Data Science Center Eindhoven

The Mathematics Behind Big Data

Alessandro Di Bucchianico

4TU AMI SRO Big Data Meeting
Big Data: Mathematics in Action!
November 24, 2017

TU/e Technische Universiteit **Eindhoven** University of Technology

Where innovation starts

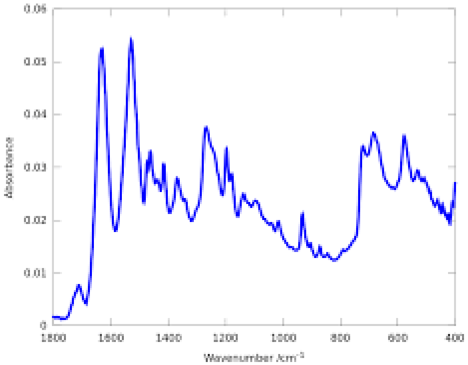
Outline

- “Big Data”
- Some real-life examples with “hidden” mathematics
- Some mathematical developments
- Conclusions

What is big data?

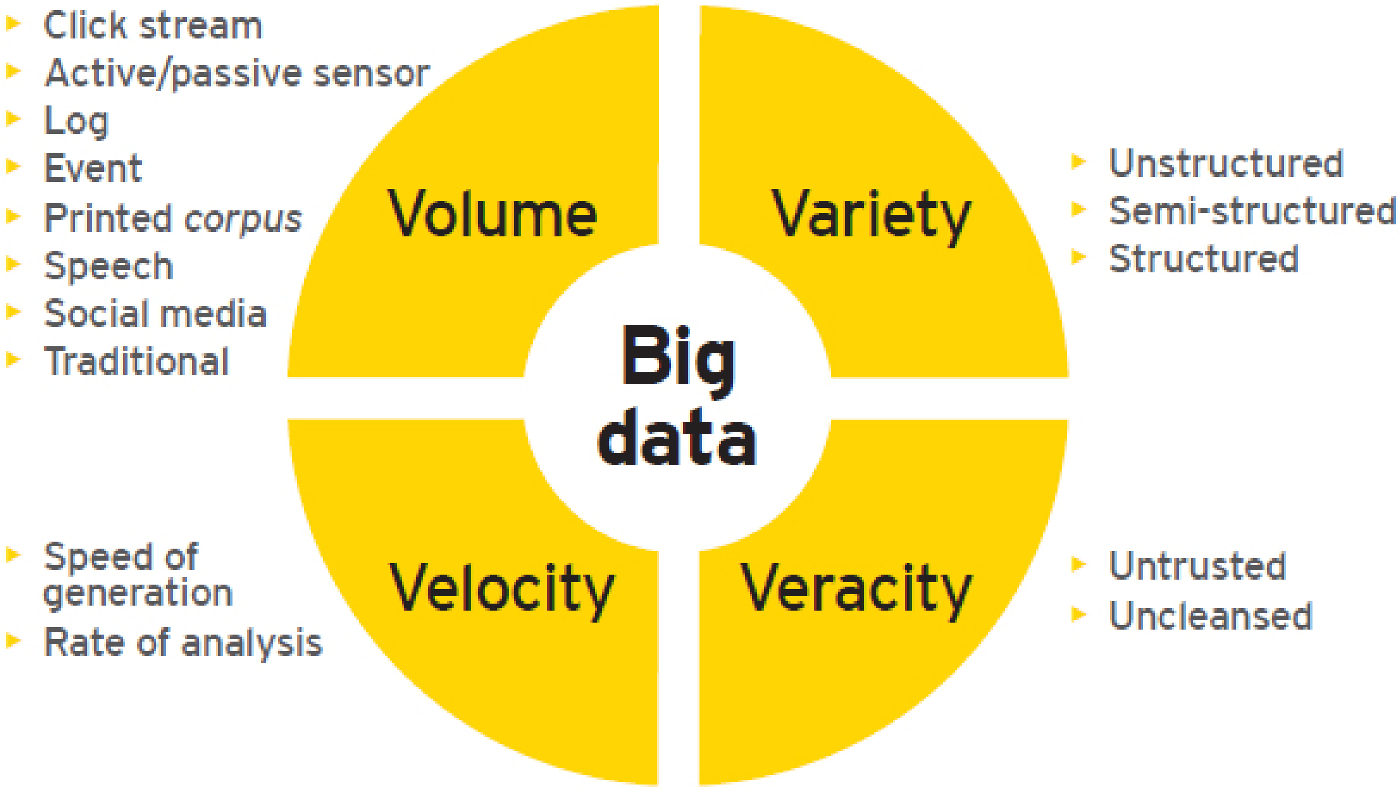
Term coined by John Mashey, chief scientist at Silicon Graphics in the 1990's

…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…

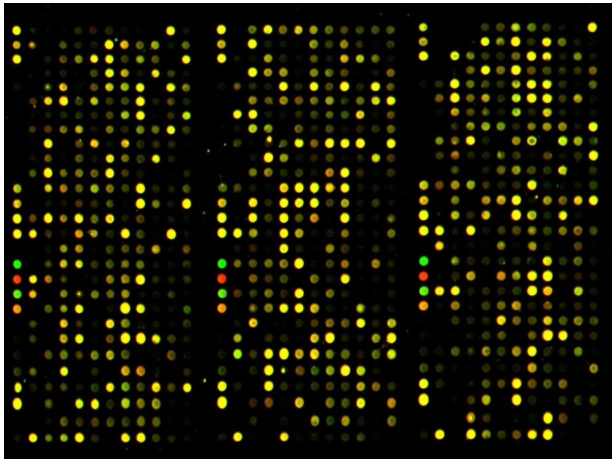


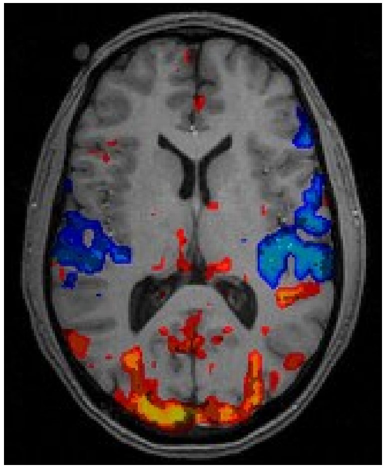
But : chemometrics has a long history of analyzing "large" data sets

Four V's of Big Data

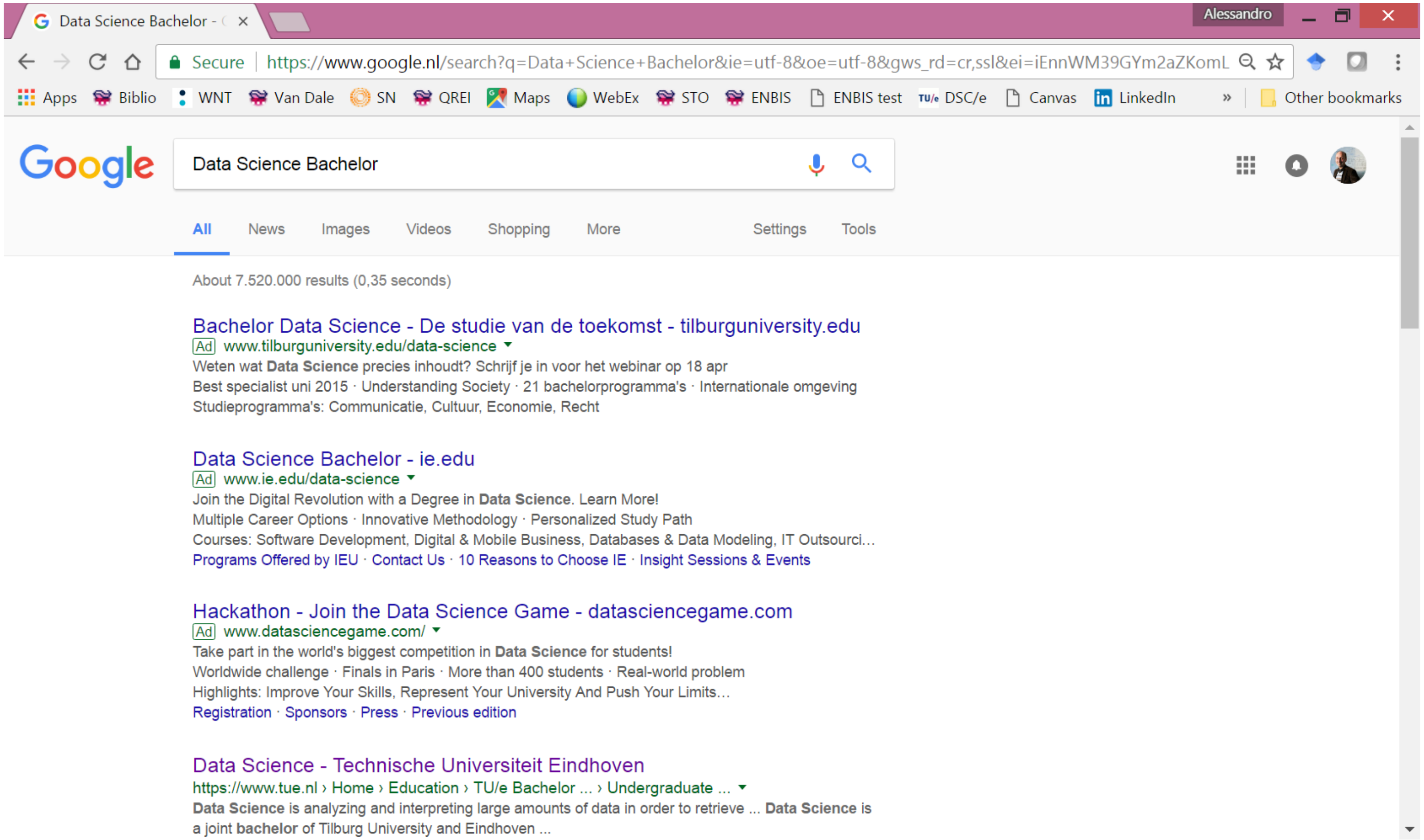


High-dimensional data: “$n \ll p$”





Topic: Google Search



Topic: Google Search

- **How can Google search so fast on its over 250 billion pages?**

- **What are proper ways to rank web pages?**

- **First idea: count words in pages.**

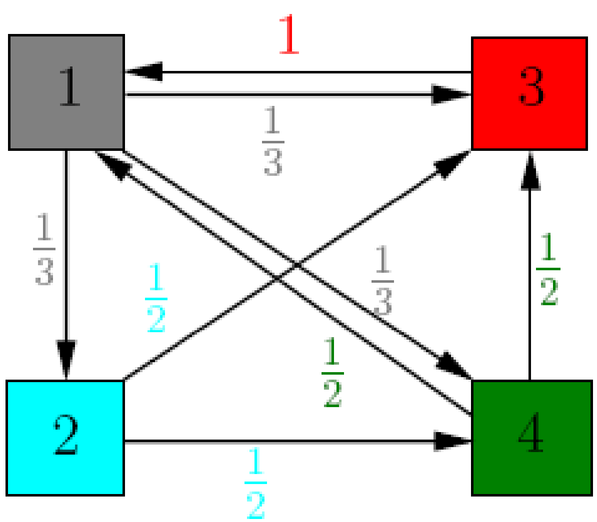
Topic: Google Search

Google PageRank

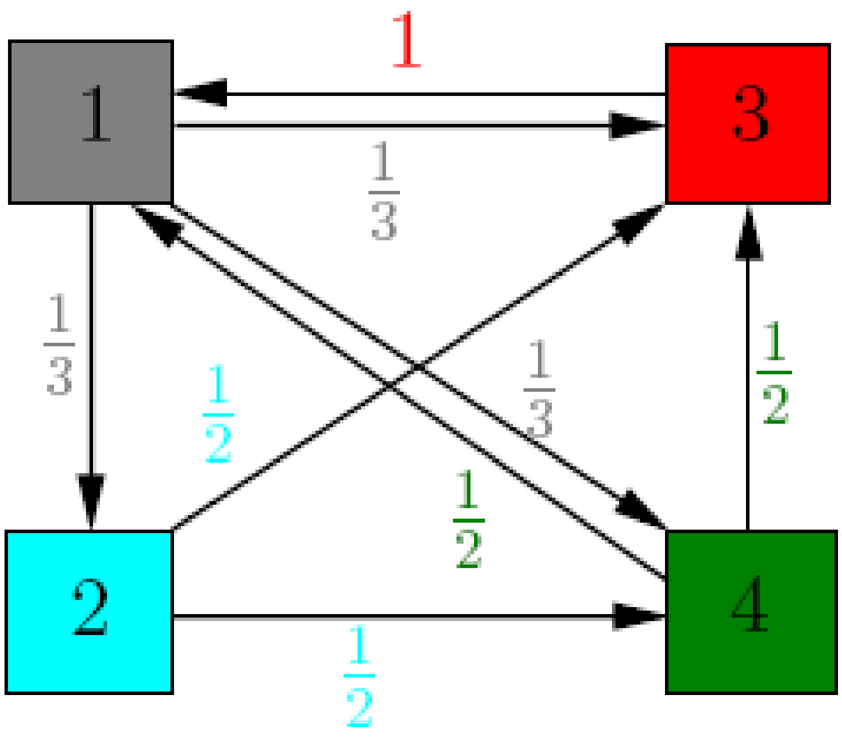
Model the WWW as a graph:

- Squares (nodes, vertices) denote web pages
- Arrows (edges) denote links from one page to another

Relevance is translated into number of incoming links weighted by importance of referring pages

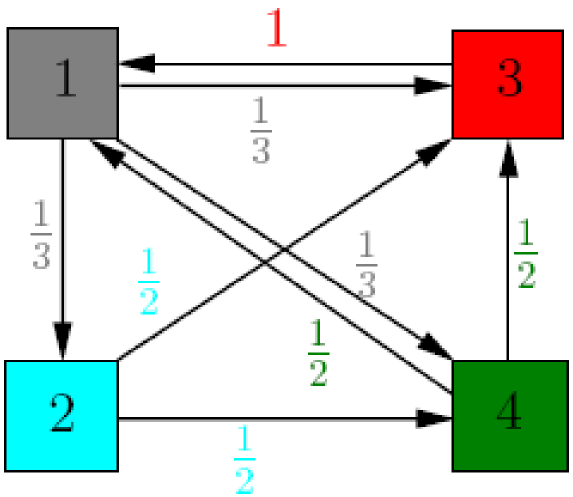


Topic: Google Search



$$
\begin{cases}
x_1 = 1 \cdot x_3 + \frac{1}{2} \cdot x_4 \\
x_2 = \frac{1}{3} \cdot x_1 \\
x_3 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2 + \frac{1}{2} \cdot x_4 \\
x_4 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2
\end{cases}
$$

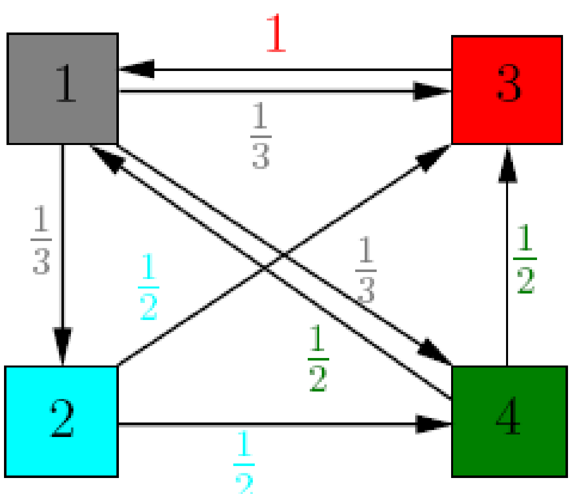
Topic: Google Search



$$A = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

$$\begin{cases} x_1 = 1 \cdot x_3 + \frac{1}{2} \cdot x_4 \\ x_2 = \frac{1}{3} \cdot x_1 \\ x_3 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2 + \frac{1}{2} \cdot x_4 \\ x_4 = \frac{1}{3} \cdot x_1 + \frac{1}{2} \cdot x_2 \end{cases} \Rightarrow A \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}$$

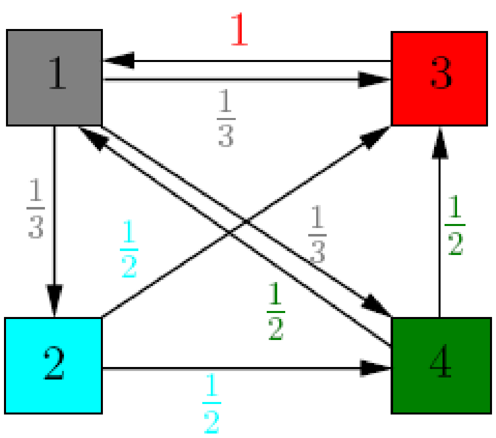
Topic: Google Search



$$\frac{1}{31}\cdot\begin{bmatrix}12\\4\\9\\6\end{bmatrix}\sim\begin{bmatrix}0.38\\0.12\\0.29\\0.19\end{bmatrix}$$

Equivalent: Random Surfer Model

- Start at any point on the graph
- Assign equal probabilities to all outgoing links
- Choose new node with probabilities given by the links
- Relevance is proportion of visits after surfing for a very long time (“stationary distribution”)



Google Matrix

- **Choose a “damping factor” *p* (0<*p*<1)**
- **Google’s *p* is secret but around 0.15**

$$M = (1-p)A + pB$$

$$B = \frac{1}{n}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

Interpretation for random surfer:

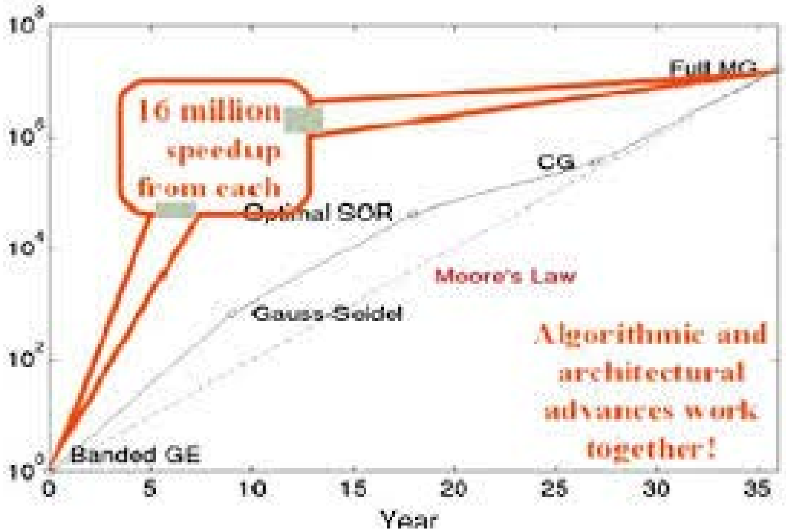
- **choose next link according to A with probability 1- p**
- **“teleport” to random page with probability p**

Choice of Matrix *B*

Google choose other *B* to avoid unwanted boosting of pages

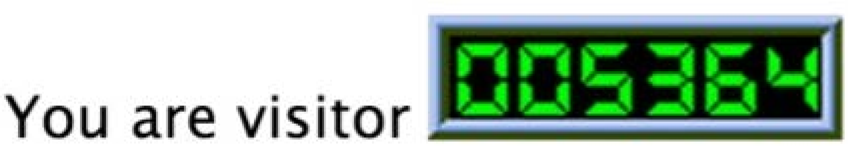
Practical Issues

- The real Google matrix has size in the order of 30 billion columns x 30 billion rows
- Every day around 250,000 new pages are added
- Matrix has many zeros (100 links to a page) which speeds up calculations
- Doing the matrix calculation requires fast computers but also advanced math !



Topic: Streaming algorithms

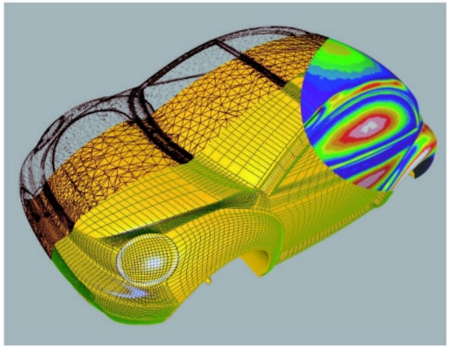
High volume, high velocity data makes *exact* counting of frequencies or number of *different* items practically infeasible but *approximate* answers suffice in several applications (web site counters, customer counts in retail transactions,…).



New algorithms using advanced probabilistic methods (random hash functions):

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

Topic: Uncertainty Quantification



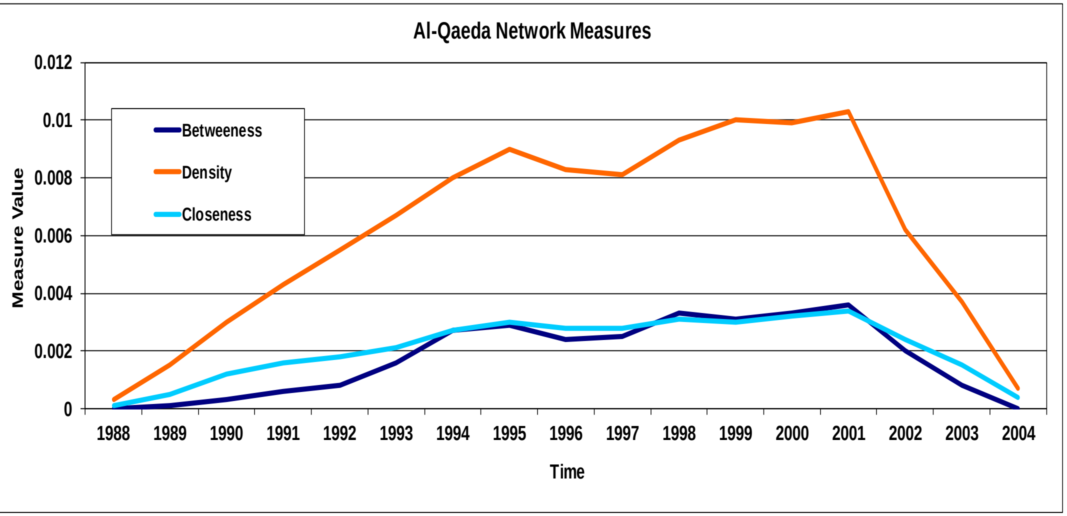
Virtual prototyping using mathematical models. UQ does not deal with

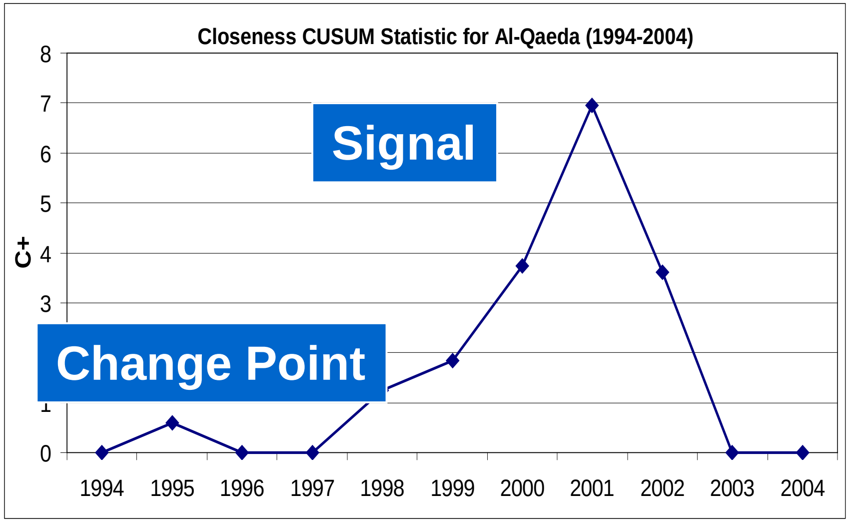
1. unknown uncertainty in the initial conditions of parameters
2. parametrisation of design building blocks / dimension reduction

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

For 2): Model Order Reduction

Topic : Network monitoring

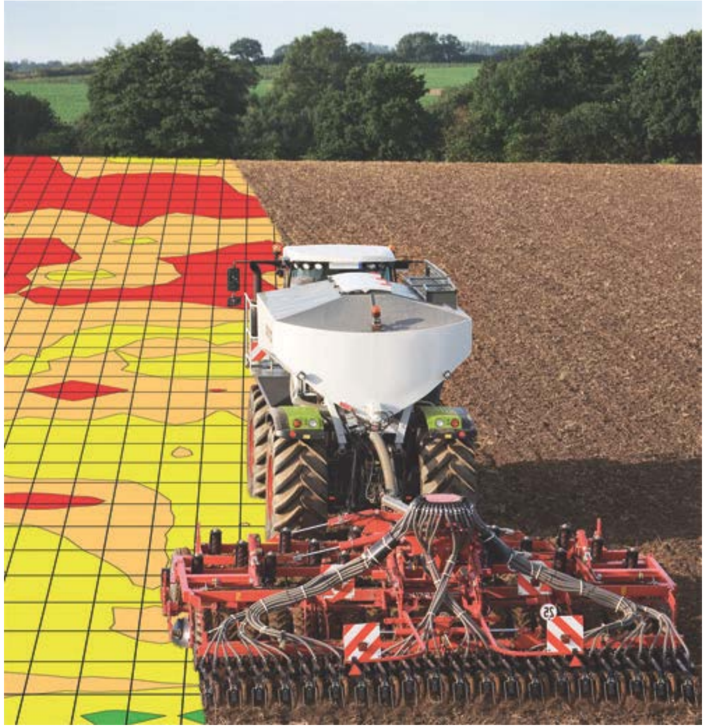




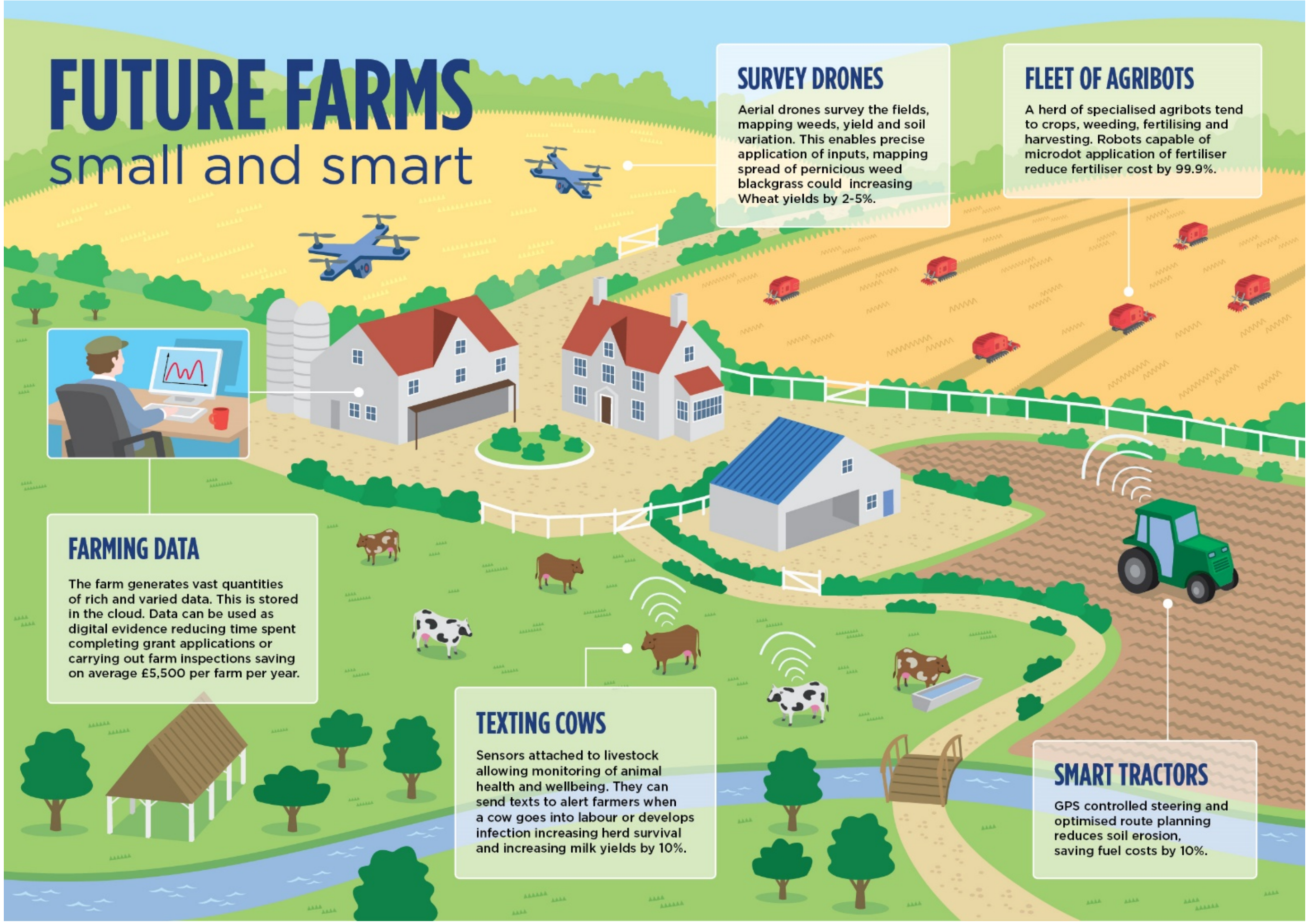
Challenges:

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

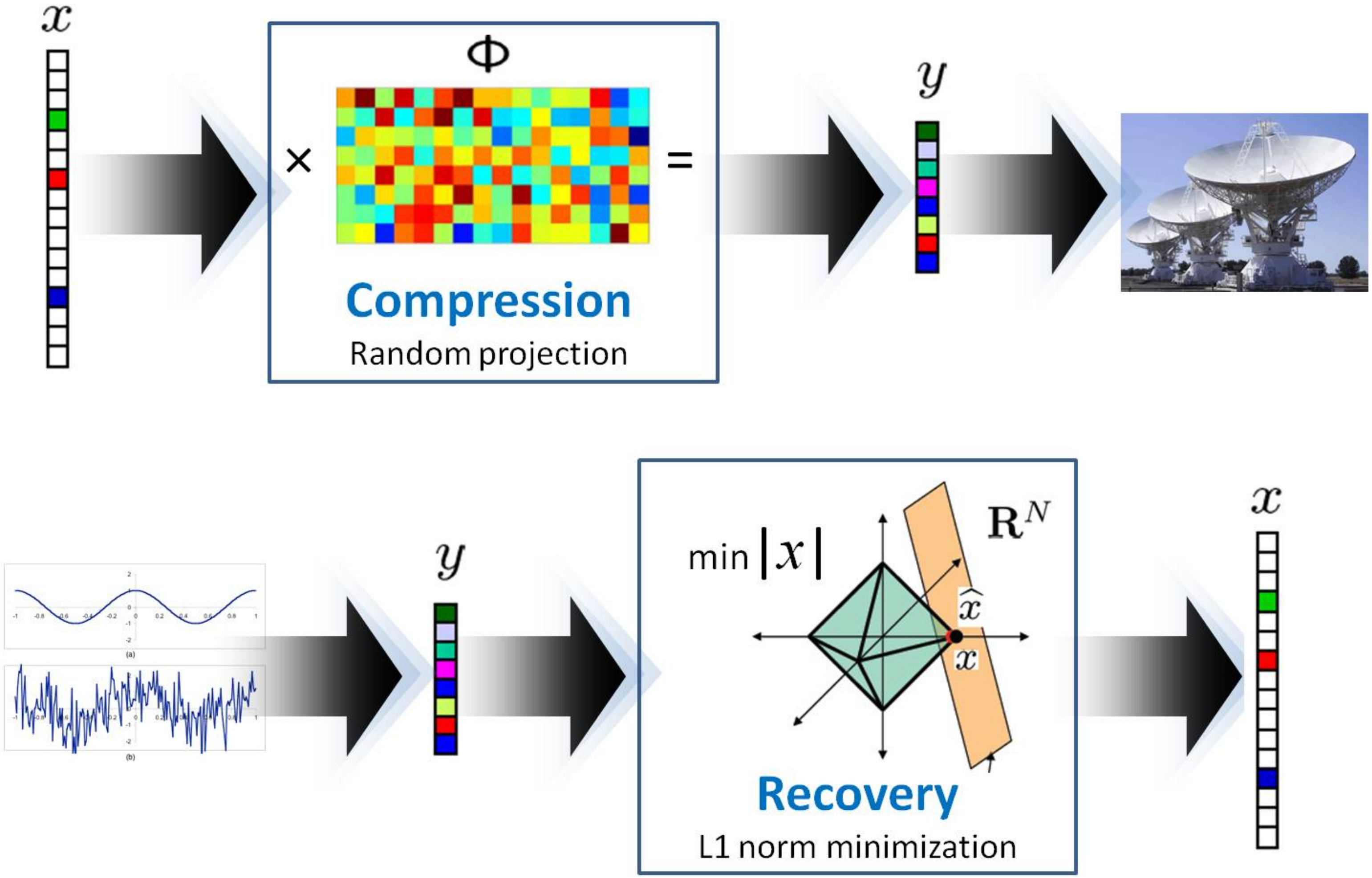
Topic: Precision farming



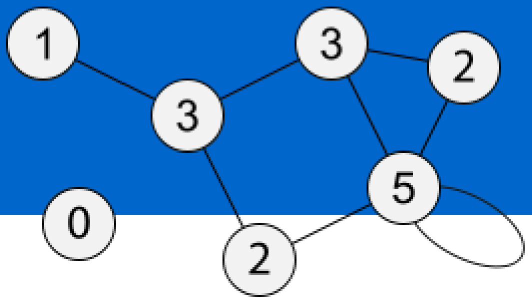
Topic: Precision farming



Topic: Compressed Sensing



Topic: Network structure



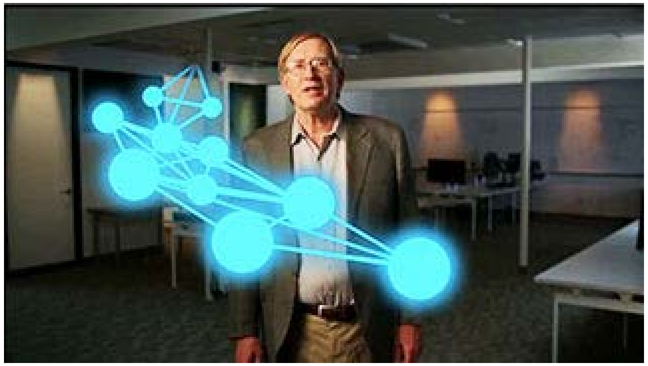
- **dependencies between nodes in networks measured by degrees of direct neighbours**
- **assortativity coefficient of Newman is nothing but Pearson's correlation statistic**
- **inconsistent estimator when variances are infinite**
- **Spearman rank correlation behaves better but calculation is computationally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

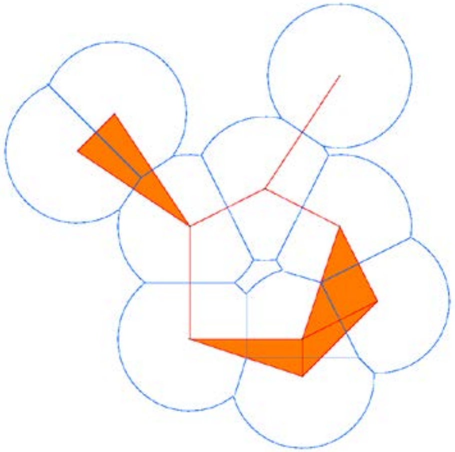
Van der Hofstad, R. and Litvak, N. (2014) *Degree-Degree Dependencies in Random Graphs with Heavy-Tailed Degrees.* Internet Mathematics, 10 (3-4). pp. 287-334

Topic: Topological Data Analysis

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

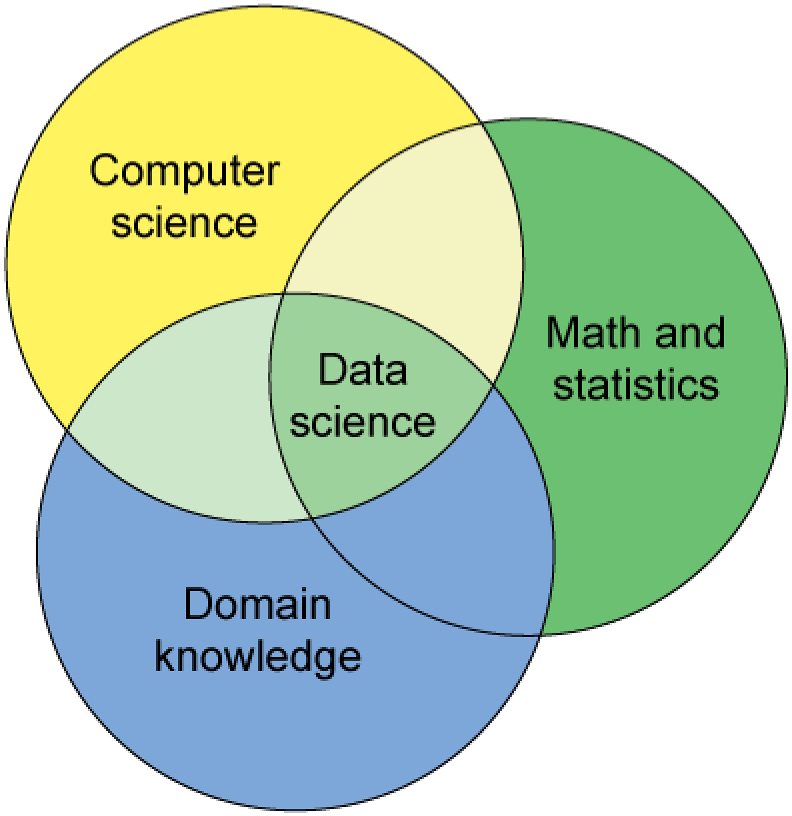
Combination of machine learning with topological tools (simplices, cohomology) yields new algorithms for finding patterns and clustering





Mathematical contributions in general

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



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

Conclusions

1. Mathematical methods are an important enabler in big data settings

2. Developments in big data settings not only require more computer speed and memory capacity, but also new advanced mathematics