



UNIVERSITEIT  
VAN AMSTERDAM



Universiteit  
Leiden

# Big Data

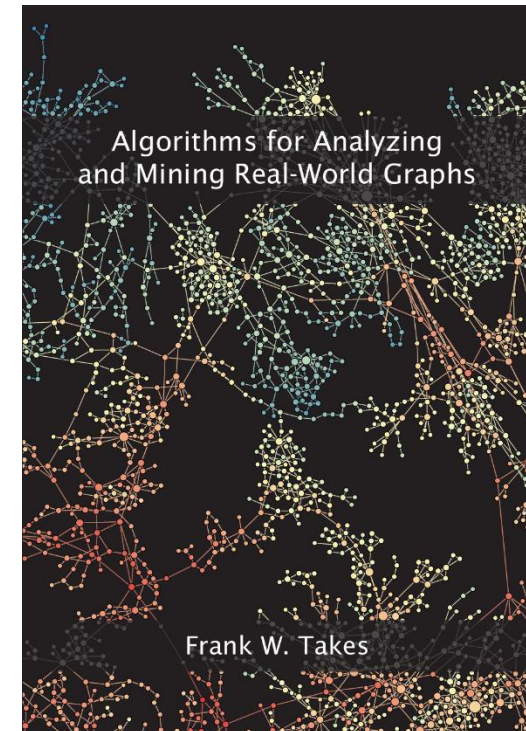
Frank Takes

LIACS, Universiteit Leiden  
AISSR, Universiteit van Amsterdam

AcceptEmail, 28 augustus 2015

# Introduction

- Frank Takes
- BSc: Informatica & Bedrijfswetenschappen (2008)
- MSc: Computer Science (2010)
- PhD: Network Analysis (2014)
- Research: Network Science
- Teaching:
  - Social Network Analysis for Computer Scientists
  - Business Intelligence and Process Modelling
- Academic webpage: <http://franktakes.nl>

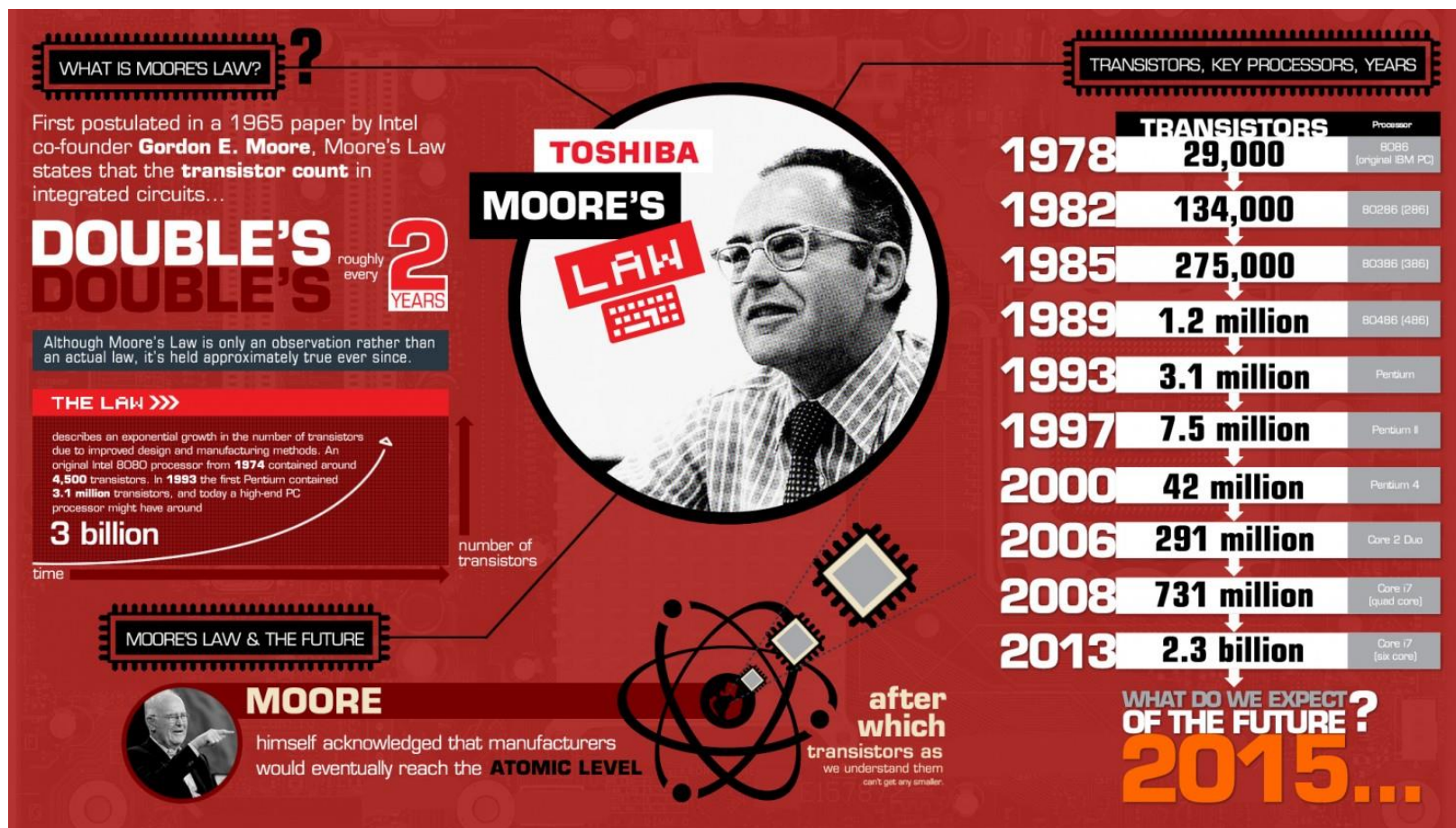




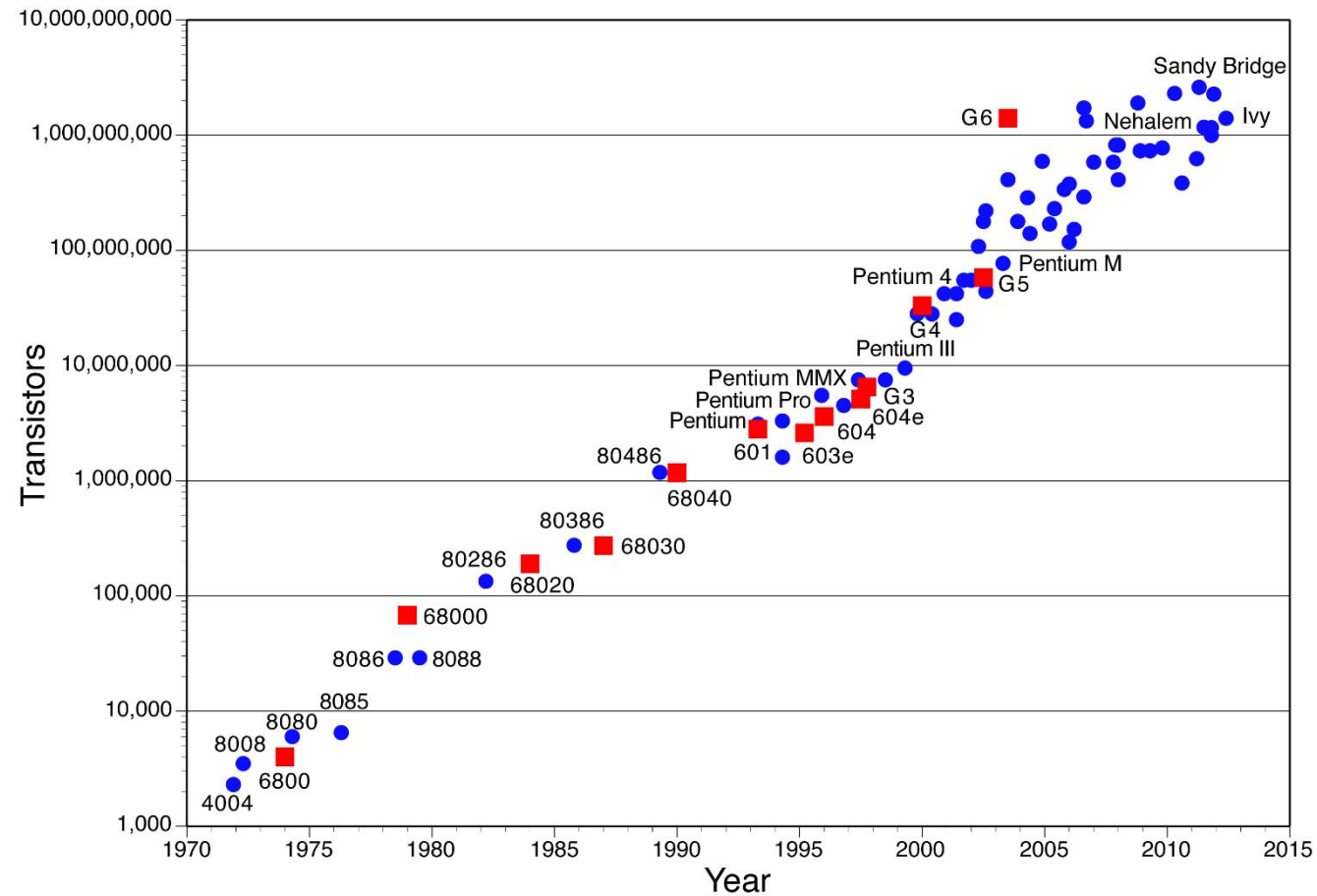
# Overview

- Data
- Big Data
- Knowledge Discovery
- Machine Learning
- Examples
- Outlook

# Moore's Law & Computation

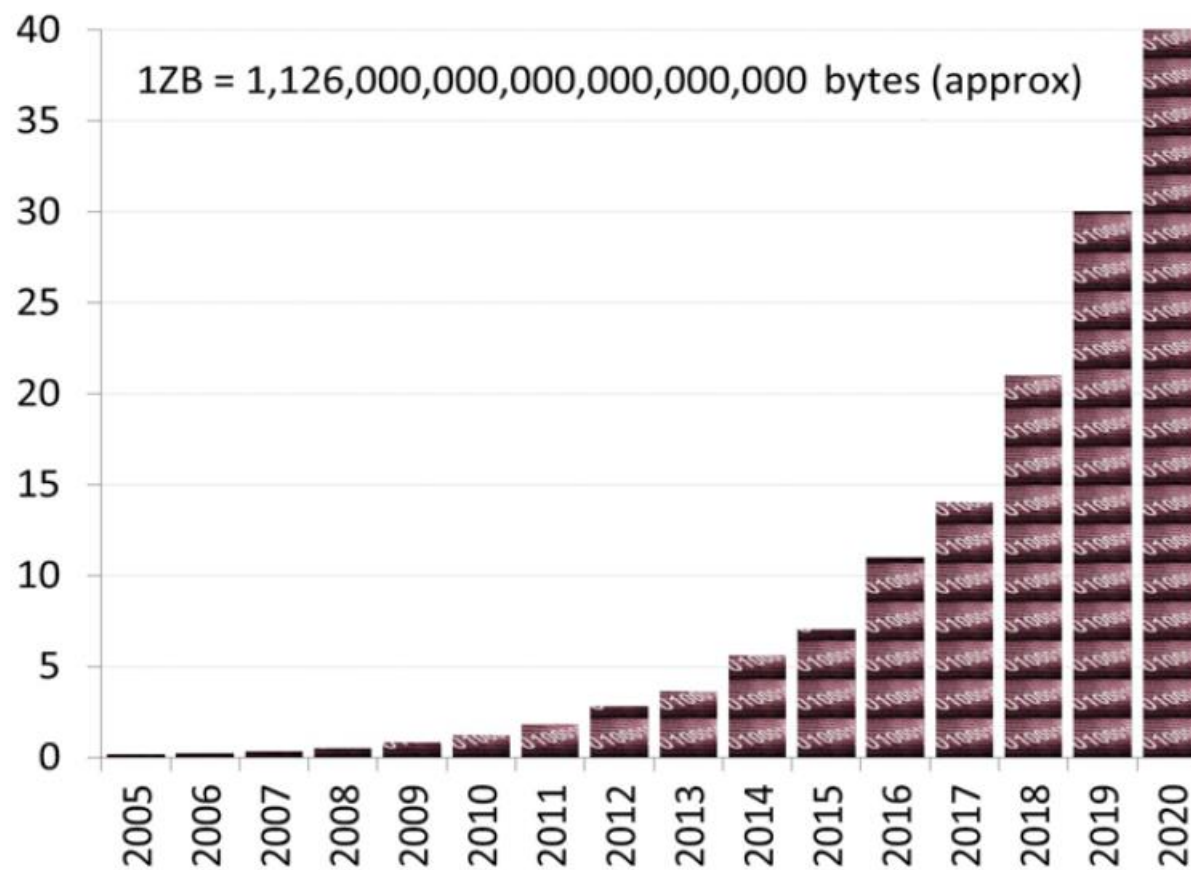


# Moore's Law & Computation



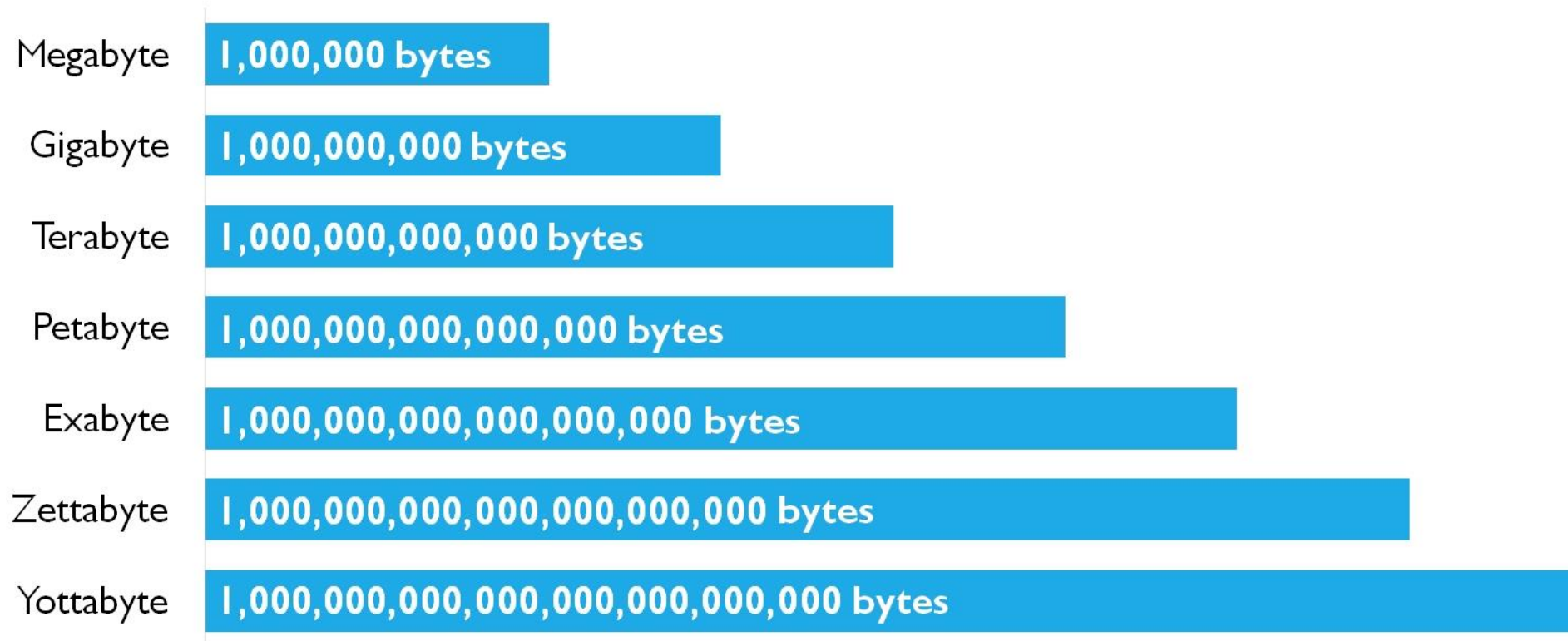


# Moore's Law & Data



Source: <http://www1.unece.org/stat/platform/display/msis/Big+Data>

# Bits and Bytes





# Data

- **Data:** facts, measurements or text collected for reference or analysis (Oxford dictionary, 2014)
- Computer science: data is input
- Industry: data is value
- Data over time:
  - Sensus data (60s)
  - Transactional data (80s)
  - Micro event data (00s)

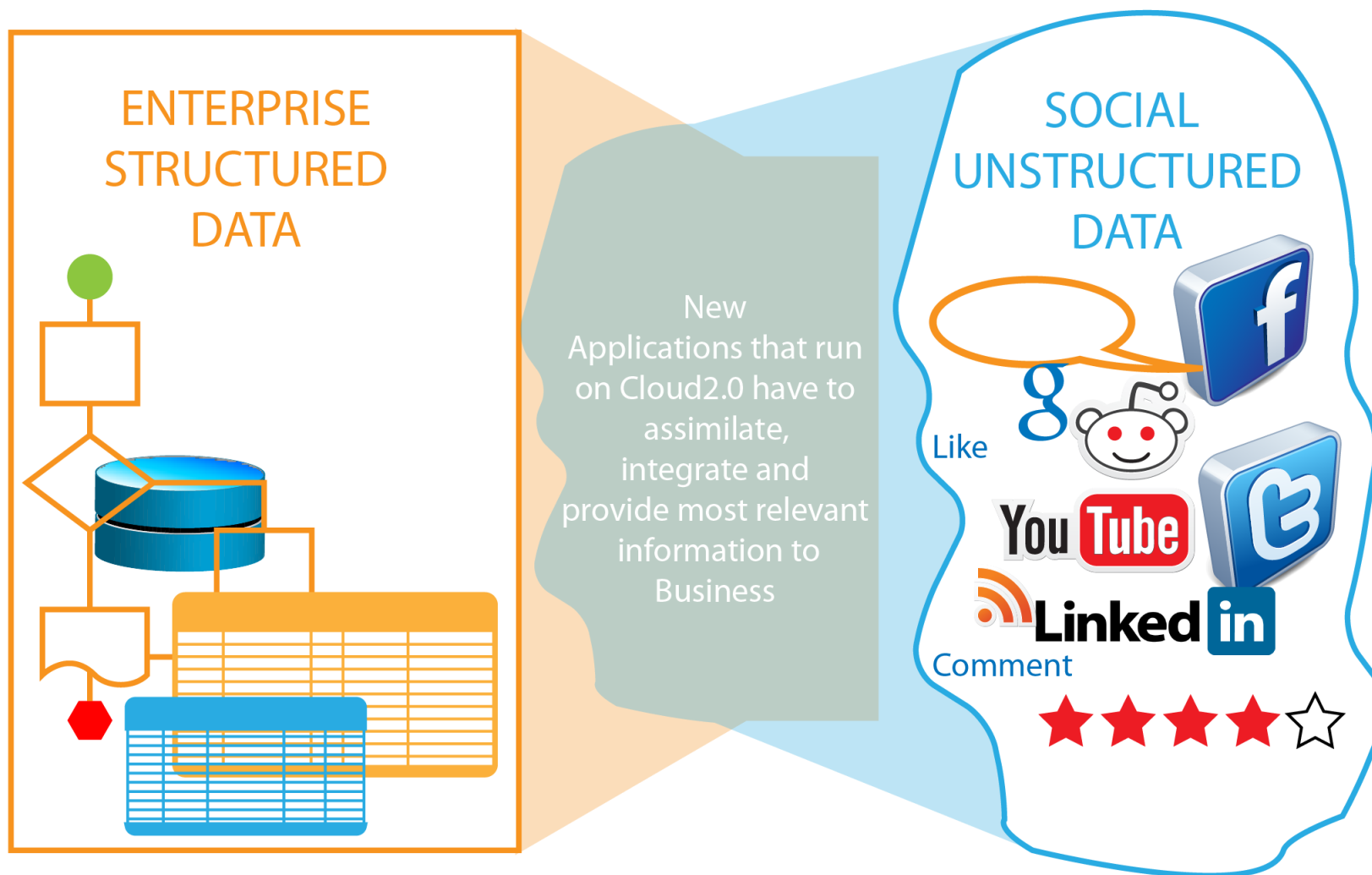




# Structured vs. Unstructured Data

- Popular definition:
  - **Unstructured data:** tabular data or text
  - **Structured data:** network/relational data, data from (social) media
- Traditional definition:
  - **Unstructured data:** data that does not fit a certain data structure (text, a list of numeric measurements)
  - **Structured data:** data that fits a certain data structure (table, tree, graph/network, etc.)

# Structured vs. Unstructured Data

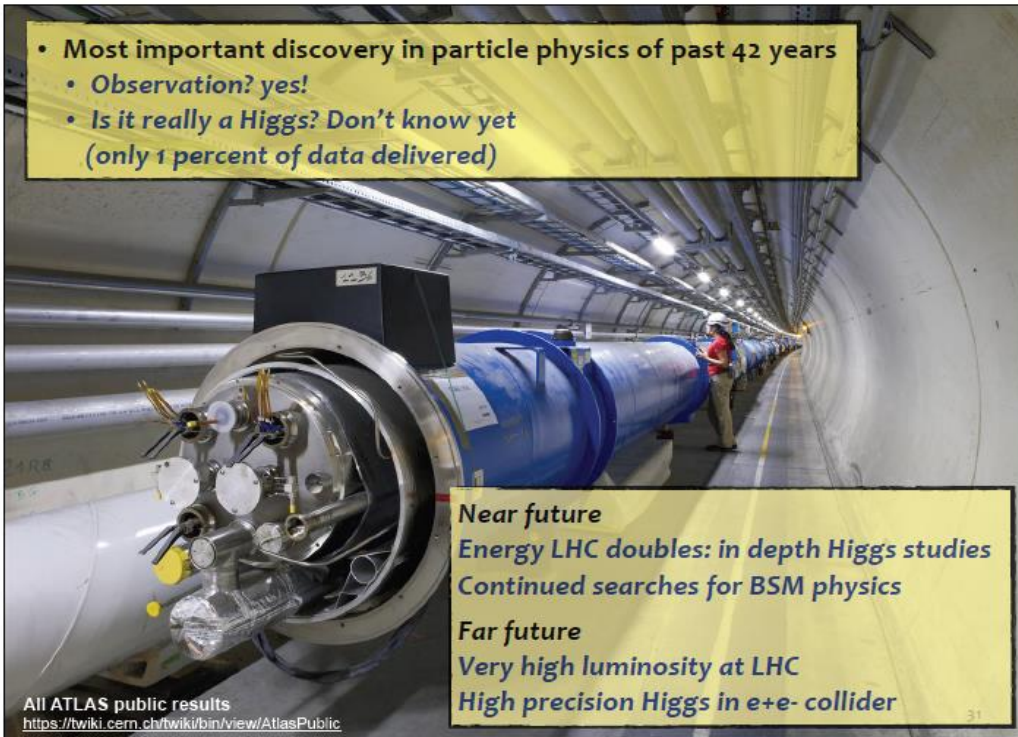


# Example: LHC at CERN





# Breakthrough of the century



- Most important discovery in particle physics of past 42 years
  - Observation? yes!
  - Is it really a Higgs? Don't know yet (only 1 percent of data delivered)

**Near future**  
Energy LHC doubles: in depth Higgs studies  
Continued searches for BSM physics

**Far future**  
Very high luminosity at LHC  
High precision Higgs in  $e+e^-$  collider

All ATLAS public results  
<https://twiki.cern.ch/twiki/bin/view/AtlasPublic>

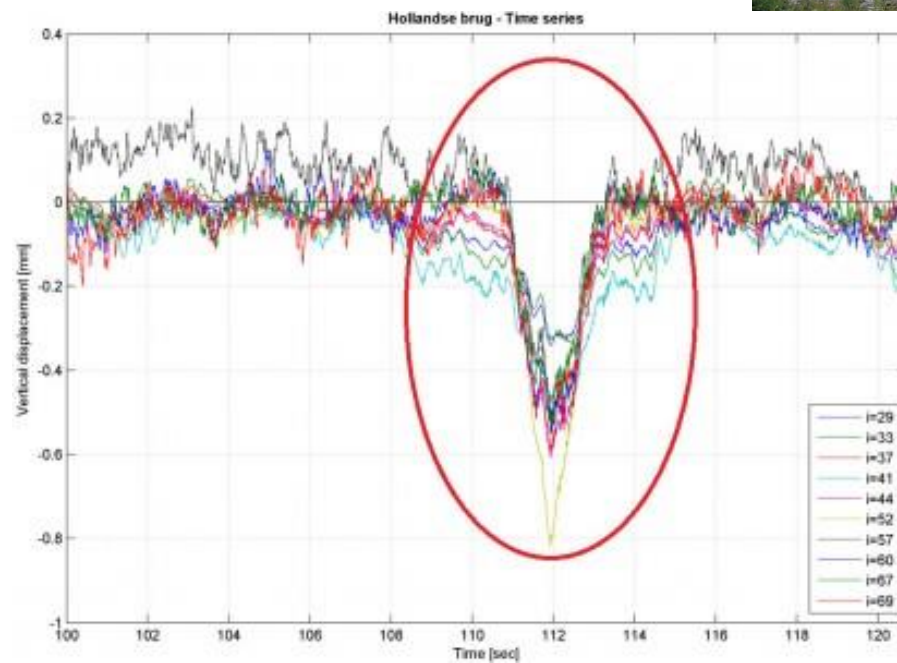
31

## Public announcement, July 4<sup>th</sup> 2012

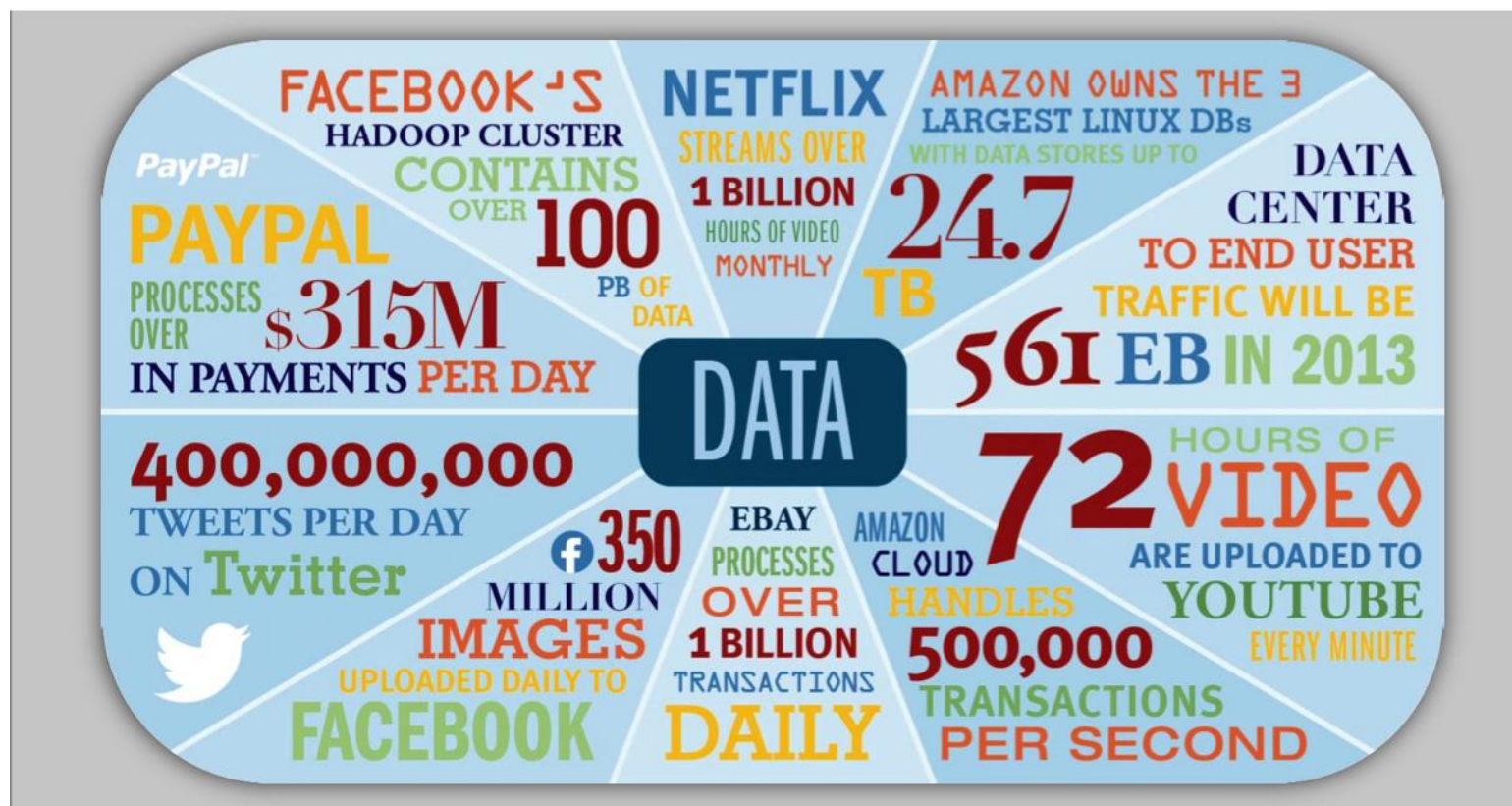




# Example: Hollandse Brug

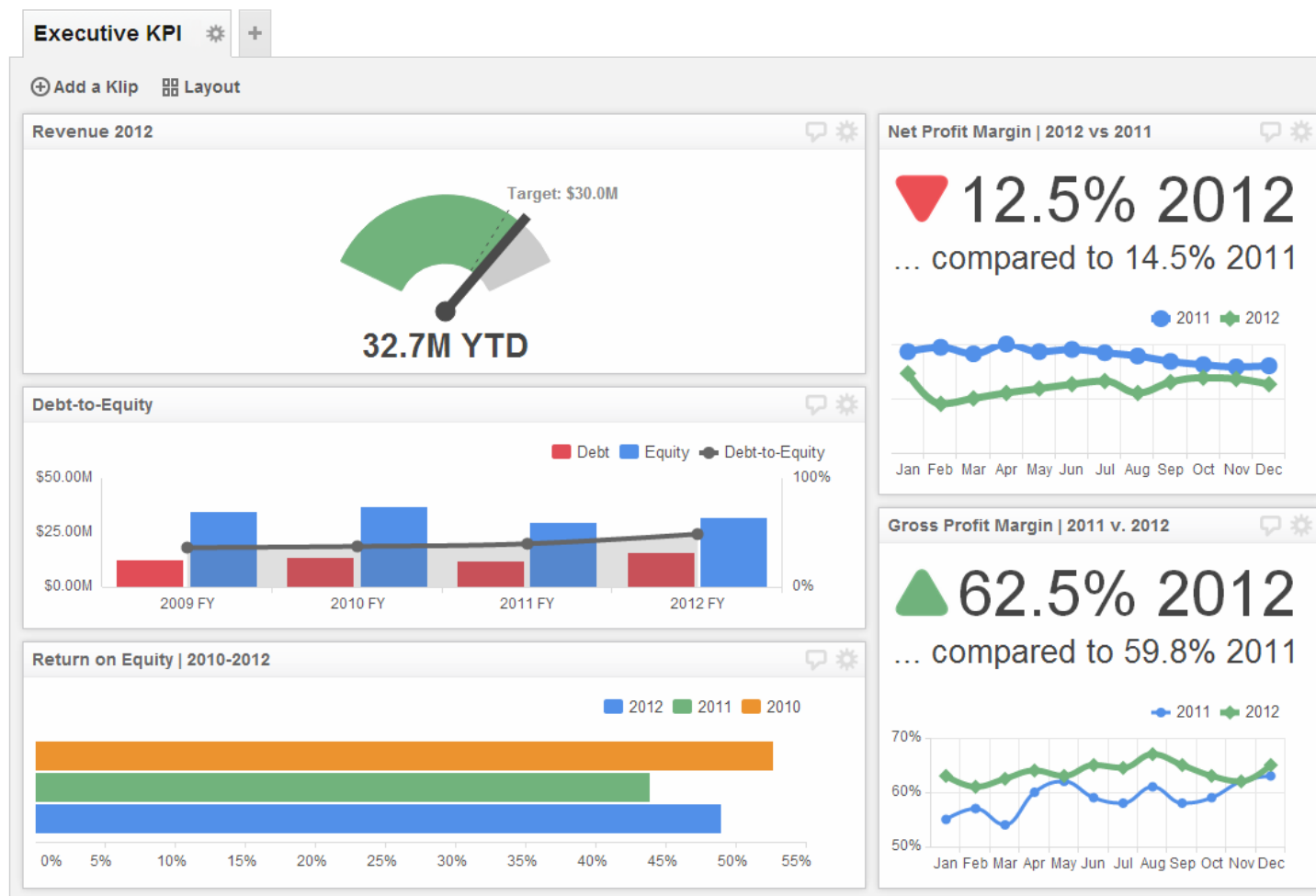


# Social & Web Data



# What can we do with data?

# Visualization



# What can we do with data?

- Business Intelligence 1.0
- OLAP queries
- KPI's
- Dashboards
- Visualizations
- Counting, averaging, sampling, measuring, ...
- But nothing smart, really.





# Privacy

- Data is about people
- People value privacy
- Data analysis may violate privacy
- People are not always happy about data analysis
- So, privacy is an interesting aspect.





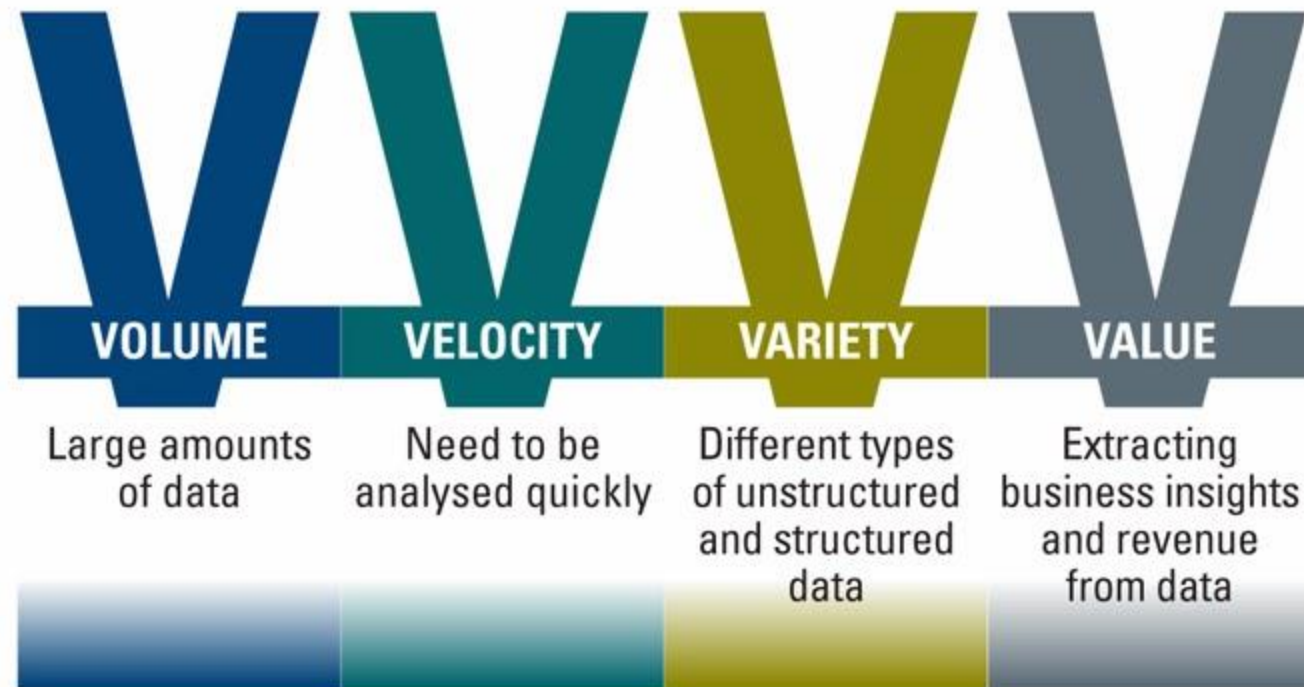
# Big Data

- “**Big data** is the term for a collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.” (White, 2012)
- “**Big data** is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of *predictive analytics* or other certain advanced methods to extract value from data...” (Wikipedia, 2015)

# Big Data

## Big Data: The four Vs

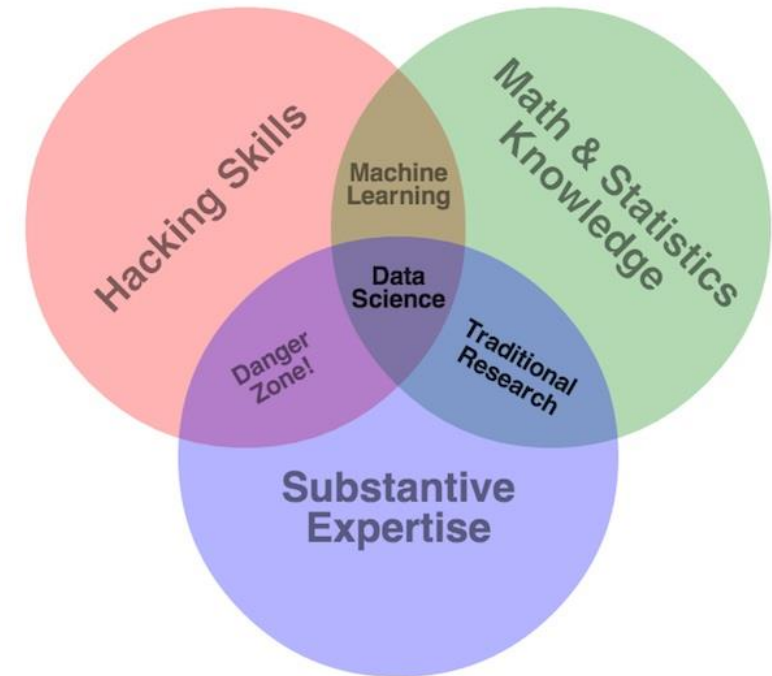
Volume, Velocity, Variety and Value



© World Newsmedia Network 2013

# Data Science

- Data
  - Data Analysis
  - Data Mining
  - Data Science
- 
- Data-driven science vs. Model-driven science
  - Generating hypotheses vs. Validating hypotheses
  - Multidisciplinary aspect

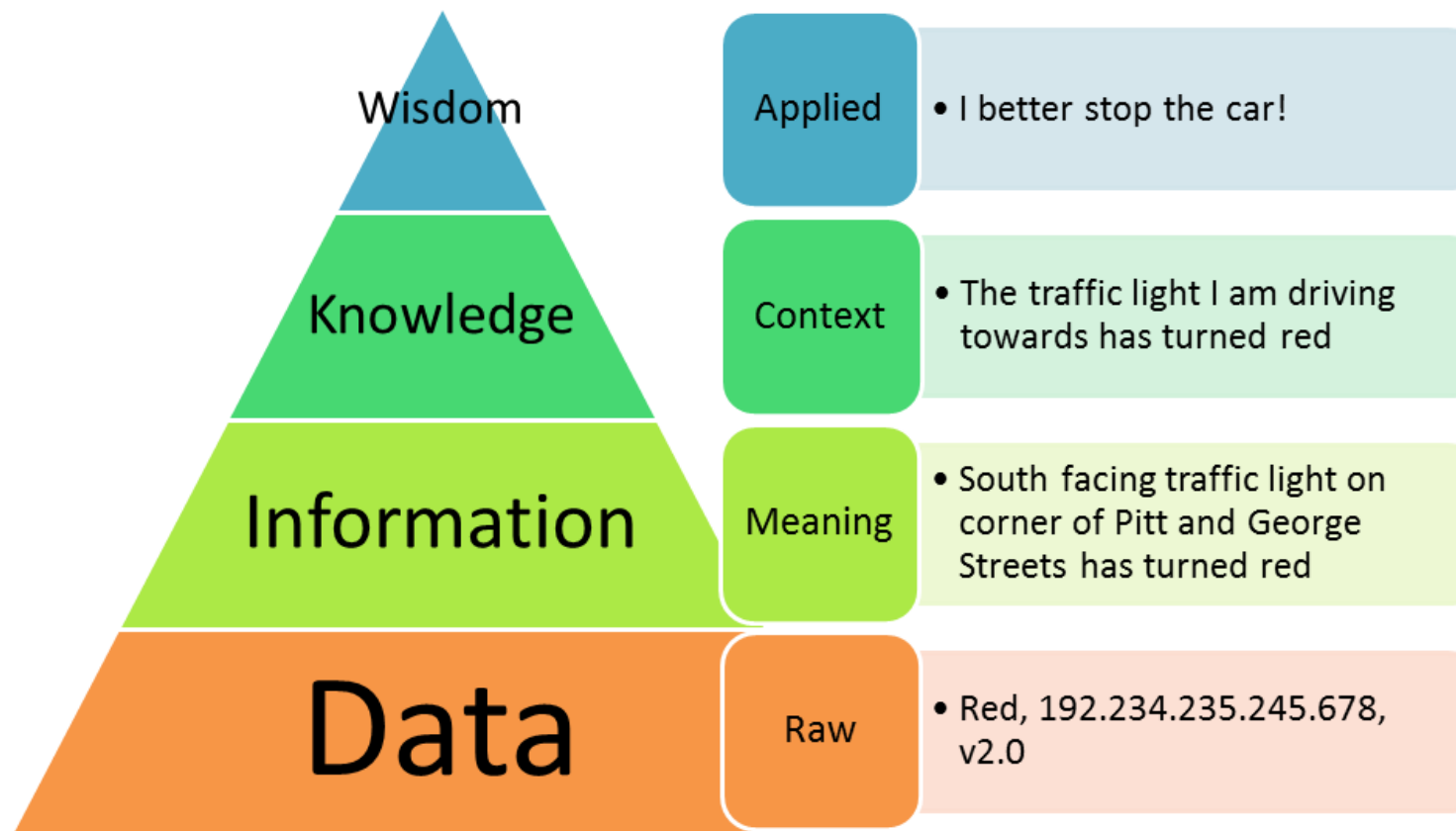


# Knowledge Discovery

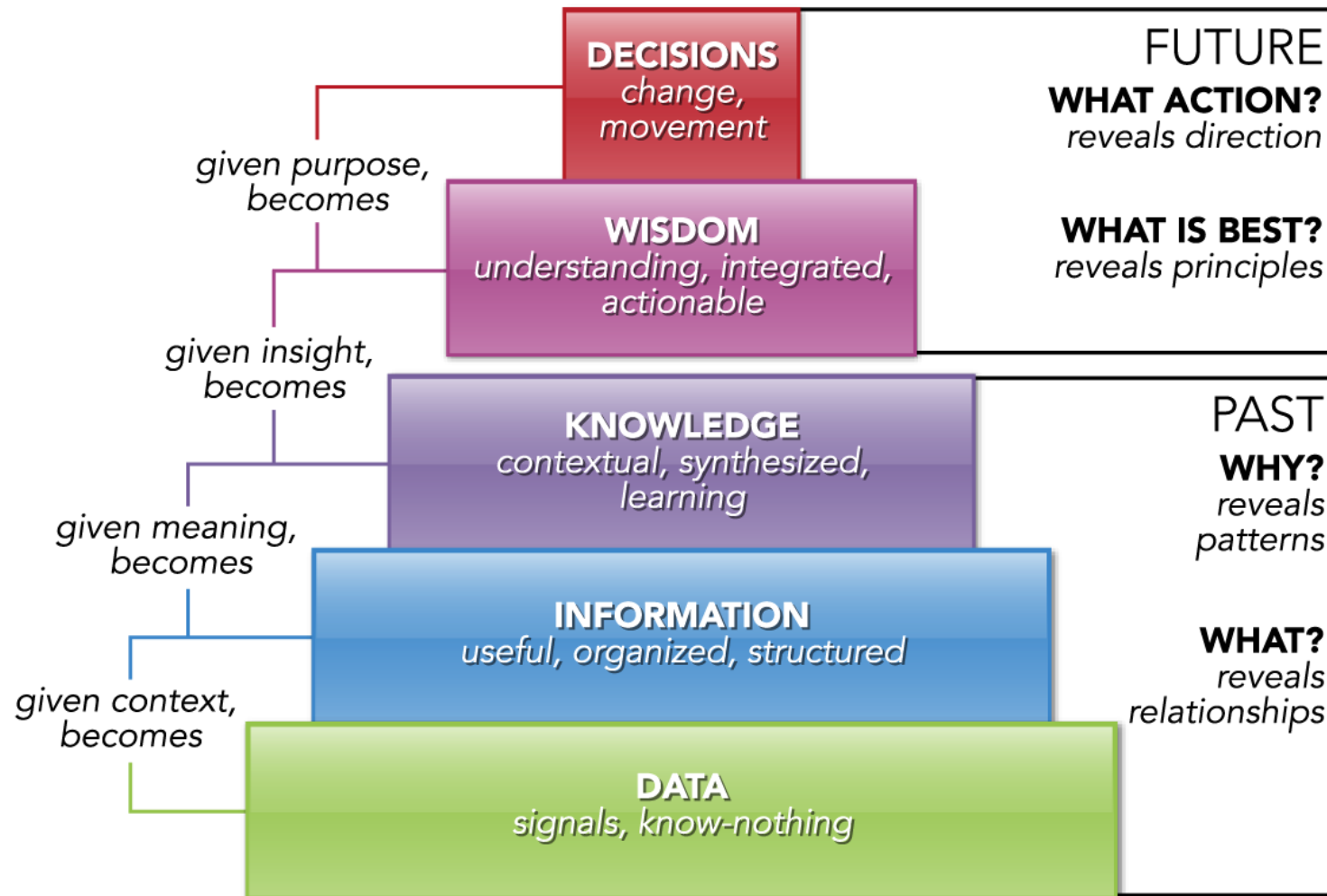
- Data explosion
- *We are drowning in data, but starving for knowledge!*
- Data != Information != Knowledge
- **Knowledge Discovery** in data is the
  - non-trivial process of identifying
  - valid,
  - novel,
  - potentially useful
  - and ultimately understandablepatterns in data.



# DIKW Pyramid



# DIKWD Pyramid





# Example: Facebook



*Wetenschappelijk bewezen*

**Facebook 'begrijpt' jou beter dan je vrienden - en je vrouw**

# Artificial Intelligence



Watson Prepares to Dominate Jeopardy!



# Knowledge Discovery

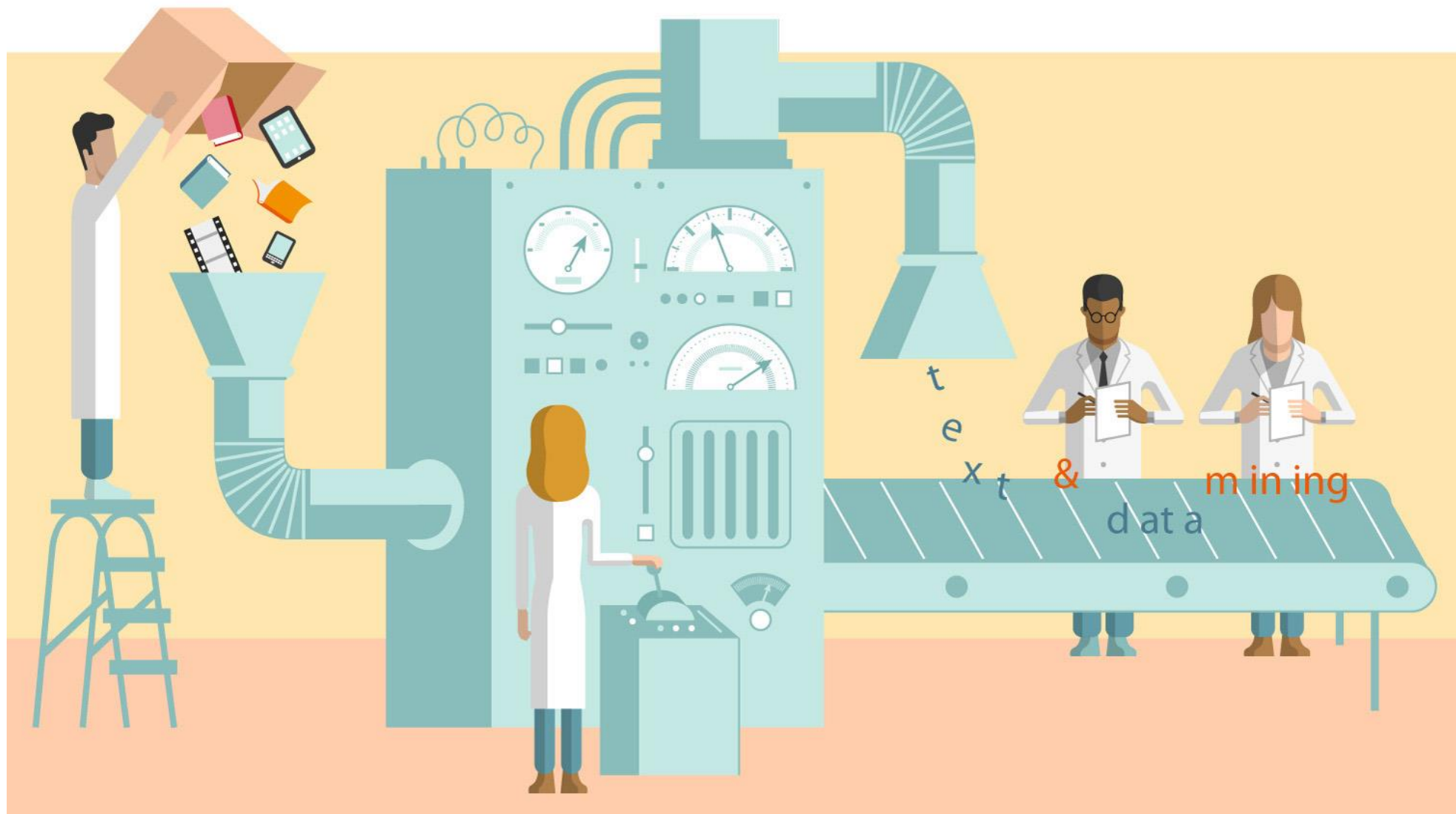
- Data Mining
  - **Descriptive data mining**: clustering, pattern mining, outlier detection, etc.
  - **Predictive data mining**: classification, prediction, recommendation, etc.
- Machine Learning
  - **Supervised** learning: learning on labeled data
  - **Unsupervised** learning: learning/mining on unlabeled data
  - **Semi-supervised** learning: partially labeled data
  - **Reinforcement** learning: agents learning to act in an environment

# Data Science in Leiden

- Bachelor Informatica
  - Minor Data Science
- Master Computer Science
  - Specialization Data Science
- Leiden Center of Data Science (LCDS)
- Big Data summerschools
- Big Data master classes
- Public-private sector collaborations
- External PhD students (e.g. at Belastingdienst, T-Mobile, BMW, etc.)



# Break?



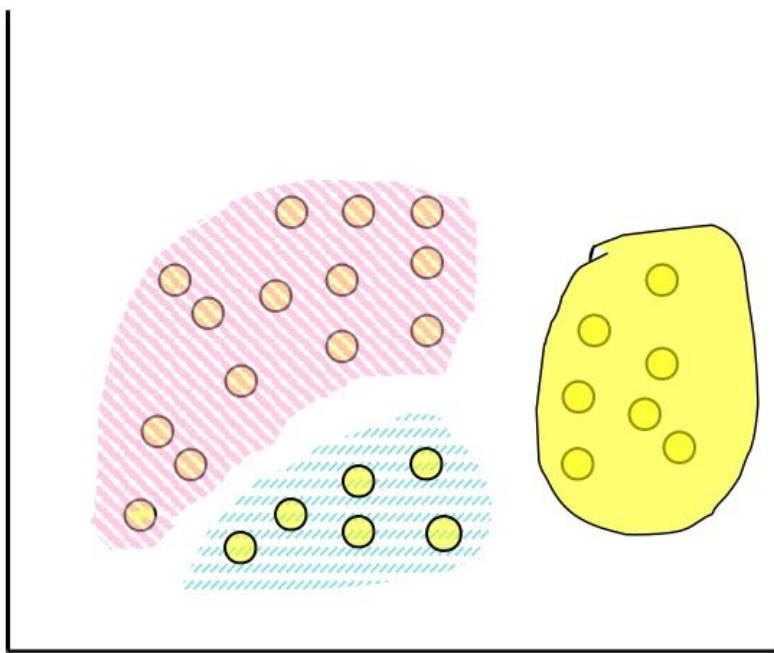


# Steps in Data Mining

- Domain exploration
- Data acquisition
- Data quality assurance
- Data cleaning and preprocessing
- Visualization
- Descriptive analytics
- Predictive analytics
- Interpretation
- Action



# Clustering



- Data is unlabeled
- Grouping based on similar attributes: relatively close "neighbors" in  $n$ -dimensional space belong to the same cluster
- Each of the  $n$  dimensions is a variable (feature)

# Clustering Example: Rijkswaterstaat



# Clustering Example: Rijkswaterstaat

- BSc project J. Kalmeijer in cooperation with “Rijkswaterstaat”
- Total of 254 objects all over the Netherlands
- Energy expenditure over 3 years known for each object
- Measurements every 15 minutes:  
 $365 \text{ days} \times 24 \text{ hours} \times 4 \text{ measurements} = 35.000 \text{ yearly measurements per sensor}$

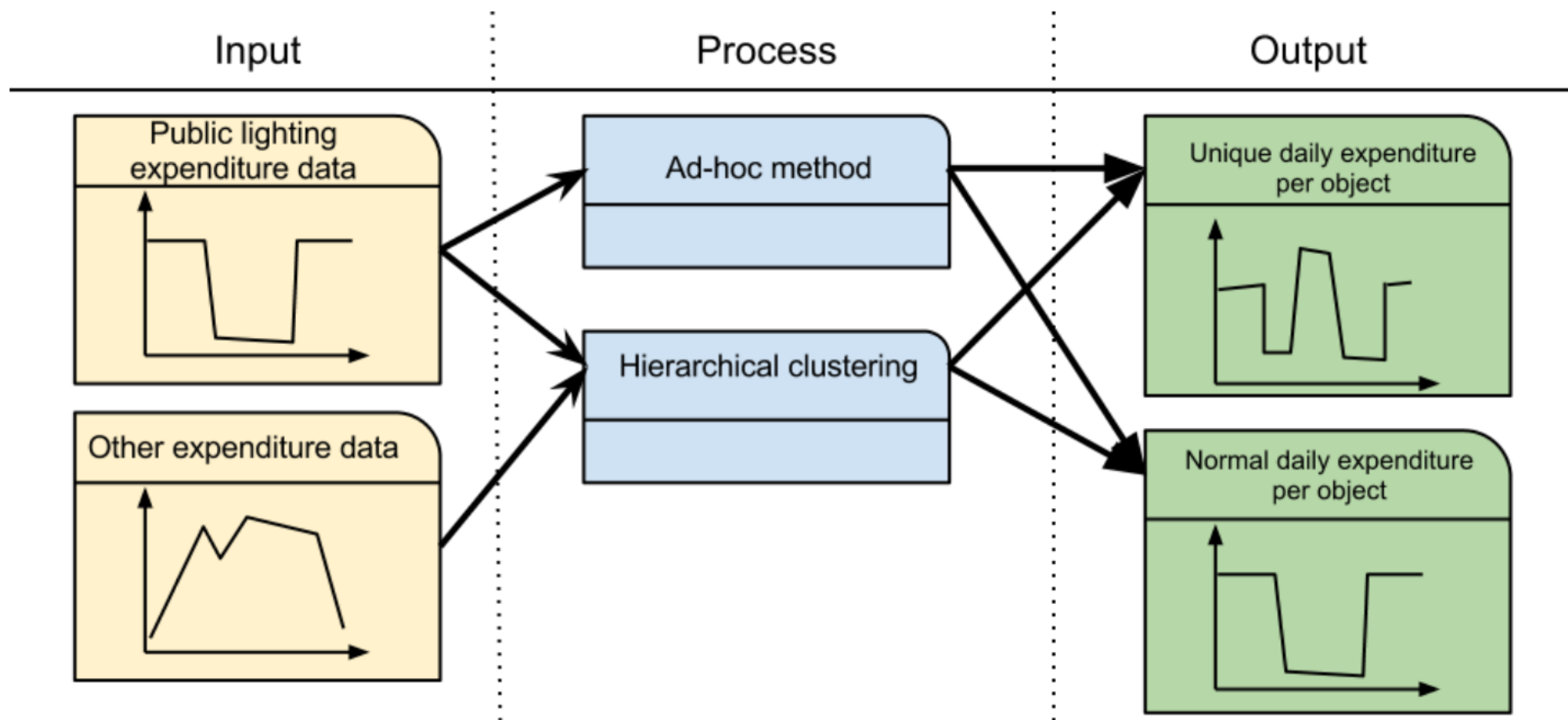


Rijkswaterstaat

# Project Goal

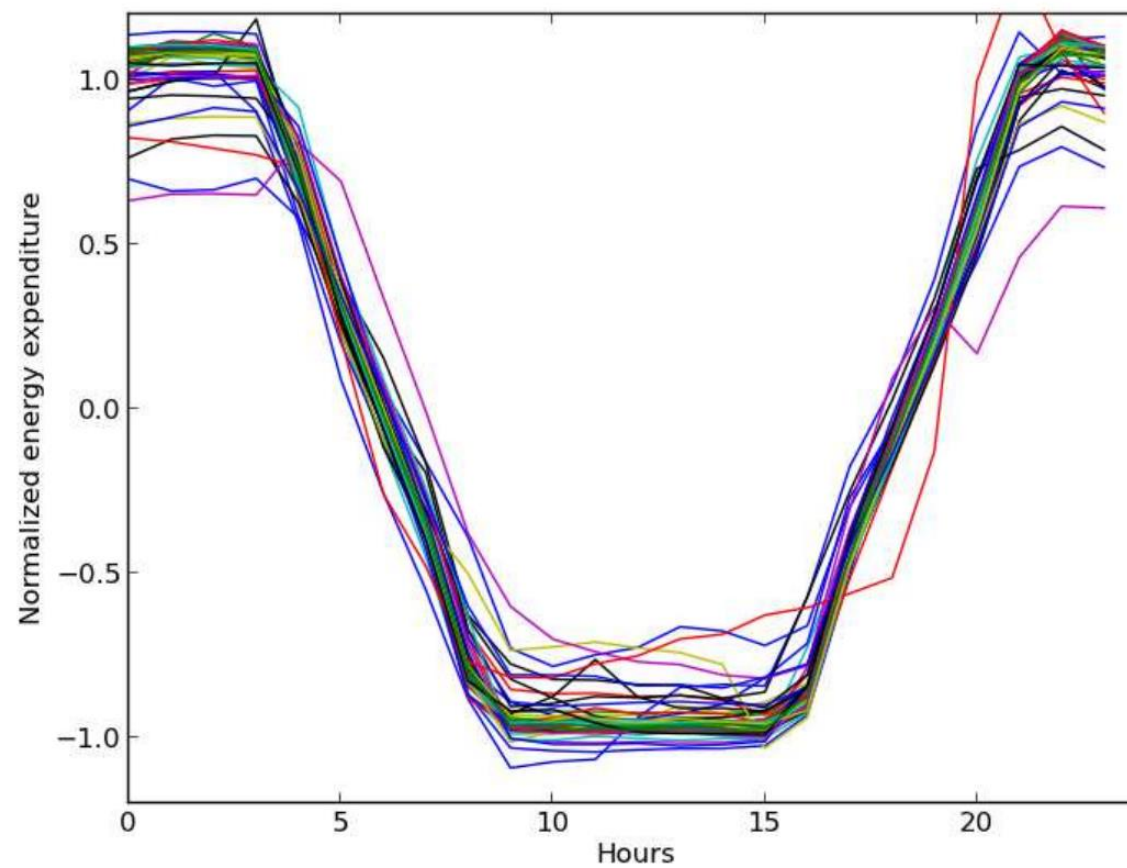
- Clustering on all data:
  - Public lighting
  - All other objects
- Clustering to detect object groups
- Identify regular energy usage pattern of objects
- Objects are of different types
- **Detect anomalies in energy usage per object type**
- Data-driven!

# Approach

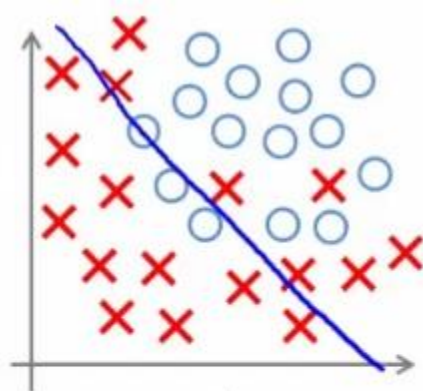




# Clustering

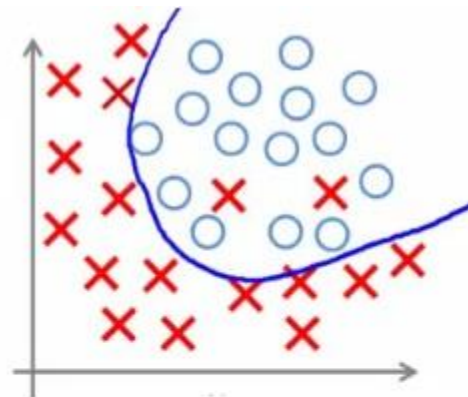


# Overfitting and Underfitting

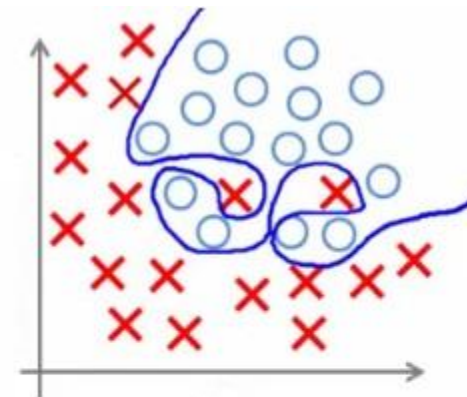


**Under-fitting**

(too simple to  
explain the  
variance)



**Appropriate-fitting**



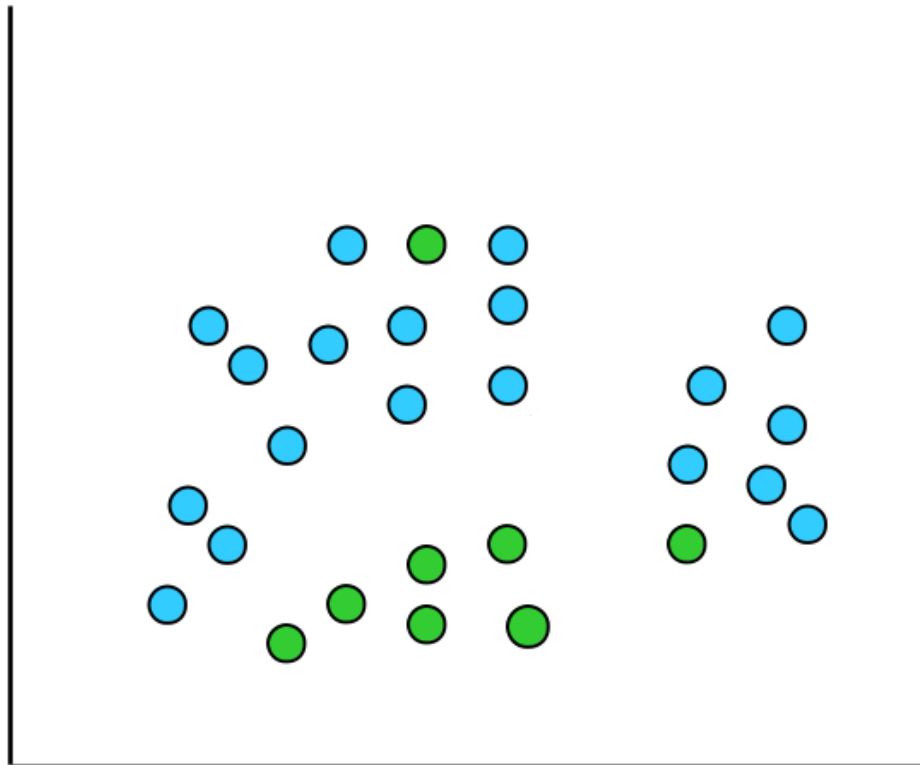
**Over-fitting**

(forcefitting -- too  
good to be true)

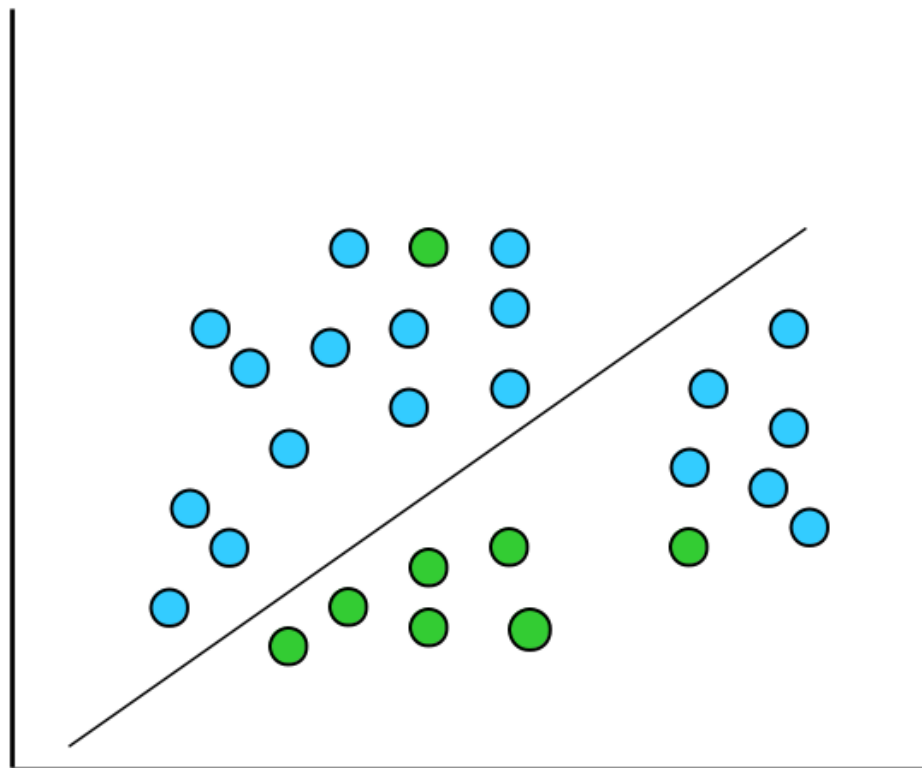
# Predictive Analytics

- So far: descriptive analytics
- **Predictive analytics**
- **Classification**: predict class attributes of data objects
- **Recommendation**: predict future events (e.g., purchases)
- Training data is labeled
- Test and validation data is unlabeled

# Classification



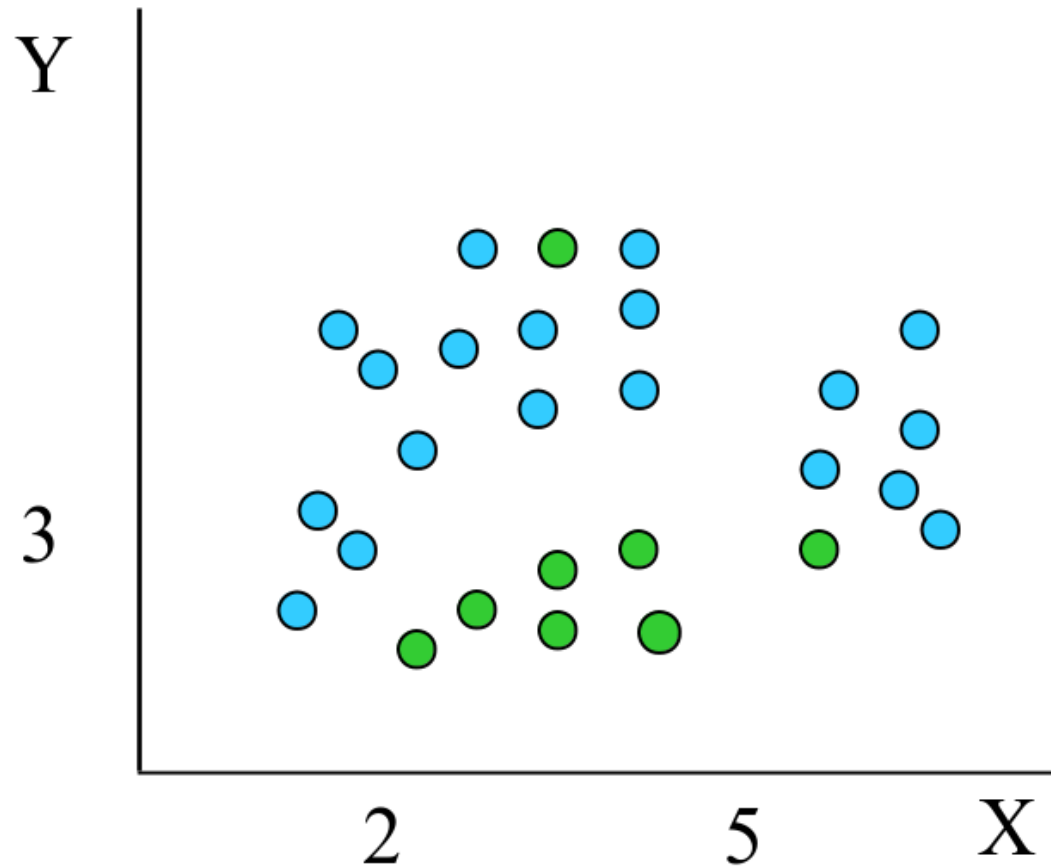
# Regression



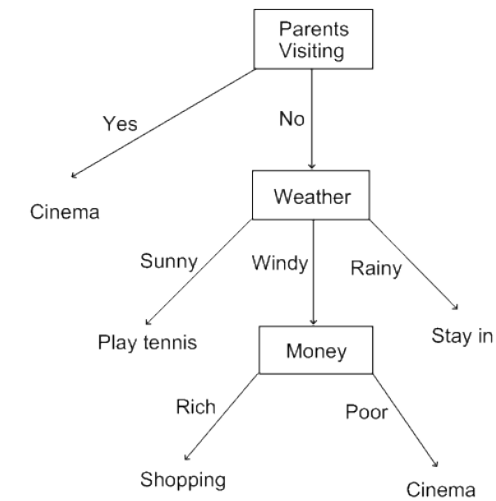
- Separate cases by drawing a line  
 $y = ax + b$



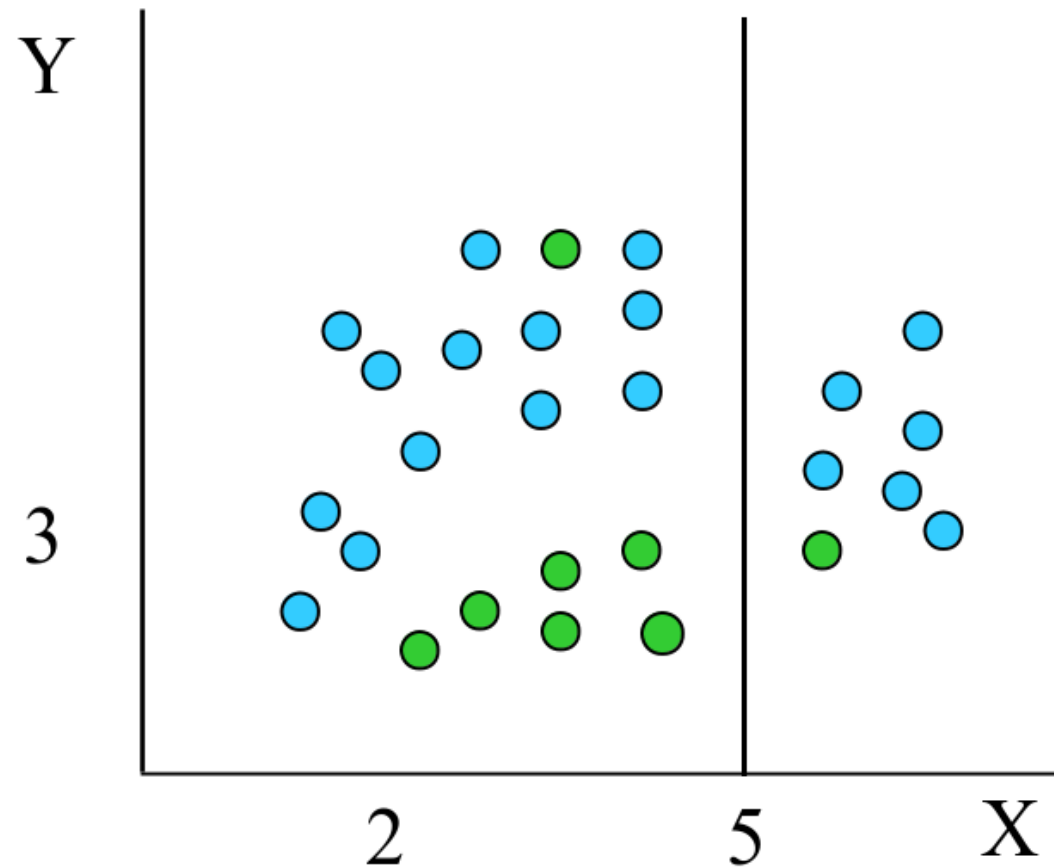
# Decision Trees



- Separate cases by drawing multiple lines, or: repeatedly make a choice.



# Decision Trees



## Decision Tree ( $d = 1$ )

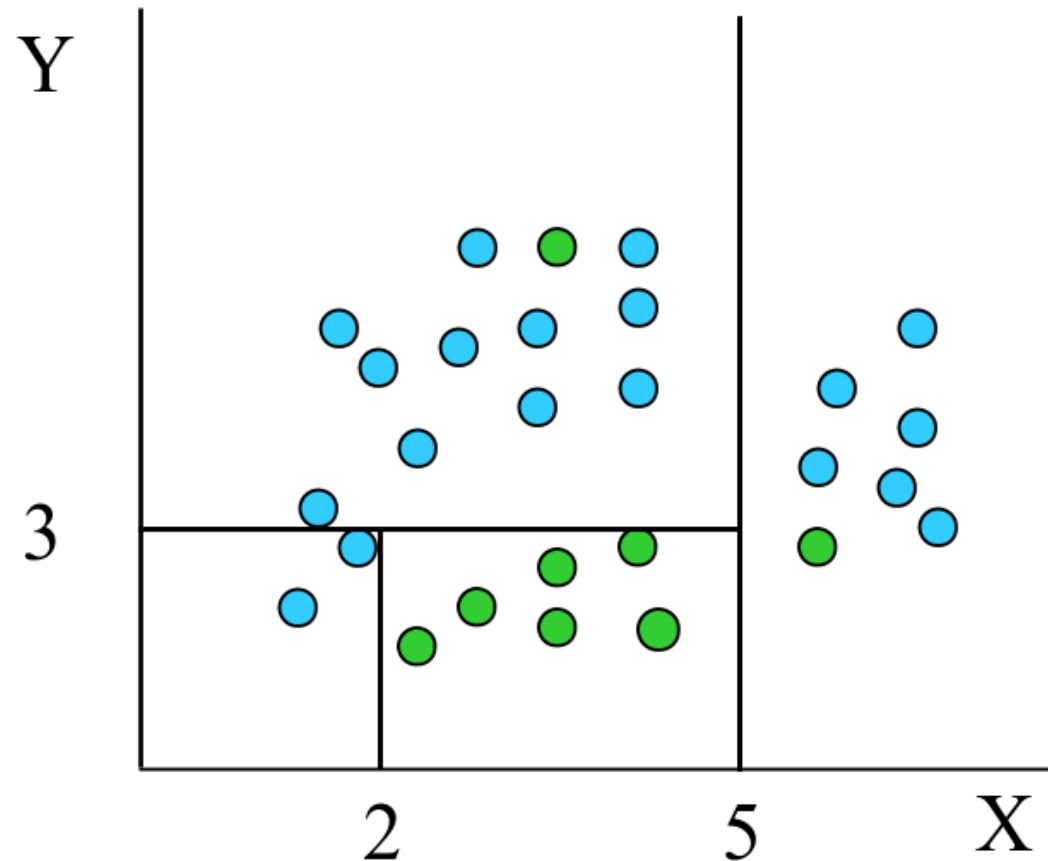
```
if (X > 5) return BLUE;  
else return GREEN; // oops!
```

# Decision Trees

## Decision Tree ( $d = 2$ )

```
if( $X > 5$ ) return BLUE;  
elseif( $Y > 3$ ) return BLUE;  
else return GREEN;
```

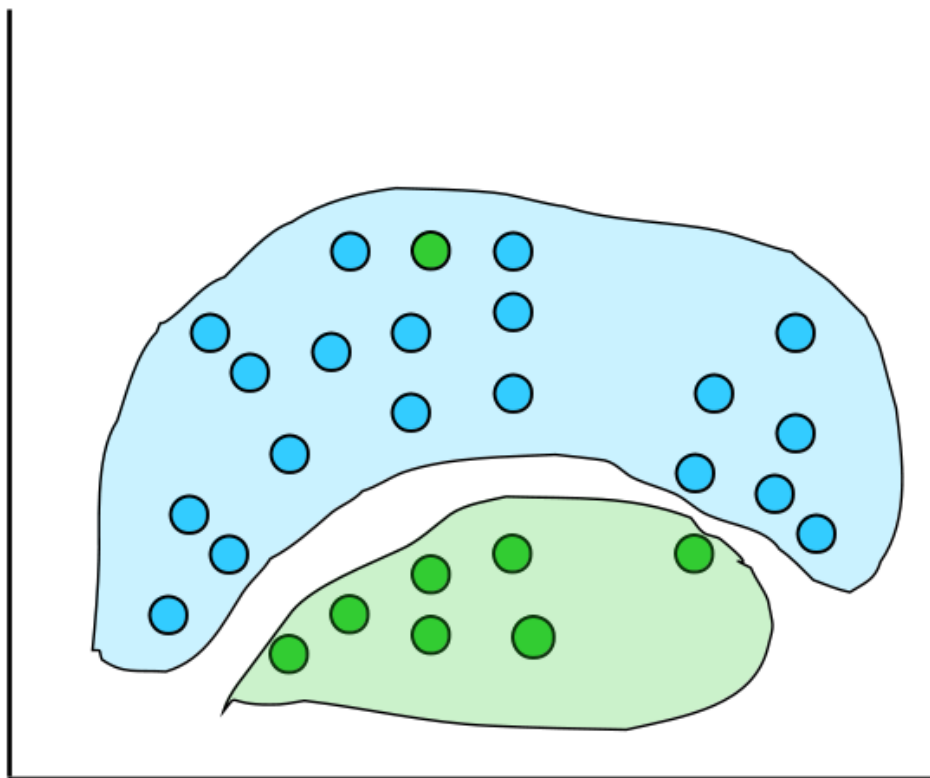
# Decision Trees



## Decision Tree ( $d = 3$ )

```
if(X > 5) return BLUE;  
elseif(Y > 3) return BLUE;  
elseif(X > 2) return GREEN;  
else return BLUE;
```

# Neural Networks



- **Neural Networks**
- Perceptrons
- Multi-level
- Backpropagation
- **Deep learning**

# Example: Cats

- Experiment at Google
  - 1000 machines
  - 16.000 cores
  - 200x200 pixel sample images
  - 10 million images
- 
- 70% accuracy (after **a lot**) of training





# Recommendation: Amazon

Louis, Welcome to Your Amazon.com (If you're not Louis Lazaris, click here.)

## Today's Recommendations For You

Here's a daily sample of items recommended for you. Click here to [see all recommendations](#)



[Algorithms in a Nutshell \(In...](#)  
(Paperback) by George Heineman  
★★★★☆ (5) \$31.49  
[Fix this recommendation](#)



[Simply JavaScript \(Paperback\)](#)  
by Kevin Yank  
★★★★☆ (19) \$26.37  
[Fix this recommendation](#)



[The Art & Science of JavaScript](#)  
(Paperback)  
★★★★☆ (19) \$26.37  
[Fix this recommendation](#)

[Any Category](#) [Business & Culture](#) [C](#) [Databases](#) [Drama](#) [Engineering](#)  
[Mystery & Suspense](#) [Networking](#) [Networks, Protocols](#)

# Recommendation: Netflix

A woman with curly hair is sitting on a couch, smiling while watching a laptop. In the background, there is a bookshelf filled with books.

**NETFLIX**

Sign In

**Watch TV shows &  
movies anytime,  
anywhere.**

Plans from EUR7.99 a month.

Start Your Free Month

# Example: Churn Prediction



# Example: Churn Prediction

- Master Computer Science project by P. Kusuma
- **Churn:** customer switching to competitor
- Dutch telecom provider
- 700 million call records
- 1.2 million customers
- Historical data
- Use (data mining) techniques to predict churn
- Class imbalance

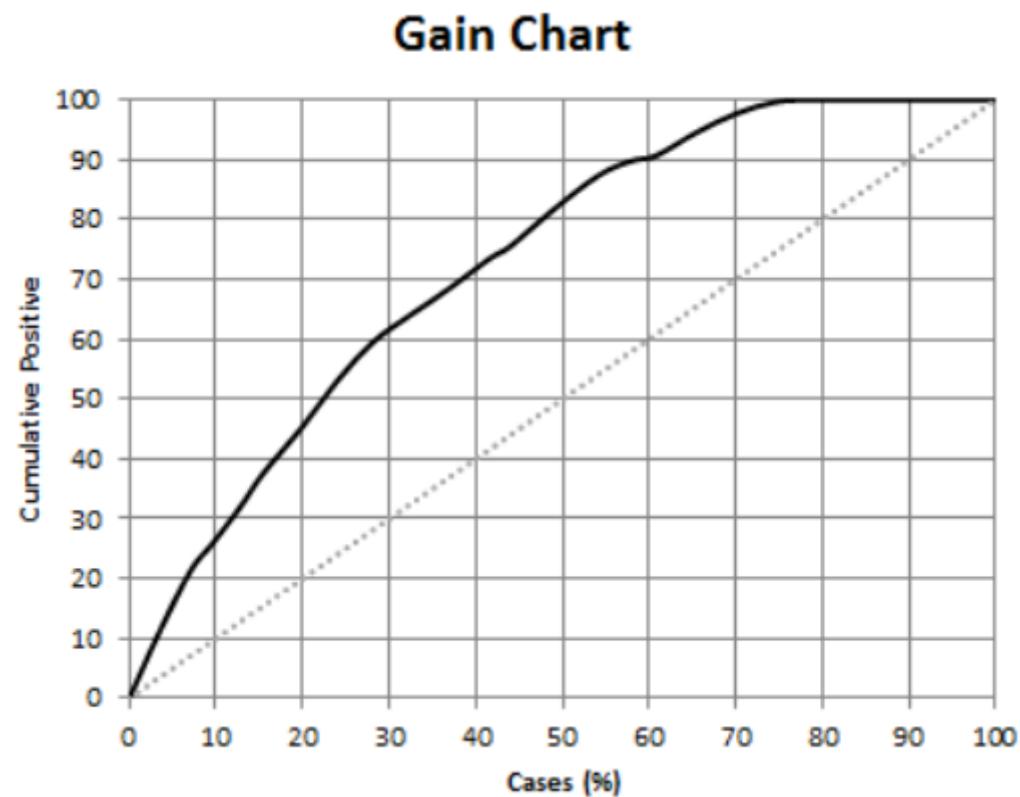


# Example: Churn Prediction

- Perform extra marketing targeted at high probability churners
- Use a set of customer features for classification:
  - Demographic characteristics (age)
  - Contractual information (type of subscription and package plan)
  - Handset information (handset model and manufacturer)
  - Service usage (voice call duration, SMS count and data usage)
  - Churn identification (churner or non-churner)
- Baseline: calling 1% of your customers will reach 1% of the total churners (assuming population uniformity)



# Churn Prediction Results





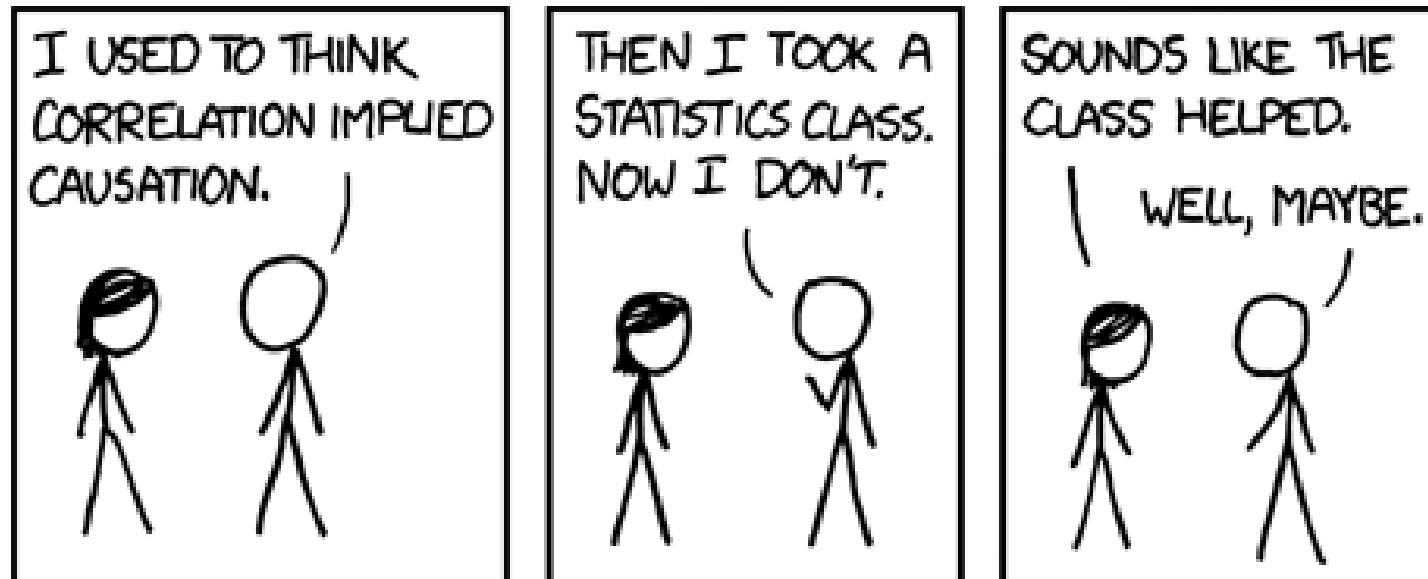
# Result Validation

- Domain experts needed
- Ground truth
- Correlation vs. causation
- Outlier or data error
- Manual inspection vs. numeric measures

# Remarks

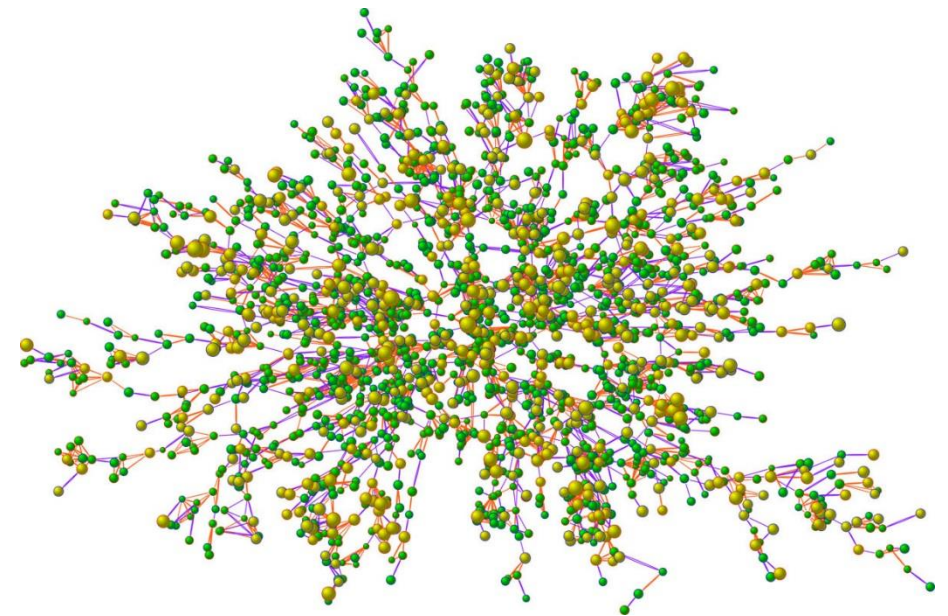
- Data is almost always biased
- Data mining may find regularities from history, but history is inherently not the same as the future
- “Patterns are not people”
- Association (correlation) does not dictate trend nor causality
- Some abnormal data (outlier) could be caused by human error
- Interpretation is crucial
- The P-word is screaming for attention . . .

# Correlation and Causation



# Network Science

- **Network:** real-world data based on interaction between objects
- (online) social networks, webgraphs, scientific collaboration citation networks, biological networks, economic trade networks, ...
- **Centrality analysis:** find important actors
- **Community detection:**  
find clusters of connected nodes
- **Complex systems:** system behavior can not be explained by looking at individual components, only by the whole.

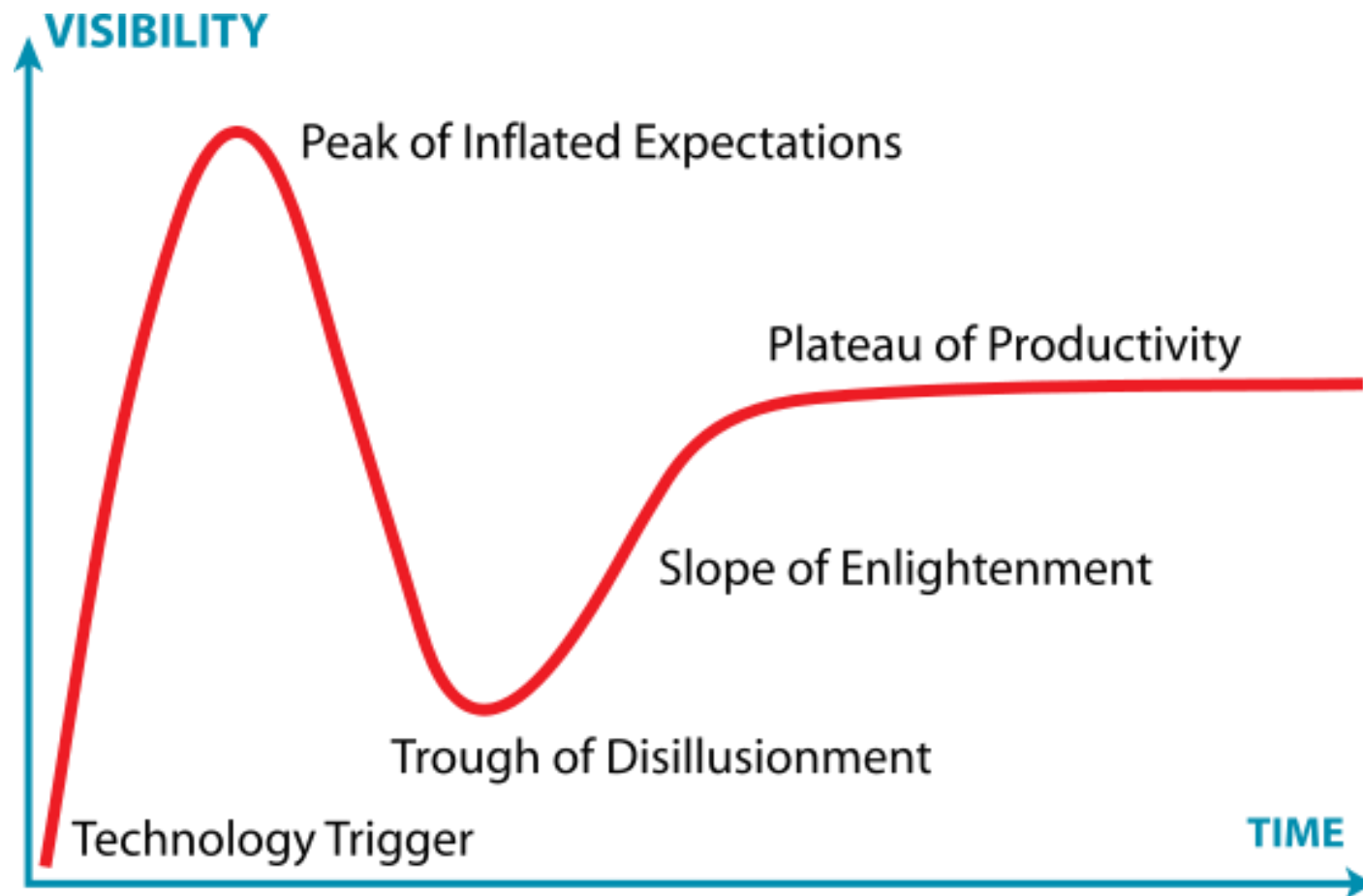


# Tooling

- WEKA, Orange, Rattle, SCAVis, RapidMiner
- Hadoop, MapReduce, Accumulo
- SPSS, SAS, KXEN
- For network science: Gephi, NetworKit, NetworkX
- Python: scikit-learn, pandas, matplotlib  
<http://scikit-learn.org/>



# Hype?





# Conclusions

- Data
  - Data
  - Data
- 
- There is value in data (analysis)
  - Big Data is a buzz-word used in the wrong context far too often
  - Knowledge discovery in data is a nontrivial process
  - The interpretation of data analysis results is crucial

## MY HOBBY: EXTRAPOLATING

