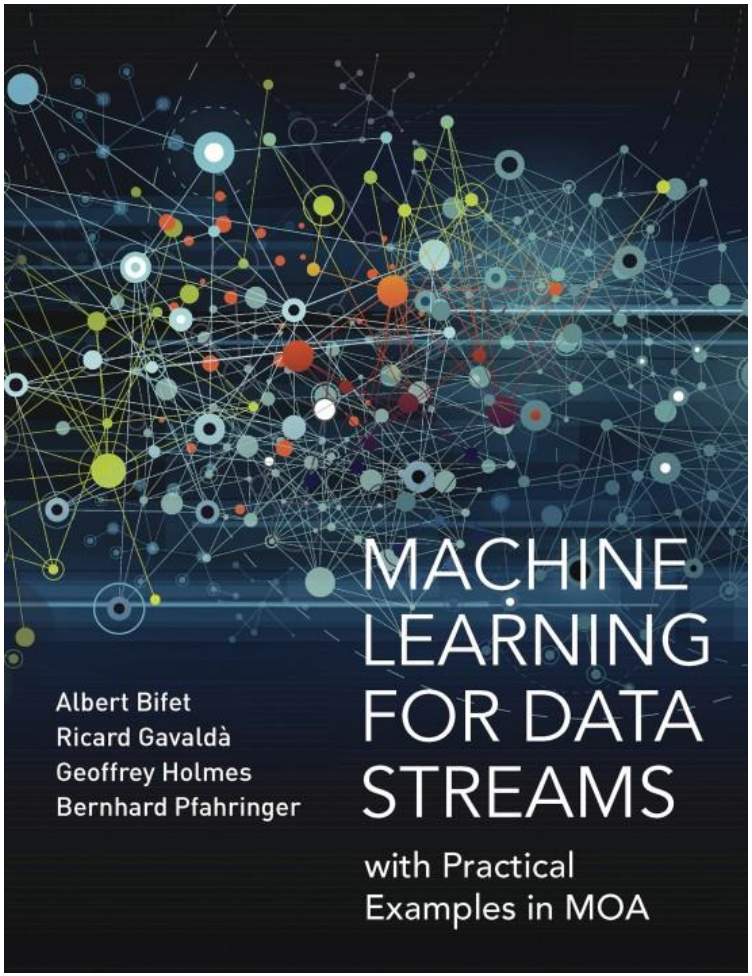
3. Arriba el valor real de F(x)
4. Revisem el model

Poden aparèixer variables, clases, …
Poden variar els bons hiperparàmetres

El procés que feia l'analista s'ha de fer autònomament

https://mitpress.mit.edu/books/machine-learning-data-streams

https://moa.cms.waikato.ac.nz/book/



closed ones, or maximal ones. Of course, the summary is also designed from efficiency considerations. The key operations are *add* (C, C') and *remove* (C, C'). Here, C and C' are summary structures representing two pattern datasets D and D'. The operations should return the summaries corresponding to $D \cup D'$ and $D \setminus D'$, or approximations thereof. Here $D \setminus D'$ denotes multiset difference, the multiset that contains $\max\{0, supp_D(p) - supp_{D'}(p)\}$ copies of each pattern p . As mentioned already, exact algorithms pay a high computational price for exactness, so *add* and *remove* operations that are not exact but do not introduce many false positives or false negatives may be of interest. No comments

Be the first to comment on this paragraph!

Comment

Your name **Required**

Your URL **Optional:** link to blog, home page, *etc*.

Remember you?

Submit Comment

BATCHUPDATEPATTERNMINER(*Stream*, b, w, σ)

Input: a stream of patterns, a batch size b,

¡Moltes gràcies!

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