Data Mining for the XXI Century
PART III

João Gama
jgama@fep.up.pt

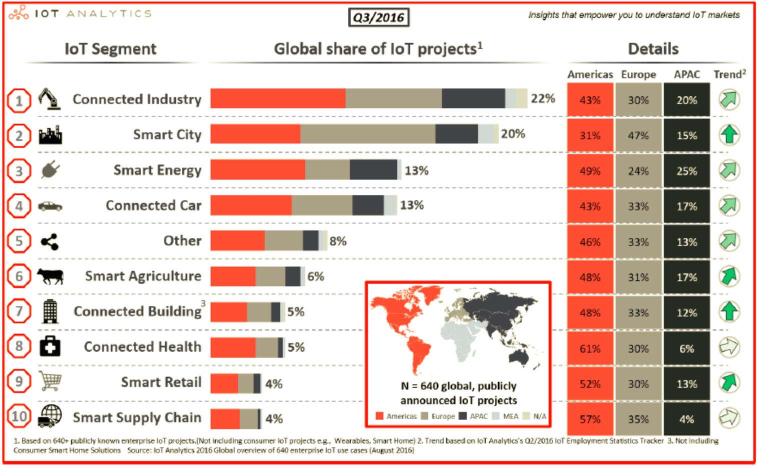
INESC TEC, FEP-University of Porto, Portugal

Outline

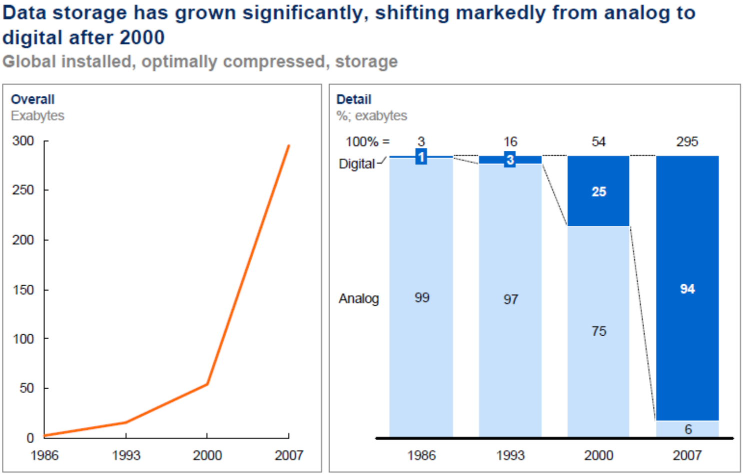
Industry 4.0

We have machines that collect, process, and send information to other machines

Internet of Things



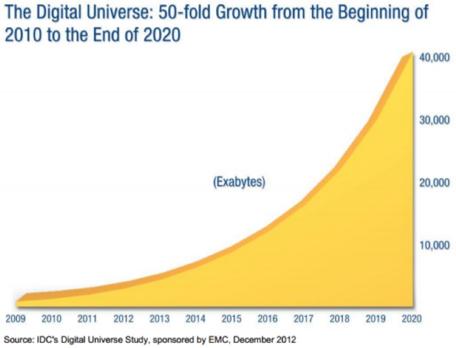
The Big Bang of digital data ...



NOTE: Numbers may not sum due to rounding.

SOURCE: Hilbert and López, "The world's technological capacity to store, communicate, and compute information," *Science*, 2011

The Growth of Digital Data...



Memory unit	Size	Binary size
kilobyte (kB/KB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

Tools: time ago ...

Tools seemed quite powerful

Tools

Problems

Tools: nowadays ...

Last few years

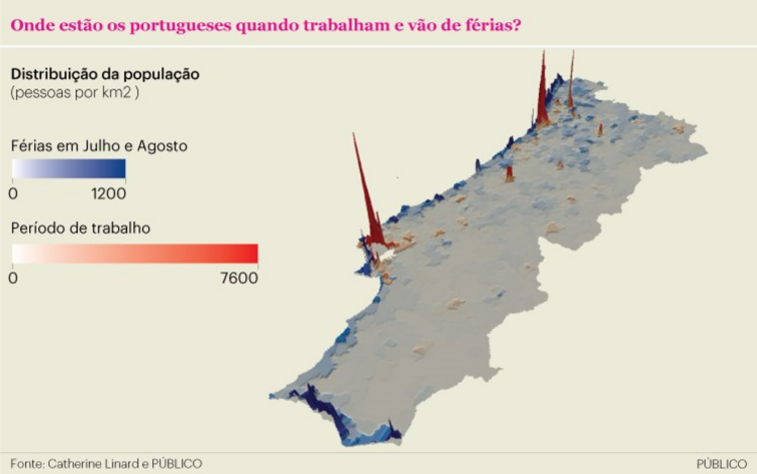
The Model has Changed ...

- **The Model of Generating/Consuming Data has Changed**

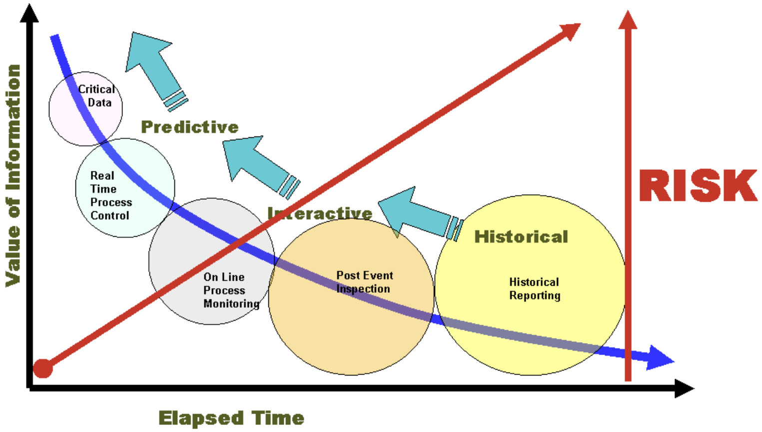
Old Model: Few companies are generating data, all others are consuming data

New Model: all of us are generating data, and all of us are consuming data

An Illustrative Example: Real-time Census ...



The Value of Information ...



Main Goal: Understanding Data

A brief history of big data, the Noam Chomsky way

CNBC

Text Size

Published: Saturday, 23 Nov 2013 | 7:00 AM ET

By: Eric Rosenbaum | CNBC.com

Recommend 47 Twitter 104 +1 6 LinkedIn 15 Share

ChinaFotoPress | Getty Images

Noam Chomsky

The latest news from the fast-evolving world of the ***Data Economy****:*

For those familiar with Noam Chomsky, the pioneering linguist whose theory of recursion seeks to find the universal in all human languages, you probably also know that Chomsky often has not-so-nice things to say about the U.S. government, and has also made a career of finding the universal

Big data is a step forward, but our problems are not lack of access to data, but understanding them. Big data is very useful if I want to find out something without going to the library, but I have to understand it, and that's the problem.

A World in Movement

- The new characteristics of data:
  - **Time and space**: The objects of analysis exist in time and space. Often they are able to move.
  - **Dynamic environment**: The objects exist in a dynamic and evolving environment.
  - **Information processing capability**: The objects have limited information processing capabilities
  - **Locality**: The objects know only their local spatio-temporal environment;
  - **Distributed Environment**: Objects will be able to exchange information with other objects.
- Main Goal:
  - **Real-Time Analysis**: decision models have to evolve in correspondence with the evolving environment.

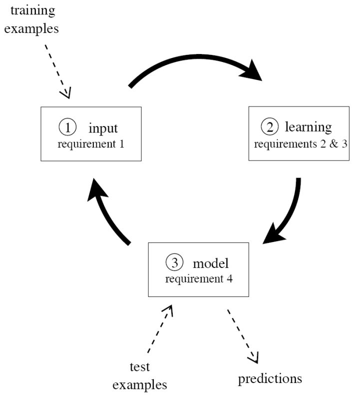
The Challenges of Real Time Data Mining

These characteristics imply:

- Switch from **one-shot learning** to continuously learning **dynamic models** that evolve over time.
- In this context, *finite training sets, static models, and stationary distributions* will have to be completely thought anew.
- Computational resources are finite. Algorithms will have to use *limited computational resources* (in terms of computations, memory, space and time, communications).

Data Stream Computational Model

1. One-pass algorithms:
   random access to data has
   high cost
2. Limited computational
   resources:
   time, memory, bandwidth
3. Anytime prediction

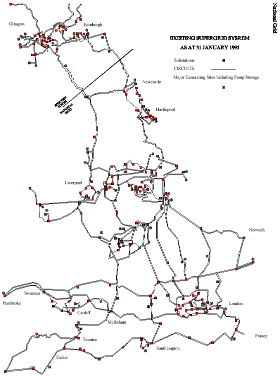


Sparkling Ideas

- Summarization:
  Compact summaries to store sufficient statistics
  and fast update rules
- Approximation:
  How much data we need to learn an hypothesis $\hat{H}$ that, with high probability, is within small error of the true hypothesis ?
  $Pr(|H - \hat{H}| < \epsilon|H|) > 1 - \delta$
- Monitoring the learning process: Estimation and Change detection

Outline

Scenario



Electrical power Network: Sensors all around network monitor measurements of interest.

Scenario

- Sensors produce continuous flow of data at high speed:
  - Send information at different time scales;
  - Act in adversary conditions: they are prone to noise, weather conditions, battery conditions, etc;
- Huge number of Sensors, variable along time
- Geographic distribution:
  - The topology of the network and the position of the sensors are known.

Illustrative Learning Tasks:

- Cluster Analysis
  - Identification of Profiles: Urban, Rural, Industrial, etc.
- Predictive Analysis
  - Predict the value measured by each sensor for different time horizons.
  - Prediction of peaks on the demand.
- Monitoring Evolution
  - Change Detection
    - Detect changes in the behavior of sensors;
    - Detect Failures and Abnormal Activities;
  - Extreme Values, Anomalies and Outliers Detection
    - Identification of **critical points** in load evolution;

Standard Approach:

This problem has been addressed time ago:

Strategy

- Select a finite sample
- Generate a static model (cluster structure, neural nets, Kalman filters, Wavelets, etc)
- Very good performance in next month!
- Six months later: Retrain everything!

Standard Approach:

This problem has been addressed time ago:

Strategy

- Select a finite sample
- Generate a static model (cluster structure, neural nets, Kalman filters, Wavelets, etc)
- Very good performance in next month!
- Six months later: Retrain everything!

What is the Problem?

The world is not static!
Things change over time.

The Data Stream Phenomenon

- Highly detailed, automatic, rapid data feeds.
  - Internet: traffic logs, user queries, email, financial,
  - Telecommunications: phone calls, sms,
  - Astronomical surveys: optical, radio,.
  - Sensor networks: many more *observation points* ...
- Most of these data will never be seen by a human!
- Need for near-real time analysis of data feeds.
- Monitoring, intrusion, anomalous activity Classification, Prediction, Complex correlations, Detect outliers, extreme events, etc

Data Streams

Continuous flow of data generated at **high-speed** in **Dynamic**, **Time-changing** environments.

The usual approaches for *querying*, *clustering* and *prediction* use **batch procedures** cannot cope with this streaming setting.

Machine Learning algorithms assume:

- Instances are independent and generated at random according to some probability distribution $\mathcal{D}$.
- It is required that $\mathcal{D}$ is stationary

Practice: *finite* training sets, *static* models.

Data Streams

We need to maintain **Decision models** in **real time**.
Decision Models must be capable of:

- **incorporating** new information at the speed data arrives;
- **detecting** changes and **adapting** the decision models to the most recent information.
- **forgetting** outdated information;

Unbounded training sets, dynamic models.

Outline

Clustering Time Series Data Streams

Goal: Continuously maintain a clustering structure from evolving time series data streams.

- Ability to Incorporate new Information;
- Process new Information at the rate it is available.
- Ability to Detect and React to *changes* in the Cluster's Structure.

Clustering of *variables* (sensors) not examples!
The standard technique of transposing the working-matrix does not work: transpose is a blocking operator!

Online Divisive-Agglomerative Clustering

Online Divisive-Agglomerative Clustering, Rodrigues & Gama, 2008

Goal: Continuously maintain a hierarchical cluster's structure from evolving time series data streams.

- Performs hierarchical clustering
- Continuously monitor the evolution of **clusters' diameters**
- Two Operators:
  - Splitting: expand the structure
    more data, more detailed clusters
  - Merge: contract the structure
    reacting to changes.
- Split and merge criteria are supported by a confidence level given by the **Hoeffding bounds**.

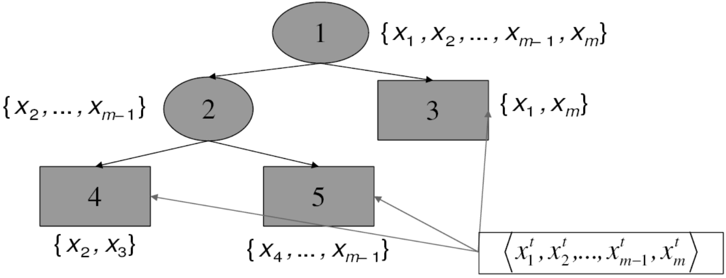
Main Algorithm

- ForEver
  - Read Next Example
  - For all the clusters
    - Update the sufficient statistics
  - Time to Time
    - Verify Merge Clusters
    - Verify Split Cluster

Feeding ODAC

Each example is processed once.
Only sufficient statistics **at leaves** are updated.
Sufficient Statistics: a triangular matrix of the correlations between variables in a leaf.

Released when a leaf expands to a node.



$$C_1 = \{x_2, x_3\}, C_2 = \{x_4, \ldots, x_{m-1}\}, C_3 = \{x_1, x_m\}$$

Similarity Distance

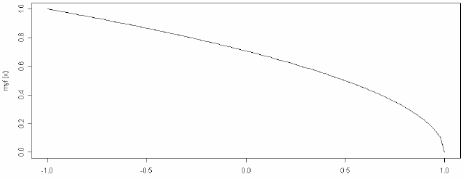
Distance between time Series: $rnomc(a, b) = \sqrt{\frac{1-corr(a,b)}{2}}$

where $corr(a, b)$ is the Pearson Correlation coefficient:

$corr(a, b) = \frac{P - \frac{AB}{n}}{\sqrt{A_2 - \frac{A^2}{n}}\sqrt{B_2 - \frac{B^2}{n}}}$

The *sufficient statistics* needed to compute the correlation are easily updated at each time step:

$A = \sum a_i,\ B = \sum b_i,\ A_2 = \sum a_i^2,\ B_2 = \sum b_i^2,\ P = \sum a_i b_i$

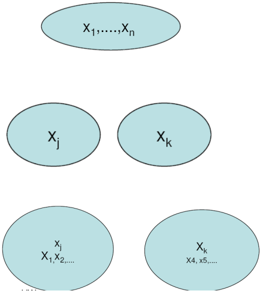


The Splitting Operator: Expanding a Leaf

Step 1 Find Pivots:
$x_j, x_k : d(x_j, x_k) > d(a, b)$
$\forall a, b \neq j, k$

Step 2 If Splitting Criteria applies:
Generate two new clusters.

Step 3 Each new cluster attract nearest variables.



Splitting a Leaf

The base Idea

A small sample can often be enough to choose a near optimal decision
(*Mining High-Speed Data Streams*, P. Domingos, G. Hulten; KDD00)

- Collect sufficient statistics from a small set of examples
- Estimate the merit of each alternative

How large should be the sample?

- **The wrong idea:** Fixed sized, defined *apriori* without looking for the data;
- **The right idea:** Choose the sample size that allow to differentiate between the alternatives.

Splitting Criteria

Expanding a leaf: How large should be the sample?
Let

- $d_1 = d(a, b)$ the farthest distance
- d_2 the second farthest distance

Question:
Is d_1 a stable option?
what if we observe more examples?

Hoeffding bound:

Split if $d_1 - d_2 > \epsilon$ with $\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n}}$
where R is the range of the random variable; δ is a user confidence level, and n is the number of observed data points.

Hoeffding bound

- Suppose we have made n independent observations of a random variable r whose range is R.
- The Hoeffding bound states that:
  - With probability $1 - \delta$
  - The true mean of r is in the range $\bar{r} \pm \epsilon$ where $\epsilon = \sqrt{\frac{R^2 \mathit{ln}(1/\delta)}{2n}}$
- Independent of the probability distribution generating the examples.

McDiarmid's Bound

- Hoeffding bound requires *independent* random variables
- Analyzing similar objects where the differences are *not independent*, use McDiarmid's Bound.

Rutkowski, L. et al. *Decision Trees for Mining Data Streams Based on the McDiarmid's Bound*, TKDE 2014

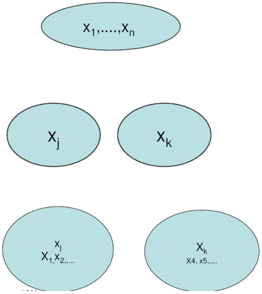
- $Pr(f(Z) - E[f(Z)] > \epsilon) \geq 1 - \delta$
  - Information Gain: $\epsilon = 6(log_2(eN) + log_2(2N))\sqrt{\frac{ln(1/\delta)}{2N}}$
  - Gini: $\epsilon = 8 \times \sqrt{\frac{ln(1/\delta)}{2N}}$

The Expand Operator: Expanding a Leaf

Step 1 Find Pivots:
$x_j, x_k : d(x_j, x_k) > d(a, b)$
$\forall a, b \neq j, k$

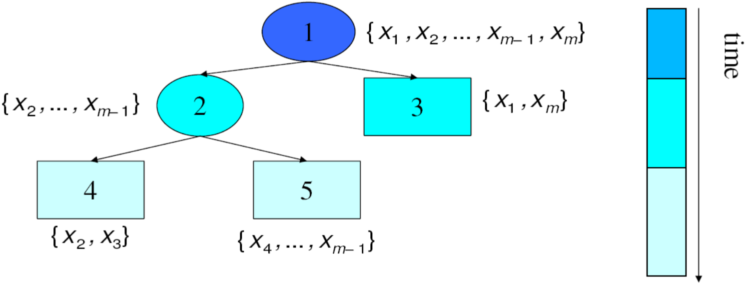
Step 2 If the Hoeffding bound applies:
Generate two new clusters.

Step 3 Each new cluster attract nearest variables.



Multi-Time-Windows

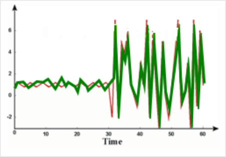
A multi-window system: each node (and leaves) receive examples from different time-windows.



The Merge Operator: Change Detection

Time Series Concept Drift:

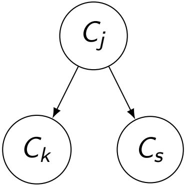
- Time evolving time-series
- Changes in the distribution generating the observations.
- Clustering Concept Drift
  - Changes in the way time series correlate with each other
  - Change in the cluster Structure.



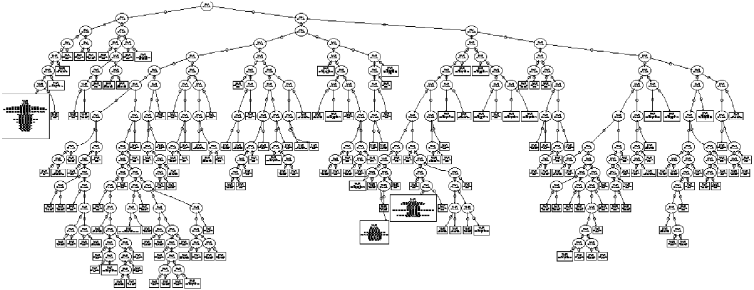
The Merge Operator: Change Detection

The Splitting Criteria guarantees that cluster's diameters monotonically decrease.

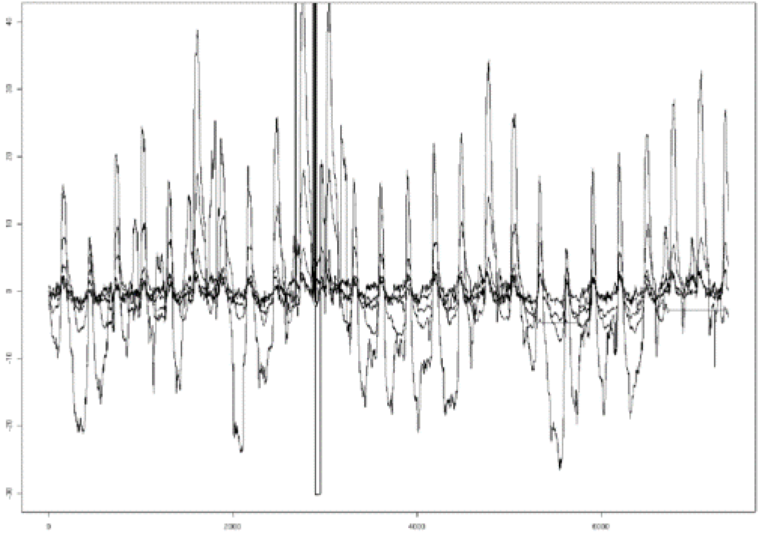
- Assume Clusters: c_j with descendants c_k and c_s.
- If $diameter(c_k) - diameter(c_j) > \epsilon$ OR $diameter(c_s) - diameter(c_j) > \epsilon$
  - Change in the correlation structure!
  - Merge clusters c_k and c_s into c_j.



The Electrical Load Demand Problem



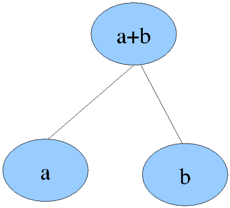
The Electrical Load Demand Problem



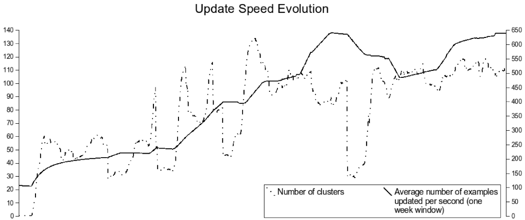
Properties of ODAC

- For stationary data the cluster's diameters monotonically decrease.
- **Constant update time/memory consumption** with respect to the number of examples!
- Every time a **split** is reported
  - the **time** to process the next example **decreases**, and
  - the **space** used by the new leaves is **less than** that used by the parent.

$$(a + b)^2 > a^2 + b^2$$



Evolution of Processing Speed



Hoeffding Algorithms

- Classification:
  Mining high-speed data streams, P. Domingos, G. Hulten, KDD, 2000
- Regression:
  Learning model trees from evolving data streams; Ikonomovska, Gama, Dzeroski; Data Min. Knowl. Discov. 2011
- Decision Rules:
  Learning Decision Rules from Data Streams, J. Gama, P. Kosina; IJCAI 2011
- Regression Rules
  E. Almeida, C. Ferreira, J. Gama: Adaptive Model Rules from Data Streams. ECML/PKDD 2013
- Clustering:
  Hierarchical Clustering of Time-Series Data Streams. Rodrigues, Gama, IEEE TKDE 20(5): 615-627 (2008)
- Multiple Models:
  Ensembles of Restricted Hoeffding Trees. Bifet, Frank, Holmes, Pfahringer; ACM TIST; 2012
  J. Duarte, J. Gama, Ensembles of Adaptive Model Rules from High-Speed Data Streams. BigMine 2014.
- ...

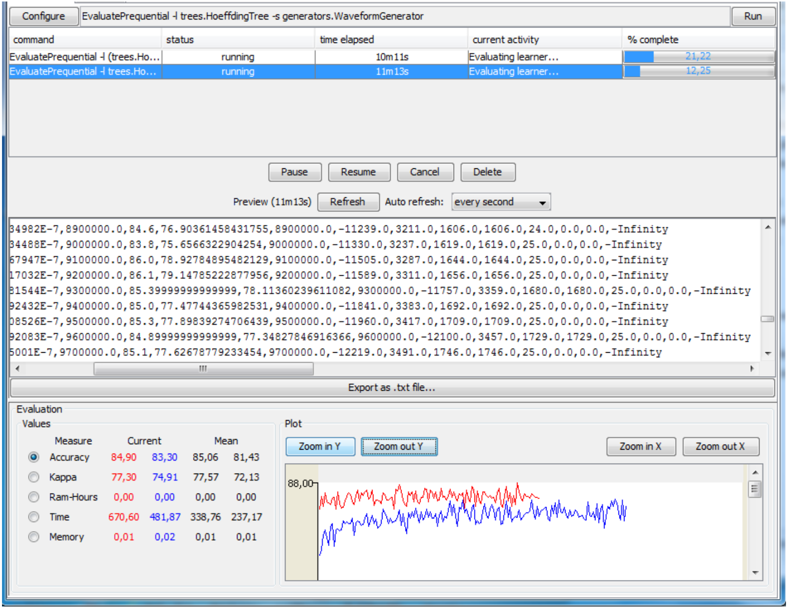
Hoeffding Algorithms: Analysis

The number of examples required to expand a node only depends on the Hoeffding bound.

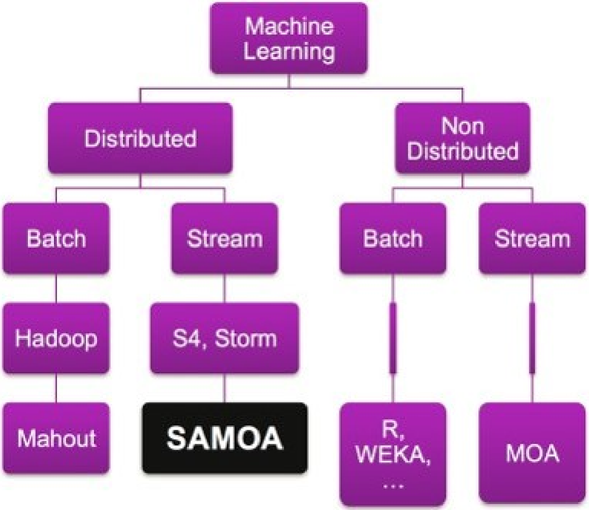
- Low variance models:
  Stable decisions with statistical support.
- Low overfiting:
  Examples are processed only once.
- No need for pruning;
  Decisions with statistical support;
- **Convergence**: Hoeffding Algorithms becomes asymptotically close to that of a batch learner. The expected disagreement is δ/p; where p is the probability that an example fall into a leaf.

Outline

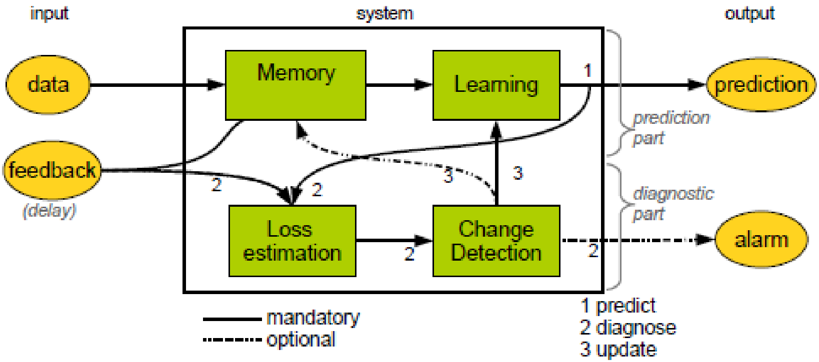
Massive Online Analysis



New Tools Emerge



A Generic Model for Adaptive Learning Algorithms



A generic schema for an online adaptive learning algorithm.

(*A survey on concept drift adaptation*, J.Gama et al, ACM-CSUR 2014)

Lessons Learned

Learning from data streams:

- Learning is not *one-shot*: is an evolving process;
- We need to monitor the learning process;
- Opens the possibility to reasoning about the learning

New Challenges

- What changed in the decision structure last week?
- Which patterns disappeared/ appeared last week?
- Which patterns are growing/shrinking this month?
- Mine the evolution of decision structures.

Reasoning about the Learning Process

Intelligent systems must:

- be able to adapt continuously to **changing environmental conditions** and evolving user habits and needs.
- be capable of **predictive self-diagnosis**.

The development of such self-configuring, self-optimizing, and self-repairing systems is a major scientific and engineering challenge.

Real-time learning: An existential pleasure!

Thank you!