le cnam

Predictive versus Generative Modelling: a Challenge for (Social) Sciences

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

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

Outline

1. The two cultures
2. Understand or predict?
3. Parcimony and complexity
4. Empirical validation
5. Interpreting models
6. Data science and the Data revolution
7. Conclusion: understanding to better predict

1. The two cultures

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

- The generative modelling culture
  - seeks to develop stochastic models which fits the data, and then make inferences about the data-generating mechanism based on the structure of those models. Implicit (…) is the notion that there is a true model generating the data, and often a truly `best' way to analyze the data.
- The predictive modelling culture
  - is silent about the underlying mechanism generating the data, and allows for many different predictive algorithms, preferring to discuss only accuracy of prediction made by different algorithm on various datasets. Machine Learning is identified by Breiman as the epicenter of the Predictive Modeling culture.

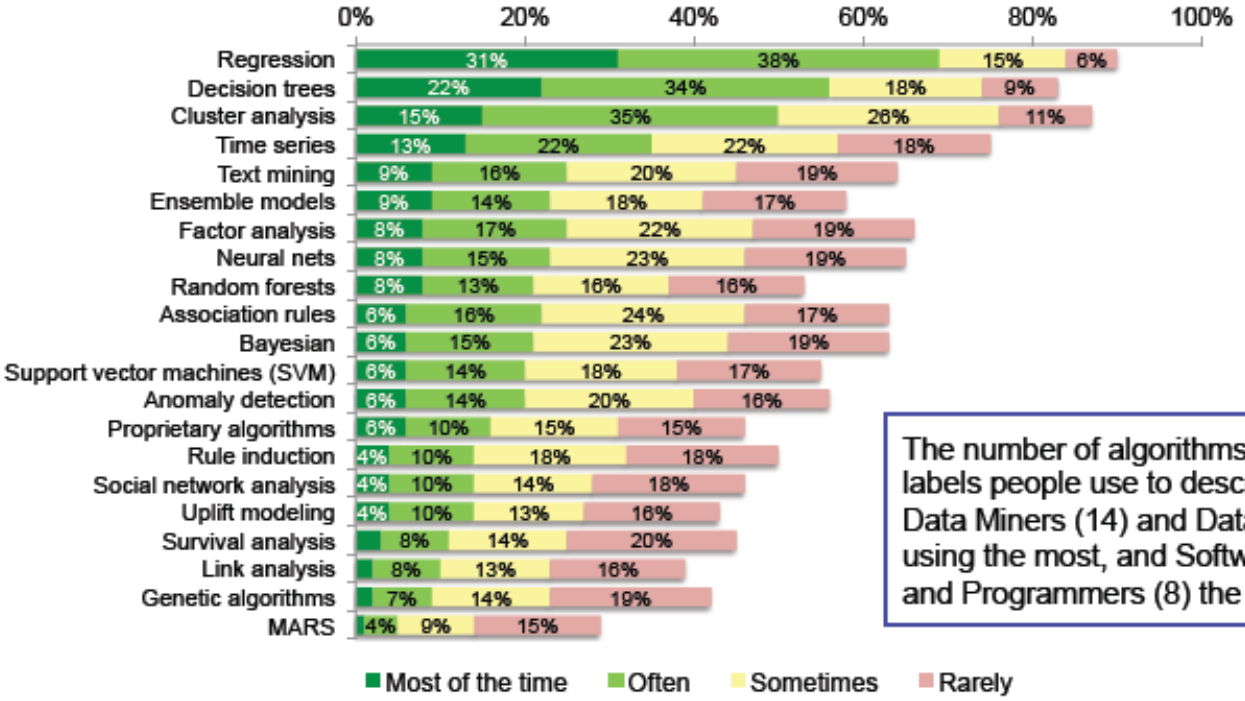
From Donoho, 2015

- 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.
- In « Big Data Analytics » one focus on prediction
  - For new observations: generalization
  - Models are merely algorithms

Cf GS, compstat 2008

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)

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

- **Generative modelling**
  - Underlying theory
  - Narrow set of models
  - Focus on parameter estimation and goodness of fit: predict the past
  - Error: white noise
- **Predictive modelling**
  - Models come from data
  - Algorithmic models
  - Focus on control of generalization error : predict the future
  - Error: minimal

2. Predict without understanding?

- Paradoxes
  - a model with a good fit may provide poor predictions at an individual level (eg epidemiology)
  - Good predictions may be obtained with uninterpretable models (targetting customers or approving loans, do not need a consumer theory)

According to Bottou, 2013:

- Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data (Breiman, 2001).
- Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms (Vapnik, 2006).

1.2.2 THE BLACK BOX MODEL

One can describe the pattern recognition problem as follows. There exists a black box BB that when given an input vector x_i returns an output y_i which can take only two values $y_i \in \{-1, +1\}$. The problem is: given the pairs $(y_i, x_i), i = 1, \ldots, \ell$ (the training data) find a function that approximates the rule that the black box uses.

Two different concepts of what is meant by a *good approximation* are possible:

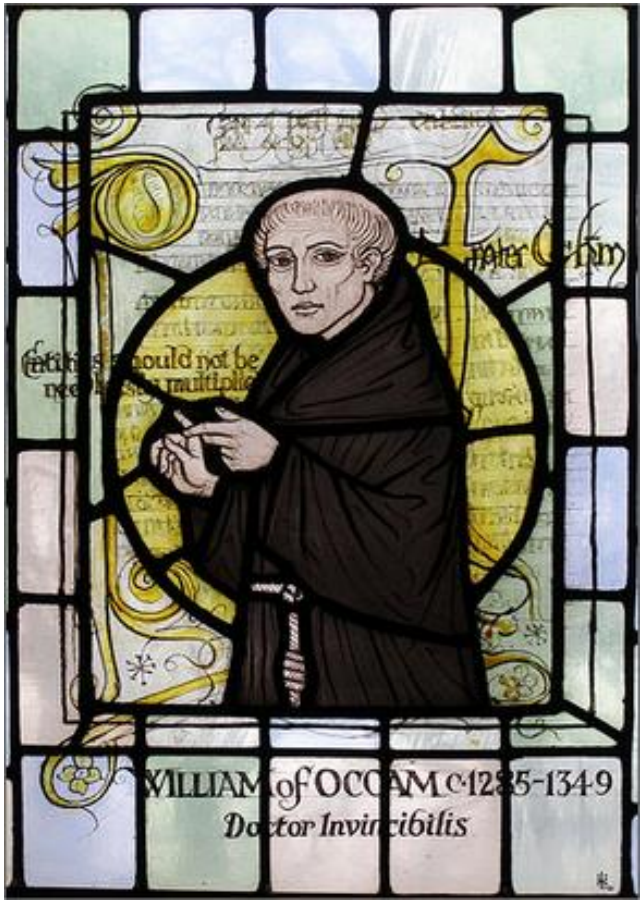
(1) A good approximation of the BB rule is a function that is close (in a metric of functional space) to the function that the BB uses. (In the classical setting often we assume that the BB uses the Bayesian rule.)

(2) A good approximation of the BB rule is a function that provides approximately the same error rate as the one that the BB uses (provides the rule that predicts the outcomes of the BB well).

In other words, in the first case one uses a concept of closeness in the sense of being close to the *true function* used by the BB (closeness in a metric space of functions), while in the second case one uses a concept of closeness in the sense of being close to the accuracy of prediction (closeness in *functionals*). These definitions are very different.

- Despite their simplicity, decision trees are not in favour of many scientists since they are not considered as generative.
- But what about linear or logistic models? Simple analytic formulas are easy to use but there are no guarantee that they represent the true mechanism
- Remind that in most sciences a good model must give good predictions, otherwise it is replaced by an other one.

3. Parsimony and complexity



- Ockham's razor *
  - *pluralitas non est ponenda sine necessitate*
  - a scientific principle for avoiding useless hypothesis

* Or Occam

- AIC, BIC and other penalized likelihood techniques 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
- **No rationale to use simultaneously AIC and BIC**

- AIC is biased : if the true model M_i belongs to the family, the probability that AIC chooses M_i does not tend to 1 when the number of observations goes to infinity. But BIC converges.

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)

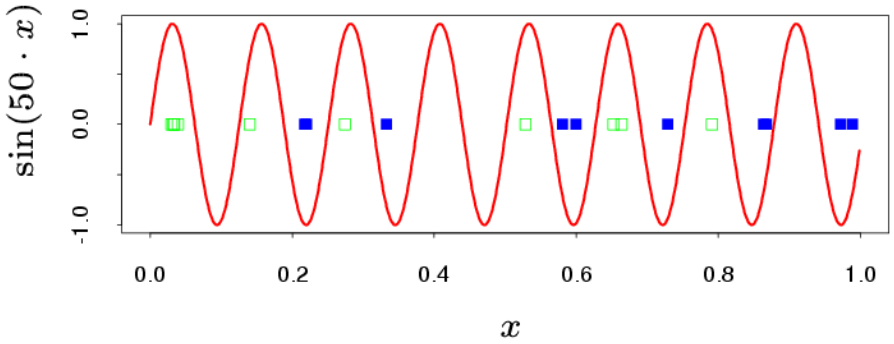
* Box, G.E.P. and Draper, N.R.: Empirical Model-Building and Response Surfaces, p. 424, Wiley, 1987

- « Occam’s Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately in prediction, accuracy and simplicity (interpretability) are in conflict » Breiman, 2011

- Vapnik's statistical learning theory

1990

h : VC dimension , a measure of model complexity, different from the number of parameters



©Hastie et al., 2009

f(x,w) = sign (sin (w.x)) one parameter but h=∞

The VC inequality between learning risk and generalization risk

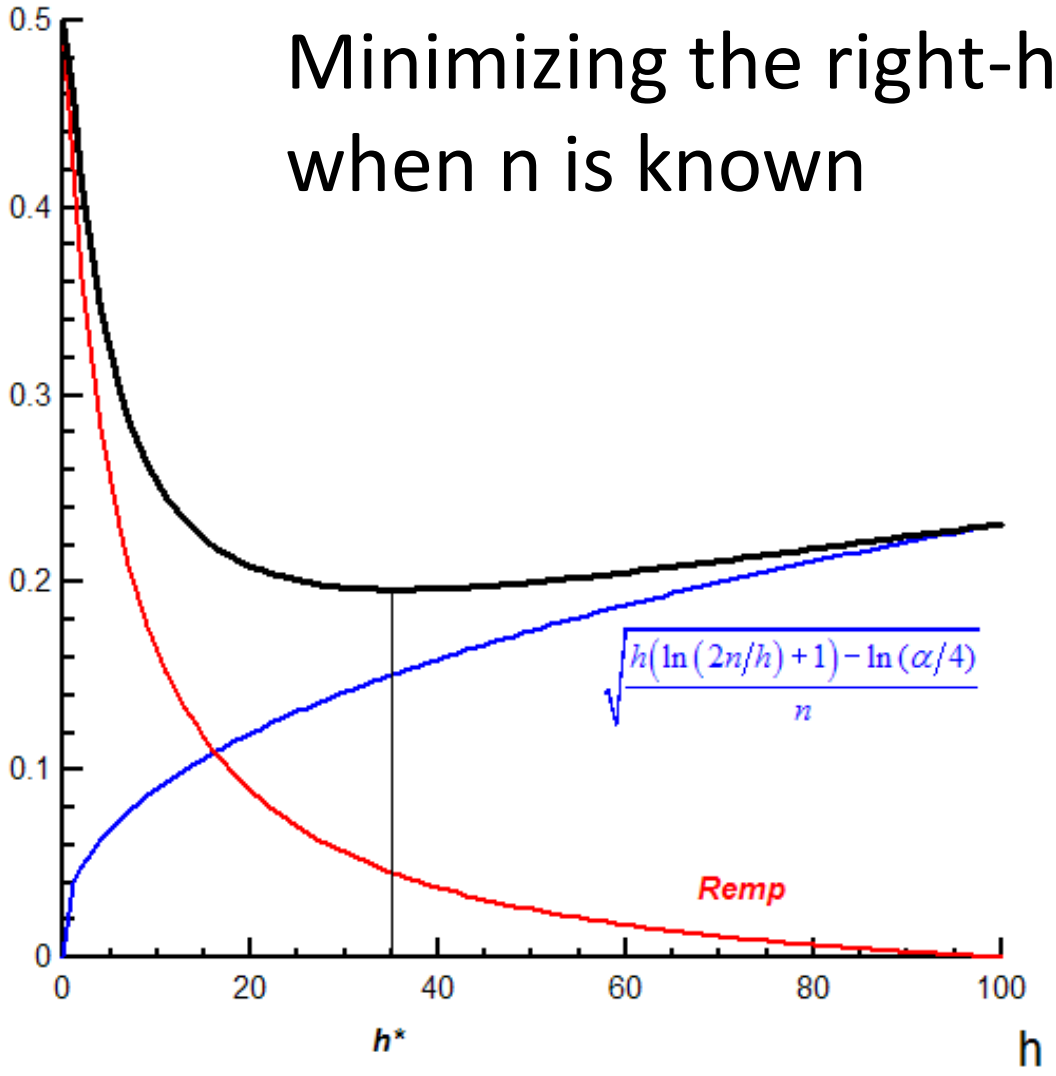
In supervised classification:

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

holds with probability1- α
h should be finite

Used to choose among models with different h

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 also big according to p
  - Solution: sparsity constraints (Lasso)

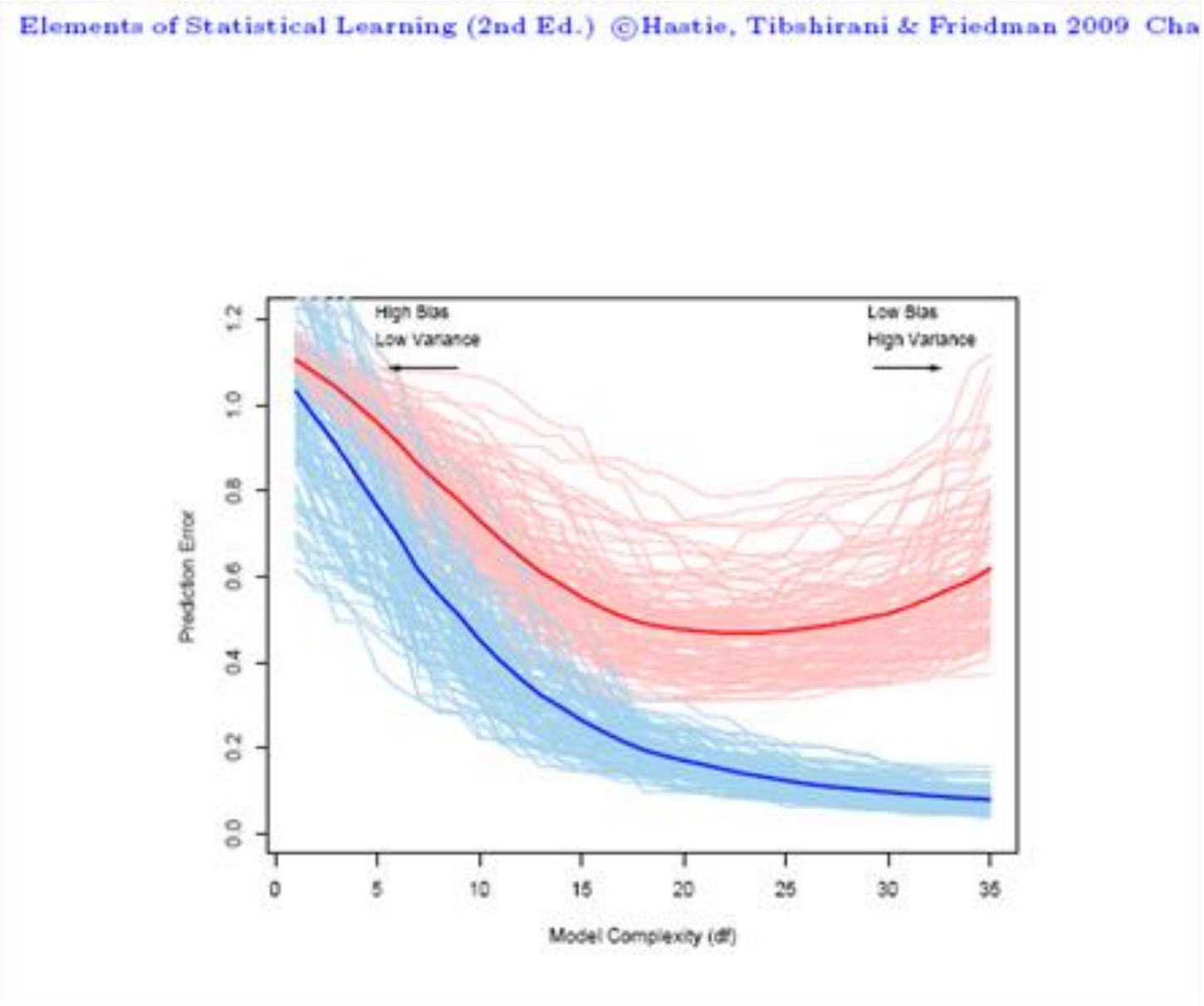
4. Empirical validation

- Combining Machine Learning and Statistics
  - A good model must give good predictions
  - Bootstrap, cross-validation, etc.
  - Learning and validation sets

The three samples procedure for selecting a a model inside a family of models

- Learning set: estimate parameters for all models in competition
- Test set : choice of the best model in terms of prediction
  - NB 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, .. prediction is rarely checked on a hold-out sample (except in time series forecasting)

5. Interpreting models

- A common belief is that simple models, like linear or logistic regression are easily interpretable
- Generally untrue!
- Except in case of orthogonal designs, parameter values hardly reflect variable importance

- More than 11 methods of quantifying variable importance in linear models! (Grömping, 2015, Wallard, 2015) including Fabbris, 1980
  - Eg Shapley value: a subset of predictors is a coalition
- Simple models are not that simple! Why not using complex ones?
  - Random forests outperforms almost all predictive algorithms
  - Deep learning, a variant of neural networks, is appropriate for Big Data. (LeCun & al, 2015)

Symposium: Big Data

Big Data: New Tricks for Econometrics (pp. 3-28)

Hal R. Varian

Abstract/Tools | Full-text Article (Complimentary) | Download Data Set (2.62

AMERICAN ECONOMIC ASSOCIATION

AEA American Economic Association

AEA | Journals | Annual Meeting | EconLi

Journal of Economic Perspectives: Vol. 28 No. 2 (Spring 2014)

H.Varian writes:

- "When confronted with a prediction problem of this sort an economist would think immediately of a linear or logistic regression. However, there may be better choices, particularly if a lot of data is available. These include nonlinear methods such as 1) classification and regression trees (CART); 2) random forests; and 3) penalized regression such as LASSO, LARS, and elastic nets. (There are also other techniques, such as neural nets, deep learning, and support vector machines,which I do not cover in this review.)"
- "Data manipulation tools and techniques developed for small datasets will become increasingly inadequate to deal with new problems. Researchers in machine learning have developed ways to deal with large datasets and economists interested in dealing with such data would be well advised to invest in learning these techniques."

- Another idea (Breiman, once again!):
  - « A variable might be considered important if deleting it seriously affects prediction accuracy. »
  - Applicable to any model including Random Forests: the values of the m^{th} variable are randomly permuted in all of the cases left out in the current bootstrap sample. Then these cases are run down the current tree and their classification noted. At the end of a run consisting of growing many trees , the percent increase in misclassification rate due to noising up each variable is computed.
- Also sensitivity analysis (partial derivatives or variance based methods) (Saltelli & al, 2000)

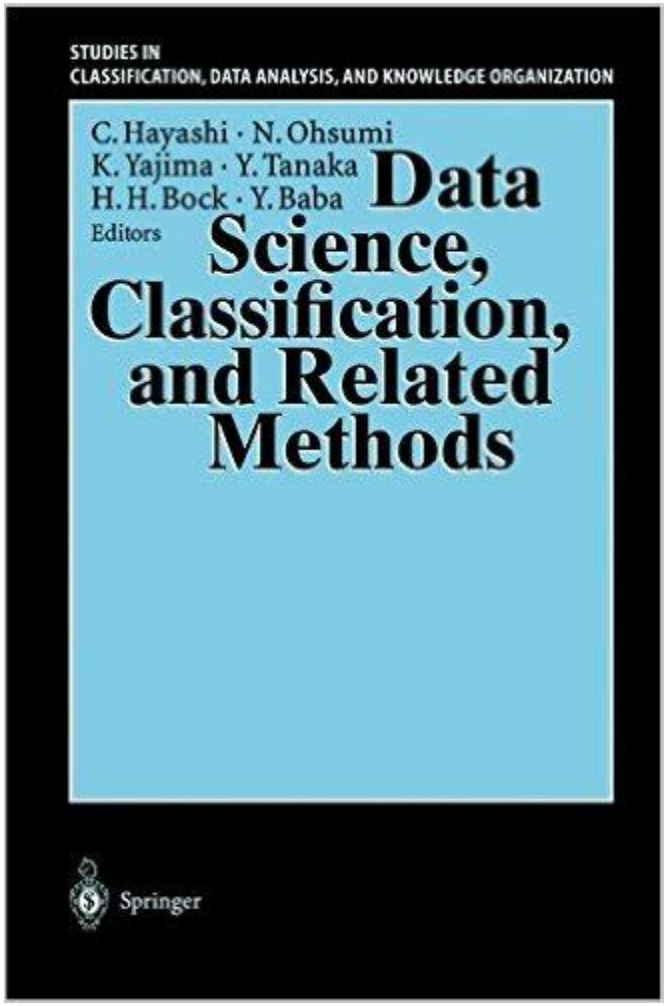
- However:
  - « holding all other variables fixed » is nonsense
  - When a predictor changes , it implies that other do : intervention (Bühlmann, 2013)
  - Causal schemes are necessary, see further

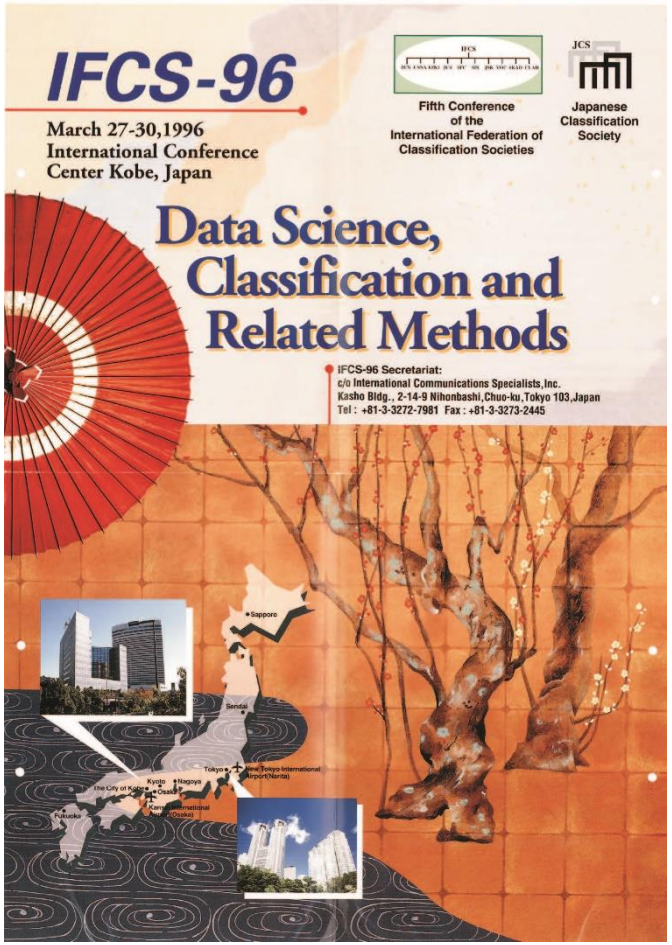
6. Data Science and the Data revolution

DEFINING THE DATA REVOLUTION

'The data revolution is: an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the 'Internet of Things,' and from other sources, such as qualitative data, citizen-generated data and perceptions data; A growing demand for data from all parts of society.'

UN Secretary-General's Independent Expert Advisory Group on a Data Revolution (A World That Counts report, page 6)


STUDIES IN
CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION
C. Hayashi · N. Ohsumi
K. Yajima · Y. Tanaka
H. H. Bock · Y. Baba
Editors
Data Science, Classification, and Related Methods
Springer


IFCS-96
March 27-30,1996
International Conference
Center Kobe, Japan
Fifth Conference of the International Federation of Classification Societies
Japanese Classification Society
Data Science, Classification and Related Methods
IFCS-96 Secretariat:
c/o International Communications Specialists,Inc.
Kasho Bldg., 2-14-9 Nihonbashi,Chuo-ku,Tokyo 103,Japan
Tel : +81-3-3272-7981 Fax : +81-3-3273-2445
Sapporo
Sendai
Tokyo
Nagoya
Kyoto
Osaka
Fukuoka

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.

Hopes

- Use of social media, web and smartphone data to improve health, quality of life, public statistics etc.
- UN, UNECE, Eurostat develop Big Data projects

WFP And UN Global Pulse Show How Big Data Can Save Lives And Fight Hunger
By Anoush Rima Tatevossian Apr 9, 2015

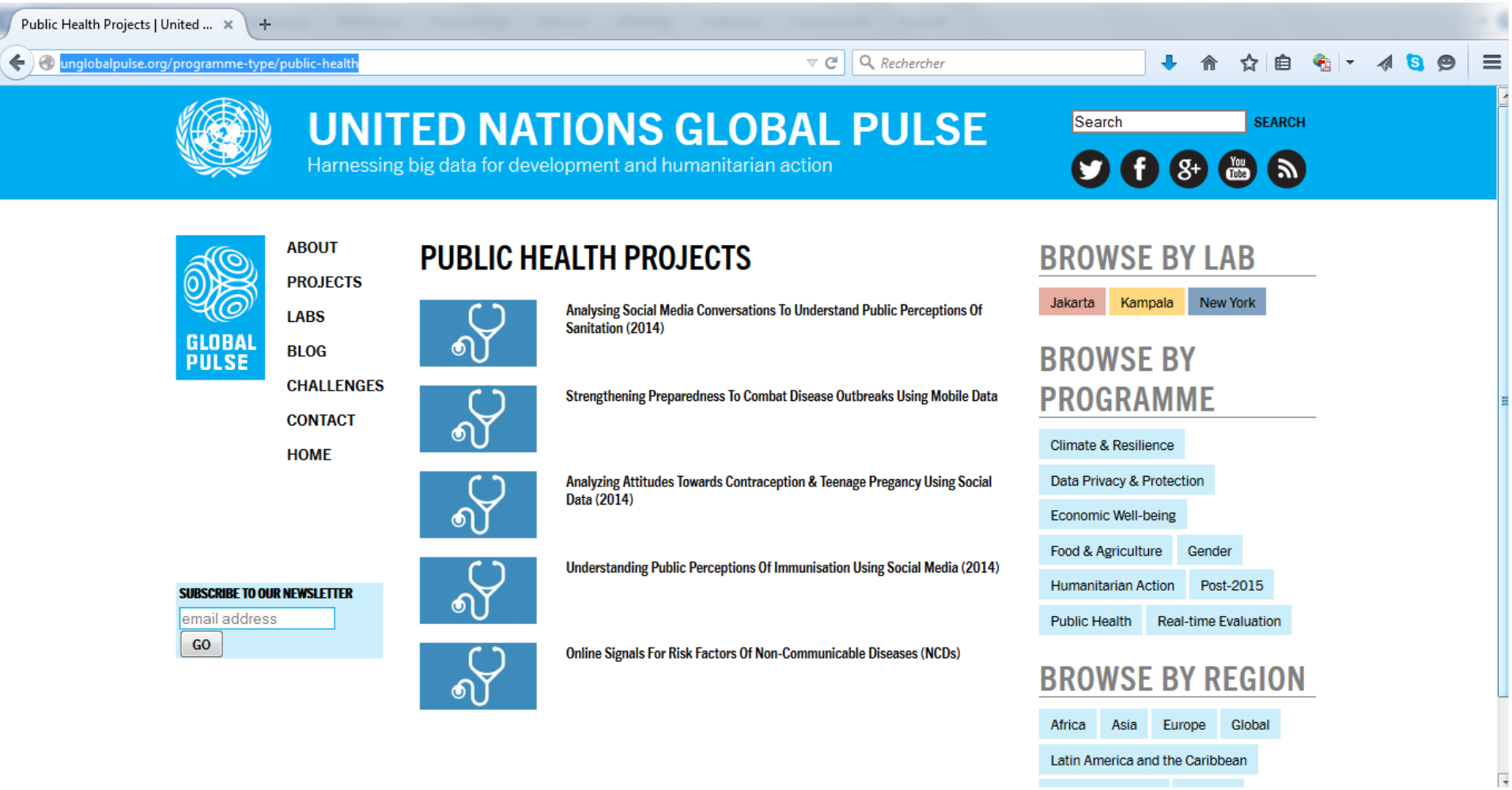
UNITED NATIONS GLOBAL PULSE

Harnessing big data for development and humanitarian action

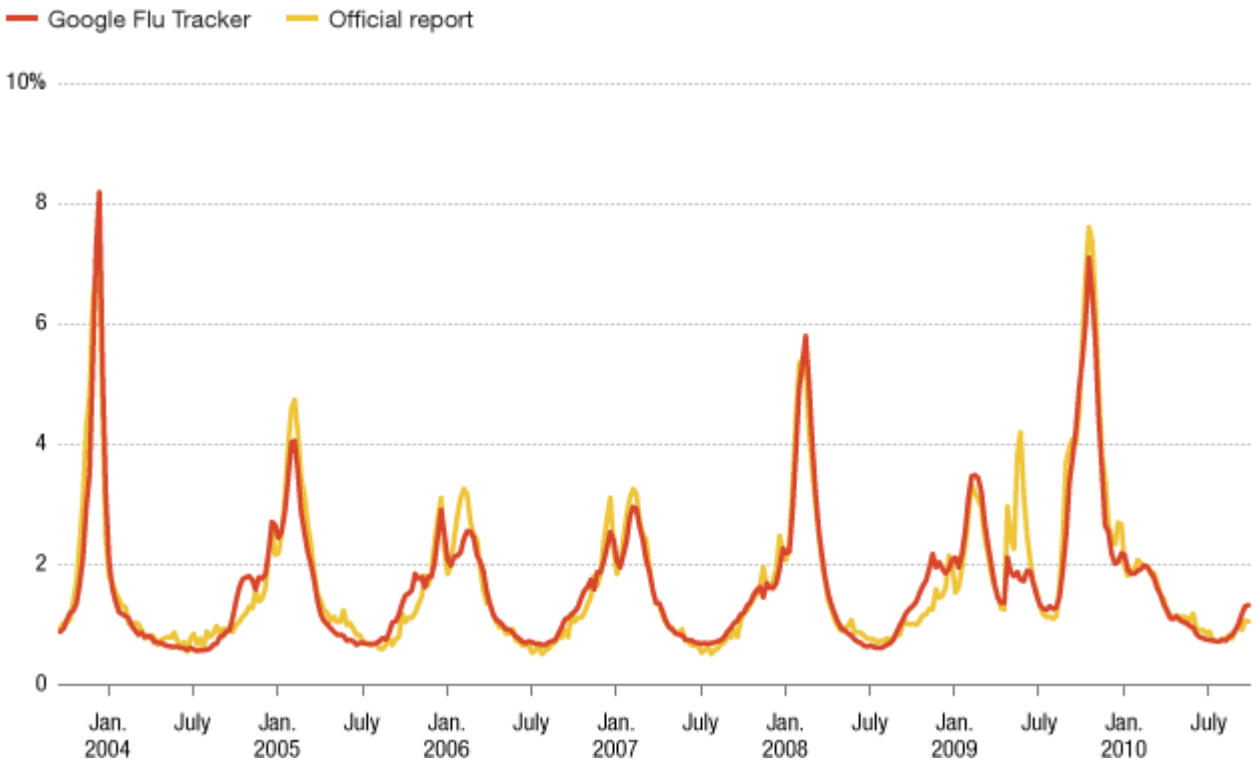
Global Pulse is a flagship innovation initiative of the United Nations Secretary-General on big data. Its vision is a future in which big data is harnessed safely and responsibly as a public good. Its mission is to accelerate discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action.
The initiative was established based on a recognition that digital data offers the opportunity to gain a better understanding of changes in human well-being, and to get real-time feedback on how well policy responses are working.

To this end, Global Pulse is working to promote awareness of the opportunities Big Data presents for relief and development, forge public-private data sharing partnerships, generate high-impact analytical tools and approaches through its network of Pulse Labs, and drive broad adoption of useful innovations across the UN System.

http://unglobalpulse.org/programme-type/public-health

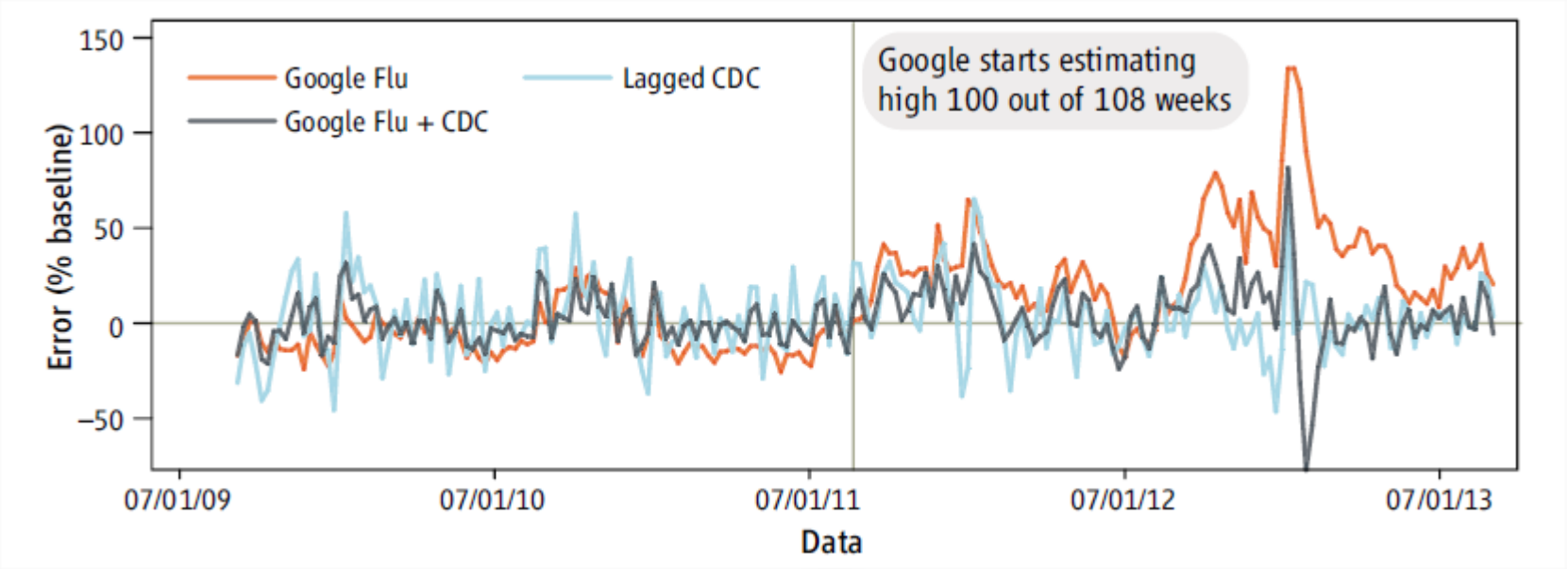


- Google FluTrends



And disappointments…

Overestimation by 50% in 2012-2013



BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

www.sciencemag.org SCIENCE VOL 343 14 MARCH 2014

- Big Data : no sampling errors but quality issues
  - Web data are produced by commercial companies mainly for their own business
  - Representativeness?
    - Companies constantly make modifications to their data collection algorithms in order to increase profits and support their business model. As a consequence, commercial data often are endogenously affected by a company's business decisions rather than exogenously determined which compromises their validity as a source of information. (Titiunik, 2015)

7. Understanding to better predict

- Correlation is not causality
  - Diapers and beer urban legend
- Causal inference from observational and interventional data is a hot topic (Bühlmann, 2013) as well as counterfactual inference
- Convergence between ML and computer science people, and statisticians.
  - See the NAS recent colloquium featuring Michael Jordan, Judea Pearl, Berhard Schölkopf, Peter Bühlmann, Léon Bottou, Hal Varian among many others

PROGRAMS

- Awards
- Koshland Science Museum
- Cultural Programs
- Sackler Colloquia
  - About Sackler Colloquia
  - Upcoming Colloquia
  - Completed Colloquia
  - Video Gallery
  - Connect with Sackler Colloquia
  - Give to Sackler Colloquia
- Kavli Frontiers of Science
- Distinctive Voices
- Sackler Forum
- Keck Futures Initiative

Arthur M. Sackler COLLOQUIA

Drawing Causal Inference from Big Data

This meeting was held March 26-27, 2015 at the National Academy of Sciences 2101Constitution Ave. NW in Washington, D.C.
Organized by Richard M. Shiffrin (Indiana University), Susan Dumais (Microsoft Corporation), Mike Hawrylycz (Allen Institute), Jennifer Hill (New York University), Michael Jordan (University of California, Berkeley), Bernhard Schölkopf (Max Planck Institute) and Jasjeet Sekhon (University of California, Berkeley)
Graduate Student / Postdoctoral Researcher travel awards sponsored by the National Science Foundation and the Ford Foundation.

Overview

This colloquium was motivated by the exponentially growing amount of information collected about complex systems, colloquially referred to as "Big Data". It was aimed at methods to draw causal inference from these large data sets, most of which are not derived from carefully controlled experiments. Although correlations among observations are vast in number and often easy to obtain, causality is much harder to assess and establish, partly because causality is a vague and poorly specified construct for complex systems. Speakers discussed both the conceptual framework required to establish causal inference and designs and computational methods that can allow causality to be inferred. The program illustrates state-of-the-art methods with approaches derived from such fields as statistics, graph theory, machine learning, philosophy, and computer science, and the talks will cover such domains as social networks, medicine, health, economics, business, internet data and usage, search engines, and genetics. The presentations also addressed the possibility of testing causality in large data settings, and will raise certain basic questions: Will access to massive data be a key to understanding the fundamental questions of basic and applied science? Or does the vast increase in data confound analysis, produce computational bottlenecks, and decrease the ability to draw valid causal inferences?

PS Political Science & Politics — American Political Science Association

Symposium: Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?

8 papers

Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science? - Introduction

William Roberts Clark and Matt Golder

PS: Political Science & Politics / Volume 48 / Issue 01 / janvier 2015, pp 65 - 70

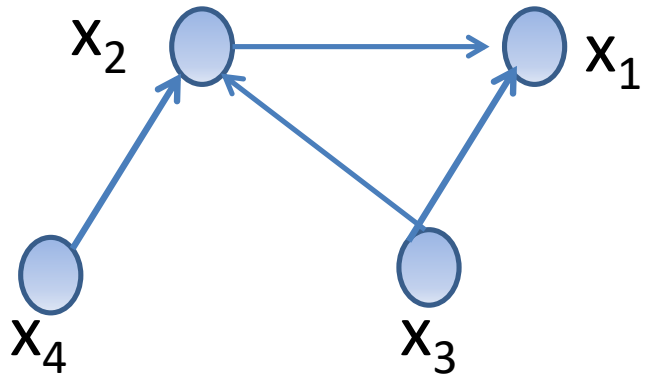
Can Big Data Solve the Fundamental Problem of Causal Inference?

Rocío Titiunik

PS: Political Science & Politics / Volume 48 / Issue 01 / janvier 2015, pp 75 - 79

- An hybrid model: complementing a regression scheme (linear or not) with a causal diagram

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



DAG: Directed Acyclic Graph

Conclusions

- Simple models
  - are not that simple
  - Give often poor prediction accuracy compared to ML algorithms
  - Truly generative models are rare
- Scientists should not be afraid using non explicit models
- An accurate predictive model improves knowledge
- Big Data urges causal inference and empirical inference to converge

Thanks for your attention

References

- C.Anderson (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, http://www.wired.com/2008/06/pb-theory/
- L.Bottou et al. (2013) Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, *Journal of Machine Learning Research*, 14, 3207–3260,
- L.Breiman (2001) Statistical Modeling: The Two Cultures, *Statistical Science*, 16, 3, 199–231
- P.Bühlmann (2013) Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research,* 77, 357-370
- D.Donoho (2015) 50 years of Data Science, *Tukey Centennial workshop*, https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf
- L.Fabbris (1980) Measures of regressor importance in multiple regression: an additional suggestion. *Qual Quant* 4:787–792

- U.Grömping, (2015). Variable importance in regression models. *WIREs Comput Stat* **7**, 137-152.
- Y.LeCun, Y.Bengio, G.Hinton (2015) Deep Learning, *Nature ,* 521, 436–444
- A.Saltelli, A Saltelli, K Chan, EM Scott (2000) *Sensitivity analysis*, Wiley
- G.Saporta (2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
- V.Vapnik (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer
- H.Varian (2014) Big Data: New Tricks for Econometrics, *Journal of Economic Perspectives*, 28, 2, 3–28
- H.Wallard (2015) Using Explained Variance Allocation to analyse Importance of Predictors, *16th ASMDA conference proceedings*, 1043-1054