

# Mining Official Data: what's new?

Gilbert Saporta

Conservatoire National des Arts et Métiers, Paris

[gilbert.saporta@cnam.fr](mailto:gilbert.saporta@cnam.fr)

<http://cedric.cnam.fr/~saporta>

**Data Mining Conference, Anaheim, 2012**

- This talk is about:
  - Data Mining in National Statistical Institutes
  - Not about mining official data in general
- An update of my 2001 paper presented at ISTAT meeting
  - Few applications of DM in NSIs

- Which kind of Data Mining?

Classical definitions stress upon exploratory (unsupervised) DM

- U.M.Fayyad, G.Piatetski-Shapiro (1996) :  
" Data Mining is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data "
- D.J.Hand (2000): " I shall define Data Mining as the discovery of interesting, unexpected, or valuable structures in large data sets"

- The metaphor of Data Mining means that there are **treasures** (or **nuggets**) hidden under mountains of data, which may be discovered by specific tools.
- Data Mining is concerned with data which were collected for another purpose: it is a **secondary analysis** of data bases that *are collected Not Primarily For Analysis*, but for the management of individual cases (Kardaun, T.Alanko,1998) .
- Data Mining is not concerned with efficient methods for collecting data such as **surveys** and experimental **designs** (Hand, 2000)

## Supervised DM:

- The purpose of Data Mining is to find structures in data. Two kinds of structures : **models** and **patterns**
- Predictive modelling:
  - Looking for models
  - But which kind of models?
    - Building models has always been a major activity for statisticians and econometricians:

- Statistical modelling aims at:
  - Providing some **understanding** of data and of its underlying mechanism through a **parsimonious** representation of a random phenomenon. A global summary of relationships between variables
  - **Predicting** new observations with a **high accuracy**.
  - Usually needs both a statistician and an expert of the application field.

## Algorithms or models?

- specific techniques of DM:
  - Decision trees, SVM, boosting, neural networks etc.
  - DM is not concerned with estimation and tests, of prespecified models, but with **discovering** models through an algorithmic search process exploring linear and non-linear models, explicit or not:  
**Models do not come from a theory, but from data exploration.**

- Paradox n°1

- A « good » statistical model should give insights in the nature of a stochastic phenomenon, not necessarily gives accurate predictions. In epidemiology eg, it is more important to find risk factors than having a prediction of getting some disease at an individual level.
- Different from physics where a good model must give good predictions, otherwise it is replaced by an other one.
- Is statistics a science or only technology?  
(C.R.Rao)



- Paradox n°2

## One may predict without understanding

- In Customer Relationship Management or pattern recognition, understanding is often a vain task: a banker does not need a theory for predicting if a loan will at risk or not, but only a good score function
- Here models are just algorithms, even black-boxes, and the quality of a model is assessed by its performance for predicting new observations.

Same formula:  $y = f(\mathbf{x}; \theta) + \varepsilon$

- **Classical framework**
  - Underlying theory
  - Narrow set of models
  - Focus on parameter estimation and goodness of fit
  - Error: white noise
- **Data mining context**
  - Models come from data
  - Algorithmic models
  - Focus on control of generalization error
  - Error: minimal
  - **An empirical conception of models**

- Today and tomorrow trends in DM

according to <http://www.kdnuggets.com>

## Industries / Fields where you applied Analytics / Data Mining in 2011

[f](#)
[in](#)
[+](#) 0
 [+](#) 5
 [Tweet](#) 8
 [comments](#)

### Industries / Fields where you applied Analytics / Data Mining in 2011?

[228 voters]

■ 2011 % of voters
 ■ 2010 % of voters

CRM/ consumer analytics (57)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 25.0%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 26.8%</div> </div>
Banking (43)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 18.9%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 19.2%</div> </div>
Health care/ HR (38)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 16.7%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 13.1%</div> </div>
Education (37)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 16.2%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 9.9%</div> </div>
Fraud Detection (32)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 14.0%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 12.7%</div> </div>
Science (31)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 13.6%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 10.3%</div> </div>
Social Networks (30)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 13.2%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 6.6%</div> </div>
Credit Scoring (29)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 12.7%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 8.0%</div> </div>
Direct Marketing/ Fundraising (28)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 12.3%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 11.3%</div> </div>
Insurance (28)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 12.3%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 10.3%</div> </div>
Finance (26)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 11.4%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 11.3%</div> </div>
Telecom / Cable (25)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 11.0%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 10.8%</div> </div>
Retail (24)	<div style="display: flex; justify-content: space-between;"> <div style="width: 40%;"><span style="color: orange;">■</span> 10.5%</div> <div style="width: 40%;"><span style="color: purple;">■</span> 8.0%</div> </div>

## Algorithms for data analysis / data mining



14














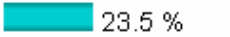
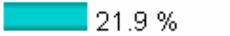
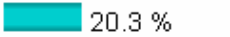
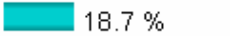
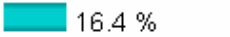
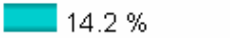
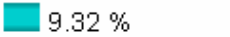
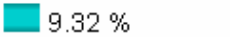
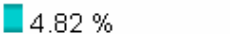
Tweet



14

[comments](#)

Which methods/algorithms did you use for data analysis in 2011? [311 voters]

Decision Trees/Rules (186)	 59.8 %
Regression (180)	 57.9 %
Clustering (163)	 52.4 %
Statistics (descriptive) (149)	 47.9 %
Visualization (119)	 38.3 %
Time series/Sequence analysis (92)	 29.6 %
Support Vector (SVM) (89)	 28.6 %
Association rules (89)	 28.6 %
Ensemble methods (88)	 28.3 %
Text Mining (86)	 27.7 %
Neural Nets (84)	 27.0 %
Boosting (73)	 23.5 %
Bayesian (68)	 21.9 %
Bagging (63)	 20.3 %
Factor Analysis (58)	 18.7 %
Anomaly/Deviation detection (51)	 16.4 %
Social Network Analysis (44)	 14.2 %
Survival Analysis (29)	 9.32 %
Genetic algorithms (29)	 9.32 %
Uplift modeling (15)	 4.82 %

## Hottest Analytics / Data Mining Topics in 2012



Tweet



### What will be the hottest analytics / data mining topics in 2012?

[366 votes total]

Big Data (183)		50.0%
Analytics in the Cloud and Hadoop (155)		42.3%
Social analytics (146)		39.9%
Text analytics (125)		34.2%
Location-aware analytics (86)		23.5%
Sensor data (61)		16.7%
Competition platforms (39)		10.7%
Game analytics (38)		10.4%
Privacy (36)		9.8%
Other (22)		6.0%

- **National Statistical Institutes and DM**
  - NSIs collect and produce mines of data for population, trade, agriculture, business...
  - But few known applications of data mining techniques for discovering new models or patterns
  - Looking for keywords “data mining”, “decision trees”, “neural networks” etc. in NSIs websites is very disappointing

– DM never or rarely occurs

- not a single paper since 1991 at the Methodology Symposium of INSEE (French NSI)
- Less than 5 working papers at the US Census Bureau
  - Aaron Gilary. (2011). Recursive Partitioning for Racial Classification Cells.
- StatCan, UK ONS: ?
- INE (Portugal) : some studies using Symbolic Analysis (see Billard & Diday, 2006)
- If yes on very specific topics: missing values, survey quality, mostly by computer science people



- Modern regression techniques are ignored even if multicollinearity is often addressed in econometrics
  - $n$  should always be larger than  $p$
  - Econometrica
    - No entry for « lasso »
    - « neural network » :3 papers in the 90's , none after, one by Nobel Prize C.W.J.Granger...
- Needless to speak of exotic methods like boosting, SVM, association rules !

## Eurostat's pioneering efforts

- 1998-2002 : European Plan of Research in Official Statistics (EPROS)
  - ASSO – Analysis System of Symbolic Official data
  - SPIN! – Spatial Mining for Data of Public Interest
  - VITAMIN S - system for statistical visualization with a data mining perspective.
- 1996 : KESO (Knowledge Extraction for Statistical Offices)
- NTTS conferences 1992, 1995, 1998, 2001, 2009, 2011, 2013

- Why such a reluctance?

My personal view:

- Main task of most NSIs is data production. Analysis are often done by other institutes
- Exploring a database with the objective of finding unexpected patterns or models is unfamiliar to official statisticians who have to answer precise questions and make forecasts
- NSIs are often ruled by economists who believe in their science, and DM is not « science » for them. Researchers dislike automatic process

**A cultural change is necessary!**

## A new way of doing inference

- Classical inference no longer holds for huge data-sets: any null hypothesis  $H_0$  is rejected when  $n$  is large
  - a correlation 0.002 is significantly different from zero with one million units.
- Significance tests should be replaced by crossvalidation, subsampling, resampling...
- Embedded in Statistical Learning Theory, DM is empirical inference science (Vapnik, 2006)

## Conclusions

- Data Mining is still a minor or inexistant activity in National Statistical Institutes due to a cultural gap and a conservative attitude.
- Mines of unexploited data by NSIs, but exploited elsewhere
- Is there a hope to change? pressure of users



Login Register



HOME

COMMUNITIES

SEMANTIC ASSETS

SOFTWARE

NEWS



## Blogs and news

Create, read and comment on news and blogs about interoperability solutions for public administrations.

### News

Browse all

### News



### Blogs

Recommended

Editor's choice

### Newsletter

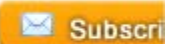
Archives

Submitted by [Gijs HILLENIOUS](#) on December 23, 2010  
Rating: 0/5 (based on 1 votes) | 215 reads

## EU External Affairs publishes crisis data mining tools as open source

The crisis room at the European Union's Directorate-General for External Relations has published several open source data mining tools, meant to provide real-time support for early warning and crisis response.

The main tool, Tarîqa, version 3.0, is an open source search platform that uses and combines information from search engines, information databases and geographic information sources including satellite images. "Tarîqa's advanced information retrieval tools makes it possible to gain useful knowledge from the masses of information that are available", the developers write in their introduction.



21 August 2010 |  
[EU: Guide on revised](#)  
13 December 2010

THANK YOU