

# Big data : challenges and opportunities for mathematicians

***Citation for published version (APA):***

Di Bucchianico, A. (2015). *Big data : challenges and opportunities for mathematicians*. 51ste Nederlands Mathematisch Congres (NMC 2015), 14-15 April 2015, Leiden, Nederland, Leiden, Netherlands.

***Document status and date:***

Published: 01/01/2015

***Document Version:***

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

***Please check the document version of this publication:***

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

***General rights***

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

***Take down policy***

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

# 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



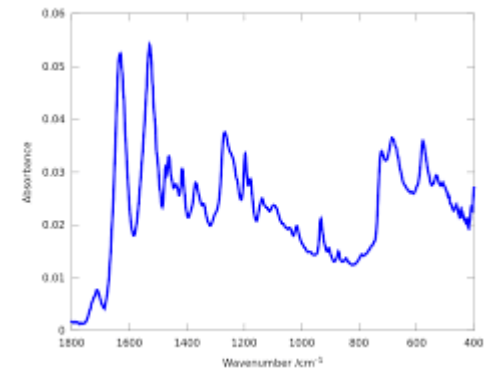
# What is big data?

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



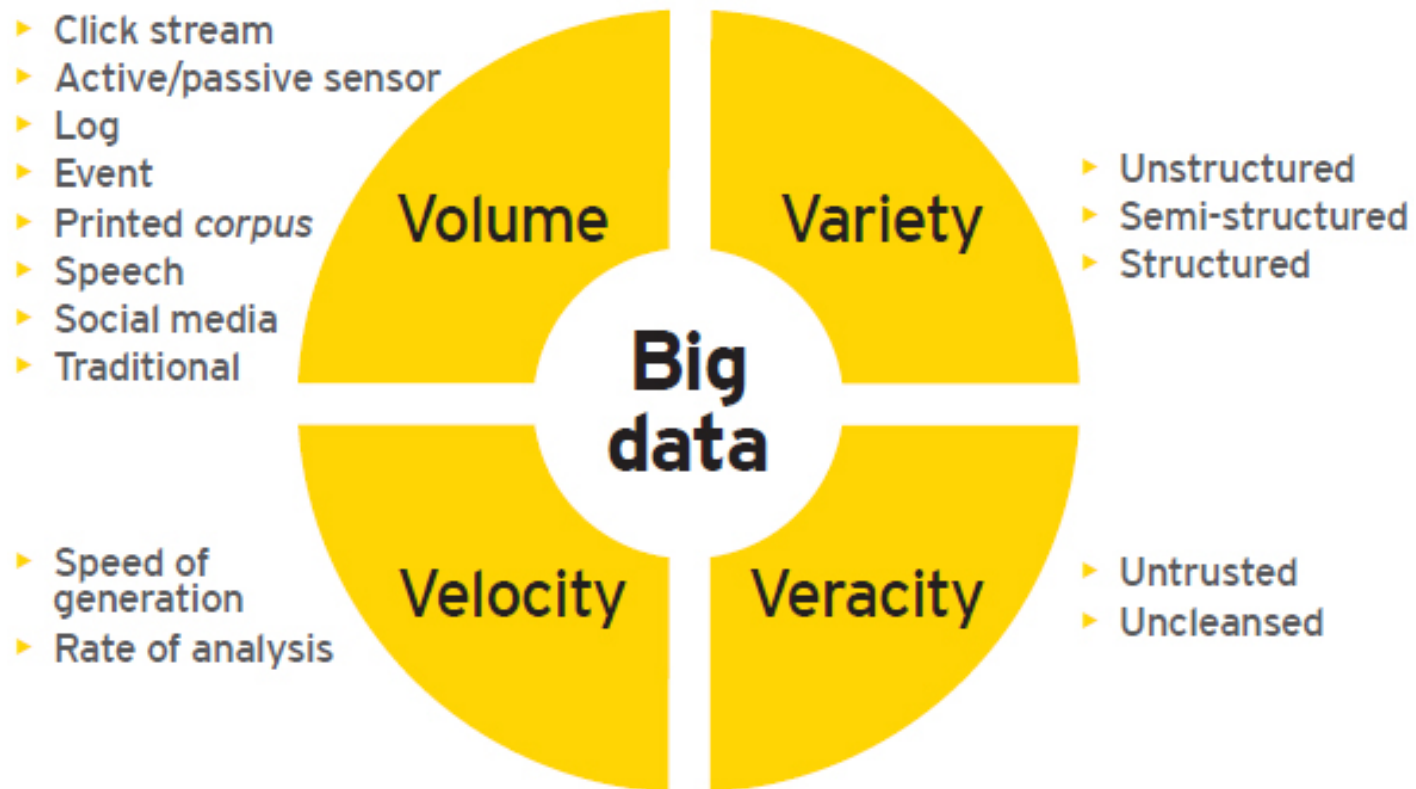
# How big is big data?

Horizon 2020 EU-ICT16-2015 call:

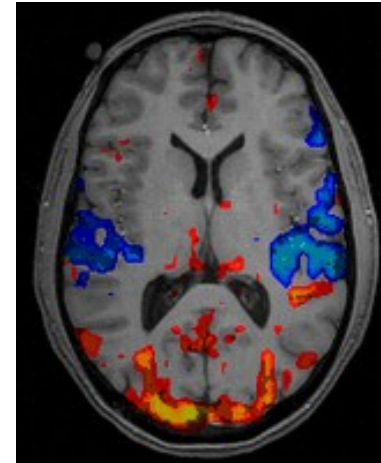
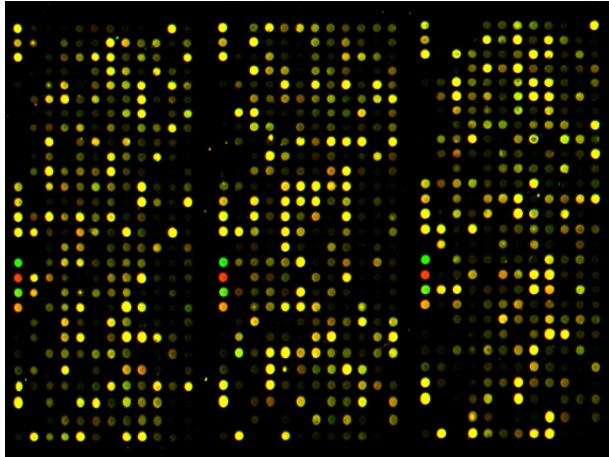
"extremely large" means "so large that today no amount of money could buy a system capable of handling it".



# 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

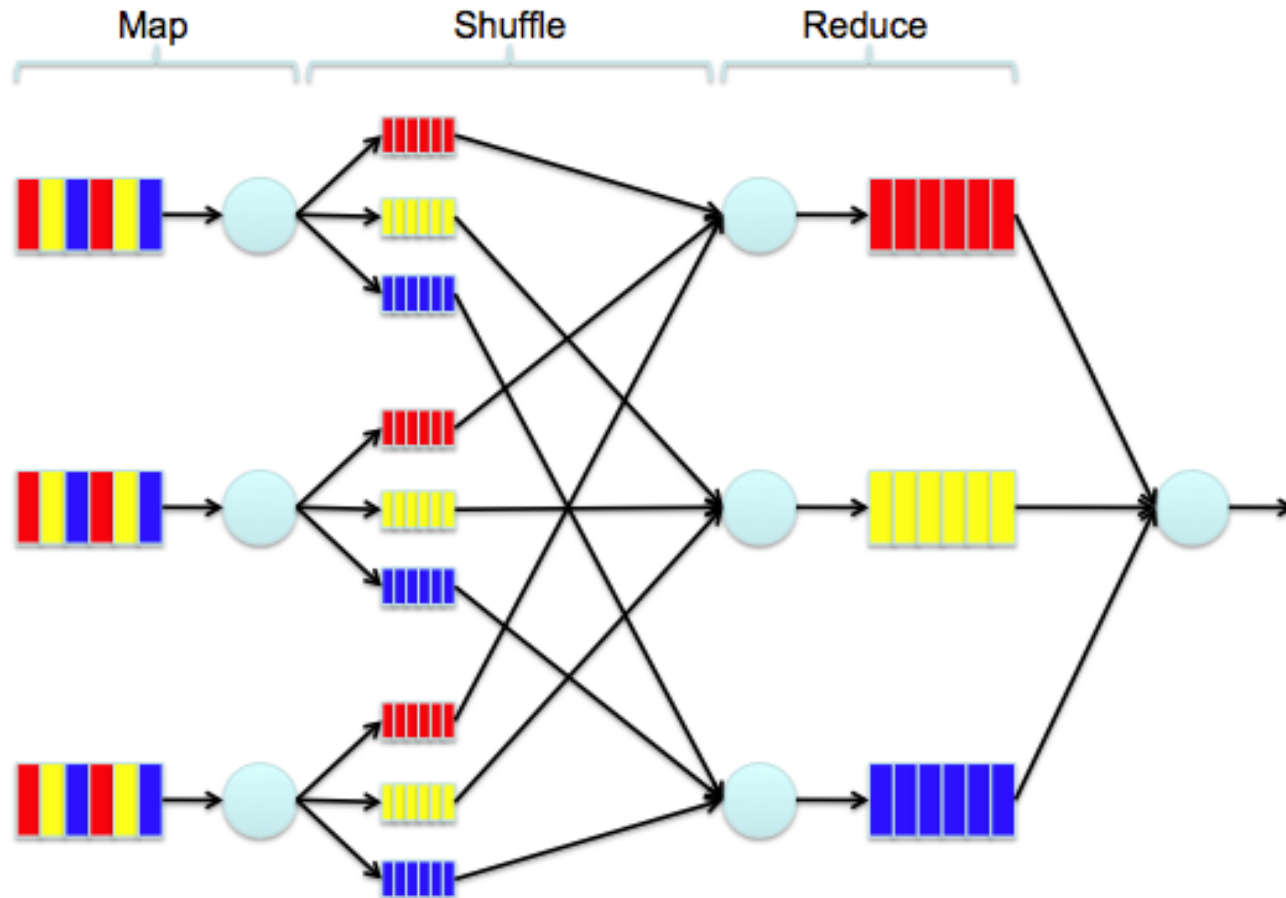
**Machine learning** focuses on prediction based on known properties “learned” from training data

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



Leo Breiman: Two cultures

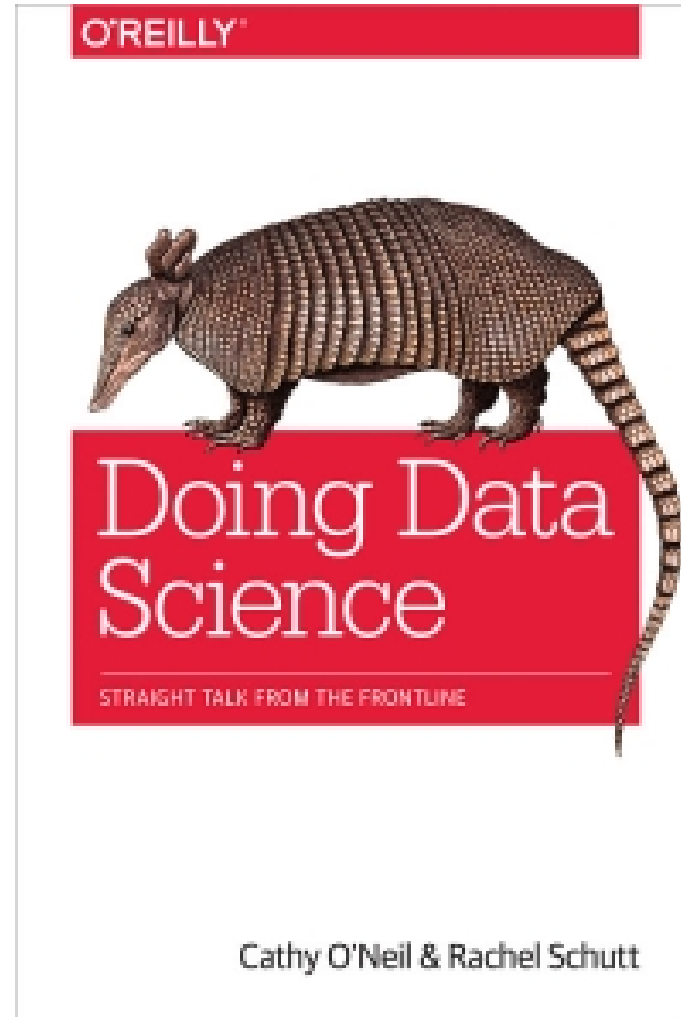
# MapReduce



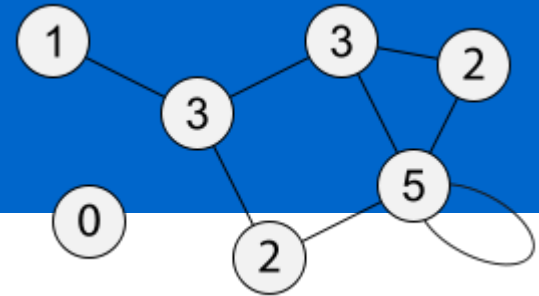
# Open source tools



# Course @ Columbia University



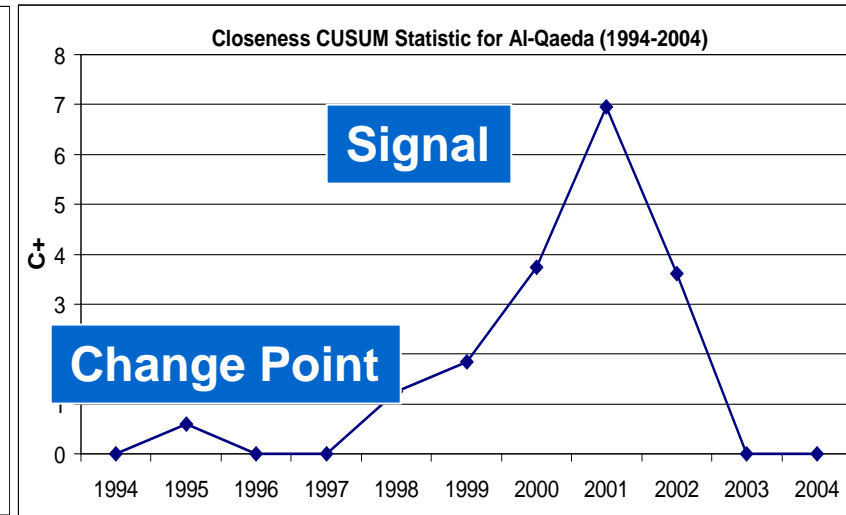
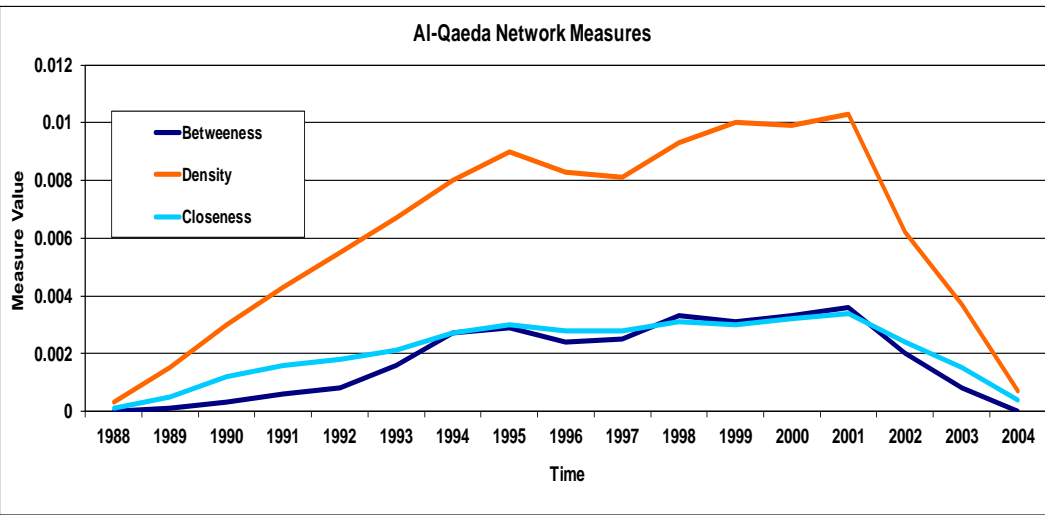
# Topic: Network structure



- 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



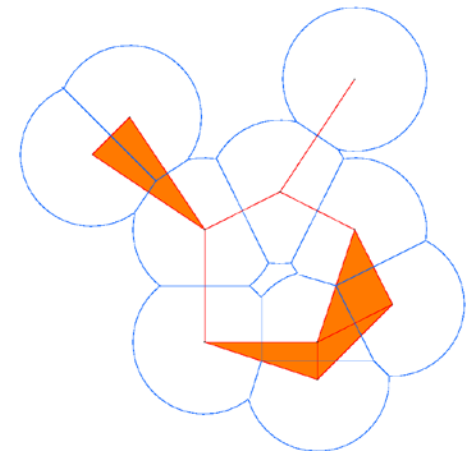
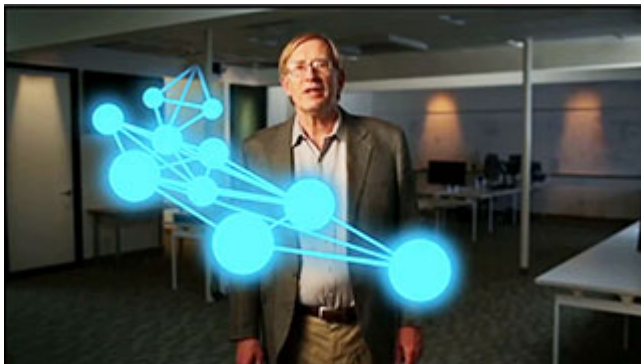
## Challenges:

- 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**

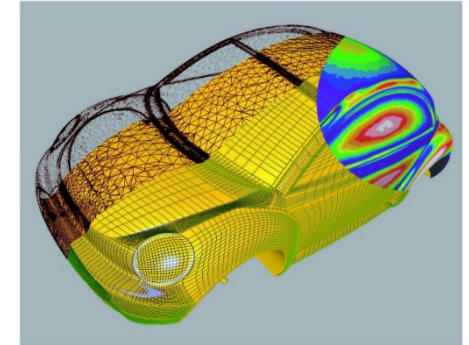
**Combination of machine learning with topological tools (simplices, cohomology) yields new algorithms for clustering**



# 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**

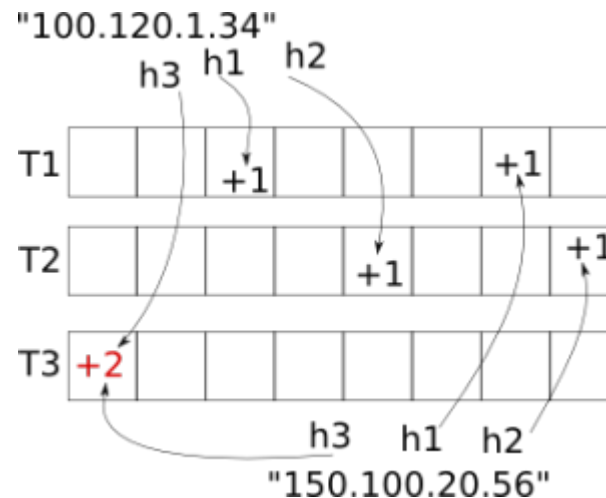


# Topic: Streaming algorithms

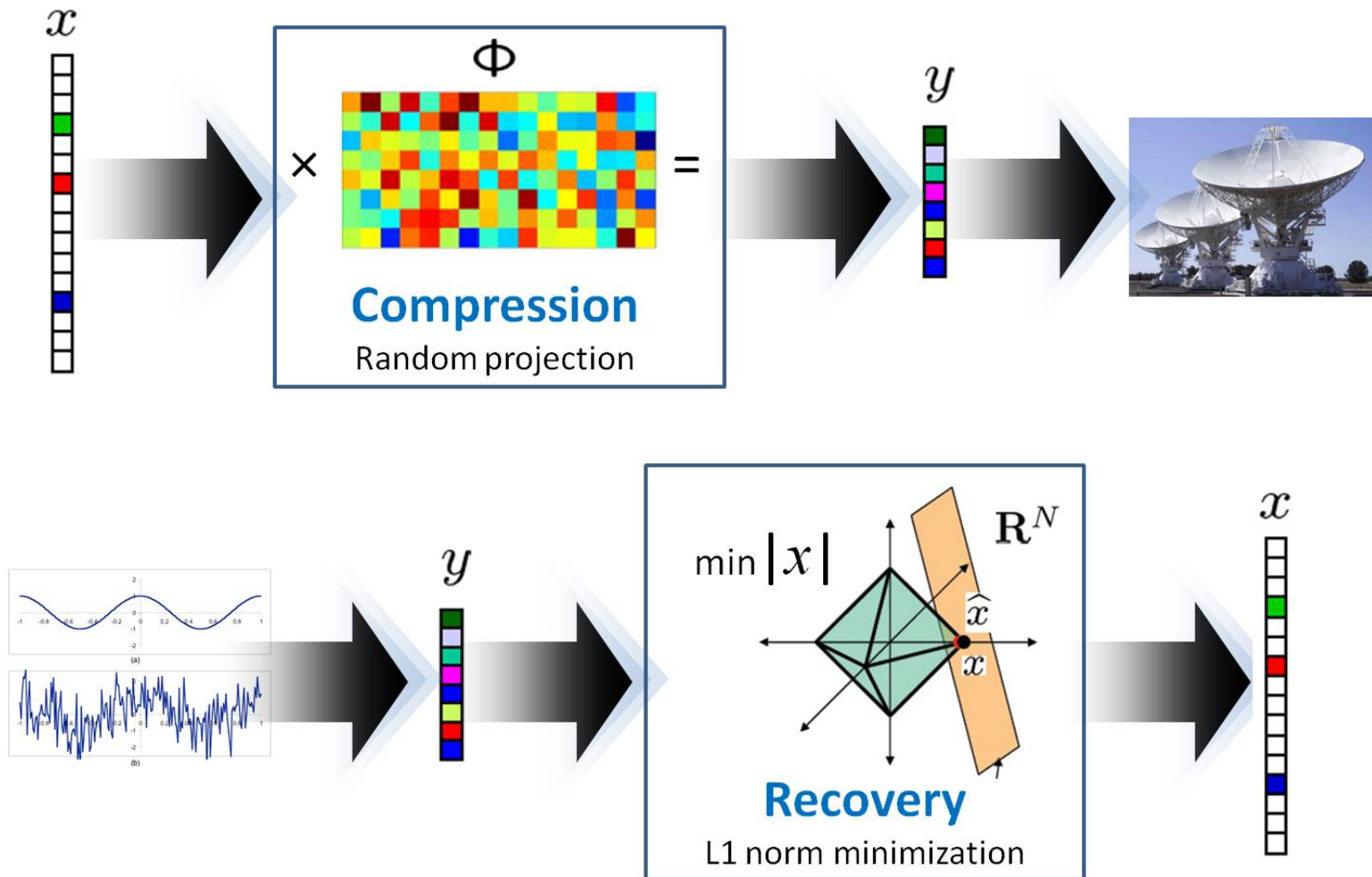
High volume high velocity data makes exact counting of frequencies or number of different items impossible but approximate answers suffice in big data.

New ideas using probabilistic means (random hash functions):

- Count-Min Sketch
- MinHash

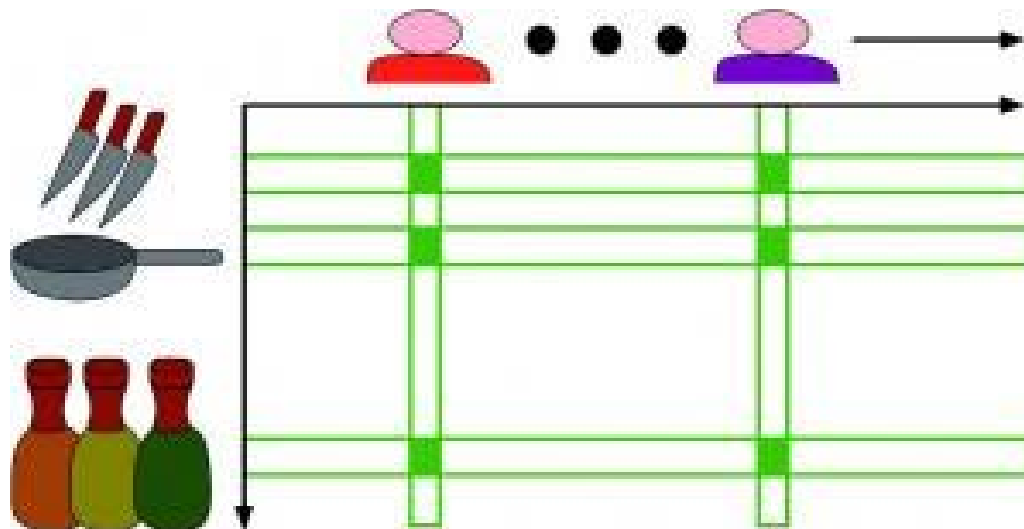


# Topic: Compressed Sensing



# Topic: Recommendation systems

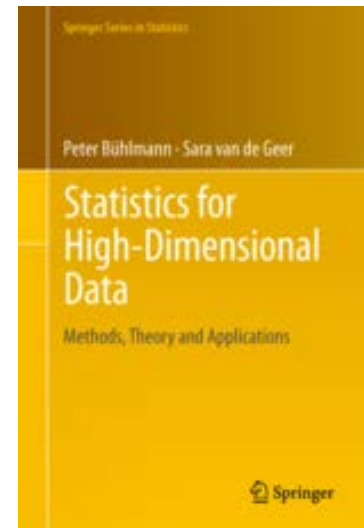
**Approach (Moitra): Use connection to existence theorems for polynomial solutions to algebraic equations and develop scalable algorithms for nonnegative matrix factorizations**



# 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**

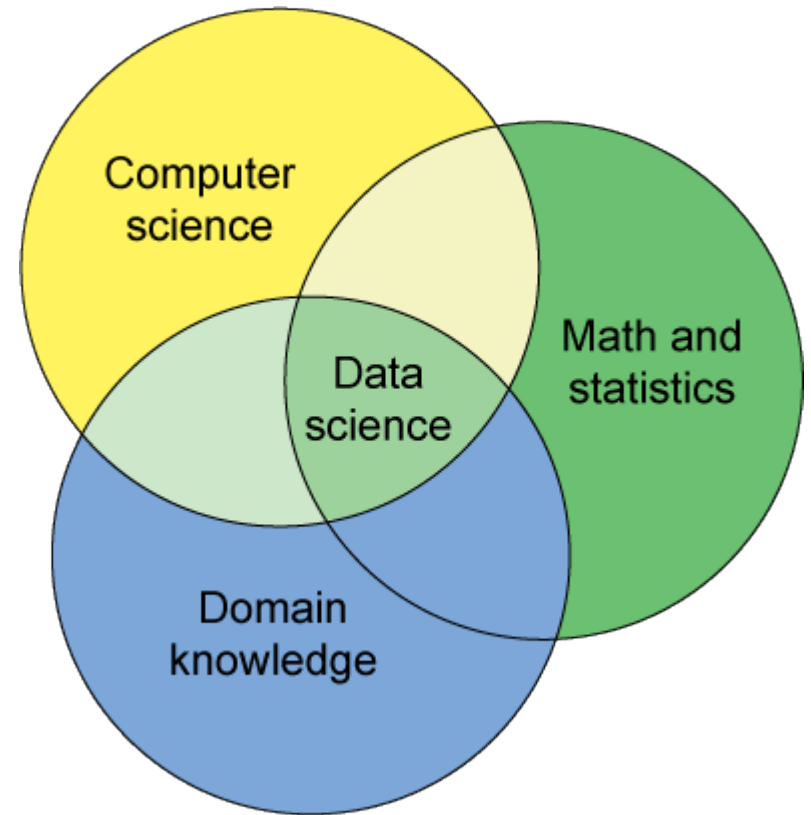
**Digital Sciences**  
**High Performance Computing**



# 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”*

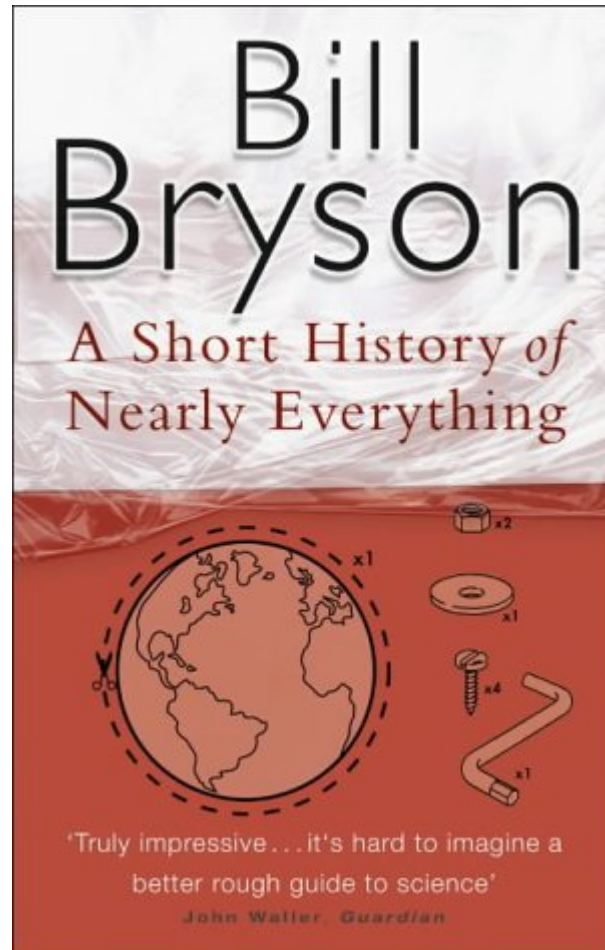


# Journals





# 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**