le cnam

# Which analytic methods for Big Data?

**Gilbert Saporta**
CEDRIC- CNAM,
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

# Outline

1. Too big ?
2. Which kind of models?
3. How to validate a model
4. Model choice and the search for sparsity
5. Sparse MCA
6. The end of theory?
7. Conclusion

# 1. Too big ?

- Estimation and tests become useless
- Everything is significant!
  - with $n=10^6$ a correlation coefficient = 0,002 is significantly different from 0 but without any interest
  - Usual distributional models are rejected since small discrepancies between model and data are significant
  - Confidence intervals have zero length

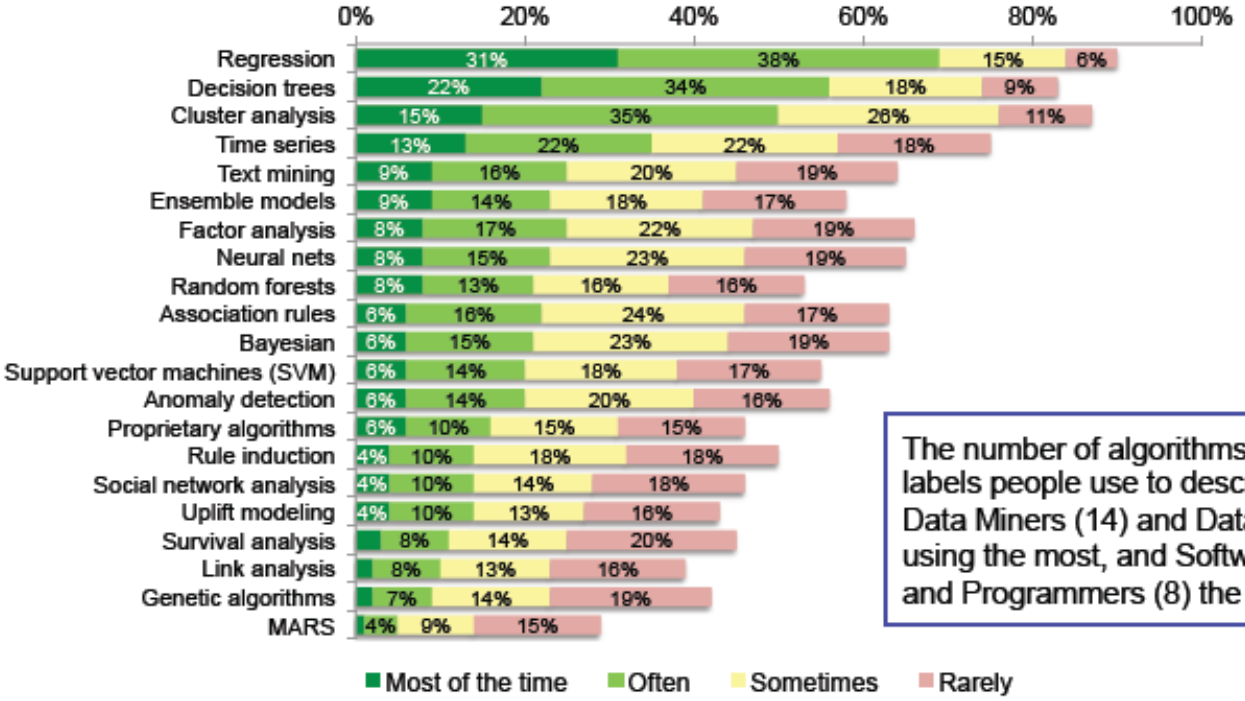
# 2. Which kind of models?

- Data Scientists and Data Miners use models in a data driven way
  - Models come from data, not from a theory
  - Quite different from classical modelling

- Toolbox: a mix of statistical and machine learning procedures

# Algorithms

- Regression, decision trees, and cluster analysis continue to form a triad of core algorithms for most data miners. This has been consistent since the first Data Miner Survey in 2007.
- The average respondent reports typically using 12 algorithms. People with more years of experience use more algorithms, and consultants use more algorithms (13) than people working in other settings (11).




The number of algorithms used varies by the labels people use to describe themselves, with Data Miners (14) and Data Scientists (14) using the most, and Software Developers (9) and Programmers (8) the fewest.

Question: What algorithms / analytic methods do you TYPICALLY use? (Select all that apply)

- Standard conception (models for understanding)
  - Provide some comprehension of data and their generative mechanism through a parsimonious representation.
  - A model should be simple and its parameters interpretable  for the specialist : elasticity, odds-ratio, etc.
- Paradoxes
  - a model with a good fit may provide poor predictions at an individual level
  - Good predictions may be obtained with uninterpretable models

- In « Big Data Analytics » one focus on prediction
  - For new observations «generalization »
  - Differs from having a good fit in the learning step (predicting the past)
    - risk of overfitting
- Models are merely algorithms

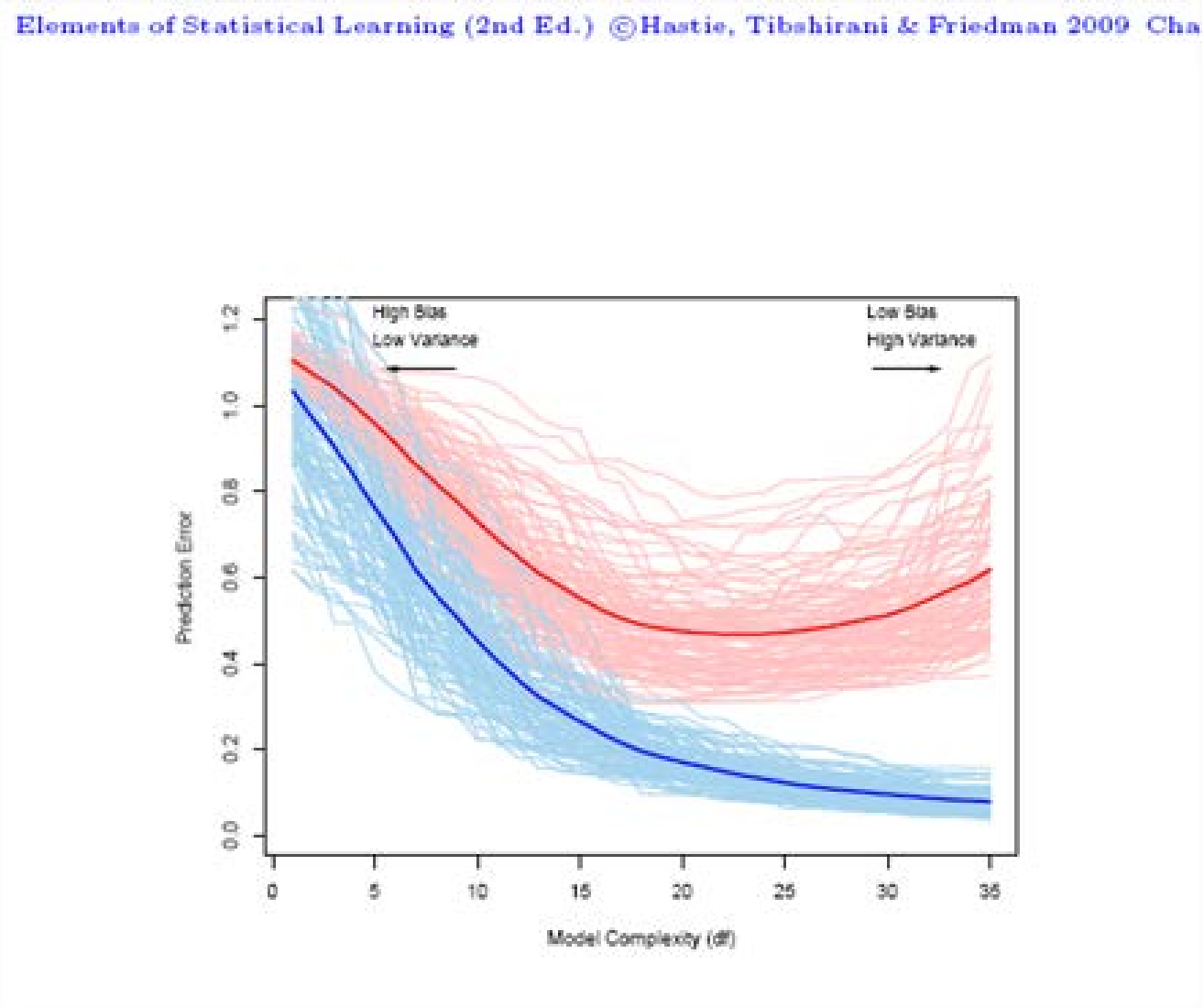
# 3. How to validate a model?

- Combining Machine Learning and Statistics
  - A good model must give good predictions
  - Goodness of fit ≠ prediction
    - Predicting the past or the future?
  - Bootstrap, cross-validation
  - Learning and validation sets

# The three samples procedure for selecting a family of models

- Learning set: estimating model parameters
- Test set : choice of the best model in terms of prediction
  - Reestimation of the final model: **with all available observations**
- Validation set : estimate the performance for future data. « Generalization »
  - Parameter estimation ≠ performance estimation

- One split is not enough!



- **Elementary?**
  - Not that sure…
  - Have a look on  publications in econometrics, epidemiology, ..

# 4. Model choice and the search for sparsity (or parsimony)

- William of Ockham
  (c. 1287 – 1347)

- Vladimir Vapnik (1990)

- Ockham's razor
  - a scientific principle for avoiding useless hypothesis

    *pluralitas non est ponenda sine necessitate*

- Ockham's razor
  - a scientific principle for avoiding useless hypothesis

    *pluralitas non est ponenda sine necessitate*

- AIC, BIC and other penalized likelihood techniques are often considered as modern versions of Ockham's razor

  $$AIC = -2\ln(L) + 2K$$

  $$BIC = -2\ln(L) + K\ln(n)$$

- A misleading similarity
- **AIC and BIC come from quite different theories**
  - AIC : approximation of the Kullback-Leibler divergence between the true distribution and the best choice inside a family
  - BIC : bayesian choice among parametric models with equal priors

- **Use AIC and BIC simultaneously is illogical**

# AIC BIC realistic?

- Likelihood not always computable: need distributional assumptions (trees, neural networks..).
- How to define the number of parameters? (trees, but also ridge, PLS..)
- Is there a « true » model?

  “Essentially, all models are wrong, but some are useful ” (G.Box,1987)

* Box, G.E.P. and Draper, N.R.: Empirical Model-Building and Response Surfaces, p. 424, Wiley, 1987

**The VC inequality** between learning risk and generalization risk

In supervised classification:

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

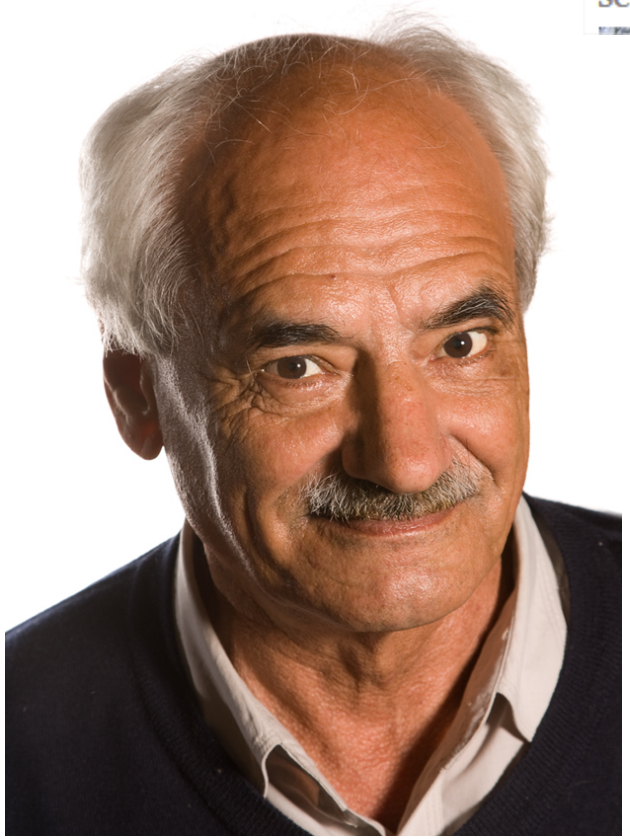
holds with probability1- $\alpha$

h : VC dimension , a measure of model complexity

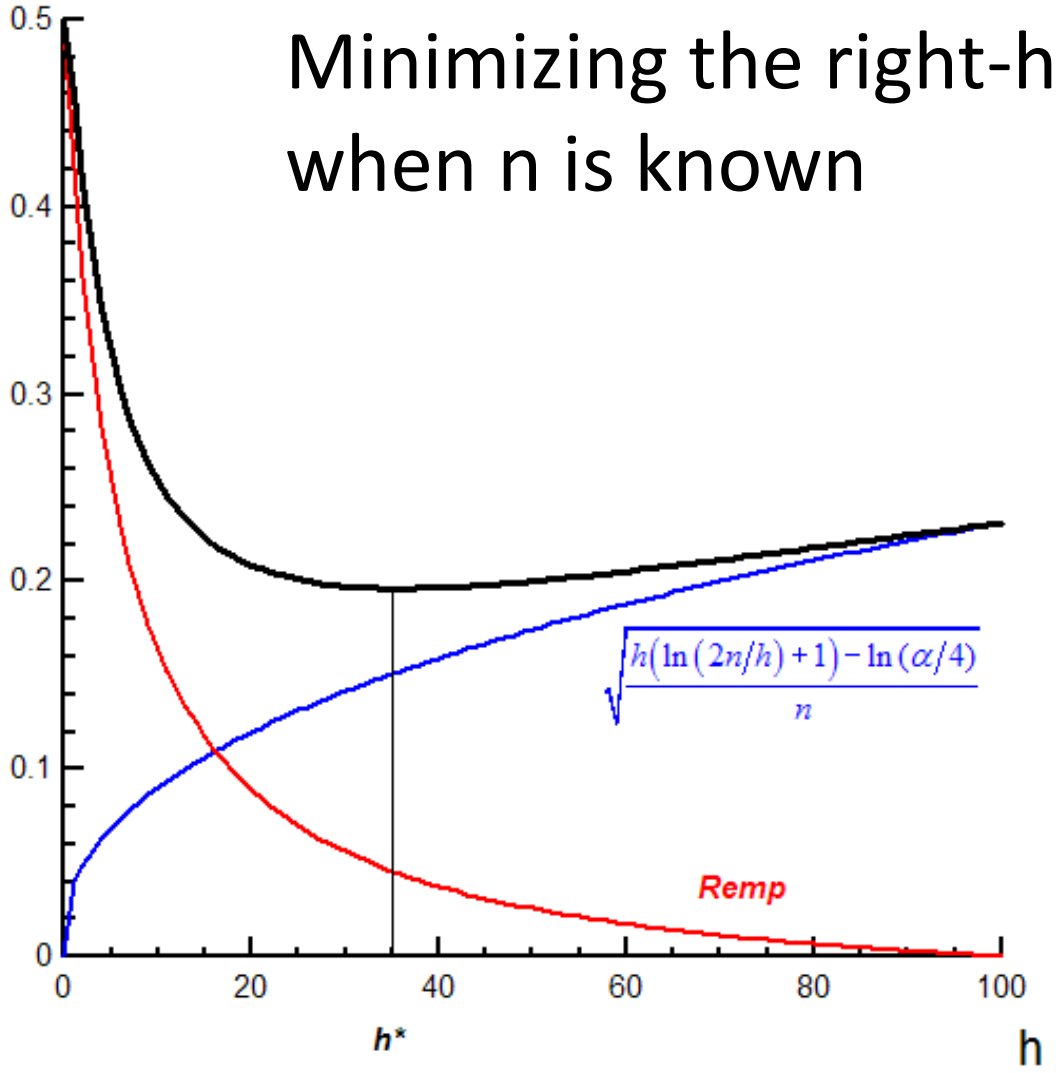
September 22, 2014

# University of London maths professor found dead in Moscow park

Alexey Chervonenkis died of hypothermia after losing his way, according to search party who found body

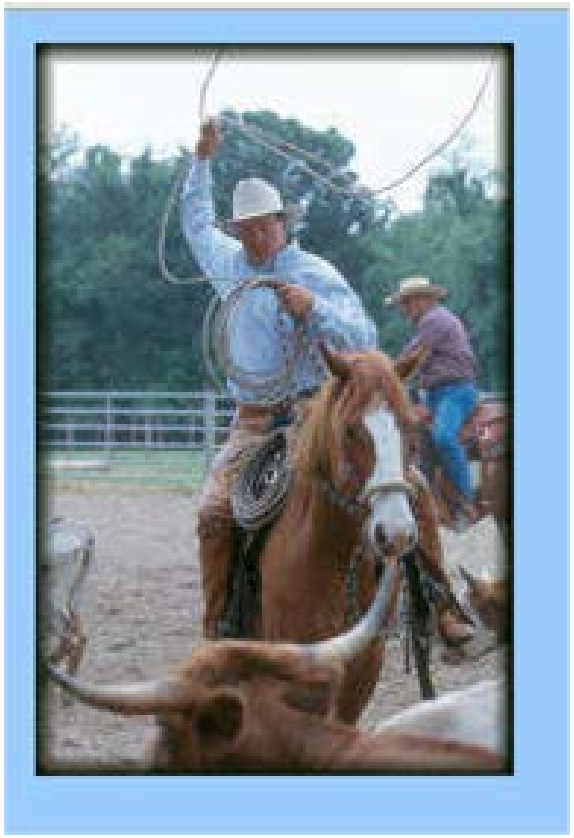


# Minimizing the right-hand side when n is known



- The upper bound depends from n/h, hence surprising results:
  - If h increases slower than n, it improves the generalization.
  - One may use more and more complex models when n is big!
- Not necessarily a good idea mainly if Data are Big according to p

- A particular kind of regularization may solve the problem of high dimensional data



$$\|y - Xb\|^2 \quad \text{with} \sum_{j=1}^{p} |b_j| < c$$

$$\hat{\beta} = \arg\min \|y - Xb\|^2 + \lambda \sum_{j=1}^{p} |b_j|$$

When $\lambda$ increases
some coefficients vanish

http://www-stat.stanford.edu/~tibs/lasso.html

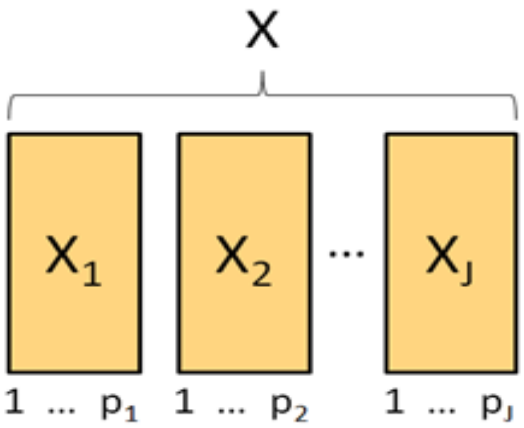
- Elastic net

Combines $L^2$ (ridge) and $L^1$ (lasso) regularization

$$\min\left(\left\|y - Xb\right\|^2 + \lambda_2 \left\|b\right\|^2 + \lambda_1 \left\|b\right\|_1\right)$$

# 5.Sparse MCA

Disjunctive table




**Selection of a categorical variable: selection of a block of indicators**

**technique:** use the group Lasso penalty

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \left\| \mathbf{Z} - \mathbf{X}\boldsymbol{\beta} \right\|^2 + \lambda \sum_{j=1}^{J} \sqrt{p_j} \left\| \boldsymbol{\beta}_j \right\|$$

$$\mathbf{X}\boldsymbol{\beta} = \sum_{j=1}^{J} \mathbf{X}_j \boldsymbol{\beta}_j$$

Bernard, A. , Guinot, C. , Saporta, G. *Compstat 2012*, 99-106

# Application on genetic data

## Single Nucleotide Polymorphisms

| SNP1=$X_1$ | … | SNP537=$X_{537}$ |
|---|---|---|
| AA | | AB |
| AB | | BB |
| . | . . . | . |
| . | | . |
| . | | . |
| AA | | AA |
| BB | | AA |

| SNP1=$D_{[1]}$ | | | … | SNP537=$D_{[537]}$ | | |
|---|---|---|---|---|---|---|
| AA | AB | BB | | AA | AB | BB |
| 1 | 0 | 0 | | 0 | 1 | 0 |
| 0 | 1 | 0 | | 0 | 0 | 1 |
| . | . | . | | . | . | . |
| . | . | . | . . . | . | . | . |
| . | . | . | | . | . | . |
| 1 | 0 | 0 | | 1 | 0 | 0 |
| 0 | 0 | 1 | | 1 | 0 | 0 |

**Data:**

n=502 individuals

p=537 SNPs (among more than 800 000 of the original data base, 15000 genes)

q=1554 (total number of columns)

X : 502 x 537matrix of qualitative variables

K : 502 x 1554 complete disjunctive table → $K=(K_1, \ldots, K_{1554})$

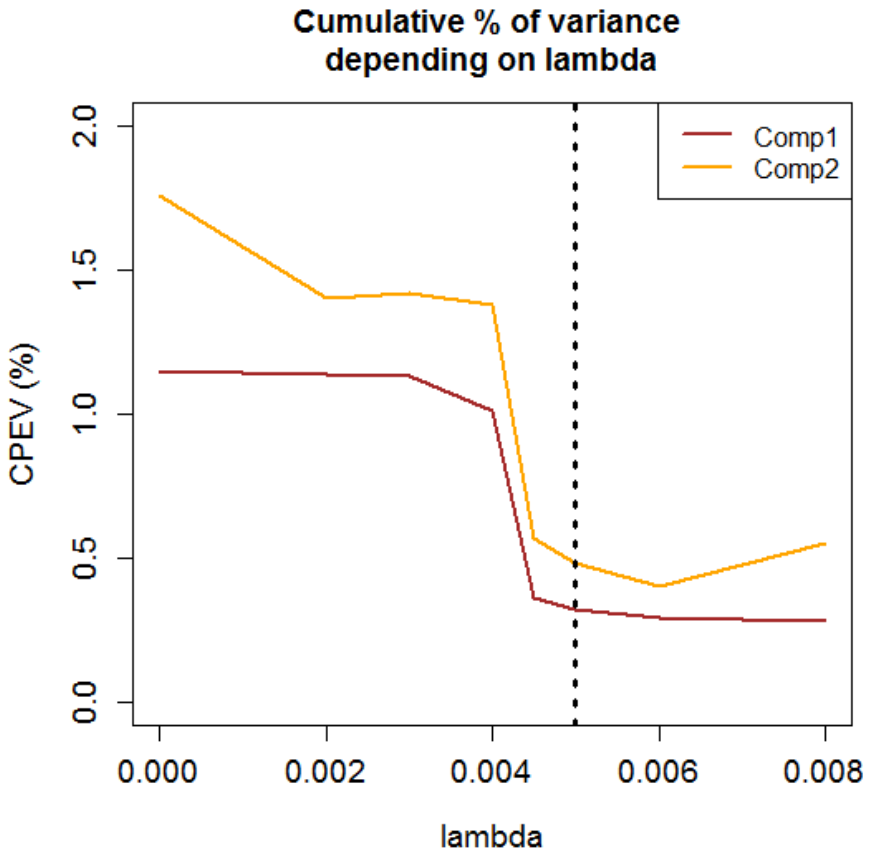
**1 block**

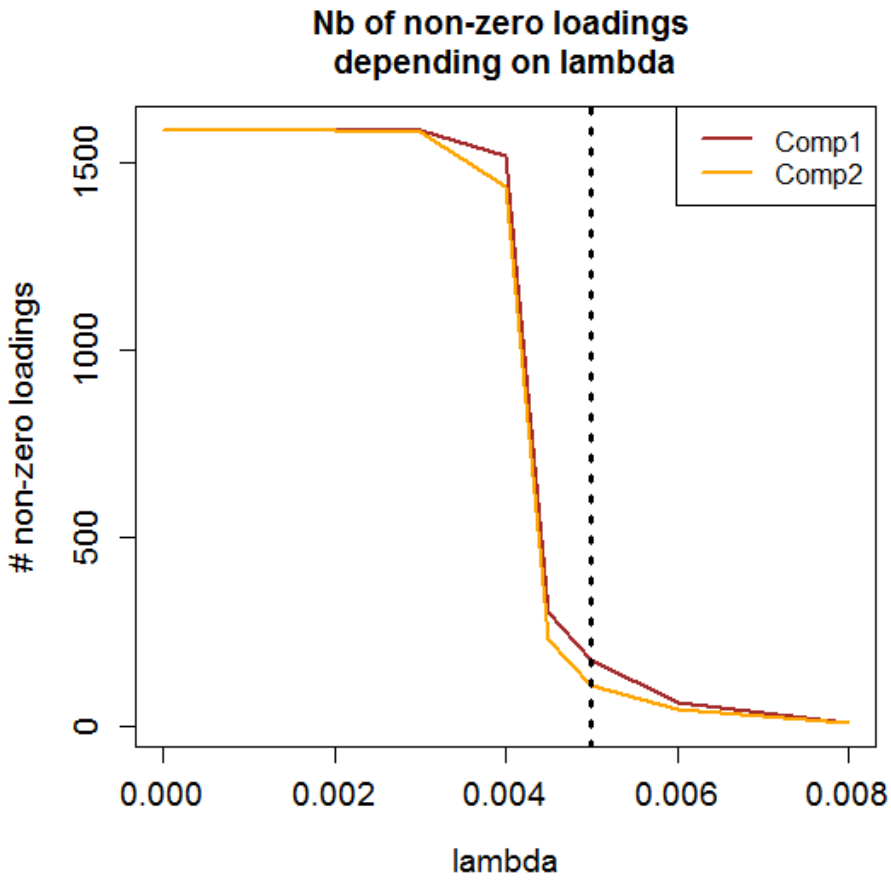
**=**

**1 SNP = 1 $K_j$ matrix**

# Application on genetic data

## Single Nucleotide Polymorphisms




$\lambda$= 0:005: CPEV= 0:32% and 174 columns selected on Comp 1

# Application on genetic data

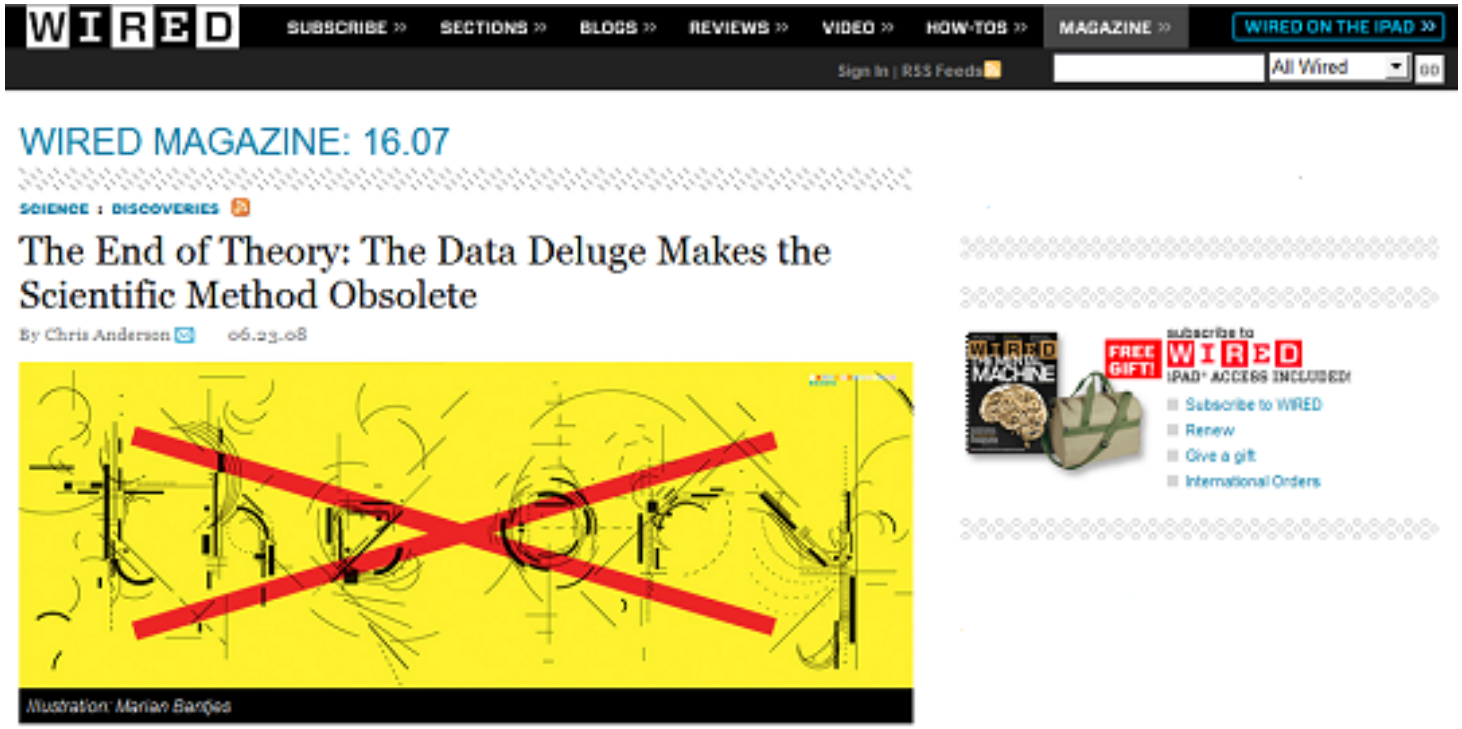
## Comparison of the loadings

| SNPs | MCA | | SMCA | |
|---|---|---|---|---|
| | Comp1 | Comp2 | Comp1 | Comp2 |
| SNP1.AA | -0.078 | 0.040 | -0.092 | 0.102 |
| SNP1.AG | -0.014 | -0.027 | -0.022 | -0.053 |
| SNP1.GG | 0.150 | -0.002 | 0.132 | -0.003 |
| SNP2.AA | -0.082 | 0.041 | -0.118 | 0.000 |
| SNP2.AG | -0.021 | -0.025 | -0.020 | 0.000 |
| SNP2.GG | -0.081 | 0.040 | -0.001 | 0.000 |
| SNP3.CC | -0.004 | 0.050 | 0.000 | 0.000 |
| SNP3.CG | 0.016 | 0.021 | 0.000 | 0.000 |
| SNP3.GG | -0.037 | -0.325 | 0.000 | 0.000 |
| SNP4.AA | 0.149 | -0.003 | 0.050 | 0.000 |
| SNP4.AG | -0.016 | -0.025 | -0.002 | 0.000 |
| SNP4.GG | -0.081 | 0.040 | -0.100 | 0.000 |
| ... | ... | ... | ... | ... |
| Nb non-zero loadings | 1554 | 1554 | 172 | 108 |
| Variance (%) | 1.14 | 0.63 | 0.32 | 0.16 |
| Cumulative variance (%) | 1.14 | 1.77 | 0.32 | 0.48 |

| Properties | MCA | Sparse MCA |
|---|---|---|
| Uncorrelated Components | TRUE | FALSE |
| Orthogonal loadings | TRUE | FALSE |
| Barycentric property | TRUE | PARTLY TRUE |
| % of inertia | $\lambda_j / tot \times 100$ | $\left\| \tilde{\mathbf{z}}_{\mathbf{j.1,\ldots,j\text{-}1}} \right\|^2$ |
| Total inertia | $\frac{1}{p}\sum_{j=1}^{p} p_j - 1$ | $\sum_{j=1}^{k} \left\| \tilde{\mathbf{z}}_{\mathbf{j.1,\ldots,j\text{-}1}} \right\|^2$ |

$\tilde{\mathbf{z}}_{\mathbf{j.1,\ldots,j\text{-}1}}$ are the residuals after adjusting $\tilde{\mathbf{z}}_{\mathbf{j}}$ for $\tilde{\mathbf{z}}_{\mathbf{1,\ldots,j\text{-}1}}$(regression projection)

# 6. The end of theory?

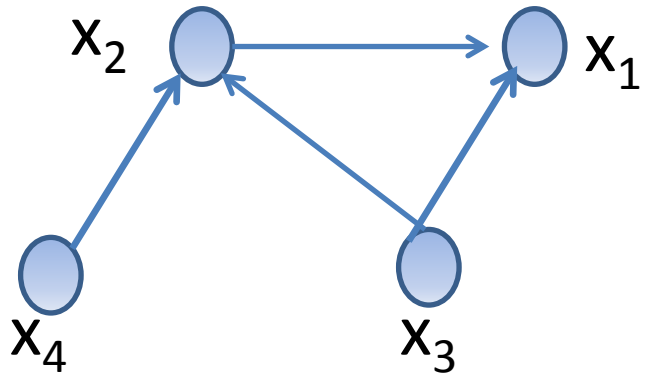



Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

- Correlation is not causality
- A regression coefficient does not measure the influence of a predictor (P.Bühlmann)
  - « holding all other variables fixed » is nonsense
  - When a predictor changes , it implies that other do (intervention vs correlation)
  - Causal schemes are necessary

- complementing a regression scheme (linear or not) with a causal diagram

$$\hat{y} = f(\mathbf{x})$$




DAG: Directed Acyclic Graph

# Conclusions

- Massive data need specific approaches
  - Models are algorithms
  - Validation
- Combine complexity and sparsity
- Good old methods (SVD, k-means) are still efficient especially in unsupervised contexts

# Thanks for your attention