iMinds Dept. MEDICAL IT

KU Leuven ESAT-STADIUS

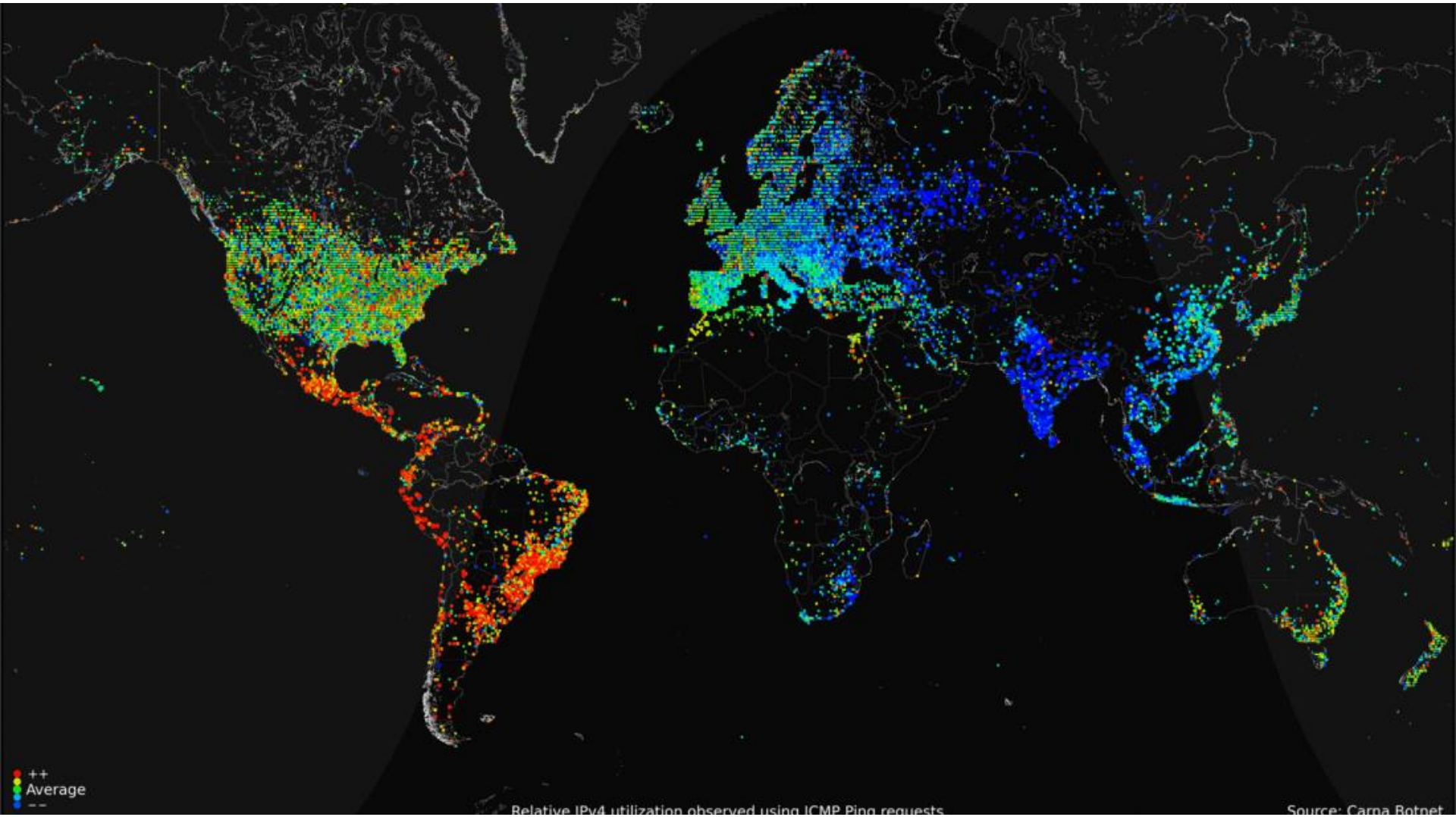
Serious Data, Serious Mining

Prof.Dr. Bart De Moor

Bart.DeMoor@iminds.be

- Big Data
  - What
  - Who
- Six dimensions
  - Data
  - Compute Infrastructure
  - Storage Infrastructure
  - Analytics
  - Visualization
  - Security & Privacy
- Machine learning as a commodity
- Expertise
  - Books & Spin-offs
  - Algorithms
  - Applications

www


++
Average
--
Relative IPv4 utilization observed using ICMP Ping requests
Source: Carna Botnet

Grains of rice the world consumes annually: **27.5 quadrillion**

Amount of data the world consumes every 30 minutes: **40.4 petabytes**

We consume more bytes on the internet in 30 minutes than grains of rice in a year.

1 million = 1 000 000
1 billion = 1 000 000 000
1 trillion = 1 000 000 000 000
1 quadrillion =
1 000 000 000 000 000

1 kB = 1 000
1 MB = 1 000 000
1 GB = 1 000 000 000
1 TB = 1 000 000 000 000
1 PB = 1 000 000 000 000 000

1 TB
= large university library
= 212 DVD discs
= 1430 CDs
= 3 year music in CD quality

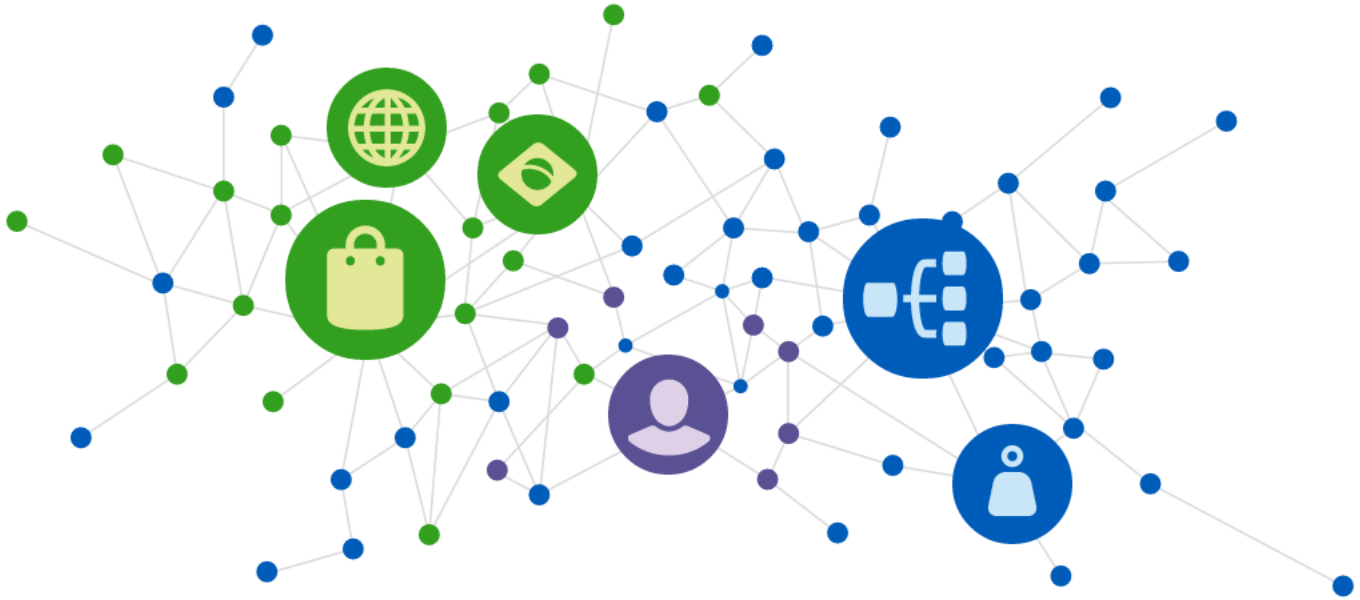
The Industrial Internet, a connected network of intelligent machines working the way they are intended, will transform business as dramatically as the consumer Internet has changed our lives.

The Industrial Internet

12 500 000 000

Connected Devices

Devices Per Capita Worldwide



Past

2020+

2000 - 2010: there was rapid growth in connectivity and its transformative impact on the world have laid the foundation for the Industrial Internet.

The Industrial Internet

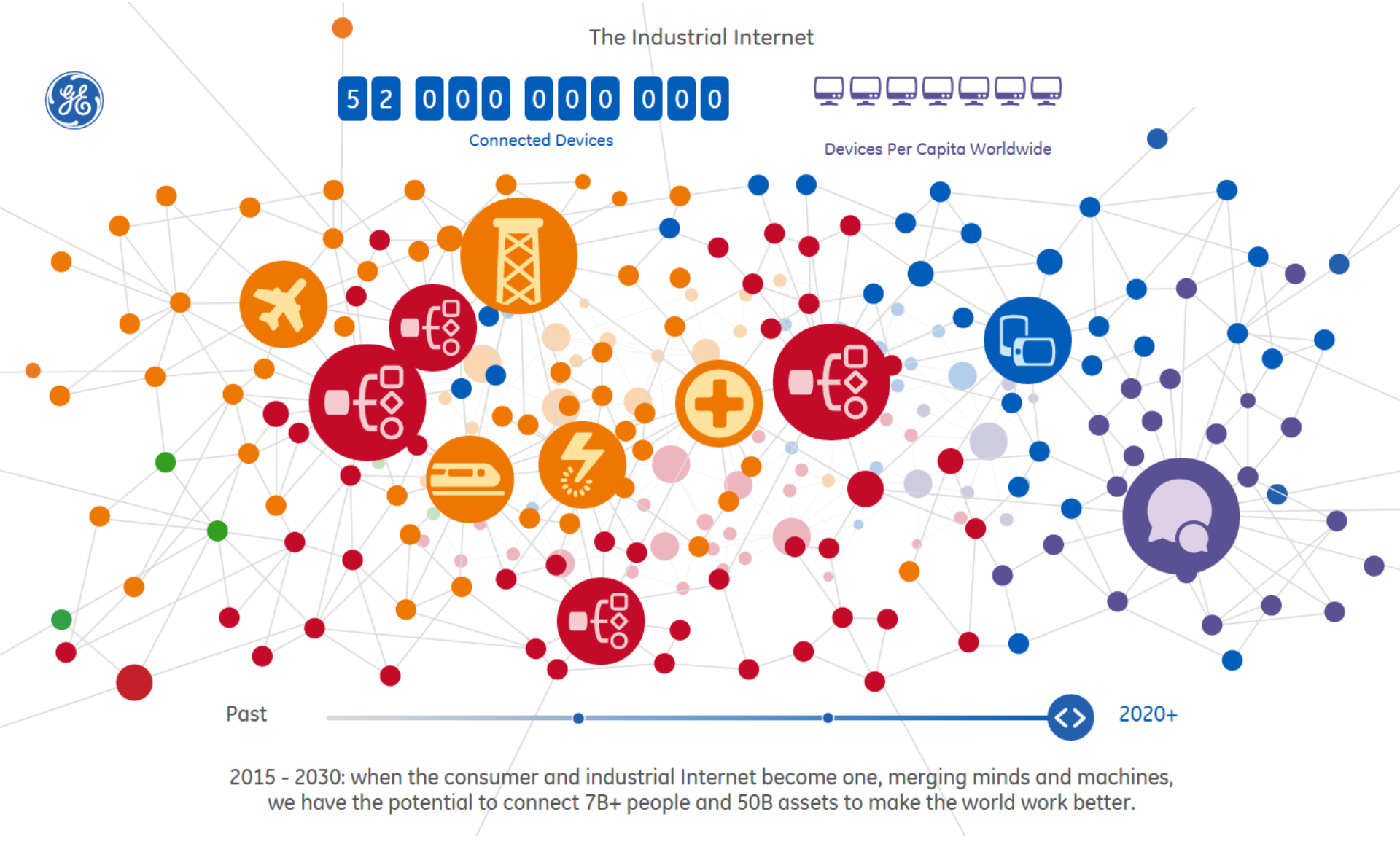
25 000 000 000

Connected Devices

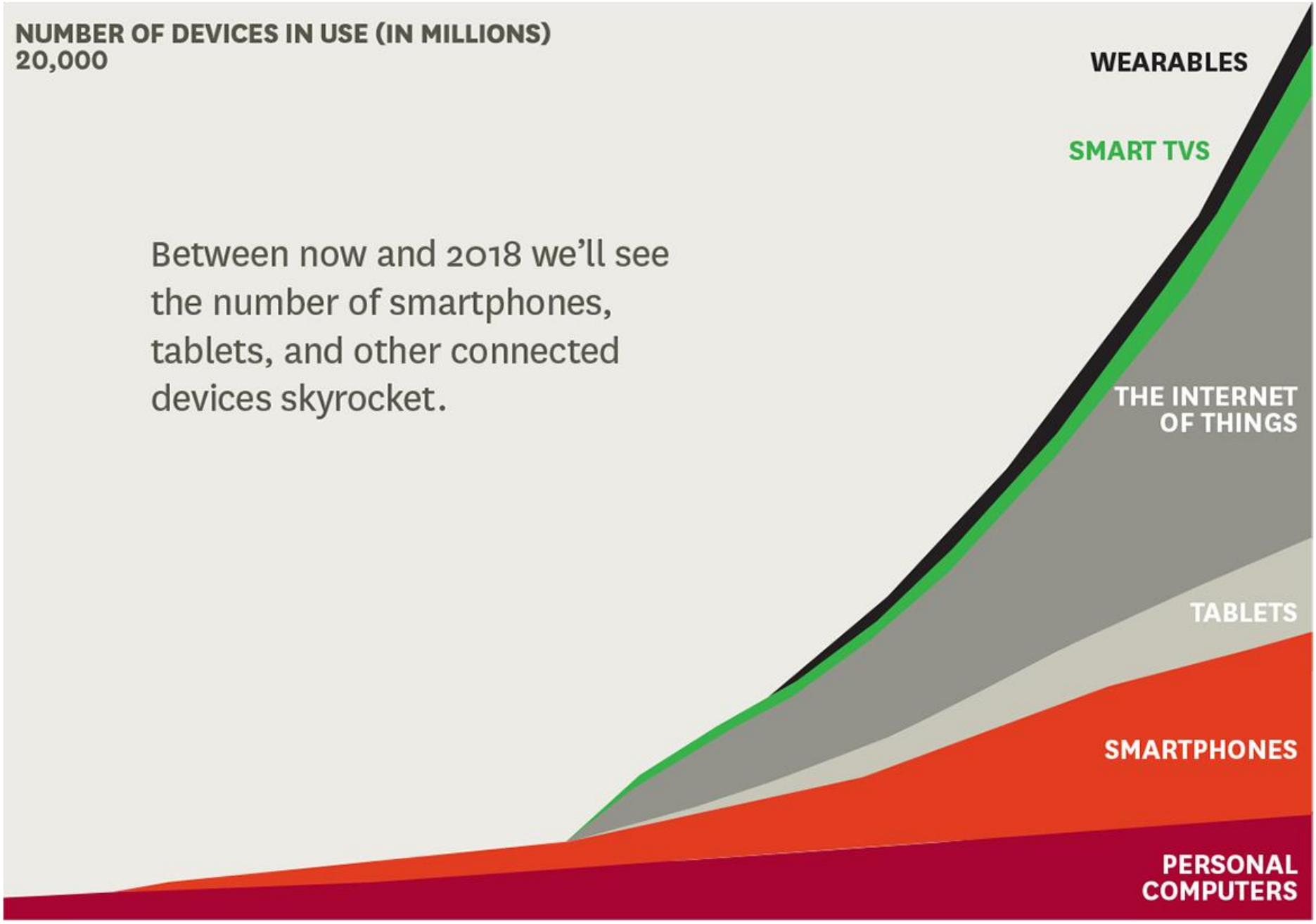
Devices Per Capita Worldwide

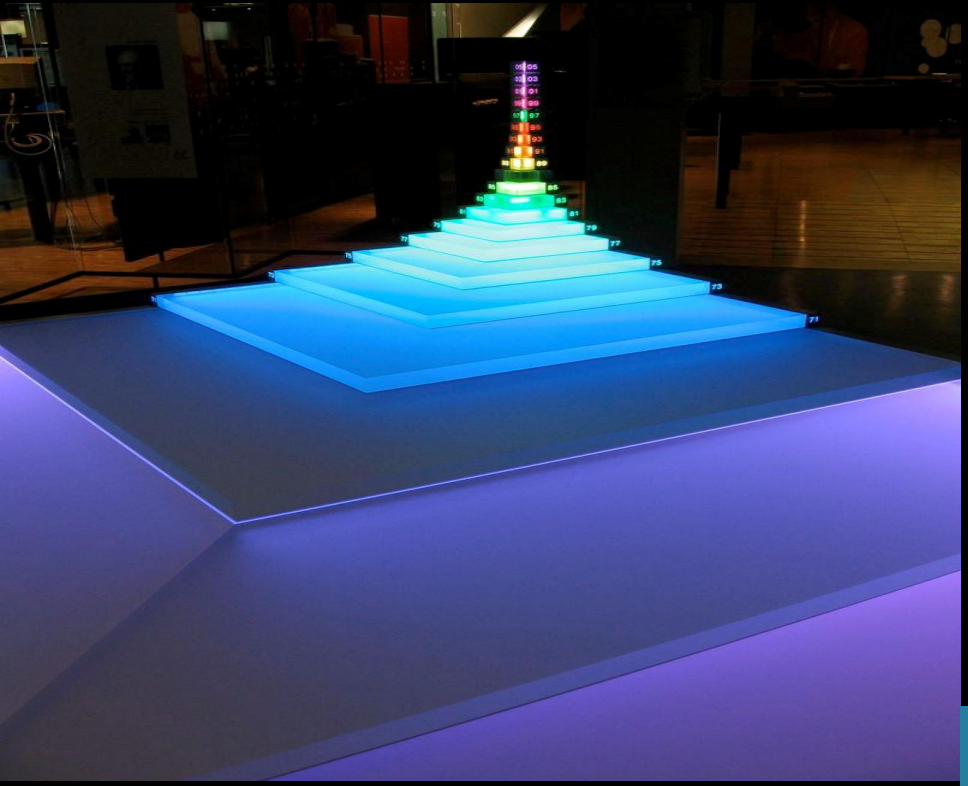
Past — 2020+

2010 - 2015: building on the foundation of the consumer Internet, the Industrial Internet is transforming business today.



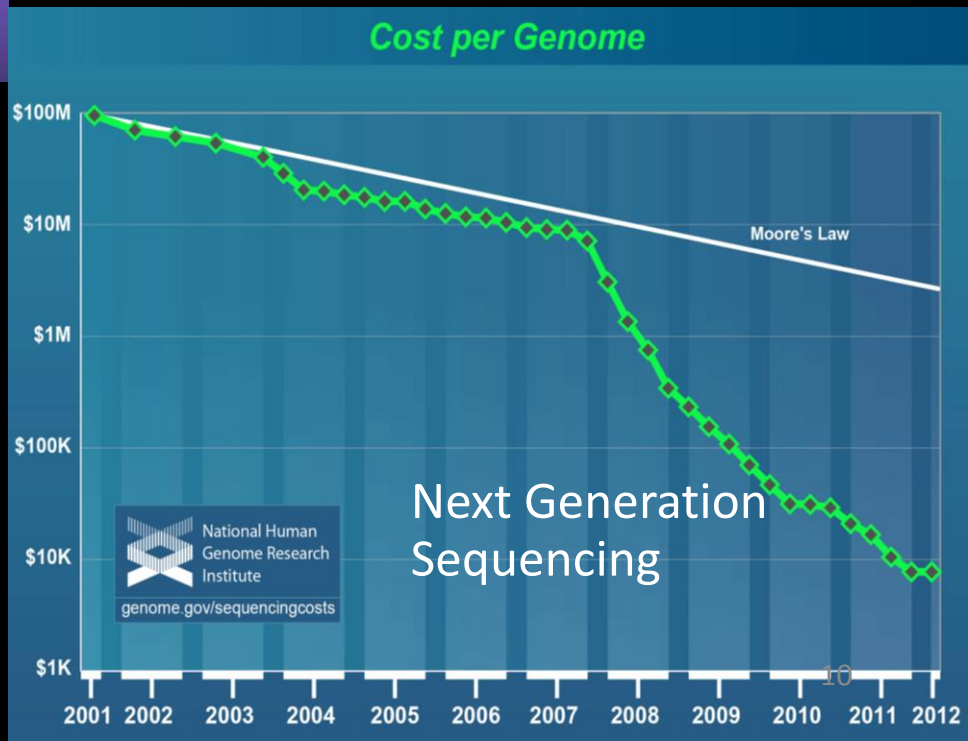
2015 - 2030: when the consumer and industrial Internet become one, merging minds and machines, we have the potential to connect 7B+ people and 50B assets to make the world work better.


NUMBER OF DEVICES IN USE (IN MILLIONS)
20,000
Between now and 2018 we'll see the number of smartphones, tablets, and other connected devices skyrocket.
WEARABLES
SMART TVS
THE INTERNET OF THINGS
TABLETS
SMARTPHONES
PERSONAL COMPUTERS
2004
2018



Moore's law:
computing power
doubles
every 18 months

Carlson's law:
complexity/cost
evolves
exponentially



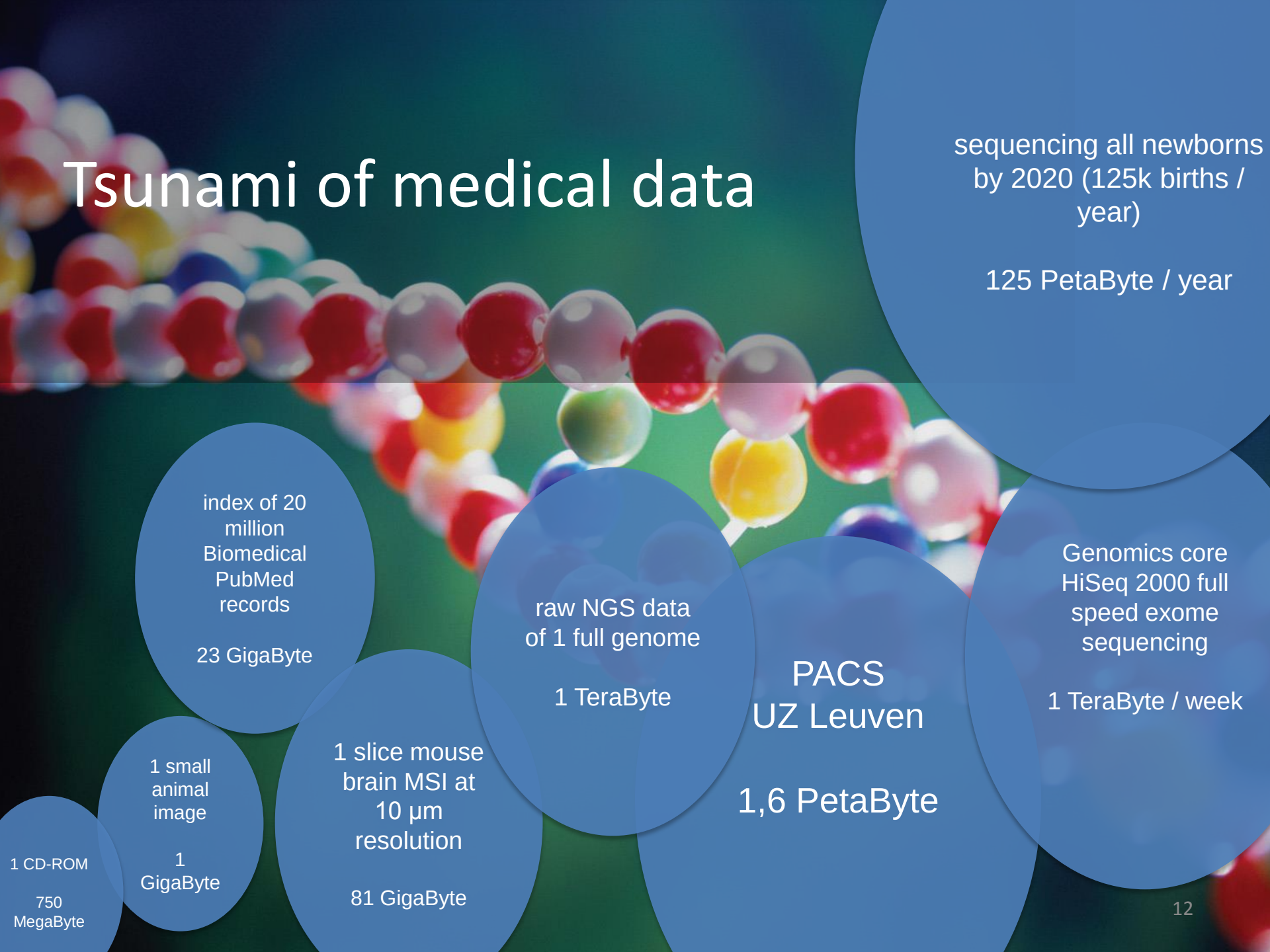
Genome data

GS-FLX Roche
Applied Science 454

Sequencers

- Human genome project (2003)
  - 13 year project
  - $300 million value with 2002 technology
- Personal genome (2007)
  - Genome of James Watson, 2 months
  - $1 000 000
- €1000-genome
  - Expected 2012-2020

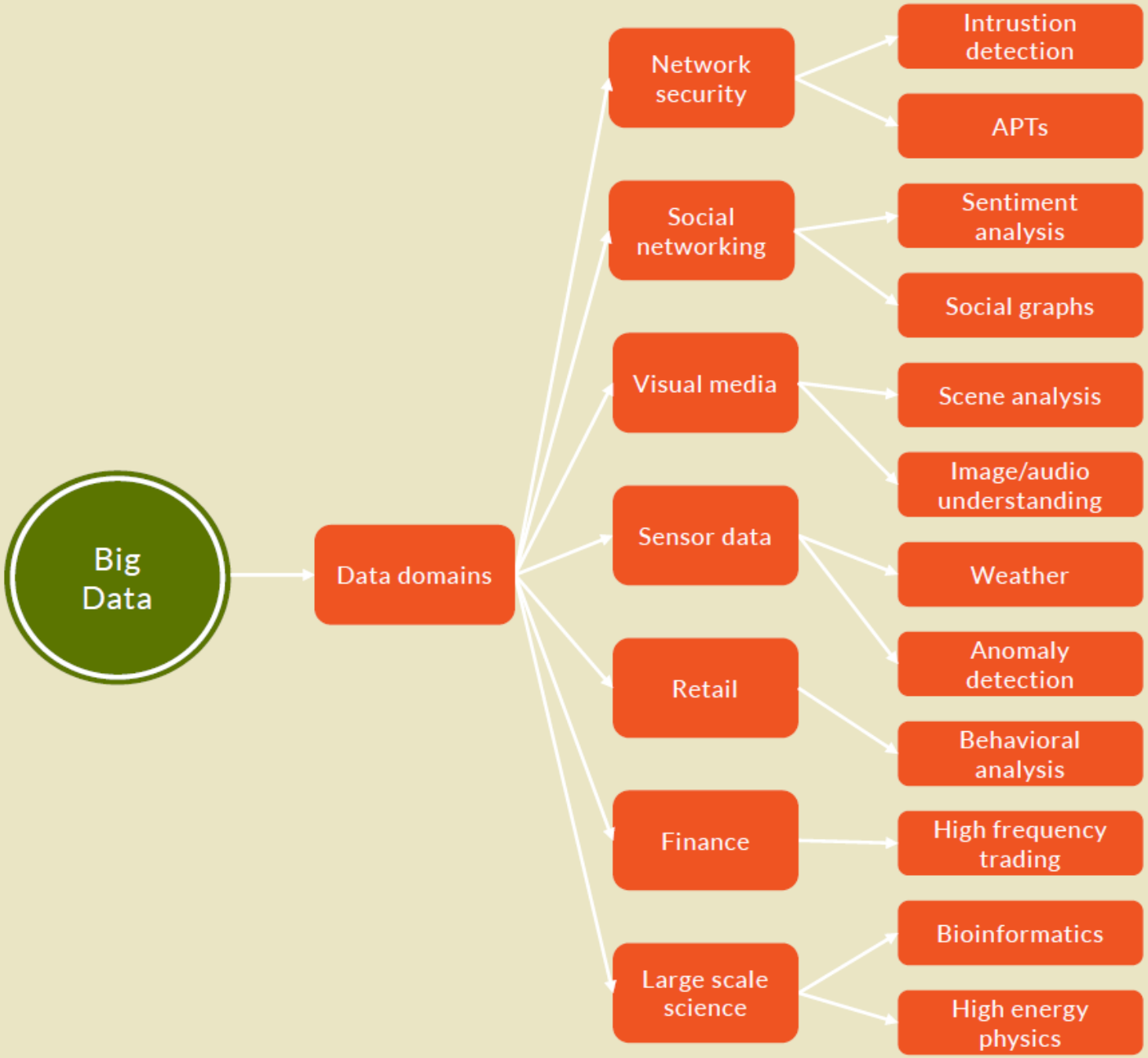
Tsunami of medical data



Data explosion in finance

Growing ~ 30-50% every year, half of this is unstructured!

Data storage became 30% cheaper, yet budgets for data storage are still rising.


Big Data
Data domains
Network security
Intrustion detection
APTs
Social networking
Sentiment analysis
Social graphs
Visual media
Scene analysis
Image/audio understanding
Sensor data
Weather
Anomaly detection
Retail
Behavioral analysis
Finance
High frequency trading
Large scale science
Bioinformatics
High energy physics

The Fourth Paradigm

Paradigm	Time Ago	Method
First	A millenium	Empirical
Second	A few centuries	Theoretical
Third	A few decades	Computational
Fourth	Today	Data-driven

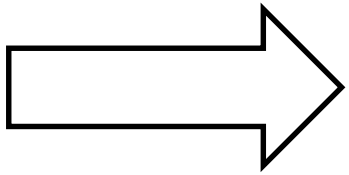
Evolution

From

I have a hypothesis

⇩

I need data to check it



To

I have data

⇩

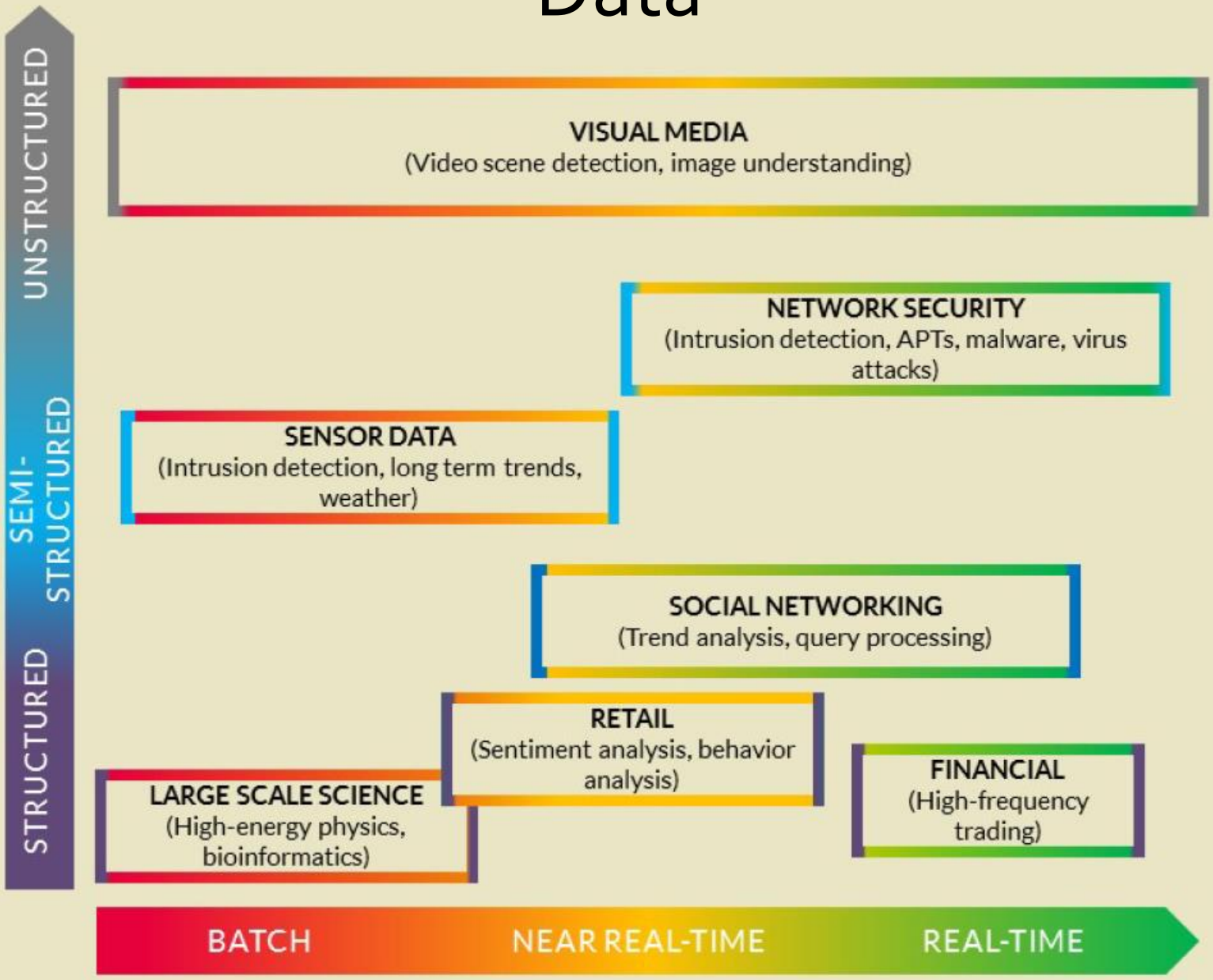
Which hypotheses can I check?

⇨ Data first!

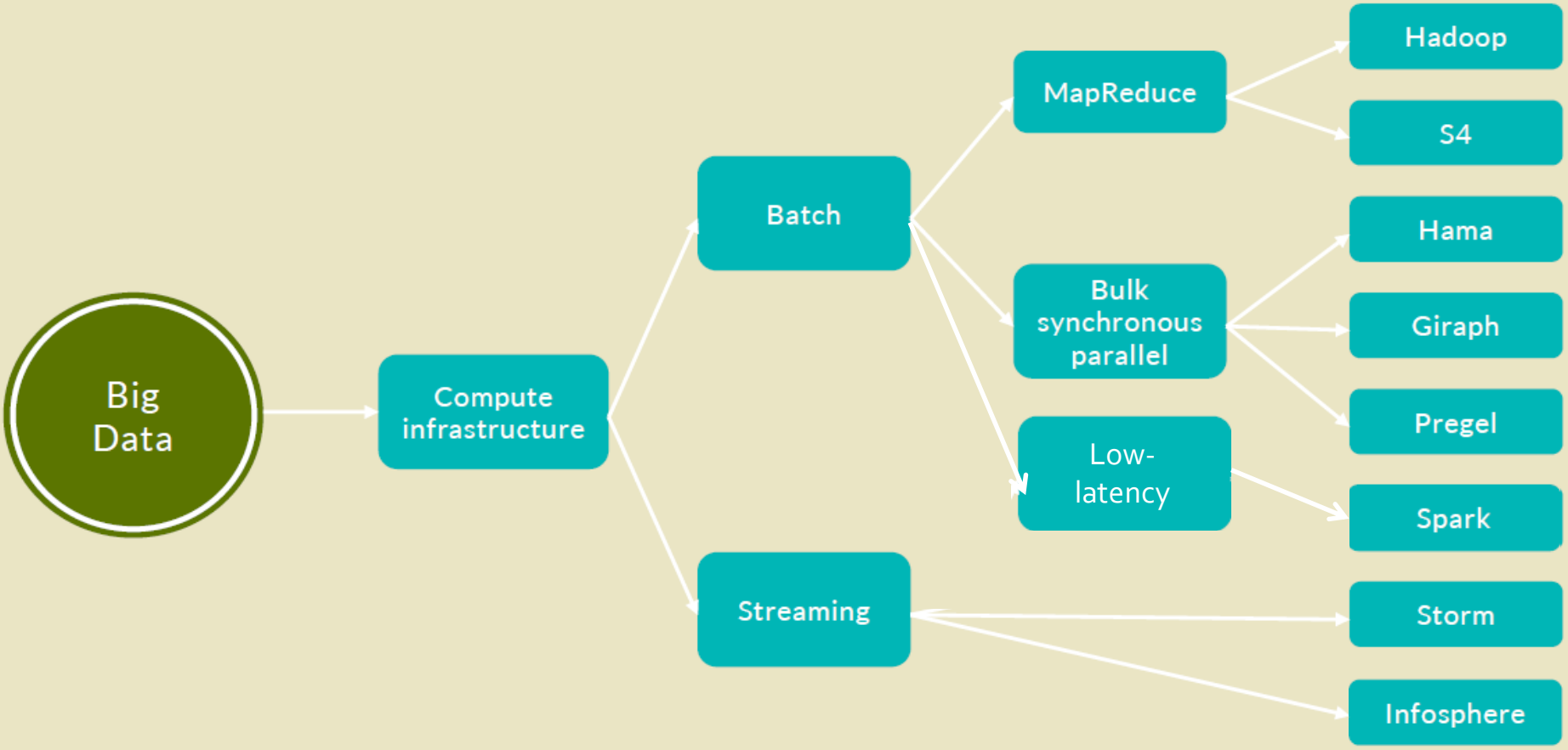
- Big Data
  - What
  - Who
- Six dimensions
  - Data
  - Compute Infrastructure
  - Storage Infrastructure
  - Analytics
  - Visualization
  - Security & Privacy
- Machine learning as a commodity
- Expertise
  - Books & Spin-offs
  - Algorithms
  - Applications


1
Data
2
Compute Infrastructure
3
Storage Infrastructure
4
Analytics
5
Visualization
6
Security & Privacy
Big Data

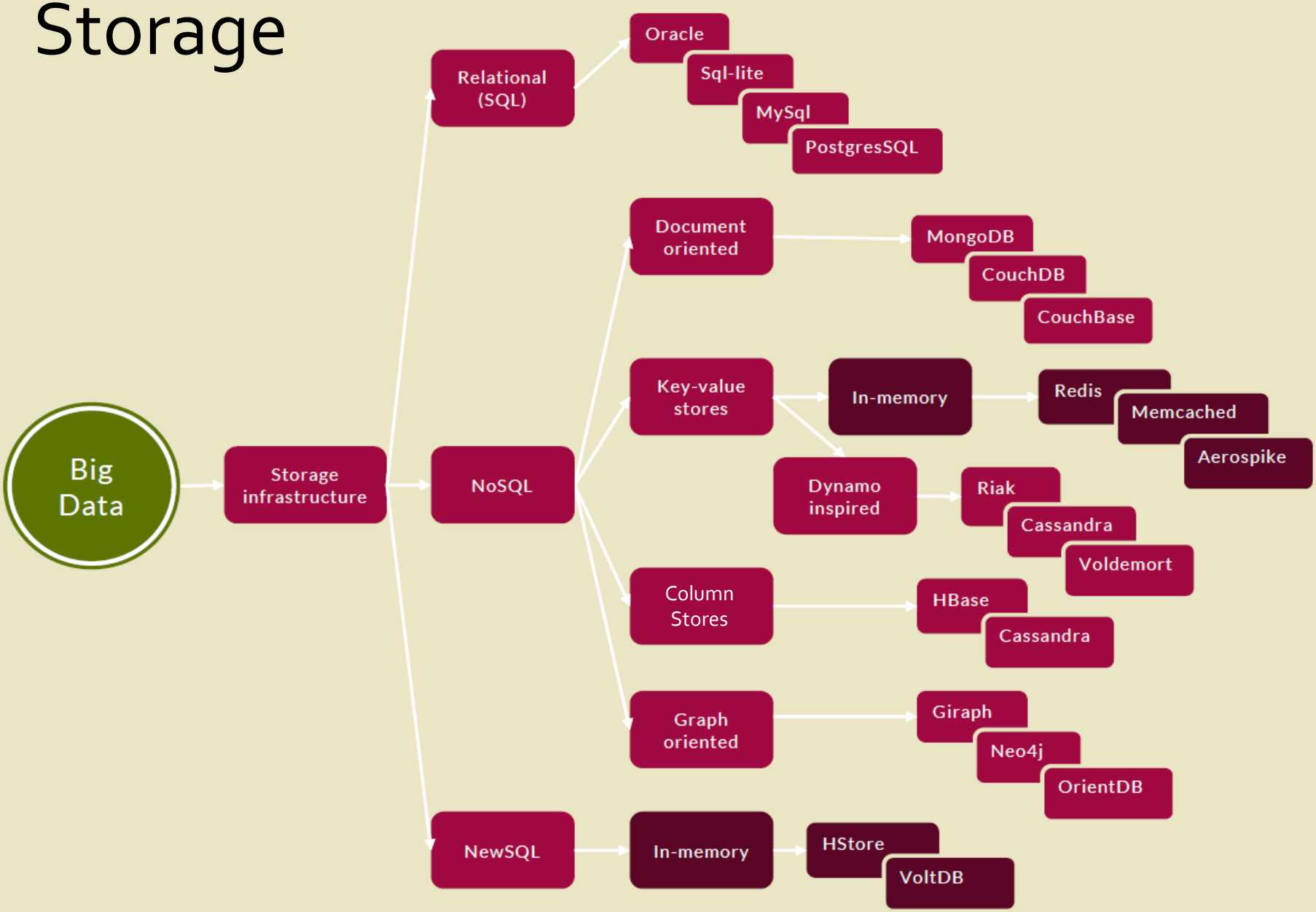
Data

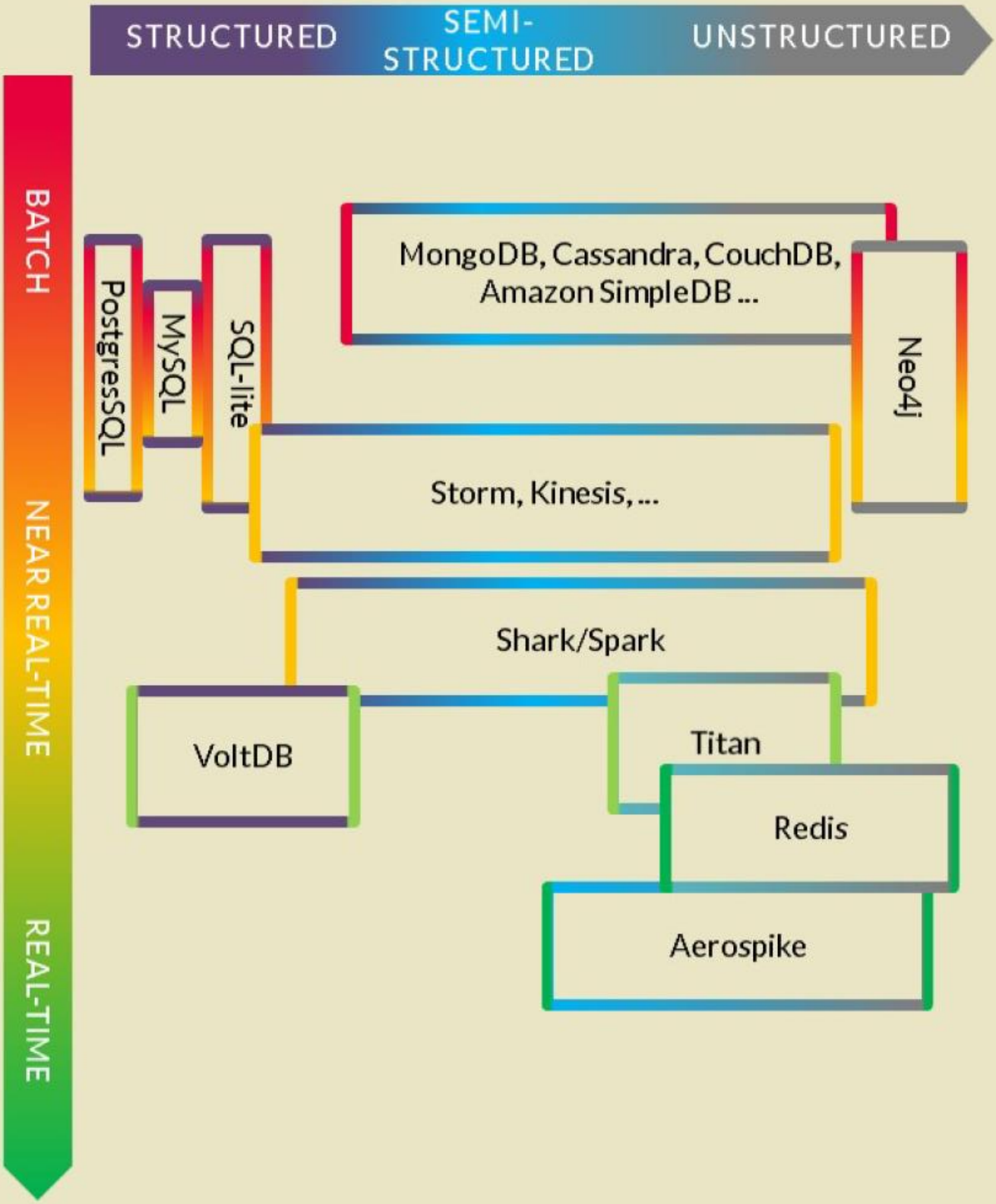

UNSTRUCTURED
SEMI-STRUCTURED
STRUCTURED
VISUAL MEDIA
(Video scene detection, image understanding)
NETWORK SECURITY
(Intrusion detection, APTs, malware, virus attacks)
SENSOR DATA
(Intrusion detection, long term trends, weather)
SOCIAL NETWORKING
(Trend analysis, query processing)
RETAIL
(Sentiment analysis, behavior analysis)
FINANCIAL
(High-frequency trading)
LARGE SCALE SCIENCE
(High-energy physics, bioinformatics)
BATCH
NEAR REAL-TIME
REAL-TIME

Compute infrastructure

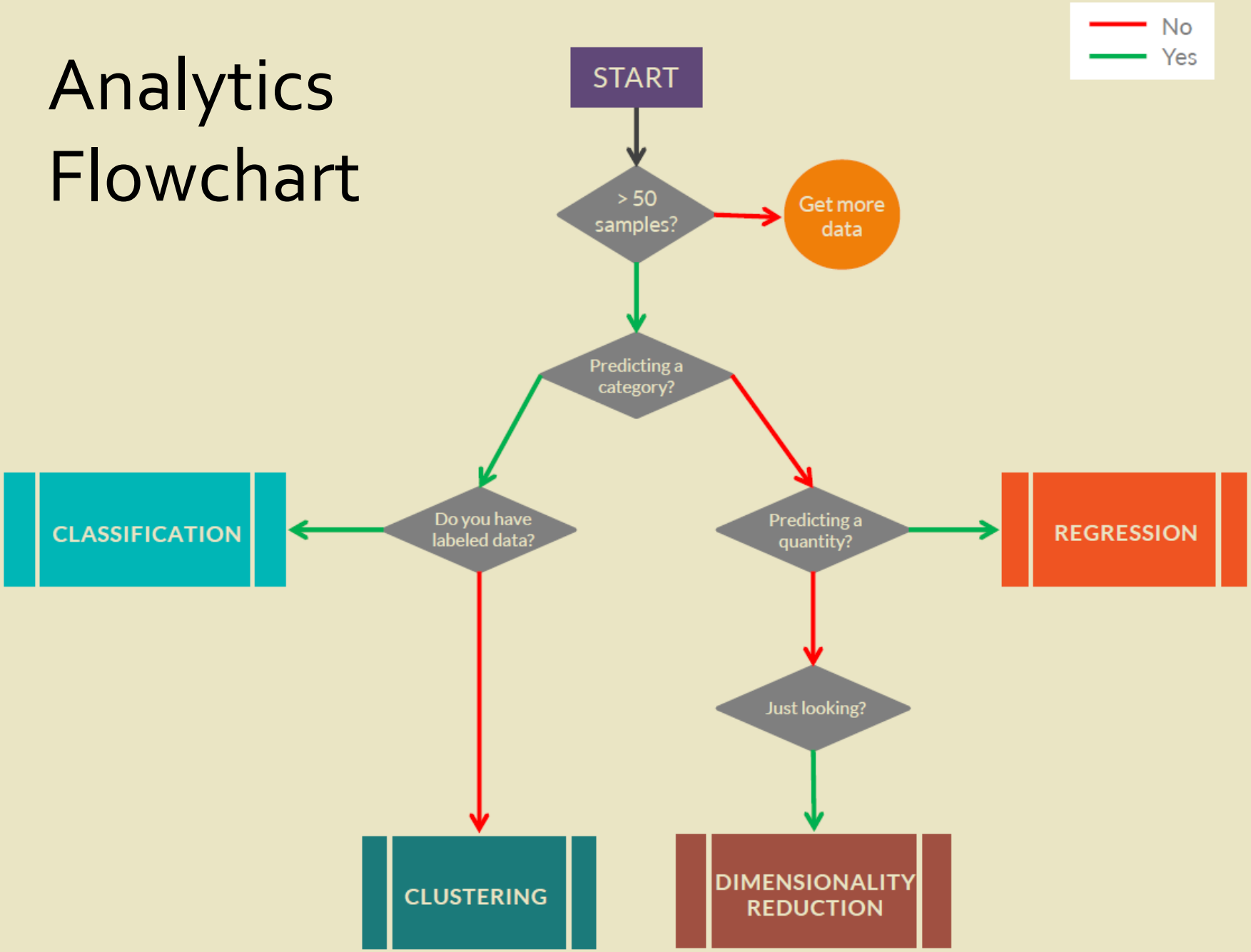


Storage

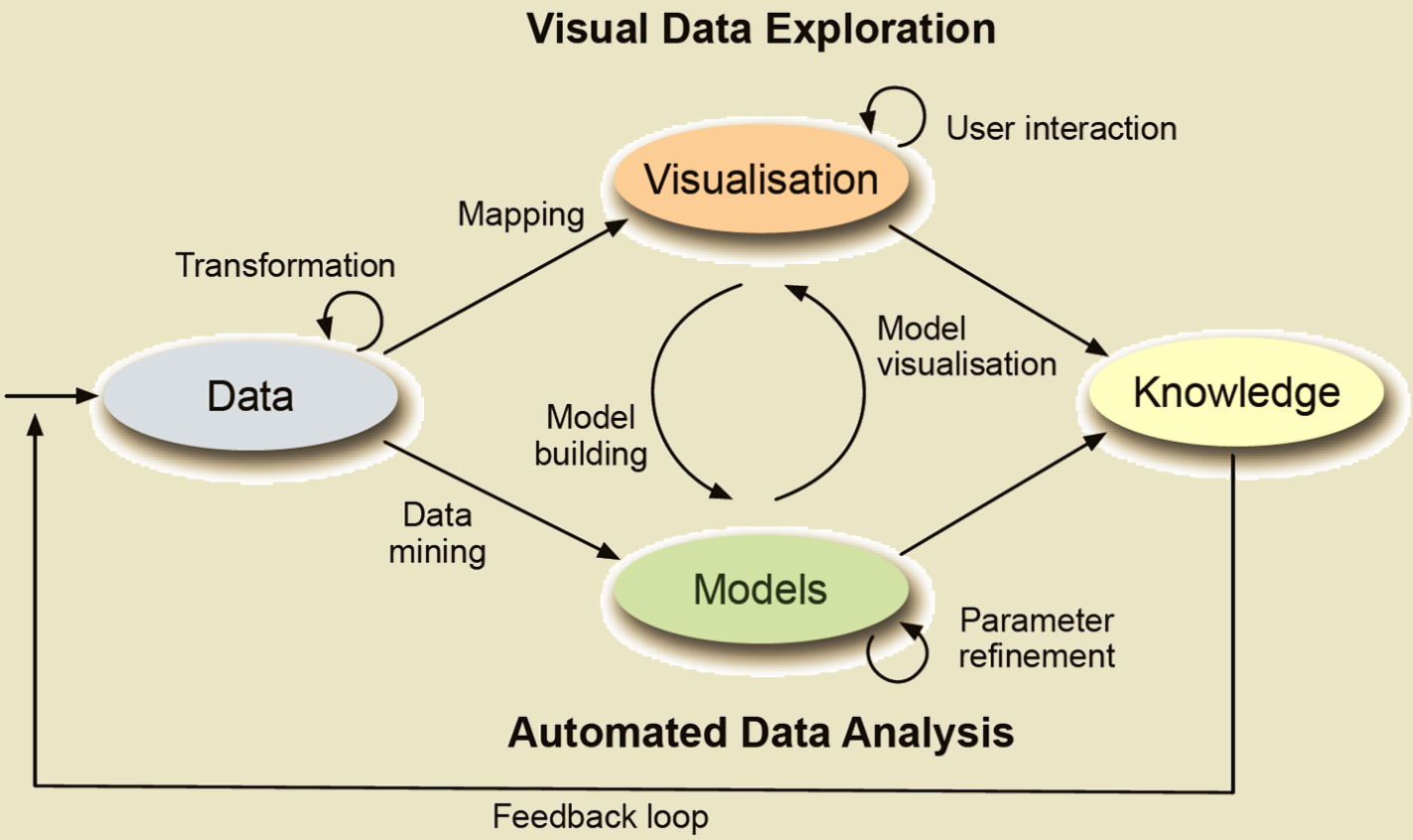

Big Data
Storage infrastructure
Relational (SQL)
Oracle
Sql-lite
MySql
PostgresSQL
NoSQL
Document oriented
MongoDB
CouchDB
CouchBase
Key-value stores
In-memory
Redis
Memcached
Aerospike
Dynamo inspired
Riak
Cassandra
Voldemort
Column Stores
HBase
Cassandra
Graph oriented
Giraph
Neo4j
OrientDB
NewSQL
In-memory
HStore
VoltDB


STRUCTURED
SEMI-STRUCTURED
UNSTRUCTURED
BATCH
NEAR REAL-TIME
REAL-TIME
PostgresSQL
MySQL
SQL-lite
MongoDB, Cassandra, CouchDB, Amazon SimpleDB ...
Neo4j
Storm, Kinesis, ...
Shark/Spark
VoltDB
Titan
Redis
Aerospike

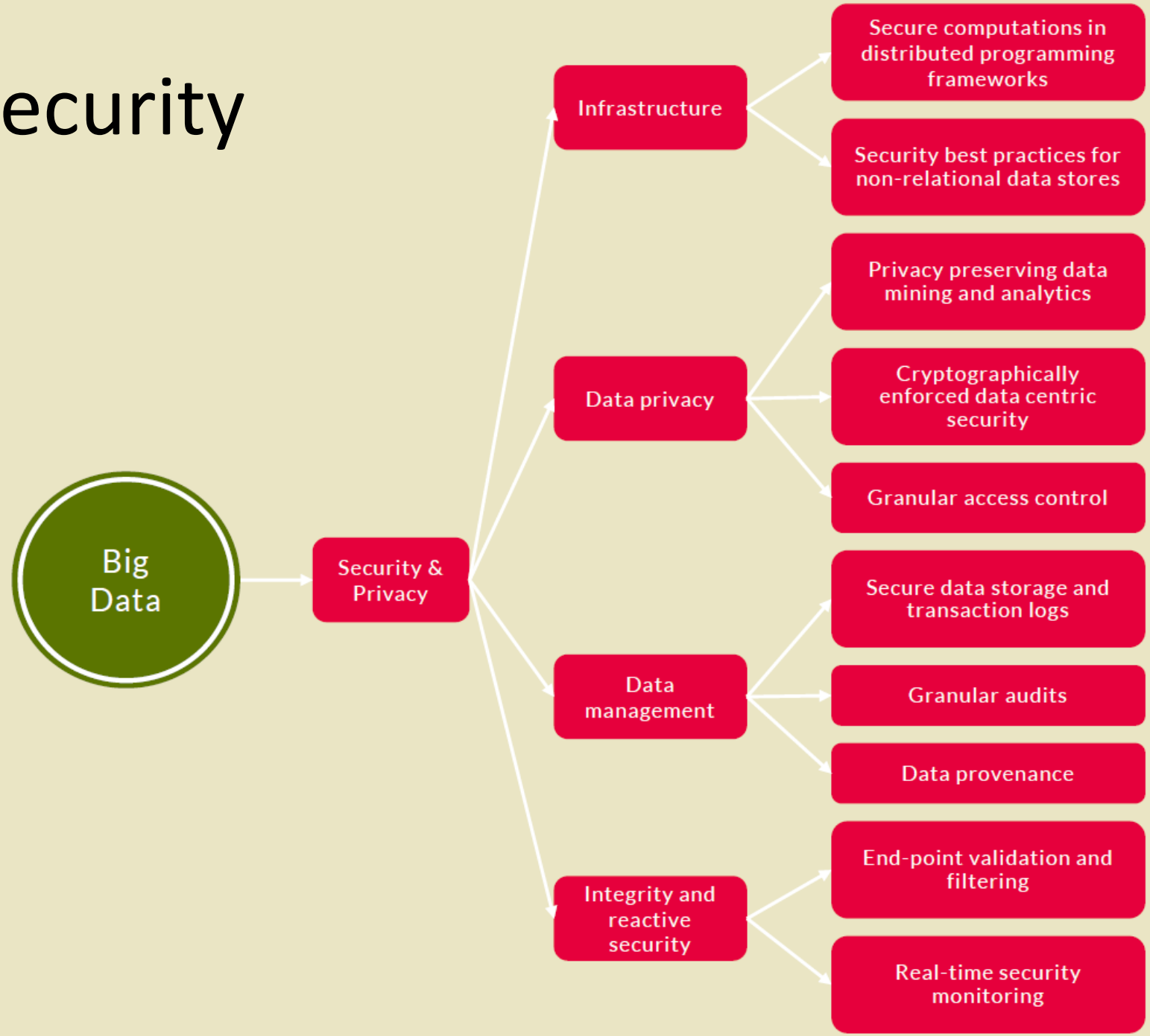
Analytics Flowchart



Visualization

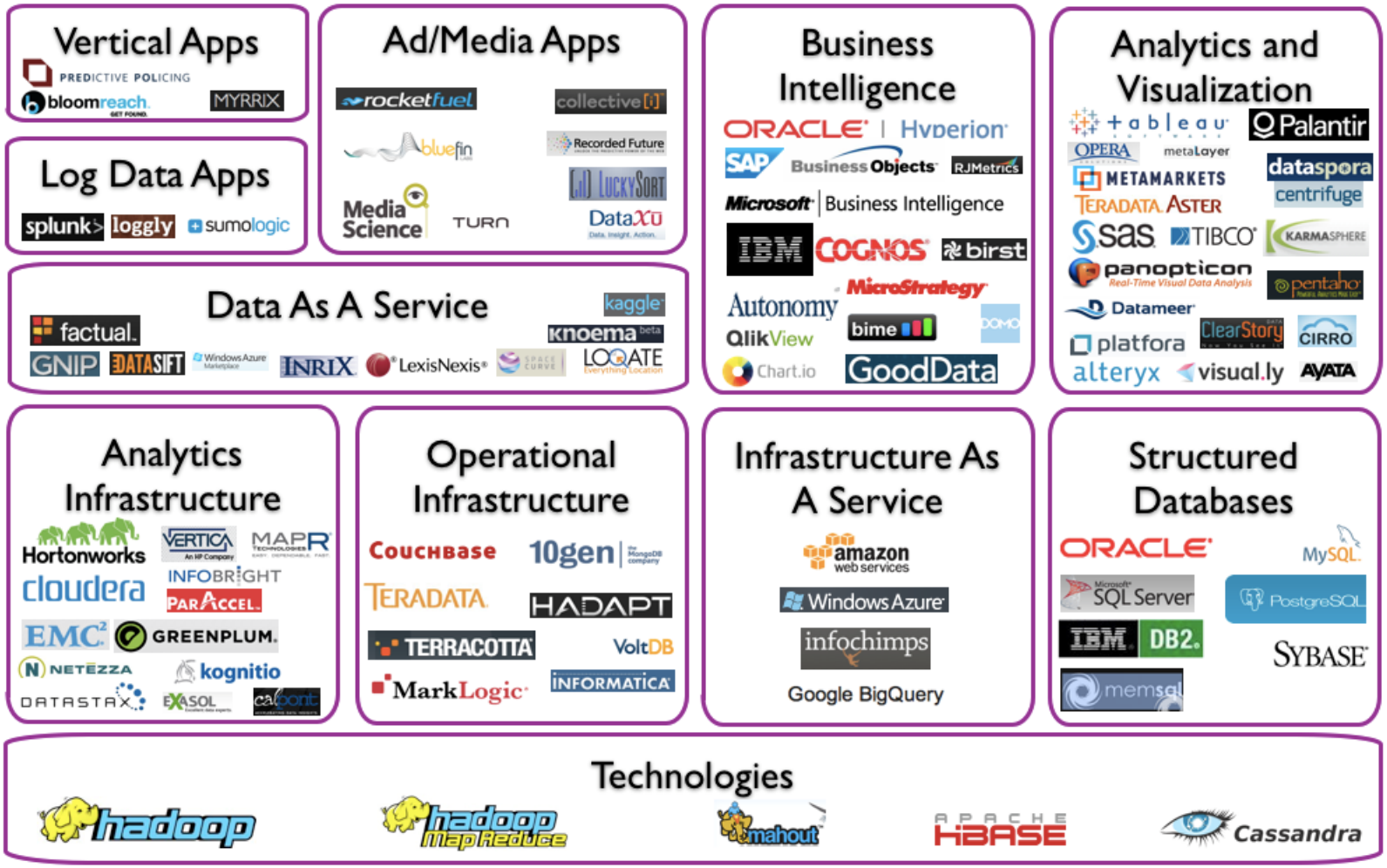


Security


Big Data
Security & Privacy
Infrastructure
Secure computations in distributed programming frameworks
Security best practices for non-relational data stores
Data privacy
Privacy preserving data mining and analytics
Cryptographically enforced data centric security
Granular access control
Data management
Secure data storage and transaction logs
Granular audits
Data provenance
Integrity and reactive security
End-point validation and filtering
Real-time security monitoring

- Big Data
  - What
  - Who
- Six dimensions
  - Data
  - Compute Infrastructure
  - Storage Infrastructure
  - Analytics
  - Visualization
  - Security & Privacy
- Machine learning as a commodity
- Expertise
  - Books & Spin-offs
  - Algorithms
  - Applications

Big Data Landscape



dave@vcdave.com

blogs.forbes.com/davefeinleib

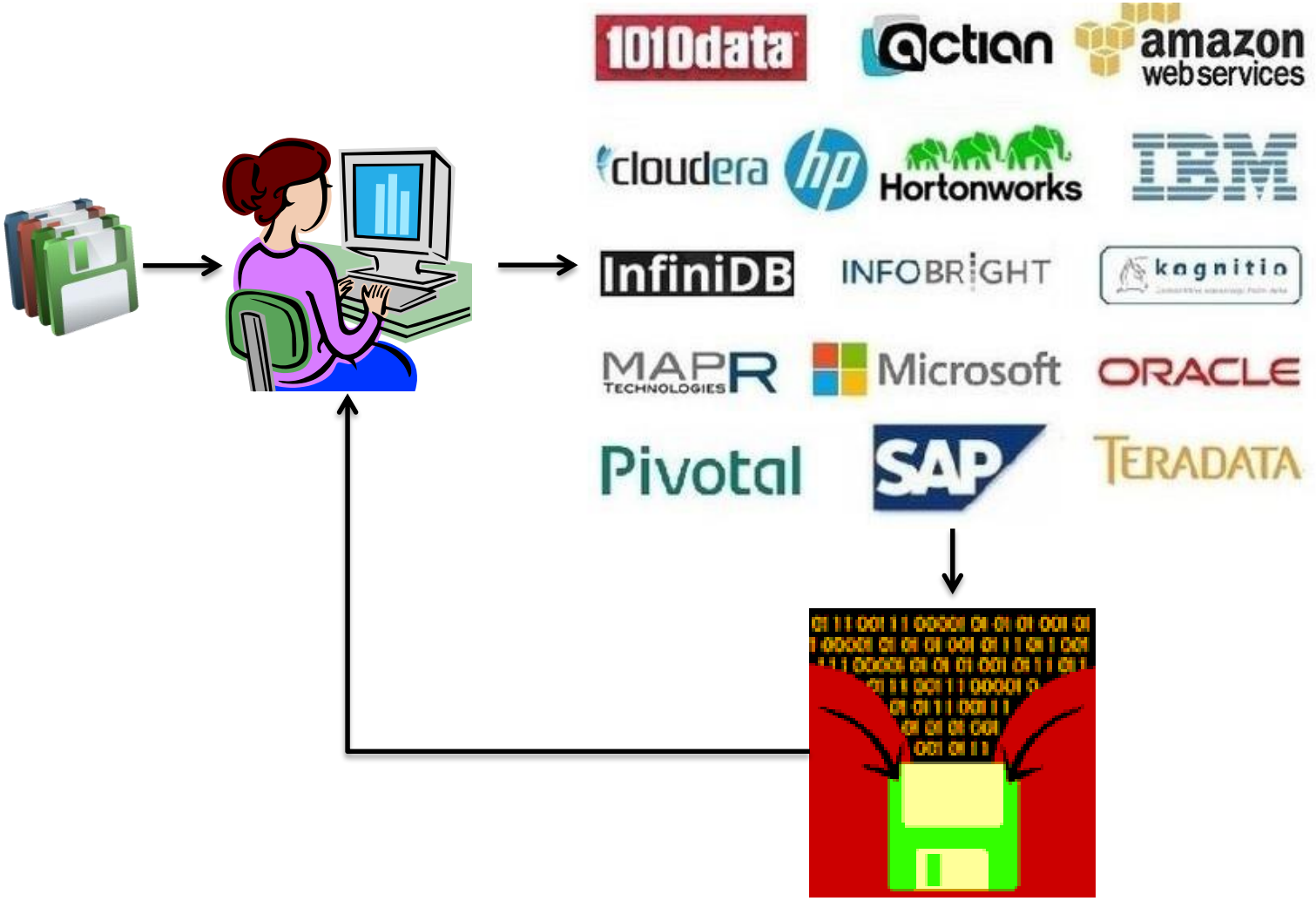
Big Data Landscape (Version 2.0)

© Matt Turck (@mattturck) and ShivonZilis (@shivonz) Bloomberg Ventures

Big Data Landscape

More and more analytics as a commodity!

Machine Learning as a commodity



Big Data Landscape

Many possible applications!

Energy

Industry

Environment

Social networks

Fraud and predictive analysis

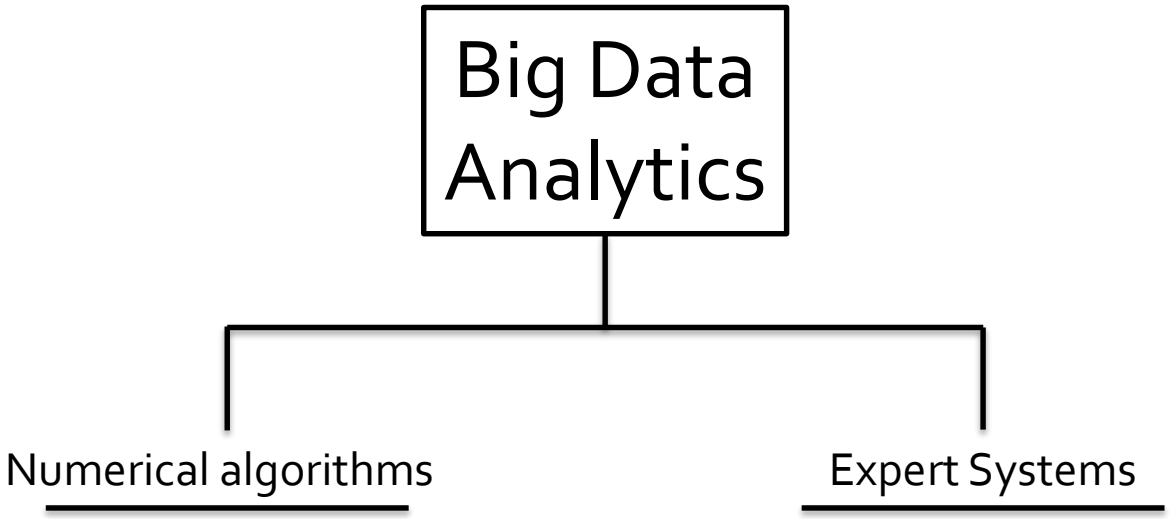
Health

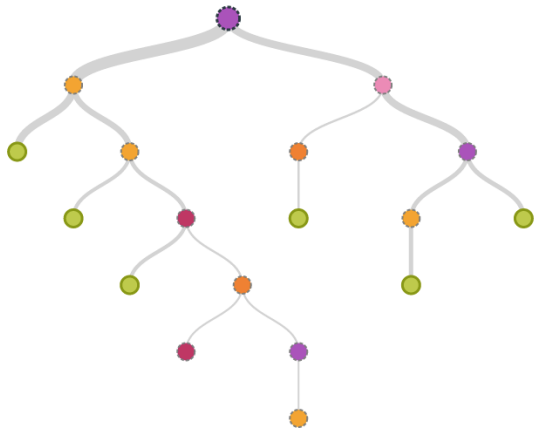
…

⇒ Focus on Serious Big Data

- Big Data
  - What
  - Who
- Six dimensions
  - Data
  - Compute Infrastructure
  - Storage Infrastructure
  - Analytics
  - Visualization
  - Security & Privacy
- Machine learning as a commodity
- Expertise
  - Books & Spin-offs
  - Algorithms
  - Applications

Analytics

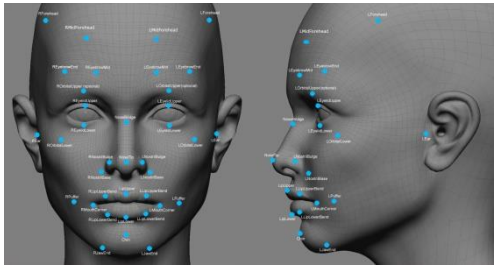




Objectives - ICT

Communication networks

Facial recognition



Home automation

Digital signing

Data center optimization

Objectives - Finance

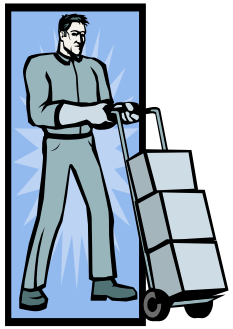
Fraud detection

Credit worthiness

Risk assessment

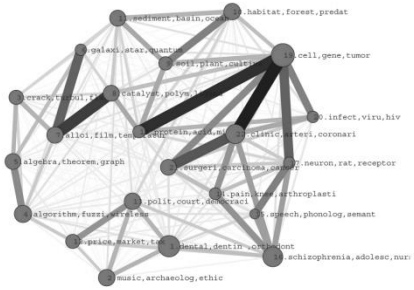
Portfolio management

Just-in-time production

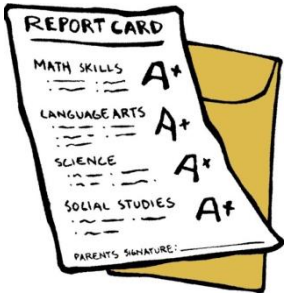


Objectives - Education

Scientometrics

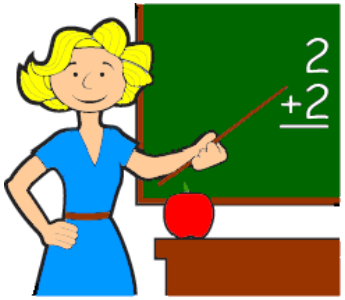


Grading



Detecting plagiarism

Teacher performance



Student performance

Objectives – Smart Cities

Predictive maintenance

Flood prediction

Smart lighting

Traffic management

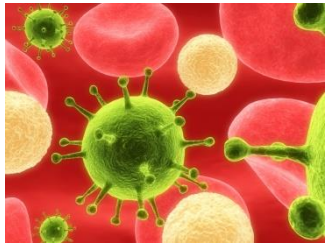
Electricity Demand

Objectives – Health

Diagnostics

Genome sequencing

Disease spreading



Tumour detection

Medical fraud detection

Main tasks

Prediction

⇩

Regression

Segmentation

⇩

Clustering

Classification

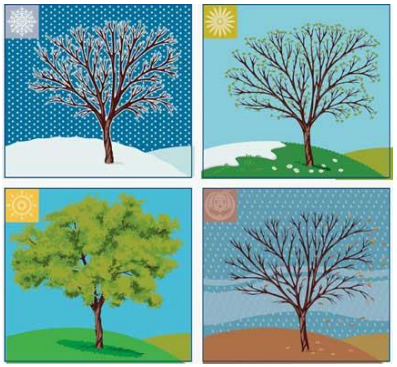
Anomalies

⇩

Detect outliers

Main tasks

Filtering effects



⇩

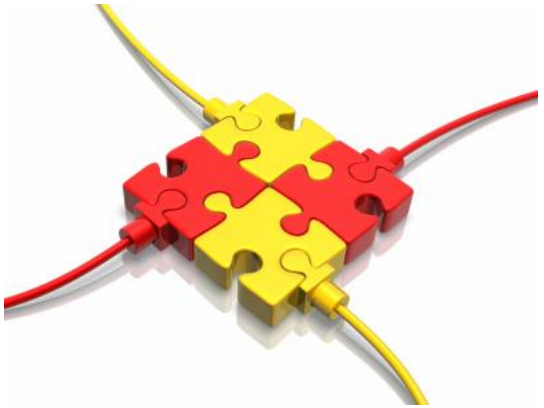
Normalization

Assess relevance

⇩

Ranking

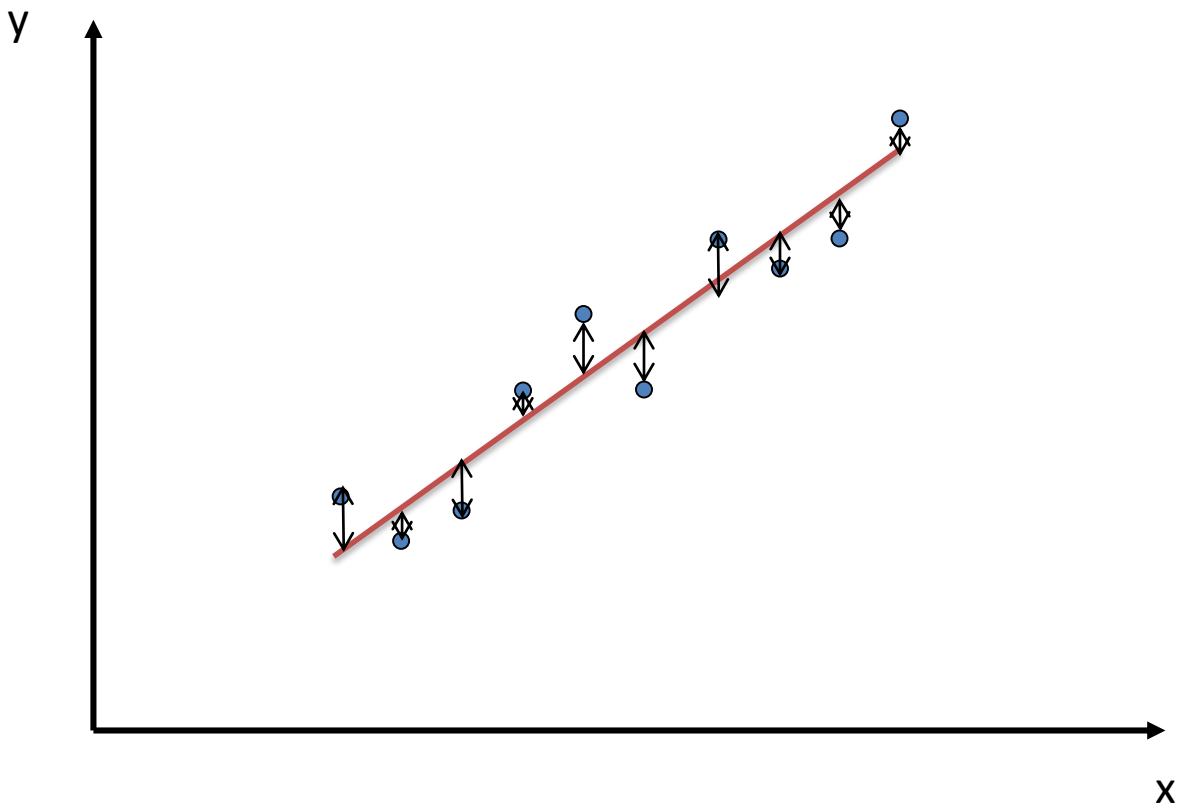
Combining info



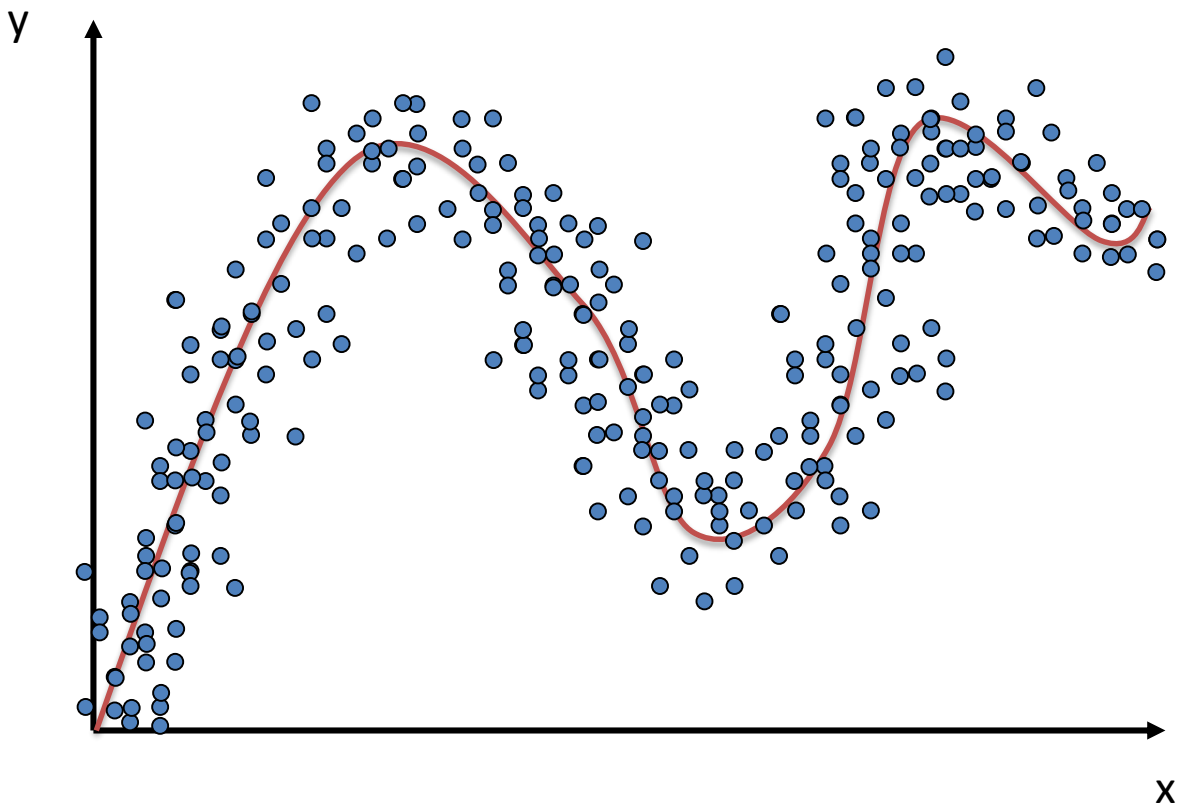
⇩

Data fusion

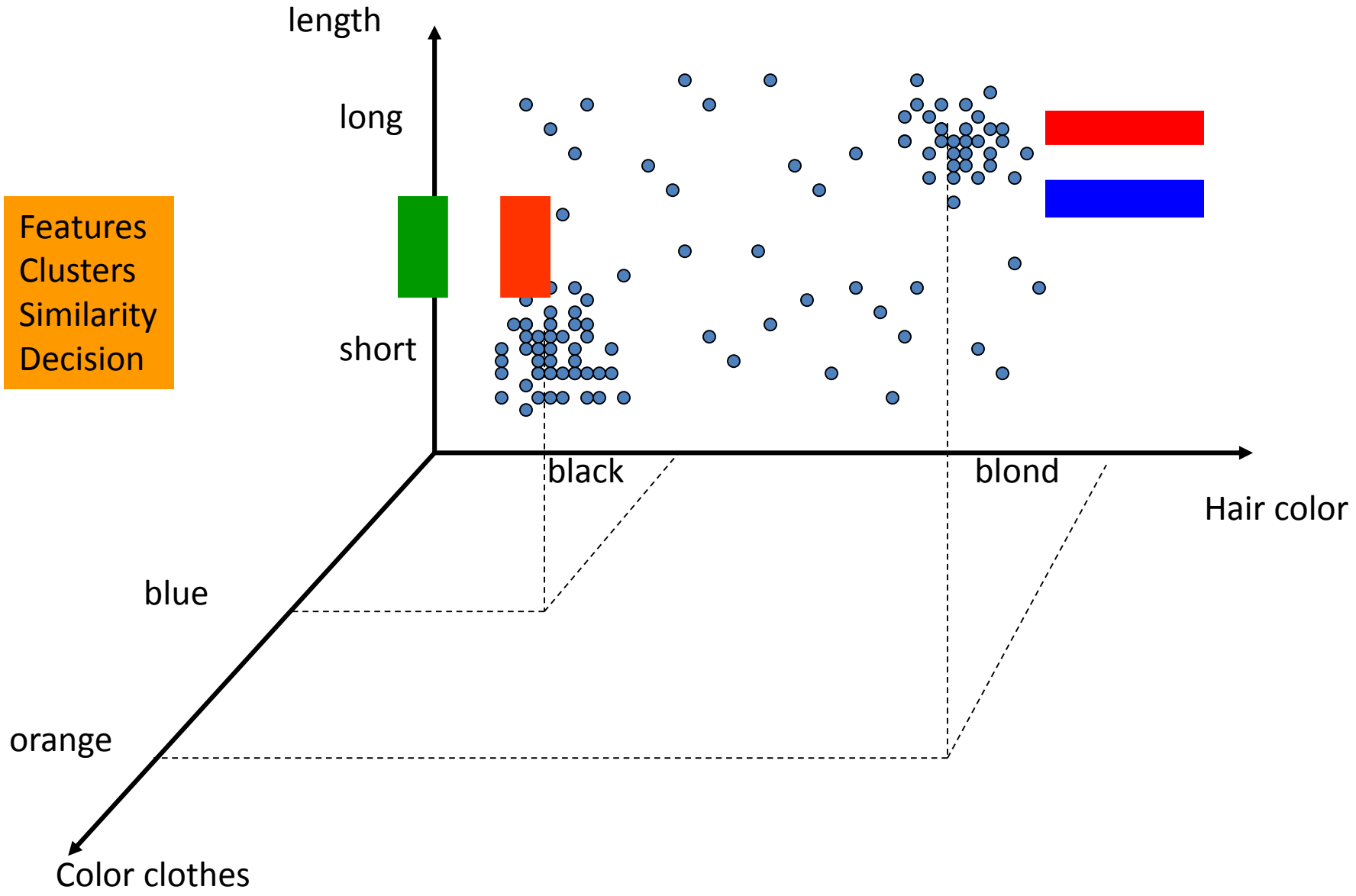
Regression



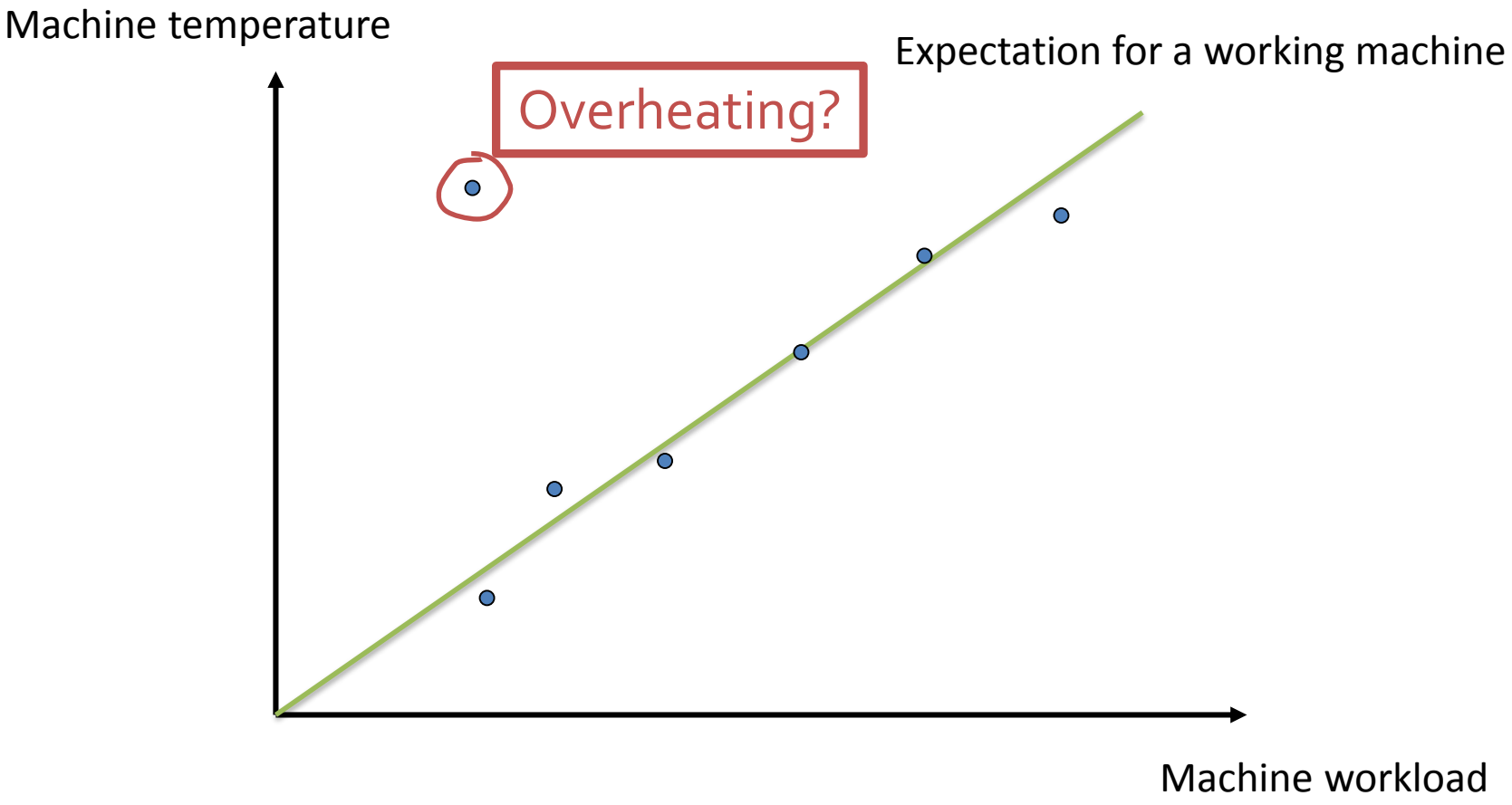
Regression



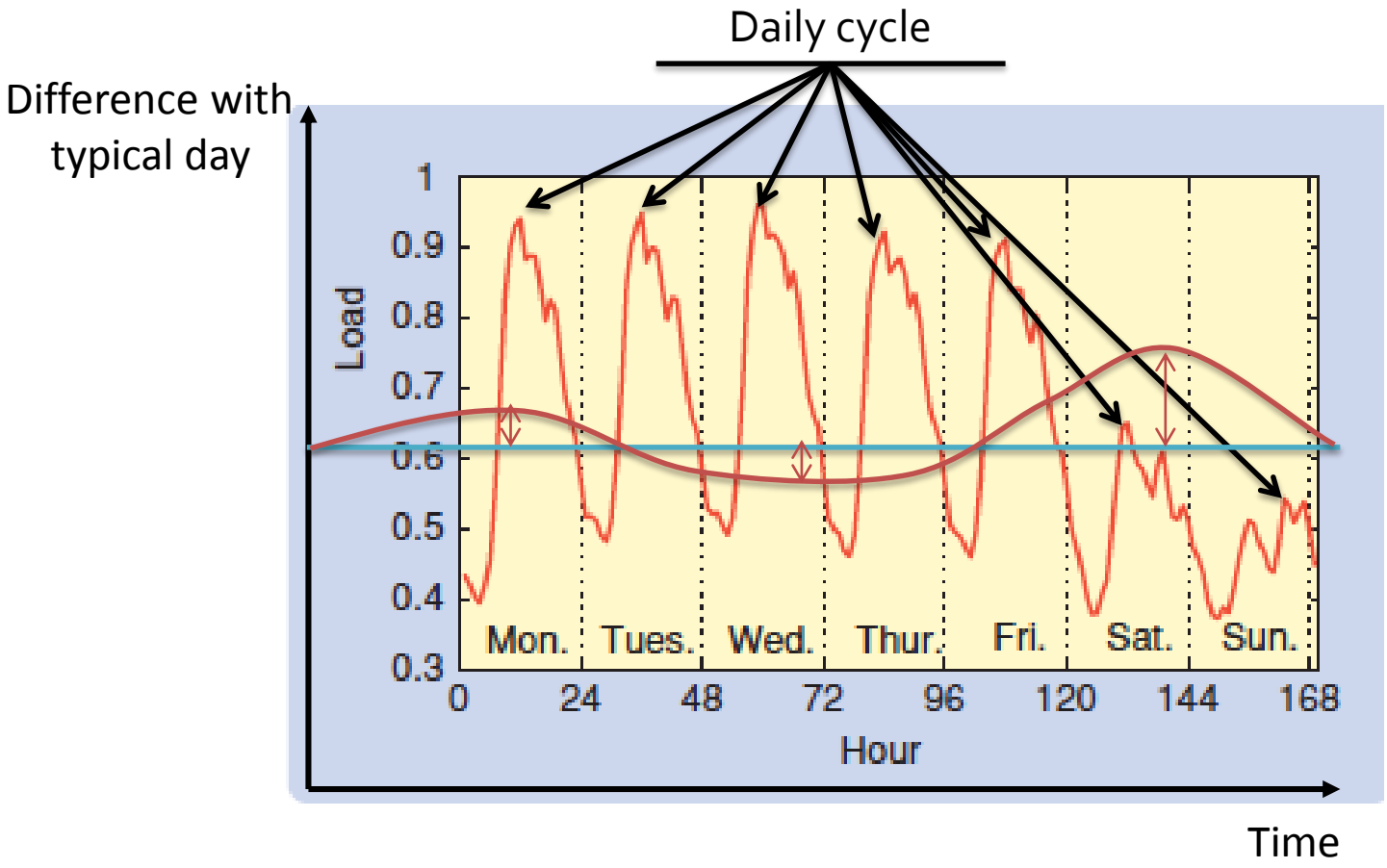
Clustering/Classification



Outlier Detection

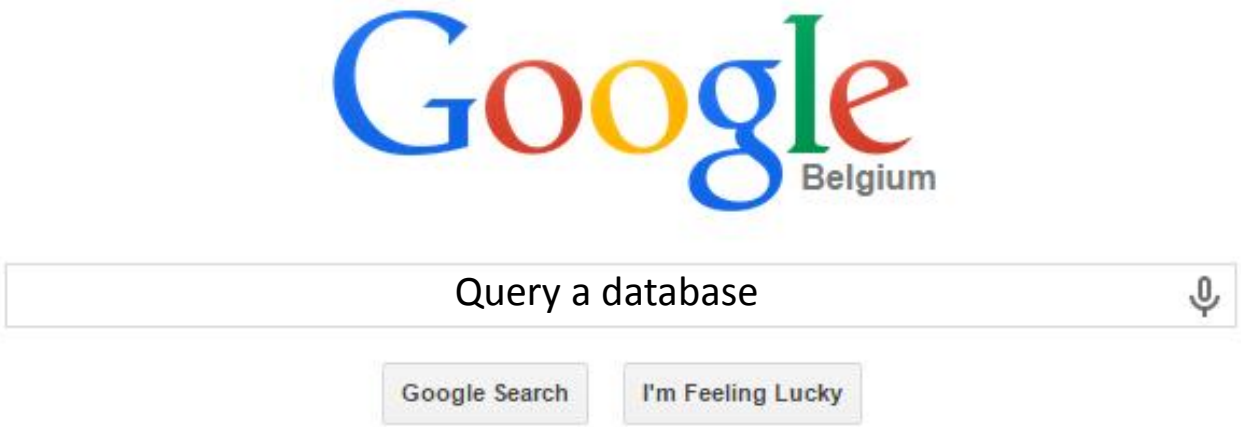


Normalization

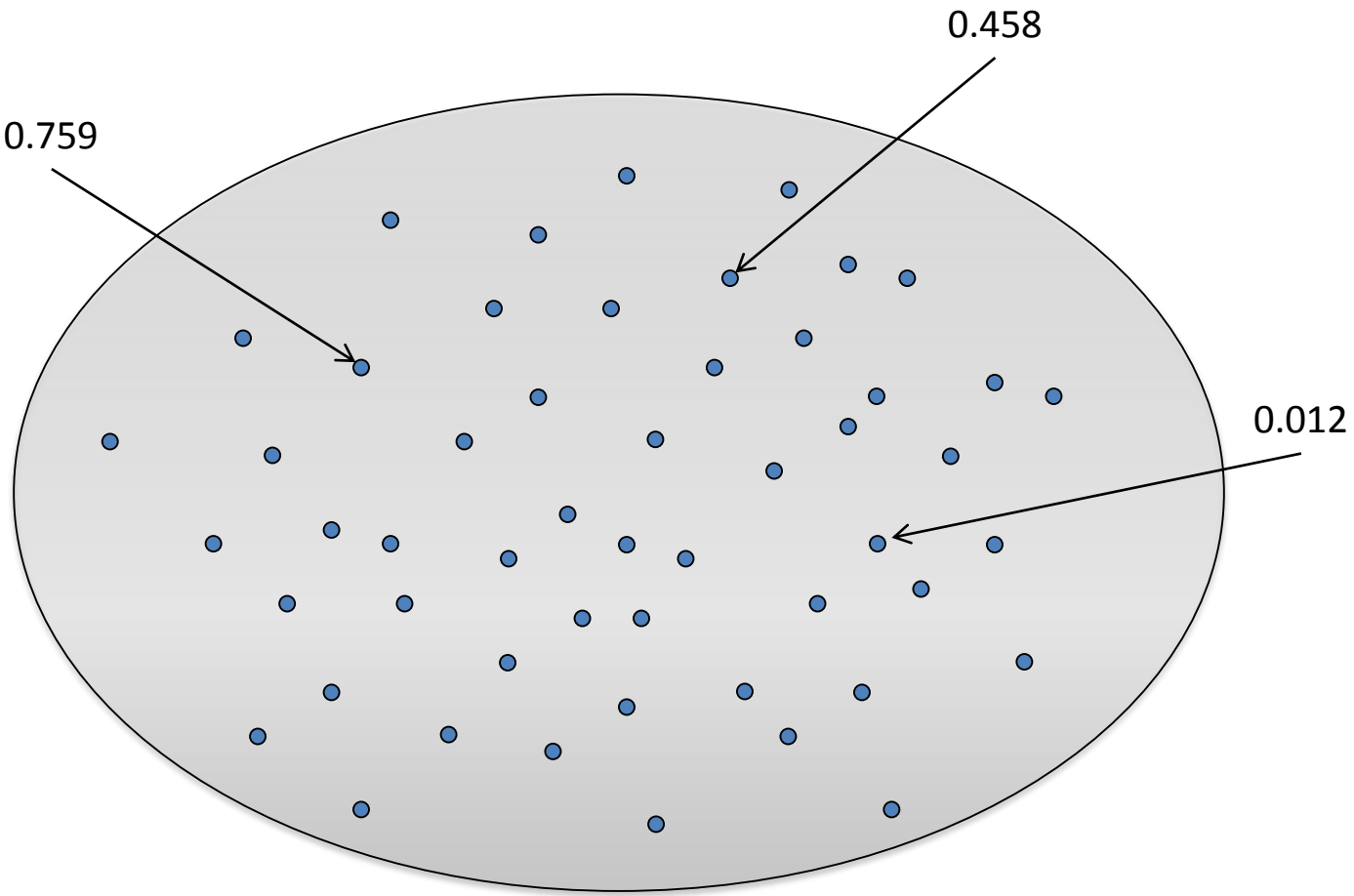


Look at divergence from
⇨ "normal" behaviour

Ranking

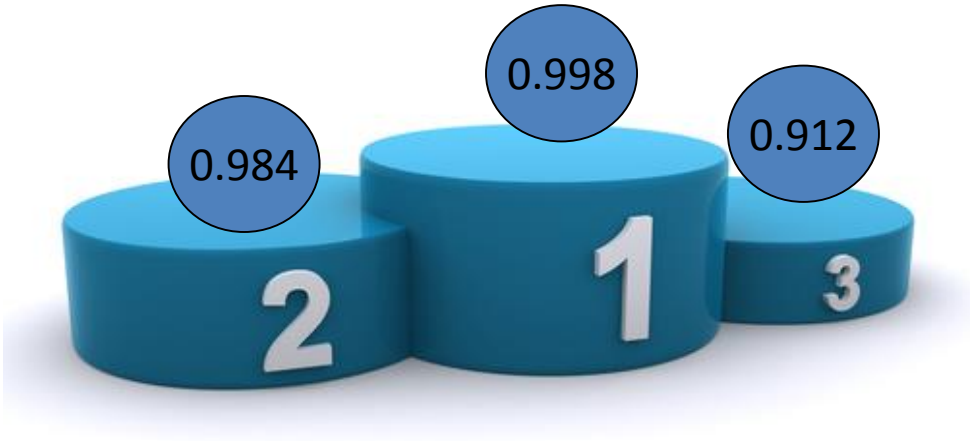

Google
Belgium
Query a database
Google Search
I'm Feeling Lucky

Ranking



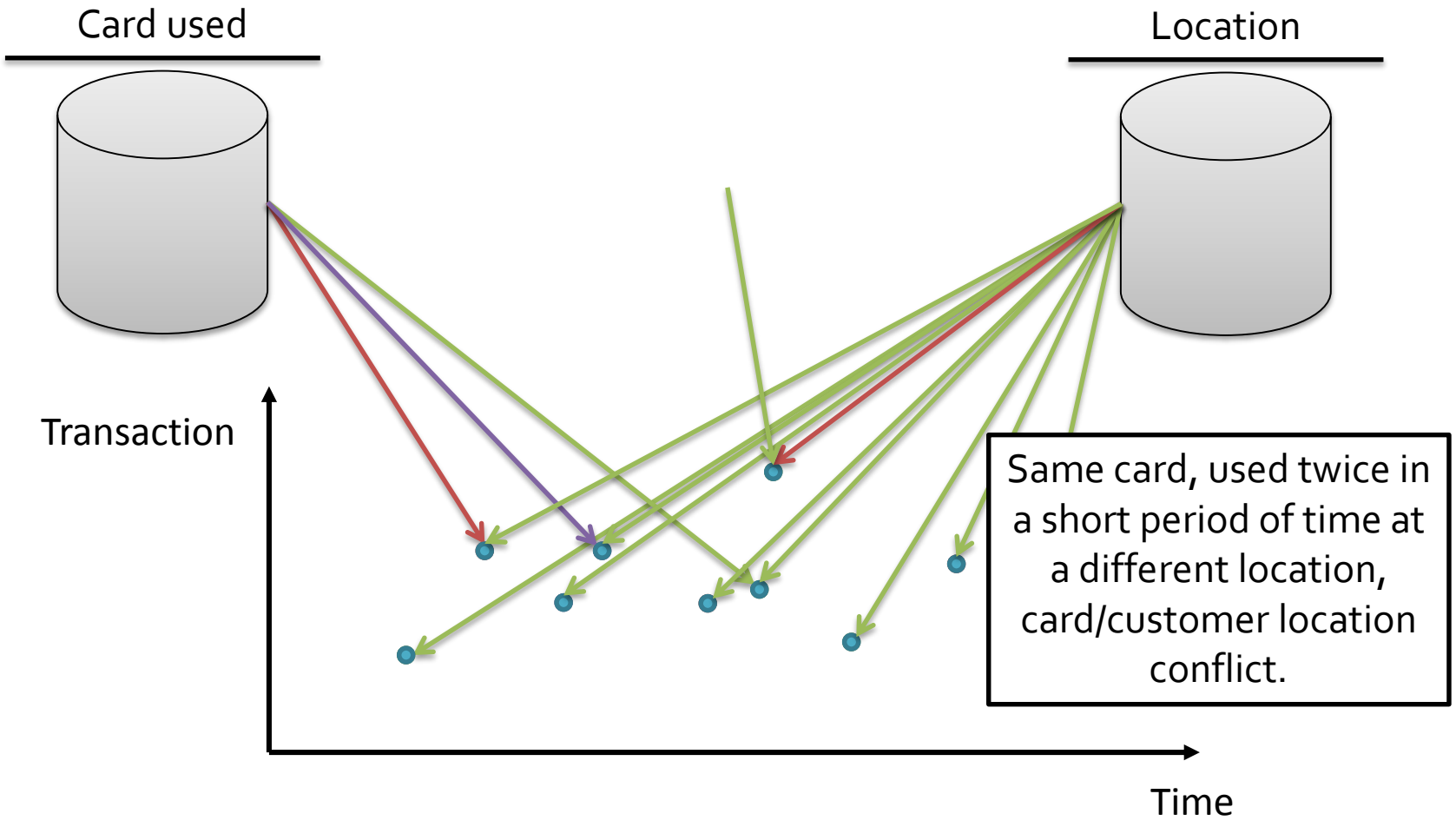
Score each point as a possible answer

Ranking



Rank all points from highest to lowest score and return this list

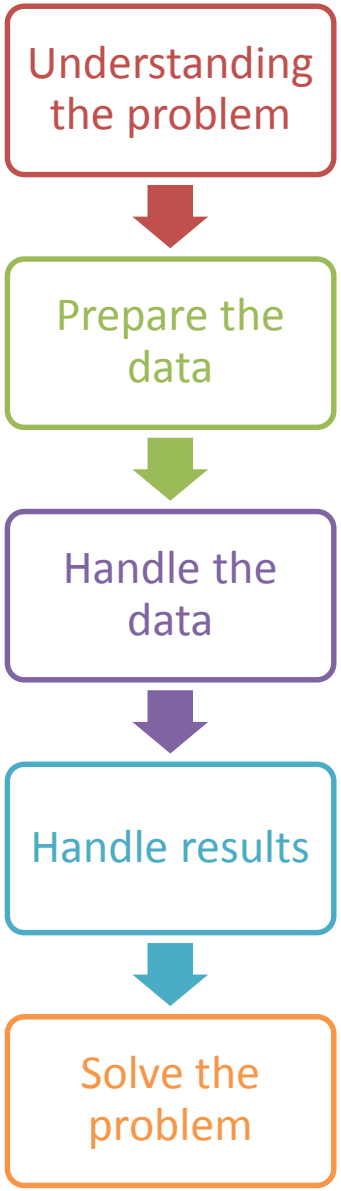
Data fusion



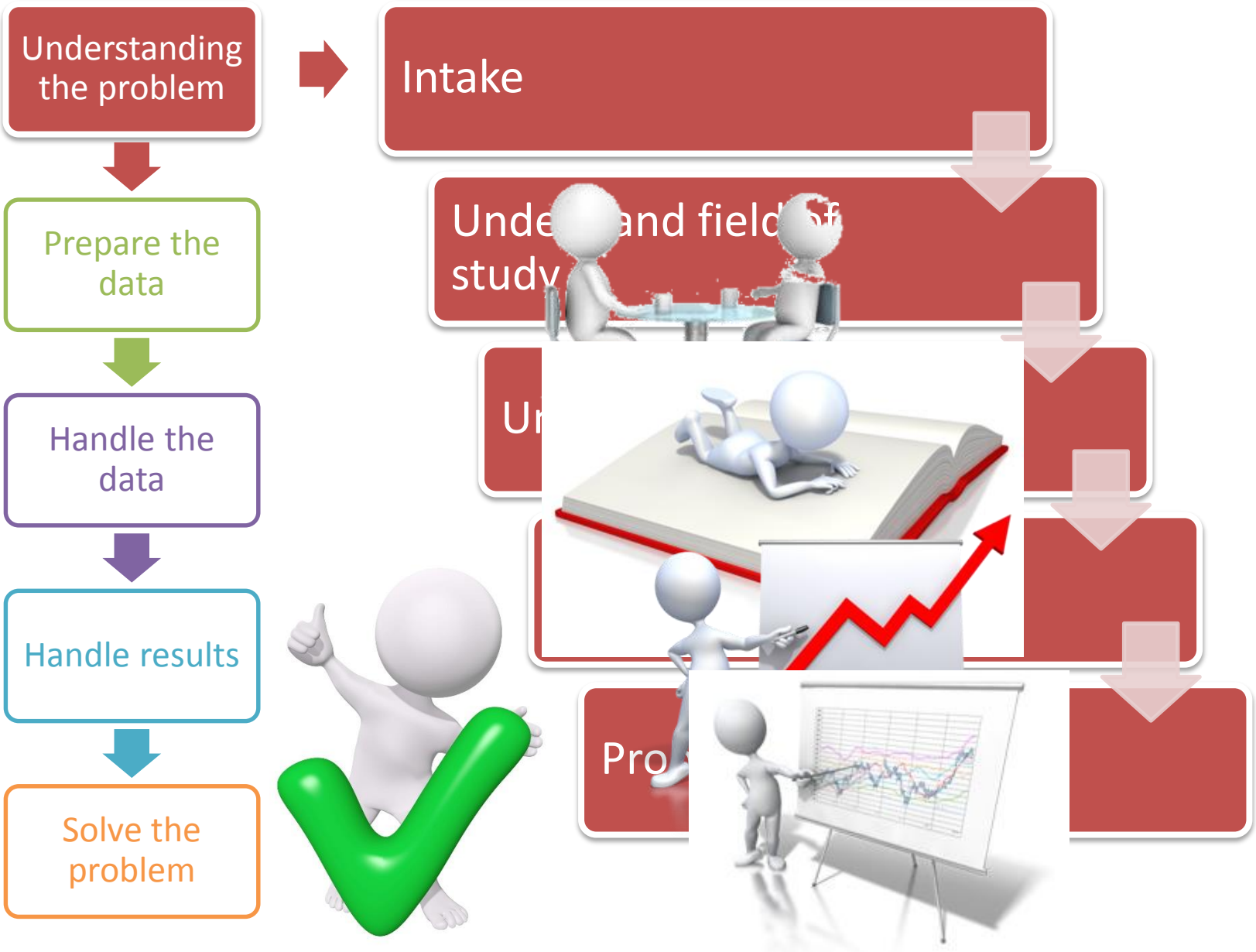
At first sight, this looks like normal behaviour

Card theft?

Taming a Data Project



Taming a Data Project



Taming a Data Project

Understanding the problem

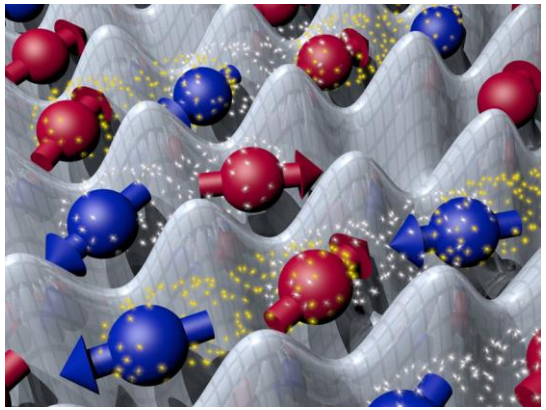
Prepare the data

Handle the data

Handle results

Solve the problem

Change this



Into this

Dim	Attribute	$\mathbf{K}^{11}$	$\mathbf{K}^{12}$	$\mathbf{K}^{22}$
1	Angle (degree)	0.0	42.9	0.0
2	Vertex distance (Å)	0.0	9.7	0.0
3	Square-root of area	4.0	4.5	5.0
4	Aspect ratio	1.0	0.8	1.0
5	Mean distance (Å)	4.2	14.0	3.9
6	Std. Dev. of distance (Å)	3.2	3.8	2.1
7	Contact pattern type	3	5	1

Data
Input module

Taming a Data Project

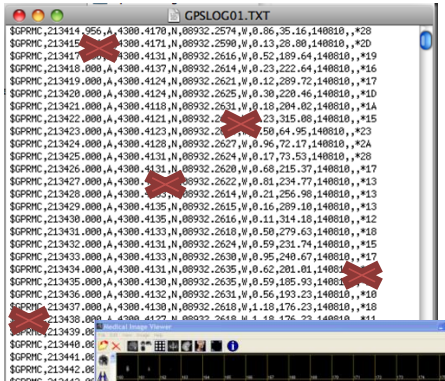
Understanding the problem

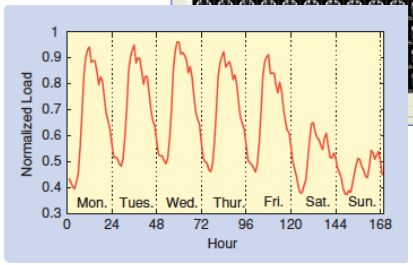
Prepare the data

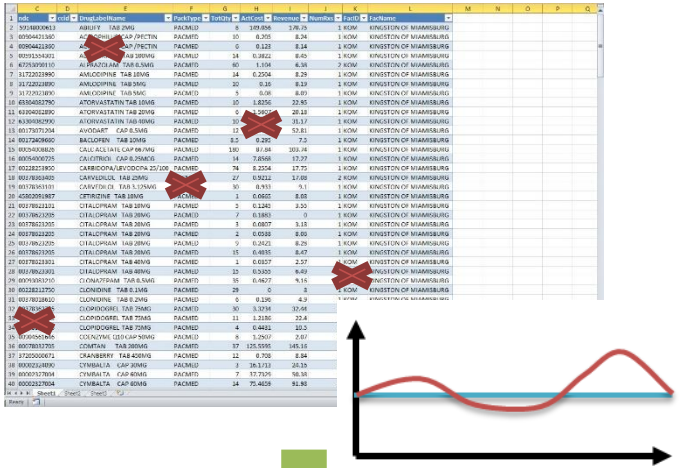
Handle the data

Handle results

Solve the problem



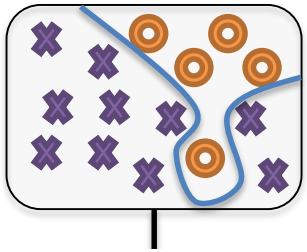


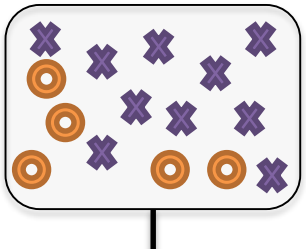


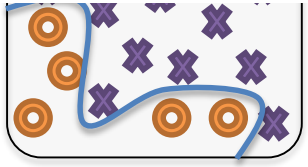
Data cleaning & translation

Taming a Data Project

- Understanding the problem
- Prepare the data
- Handle the data
  - Confidentiality & Security
  - Analytics
  - Visualization
- Handle results
- Solve the problem





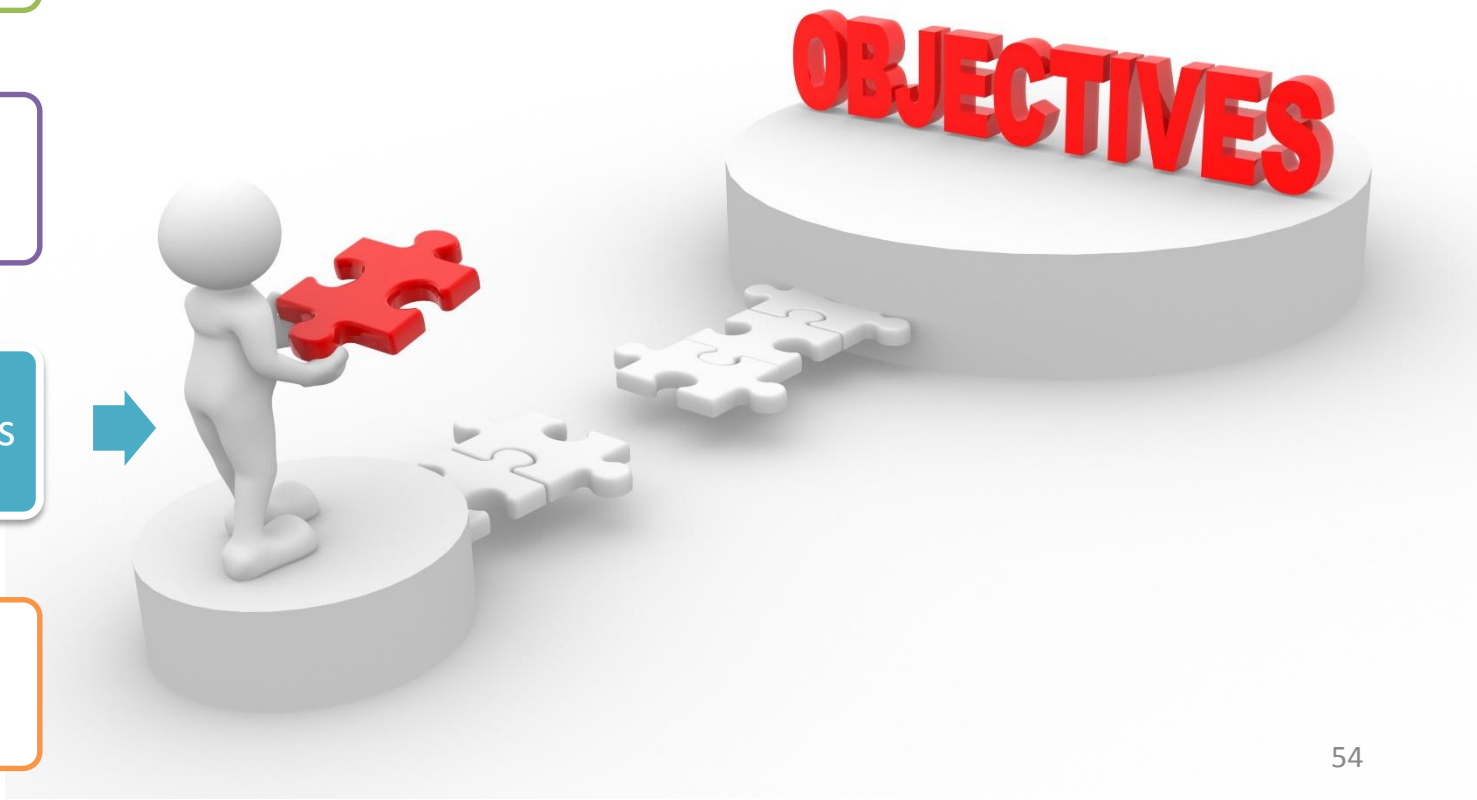


Is there enough and adequate data?

Taming a Data Project

- Understanding the problem
- Prepare the data
- Handle the data
- **Handle results**
- Solve the problem

Correlate to objectives



Taming a Data Project

- Understanding the problem
- Prepare the data
- Handle the data
- Handle results
- Solve the problem

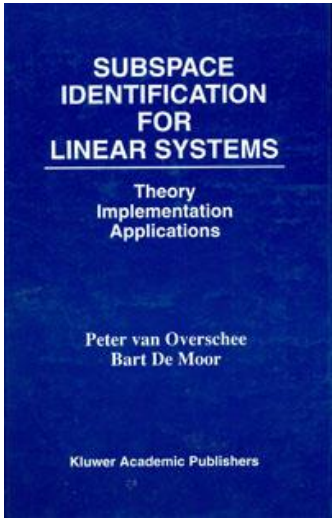
Stadius Expertise

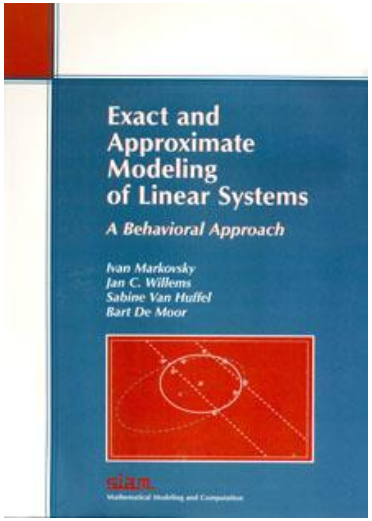
Serious Data
Serious Mining

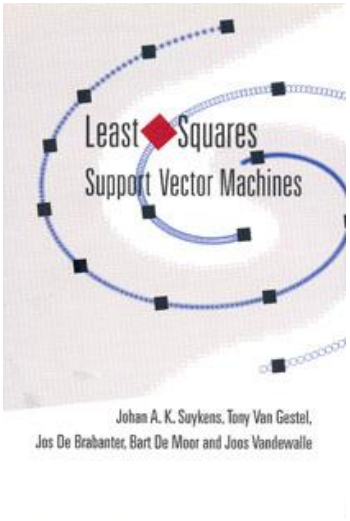
Energy
Industry
Environment
Social networks
Fraud and predictive analysis
Health
…

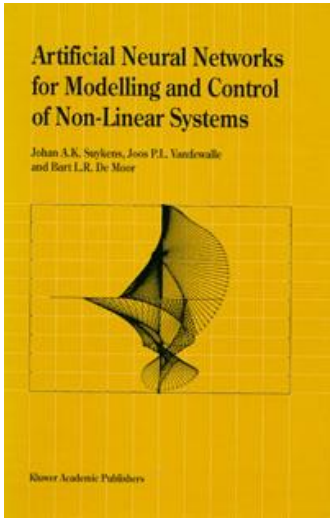
Real Quantifiable
Return on Investment

Stadius - Books









Stadius - Software

- EnsembleSVM
- TensorLab
- Clinical Data Miner
- Endeavour
- Beegle
- …

Stadius - Spin-offs

IPCOS

- K.U.Leuven spin-off company, 1995
- www.ipcos.be
- specialized in modelling and control of multivariable industrial processes (chemical and power plants, oil exploration)
- CEO : Dr. Peter Van Overschee
- Awarded with :
  - Starters award Foundation King Baudouin (1995)
  - 3-annual starters award Flanders Technoland (1996)

- K.U.Leuven, Spin-off company, 2005
- In silico drug discovery and screening
- CEO : Wilfried Langenaeker
- Silicos in the press:
  - 41ste Van Cauterenleerstoel : Maken ingenieurs het verschil ?
  - Silicos expands biology lab and announces move to new facilities (20/07/2009)
  - Silicos, a computational drug design biotech, appoints Dr. Jack Elands as Chief Business Officer (29/06/2009)

- K.U.Leuven spin-off company, 2002
- www.tmleuven.be
- Traffic and mobility consulting
- CEO : Griet De Ceuster
- TML in the press:
  - Vlaamse regering houdt meccano-tracé buitenspel
  - Meccano wil alsnog meespelen
  - Knack : Meccano superieur aan BAM-tracé
  - De Standaard : Ik mijd dure spits in de stad
  - De Tijd : Slimme kilometerheffing is mogelijk
  - Rekeningrijden : 30 % minder files
  - De Standaard : Verdubbeling Brusselse Ring kost miljard
  - AHA-wetenstappen : Ontdek Leuven via de AHA ! wetenstappen route
  - Wiskundige modellen in het verkeer

COMPANY
COMPETING O

Business Planning & Supply Network Mgmt
LEVEL 4 (ERP LAYER)

Product Definition | Production Capability | Production Schedule | Production Performance

Manufacturing Operations Management
LEVEL 3 (MES LAYER)

Continuous | Batch | Logic

Common Data/Information

Sensors, Actuators and Devices

LEVEL 0,1,2 (DCS LAYER)

COST MES PROJECT

INFORMATION INTELLIGENCE (I-SQUARE)

inversys INFUSION

APPLICATION ARCHITECTURE
OS SOFTWARE
HARDWARE

BUSINESS INTELLIGENCE USING PREDICTIVE TECHNOLOGIES ROI OF 145%

BUSINESS INTELLIGENCE WITHOUT PREDICTIVE TECHNOLOGIES: ROI OF 89%

PRODUCTION IMPROVEMENT ROI OF 237%

IDC Analyze the Future

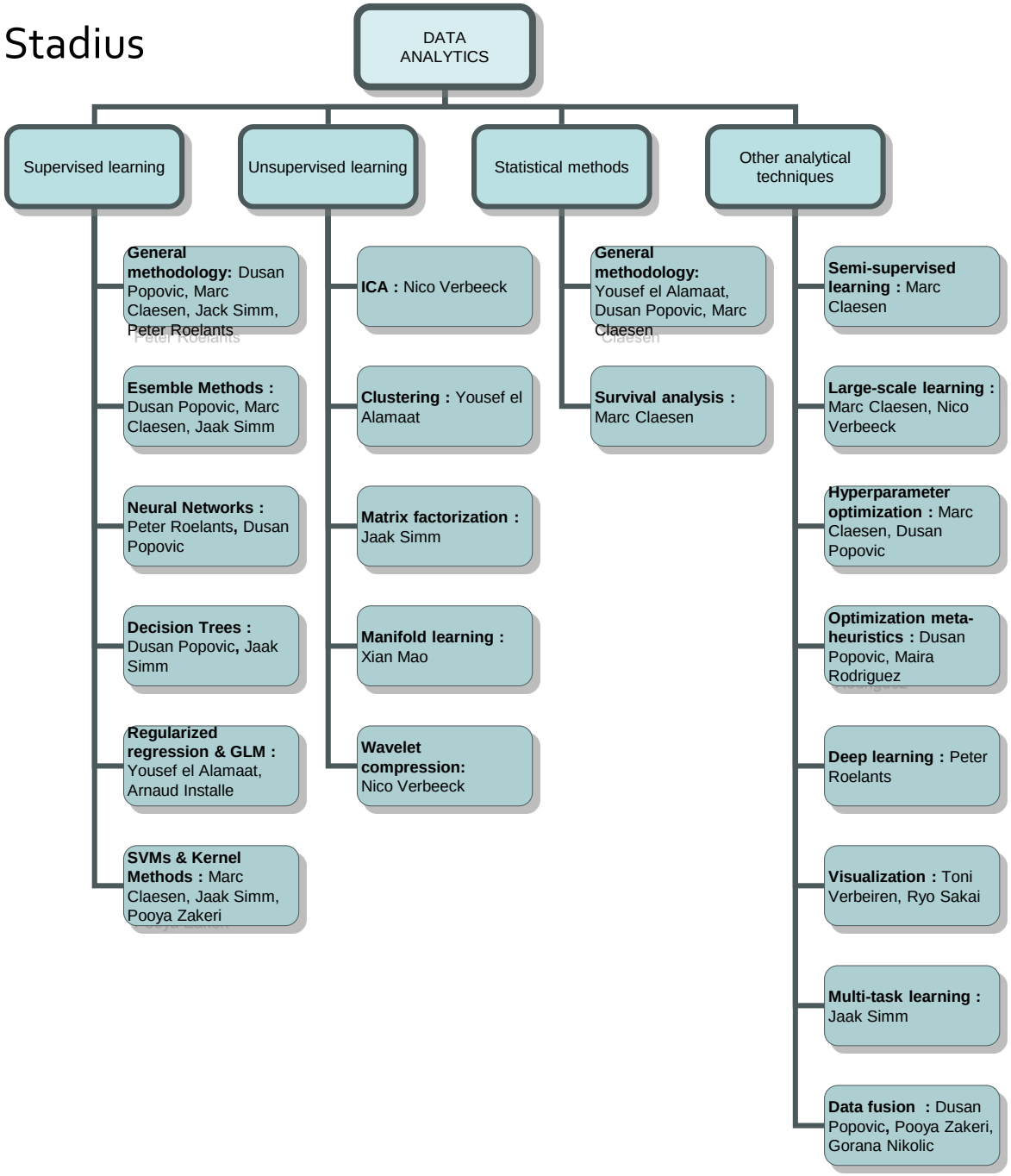
BUSINESS ADDED VALUE, COMPETITIVE ADVANTAGE VS. DEGREE OF INTELLIGENCE

INFORMATION INTELLIGENCE

- K.U.Leuven spin-off company, 2000
- www.norkom.com (acquired by Norkom Technologies, 2006)
- Fraud detection in telecom and finance
- CEO : Steven Verhoeven
- Data4s in the press:
  - Campuskrant : "Gemma Frisiusfonds voor prille spin-offs bestaat 10 jaar" (Dec. 2007)
  - LeuvenInc : Norkom technologies acquires European risk management firm data4s (Nov. 2004)
  - Norkom Technologies neemt Europese Risk Management firma DATA4s over (Nov. 2004)
  - Norkom Technologies acquires European risk management firm DATA4s (Nov. 2004)
  - Norkom and DATA4s join forces (Nov. 2004)
  - KBC bank selecteert anti-witwastechnologie van data4s (Sep. 2004)
  - LeuvenInc : DATA4s secures 1.75 million euro in second financing round (Feb. 2002)
  - Rabobank bestrijdt witwassen met DATA4s technologie
  - Clear2pay integrates data4s's fraud money laundering prevention platform
  - Data4s, tussen data-mining en Rode Duivels (Campuskrant, Jun. 2000)
  - Technoparels

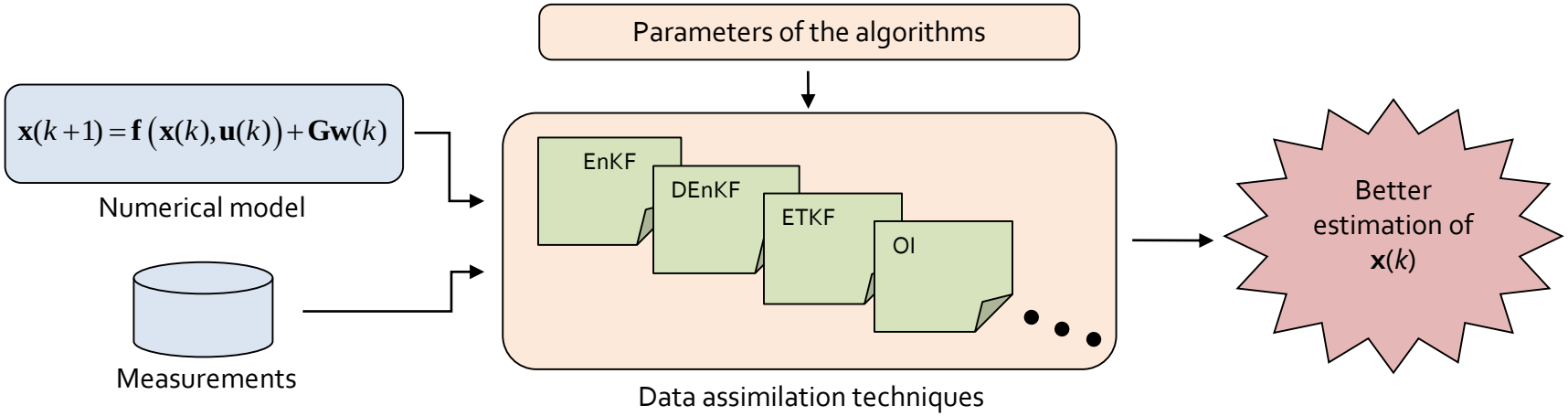
- K.U.Leuven Spin-off Company, 2008
- www.cartagenia.com
- CEO : Herman Verrelst
- Clinical applications of genetic analysis
- Cartagenia in the press:
  - Cartagenia partners major UK study on new prenatal diagnosis technology
  - Honderdste spin-off is kroon op succesverhaal LRD (Campuskrant 29/05/2013)

Algorithms in Stadius



Data Assimilation

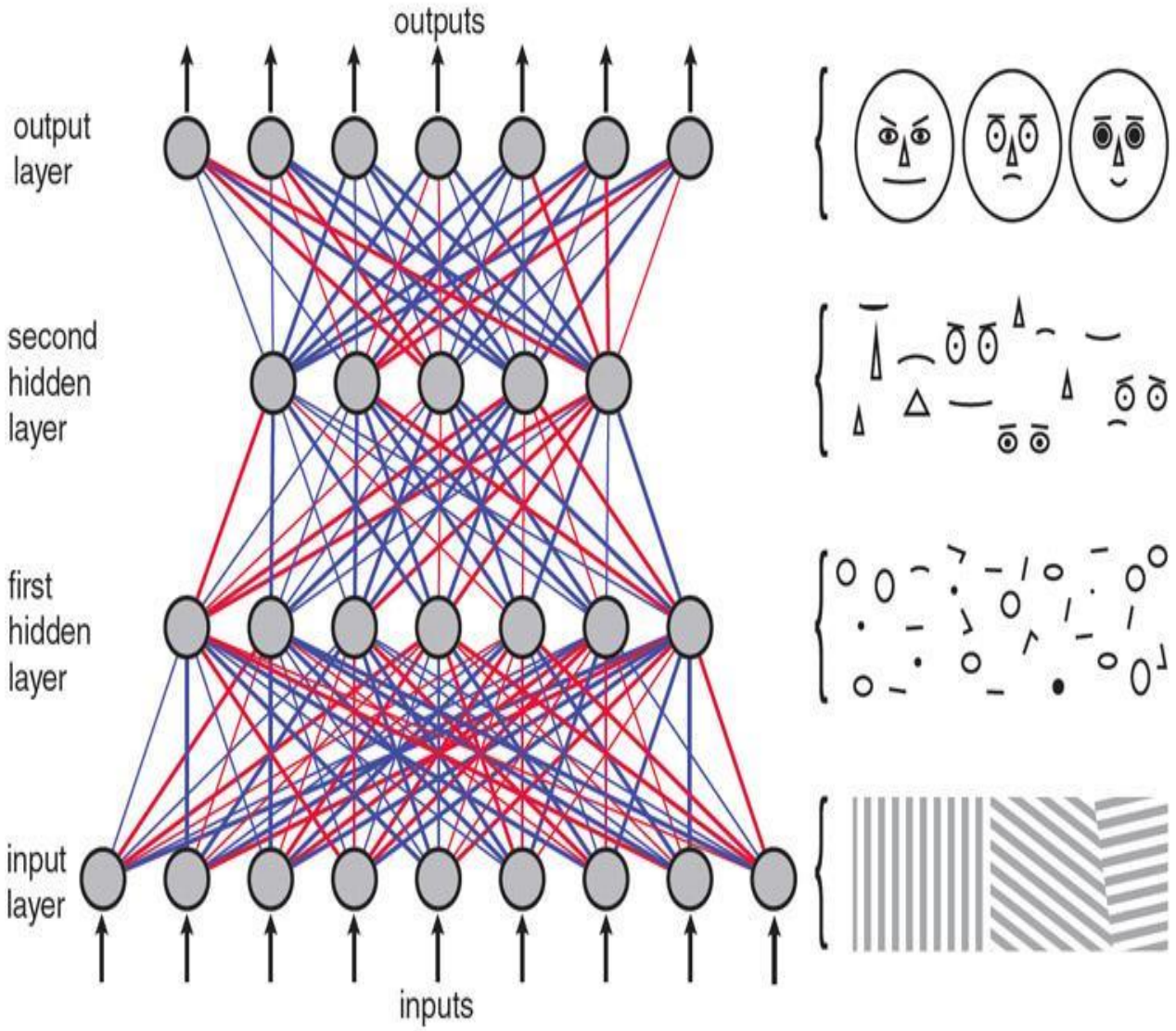
Data assimilation is the common name given to several numerical techniques that combine the outputs of a numerical model with observational data in order to improve the quality of the model predictions.



Some data assimilation techniques: 3DVAR, 4DVAR, Ensemble Kalman Filter (EnKF) and its variants, Optimal Interpolation (OI), particle filters, etc.

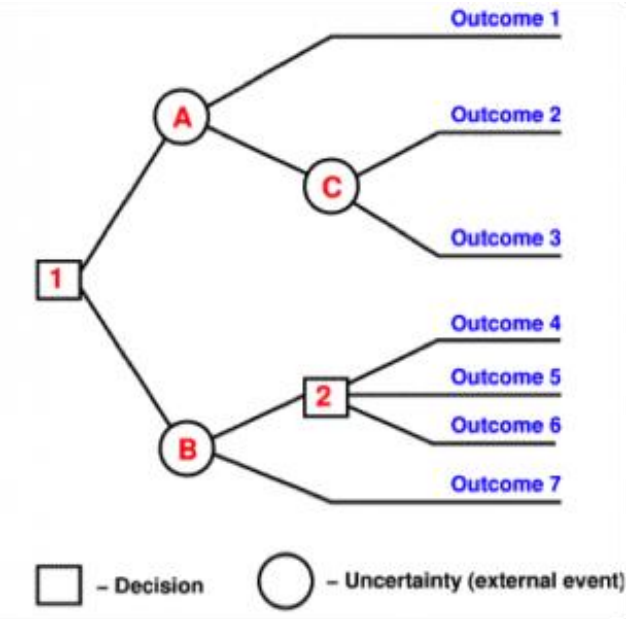
Deep Learning & Neural Networks

- Neural networks.
- New algorithms.
- Multiple layers on top of each other.
- Each layer learns a more complex representation.
- Learn feature hierarchies.



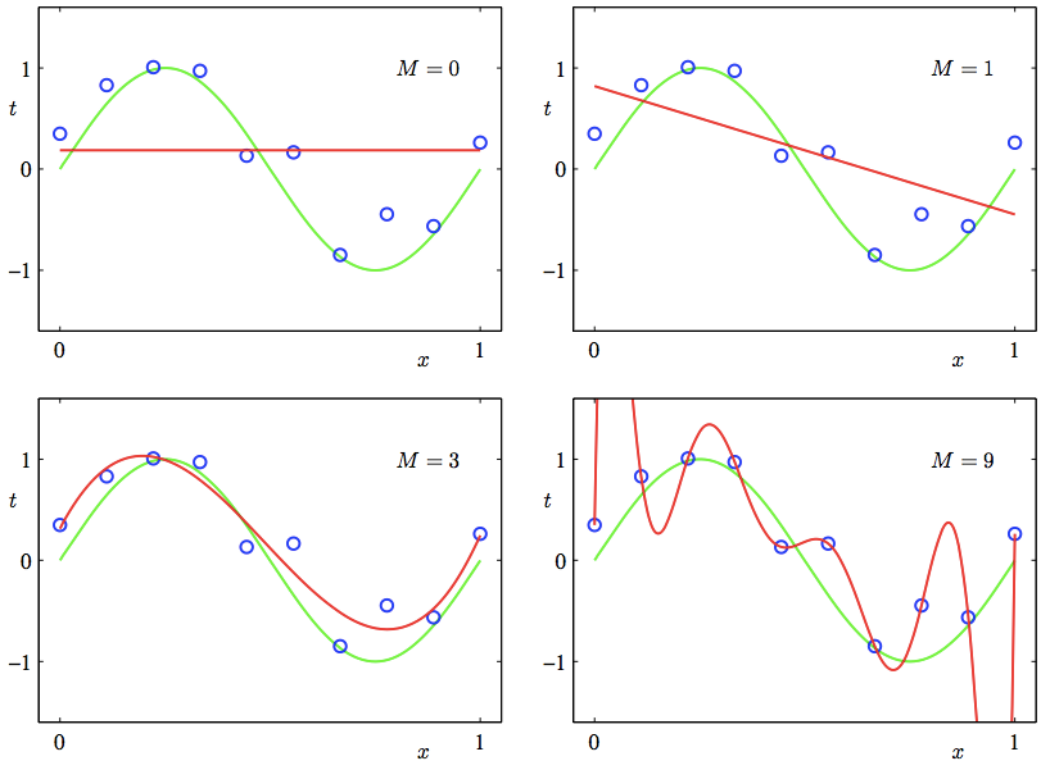
Decision trees

Decision nodes are trained according to a labeled set of data points. A new instance is given as an input and run through the tree, which then produces the most likely output.



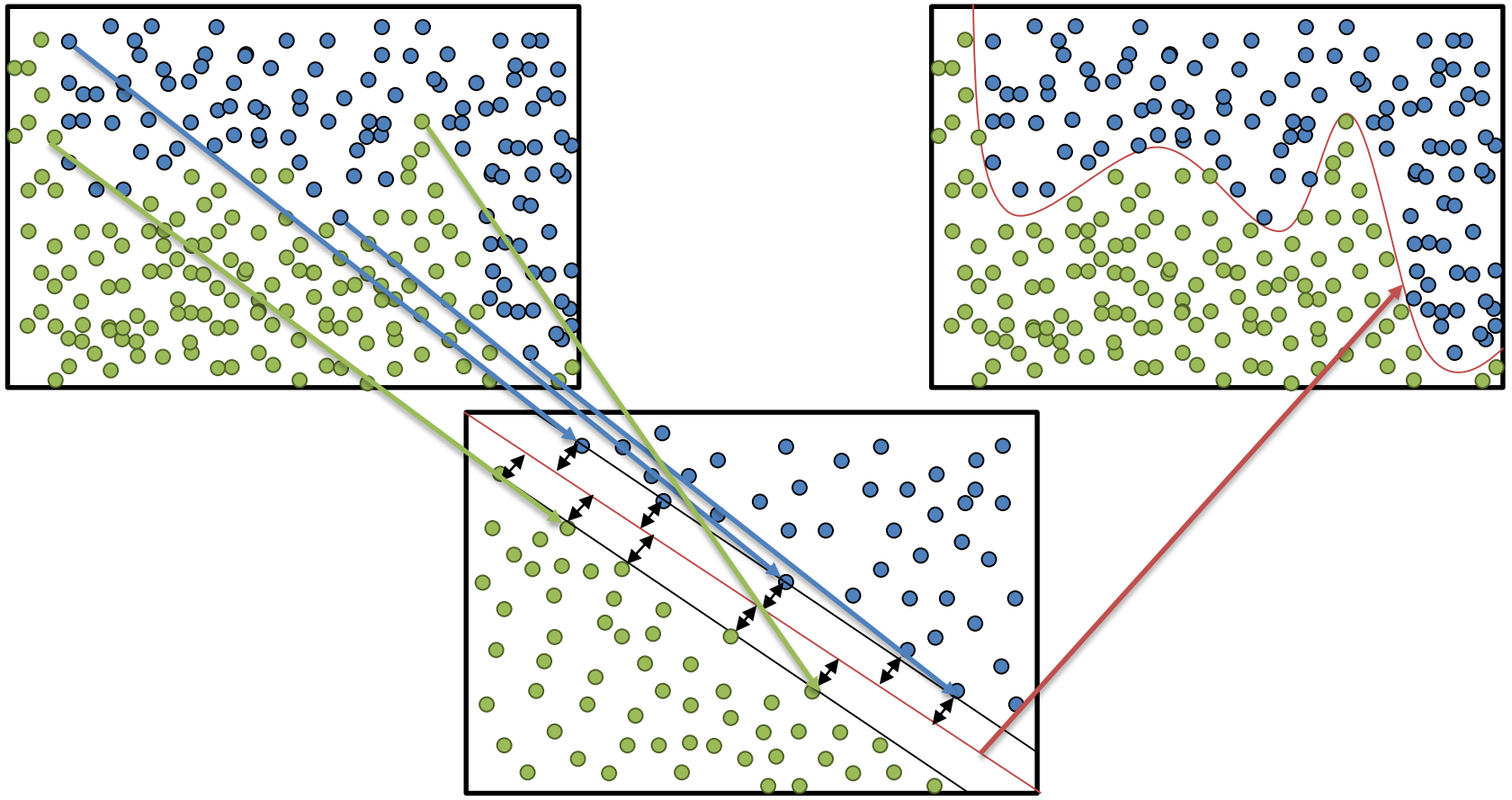
Regularized Regression

Fitting a regression function on a data set can result in overfitting: the regression fits to the data, but not to the general trend. The regression is thus not generalizable!
A solution is to punish the learner for creating a model with high complexity.

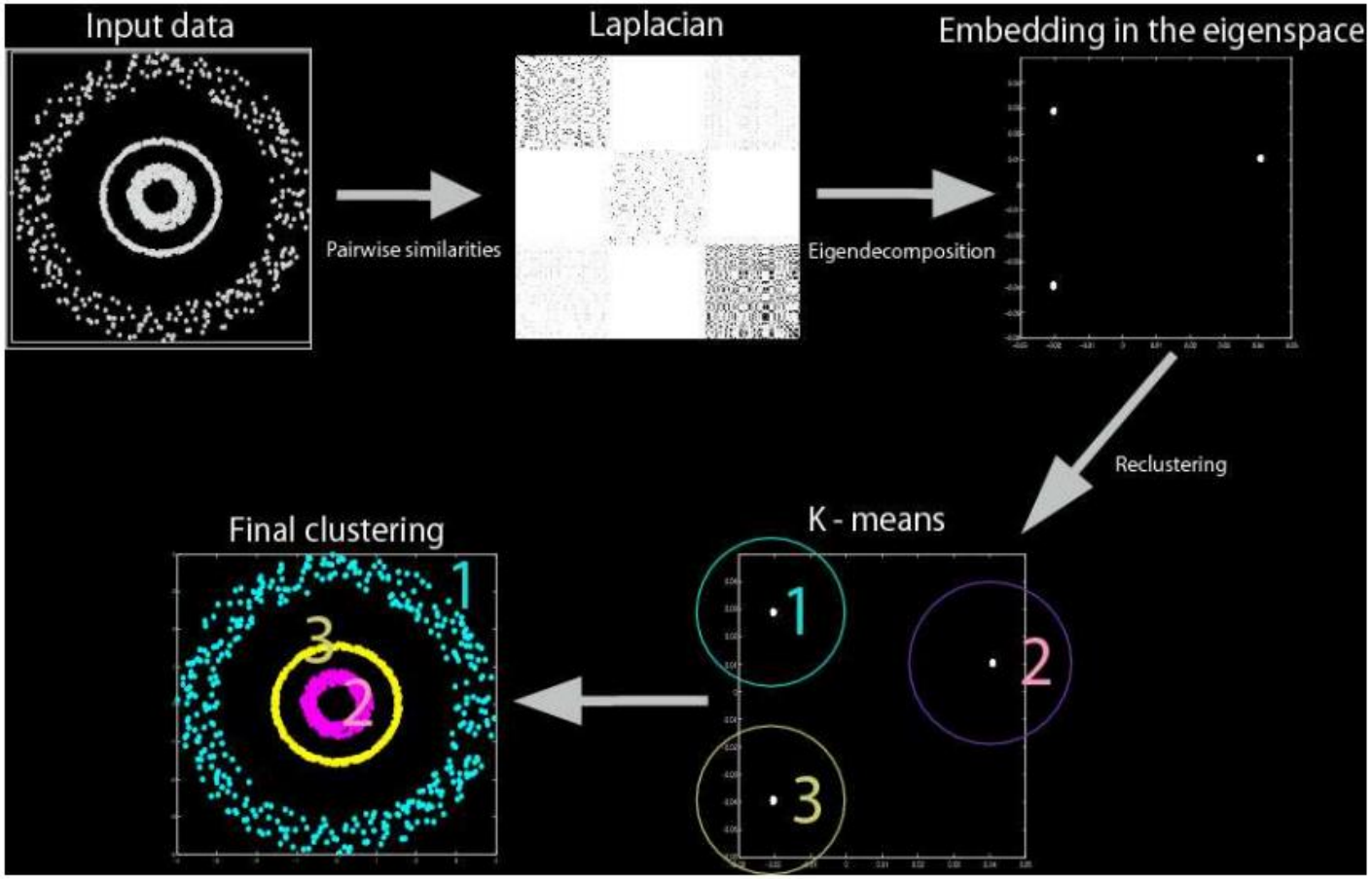


Support Vector Machine

First transform the problem to a high-dimensional form, where the solution is easily found, through the so-called 'kernel trick'. Then, transform the decision boundary back to the original form.

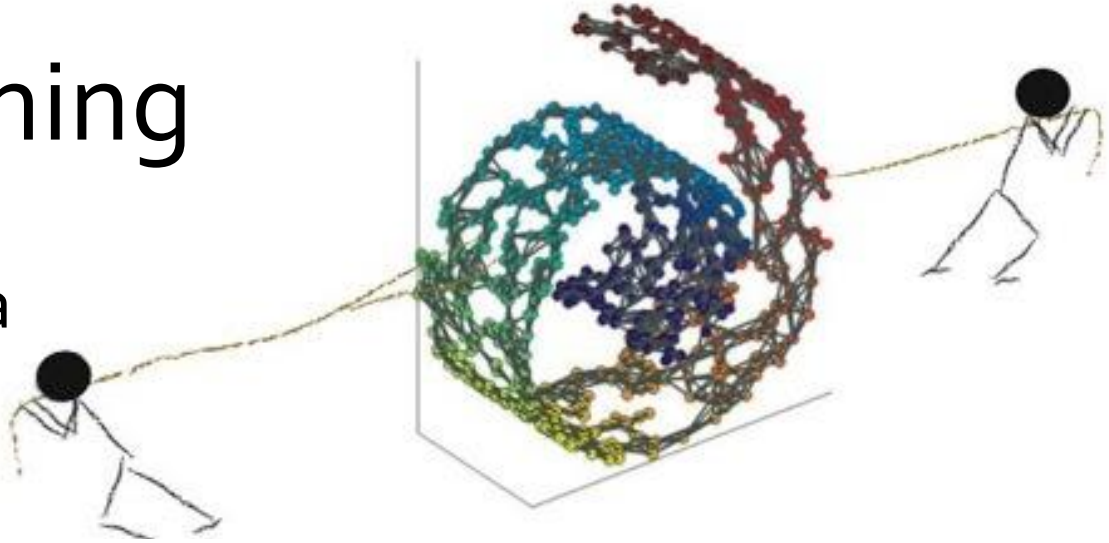


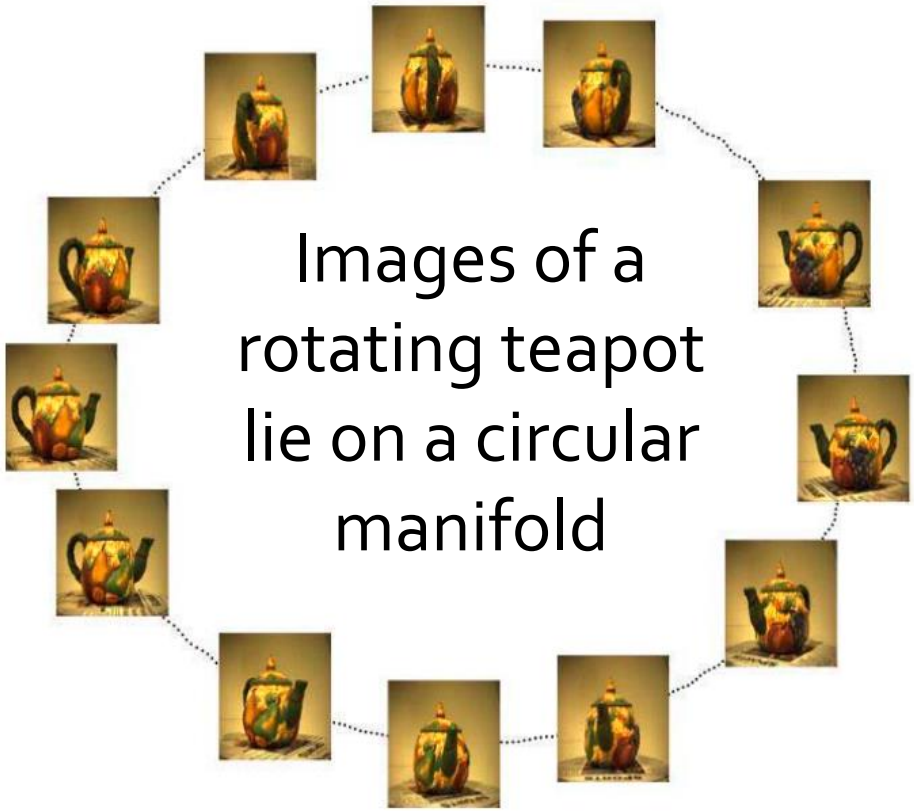
Spectral clustering



Manifold learning

A lot of datasets live on a low dimensional manifold.



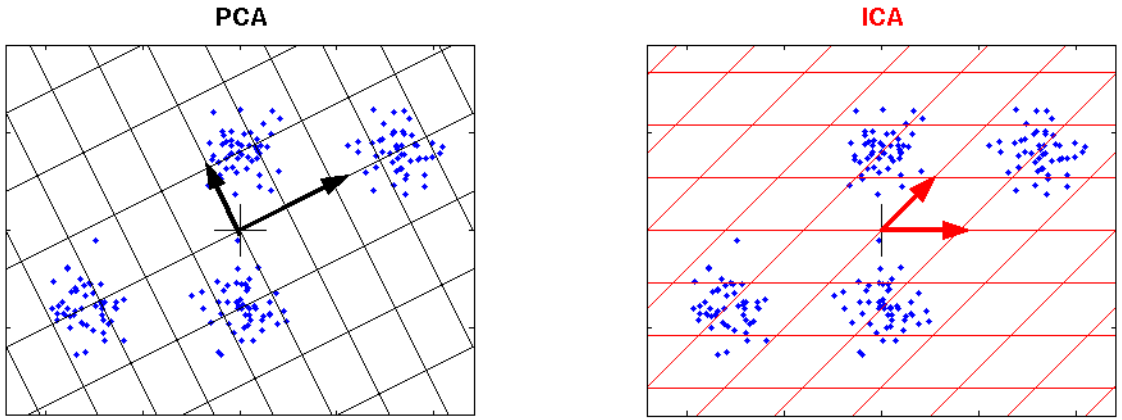


Images of a rotating teapot lie on a circular manifold

Goal: Find a low-dimensional basis for describing the high-dimensional data

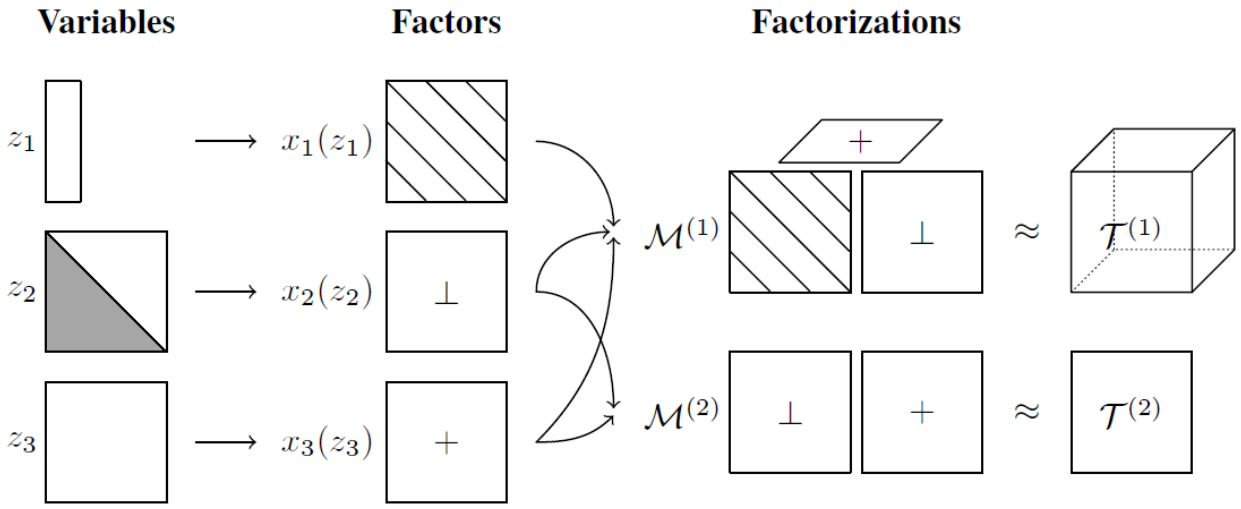
Component analysis

The data dimensionality is reduced by dividing the data set into smaller, relevant components. This can be done by maximizing the variance (principal component analysis), or by finding independent sources of data (independent component analysis).



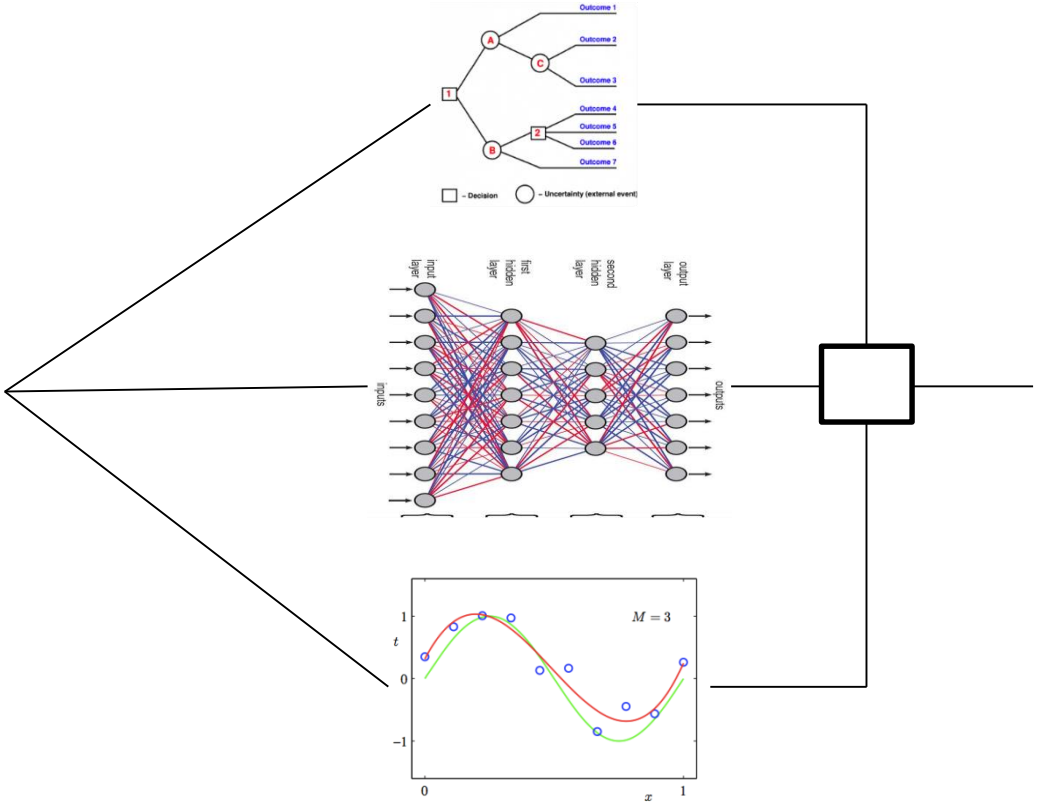
Tensor methods

The data is represented in a so-called tensor, which is a higher-dimensional extension to a vector or a list (i.e., a tensor in three dimensions is a list of lists of lists). This can be used to represent highly-dimensional data in a very concise way. We use very efficient methods to decompose these tensors, that allow us to extract common influential factors among different features of a data set. These techniques are highly scalable and work well on Big Data.



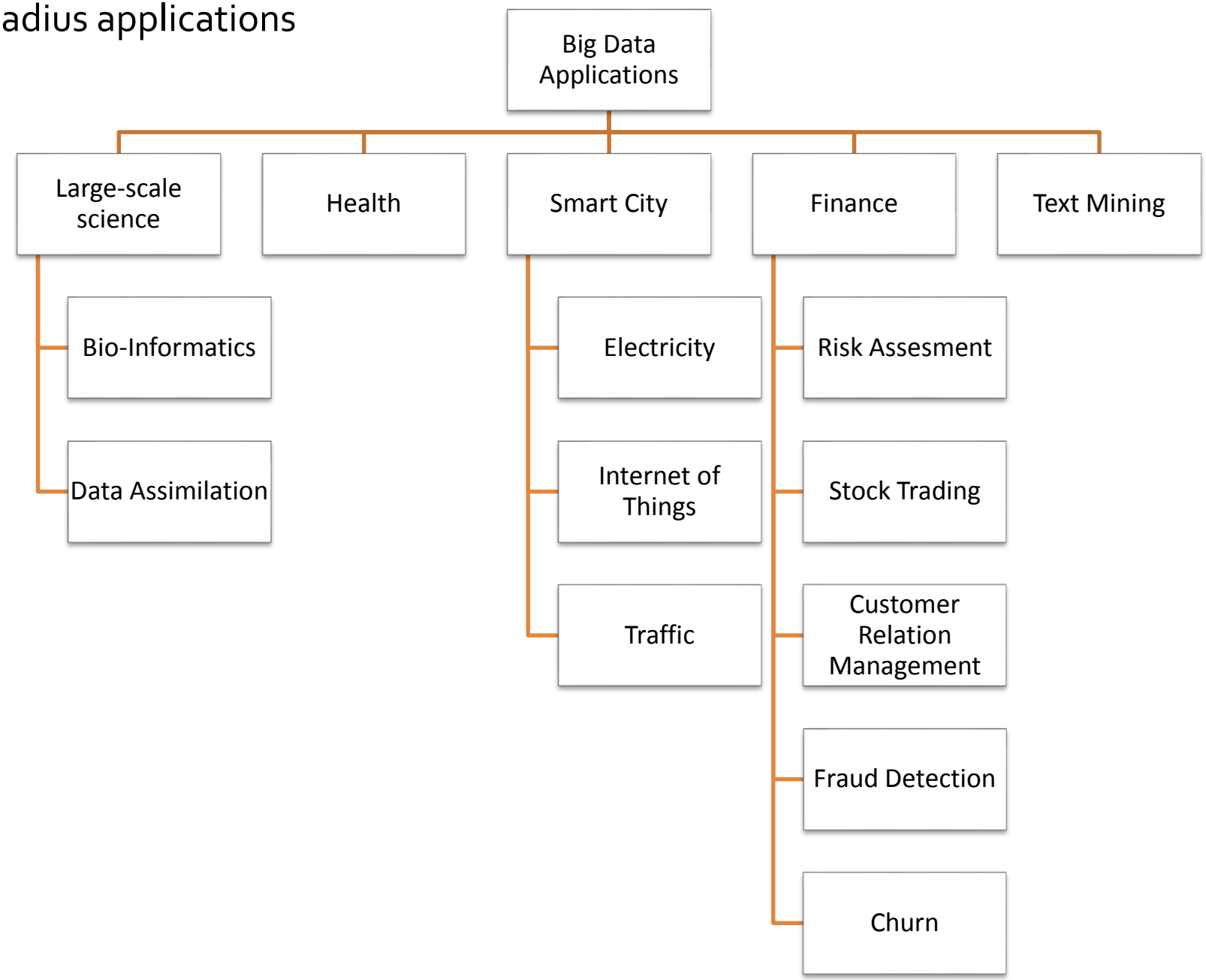
Ensemble methods

Several machine learning algorithms are implemented in parallel to each other. A decision on the outcome is then made, based on some decision rule (e.g., majority voting).



- Big Data
  - What
  - Who
- Six dimensions
  - Data
  - Compute Infrastructure
  - Storage Infrastructure
  - Analytics
  - Visualization
  - Security & Privacy
- Machine learning as a commodity
- Expertise
  - Books & Spin-offs
  - Algorithms
  - Applications

Stadius applications



Energy

Industry

Environment

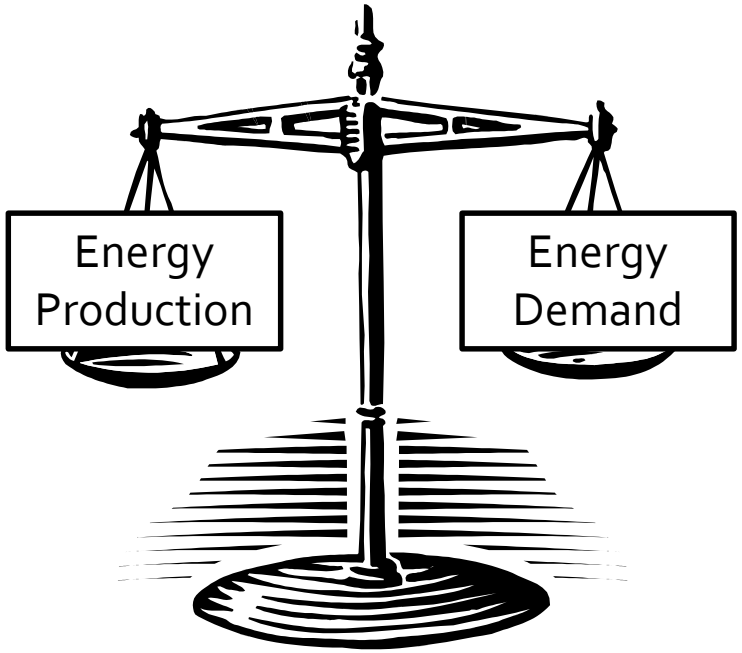
Social networks

Finance and Fraud

Health

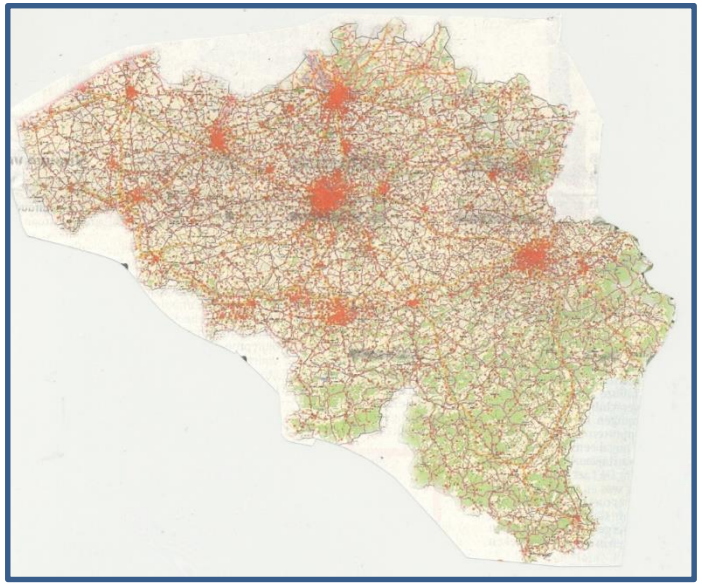
Electric load forecasting

Problem



How to forecast the demand?

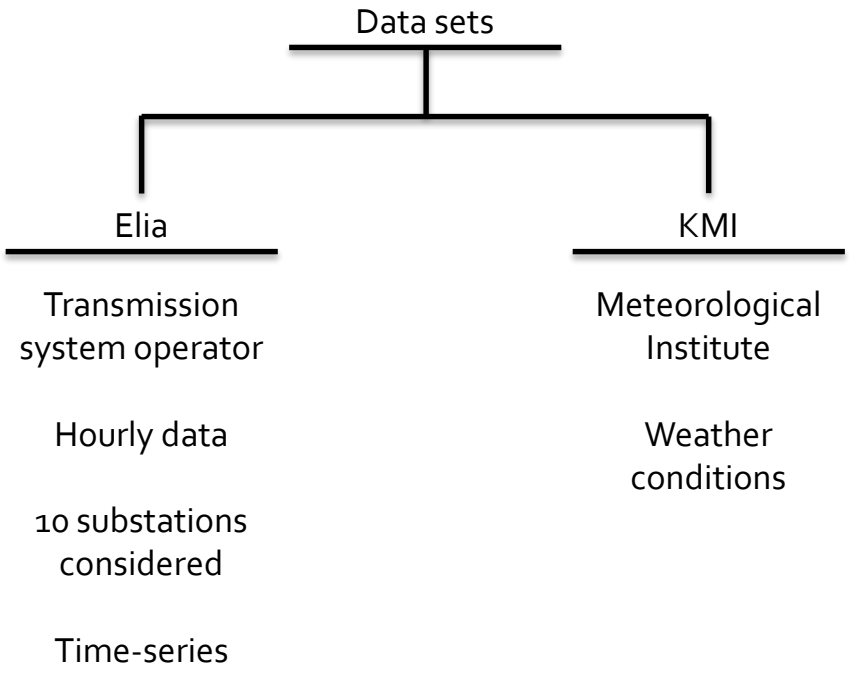
Power grid

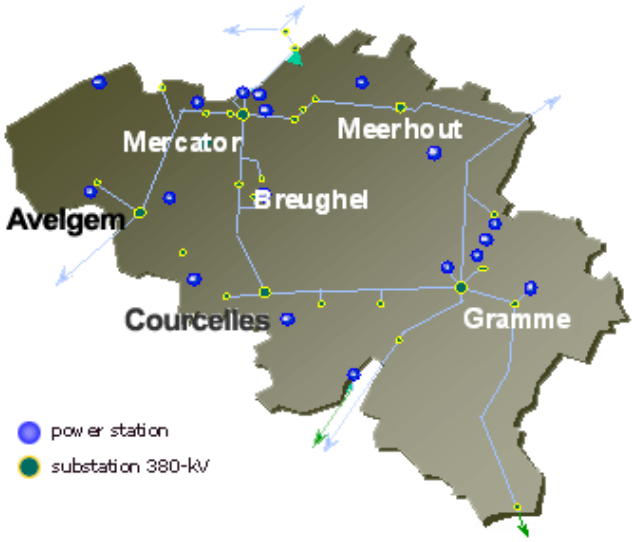



België en Europa
Het Elia-net:
knooppunt van elektriciteitverkeer in Europa
380 kV interconnectienet met hoogspanningsstations

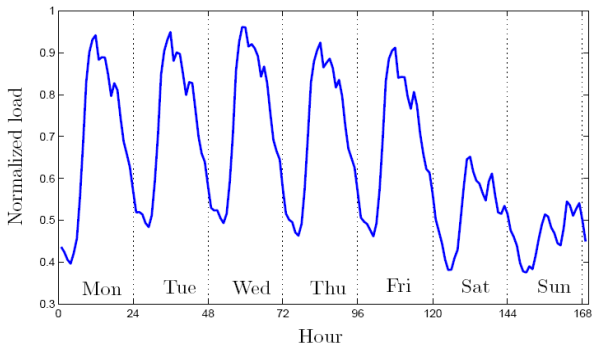
Electric load forecasting

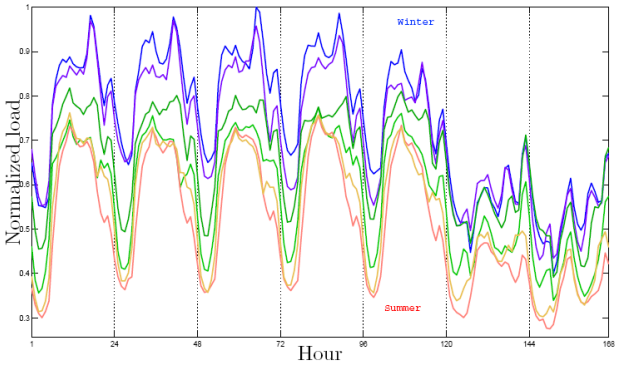
Data





250 transformer substations
Every 15 min, 5 years



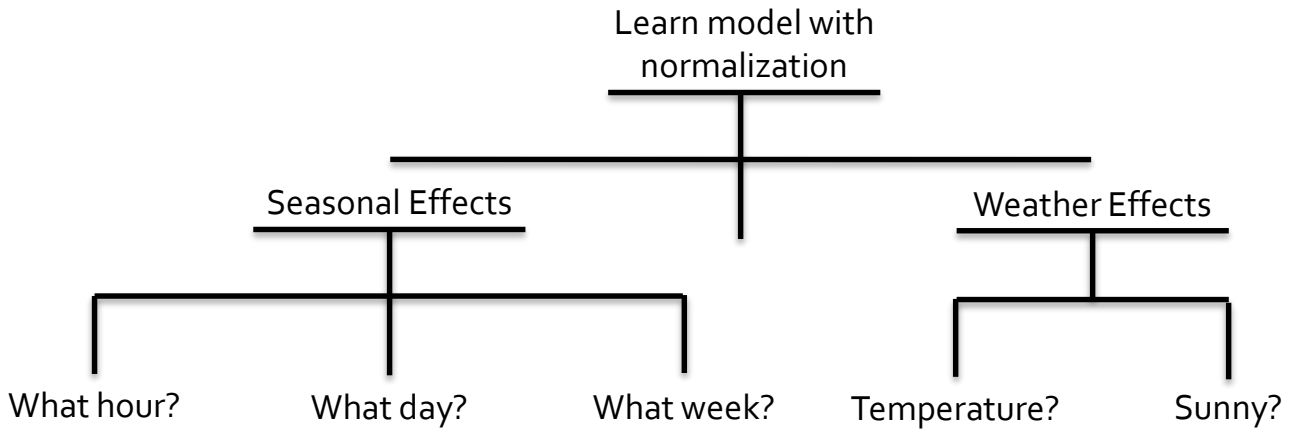


1 post, 1 week

1 post, four seasons

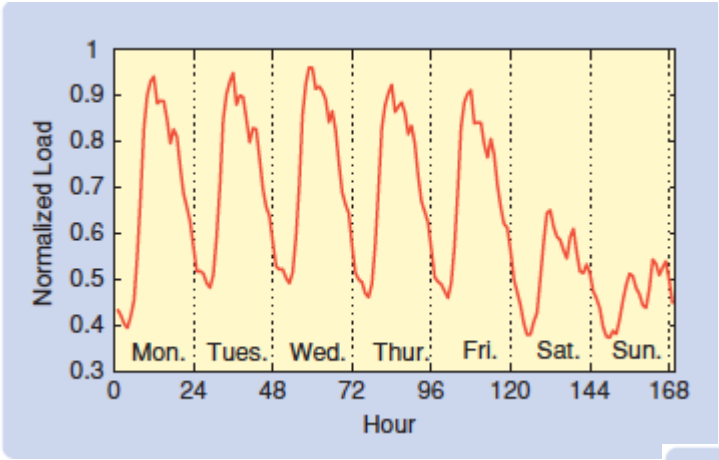
Electric load forecasting

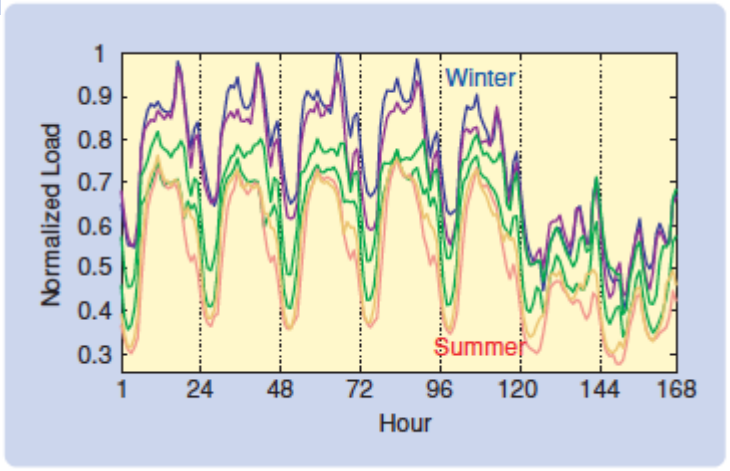
Expertise in Action

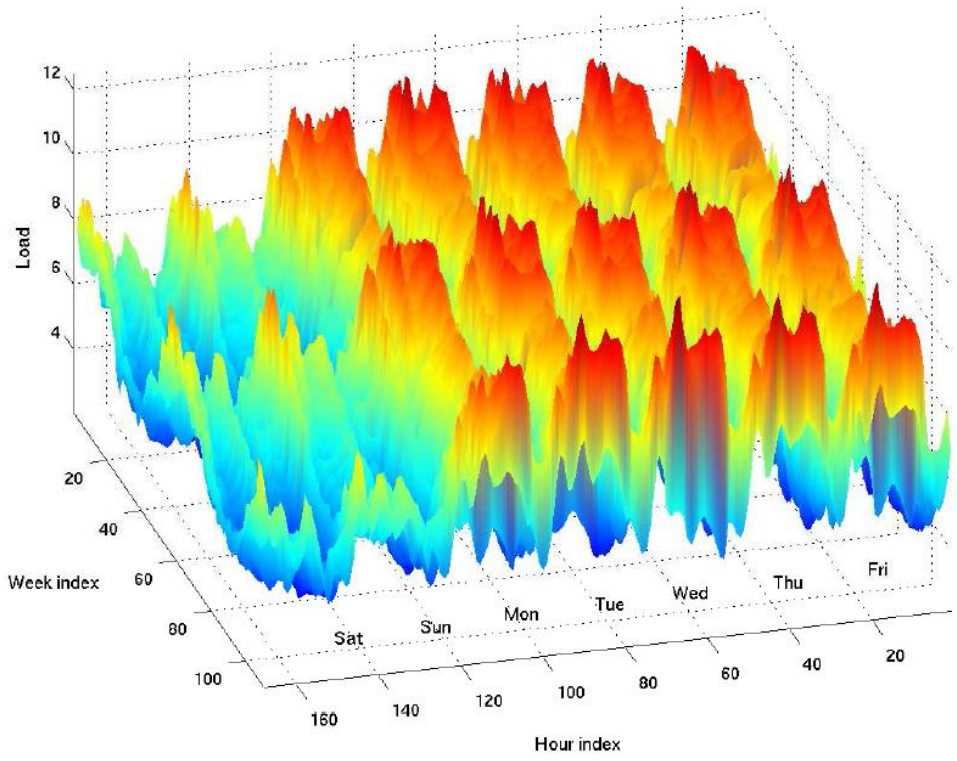


Model update: Every week!

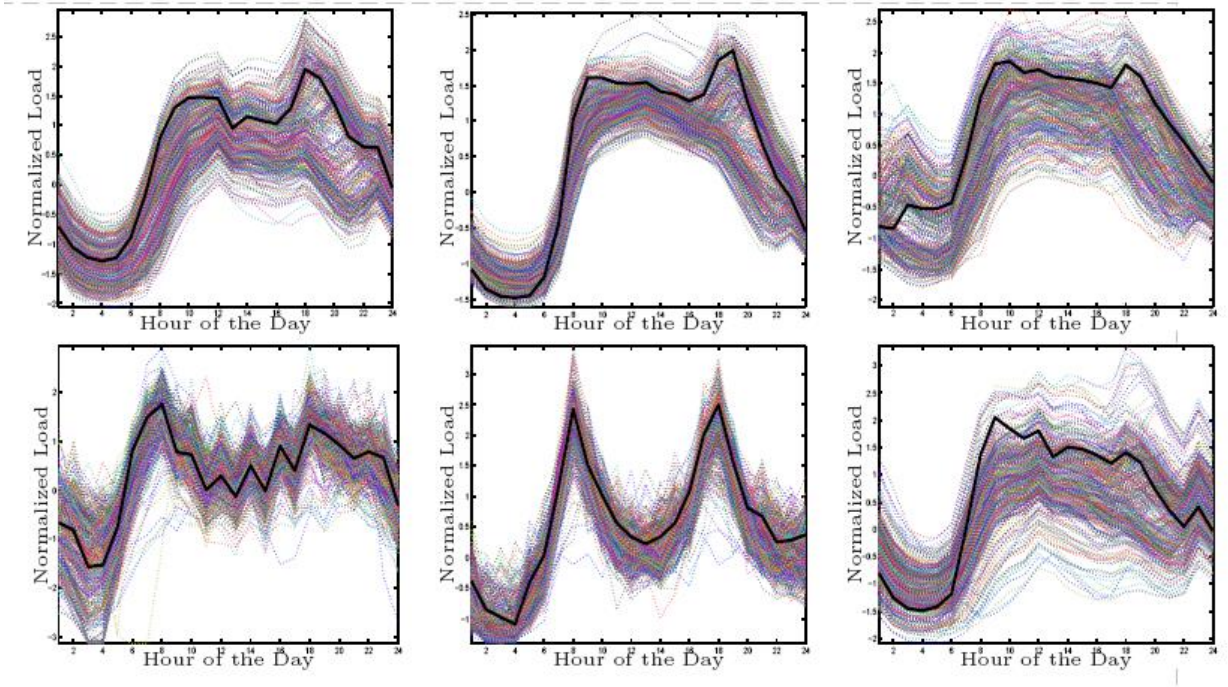
Electric load forecasting







Seasonalities in the load: day, week, year, holidays



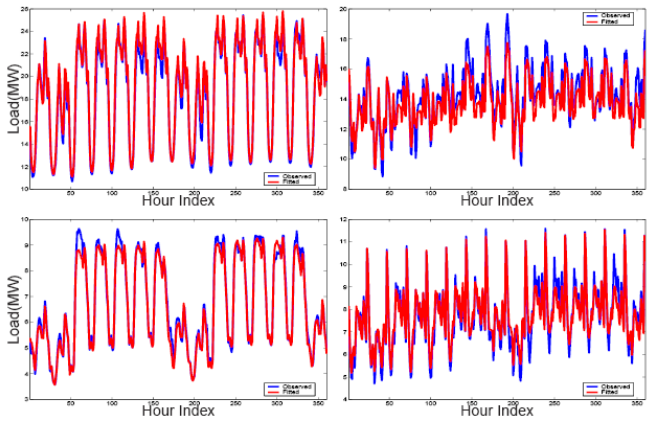
6 posts, 1 year
Seasonalities, calender holidays !

Electric load forecasting

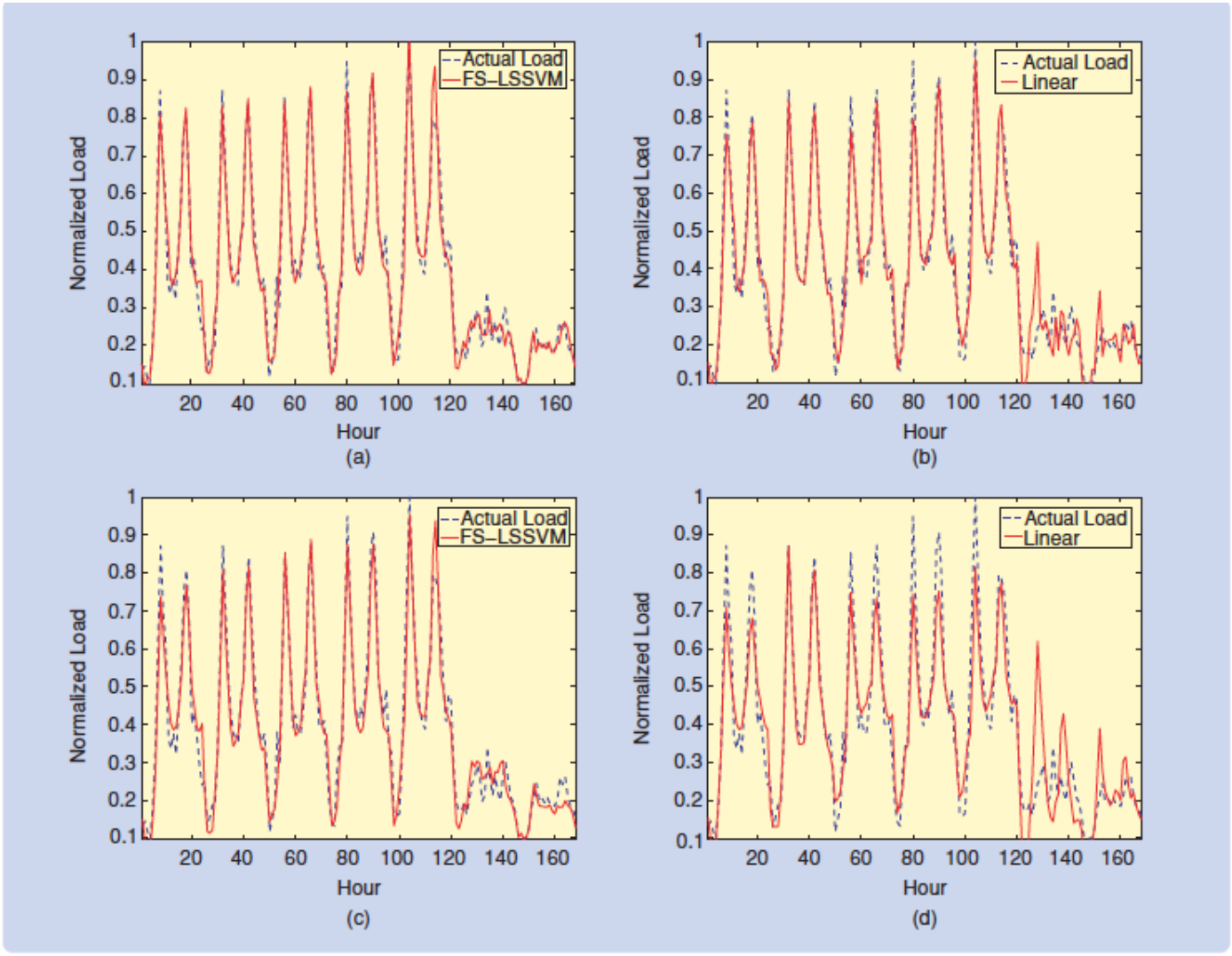
Problem Solved

- Seasonalities = a priori information (regress Monday on Monday !)
- Normalization:
  - remove effects of temperature, cloudiness,….
  - remove effect of holidays – calender days (use dummies)
- Calculate 'eigenprofiles' = daily shape per post

15-days ahead forecasts for 4 posts:



⇨ Accurate forecast


Actual Load
FS–LSSVM
Normalized Load
Hour
(a)
Actual Load
Linear
Normalized Load
Hour
(b)
Actual Load
FS–LSSVM
Normalized Load
Hour
(c)
Actual Load
Linear
Normalized Load
Hour
(d)

Energy

Industry

Environment

Social

Finance and Fraud

Health

Top 99.8 % (99.8%)
Bottom 0.31 % (0.2 %)
steam 20 T/h

Top 99.75 % (99.8%)
Bottom 0.25 % (0.2 %)
steam 20 T/h

Top 99 %
Bottom 1 %
steam 15.2 T/h

Condensor
SPLc = .5 m
Lc = .5 m
Fref = 12.66 T/h
4
OK
FF = 4 T/h
50
OK
XF = 50 %
SPLb = .7 m
Lb = .7 m
Fst = 20 T/h
Reboiler

INCAView - [Overview]
View Window Help
Log Status
Idle
Controller Status, Reason
Turned on by operator request

All | CVs | MVs | DVs | Miscellaneous

CV NAME	ENGLOW	OPERLOW	IDEAL	IDEALRANK	OPERUPP	ENGUPP
linX_destil	-5.00	-2.61	-1.61	2	0.68	5.00
linX_bottom	-5.00	-2.00	-1.61	3	0.68	5.00

MV NAME	ENGLOW	OPERLOW	IDEAL	IDEALRANK
F_reflux	6.50	6.80	9.20	4
F_steam	1.00	2.50	3.00	4

DV NAME	DESCRIPTION	UNIT	PV	USE	CRIT	AUTO	BAD	LBND	UBND
Feedflow	Feed Flow	t/h	4.00	☑	☐	☑	☐	☐	☐

Setpoint changes

Idealrank top 3 -> 2

[Time] : 3/8/0 2:56
X_bottom.pv : 0.32
X_destil.pv : 99.800
F_steam.sp : 20.00

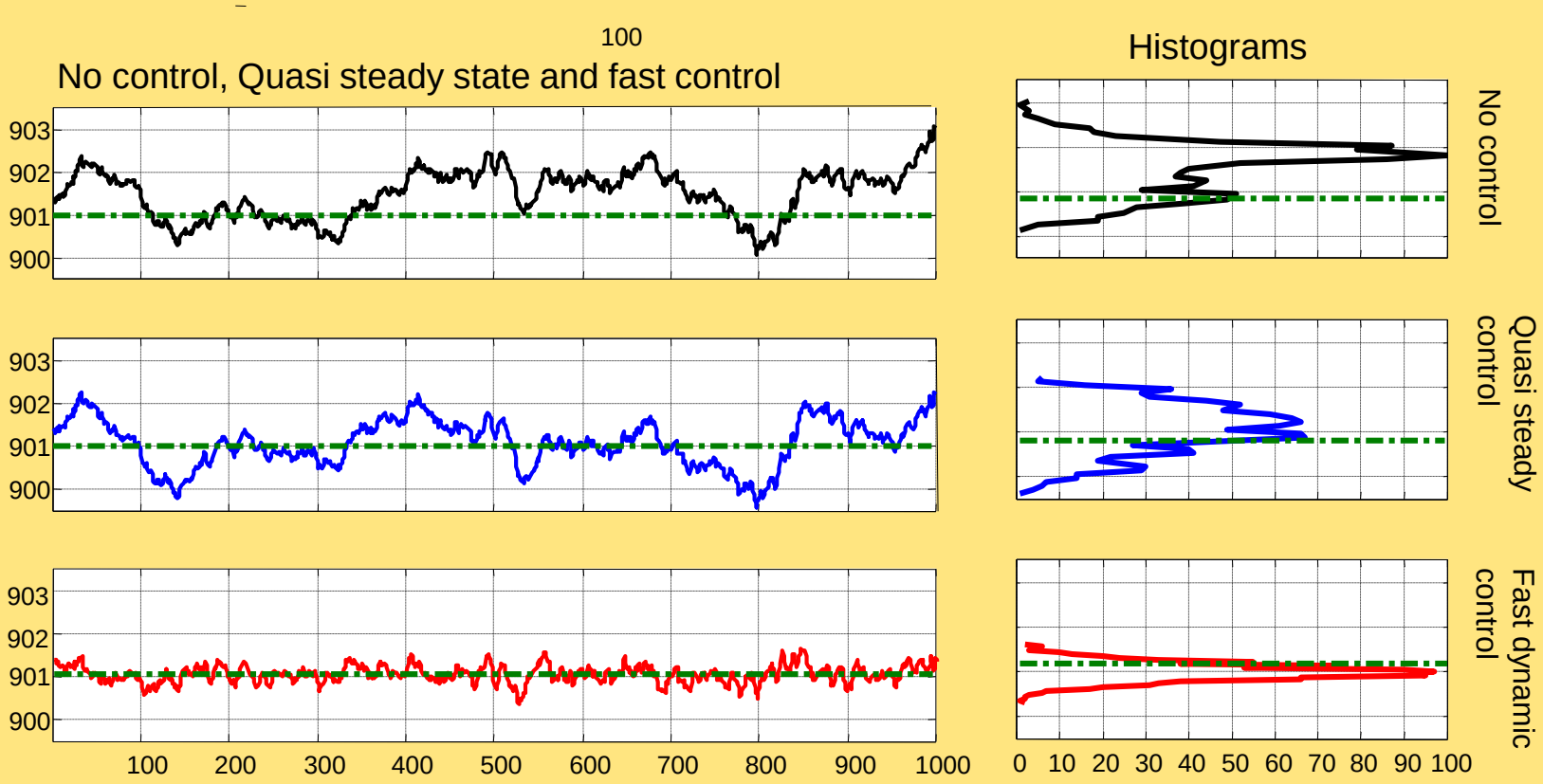
Status Idle
Test Status Idle

100.0
99.5
99.0
98.5
98.0

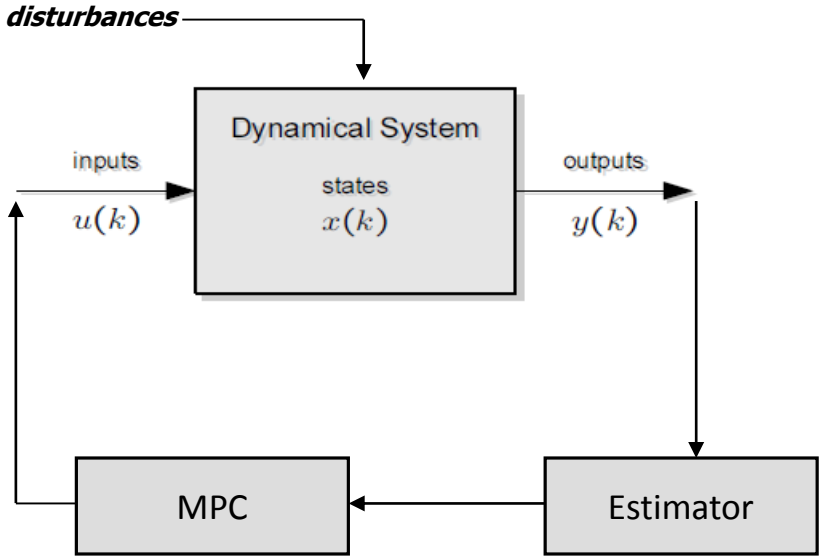
12:00 15:00 18:00 21:00 0:00 3:00
3/7/0 Time 3/8/0

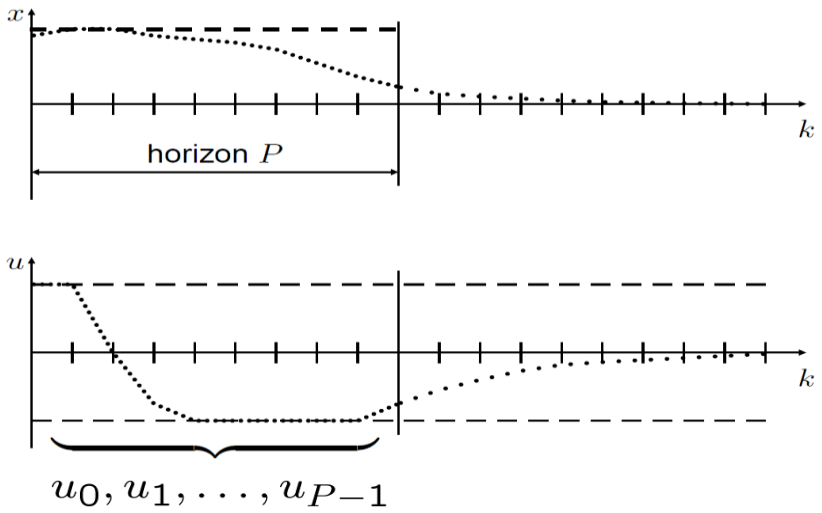
:44:15 PM | Last Run : 3/8/2000 3:05:15 AM | Count : 9322

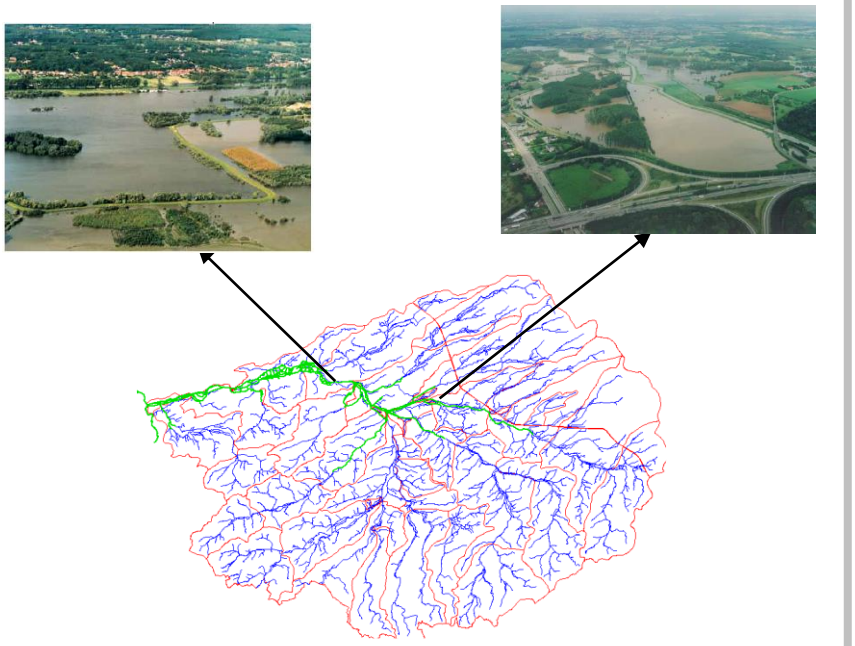
Modelling for control

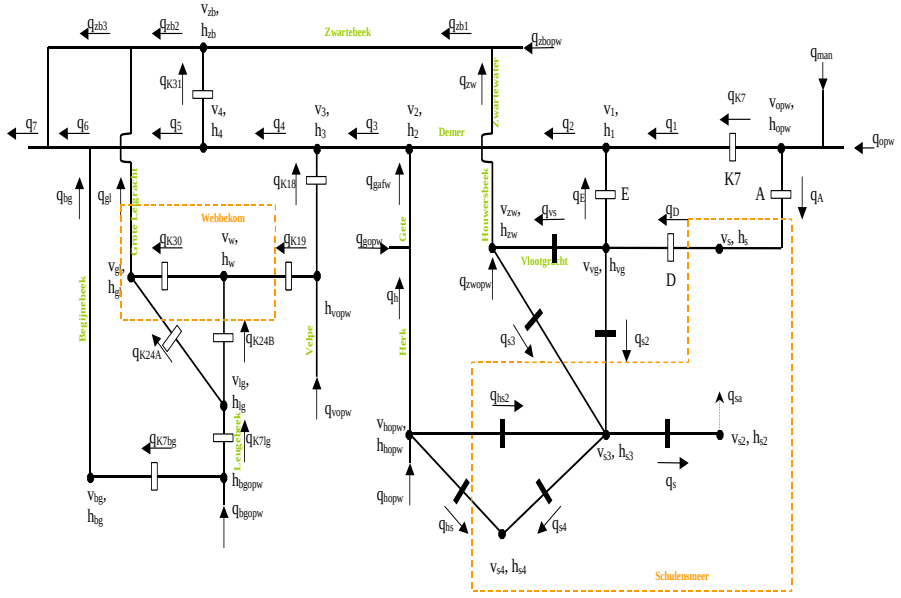


Model Based Predictive Control for Flood Regulation: Demer



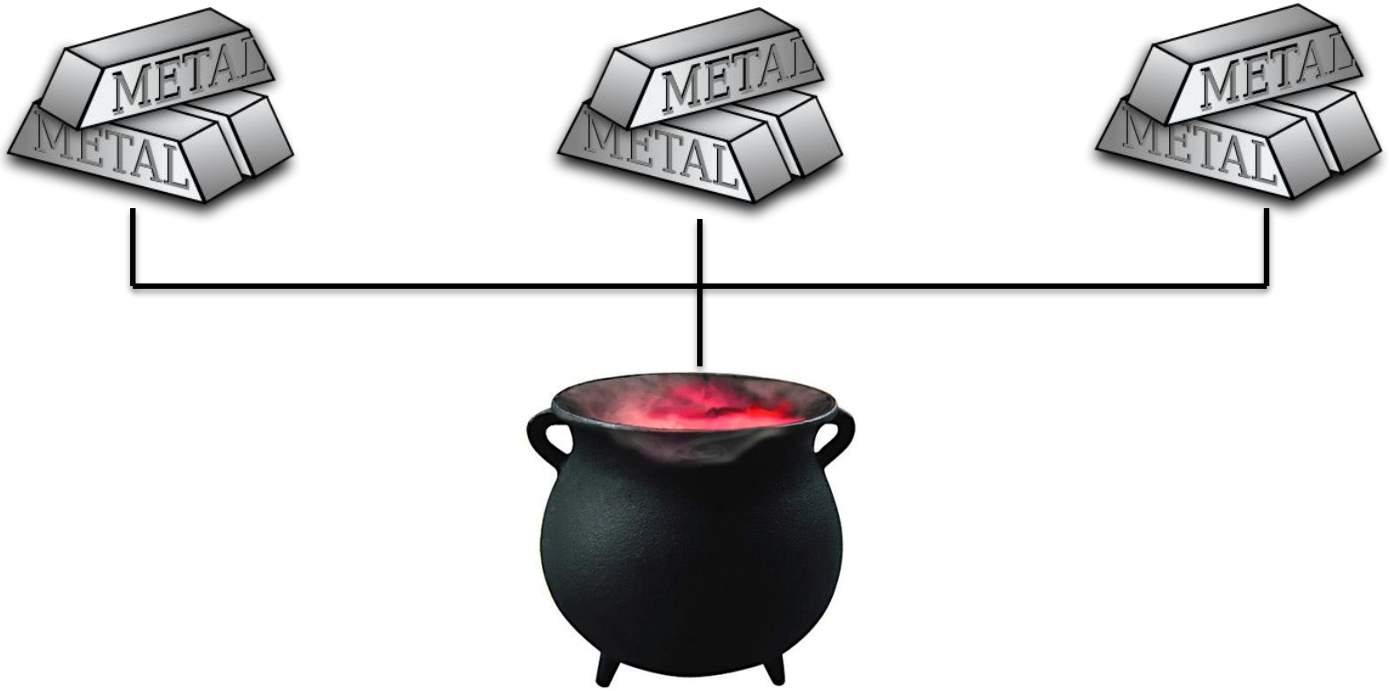






Alloy melting point modelling

Problem & Objectives



How to predict material properties when designing a new alloy?

Alloy melting point modelling

Data

InsPyro NV

Random measurements of melting point in function of constituent concentrations

10 different materials!

130 000 points sampled!

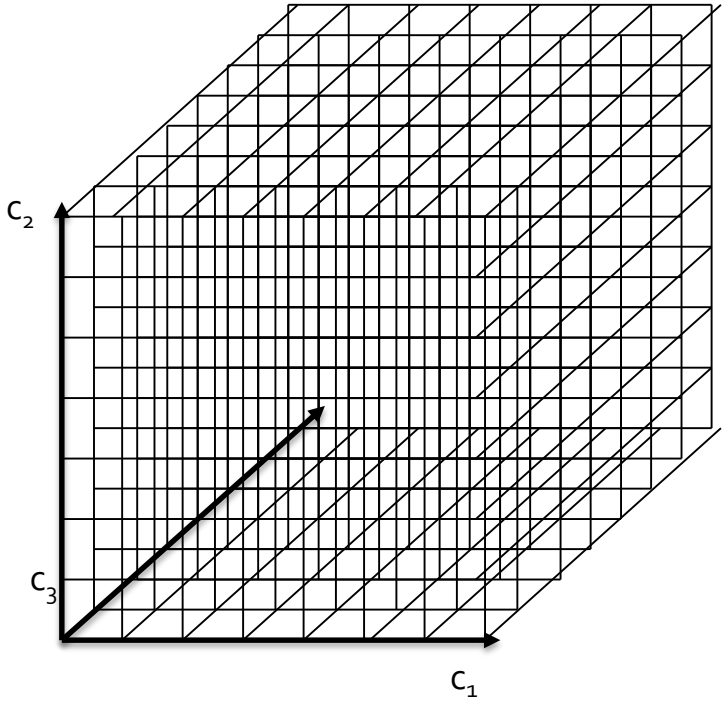
Many missing values.

Curse of Dimensionality!

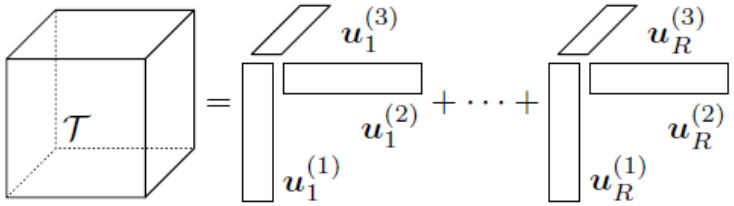
Alloy melting point modelling

Expertise In Action

Represent as tensor



Decompose in factors



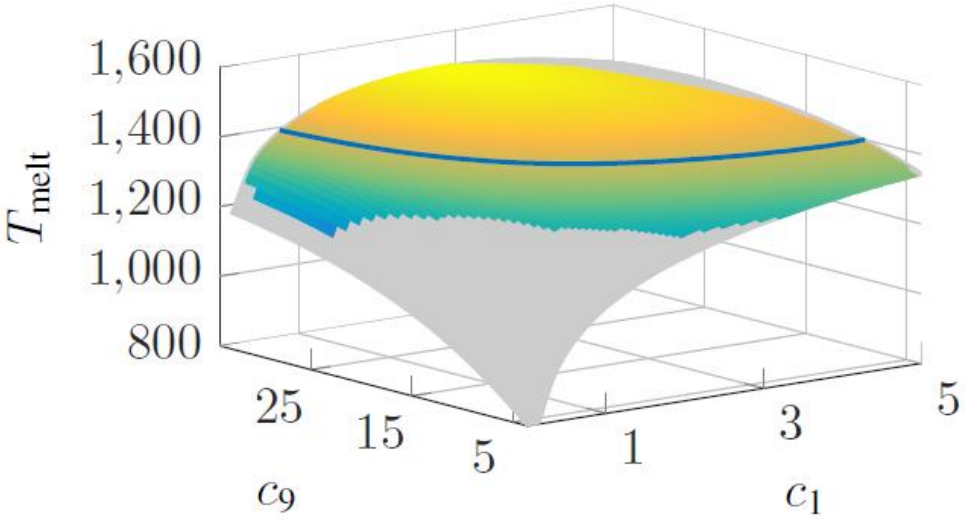
New model has only
4500 parameters

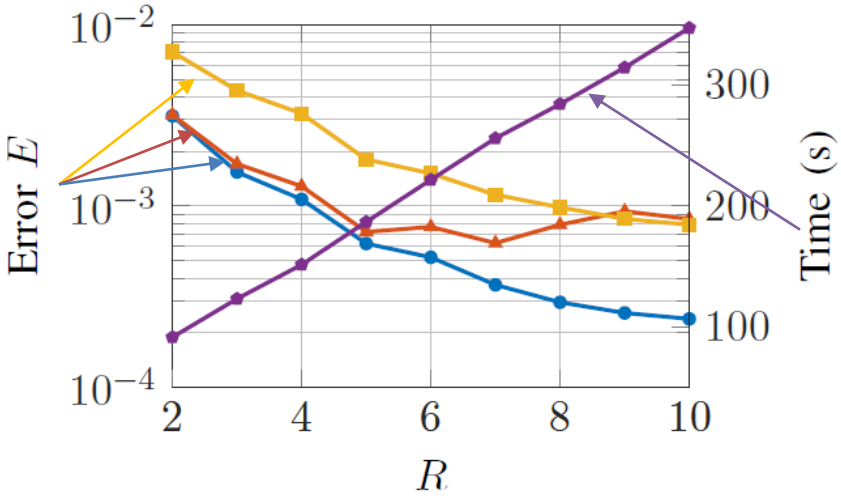
⇨ Curse of Dimensionality broken!

Alloy melting point modelling

Problem Solved

Economical prediction of new material properties





(Different Model Ranks)

Energy

Industry

Environment

Social networks

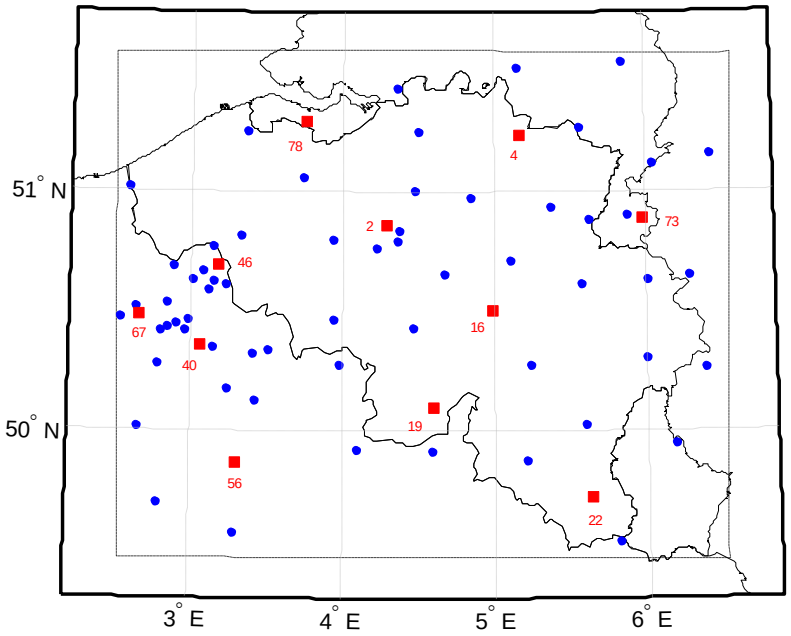
Finance and Fraud

Health

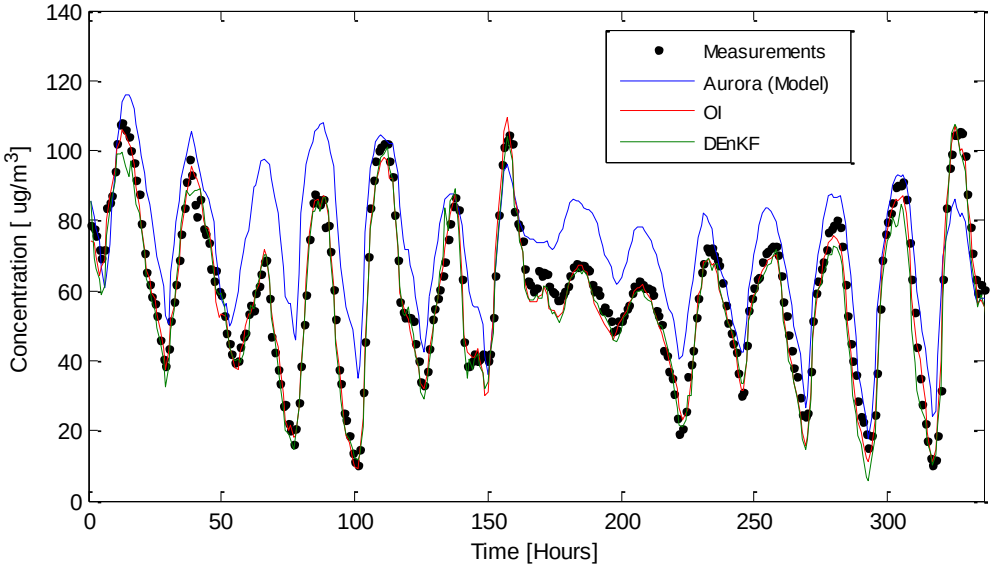
Data Assimilation

The Deterministic Ensemble Kalman Filter (DEnKF) and the OI technique have been used to improve the O_3 estimates of the Air-quality model AURORA.

O_3 air-quality stations



Average of the O_3 concentration over the validation stations

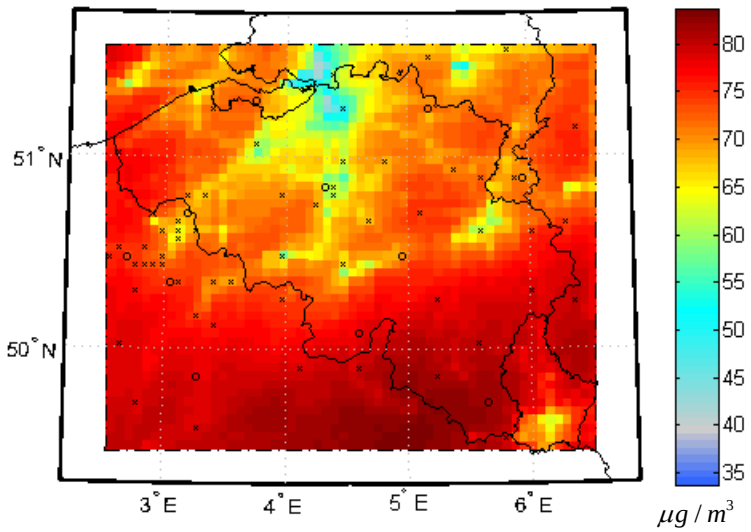


- Assimilation stations
- Validation stations

Starting date: May 28th, 2005 at midnight

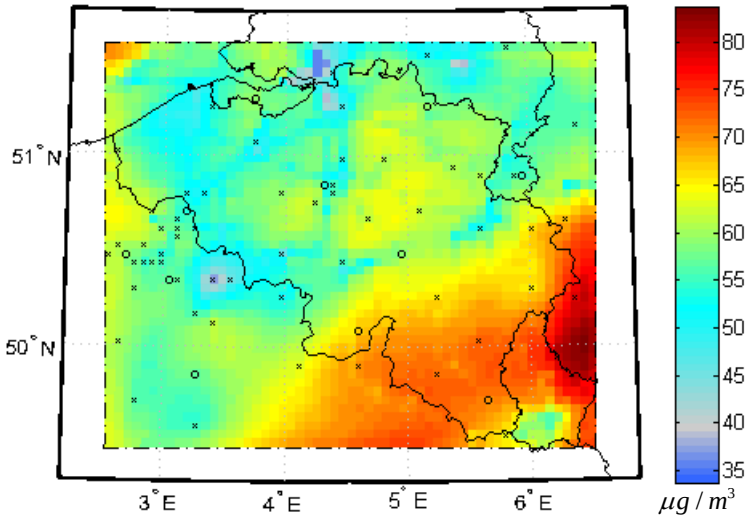
Average of the O_3 concentration field over the 14 day period

Free-run of Aurora

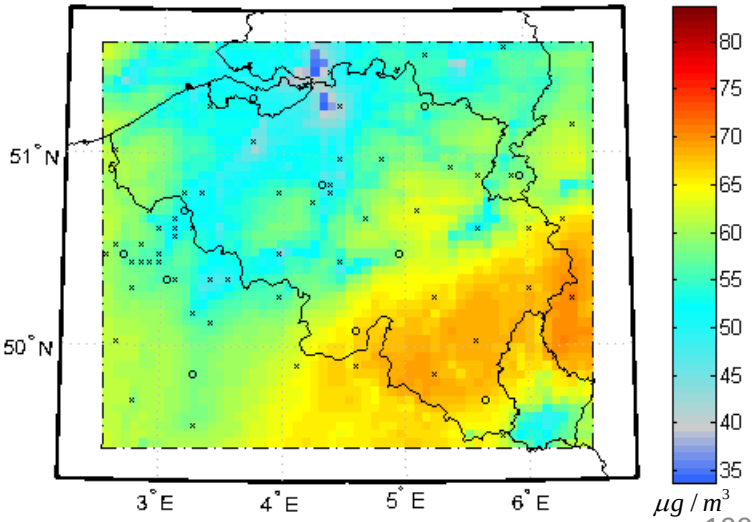


x - Assimilation station
o - Validation station

Optimal Interpolation



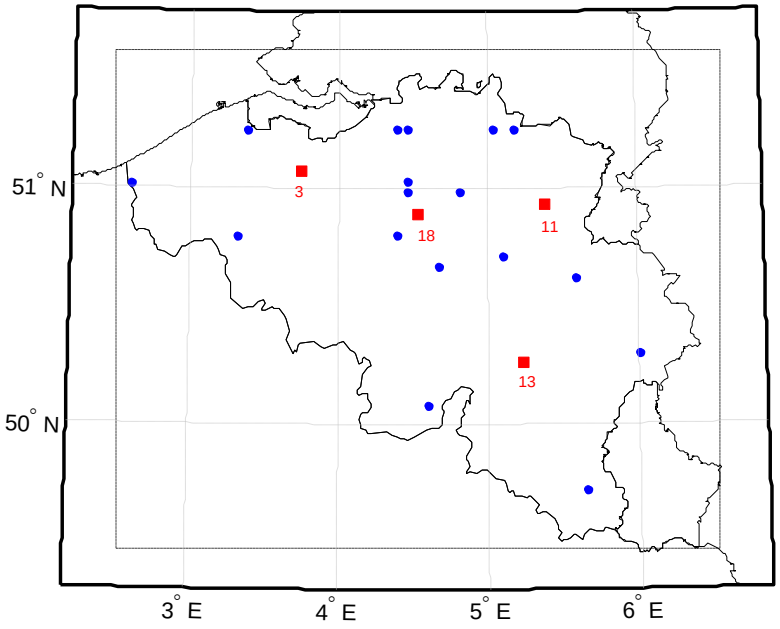
DEnKF



Data Assimilation

The Deterministic Ensemble Kalman Filter (DEnKF) and the OI technique have been used to improve the PM_{10} estimates of the Air-quality model AURORA.

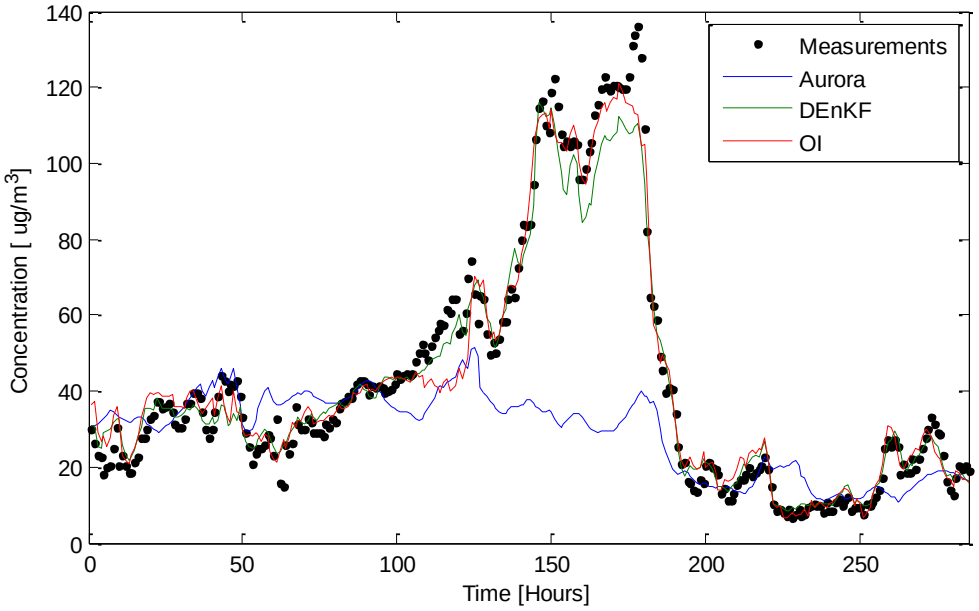
PM_{10} air-quality stations



● - Assimilation stations

■ - Validation stations

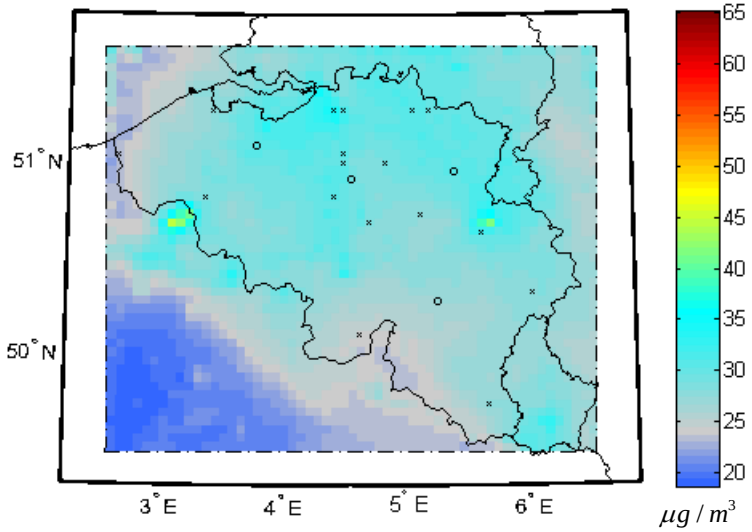
Average of the PM_{10} concentration over the validation stations



Starting date: January 20th, 2010 at midnight

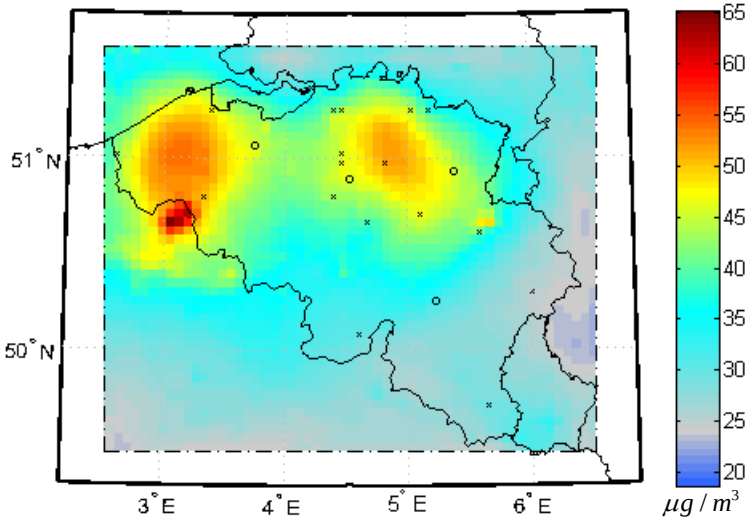
Average of the PM_{10} concentration field

Free-run of Aurora

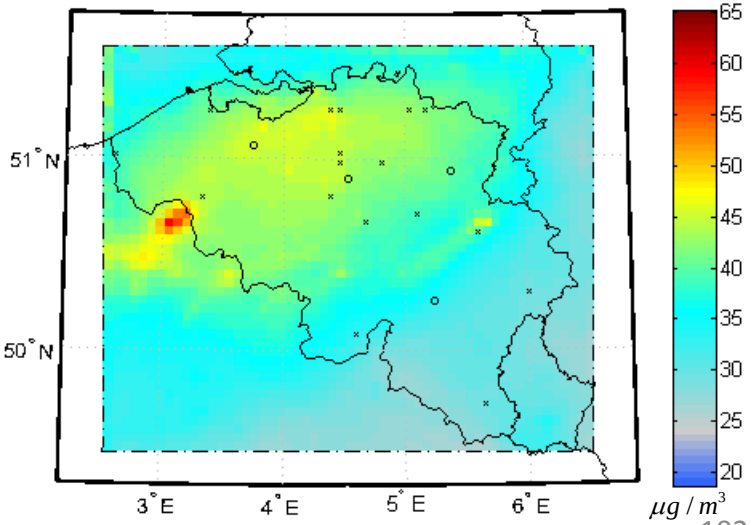


x - Assimilation station
o - Validation station

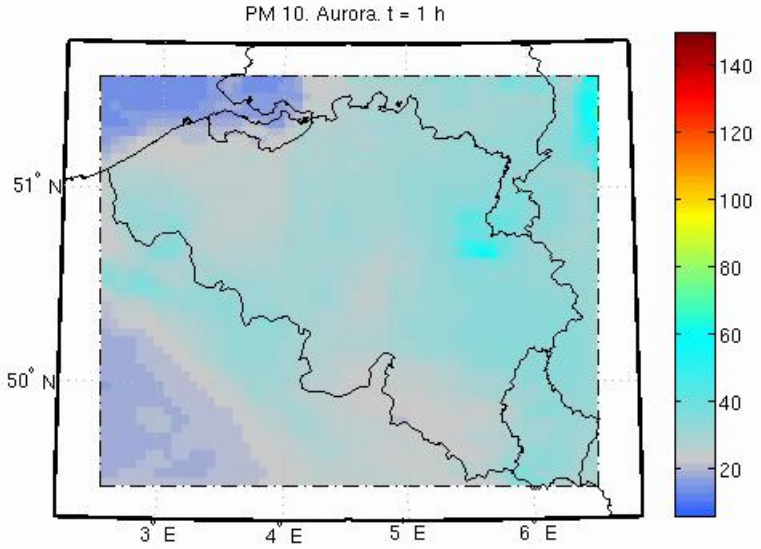
Optimal Interpolation

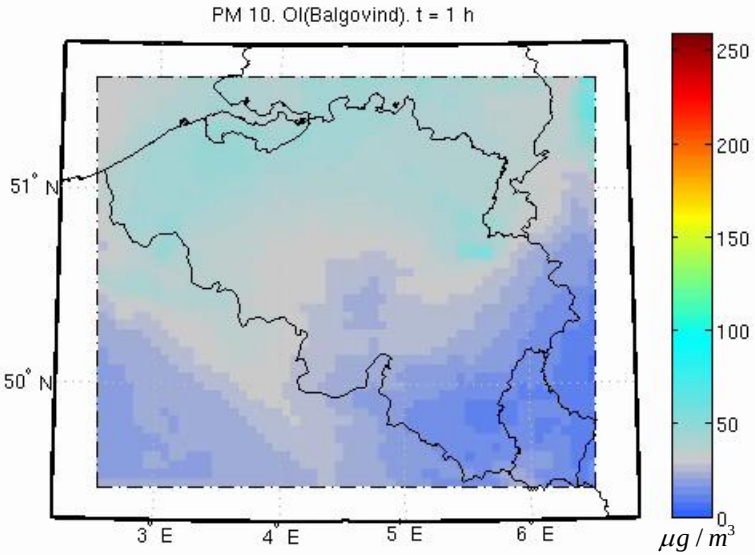


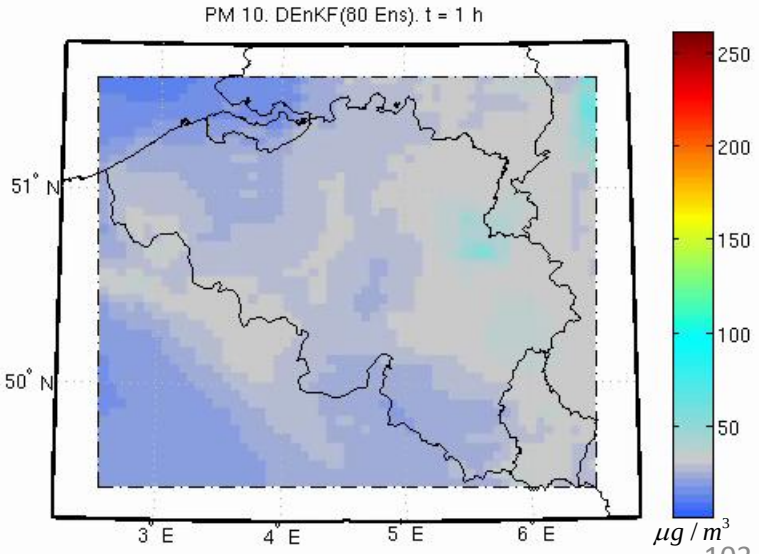
DEnKF



Average of the PM_{10} concentration field







Energy

Industry

Environment

Social networks

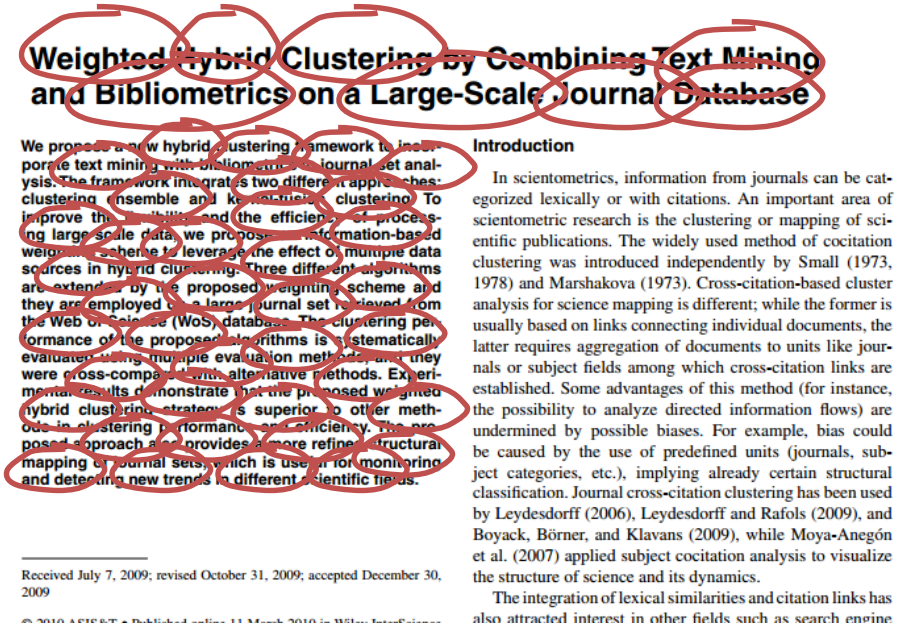
Finance and Fraud

Health

Journal Clustering

Find all about specific topic?

Journal Clustering



Weighted Hybrid Clustering by Combining Text Mining and Bibliometrics on a Large-Scale Journal Database

We propose a new hybrid clustering framework to incorporate text mining with bibliometrics for journal set analysis. The framework integrates two different approaches: clustering ensemble and kernel-fusion clustering. To improve the scalability and the efficiency of processing large scale data, we propose information-based weighting scheme to leverage the effect of multiple data sources in hybrid clustering. Three different algorithms are extended by the proposed weighting scheme and they are employed to a large journal set retrieved from the Web of Science (WoS) database. The clustering performance of the proposed algorithms is systematically evaluated using multiple evaluation methods, and they were cross-compared with alternative methods. Experimental results demonstrate that the proposed weighted hybrid clustering strategy is superior to other methods in clustering performance and efficiency. The proposed approach also provides a more refined structural mapping of journal sets, which is useful for monitoring and detecting new trends in different scientific fields.

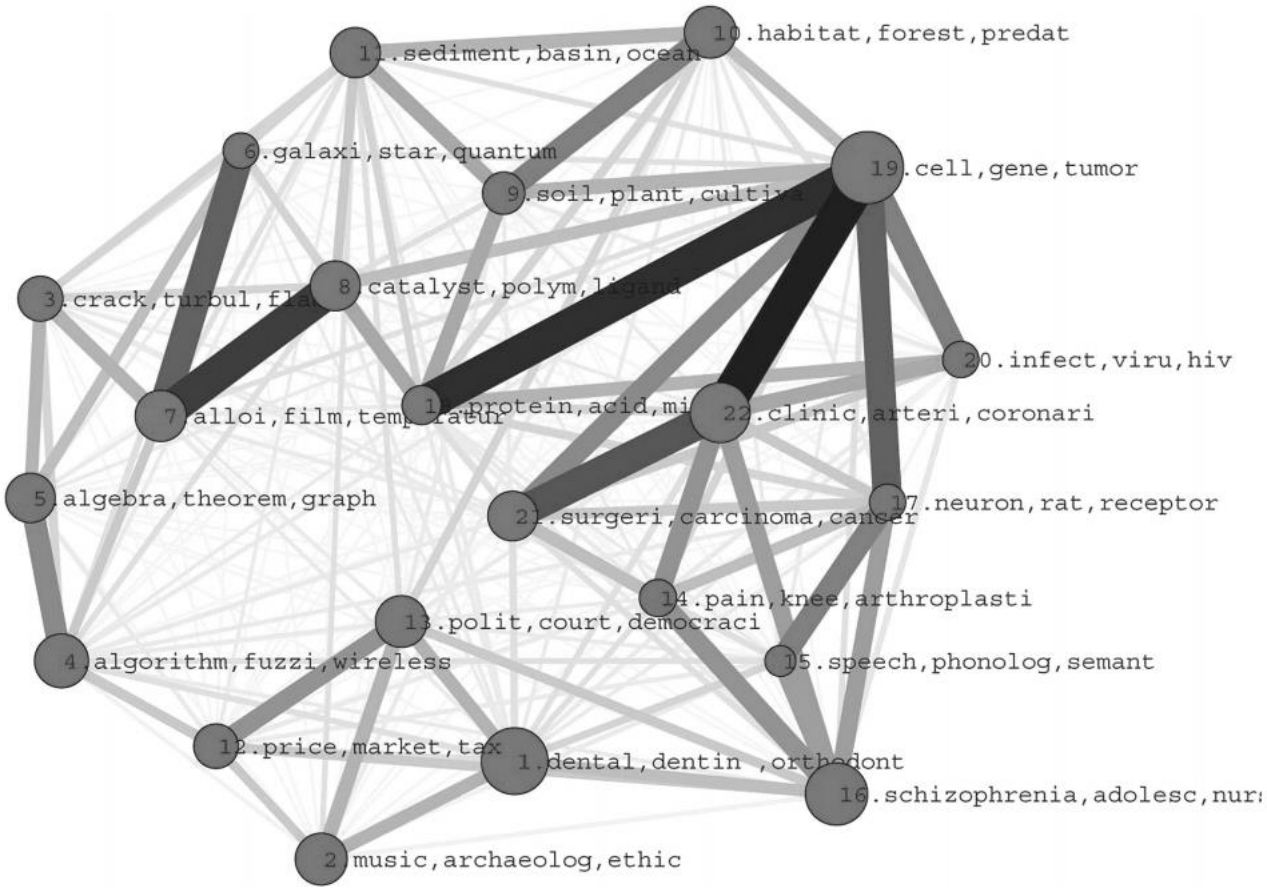
Received July 7, 2009; revised October 31, 2009; accepted December 30, 2009

Introduction

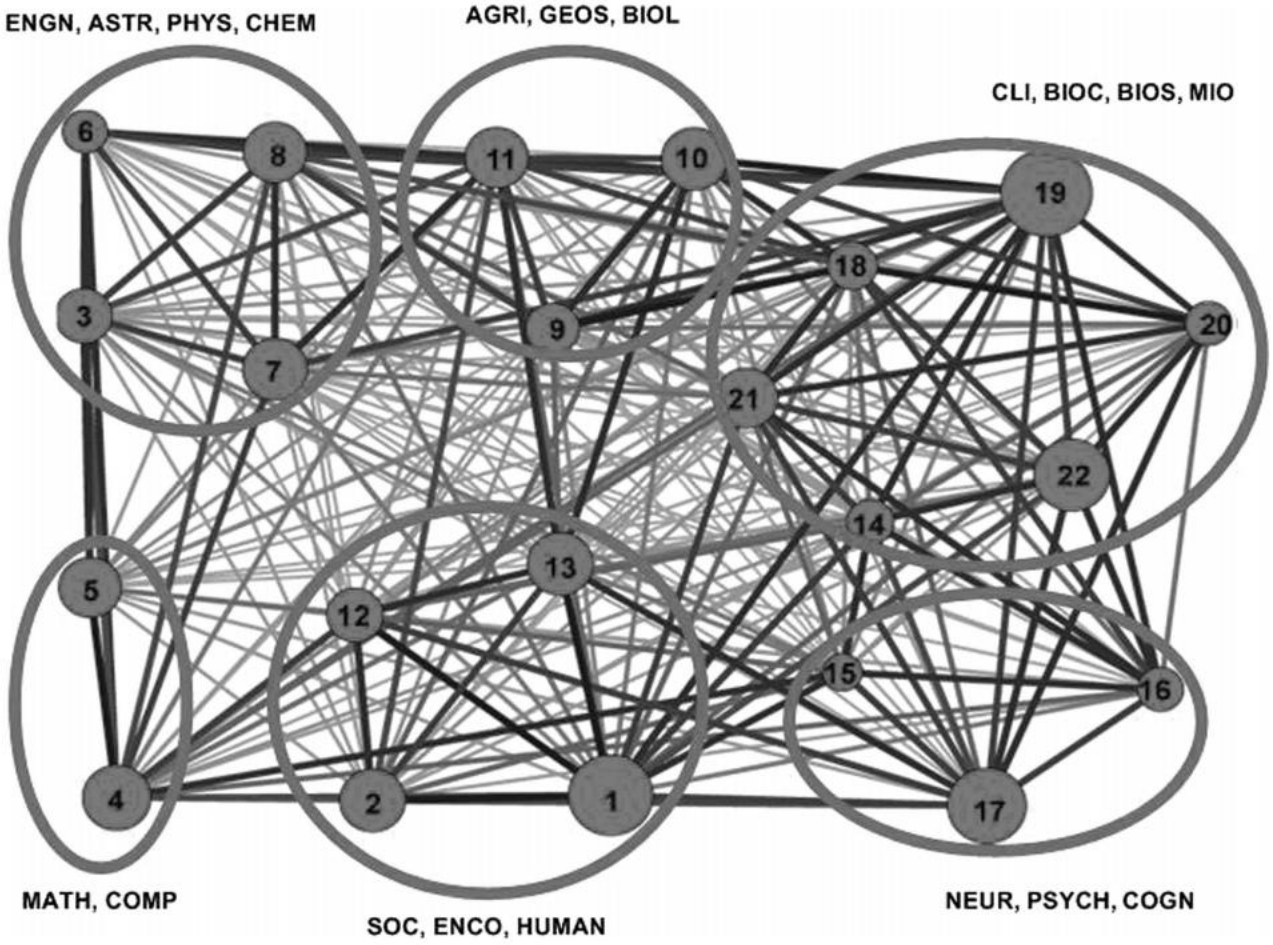
In scientometrics, information from journals can be categorized lexically or with citations. An important area of scientometric research is the clustering or mapping of scientific publications. The widely used method of cocitation clustering was introduced independently by Small (1973, 1978) and Marshakova (1973). Cross-citation-based cluster analysis for science mapping is different; while the former is usually based on links connecting individual documents, the latter requires aggregation of documents to units like journals or subject fields among which cross-citation links are established. Some advantages of this method (for instance, the possibility to analyze directed information flows) are undermined by possible biases. For example, bias could be caused by the use of predefined units (journals, subject categories, etc.), implying already certain structural classification. Journal cross-citation clustering has been used by Leydesdorff (2006), Leydesdorff and Rafols (2009), and Boyack, Börner, and Klavans (2009), while Moya-Anegón et al. (2007) applied subject cocitation analysis to visualize the structure of science and its dynamics.

The integration of lexical similarities and citation links has also attracted interest in other fields such as search engine

Journal Clustering

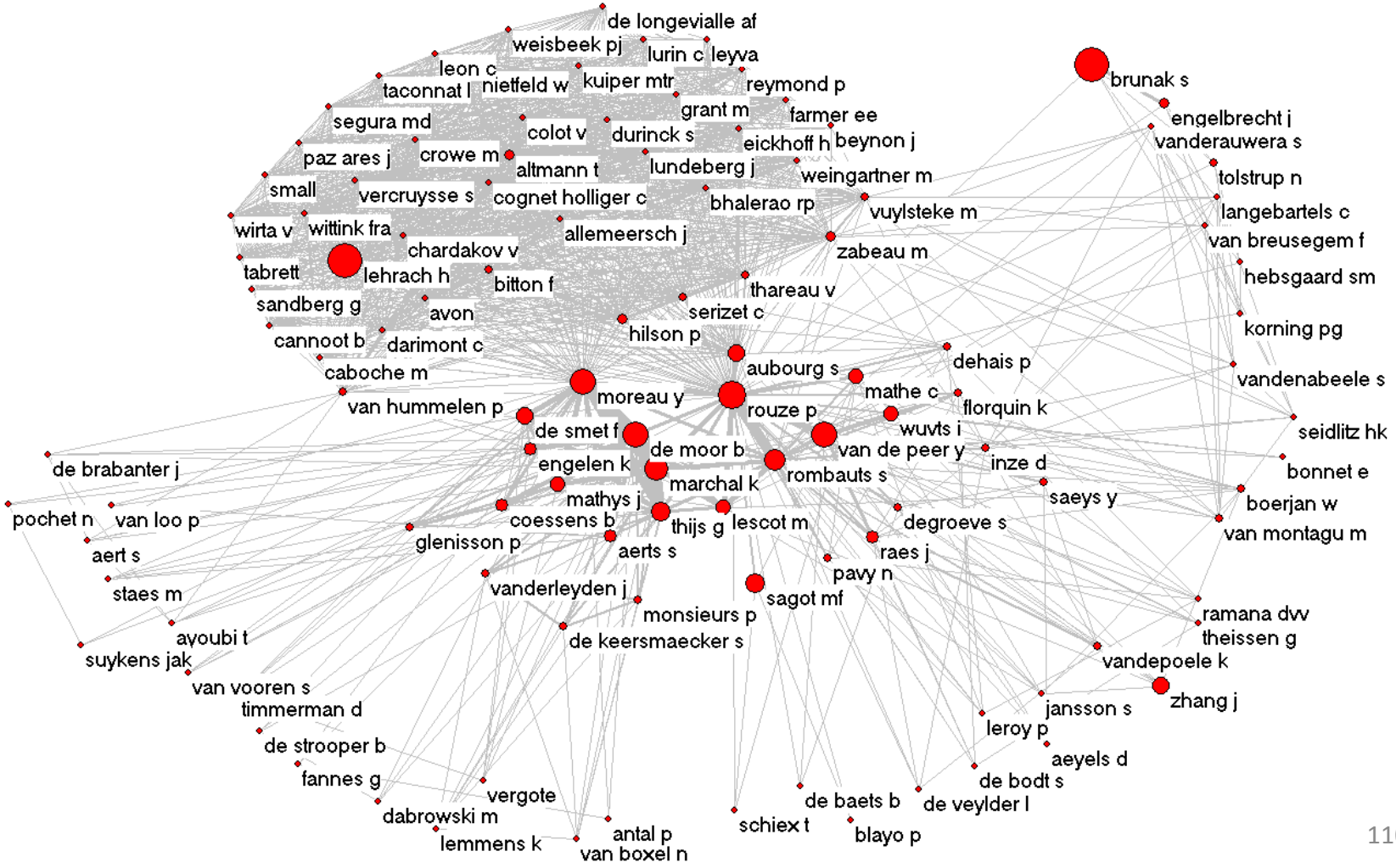


Journal Clustering

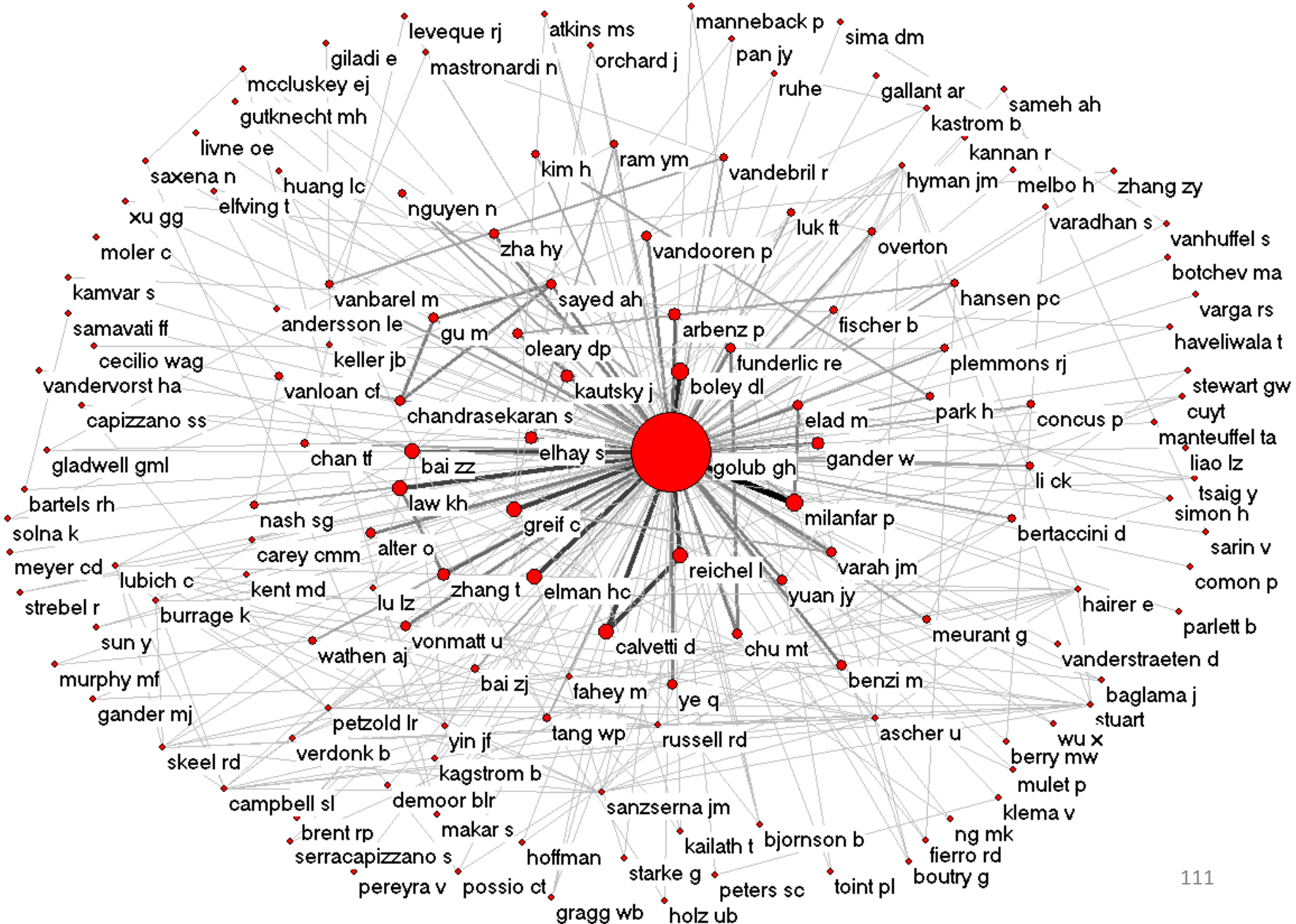


Author Collaboration Clustering

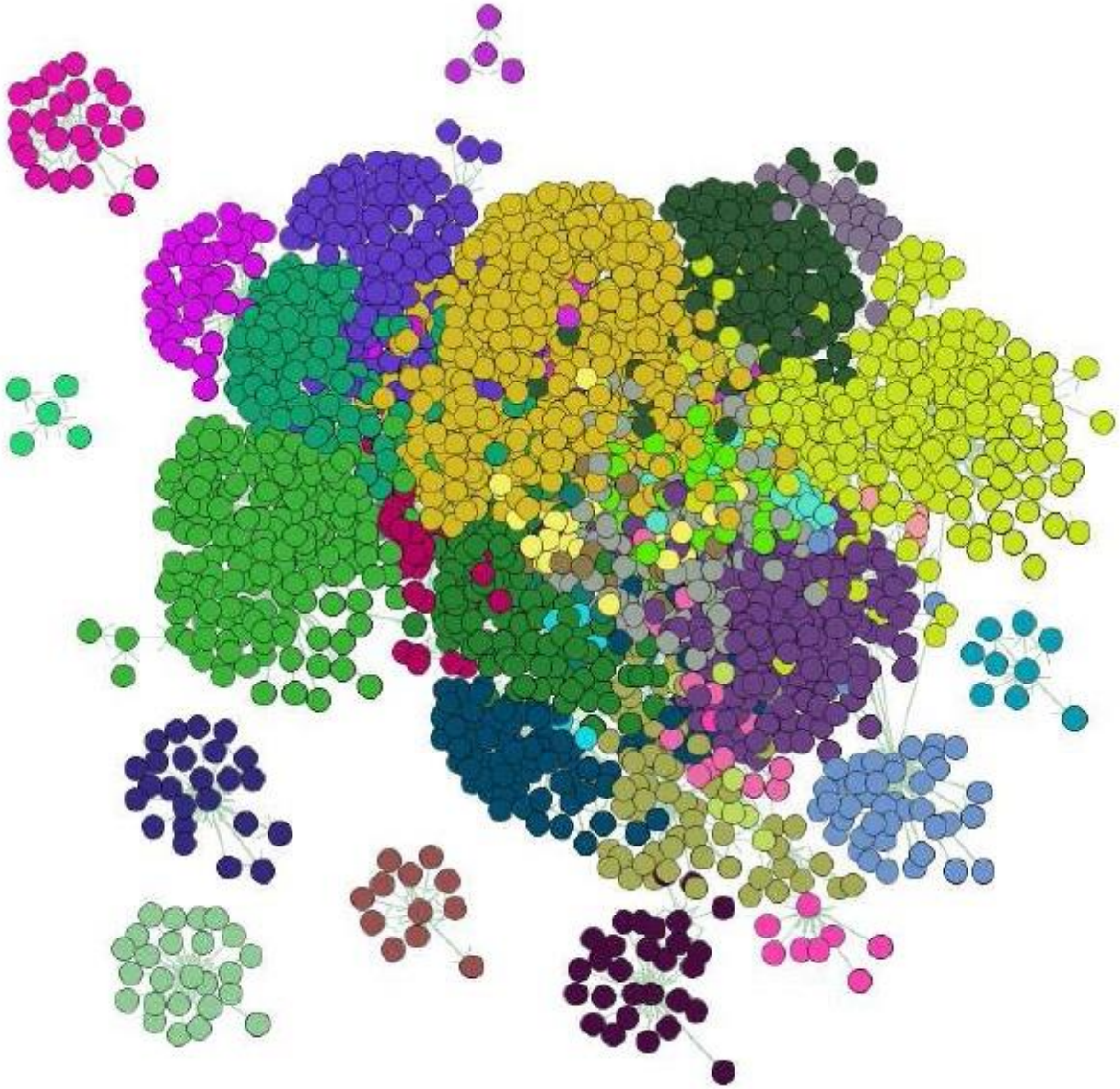
Author Collaboration Clustering



Golub around the world commemoration February 29 2008



Web of Science based literature network for Lennart Ljung



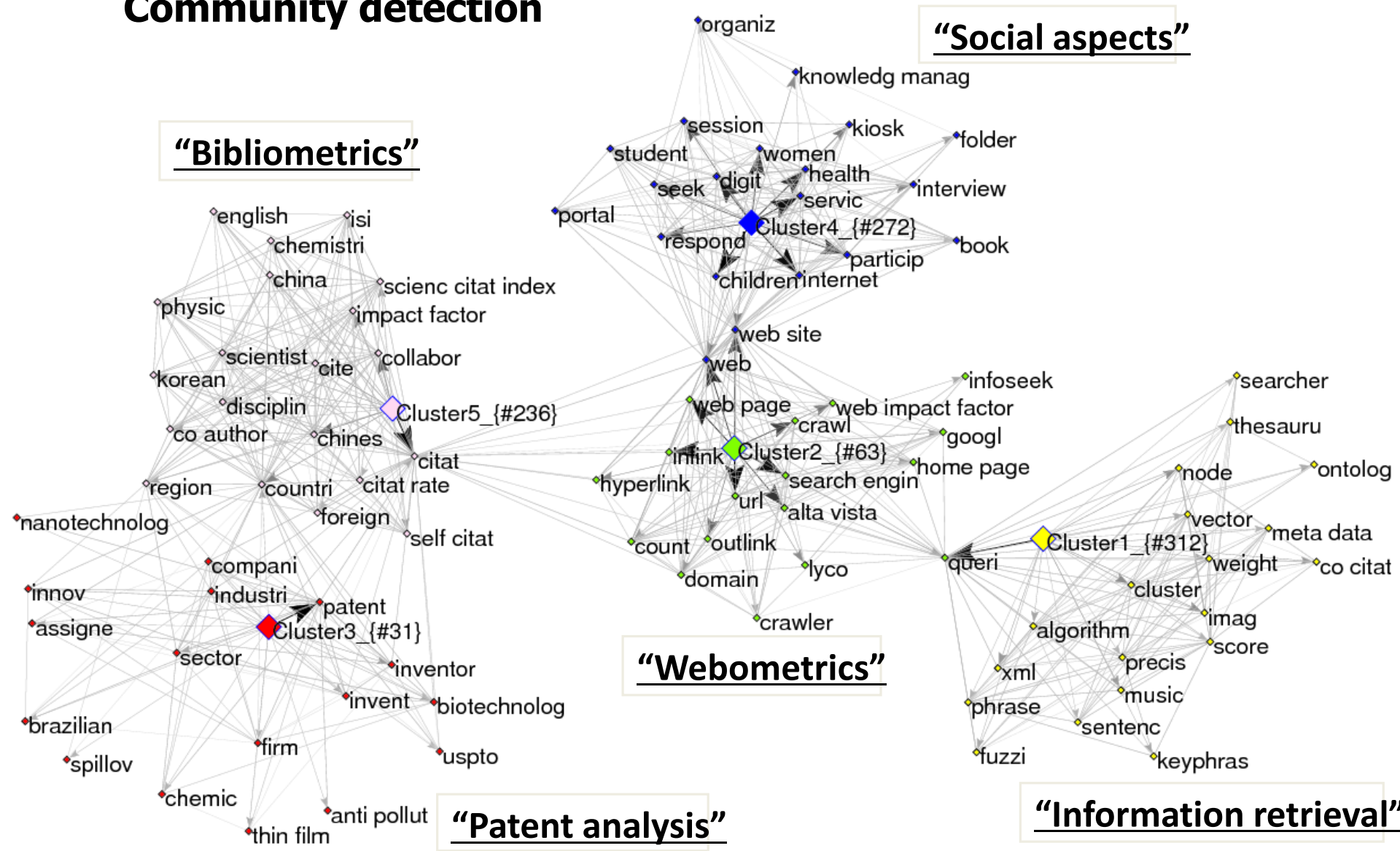
138 seed papers
+ all cited and citing publications

Result: 4943 nodes, 6216 edges

Link based clustering identifies topically homogeneous clusters.

13 papers are written by another L. Ljung.

Community detection



Energy

Industry

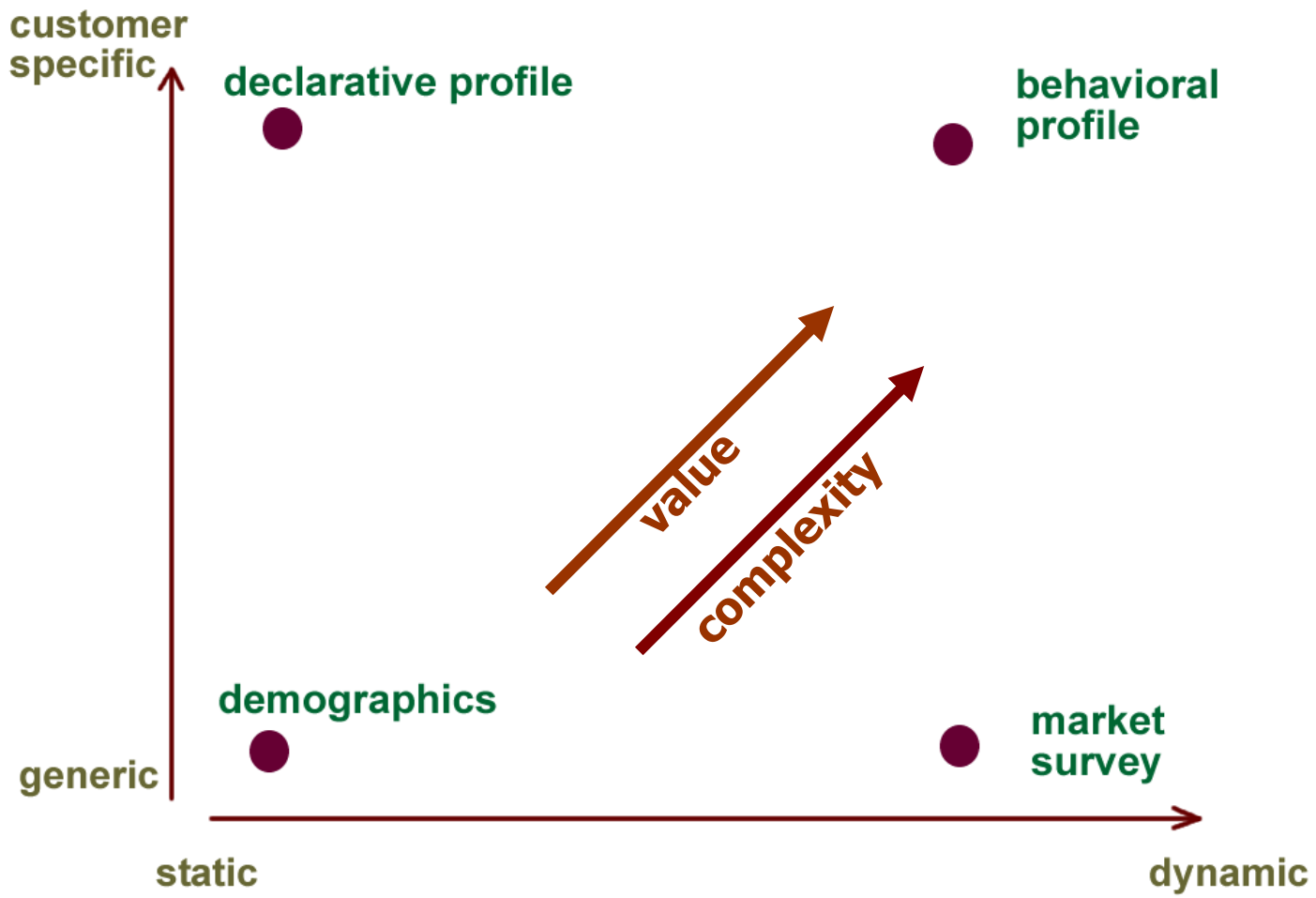
Environment

Education

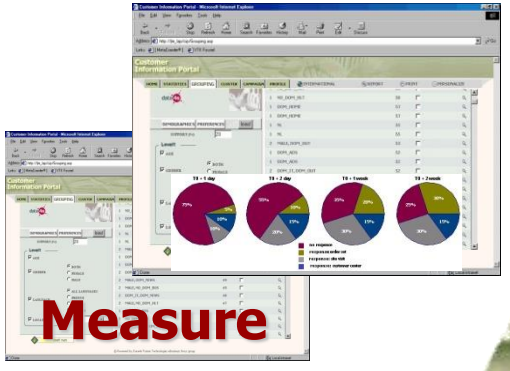
Finance and Fraud

Health

Customer Intelligence

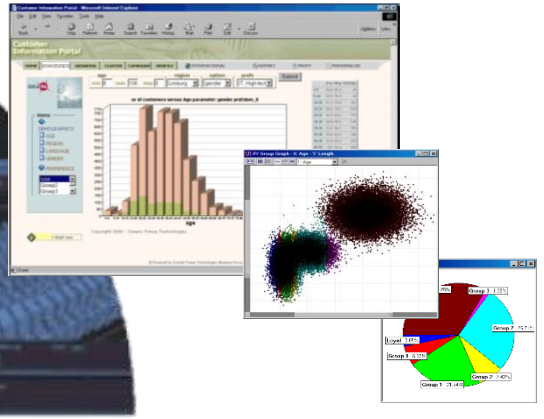


Customer Intelligence

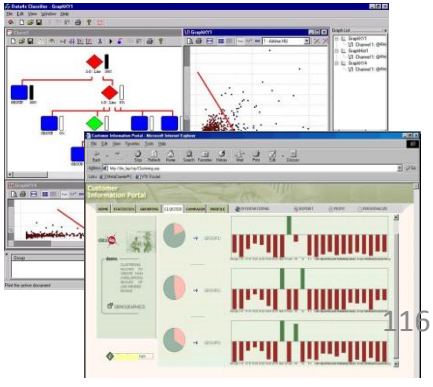


Measure response

Analyze & Model

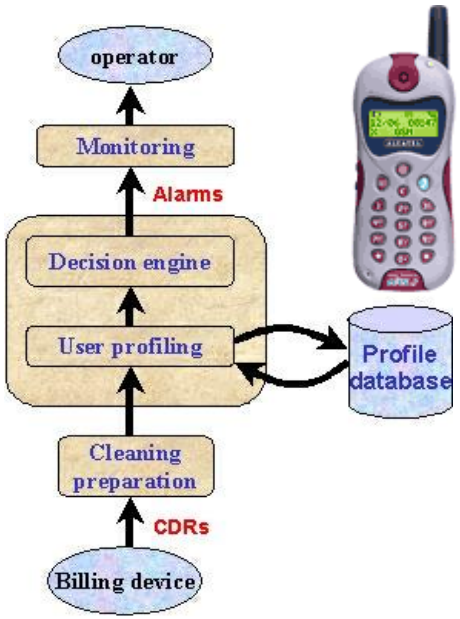


Score & Target

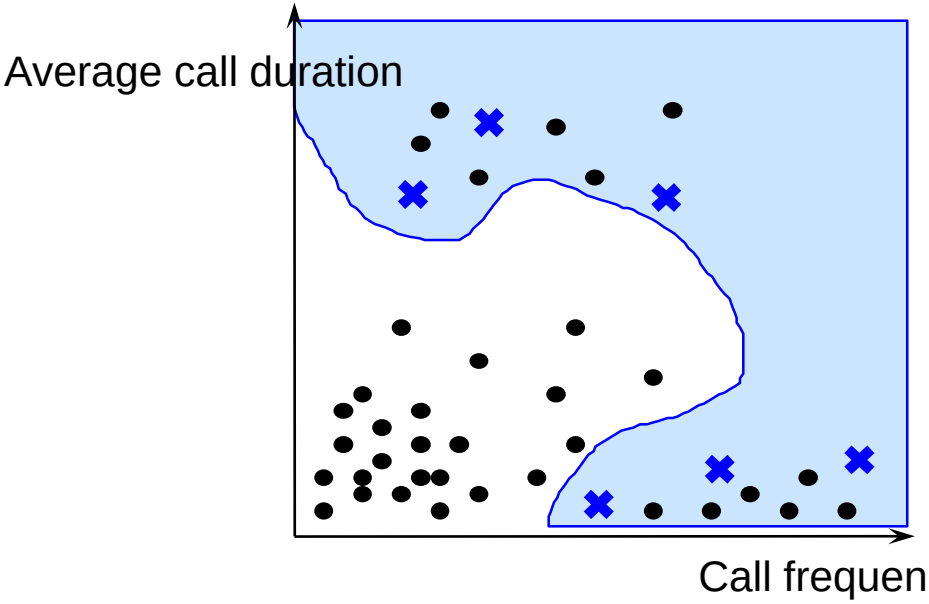


Contact

Fraud detection on mobile phone network

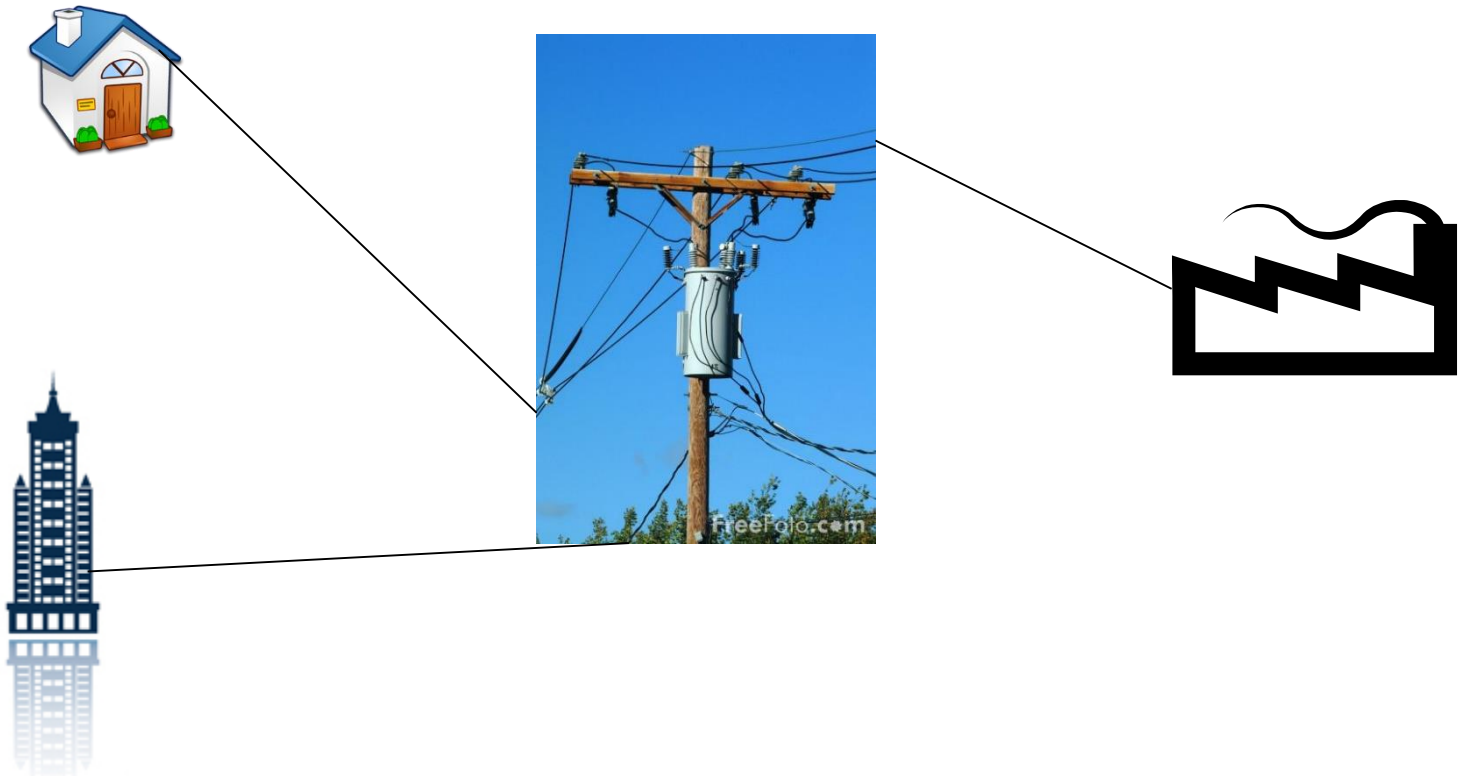


	Short Duration	Long Duration	High Frequency	International	Same Destination	Off Peak	Call Forwarding	Behaviour Change
irect call elling		X	X	X			X	
ABX fraud	X		X		X	X		X
reephone fraud	X		X		X			X
Premium rate fraud		X	X		X			X
Subscription fraud			X					
Handset theft		X	X	X	X			X



Electric Market Segmentation

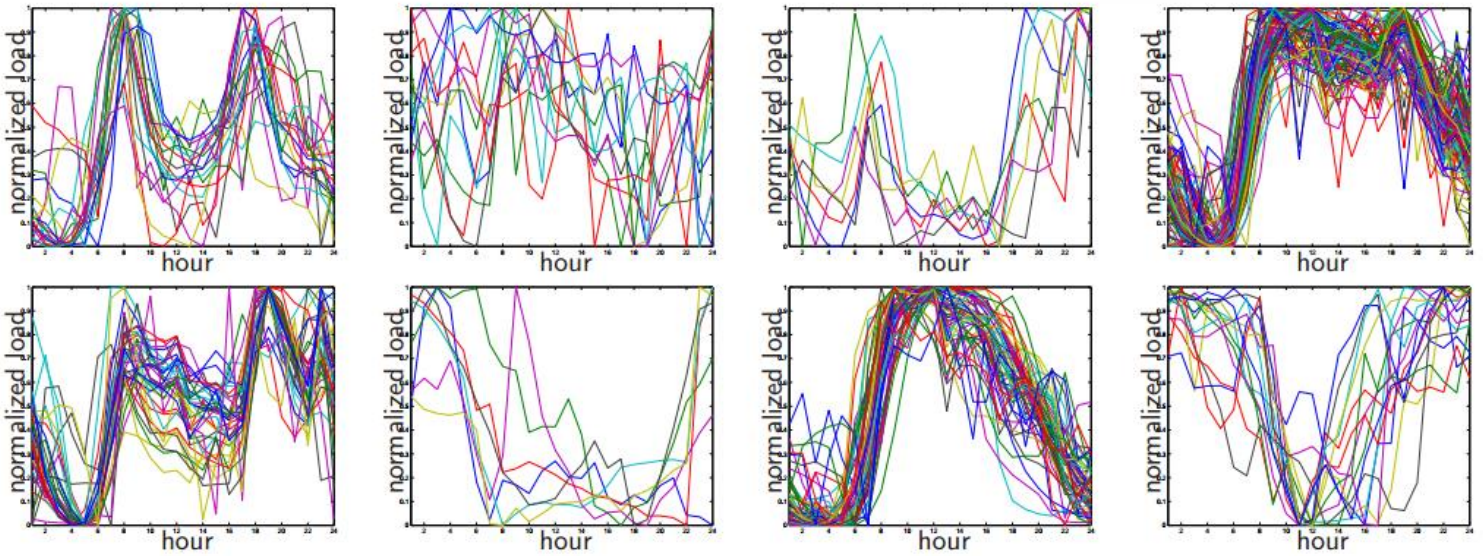
Problem & Objectives



Electric Market Segmentation

Data

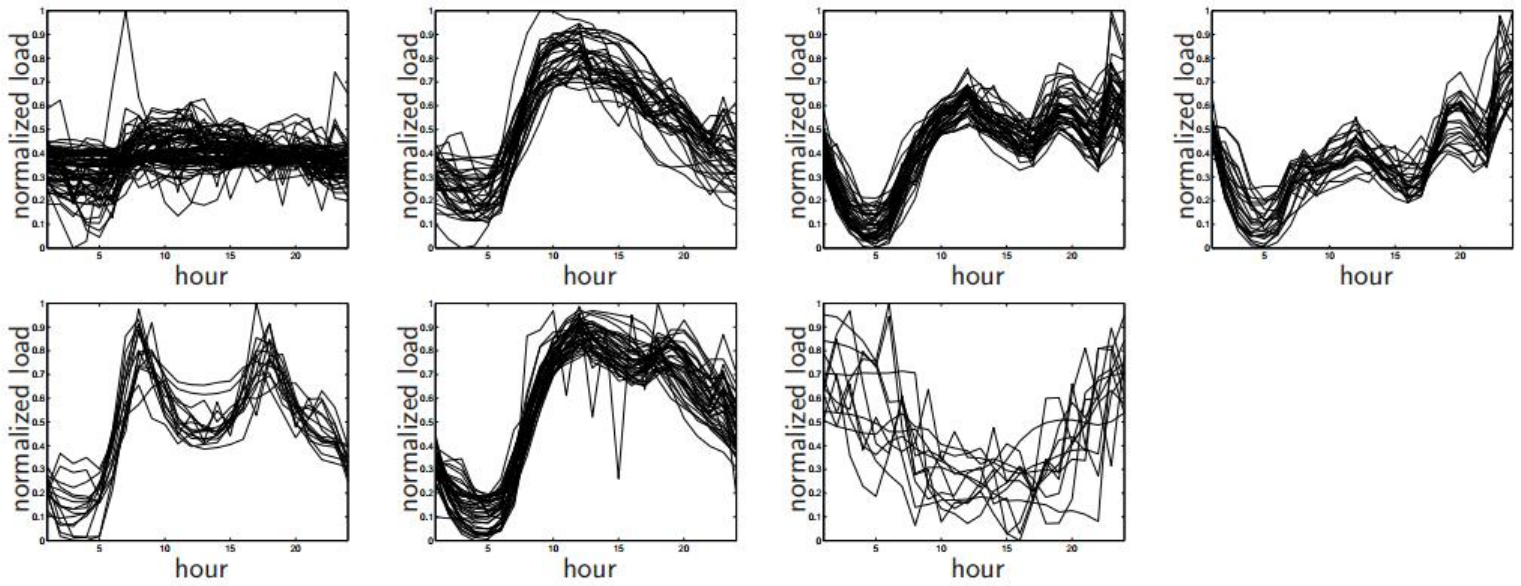
Power load: 245 substations, hourly (5 years)
Periodic AR modelling: dim reduction $43.824 \rightarrow 24$
k-means applied after dimensionality reduction



Electric Market Segmentation

Expertise In Action

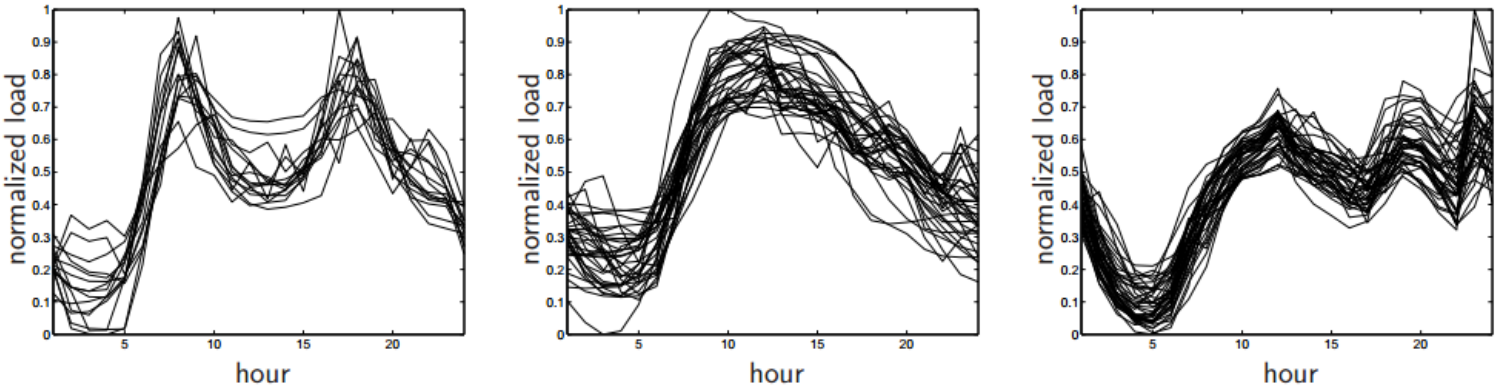
Application of kernel spectral clustering, directly in high dim $d = 43.824$
Model selection on kernel parameter and number of clusters



[Alzate, Espinoza, De Moor, Suykens, 2009]

Electric Market Segmentation

Problem Solved



Electricity load: 245 substations in Belgian grid (1/2 train, 1/2 validation)
$x_i \in \mathbb{R}^{43.824}$: spectral clustering on **high dimensional data** (5 years)

3 of 7 detected clusters:
- *1: Residential profile: morning and evening peaks*
- *2: Business profile: peaked around noon*
- *3: Industrial profile: increasing morning, oscillating afternoon and evening*

Bankruptcy Prediction

Problem

What we know!

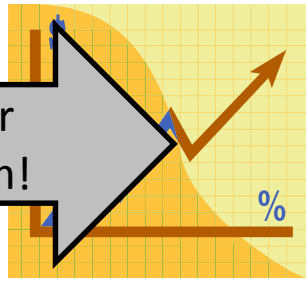
: Capital

: Credit

: Total assets

In total : 40 candidate inputs

Classifier Algorithm!



or

?

Bankruptcy Prediction

Solution

Cross-validation

Only a subset of inputs is important

⇩

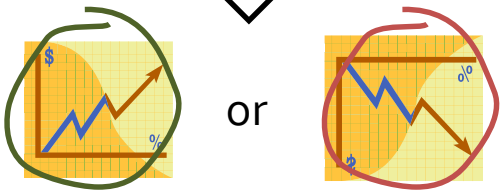
Concise classifier!

⇨ > 90% accurate

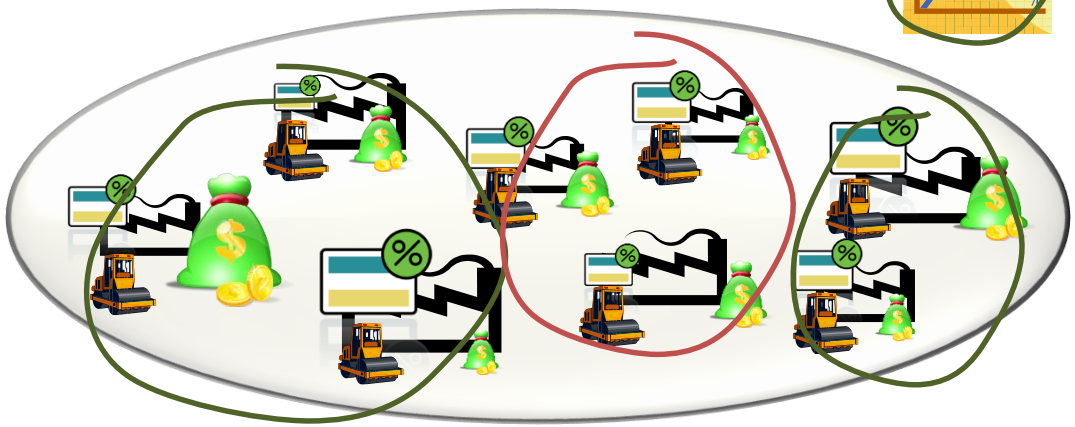
Classifier

Decides on basis of inputs

⇩



or



Energy

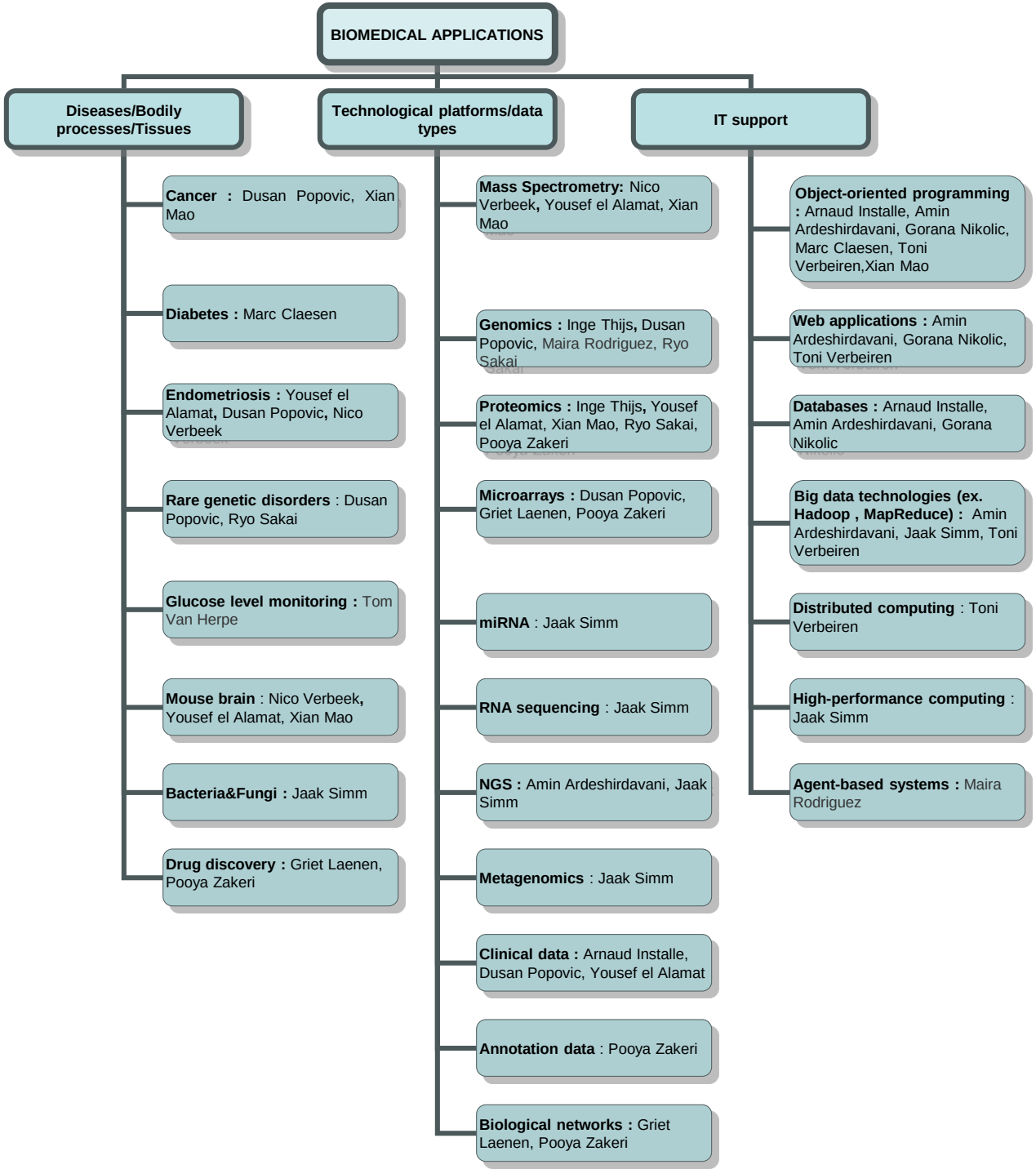
Industry

Environment

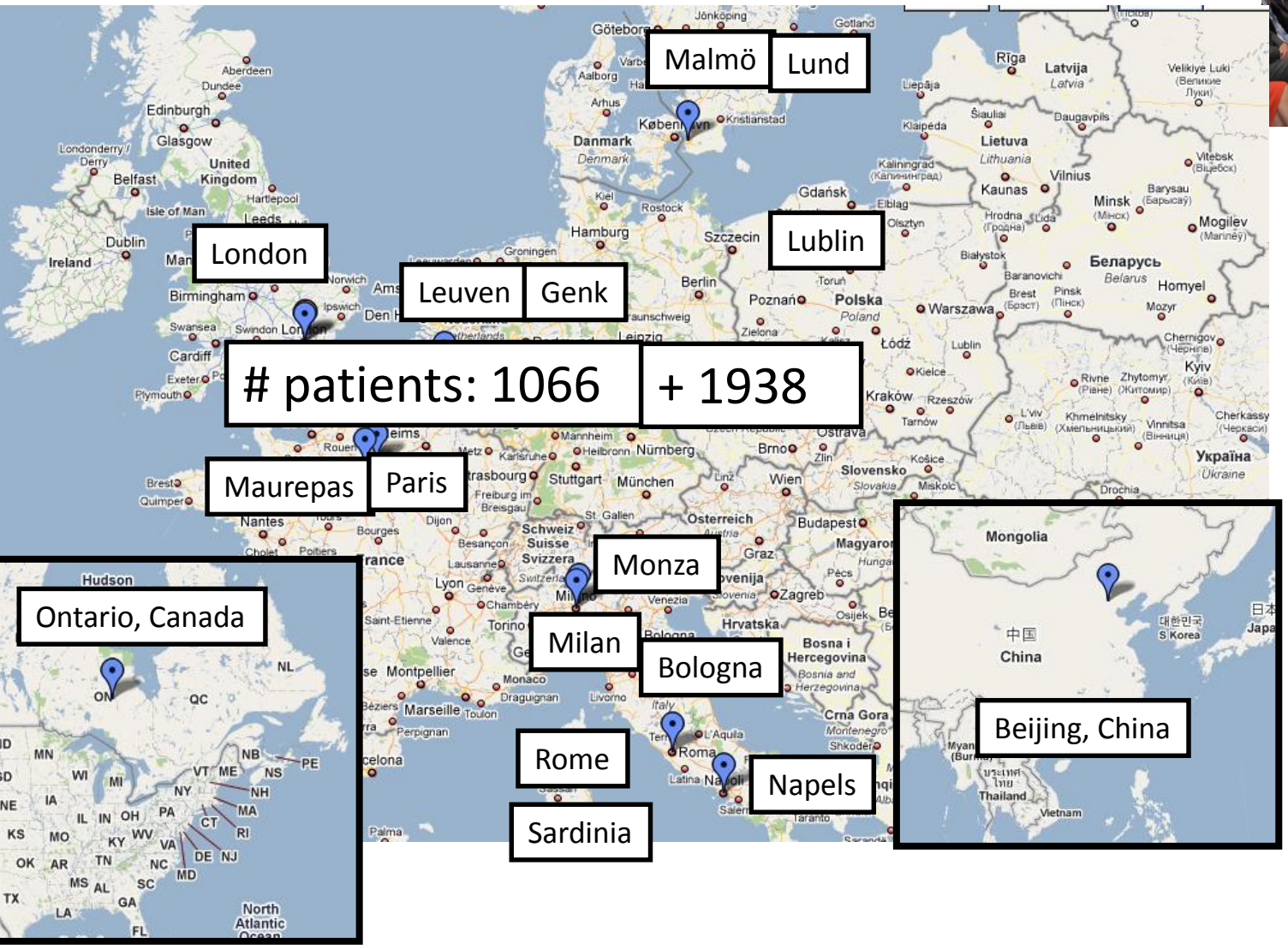
Social networks

Finance and Fraud

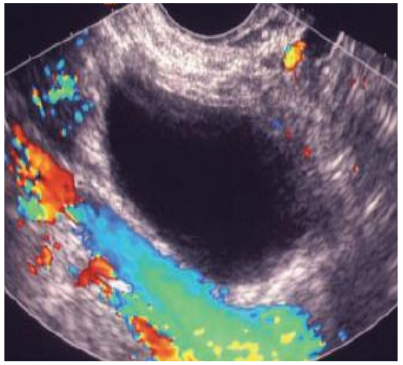
Health


BIOMEDICAL APPLICATIONS
Diseases/Bodily processes/Tissues
Technological platforms/data types
IT support
Cancer : Dusan Popovic, Xian Mao
Diabetes : Marc Claesen
Endometriosis : Yousef el Alamat, Dusan Popovic, Nico Verbeek
Rare genetic disorders : Dusan Popovic, Ryo Sakai
Glucose level monitoring : Tom Van Herpe
Mouse brain : Nico Verbeek, Yousef el Alamat, Xian Mao
Bacteria&Fungi : Jaak Simm
Drug discovery : Griet Laenen, Pooya Zakeri
Mass Spectrometry: Nico Verbeek, Yousef el Alamat, Xian Mao
Genomics : Inge Thijs, Dusan Popovic, Maira Rodriguez, Ryo Sakai
Proteomics : Inge Thijs, Yousef el Alamat, Xian Mao, Ryo Sakai, Pooya Zakeri
Microarrays : Dusan Popovic, Griet Laenen, Pooya Zakeri
miRNA : Jaak Simm
RNA sequencing : Jaak Simm
NGS : Amin Ardeshirdavani, Jaak Simm
Metagenomics : Jaak Simm
Clinical data : Arnaud Installe, Dusan Popovic, Yousef el Alamat
Annotation data : Pooya Zakeri
Biological networks : Griet Laenen, Pooya Zakeri
Object-oriented programming : Arnaud Installe, Amin Ardeshirdavani, Gorana Nikolic, Marc Claesen, Toni Verbeiren,Xian Mao
Web applications : Amin Ardeshirdavani, Gorana Nikolic, Toni Verbeiren
Databases : Arnaud Installe, Amin Ardeshirdavani, Gorana Nikolic
Big data technologies (ex. Hadoop , MapReduce) : Amin Ardeshirdavani, Jaak Simm, Toni Verbeiren
Distributed computing : Toni Verbeiren
High-performance computing : Jaak Simm
Agent-based systems : Maira Rodriguez

Participatory

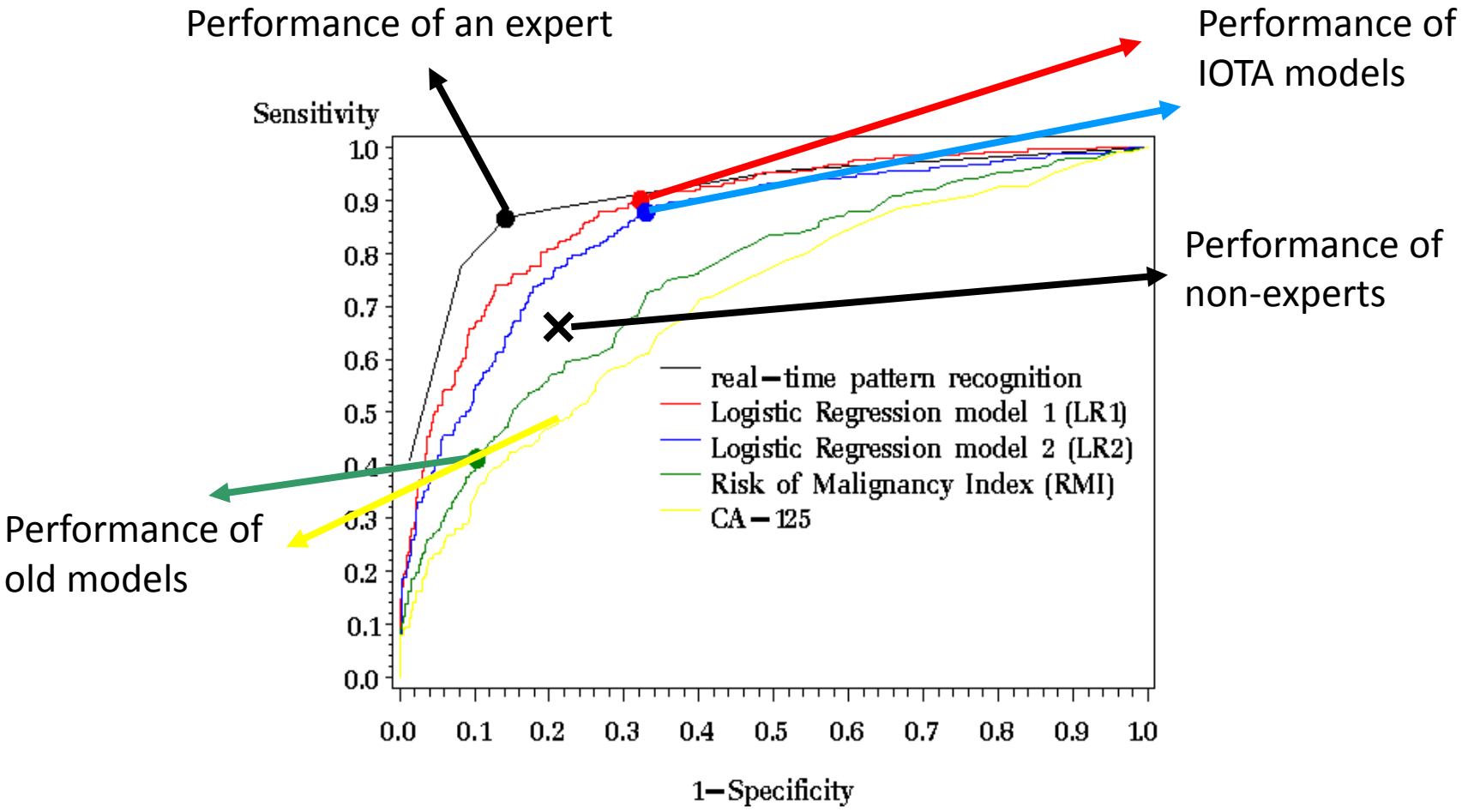


IOTA app: population based assessment of **ovarian tumor malignancy**:



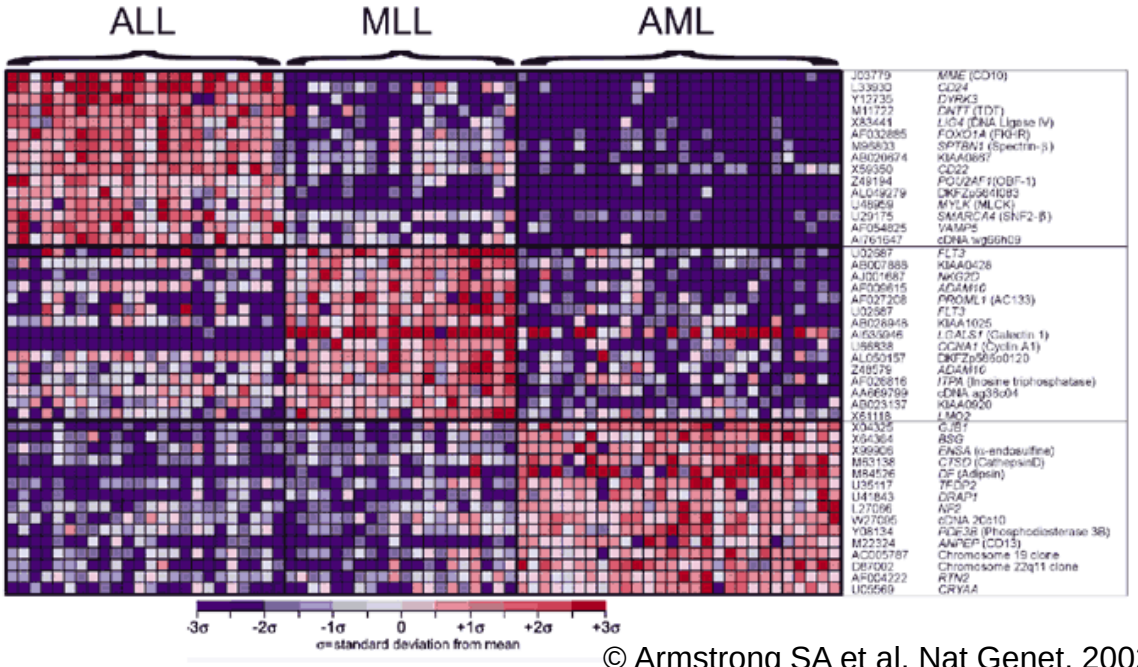
IOTA app available in iTunes app store and on http://homes.esat.kuleuven.be/~sistawww/biomed/iota/

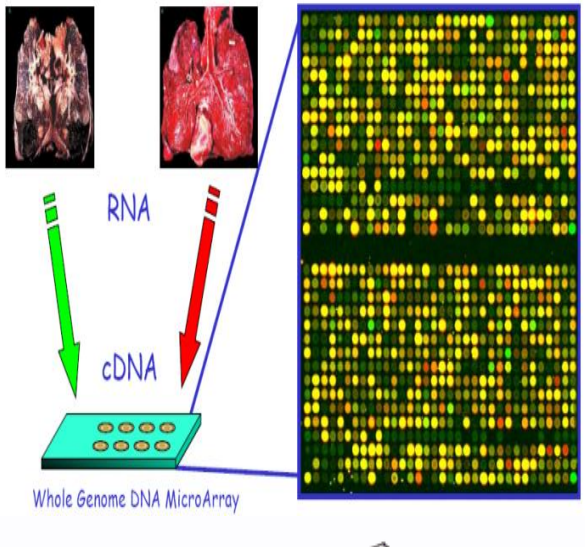
Performance



You share, we care !

Genomic markers for Leukemia

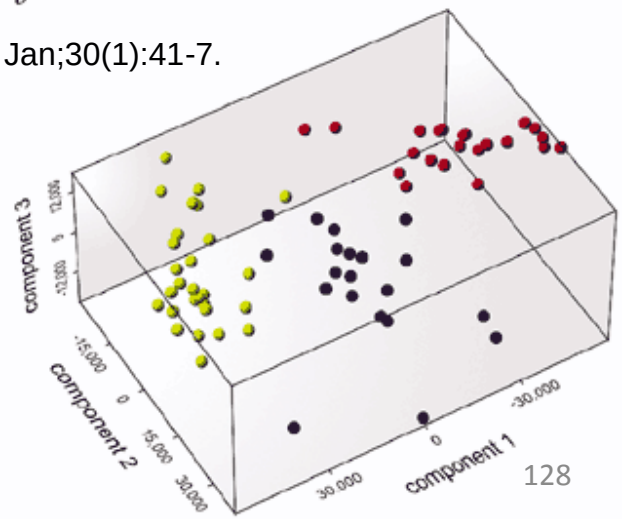




12 600 genes

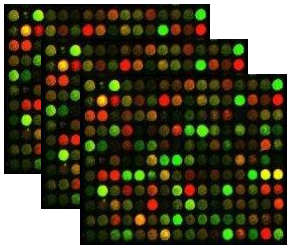
72 patients

- 28 Acute Lymphoblastic Leukemia (ALL)
- 24 Acute Myeloid Leukemia (AML)
- 20 Mixed Linkage Leukemia (MLL)

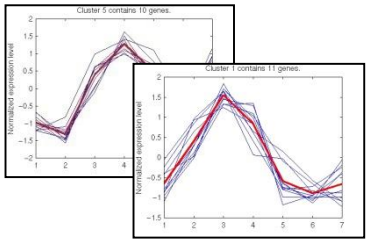


Genomic Data Fusion

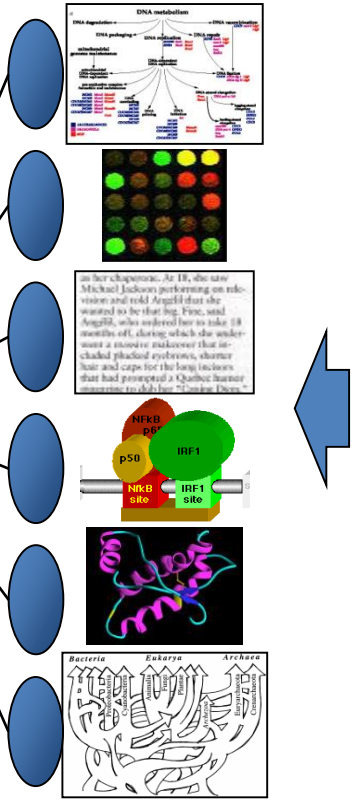
High-throughput genomics



Data analysis



Information sources



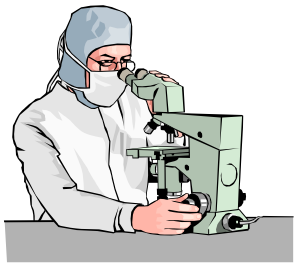
Candidate genes

Name	Ensembl
TTR	ENSG00000118271
PAH	ENSG00000171759
G6PC	ENSG00000131482
IGF1	ENSG00000017427
ALB	ENSG00000163631
CRP	ENSG00000132693
HABP2	ENSG00000148702
IF	ENSG00000138799
FST	ENSG00000134363
ARAF1	ENSG00000078061
HMGA2	ENSG00000149948
C9	ENSG00000113600
PCBP2	ENSG00000111406
HOXB6	ENSG00000108511
RERE	ENSG00000142599
HOXA11	ENSG00000005073
CLIC1	ENSG00000096238
ERCC3	ENSG00000163161
ERCC3	ENSG00000163161
TLL2	ENSG00000095587
SYT4	ENSG00000132872
SYT4	ENSG00000132872
PIK4CB	ENSG00000143393
PKD2	ENSG00000118762
	ENSG00000081026
ANKRD3	ENSG00000183421
F13A1	ENSG00000124491
BPAG1	ENSG00000151914
KCNN3	ENSG00000143603
GRIN2A|GRIN2B	ENSG00000150086
SIM1	ENSG00000112246
	ENSG00000174891
	ENSG00000089195
C14orf10	ENSG00000092020
STX8	ENSG00000170310
	ENSG00000107671
MSH5	ENSG00000096474
CRH	ENSG00000147571
MID1	ENSG00000101871
	ENSG00000184508
	ENSG00000113460
TGFB3	ENSG00000119699
C1QR1	ENSG00000125810
NR4A2	ENSG00000153234
PDGFC	ENSG00000145431
PDGFC	ENSG00000145431
NR3C2	ENSG00000151623
NFYA	ENSG00000001167
	ENSG00000101898
C8orf4	ENSG00000176907
TM4SF13	ENSG00000106537
MMP3|MMP1	ENSG00000149968

Candidate prioritization

Rank	En	Ex	Ip	Ke	GO	Te Avg	Pval
1	TTR	G6PC	PAH	G6PC	IGF1	TTR	TTR
2	IGF1	TTR	IGF1	PAH	PAH	IGF1	PAH
3	CRP	ALB	TTR	RERE	G6PC	CRP	G6PC
4	HOXB6	HABP2	ALB	ERCC3	TTR	HOXB6	IGF1
5	ALB	PAH	HDC	ERCC3		ALB	ALB
6	NR4A2	IF	TLL2	ANKRD3		HMGA2	CRP
7	PAH		C1QR1	ARAF1	HDC	NR4A2	HABP2
8	HOXA11	IGF1	G6PC	PKD2	F13A1	PAH	IF
9	NFYA	CRP	HABP2	MTMR1	KCNN3	HOXA11	C13orf7	FST
10	C9	ARAF1	IF	HDC	CLIC1	NFYA	TTR	ARAF1

Validation



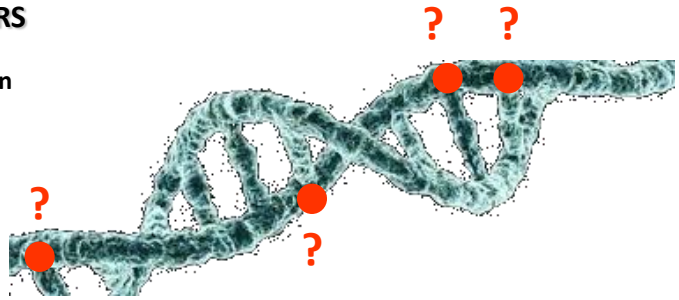
Endeavour: Aerts et al., Nature Biotechnology, 2006

Mutation Prioritization

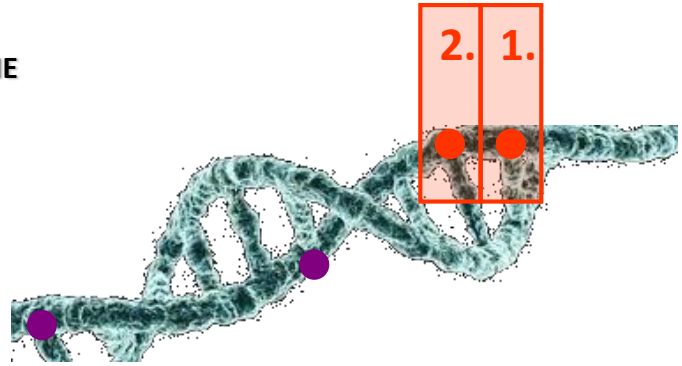
Problem & Objectives

PREVELANCE OF GENETIC DISORDERS

Biomarker discovery as a classification problem



THOUSANDS MUTATIONS IN A GENOME



NEED TO PRIORITIZE MUTATIONS

Mutation prioritization

Data

Very sensitive
and confidential

Mutation prioritization

Expertise In Action

NOVEL GENETIC DISORDER

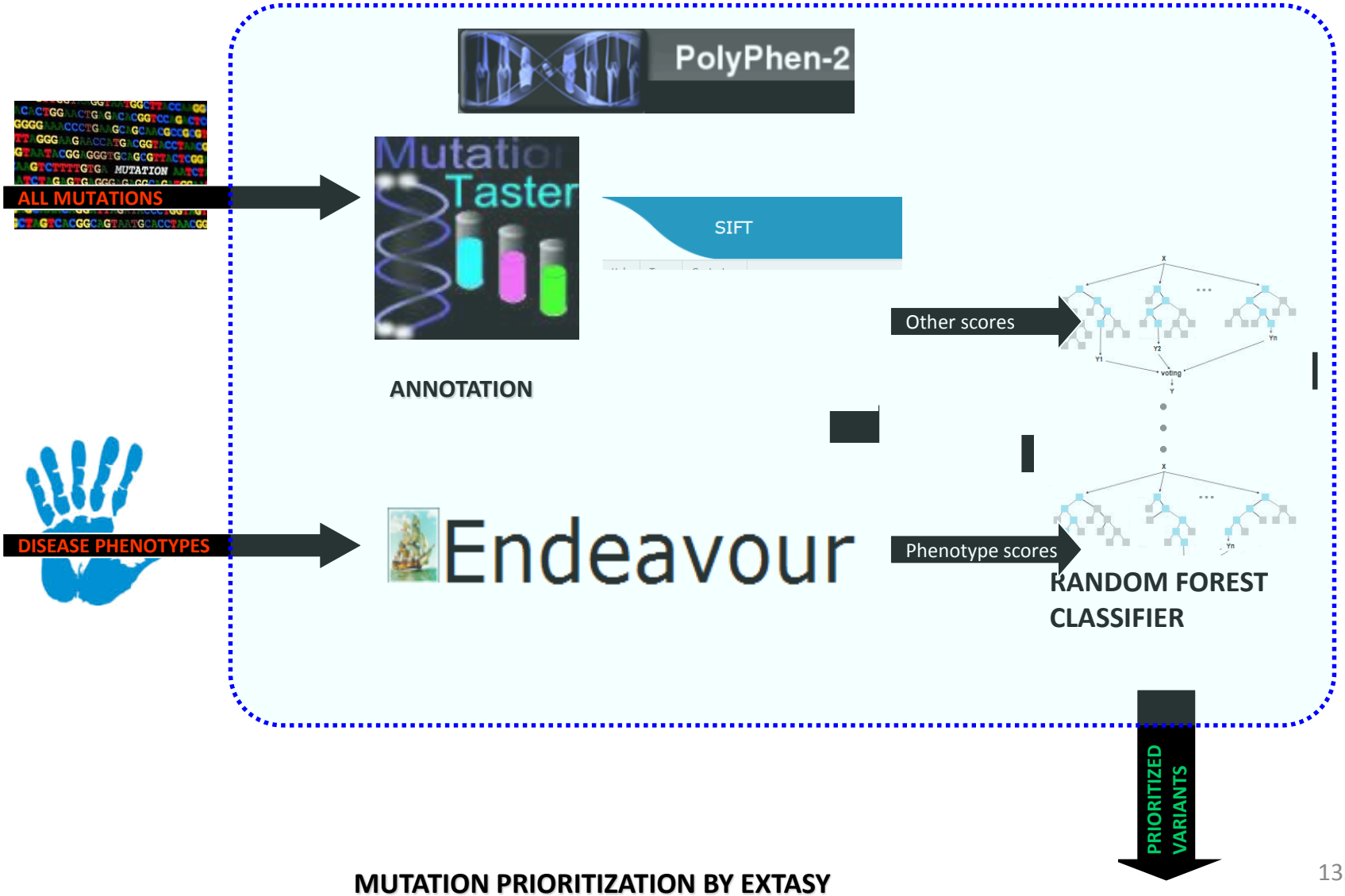
SEQUENCING

CONFIRMATORY EXPERIMENTS

PERSONALIZED THERAPY

Mutation prioritization

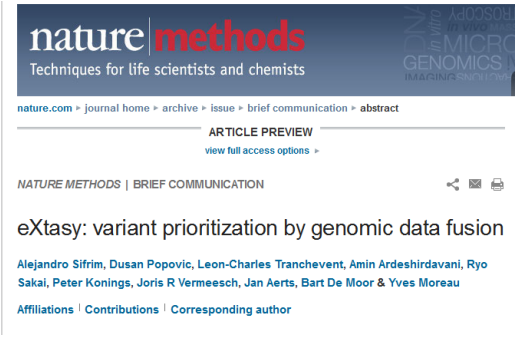
Expertise In Action

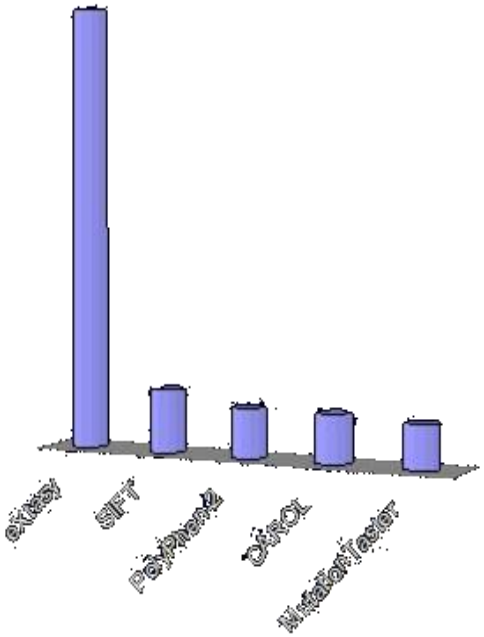


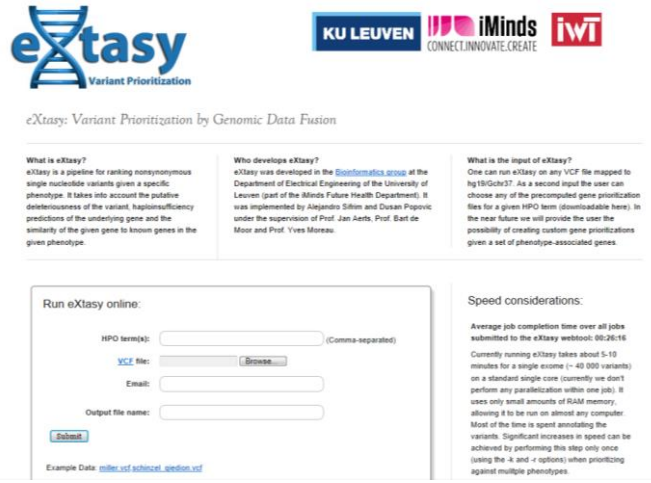
MUTATION PRIORITIZATION BY EXTASY

Mutation prioritization

Problem Solved





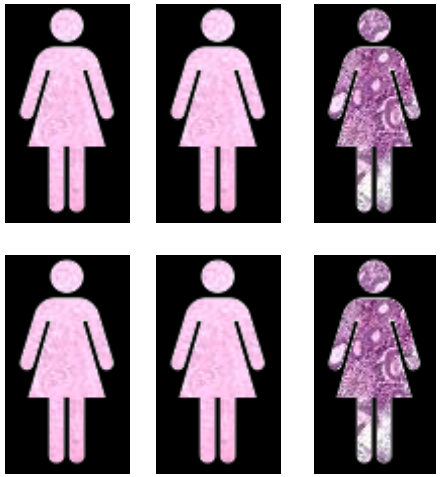


10-FOLD INCREASE IN PRECISION

FREE AND EASY-TO-USE WEB TOOL

Non-invasive diagnosis of endometriosis

Problem & Objectives



PREVELANCE OF ENDOMETRIOSIS

INVASIVE DIAGNOSTICS

2012
2013
2014
…

LONG DELAYS IN DIAGNOSIS

Non-invasive diagnosis of endometriosis

Expertise In Action

PLASMA PROTEINS LEVELS IN PATIENTS

FEATURE SELECTION

Biomarker	Cycle phase	Cut-off
CA-125	All	>12.5 U/ml
Glycodelin	All	>18 ng/ml
VEGF	All	>1.5 pg/ml
IGFBP-3	All	>210 ng/ml
sICAM-1	All	<243 ng/ml
CA 19–9	All	>9.5 IU/ml
sICAM-1	Menstrual	<254.6 ng/ml
IL-1beta	Follicular	<0.9 pg/ml
IL-6	Follicular	<10 pg/ml
IFN-γ	Follicular	<76 pg/ml
TNF-α	Follicular	<45.6 pg/ml
IGFBP-3	Follicular	>200 ng/ml
Glycodelin	Follicular	>9.0 ng/ml
CA-125	Follicular	>11.5 U/ml
CA-125	Luteal	>13.5 U/ml
CA 19–9	Luteal	>7.5 IU/ml

BIOMARKERS

- Big Data
  - What
  - Who
- Six dimensions
  - Data
  - Compute Infrastructure
  - Storage Infrastructure
  - Analytics
  - Visualization
  - Security & Privacy
- Machine learning as a commodity
- Expertise
  - Books & Spin-offs
  - Algorithms
  - Applications