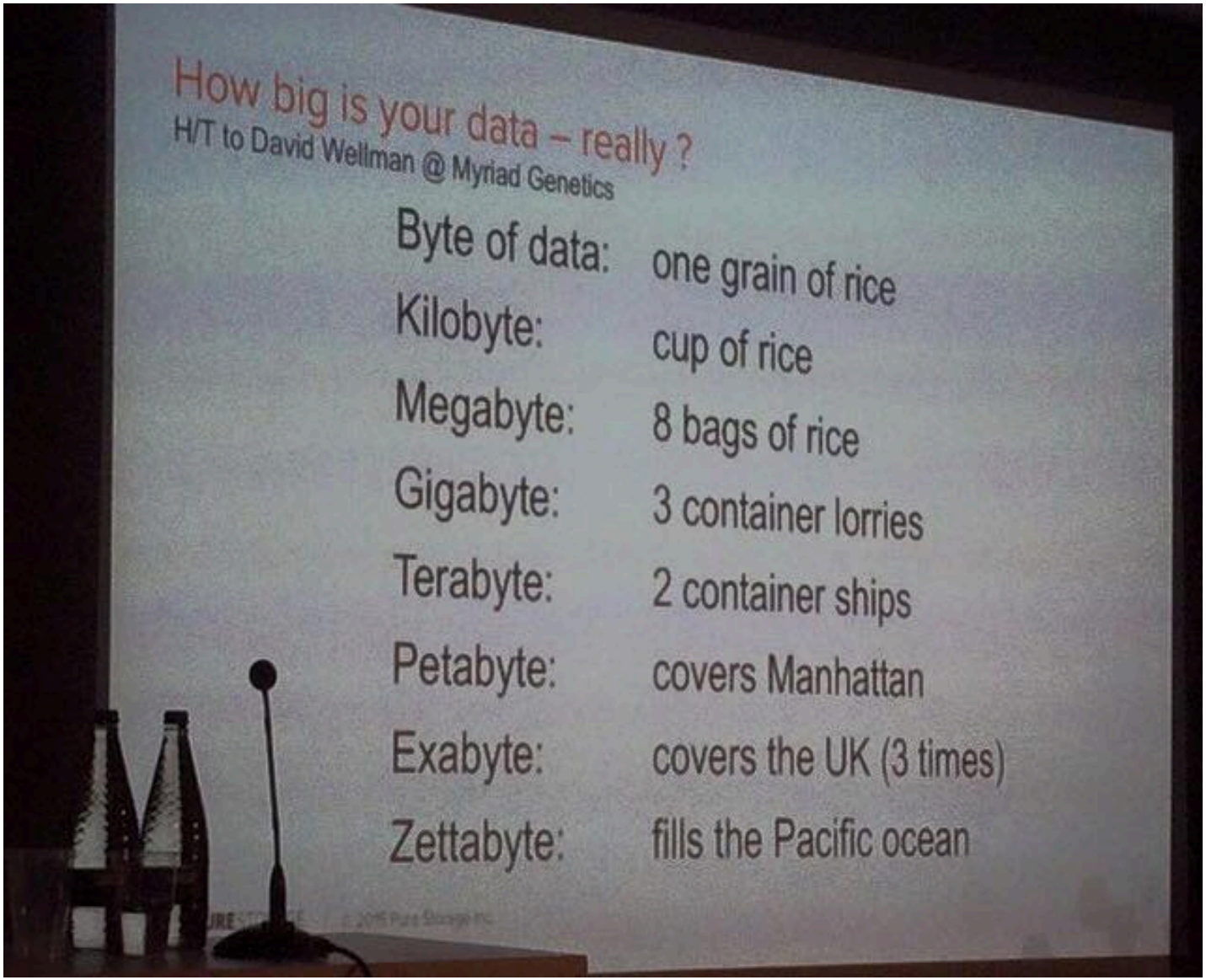
Big Data Analytics

Genoveva Vargas-Solar
http://www.vargas-solar.com/big-data-analytics
French Council of Scientific Research, LIG & LAFMIA Labs

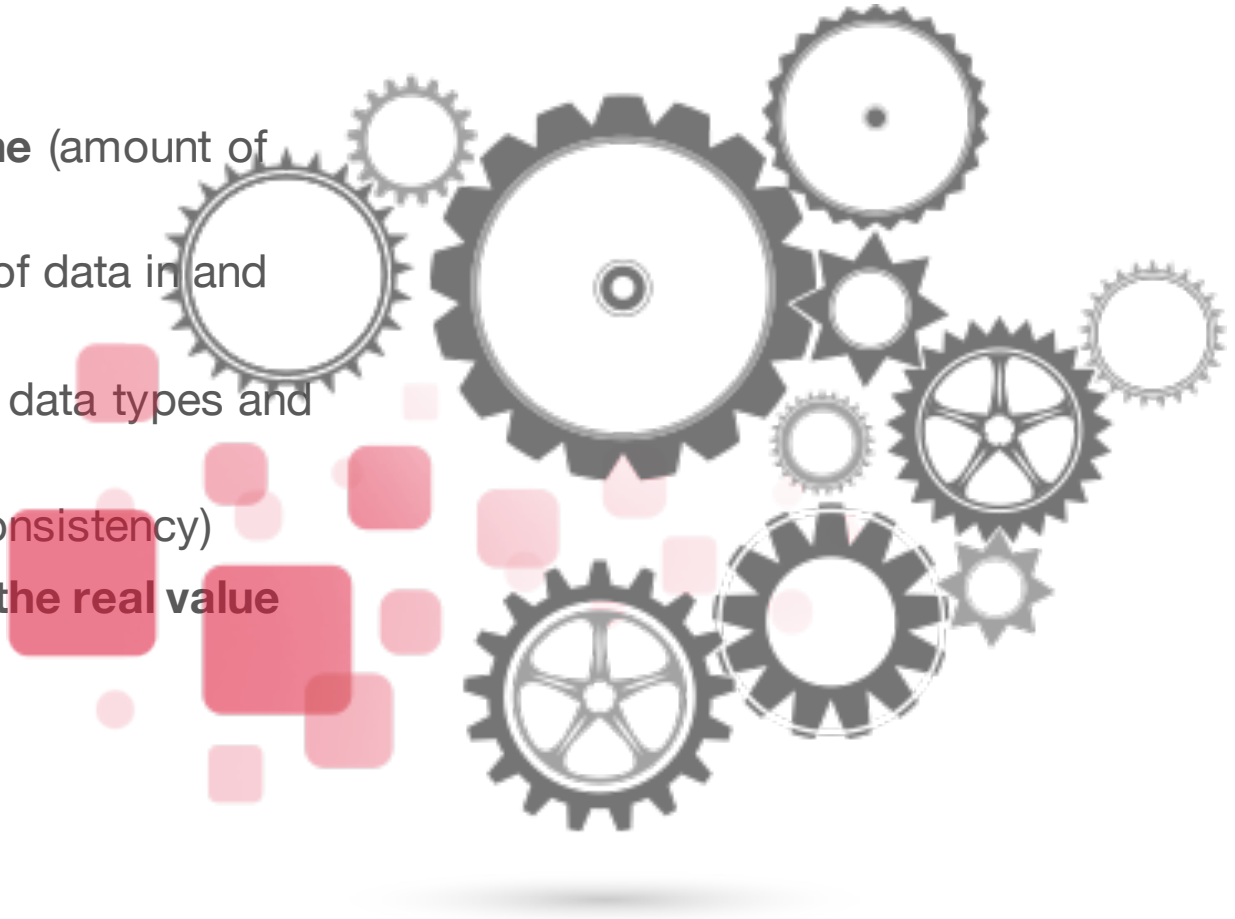
Montevideo, 22nd November – 4th December, 2015


How big is your data – really ?
H/T to David Wellman @ Myriad Genetics
Byte of data: one grain of rice
Kilobyte: cup of rice
Megabyte: 8 bags of rice
Gigabyte: 3 container lorries
Terabyte: 2 container ships
Petabyte: covers Manhattan
Exabyte: covers the UK (3 times)
Zettabyte: fills the Pacific ocean

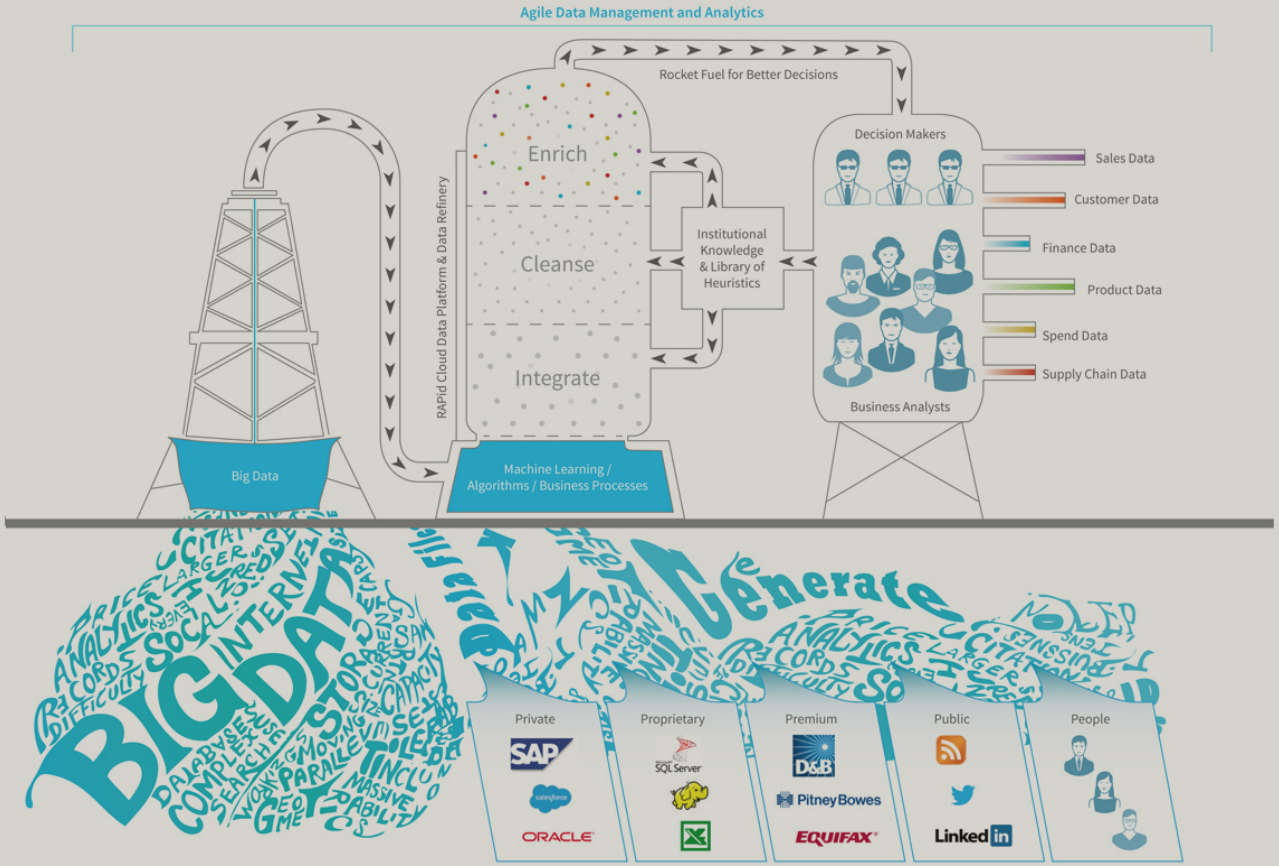
Collection of data sets **so large** and **complex** *that it becomes* **difficult** to **process** *using* **on-hand database management** *tools or* **traditional** *data* **processing** applications

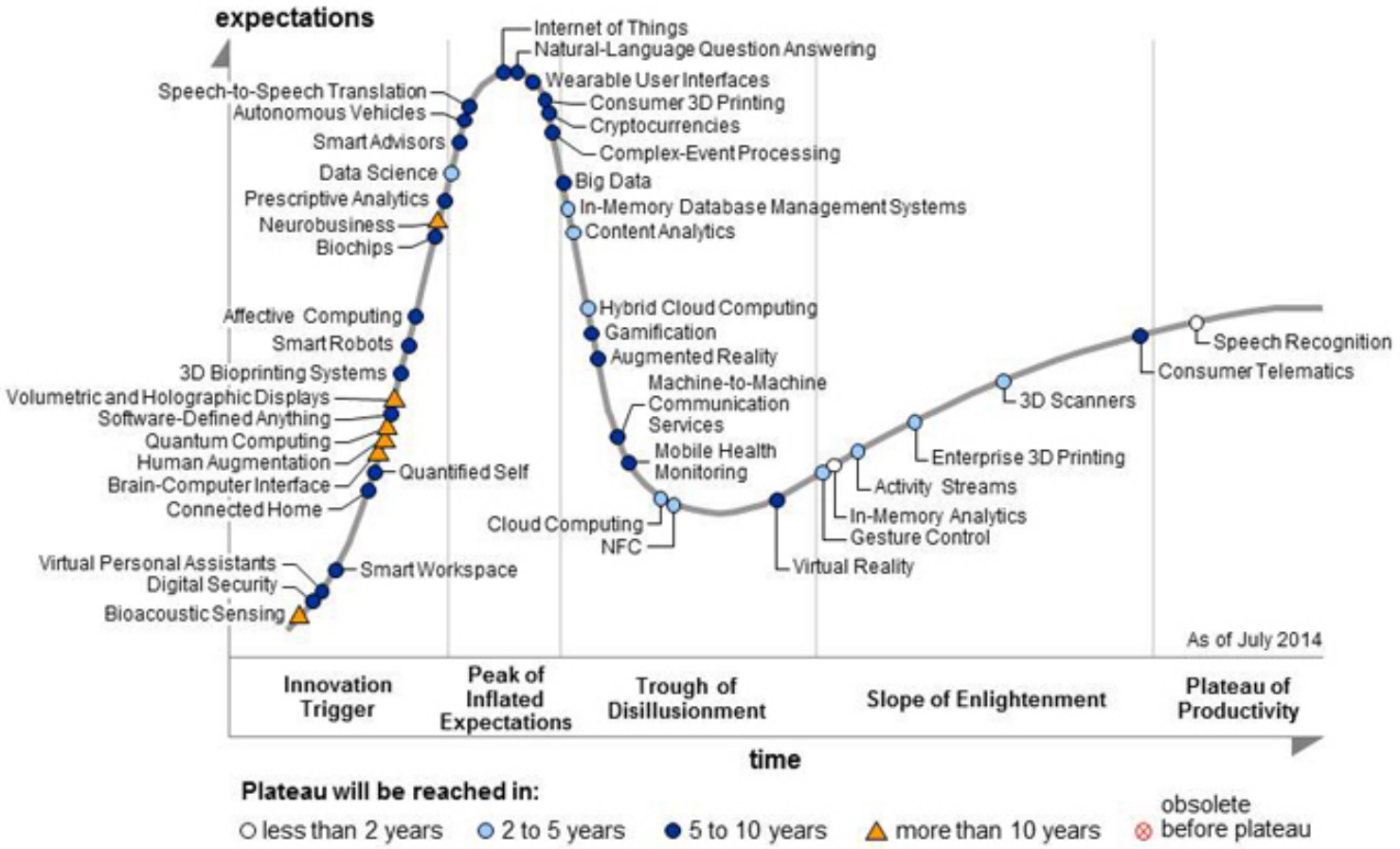
The V's & the needs of Big Data

- increasing **volume** (amount of data)
- **Velocity** (speed of data in and out)
- **Variety** (range of data types and sources)
- **Veracity** (data consistency)
- **Value (which is the real value of data?)**



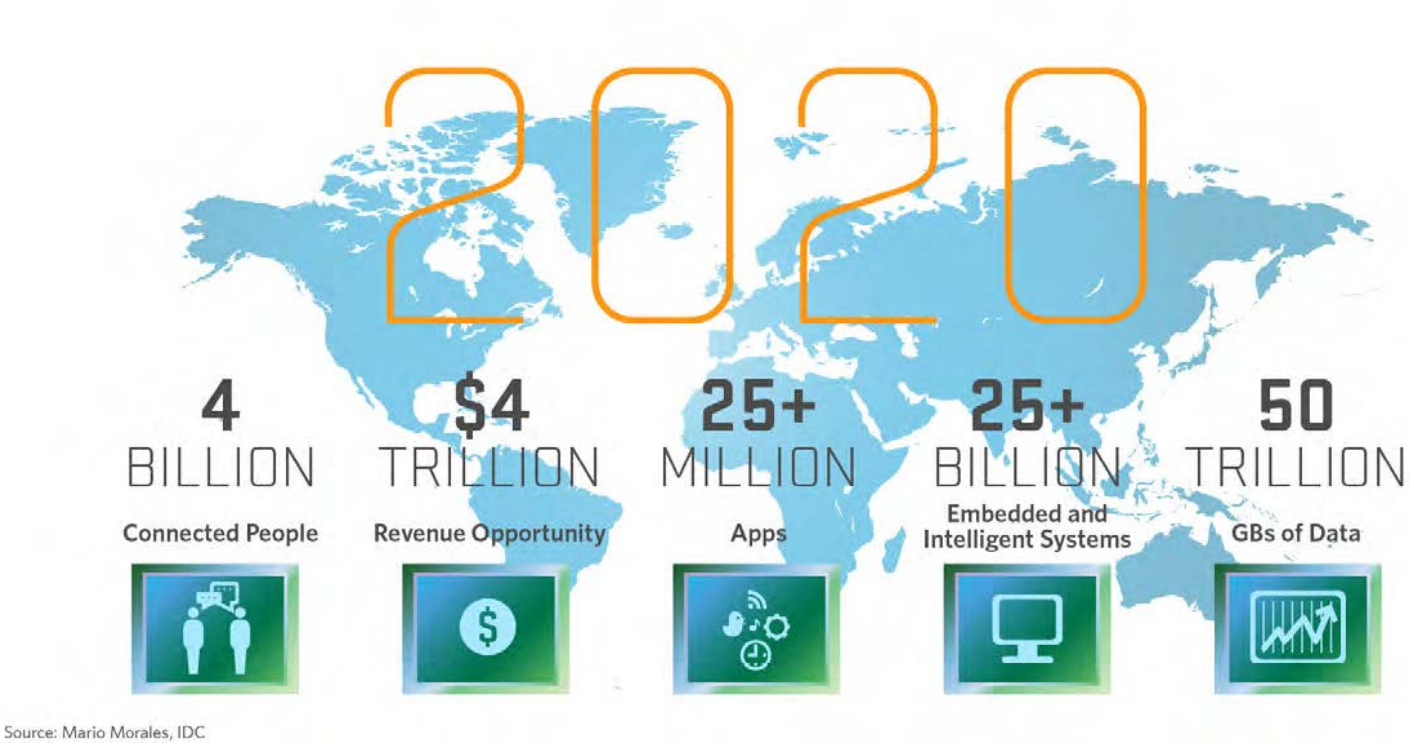
Big Data processing at glance





http://www.gartner.com/newsroom/id/2819918

Internet of Things



Source: Mario Morales, IDC

Big Data at a bronto scale

1 bit	Binary digit
8 bits	1 byte
1000 Terabytes	1 Petabyte
1000 Petabytes	1Exabyte
1000 Exabyte	1 Zettabyte
1000 Zettabytes	1 Yottabyte

We will no longer have the luxury of dealing with just "big" data

http://spectrum.ieee.org/computing/software/beyond-just-big-data

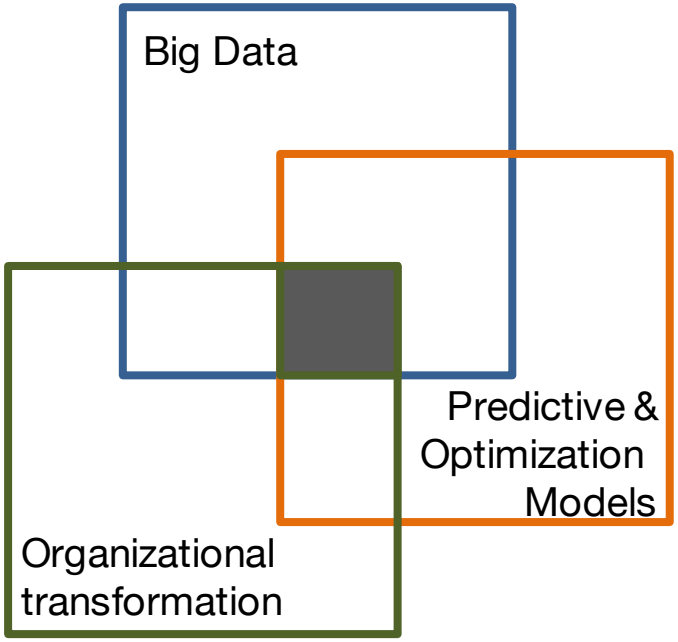
New types of huge data collections

- **Thick data**: combines both quantitative and qualitative analysis,
- **Long data**: extends back in time hundreds or thousands of years
- **Hot data**: used constantly, meaning it must be easily and quickly accessible
- **Cold data**: used relatively infrequently, so it can be less readily available

http://spectrum.ieee.org/computing/software/beyond-just-big-data

What about analytics ?

Capturing value from advanced analytics



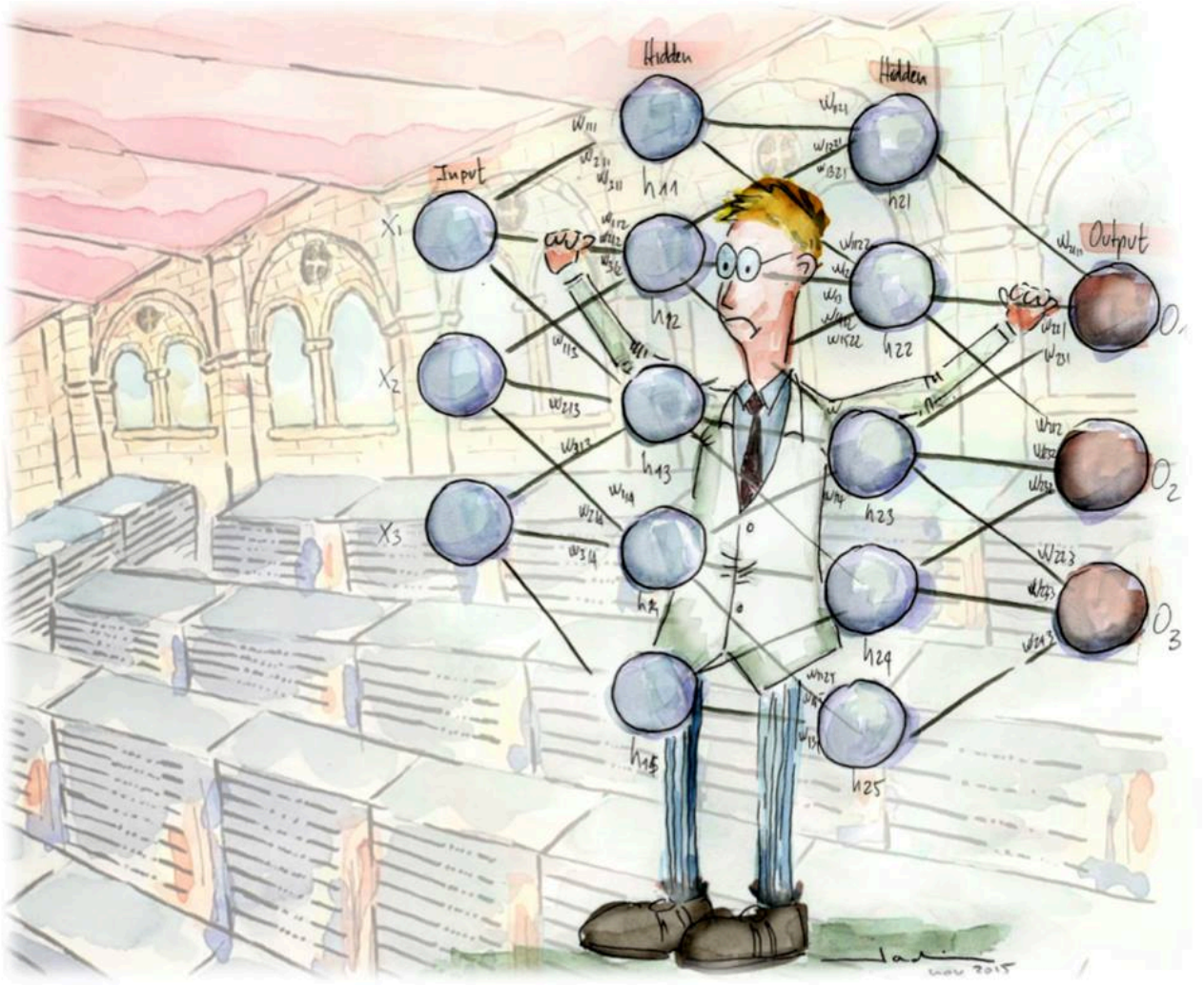
Based on three guiding principles

- Decision backwards
- Step by step
- Test and learn

Data was not stored

Beginning of the use of BDs & basic reports

Great variety of visual resources to analyse data

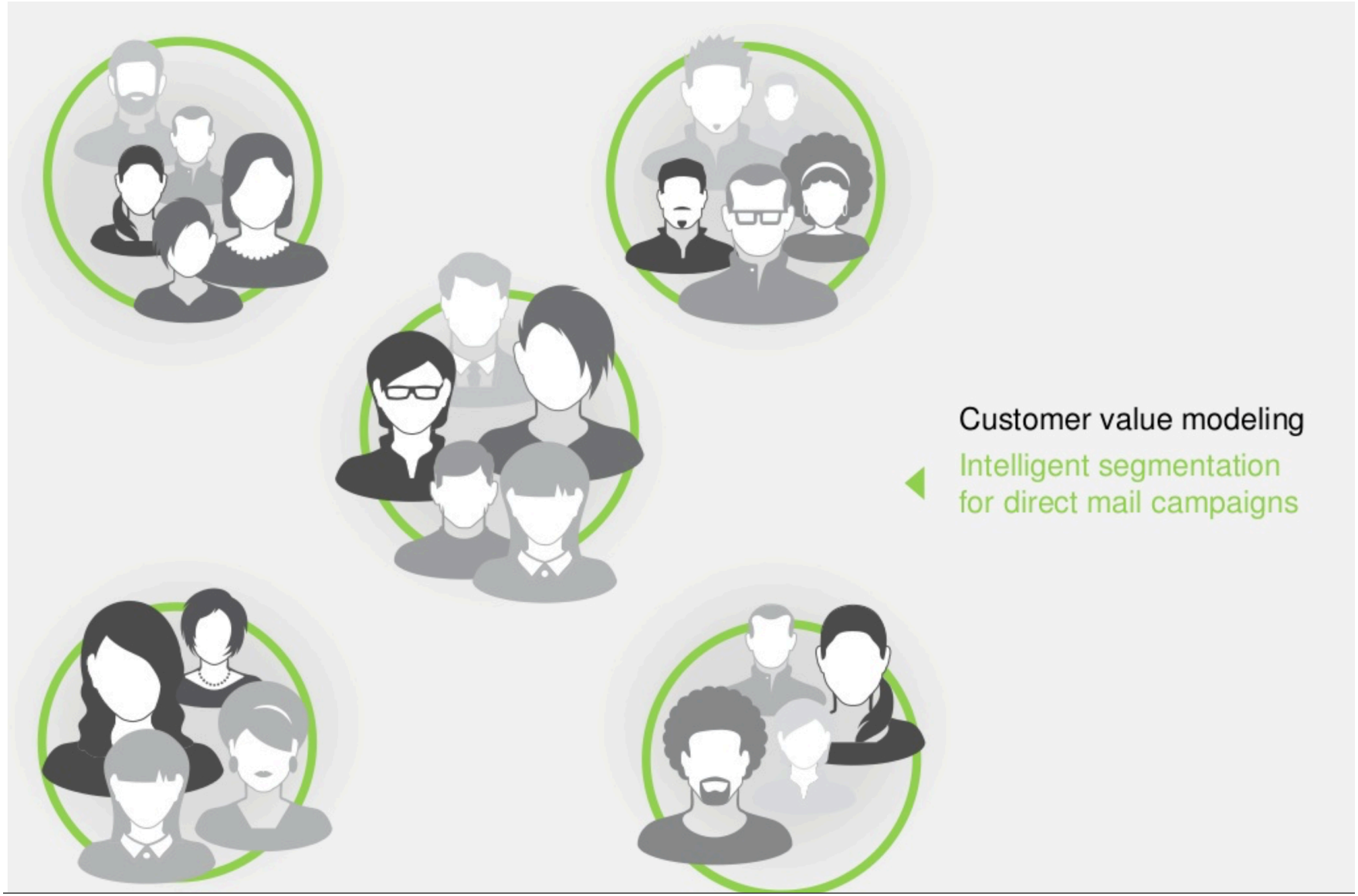


Example of High-Performance Big-Data Analytics research at BSC: One of our senior researchers is trying to deploy a neural network model into our supercomputer Marenostrum In Barcelona.

Direct mailing campaign

The traditional approach

Direct mailing based on generic segmentation



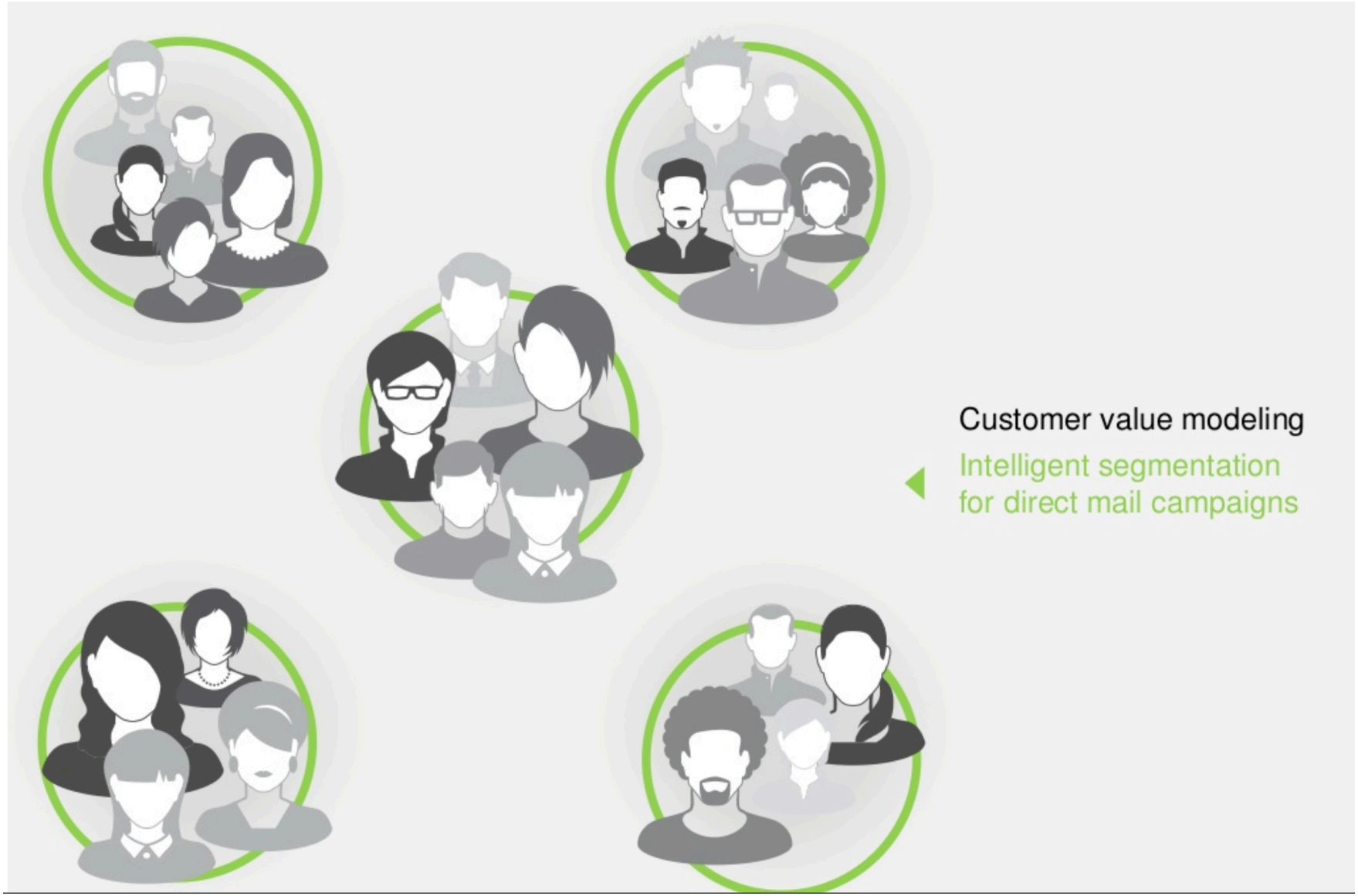
Customer value modeling

Intelligent segmentation
for direct mail campaigns

Regression + clustering based on transaction history

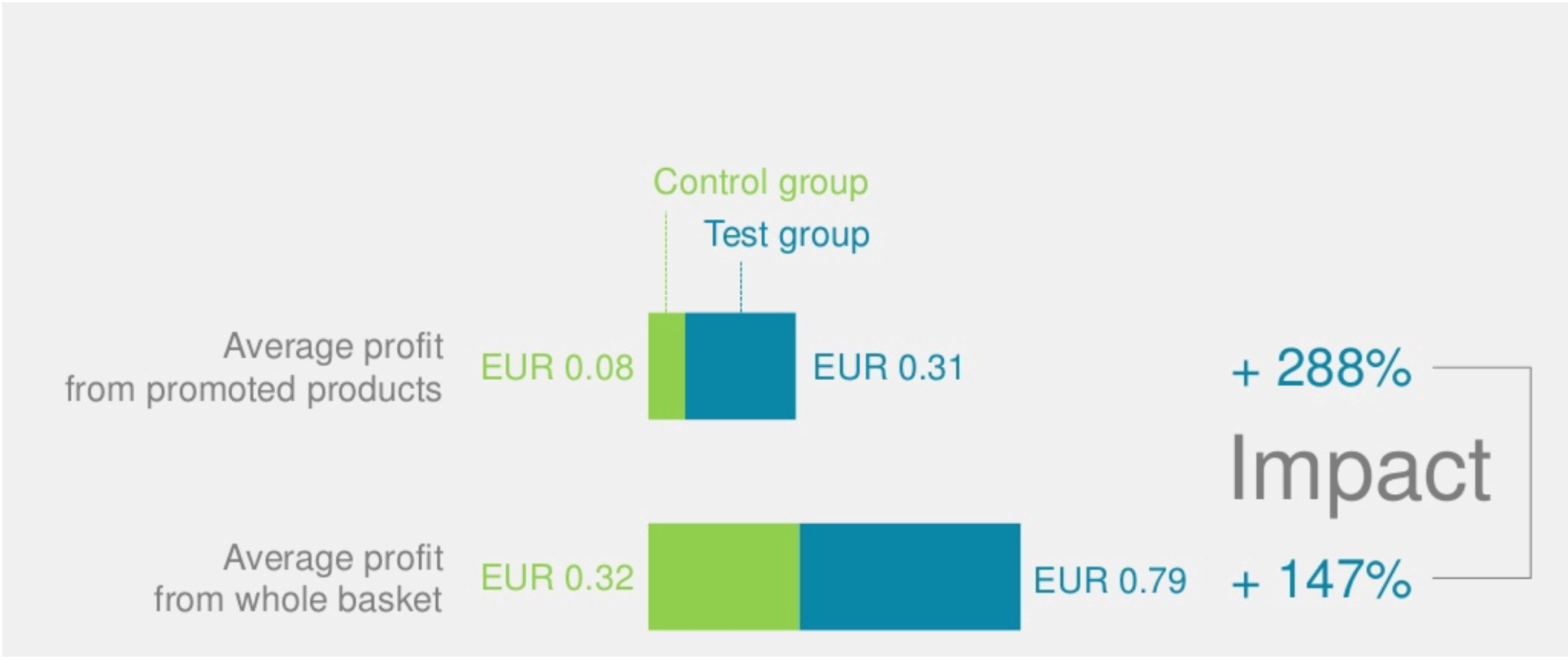
Predictive modeling based on customer preferences

> 100 campaigns per year
> 2 million customers
> 10 million transactions

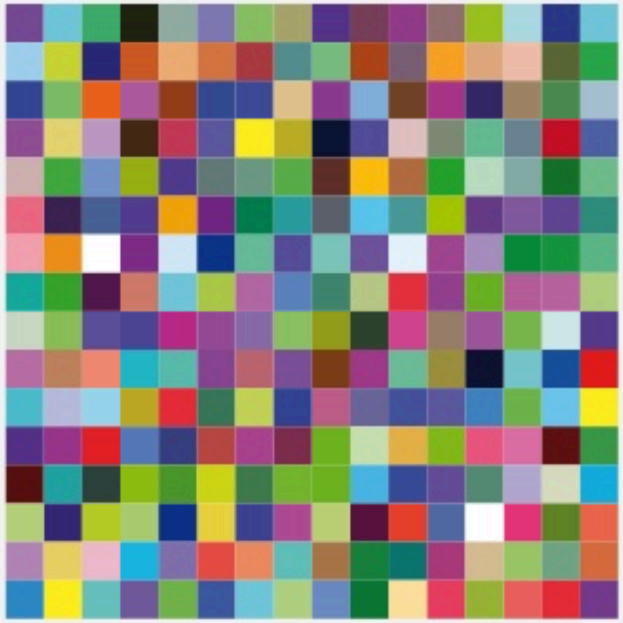


Customer value modeling

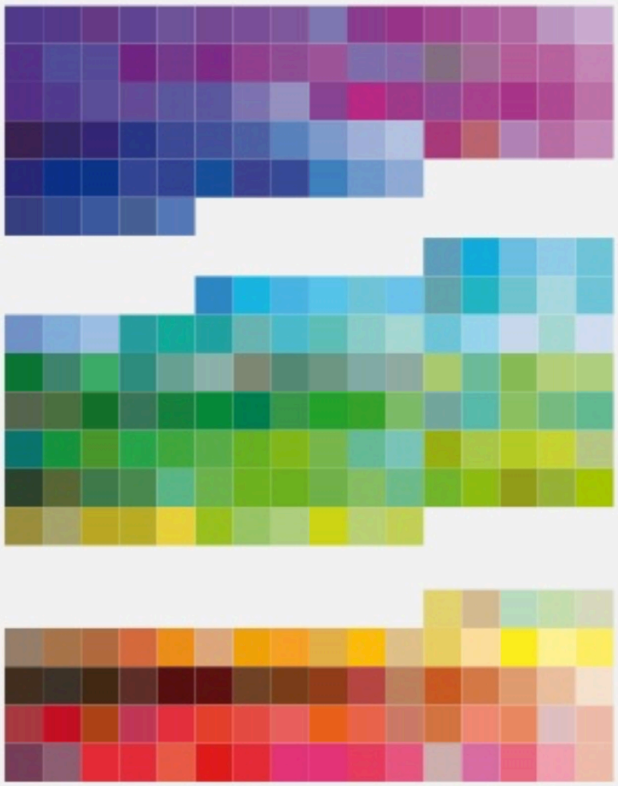
Intelligent segmentation
for direct mail campaigns


Control group
Test group
Average profit from promoted products
EUR 0.08
EUR 0.31
+ 288%
Impact
Average profit from whole basket
EUR 0.32
EUR 0.79
+ 147%

Next product to buy



80 million consumers
100 million transactions



Multivariate statistics
Association rule analysis

"infrequent shoppers"

Are being offered the most probable recommendation

"frequent shoppers"

Get recommendations to generate maximum margin

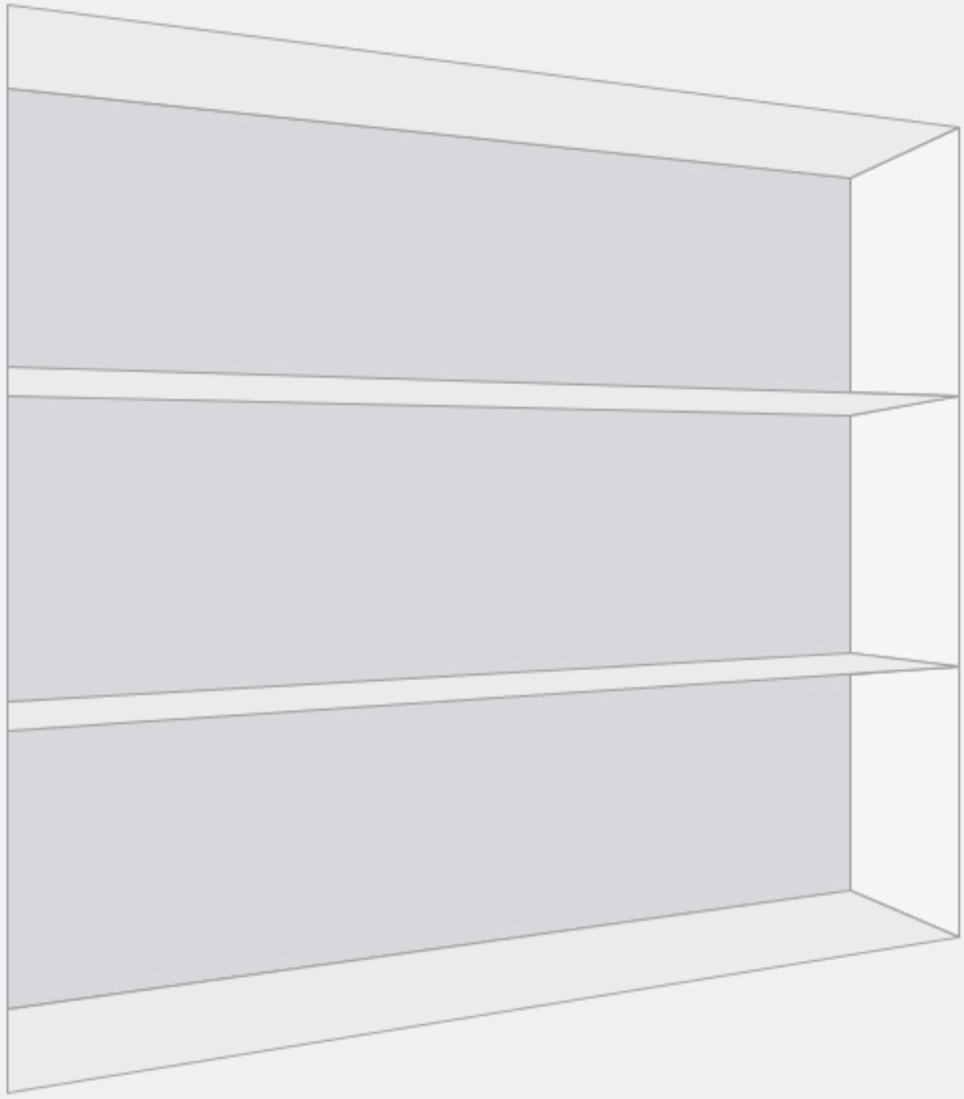
"site lovers"

Receive recommendations from other categories to broaden their purchase behavior

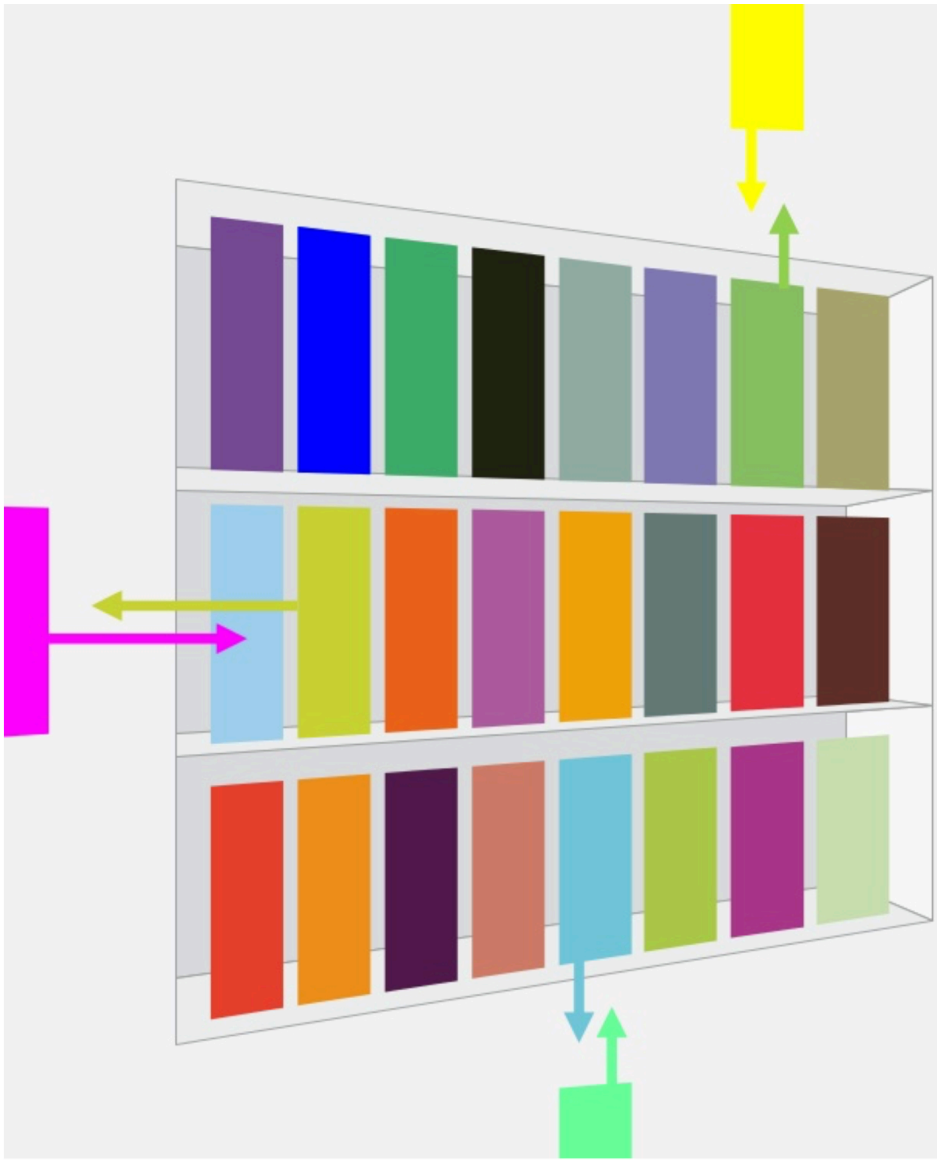
USD 1 billion identified Impact

USD 300 million already realized within 6 months

Assortment optimization



The traditional approach



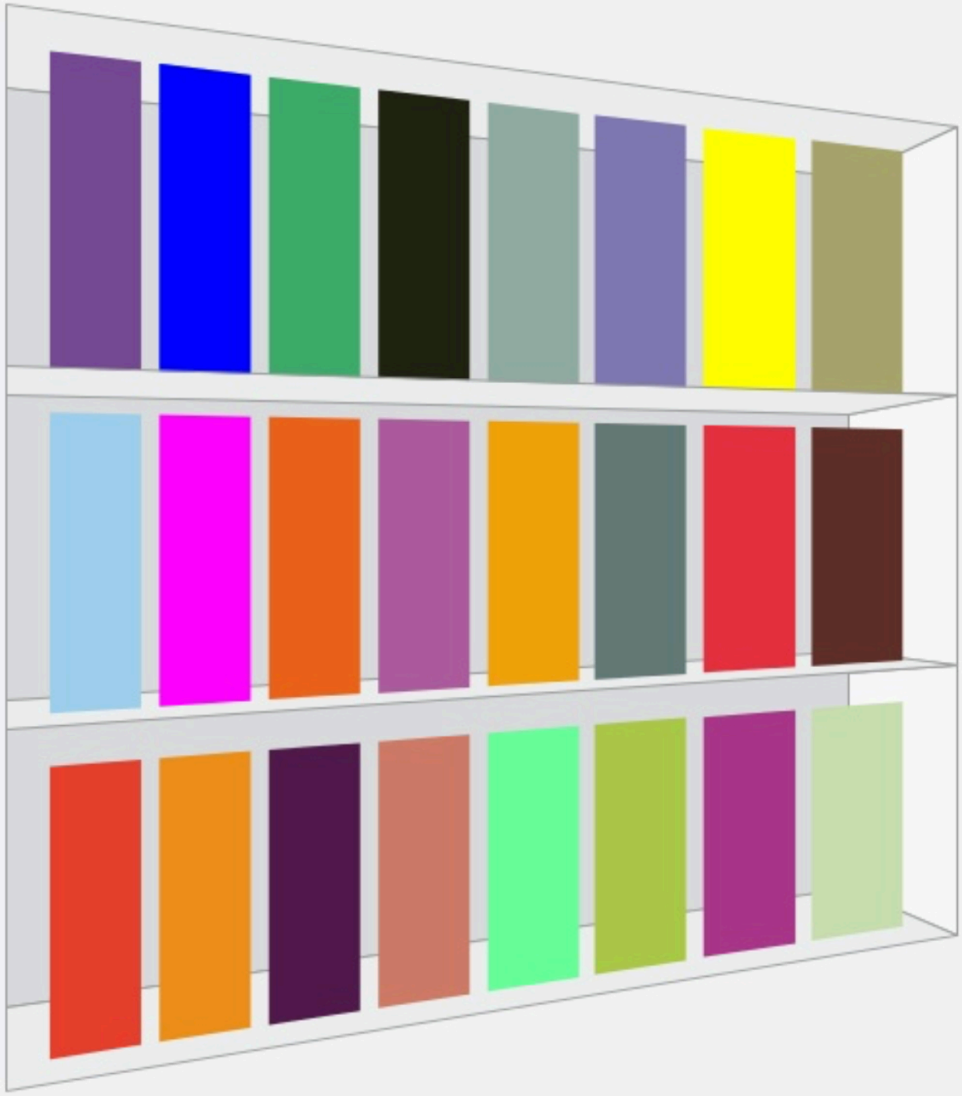
The traditional approach

Generic allocation
of limited shelf space

SKUs ranked by sales

No substitution
of SKUs considered

Limited granularity



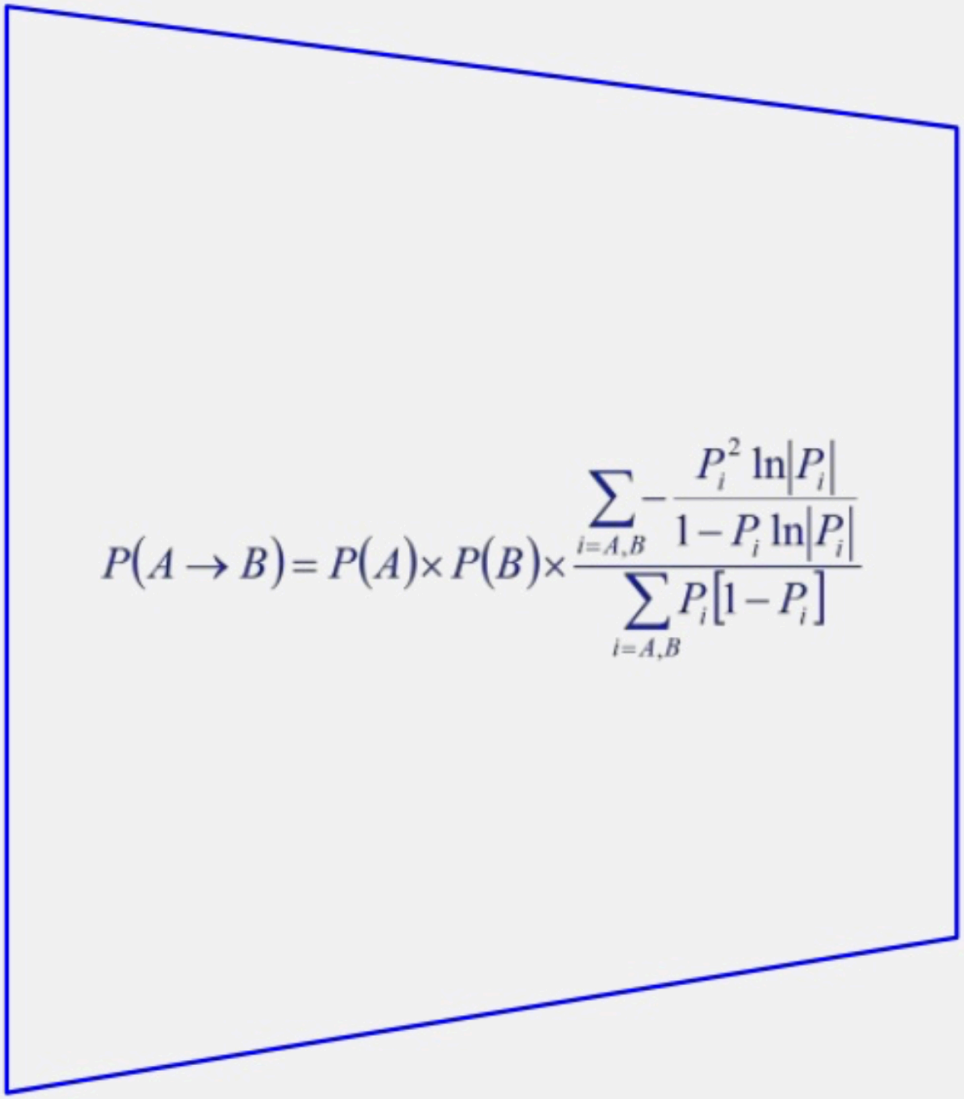
The Big Data approach

Terabytes of data

Multi-year transaction data

Consumer panel data

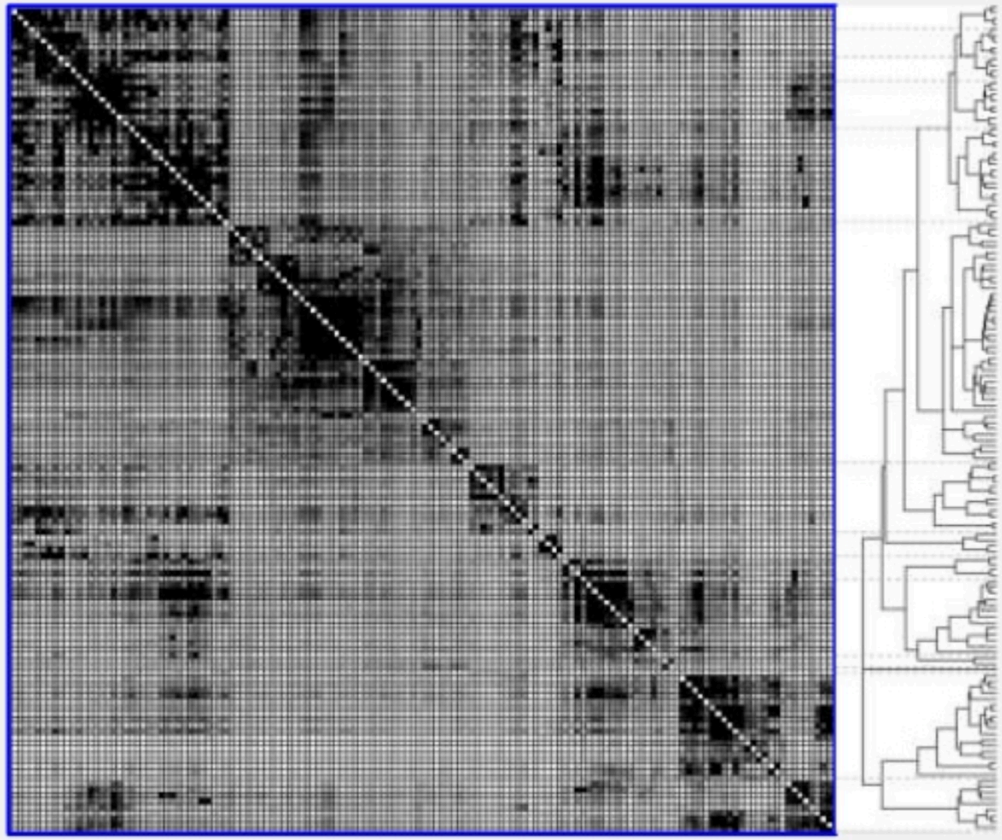
Loyalty card data



$$P(A \to B) = P(A) \times P(B) \times \frac{\sum_{i=A,B} -\frac{P_i^2 \ln|P_i|}{1 - P_i \ln|P_i|}}{\sum_{i=A,B} P_i[1 - P_i]}$$

Advanced statistical methods

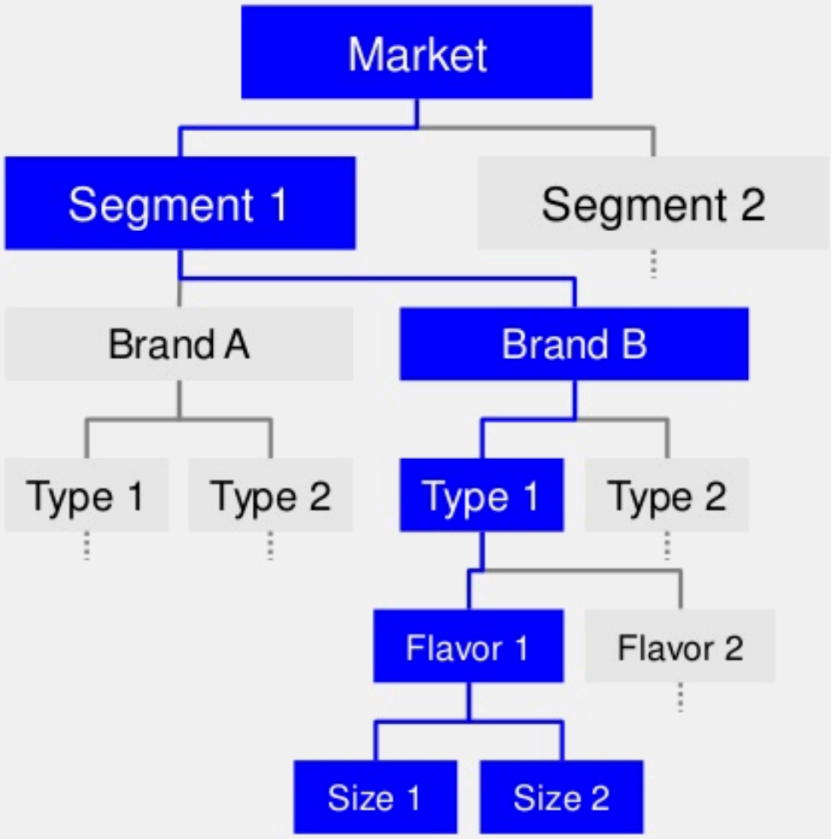
Stochastic switching model (entropy calculations)



Advanced statistical methods

Stochastic switching model (entropy calculations)

Hierarchical clustering (dendograms)



Advanced statistical methods

Multidimensional scaling (consumer decision tree)

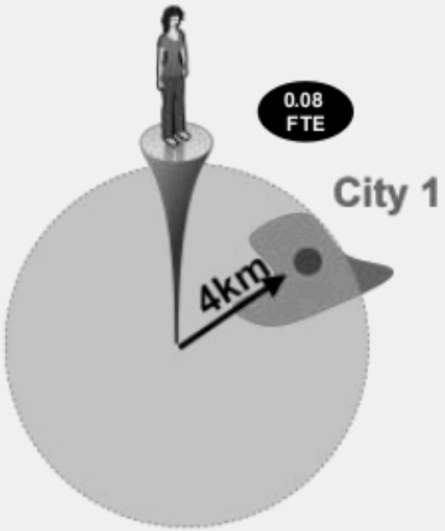
- Actual behavior (switching, walk rates)
- Statistically relevant
- Optimal SKU selection per store
- Predictive sales forecast

Revenue growth more than double the category growth in the market

Impact

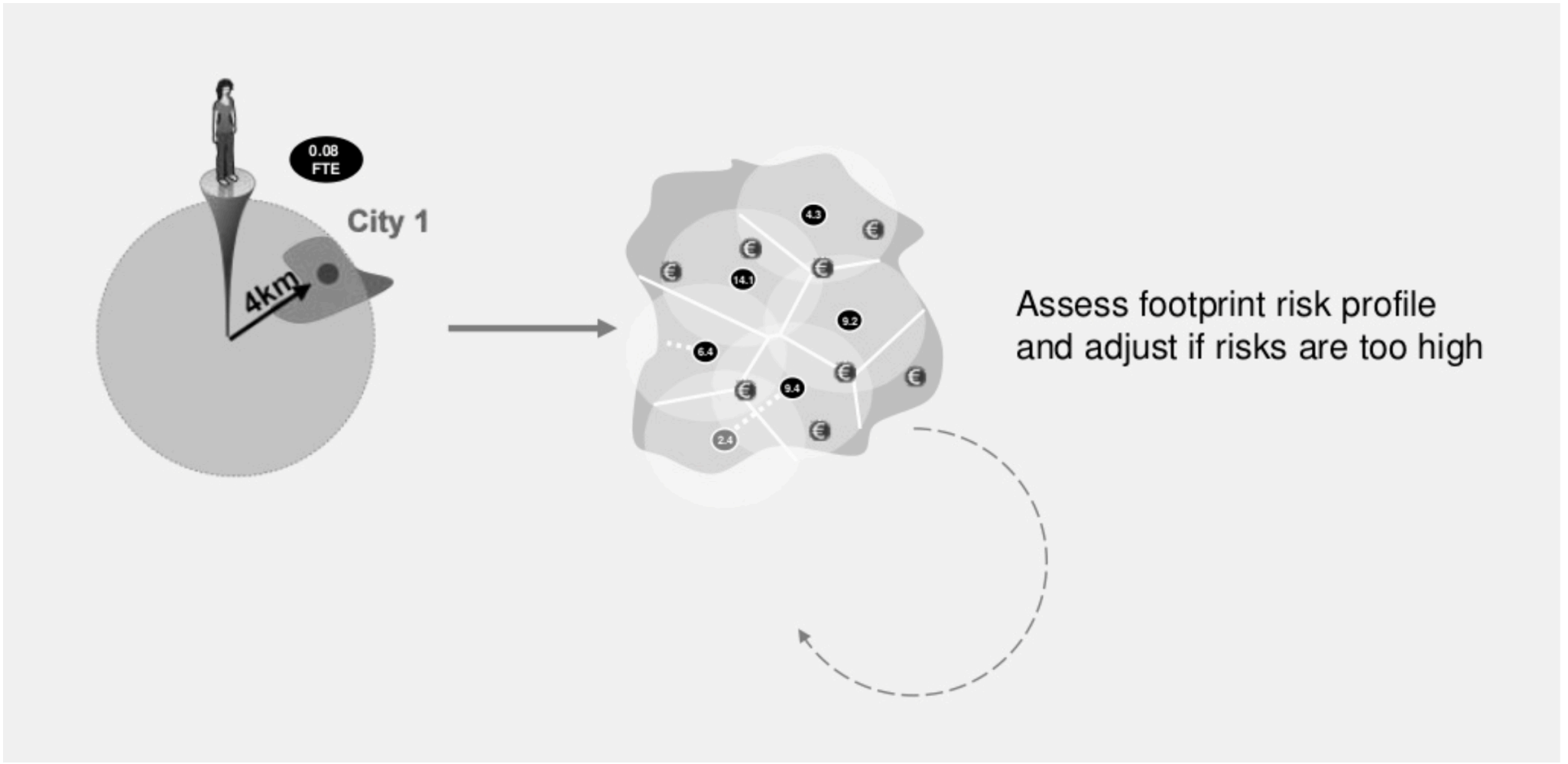
Optimizing branch networks

Segment customers based on channel behaviour and willingness to travel

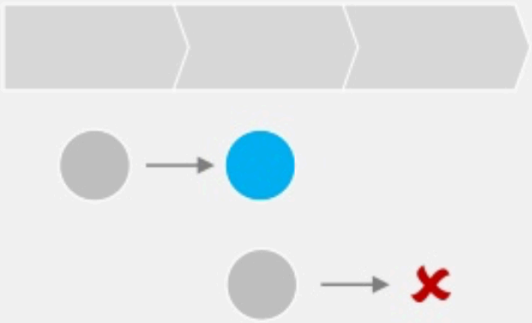


Determine required capacity by customer and plot capacity within micromarket using geo-marketing methods

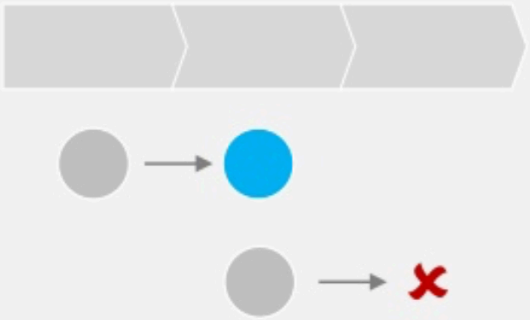
Define branch concepts (e.g., advice branch) in line with multi-channel strategy


0.08
FTE
City 1
4km
4.3
14.1
9.2
6.4
9.4
2.4
Assess footprint risk profile
and adjust if risks are too high

Optimize locations to set up branch for success



Carefully plan and execute micromarket transition



Multiple scenarios based on

90 processes

7 million customers

1,000 branches

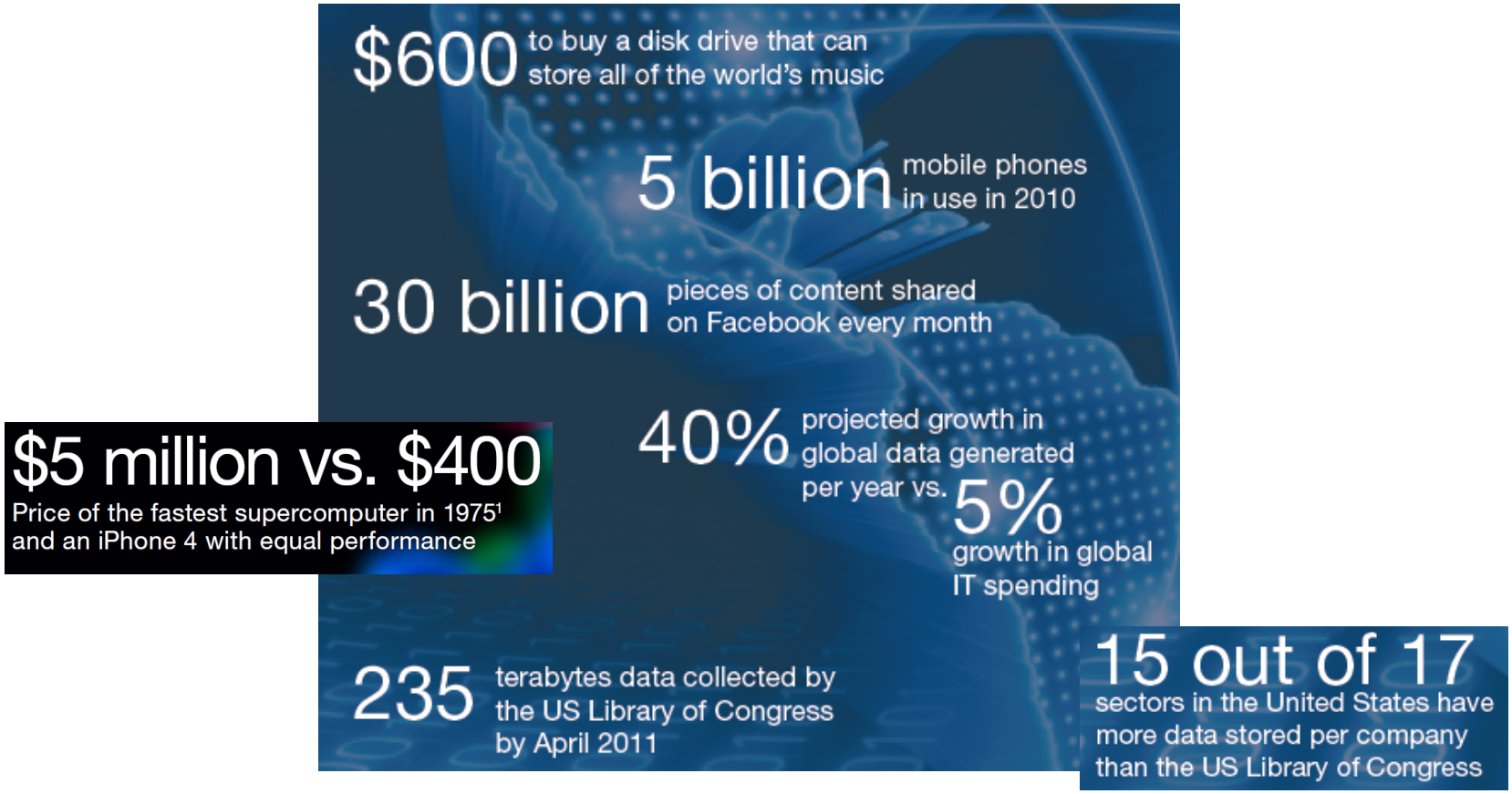
40% cost reduction Impact

< 1% revenue at risk

What is Data Mining?

Knowledge discovery from data

From the Jure Leskovec, Anand Rajaraman, Jeff Ullman Stanford University http://www.mmds.org


$600 to buy a disk drive that can store all of the world's music
5 billion mobile phones in use in 2010
30 billion pieces of content shared on Facebook every month
$5 million vs. $400
Price of the fastest supercomputer in 1975[1] and an iPhone 4 with equal performance
40% projected growth in global data generated per year vs. 5% growth in global IT spending
235 terabytes data collected by the US Library of Congress by April 2011
15 out of 17
sectors in the United States have more data stored per company than the US Library of Congress

Data contains value & knowledge

Knowledge extraction

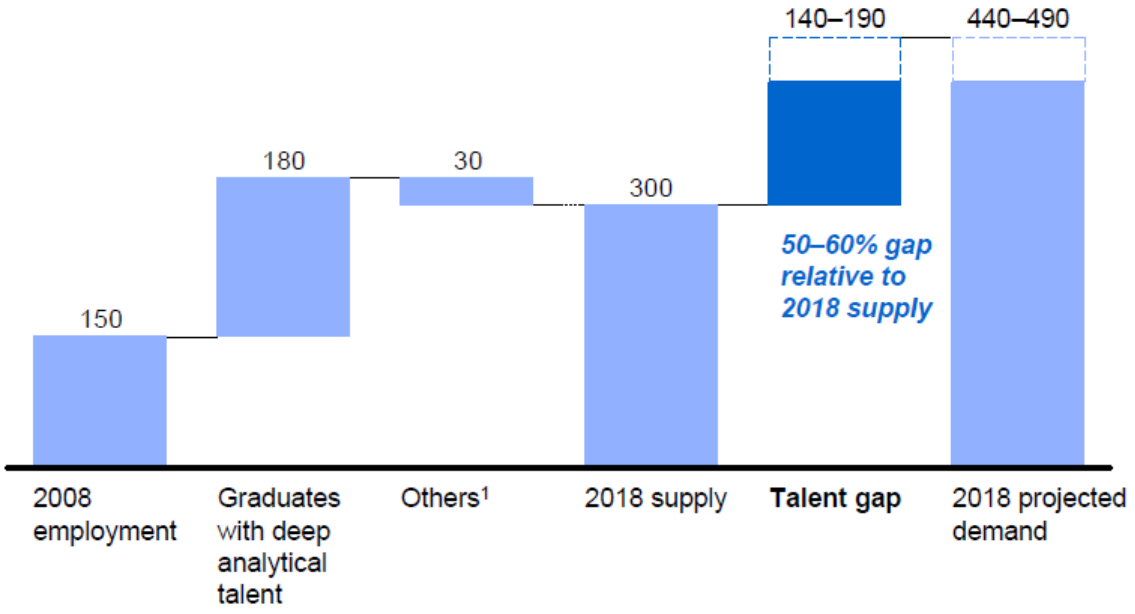
- Data needs to be
  - Stored
  - Managed
  - **ANALYZED** ← this class

Data Mining ≈ Big Data ≈ Predictive Analytics ≈ Data Science

Demand for Data Mining

Demand for deep analytical talent in the United States could be 50 to 60 percent greater than its projected supply by 2018

Supply and demand of deep analytical talent by 2018
Thousand people



1 Other supply drivers include attrition (-), immigration (+), and reemploying previously unemployed deep analytical talent (+).
SOURCE: US Bureau of Labor Statistics; US Census; Dun & Bradstreet; company interviews; McKinsey Global Institute analysis

Principle

- **Given lots of data**
- **Discover patterns and models that are:**
  - **Valid:** hold on new data with some certainty
  - **Useful:** should be possible to act on the item
  - **Unexpected:** non-obvious to the system
  - **Understandable:** humans should be able to interpret the pattern

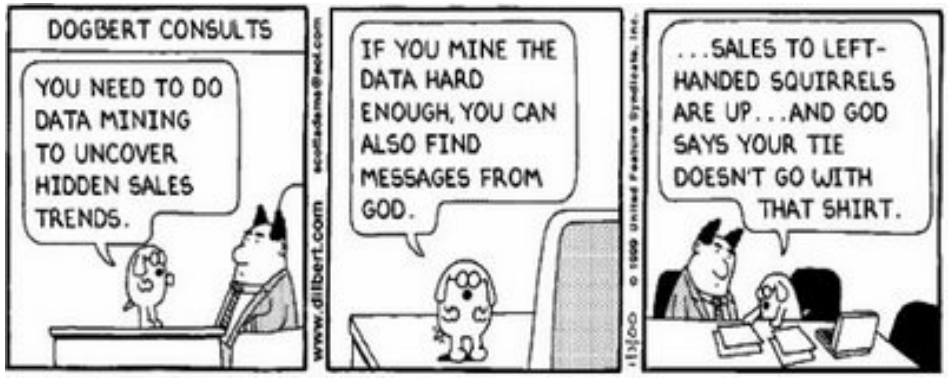
Data Mining Tasks

- **Descriptive methods**
  - Find human-interpretable patterns that describe the data
    - **Example:** Clustering
- **Predictive methods**
  - Use some variables to predict unknown or future values of other variables
    - **Example:** Recommender systems

Meaningfulness of Analytic Answers

- **A risk with "Data mining" is that an analyst can "discover" patterns that are meaningless**

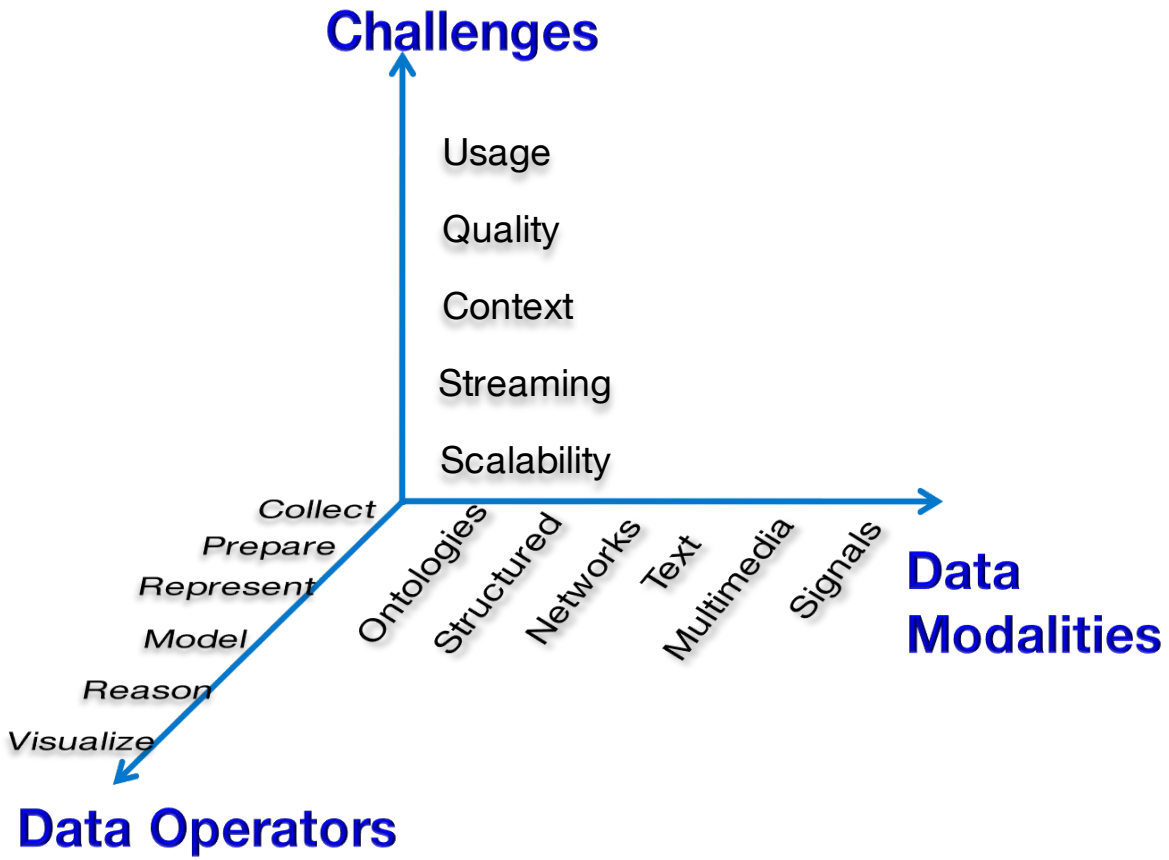
- Statisticians call it **Bonferroni's principle**:
  - Roughly, if you look in more places for interesting patterns than your amount of data will support, you are bound to find crap



Meaningfulness of Analytic Answers

- We want to find (unrelated) people who **at least twice have stayed at the same hotel on the same day**
  - 10^9 people being tracked
  - 1,000 days
  - Each person stays in a hotel 1% of time (1 day out of 100)
  - Hotels hold 100 people (so 10^5 hotels)
  - **If everyone behaves randomly (i.e., no terrorists) will the data mining detect anything suspicious?**
- **Expected number of "suspicious" pairs of people:**
  - 250,000
  - … too many combinations to check – we need to have some additional evidence to find "suspicious" pairs of people in some more efficient way

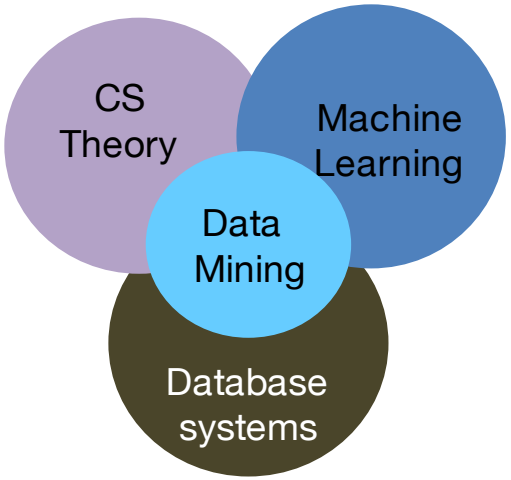
What matters when dealing with data?



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

Data Mining: Cultures

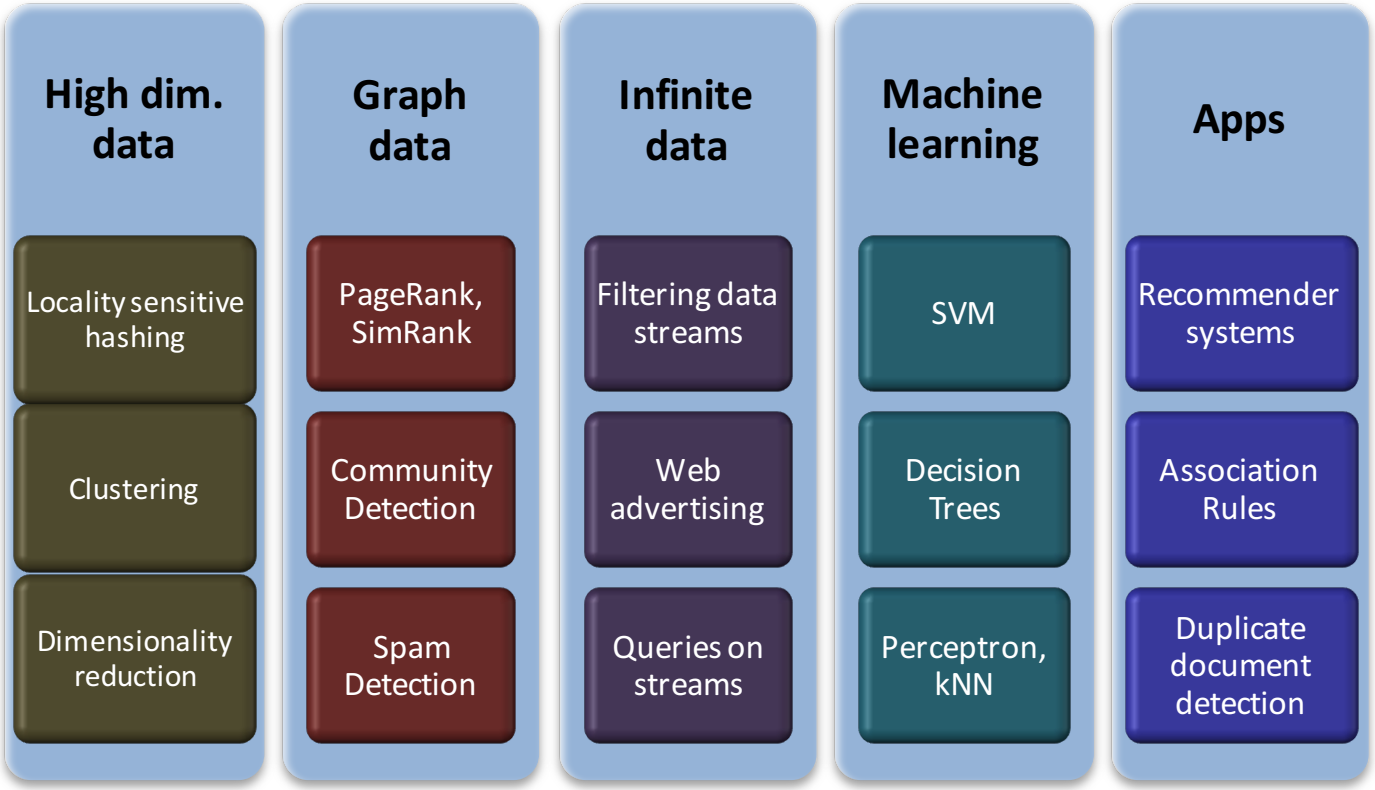
- **Data mining overlaps with:**
  - **Databases:** Large-scale data, simple queries
  - **Machine learning:** Small data, Complex models
  - **CS Theory:** (Randomized) Algorithms
- **Different cultures:**
  - To a DB person, data mining is an extreme form of **analytic processing** – queries that examine large amounts of data
    - Result is the query answer
  - To a ML person, data-mining is the **inference of models**
    - Result is the parameters of the model
- **In this class we will do both!**



What will we learn?

- **We will learn to mine different types of data:**
  - Data is high dimensional
  - Data is a graph
  - Data is infinite/never-ending
  - Data is labeled
- **We will learn to use different models of computation:**
  - MapReduce
  - Streams and online algorithms
  - Single machine in-memory

How It All Fits Together



J. Leskovec, A. Rajaraman, J. Ullman: Mining of Massive Datasets, http://www.mmds.org

How do you want that data?