Big data : challenges and opportunities for mathematicians

Data Science Center Eindhoven

Big Data: Challenges and Opportunities for Mathematicians

Alessandro Di Bucchianico

Dutch Mathematical Congress
April 15, 2015

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts

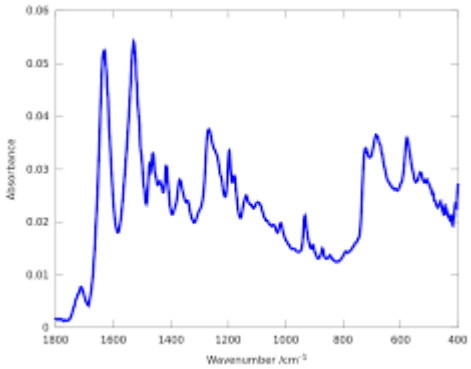
Contents

1. Big Data terminology
2. Various mathematical topics
3. Conclusions

What is big data?

Term coined by John Mashey, chief scientist at Silicon Graphics in the 1990's

…I was using one label for a range of issues, and I wanted the simplest, shortest phrase to convey that the boundaries of computing keep advancing…



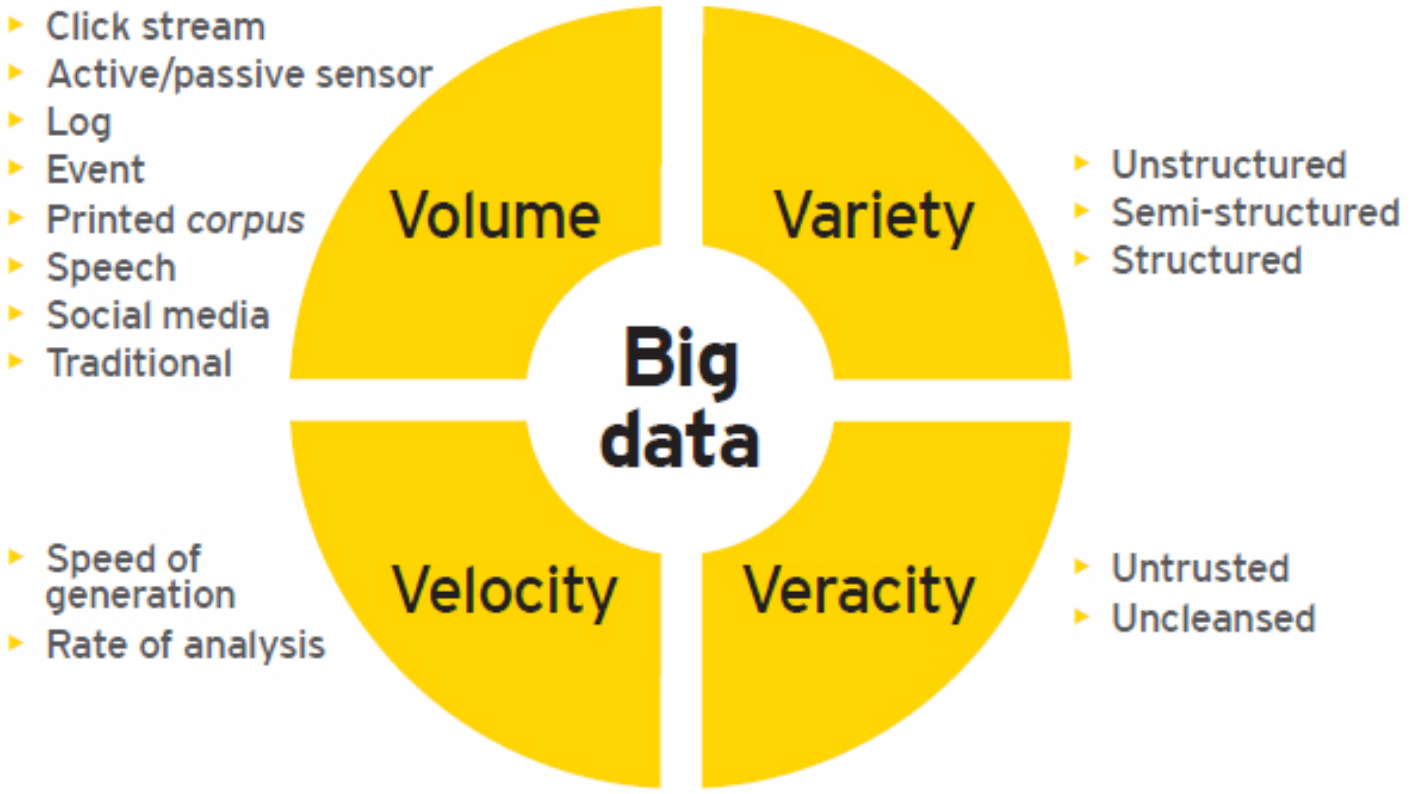
But : chemometrics has a long history of analyzing "large" data sets

How big is big data?

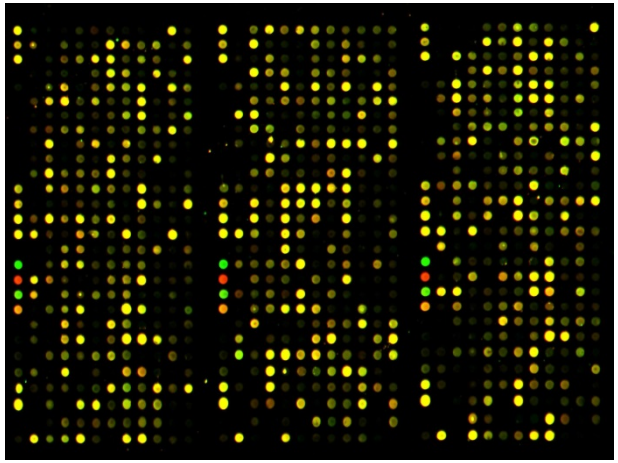
Horizon 2020 EU-ICT16-2015 call:

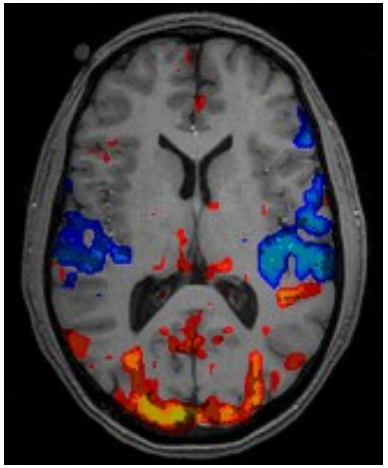
"extremely large" means "so large that today no amount of money could buy a system capable of handling it".

Four V's of Big Data



High-dimensional data: "$n \ll p$"





Big Data buzz words

- scalable algorithm
- data science / data scientist
- streaming data
- data warehousing and ETL
- in-memory database
- predictive analytics / predictive modelling
- high performance computing (exascale computing)

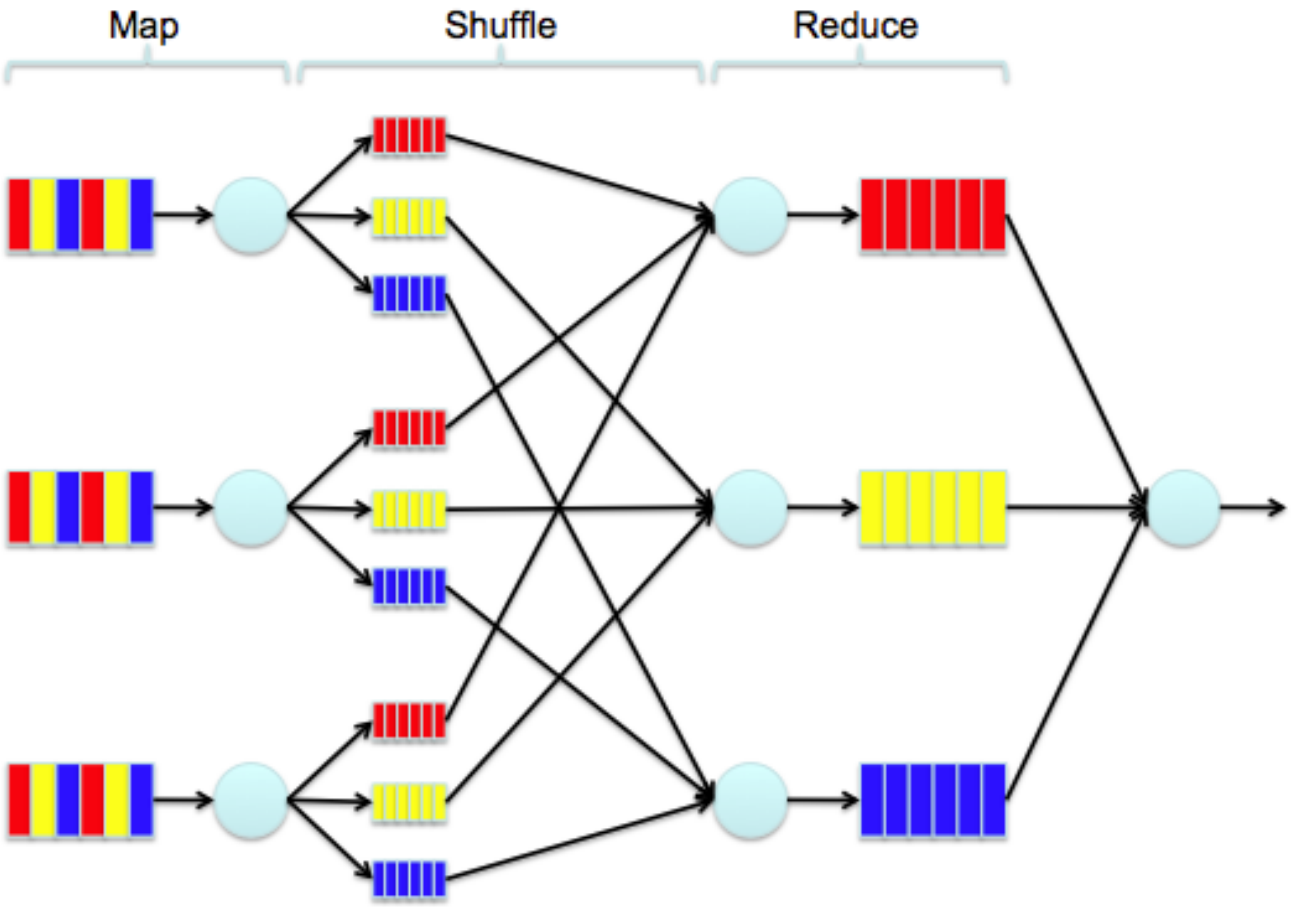
Machine learning and data mining

Machine learning focuses on prediction based on known properties “learned” from training data

Data mining focuses on the discovery of (previously) unknown properties of the data

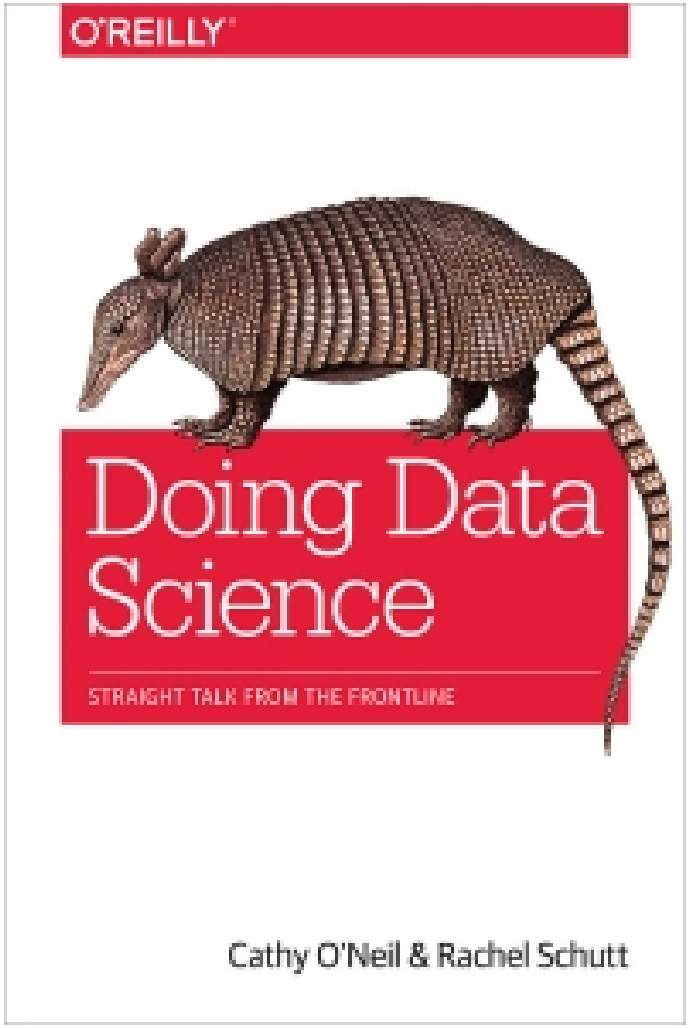
Leo Breiman: Two cultures

MapReduce

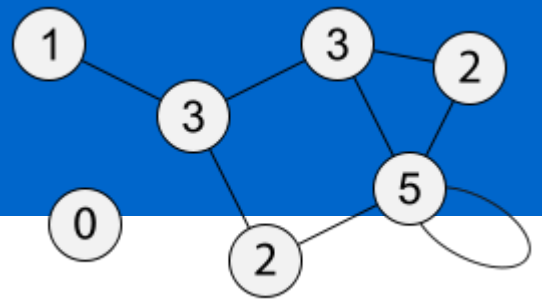


Open source tools

Course @ Columbia University



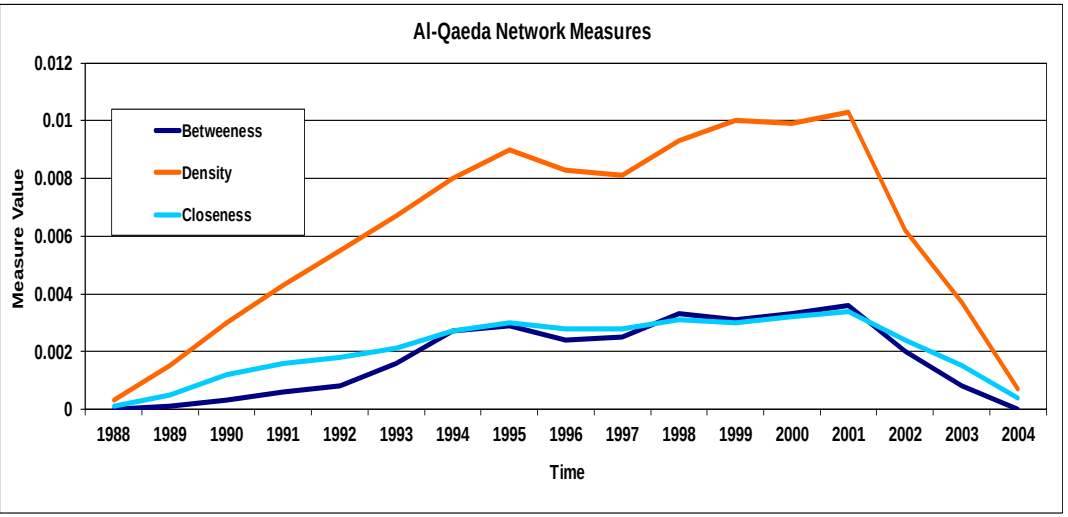
Topic: Network structure

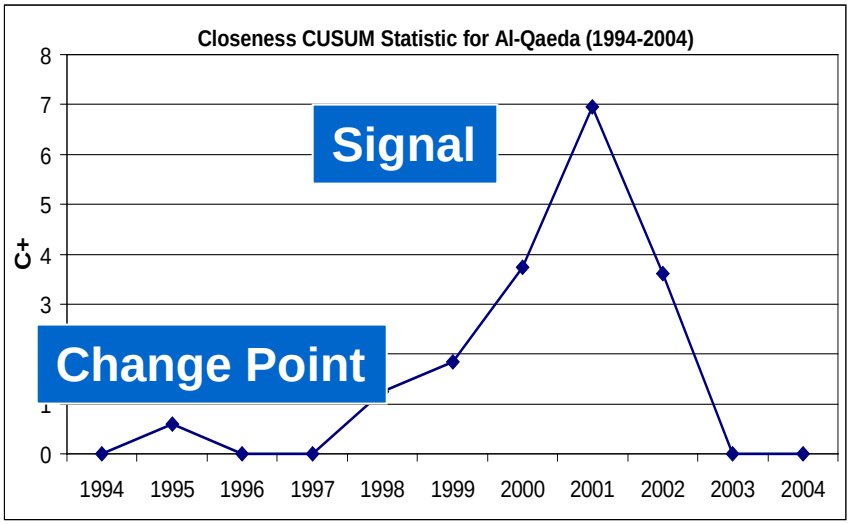


- **dependencies between nodes in networks measured by degrees of direct neighbours**
- **assortativity coefficient of Newman is nothing but Pearson's correlation statistic**
- **inconsistent estimator when variances are infinite**
- **Spearman rank correlation behaves better but calculation is computationally intensive**
- **requires heavy asymptotics**

Van der Hofstad, R. and Litvak, N. (2014) *Degree-Degree Dependencies in Random Graphs with Heavy-Tailed Degrees.* Internet Mathematics, 10 (3-4). pp. 287-334

Topic : Network monitoring





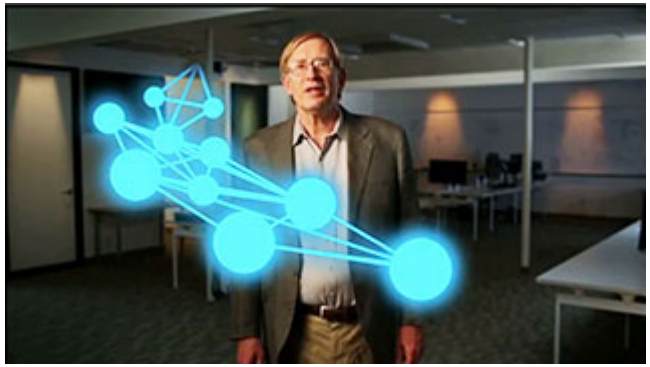
Challenges:

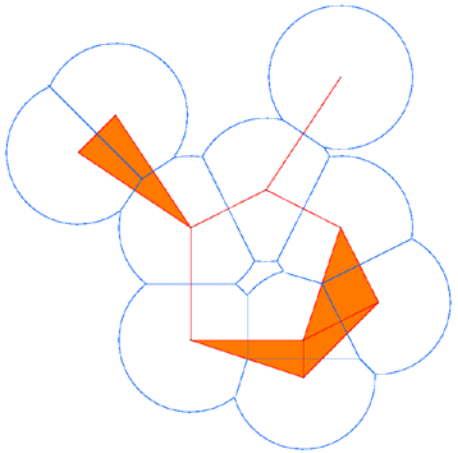
- **monitor high-number of variables**
- **models to capture structural changes**
- **scalable algorithms for likelihood ratio calculations**

Topic: Topological Data Analysis

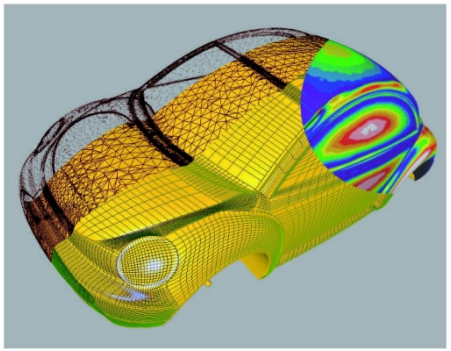
Common Big Data problem is to choose relevant "features" from high-dimensional data

Combination of machine learning with topological tools (simplices, cohomology) yields new algorithms for clustering





Topic: Uncertainty Quantification



Virtual prototyping using mathematical models. UQ does not deal with

1. unknown uncertainty in the initial conditions of parameters
2. parametrisation of design building blocks

For 1) : polynomial chaos (Wiener chaos) , inverse statistical models, Bayesian analysis (calibration)

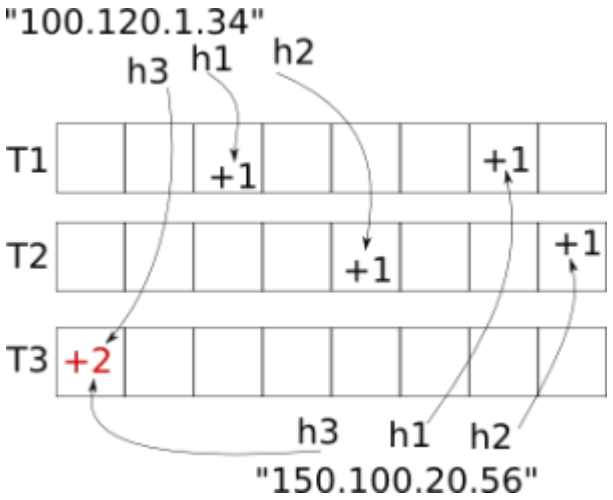
For 2): Model Order Reduction

Topic: Streaming algorithms

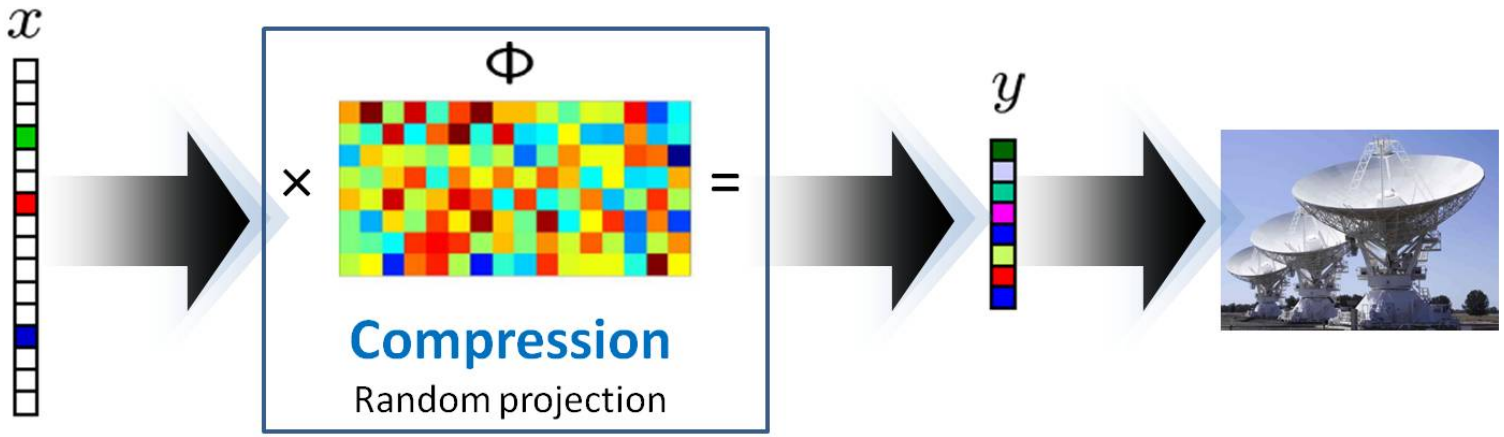
High volume high velocity data makes exact counting of frequencies or number of different items impossible but approximate answers suffice in big data.

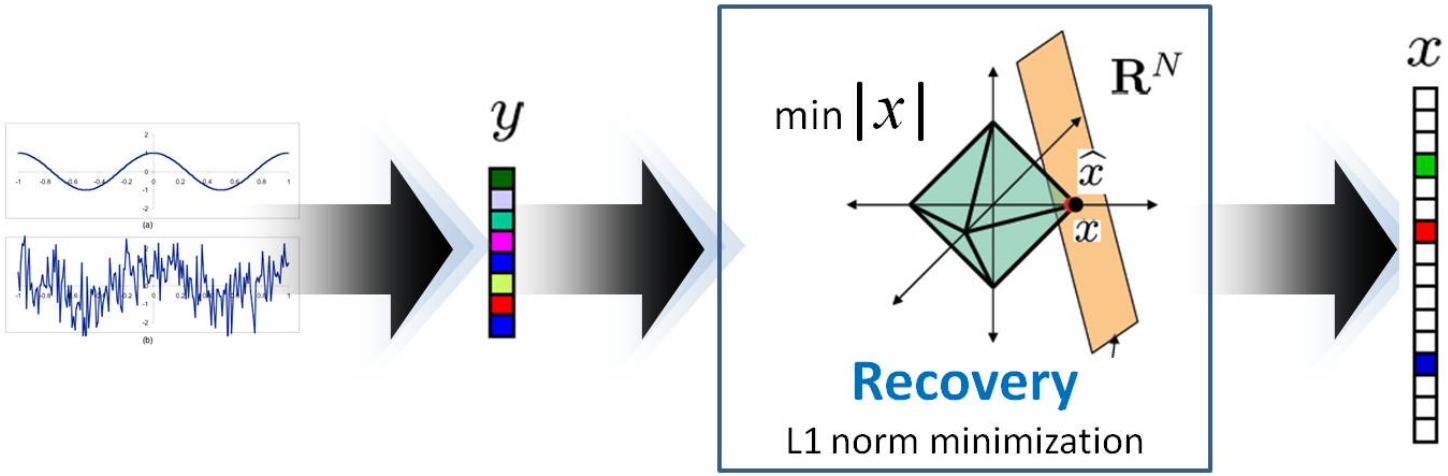
New ideas using probabilistic means (random hash functions):

- **Count-Min Sketch**
- **MinHash**



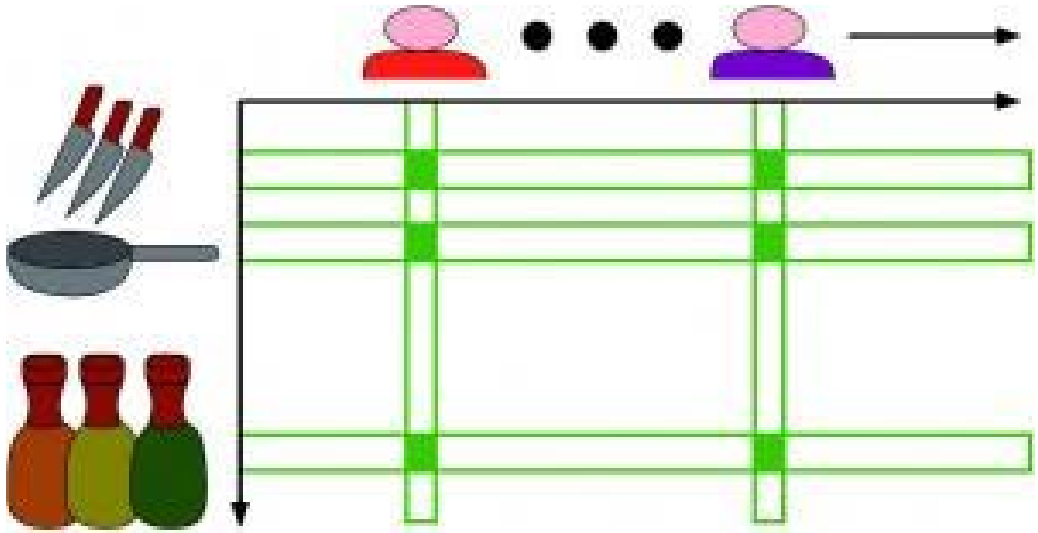
Topic: Compressed Sensing





Topic: Recommendation systems

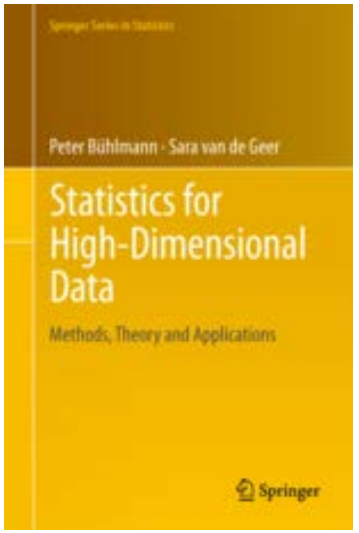
Approach (Moitra): Use connection to existence theorems for polynomials solutions to algebraic equations and develop scalable algorithms for nonnegative matrix factorizations



Statistical challenges

Big Data shows phenomena not present in “small data”:

- **heterogeneity**
- **spurious correlations**
- **noise accumulation**
- **incidental endogeneity**



This requires critically revising statistical models and developing new tools

Big Data Research Funding

Big Data Challenges

NWO

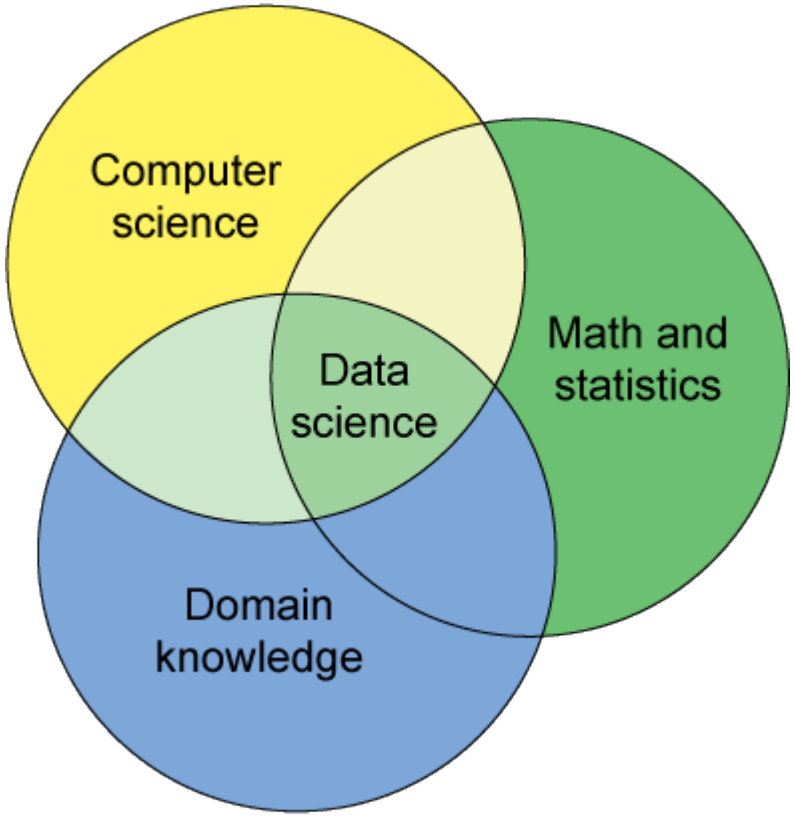
Digital Sciences

High Performance Computing

Mathematical contributions in general

- **Modelling**
- **Performance of algorithms**
- **Statistical thinking**

We need to work hard to make mathematical contributions more explicit.



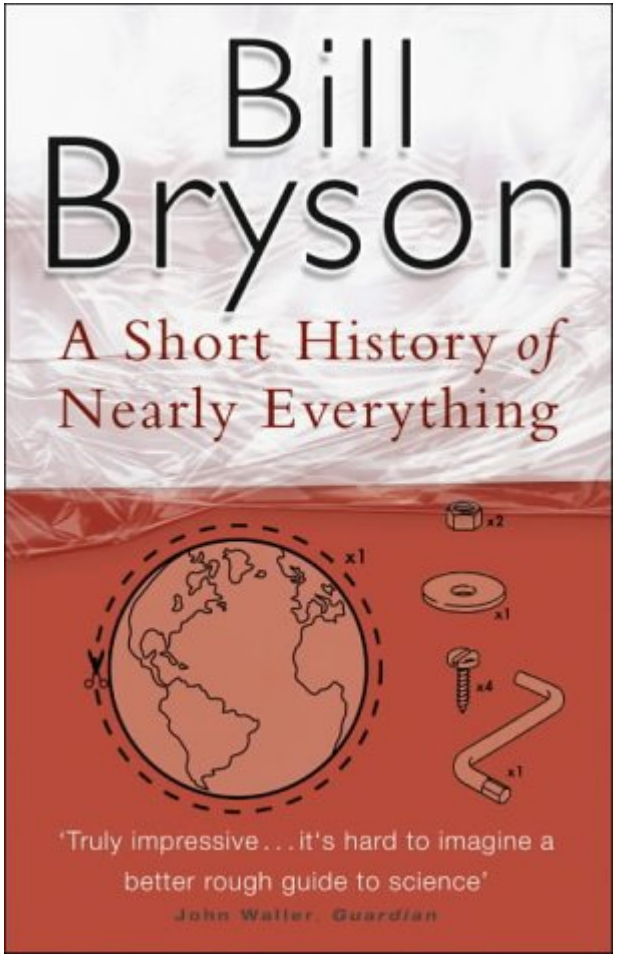
Actions

- **get involved in local data science centres**
- **get into touch with NWO**
- **get into touch with EU (e.g., Nov 6 2014 Workshop "Mathematics for Digital Sciences")**
- **2014 IMS presidential address Bin Yu:**

"***work on real problems, relevant theory will follow***"

Journals

Learn from the past



Conclusions

- interesting mathematics behind big data
  - statistics
  - combinatorics
  - numerical mathematics
  - topology
  - ...
- efforts required
  - learn big data concepts
  - get into touch with computer scientists
- actions required to ensure funding
  - data science centres
  - funding agencies