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. The Big Data phenomenon
2. Big Data Analytics
3. A new conception of models
4. New technologies
5. The validation issue
6. The end of theory?
7. Skills and training

1. The Big Data phenomenon

Le Monde

Big data

Les données, nouvelle devise

Les grands volumes d'information finiront-ils par devenir la matière première des entreprises ? Ils constituent déjà, grâce au « big data », une ressource essentielle pour créer des produits plus ciblés et augmenter la productivité

06/11/2014

- A revolution

DEFINING THE DATA REVOLUTION

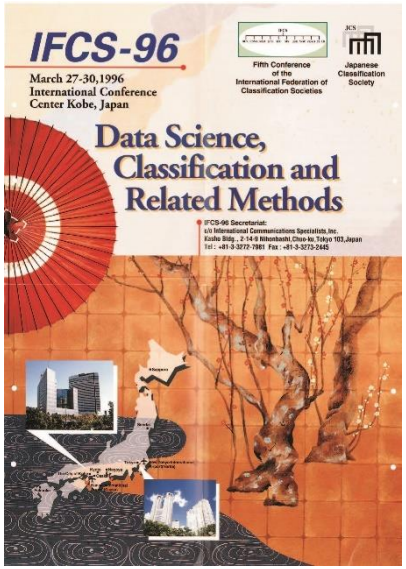
'The data revolution is: an explosion in the volume of data, the speed with which data are produced, the number of producers of data, the dissemination of data, and the range of things on which there is data, coming from new technologies such as mobile phones and the 'Internet of Things,' and from other 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)

- Origin:
  - Web, social media: digital footprints
  - Internet of things

- Big Data appears for the first time in 1997:
  - Cox & Ellsworth (NASA, not NSA!)
    « Managing Big Data for Visualisation »
    ACM SIGGRAPH '97

- Data Science :
  - P.Naur 1960
  - IFCS (Kobe, 1996) "Data Science, classification, and related methods“
  - Journal of Data Science since 2003

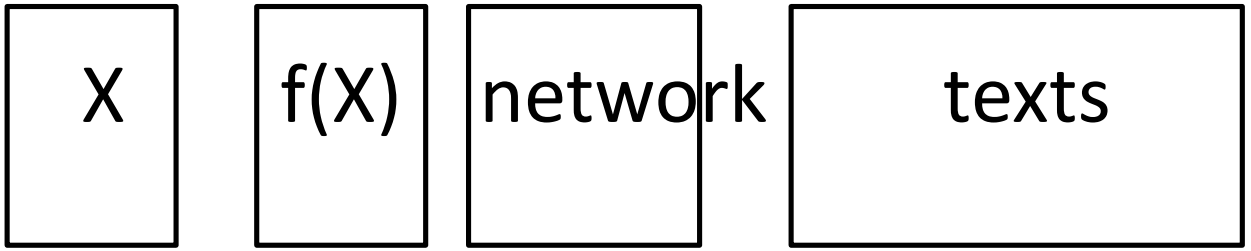


- The three V:
  - **Volume**
  - **Velocity**
  - **Variety**

  Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation. (Gartner)

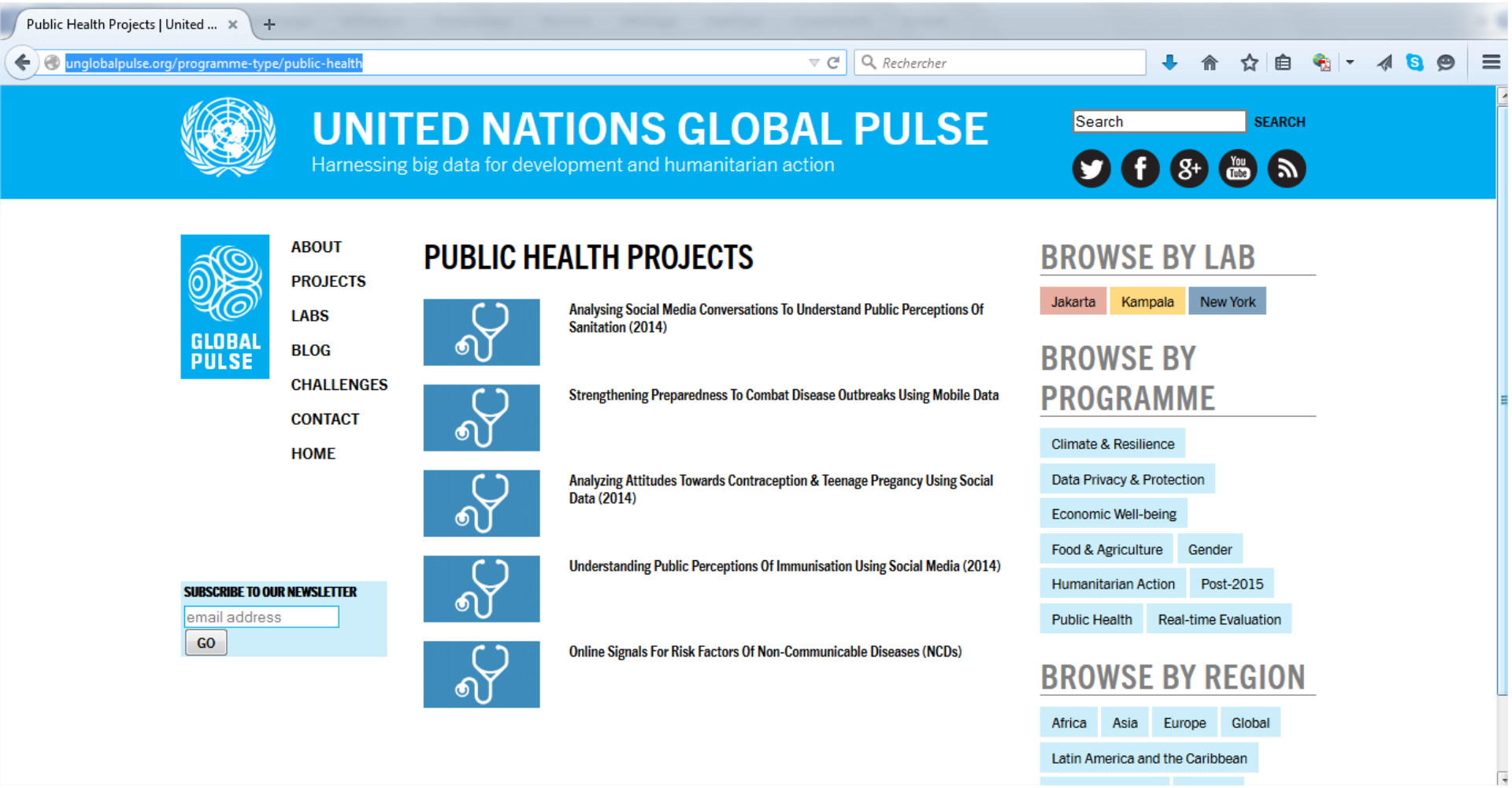
  - More V's: veracity, validity, visualisation, value..

- **Variety**: numeric, categorical, textual, network etc. data



« Feature engineering »

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

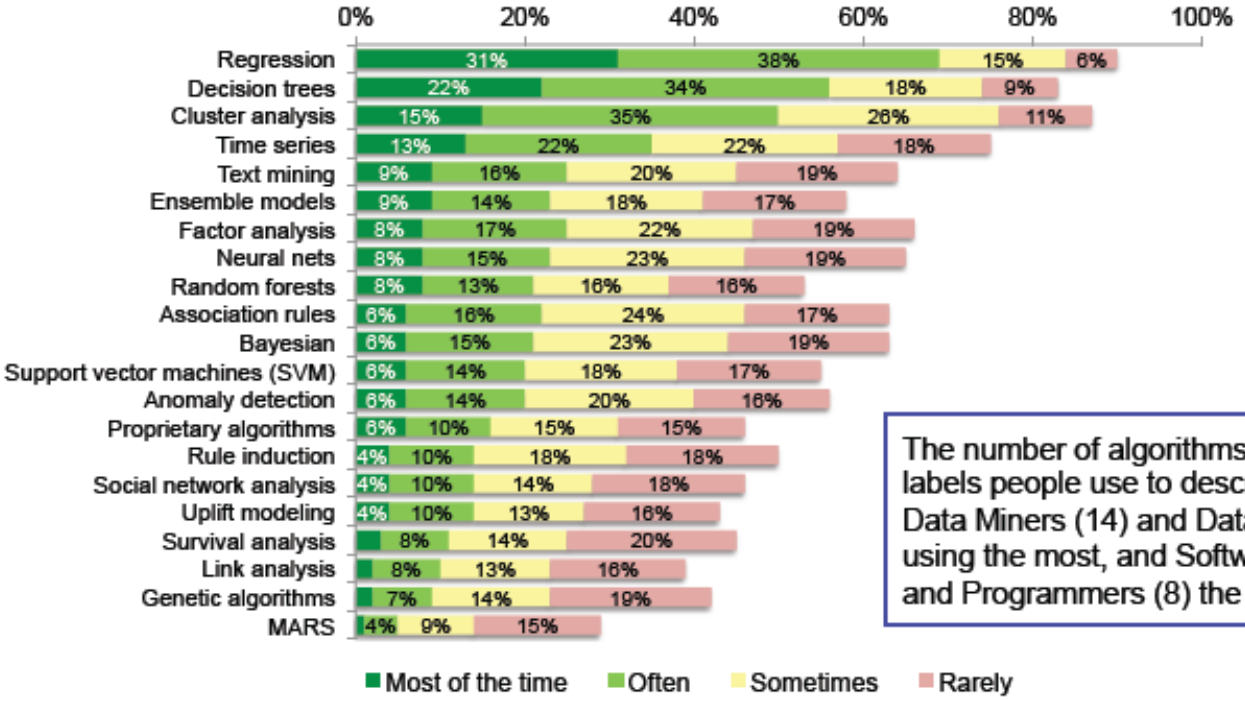


2. Big Data Analytics

- Exploratory or nonsupervised
  - Data visualisation, dimension reduction : factor analysis, k-means clustering
  - Association rules
- Predictive or supervised
  - Explicit models :regression, with or without regularisation, trees ..
  - Black boxes (neural nets, SVM, ..)

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)

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

3. A new conception of « models »

- 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. « Data driven »

- 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

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

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

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).

- « new » models coming from Machine Learning
  - Neural neworks and deep learning
  - SVM
  - Association rules and recommending systems(eg Amazon)
  - Random forests
  - Meta models and stacking

Stacking

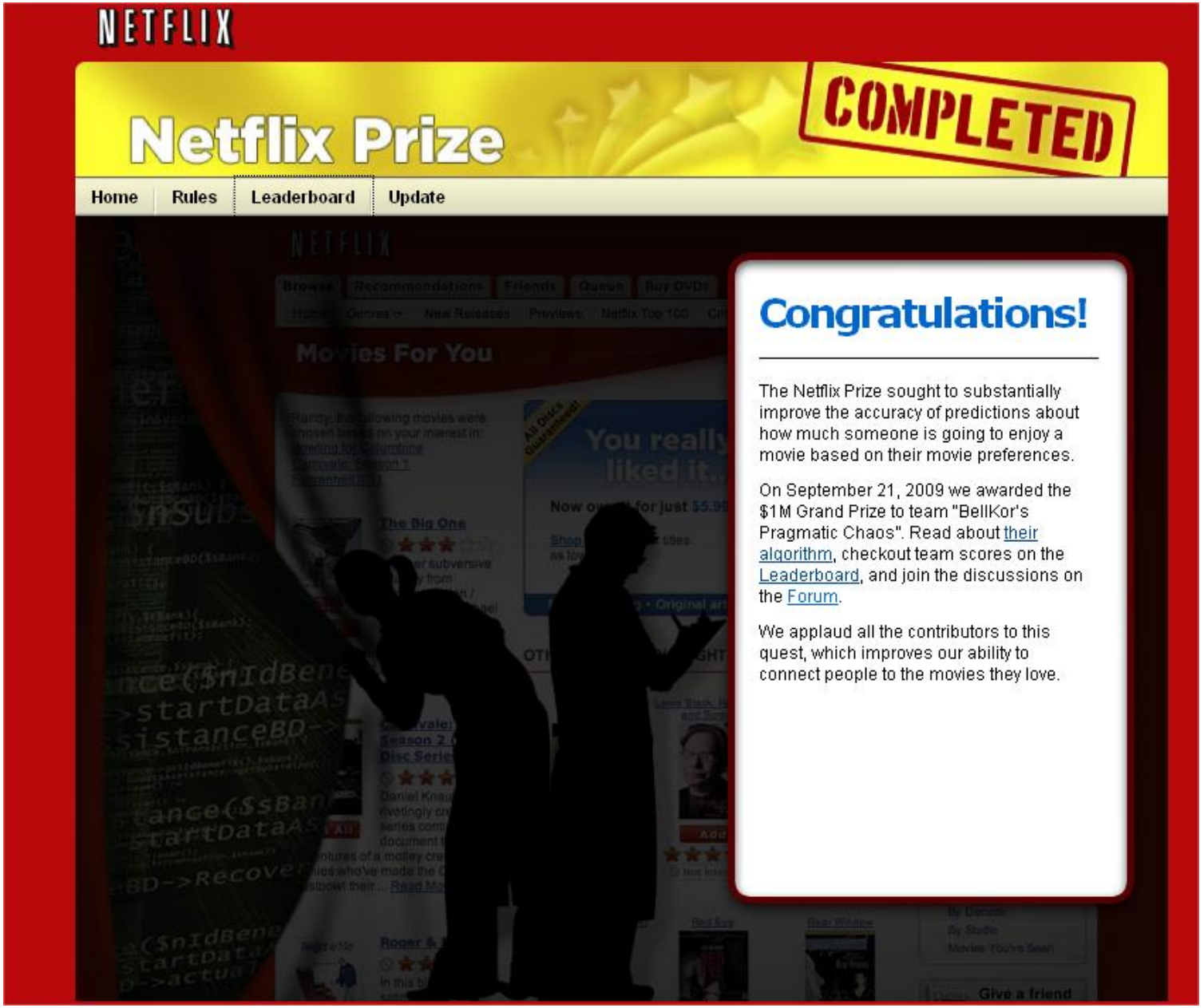
- M regression models (linear, non-linear, ...), give M predictions $\hat{f}_1(\mathbf{x}), \hat{f}_2(\mathbf{x}),..., \hat{f}_m(\mathbf{x})$
- Looking for the linear combination of $f_m(x_i)$, which gives the best prediction.
  - First idea: OLS

$$\min \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} w_j \hat{f}_j(\mathbf{x}) \right)^2$$

- Obtaining weights by OLS leads to overfitting since all models are not on the same foot (Hastie & al, 2009): the more complex a model is, the higher is its weight.
- Instead of standard predicted values, stacking uses the cross-validated prediction at x_i, not using x_i. Weights minimize:

$$\sum_{i=1}^{n}\left(y_i-\sum_{m=1}^{M}w_m f_m^{-i}\left(x_i\right)\right)^2$$

- When weights are constrained being positive and to sum 1, which is recommended, stacking looks like a frequentist version of Bayesian Model Averaging (BMA)
- Unlike BMA, stacking does not need that all models belong to the same kind, nor that the true model belongs to the family.
- One can mix k-nn, trees, Neural networks, et.
- Experiments proved that stacking outperforms BMA in a large number of cases (Clarke, 2003) involving much simpler computations


NETFLIX
Netflix Prize
COMPLETED
Home
Rules
Leaderboard
Update
Congratulations!
The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.
On September 21, 2009 we awarded the $1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their algorithm, checkout team scores on the Leaderboard, and join the discussions on the Forum.
We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

- The Netflix dataset contains more than 100 million datestamped movie ratings performed by anonymous Netflix customers between Dec 31, 1999 and Dec 31, 2005. This dataset gives ratings about m = 480 189 users and n = 17 770 movies
- The contest was designed in a training-test set format. A hold-out set of about 4.2 million ratings was created consisting of the last nine movies rated by each user (or fewer if a user had not rated at least 18 movies over the entire period).The remaining data made up the training set.

- *BellKor's Pragmatic Chaos team*. A blend of hundreds of different models
- *The Ensemble Team* . Blend of 24 predictions
- Same Test RMSE : 0.8567 (10.06%)

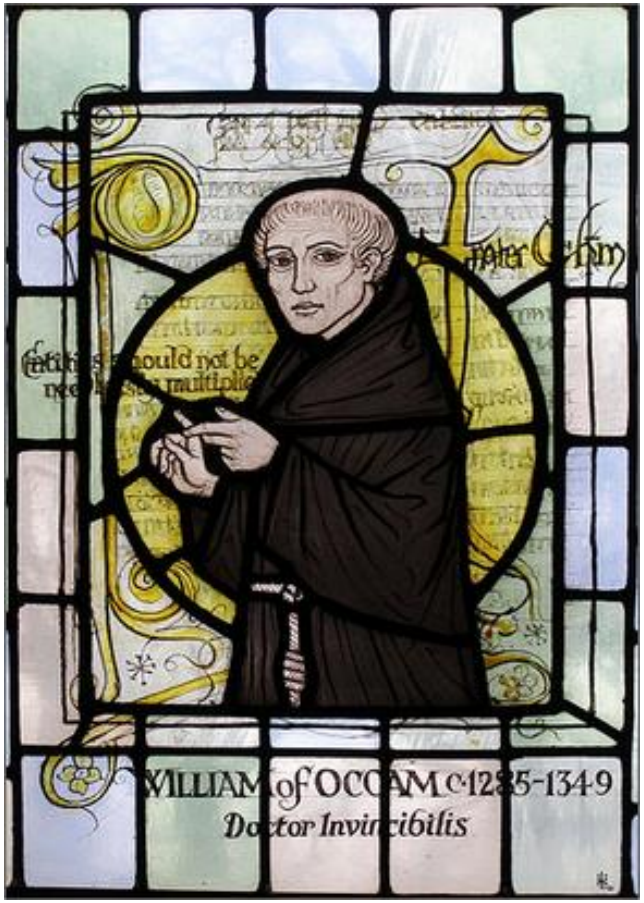
- Bellkor's Pragmatic Chaos defeated The Ensemble by submitting just 20 minutes earlier!

However Netflix did not implement the winning solution…

We evaluated some of the new methods offline but the additional accuracy gains that we measured did not seem to justify the engineering effort needed to bring them into a production environment.

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

The complexity challenge and model choice



- 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?

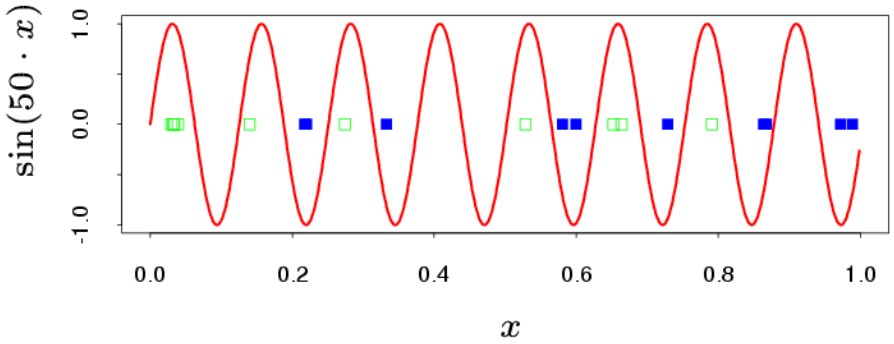
  “Essentially, all models are wrong, but some are useful ” (G.Box,1987)

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

- 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

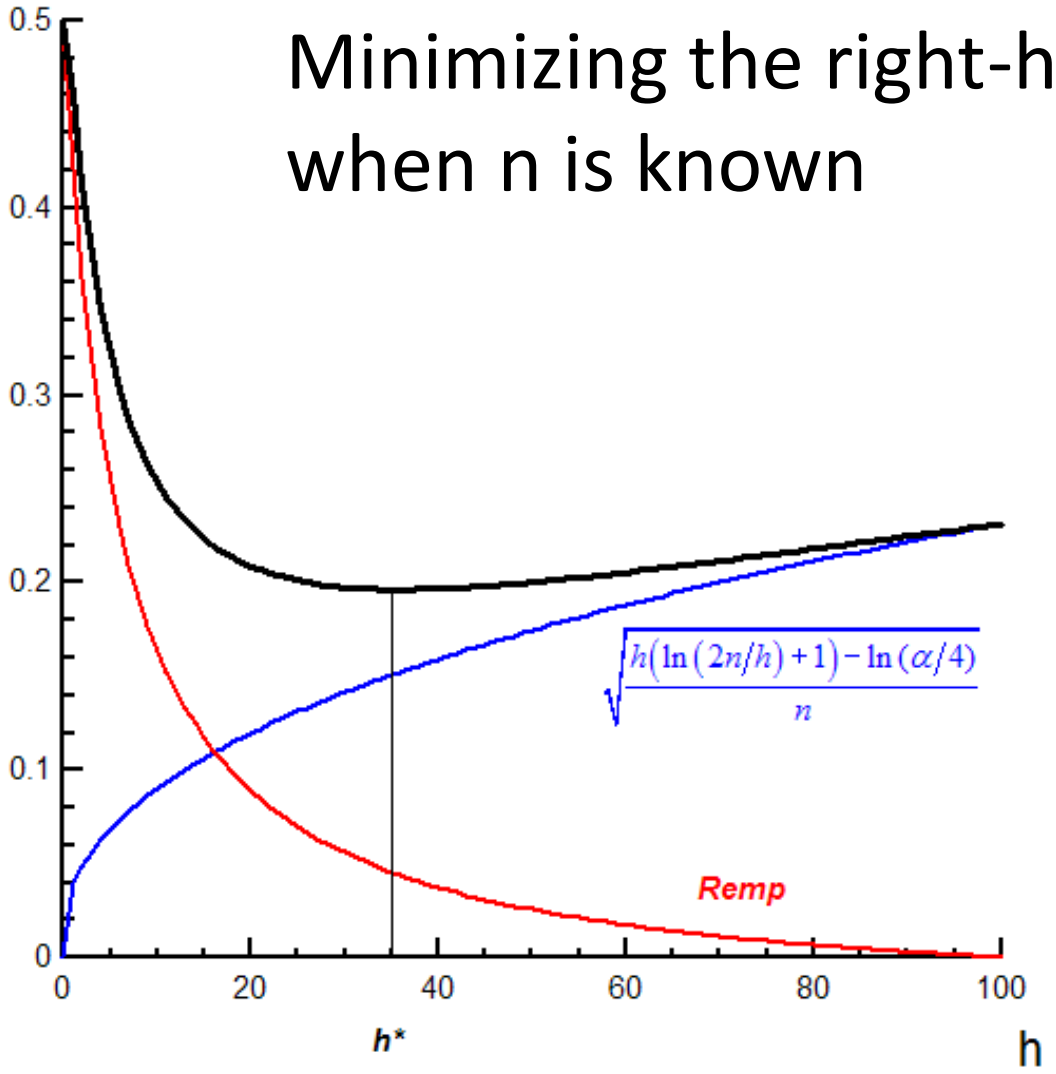
In supervised classification:

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

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 if data are also big according to p (high-dimensional data)
  - Difficult to interpret
  - Curse of dimensionality
  - Solution: sparsity constraints (Lasso)

4. New technologies

- Programming languages
  - Python versus R?
- New environments

Spark and MLLib

- Cloud computing
- NoSQL
- Solutions provided by internet big companies
  - MapReduce (Google)
  - Hadoop (Apache Foundation)
  - TensorFlow (Google)

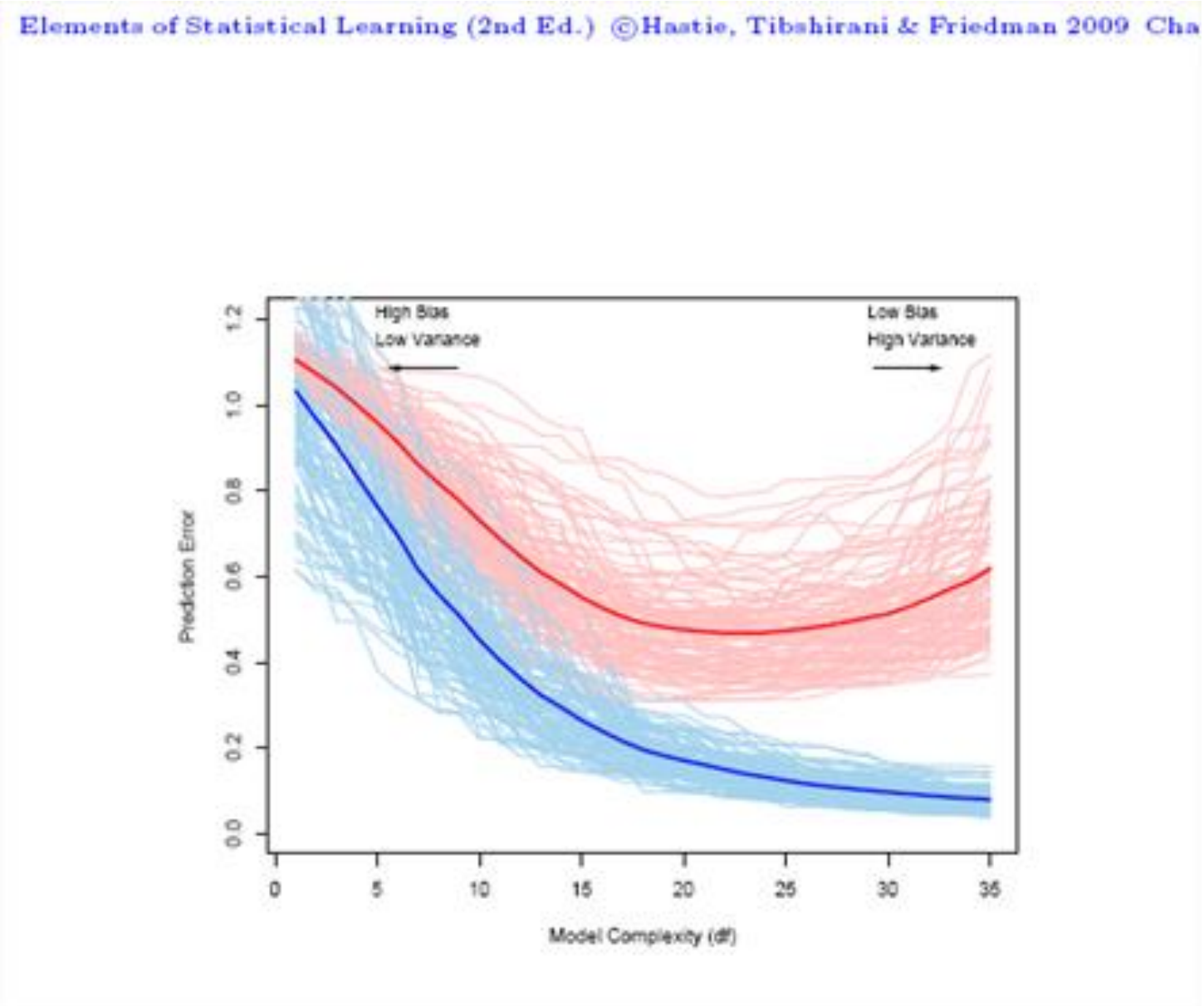
5. Empirical validation

- 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)

6. The end of theory?

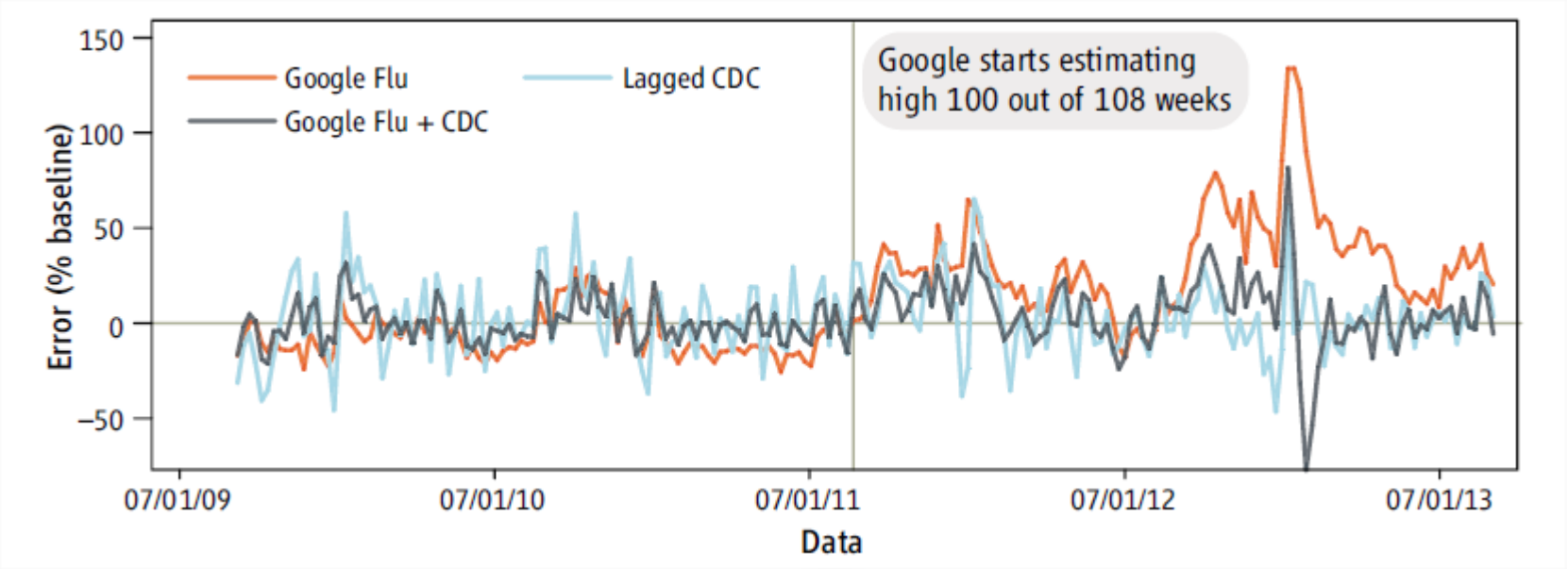
Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

- Google FluTrends

« **Google Flu Trends** was a web service operated by Google.
It provided estimates of influenza activity for more than 25 countries.
By aggregating Google search queries, it attempted to make accurate predictions about flu activity. http://www.google.org/flutrends/ »

http://esante.gouv.fr/le-mag-numero-10/decryptage-le-big-data-sante

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

- Correlation is not causality
  - Diapers and beer urban legend
- 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
- 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?

7. Skills

- Massive data need specific approaches
- Good old methods (PCA) still efficient , mainly for unsupervised problems
- **Data scientists**: a new kind of statisticians for Big Data?
  - Data Scientist (n.): Person who is better at statistics than any software engineer and better at software engineering than any statistician (Donoho, 2015)

Data Scientist:

The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.

by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

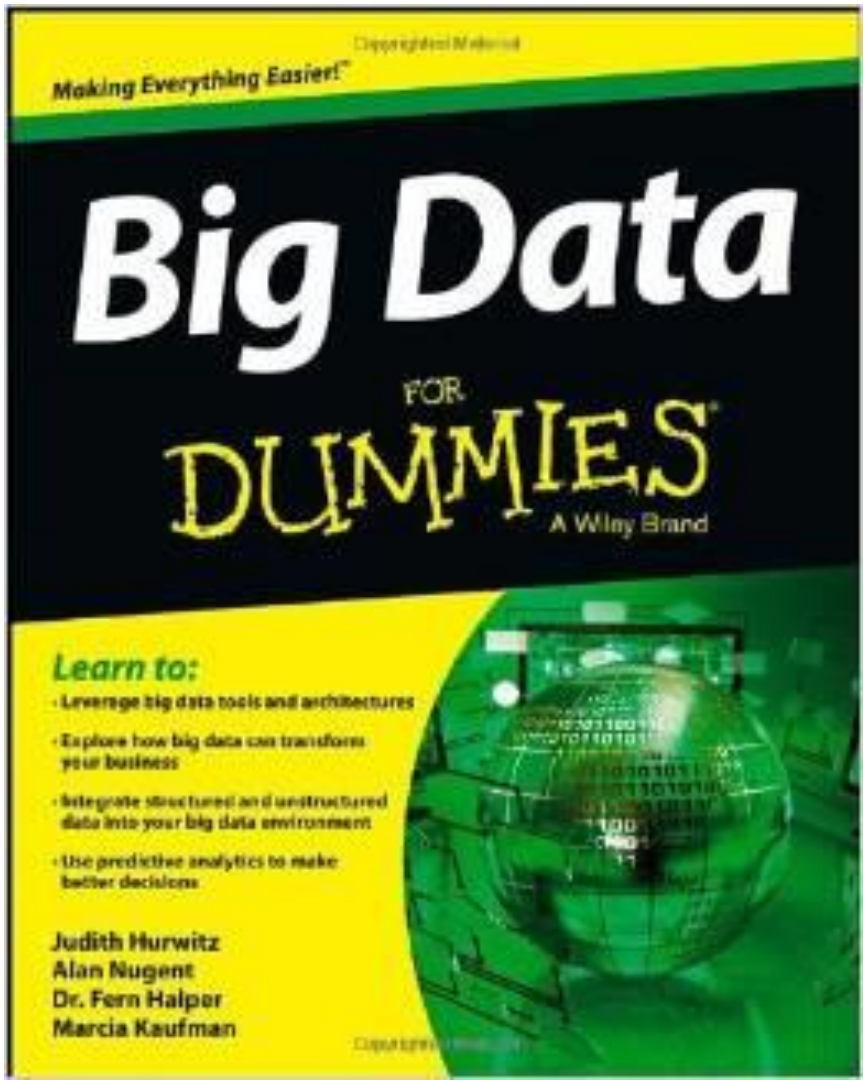
Thanks for your attention

The two cultures: a few references

- Breiman L. (2001) Statistical modeling: The two cultures. *Statistical Science,* **16,** 199–215.
- Donoho D. (2015). 50 years of Data Science, Tukey Centennial workshop, https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf
- Saporta G.(2008) Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322
- Shmueli G. (2010) To explain or to predict? *Statistical Science,* **25,** 289–310

Additional 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,
- P.Bühlmann (2013) Causal statistical inference in high dimensions. *Mathematical Methods of Operations Research,* 77, 357-370
- Y.LeCun, Y.Bengio, G.Hinton (2015) Deep Learning, *Nature ,* 521, 436–444
- 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


Making Everything Easier!™
Big Data
FOR
DUMMIES
A Wiley Brand
Learn to:
· Leverage big data tools and architectures
· Explore how big data can transform your business
· Integrate structured and unstructured data into your big data environment
· Use predictive analytics to make better decisions
Judith Hurwitz
Alan Nugent
Dr. Fern Halper
Marcia Kaufman


Tout savoir sur...
Big Data
Le cinéma avait déjà tout imaginé !
UN CRIME DANS LA TÊTE
BACK TO THE FUTURE
SCHWARZENEGGER
TERMINATOR 3
MATRIX
SCHWARZENEGGER
TOM CRUISE
THE MANCHURIAN CANDIDATE
CARREY
THE TRUMAN
DEAD ZONE
REQUIEM
MINORITY REPORT
Un livre de Xavier Perret & Guy Jacquemelle
Préface de Stéphane Richard
Une collection dirigée par Henri Kaufman
EDITIONS kawa
www.editions-kawa.com