iMinds Dept. MEDICAL IT

KU Leuven ESAT-STADIUS

Serious Data Mining

Prof.Dr. Bart De Moor

Bart.DeMoor@iminds.be

Content

Big Data
What
Who

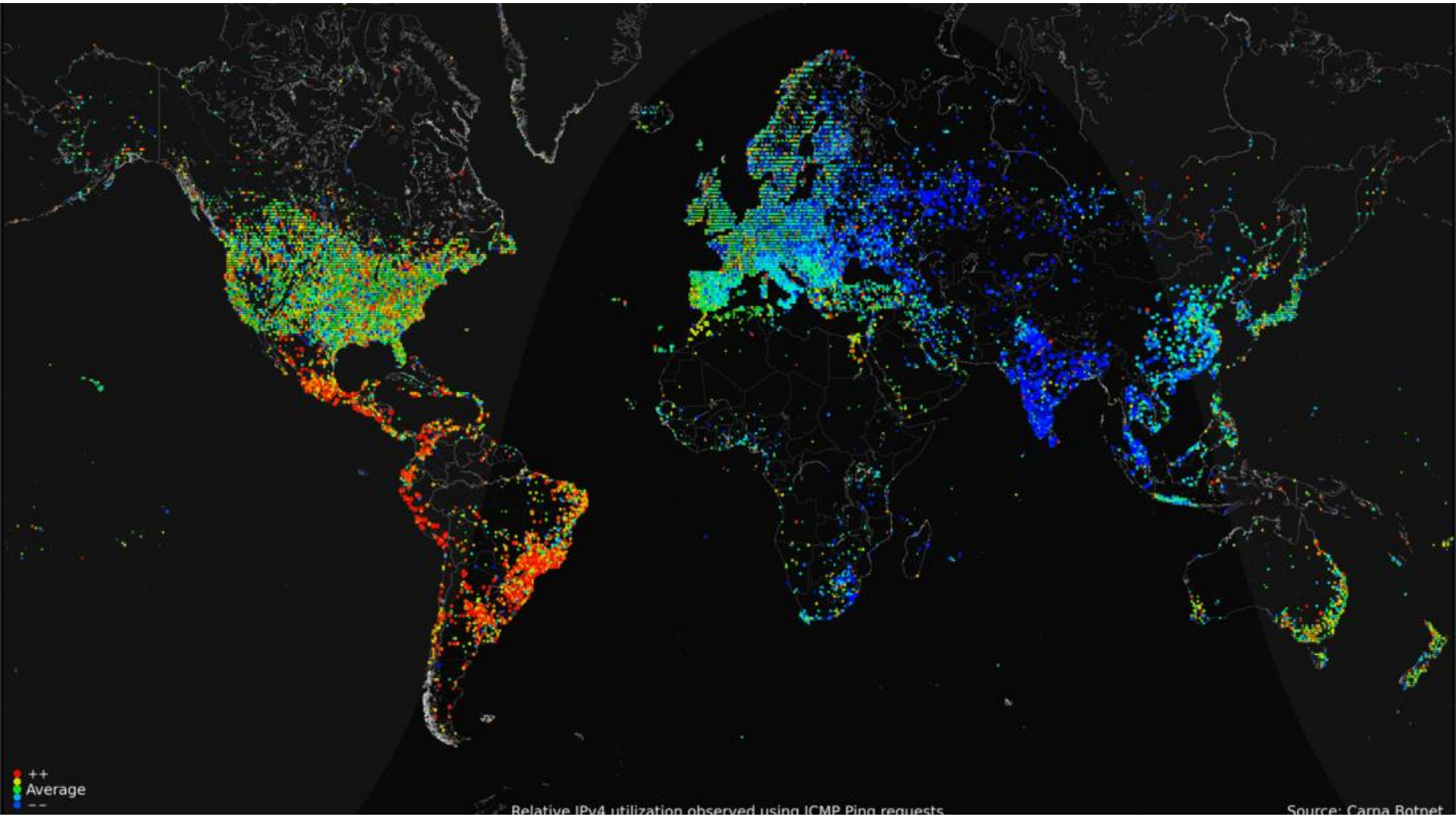
Six issues
Data
Compute Infrastructure
Storage Infrastructure
Analytics
Visualization
Security & Privacy

Machine learning as a commodity

Expertise of ESAT-STADIUS, KU Leuven

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

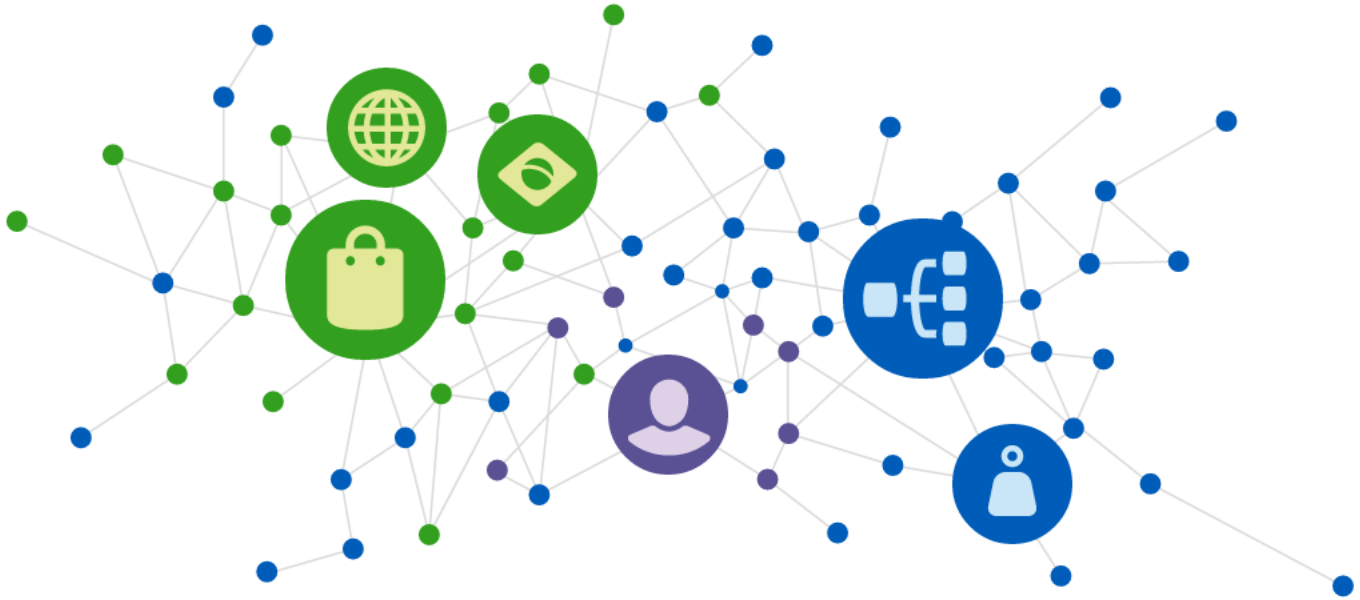
The Industrial Internet, a connected network of intelligent machines working the way they are intended, will transform business as dramatically as the consumer Internet has changed our lives.

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

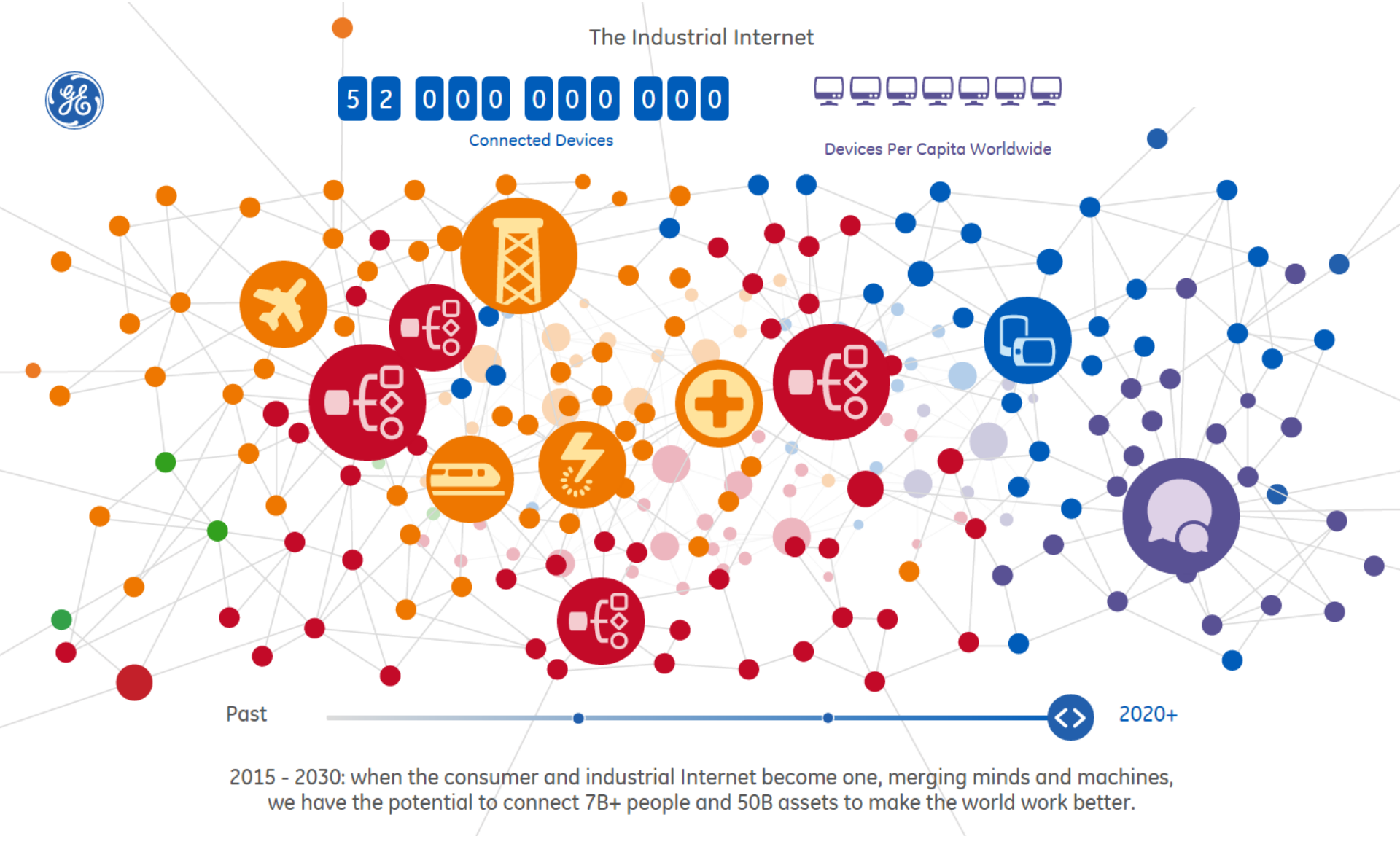
25 000 000 000

Connected Devices

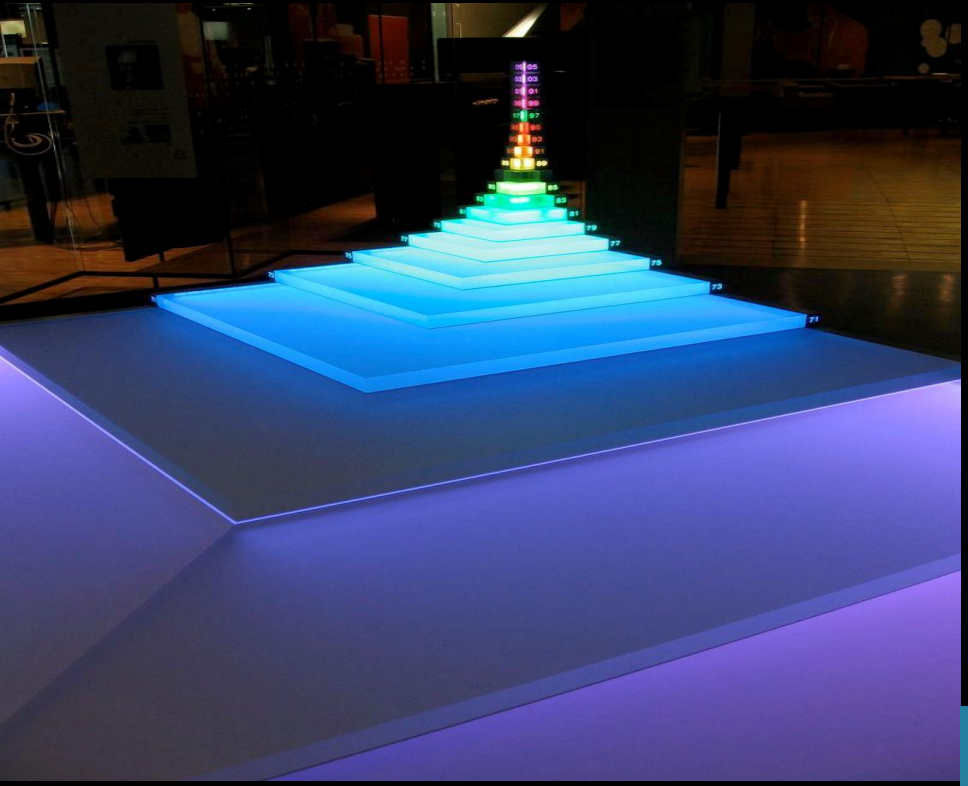
Devices Per Capita Worldwide

Past — 2020+

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

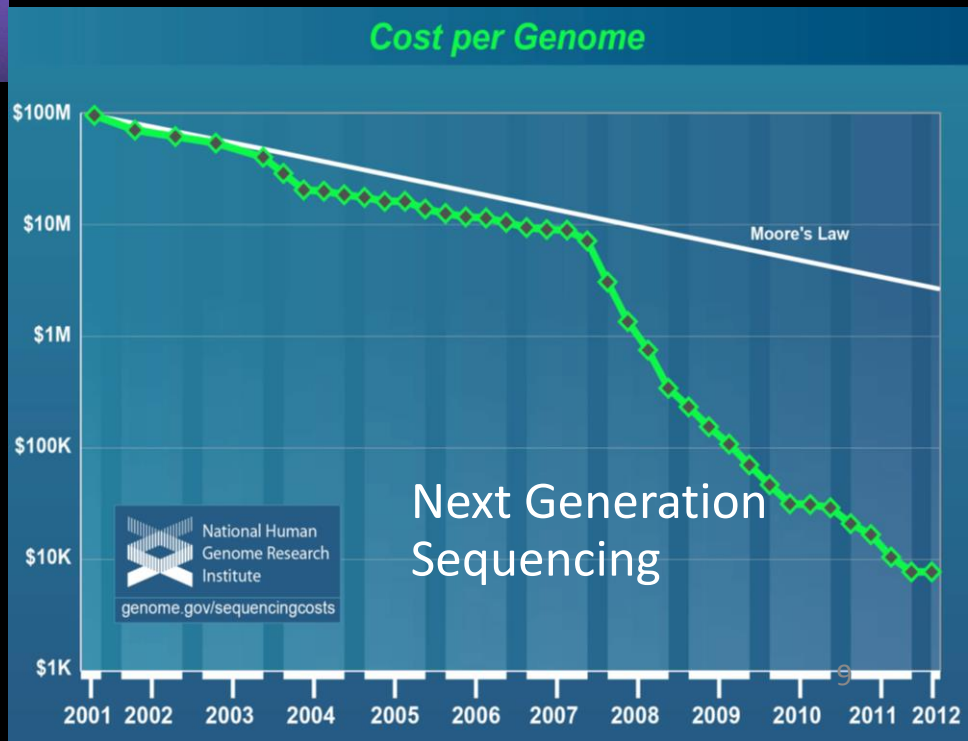


2015 - 2030: when the consumer and industrial Internet become one, merging minds and machines, we have the potential to connect 7B+ people and 50B assets to make the world work better.



Moore's law:
computing power
doubles
every 18 months

Carlson's law:
complexity/cost
evolves
exponentially



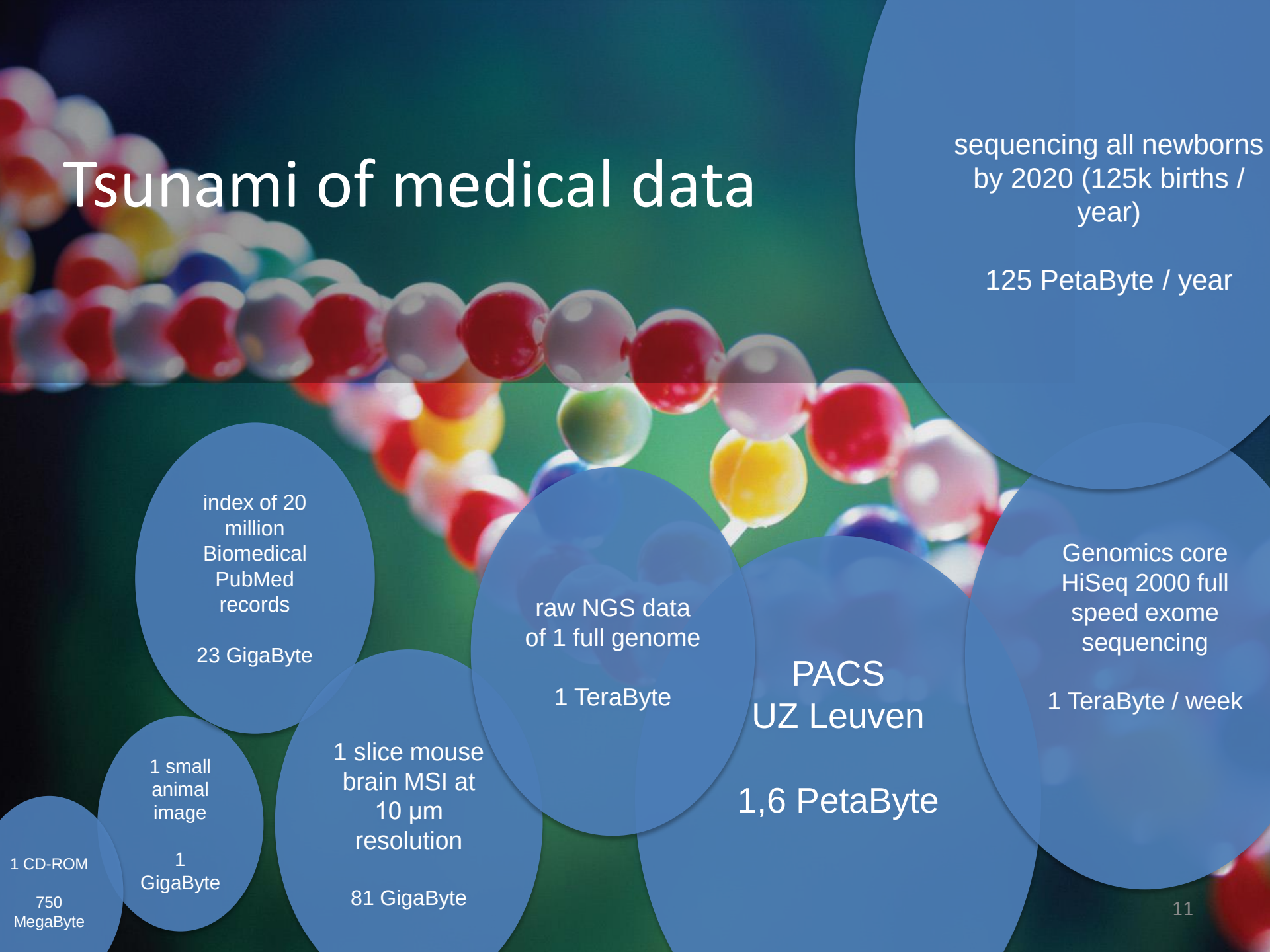
Genome data

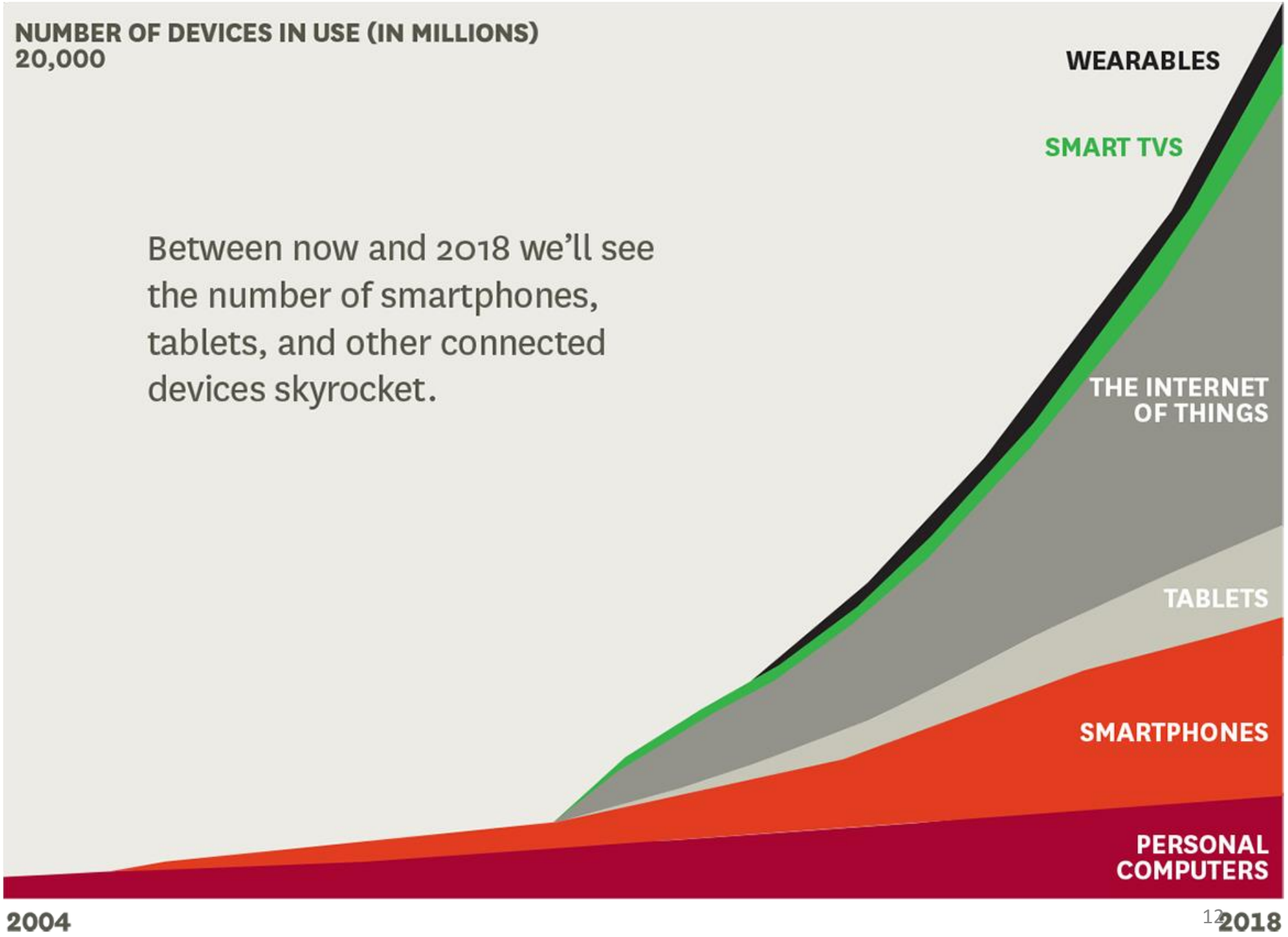
GS-FLX Roche
Applied Science 454

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




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

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.

How big is big?

Google

?

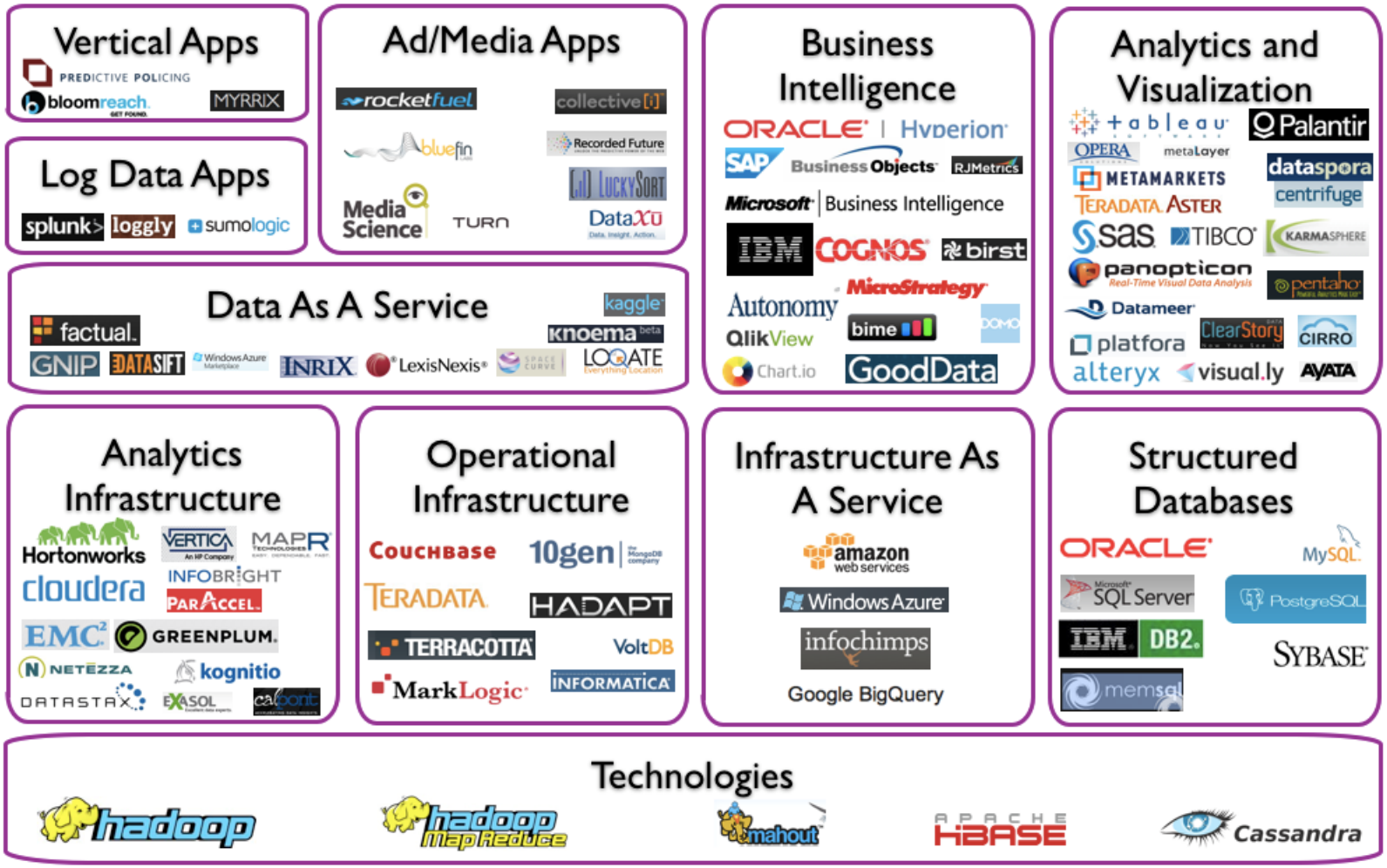
Microsoft

1 million servers (2013)

Amazon

?

Big Data Landscape

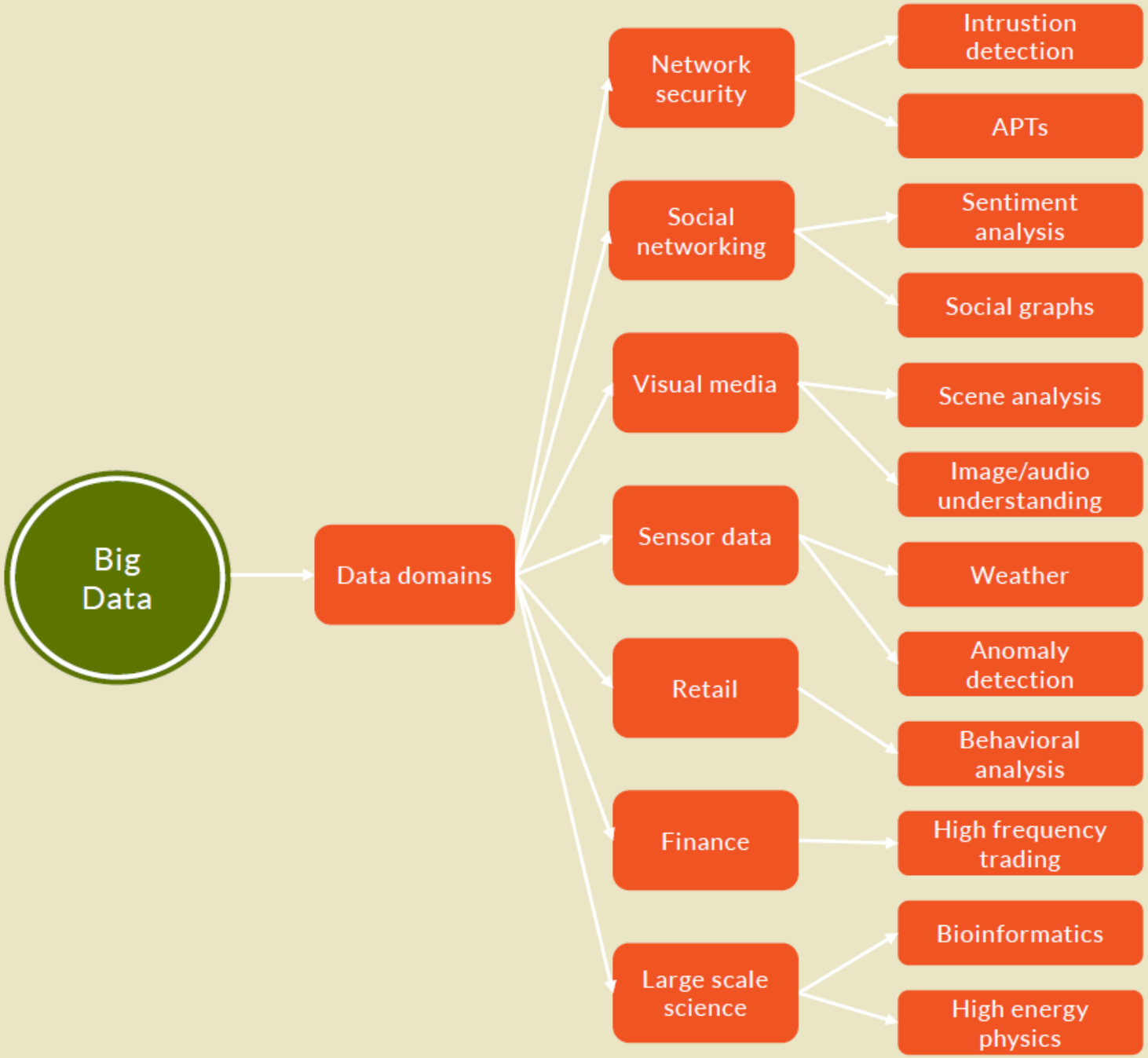


dave@vcdave.com

blogs.forbes.com/davefeinleib

Big Data Landscape (Version 2.0)

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


Big Data
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

Content

Big Data
What
Who

Six issues
Data
Compute Infrastructure
Storage Infrastructure
Analytics
Visualization
Security & Privacy

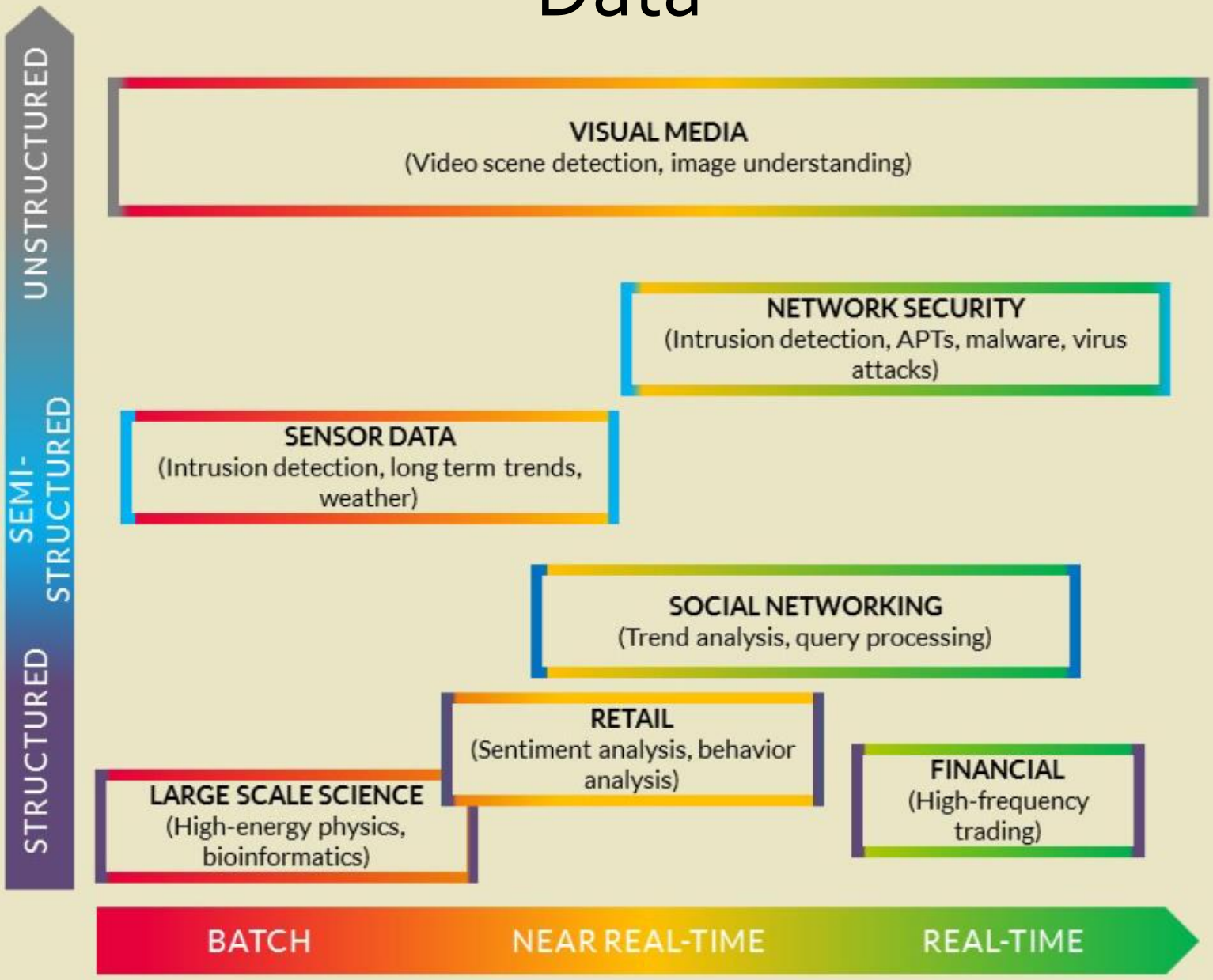
Machine learning as a commodity

Expertise of ESAT-STADIUS, KU Leuven

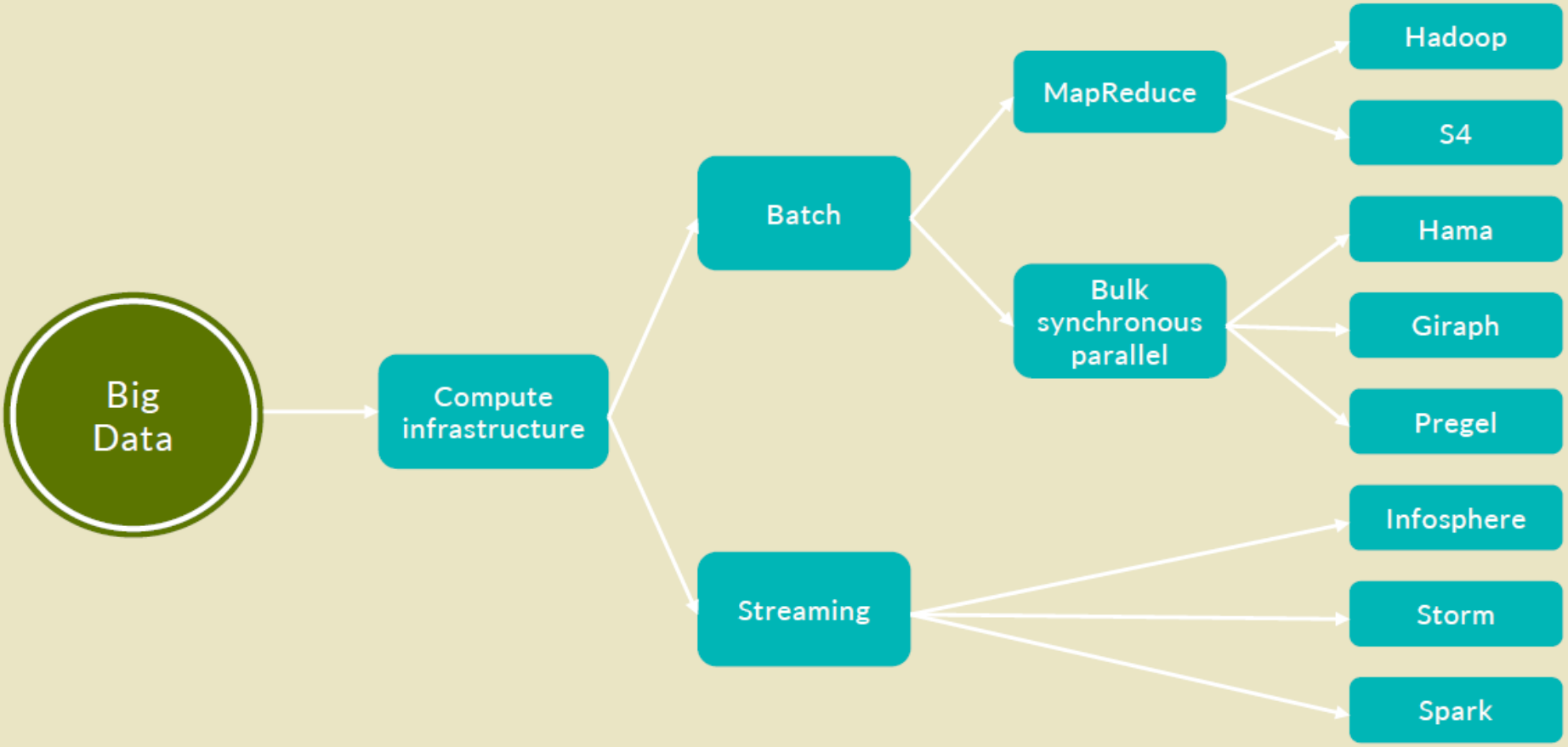
Books & Spin-offs
Algorithms
Applications


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

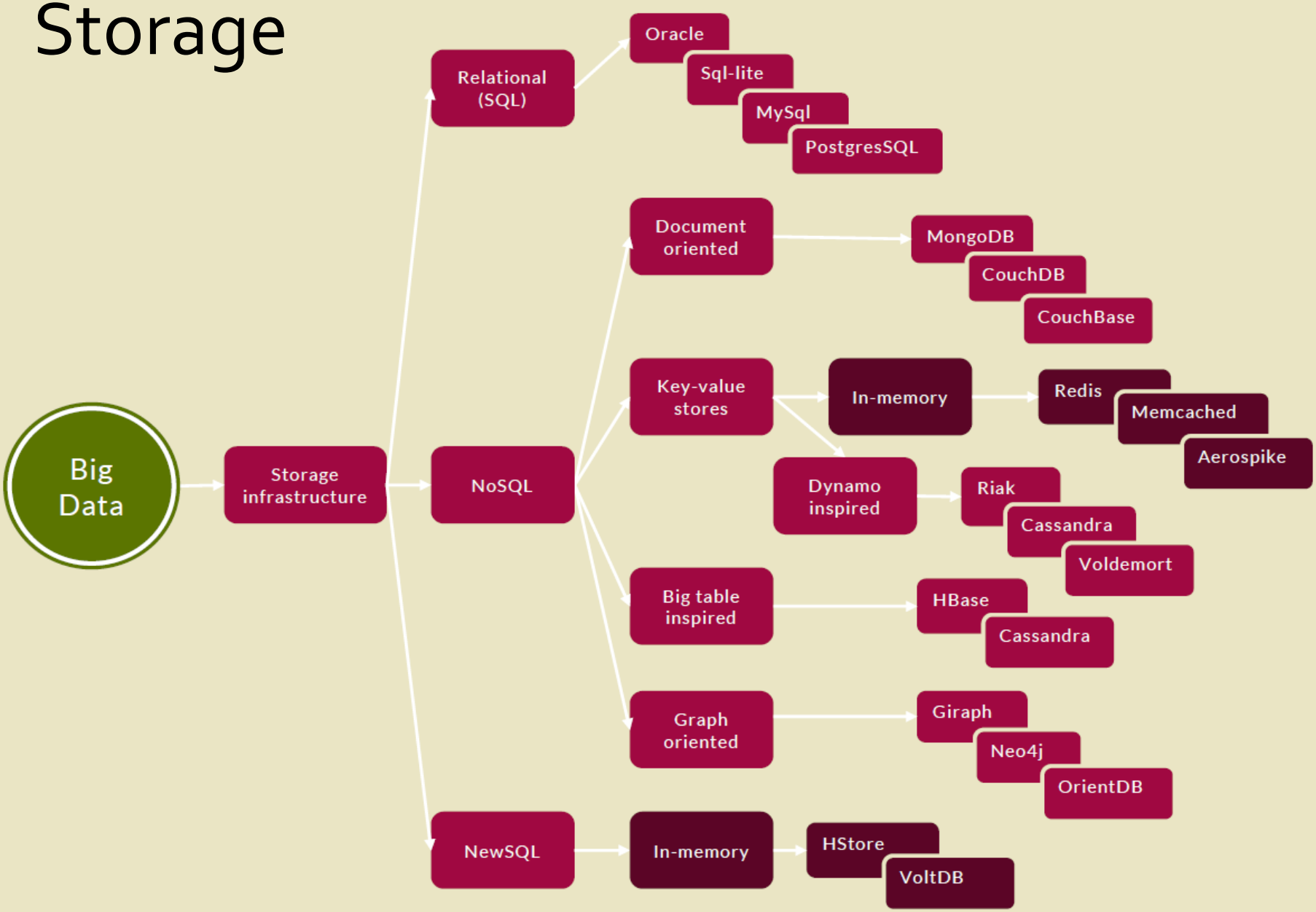
Data

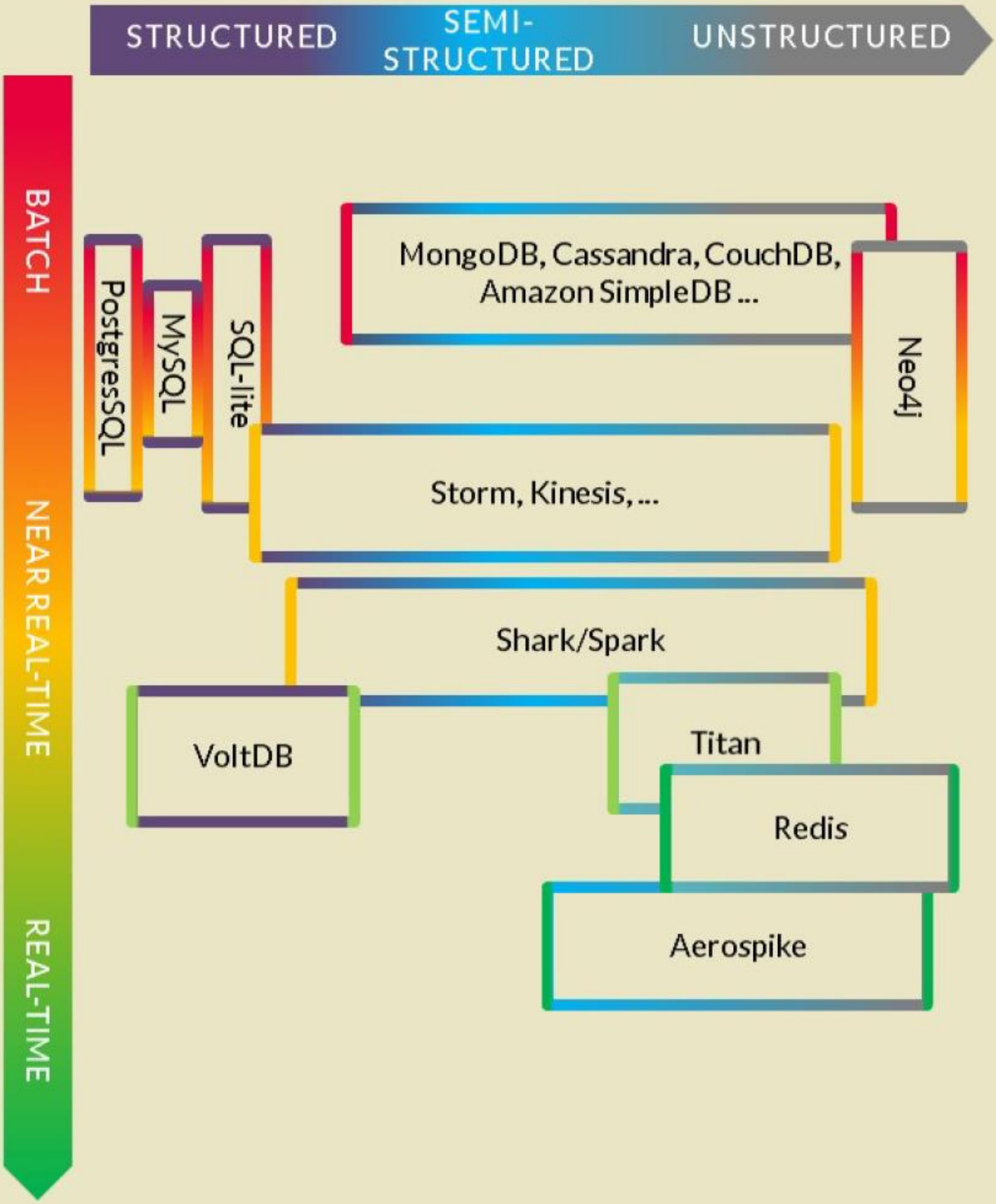

UNSTRUCTURED
SEMI-STRUCTURED
STRUCTURED
VISUAL MEDIA
(Video scene detection, image understanding)
NETWORK SECURITY
(Intrusion detection, APTs, malware, virus attacks)
SENSOR DATA
(Intrusion detection, long term trends, weather)
SOCIAL NETWORKING
(Trend analysis, query processing)
RETAIL
(Sentiment analysis,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

Compute infrastructure

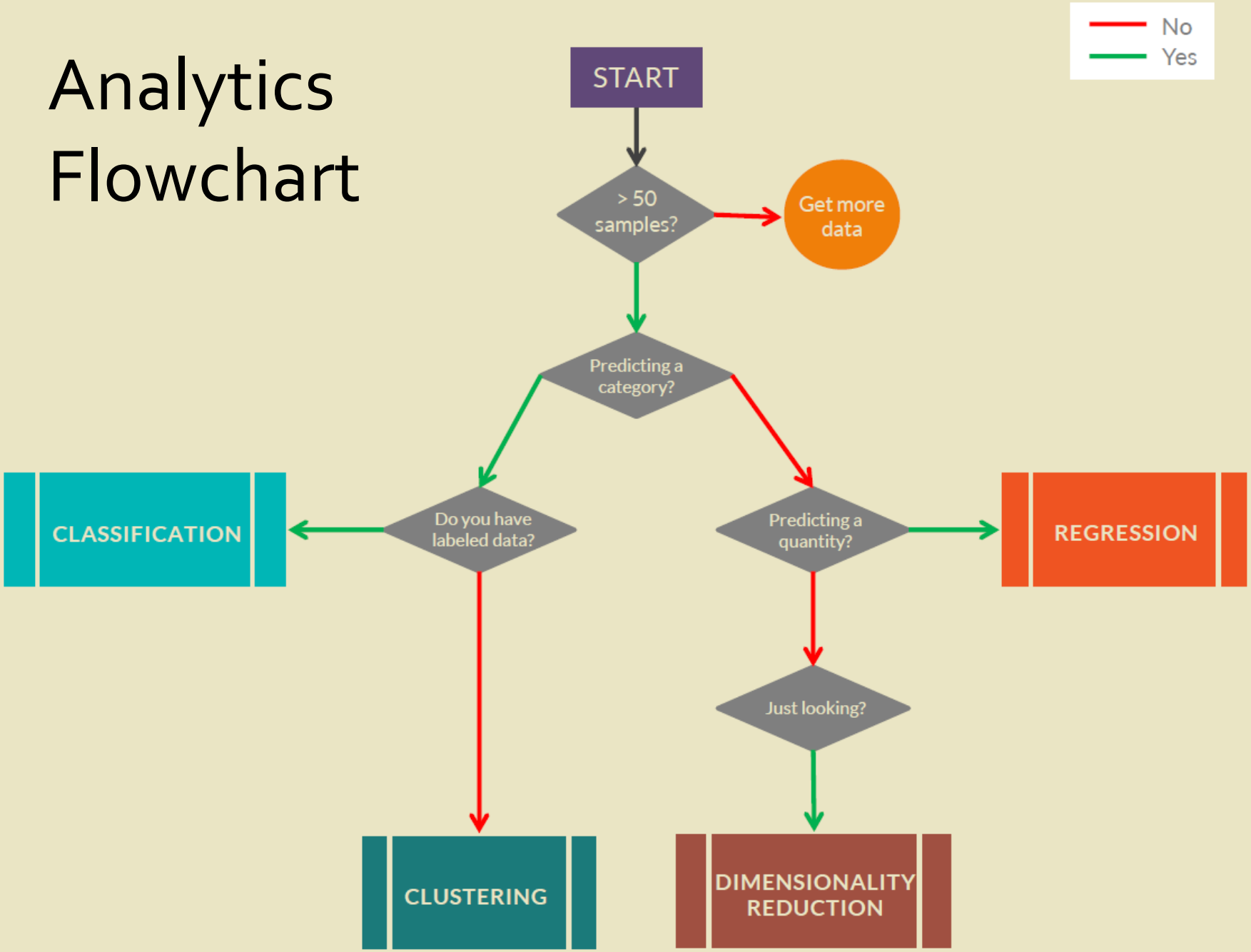


Storage

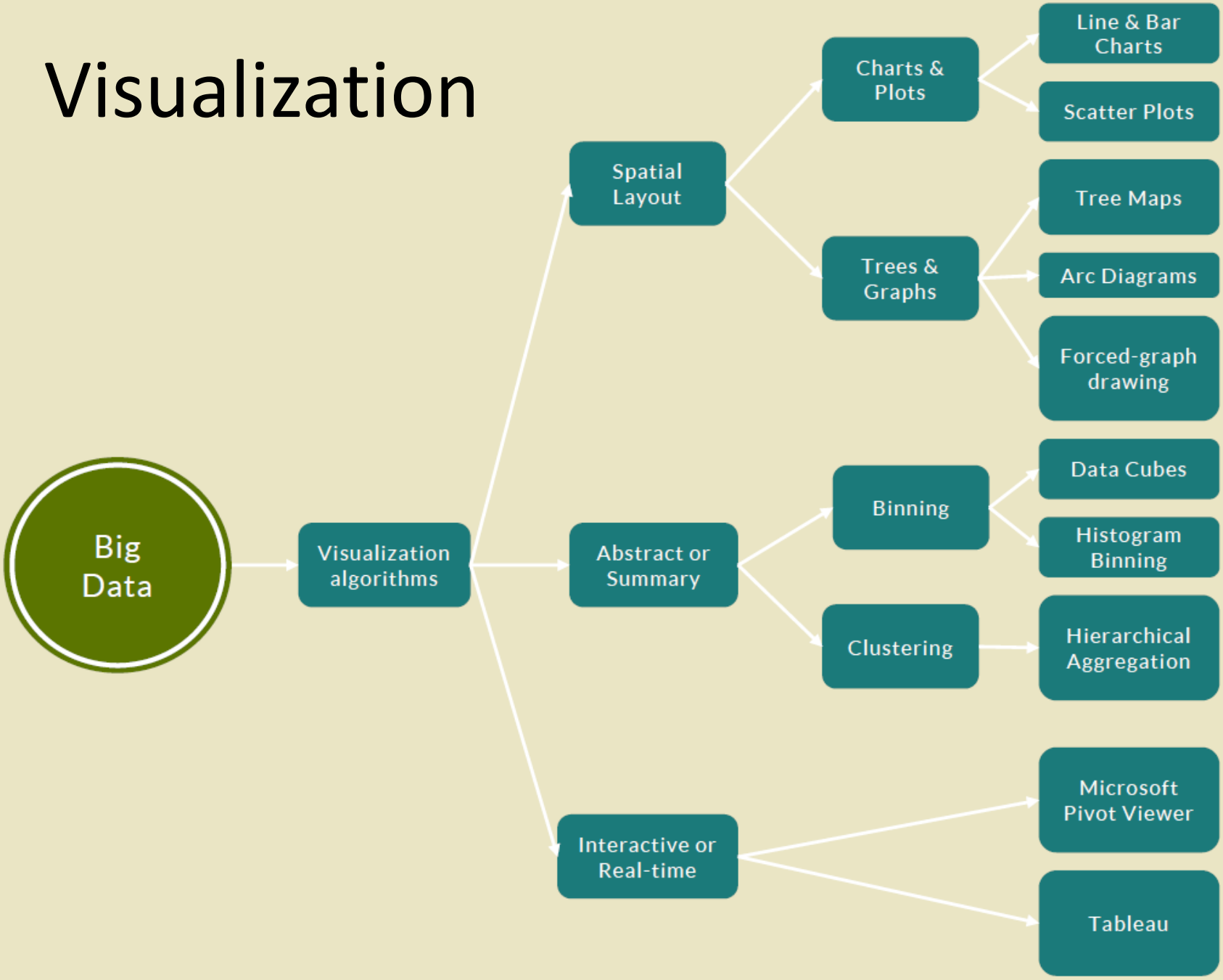

Big Data
Storage infrastructure
Relational (SQL)
Oracle
Sql-lite
MySql
PostgresSQL
NoSQL
Document oriented
MongoDB
CouchDB
CouchBase
Key-value stores
In-memory
Redis
Memcached
Aerospike
Dynamo inspired
Riak
Cassandra
Voldemort
Big table inspired
HBase
Cassandra
Graph oriented
Giraph
Neo4j
OrientDB
NewSQL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

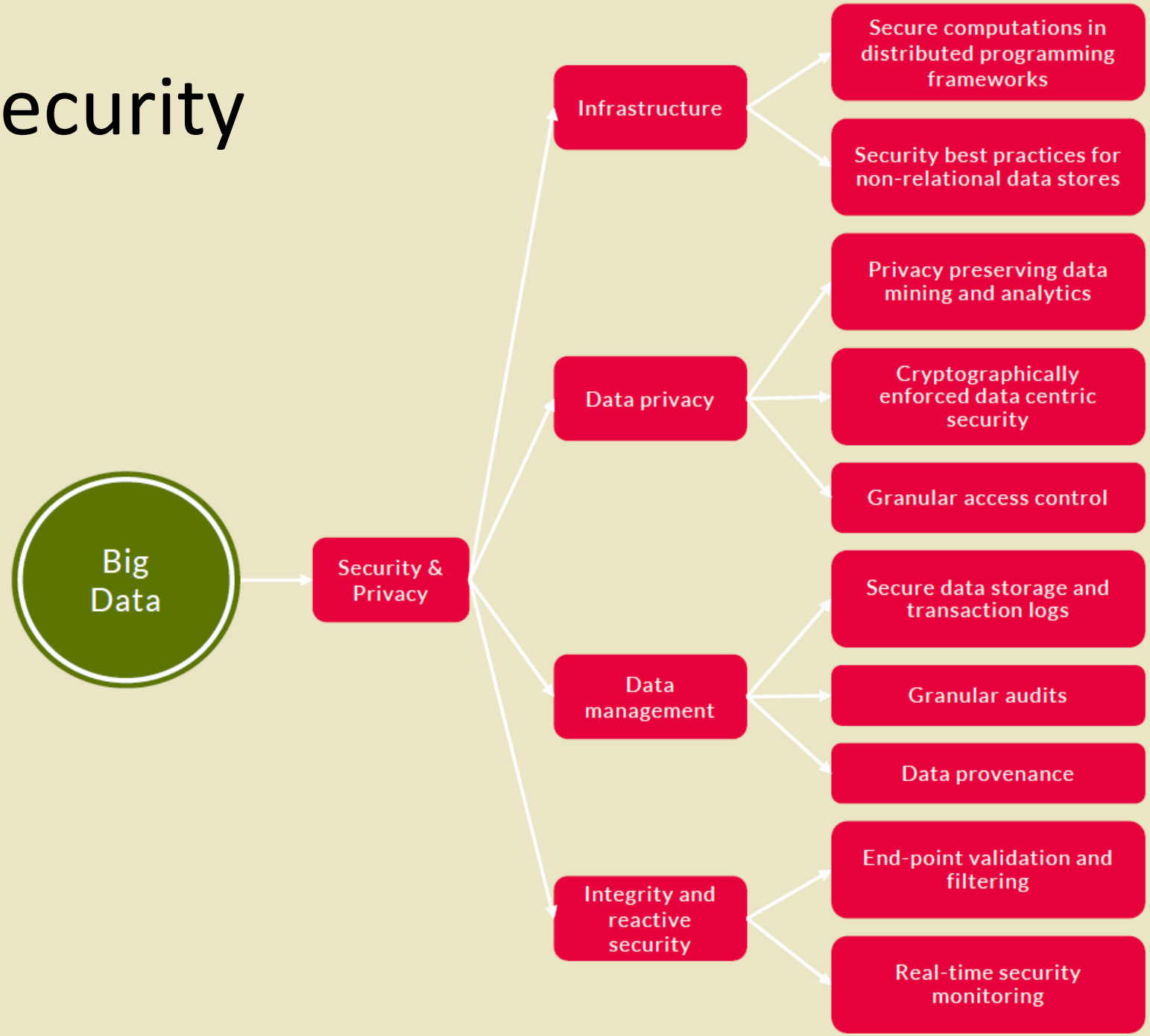
Analytics Flowchart



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

Content

Big Data
What
Who

Six issues
Data
Compute Infrastructure
Storage Infrastructure
Analytics
Visualization
Security & Privacy

Machine learning as a commodity

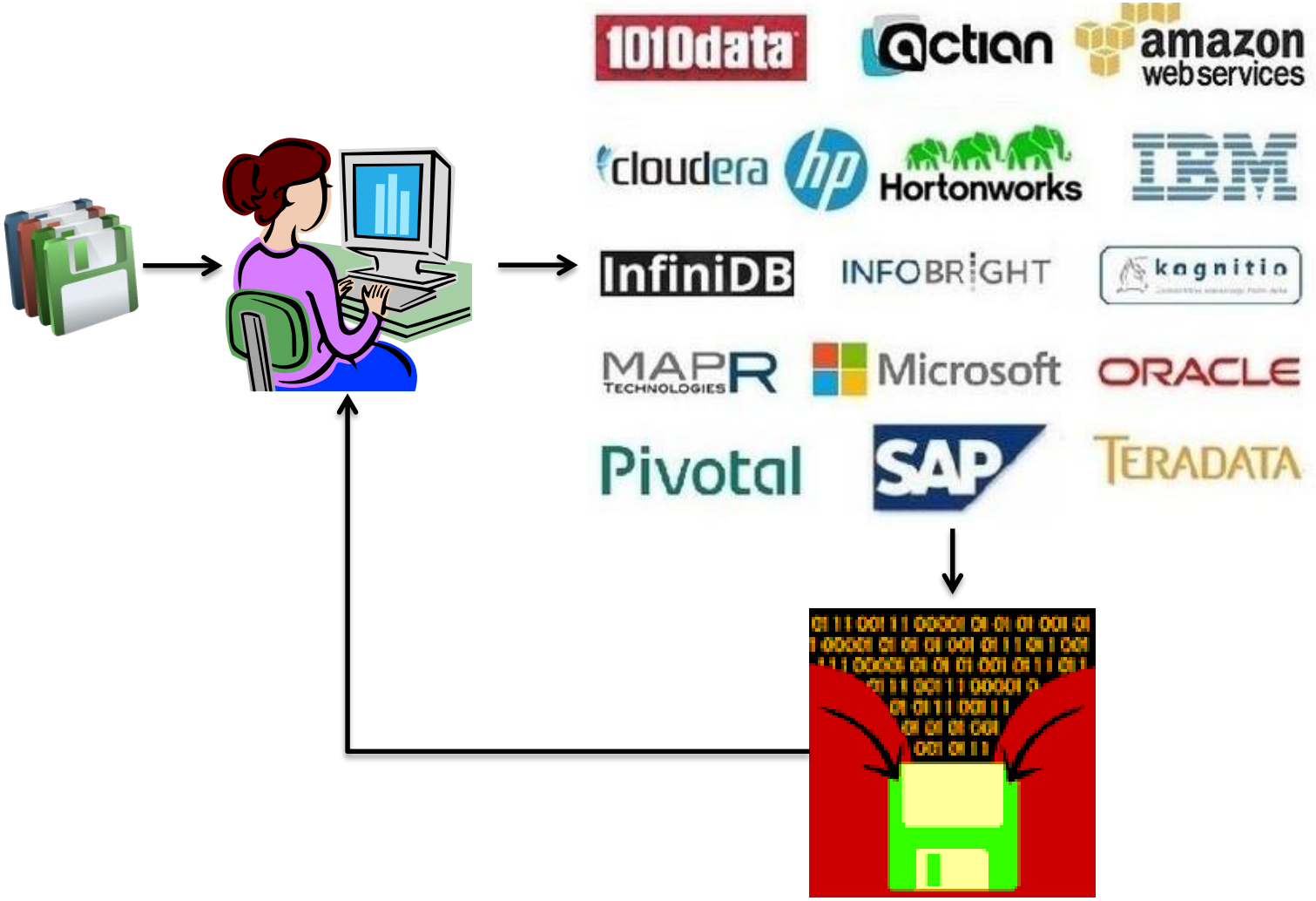
Expertise of ESAT-STADIUS, KU Leuven

Books & Spin-offs
Algorithms
Applications

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

Content

Big Data
What
Who

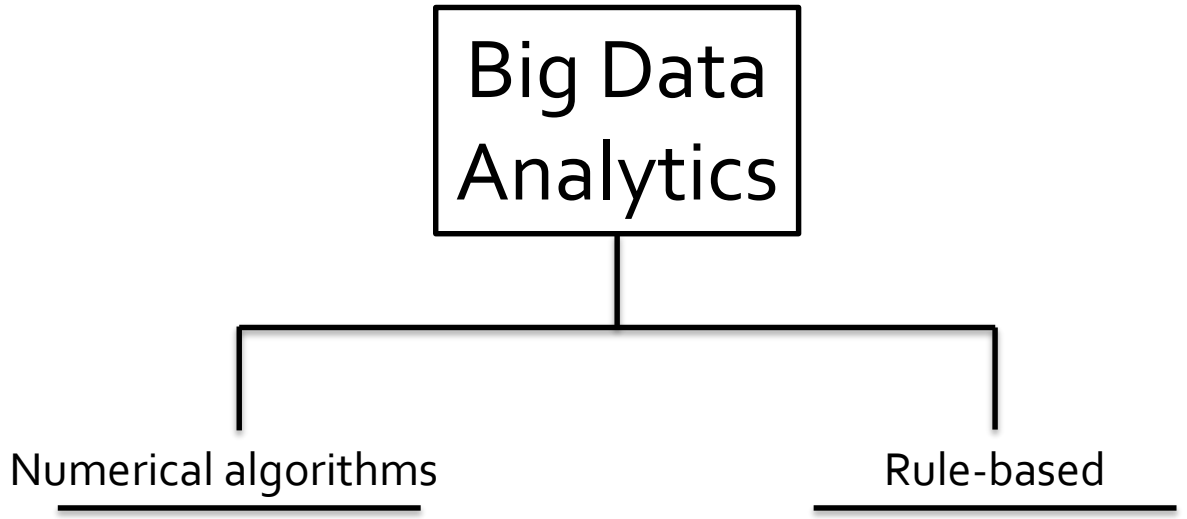
Six issues
Data
Compute Infrastructure
Storage Infrastructure
Analytics
Visualization
Security & Privacy

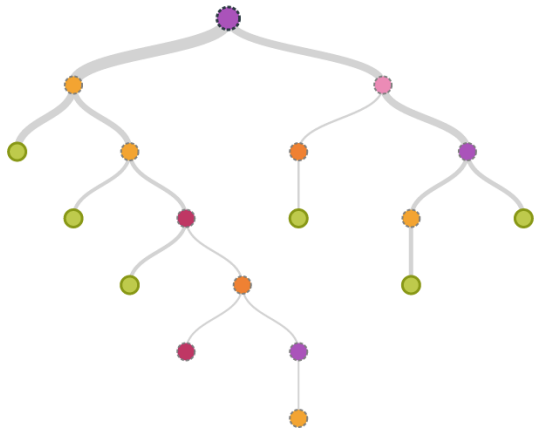
Machine learning as a commodity

Expertise of ESAT-STADIUS, KU Leuven

Books & Spin-offs
Algorithms
Applications

Analytics





Main tasks

Prediction

Regression

Segmentation

Clustering

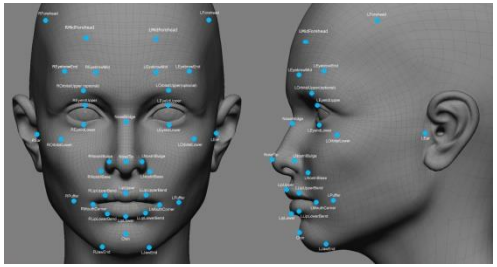
Classification

Anomalies

Outlier

What can we do?

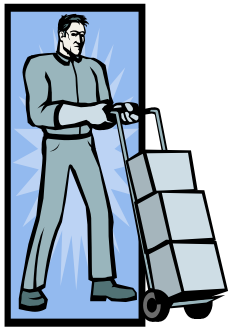
Face recognition



Fraud detection

Shopping cart analysis

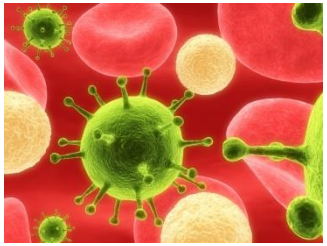
Just-in-time production



Credit worthiness

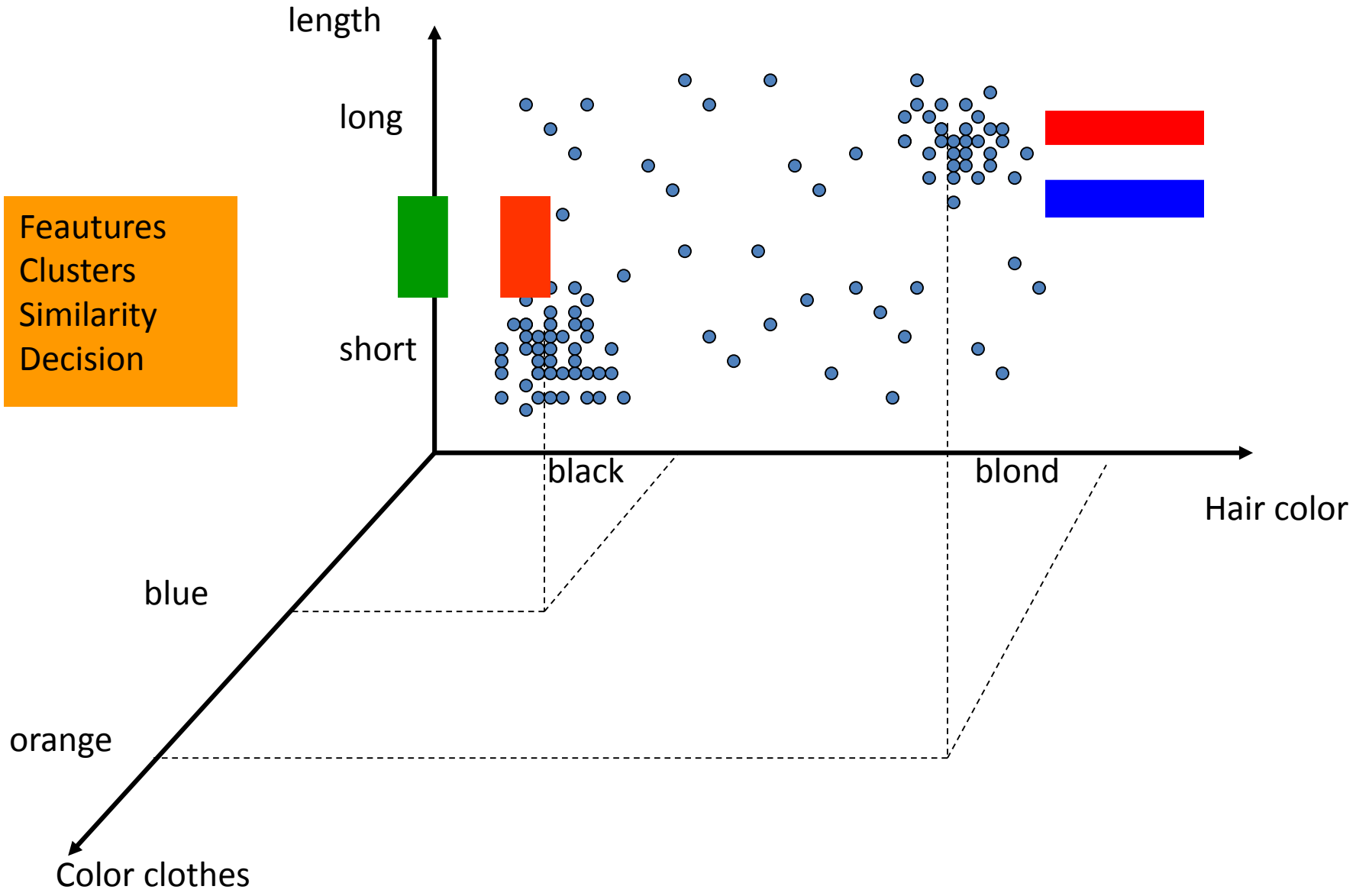
Traffic management

Disease spreading

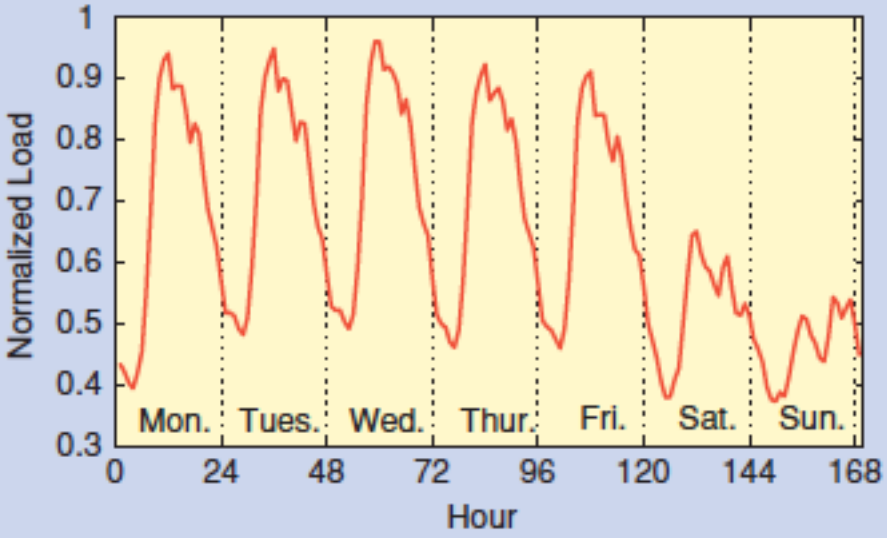


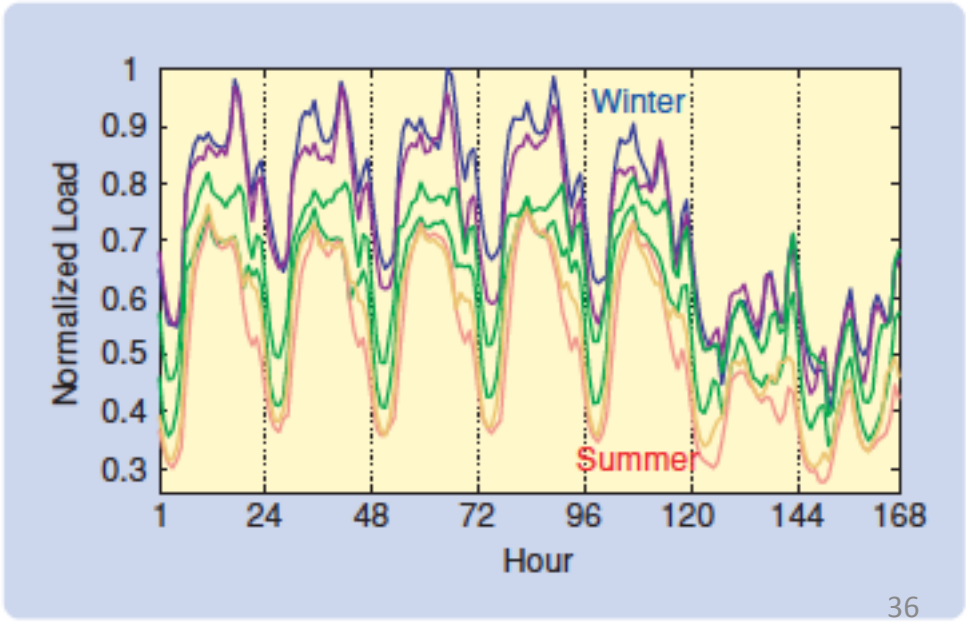
Movie reccomendation

Clustering/Classification



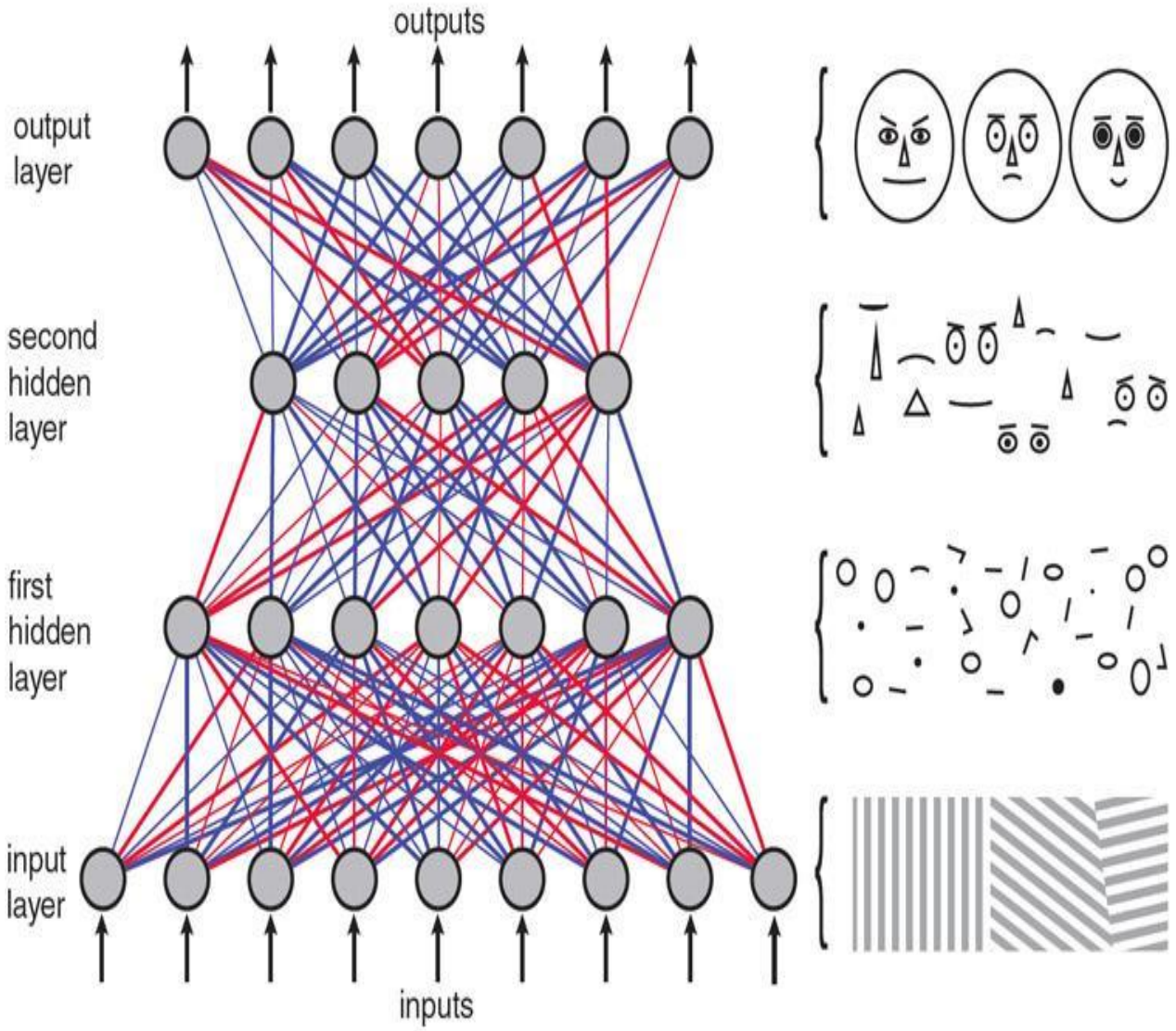
Forecasting



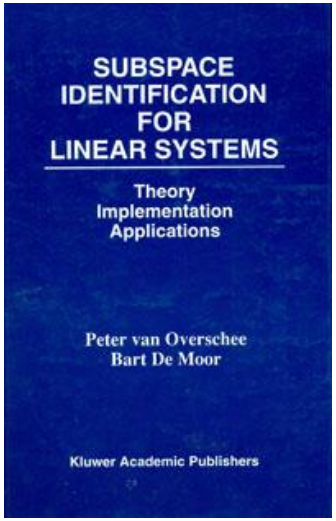


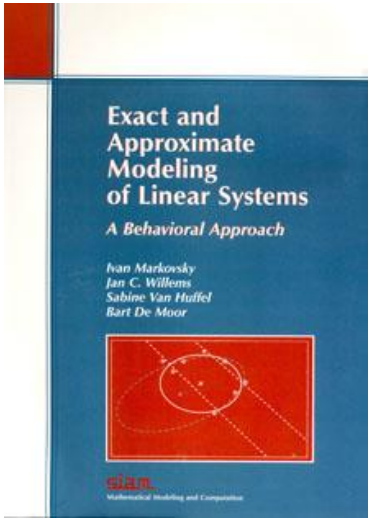
Deep Learning

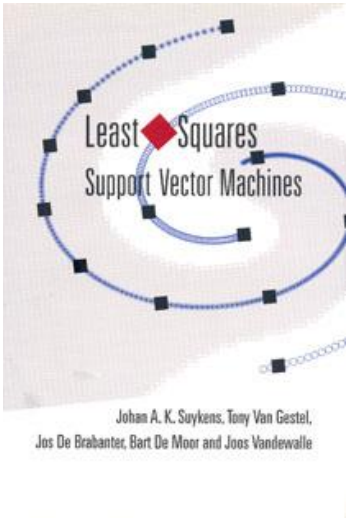
- Neural networks.
- New algorithms.
- Multiple layers on top of each other.
- Each layer learns a more complex representation.
- Learn feature hierarchies.

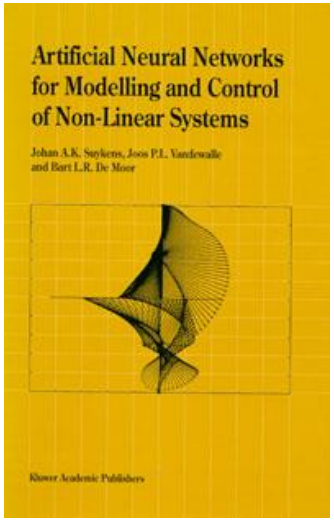


Stadius - Books



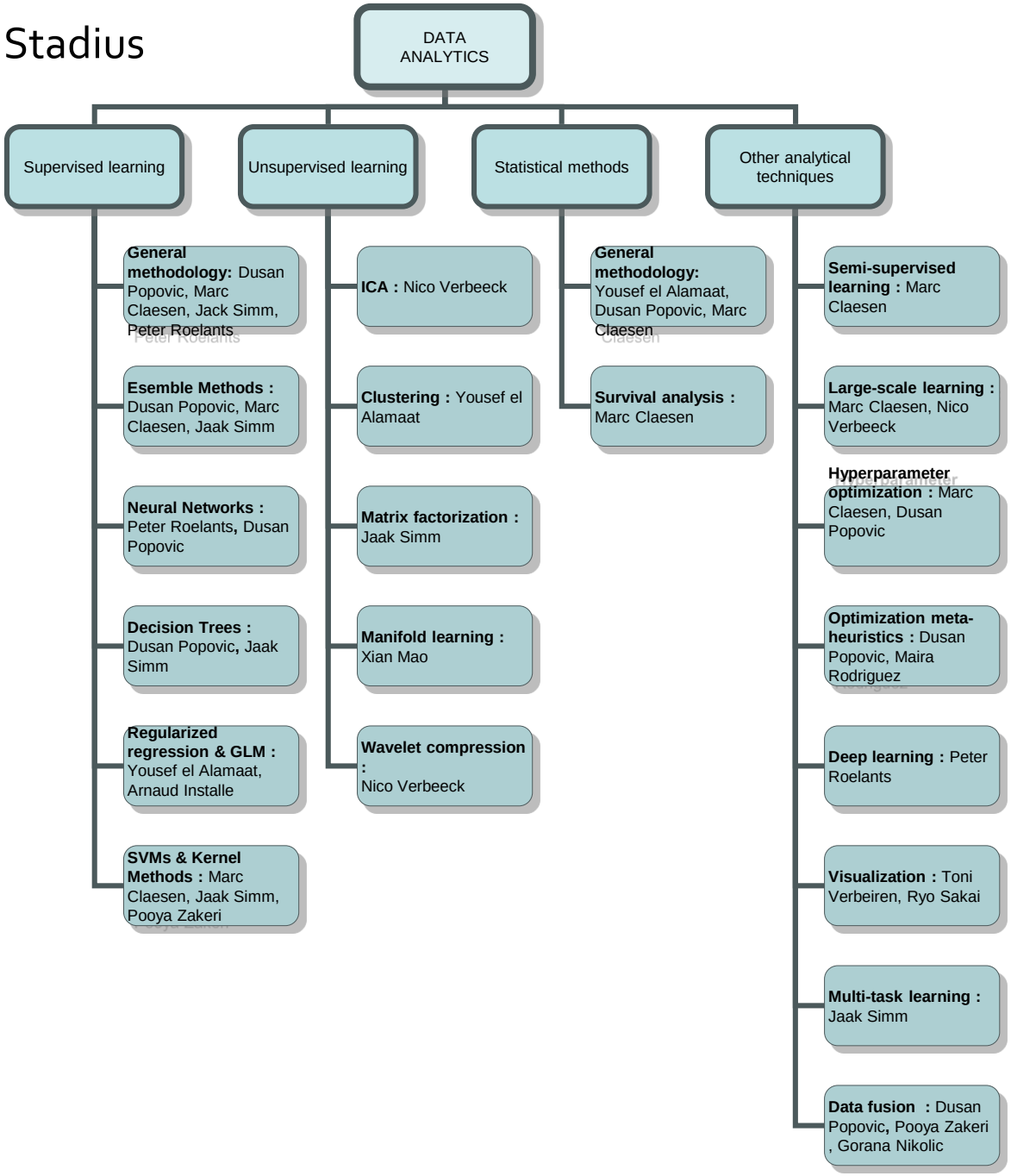




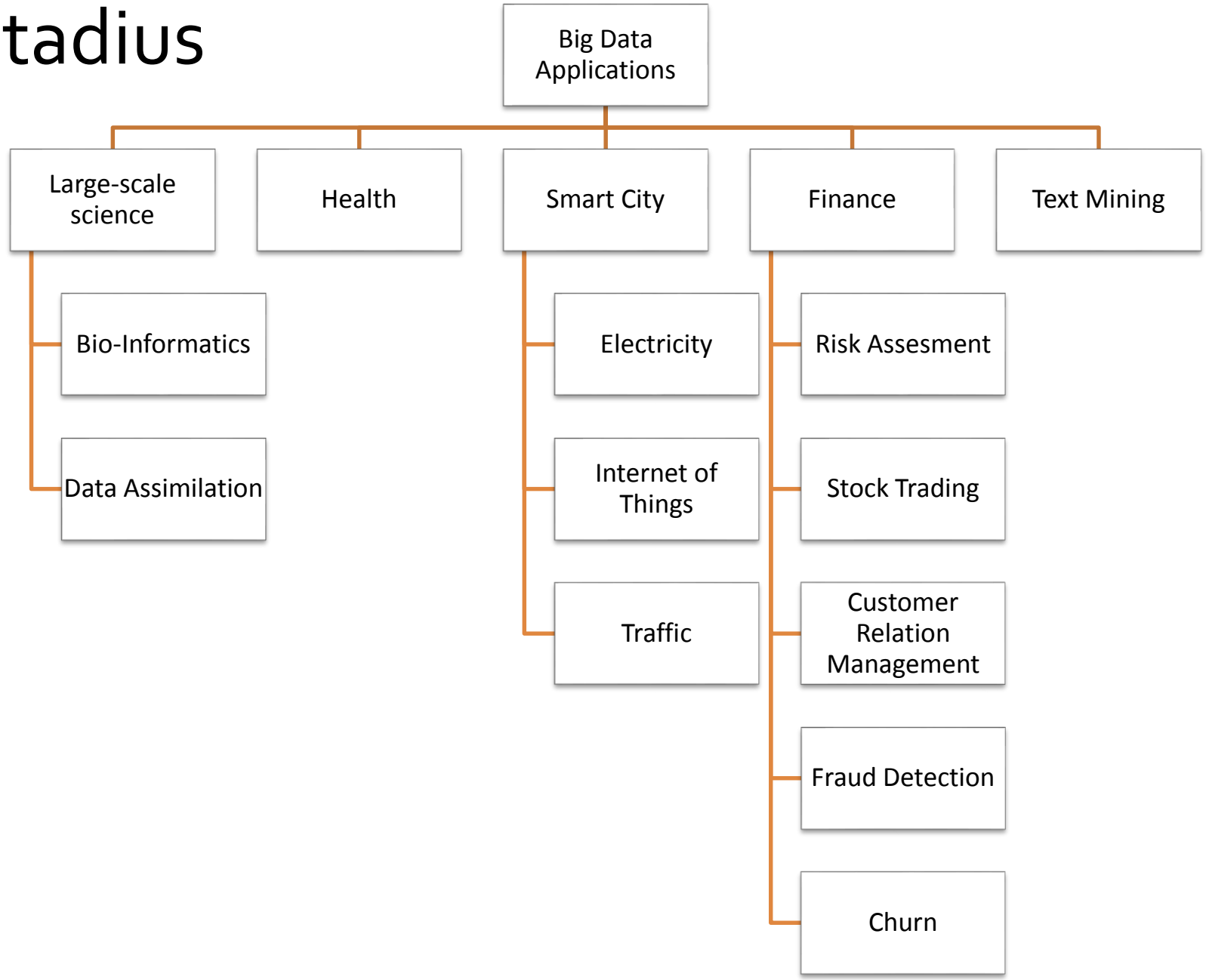


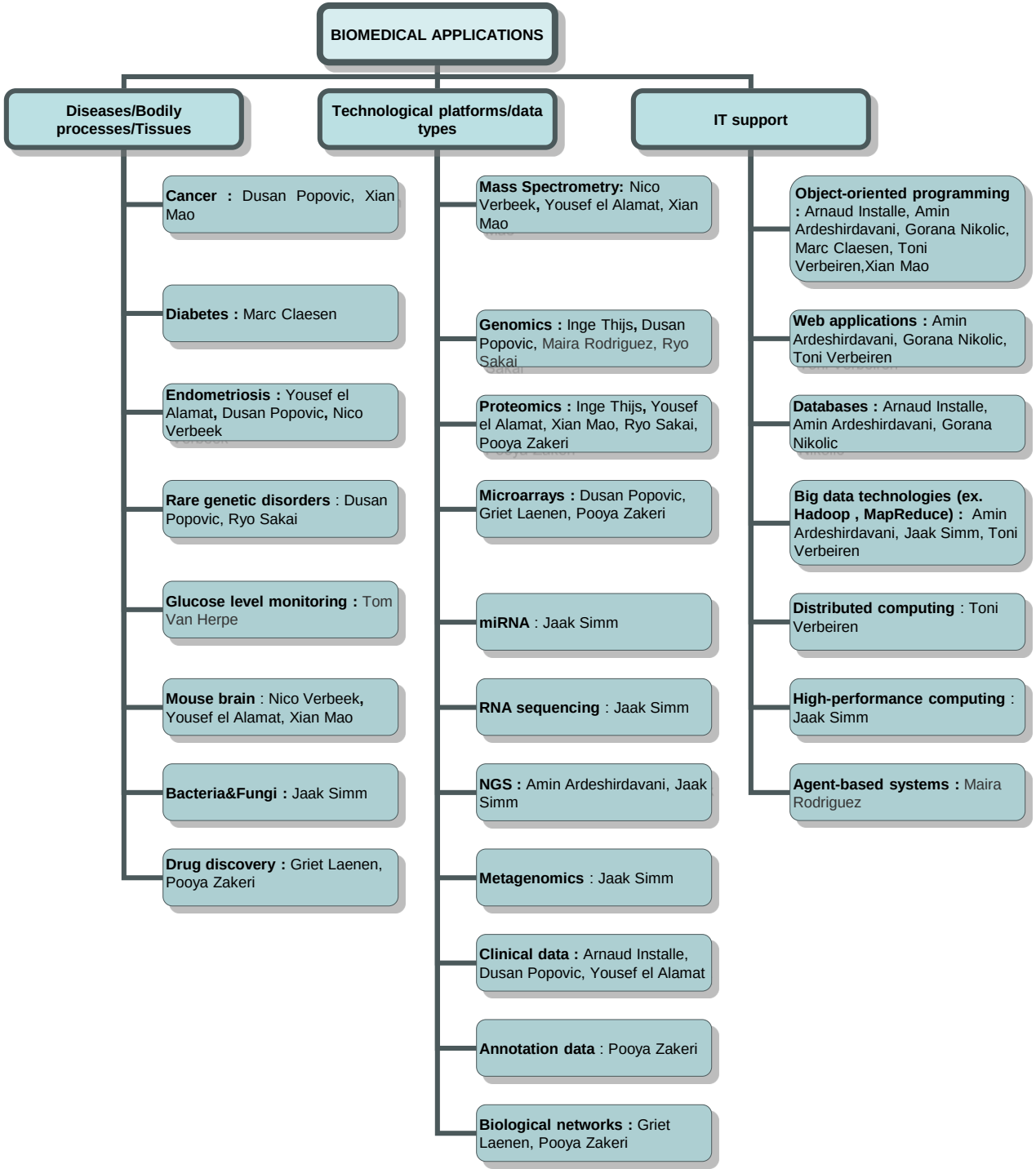
Stadius - Spin-offs

Algorithms in Stadius



Stadius


Big Data Applications
Large-scale science
Health
Smart City
Finance
Text Mining
Bio-Informatics
Data Assimilation
Electricity
Internet of Things
Traffic
Risk Assesment
Stock Trading
Customer Relation Management
Fraud Detection
Churn


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

Energy

Industry

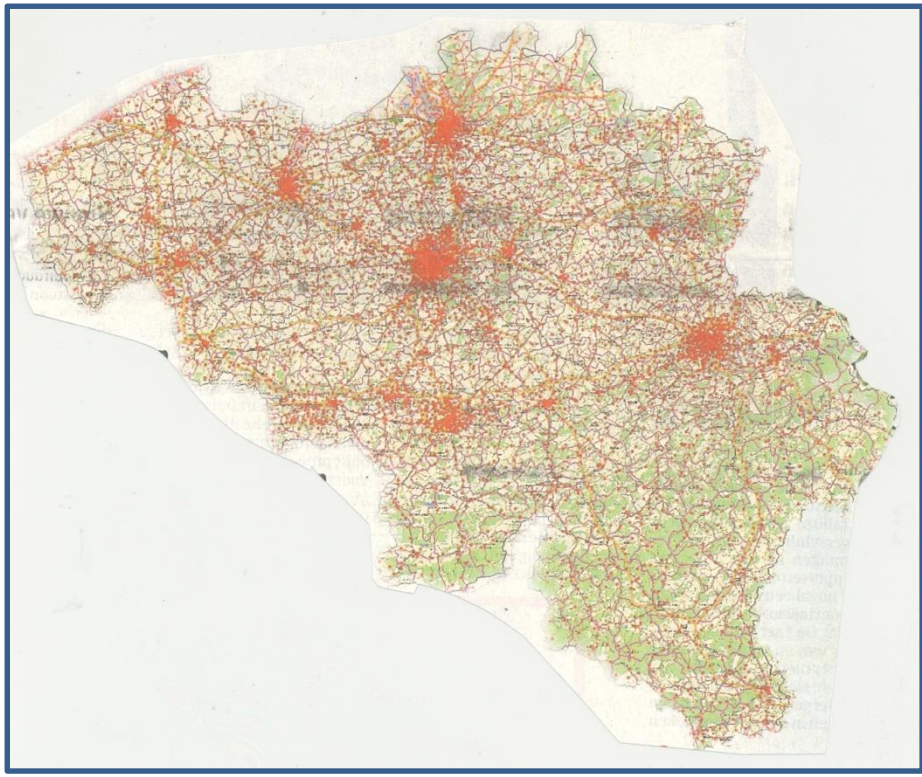
Environment

Social networks

Fraud and predictive analysis

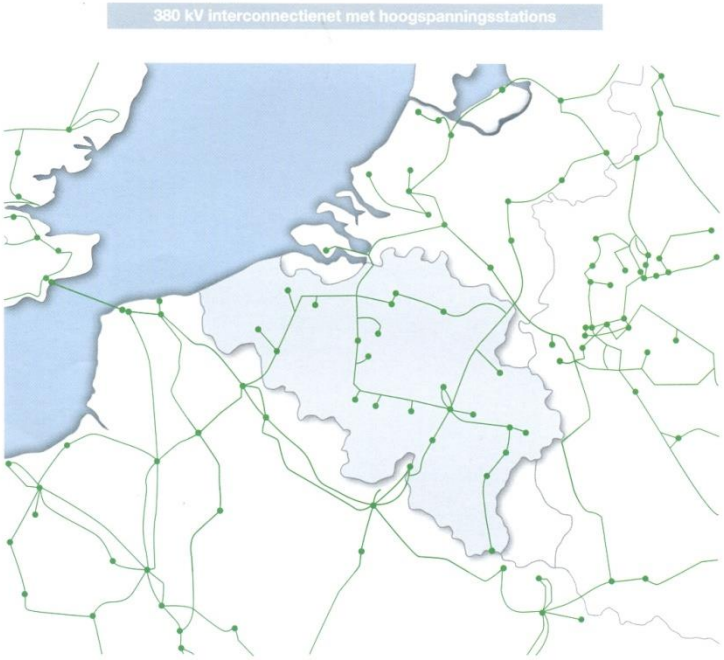
Health

Power grid



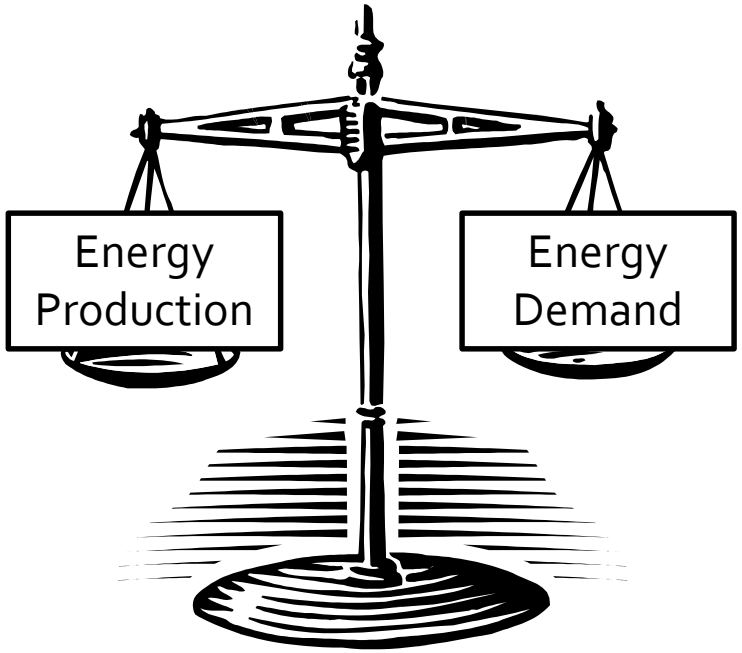
België en Europa

Het Elia-net:
knooppunt van elektriciteitverkeer in Europa



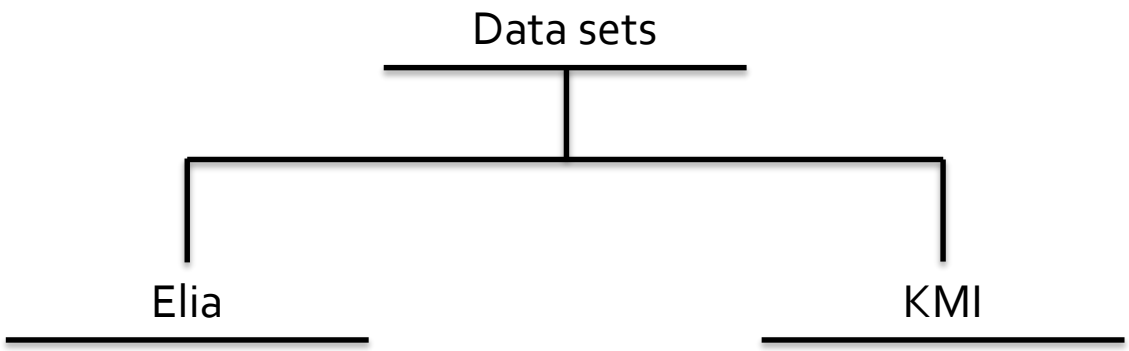
Electric load forecasting

Problem



How to forecast the demand?

Electric load forecasting



Transmission system operator

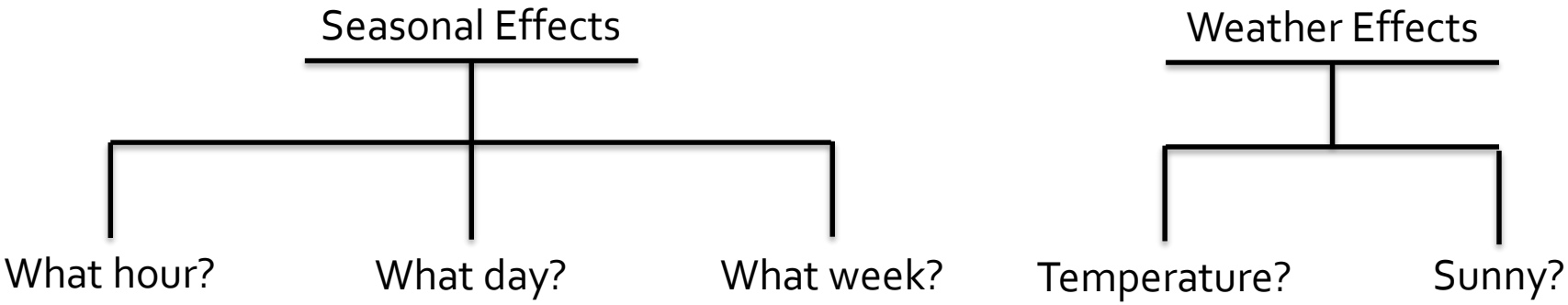
Hourly data

10 substations considered

Meteorological Institute

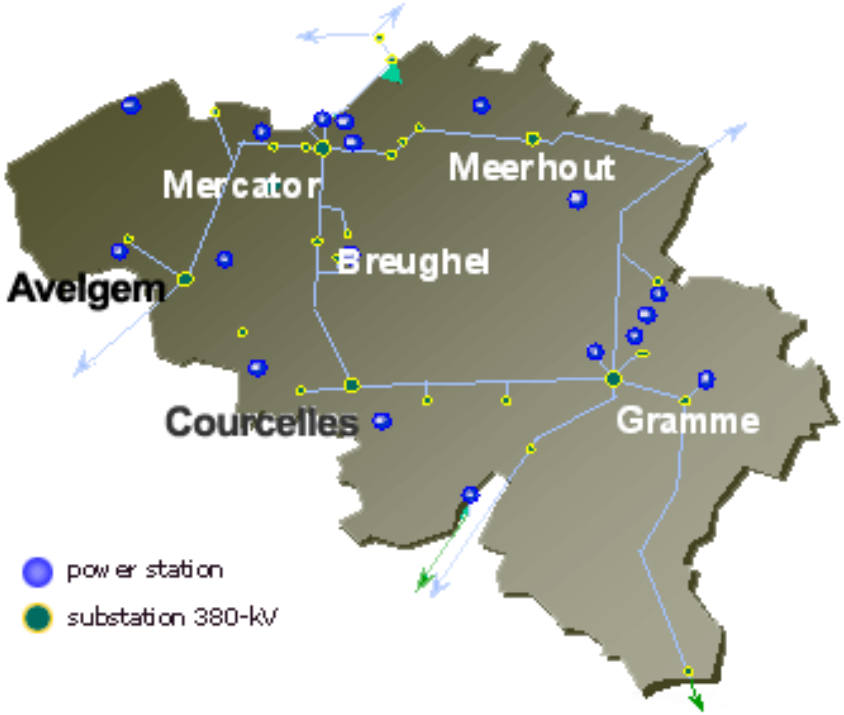
Weather conditions

Electric load forecasting

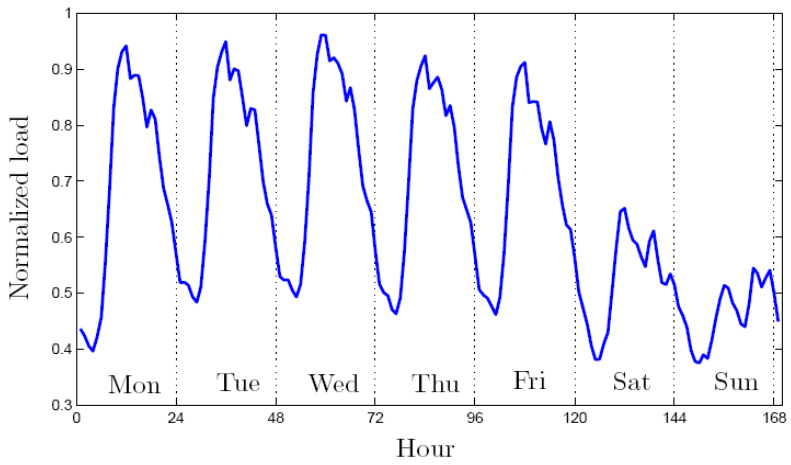


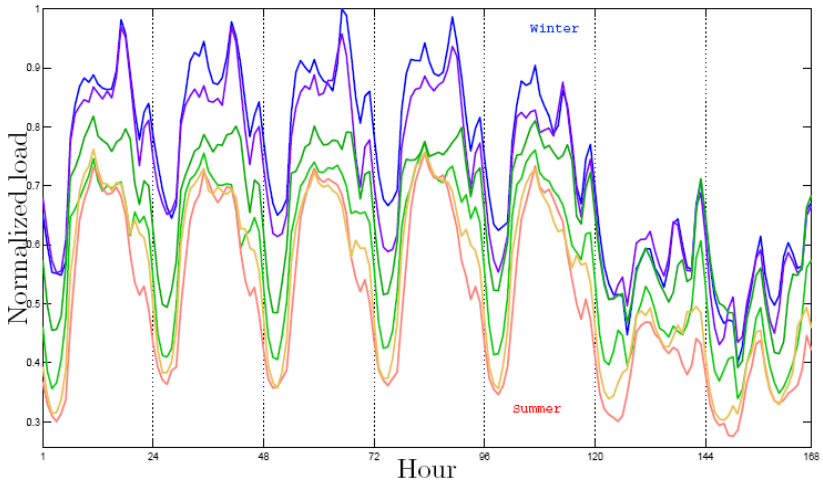
Model update: Every week!

⇨ Accurate forecast



250 transformer substations
Every 15 min, 5 years

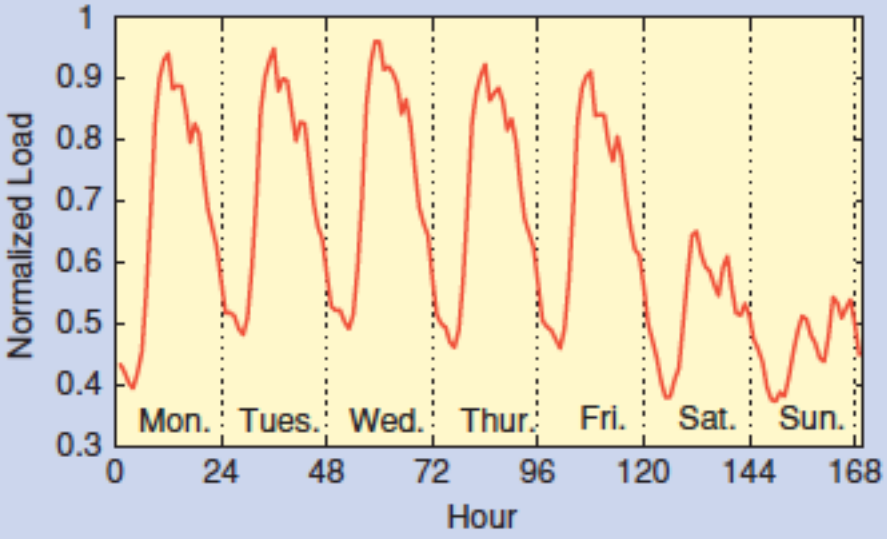


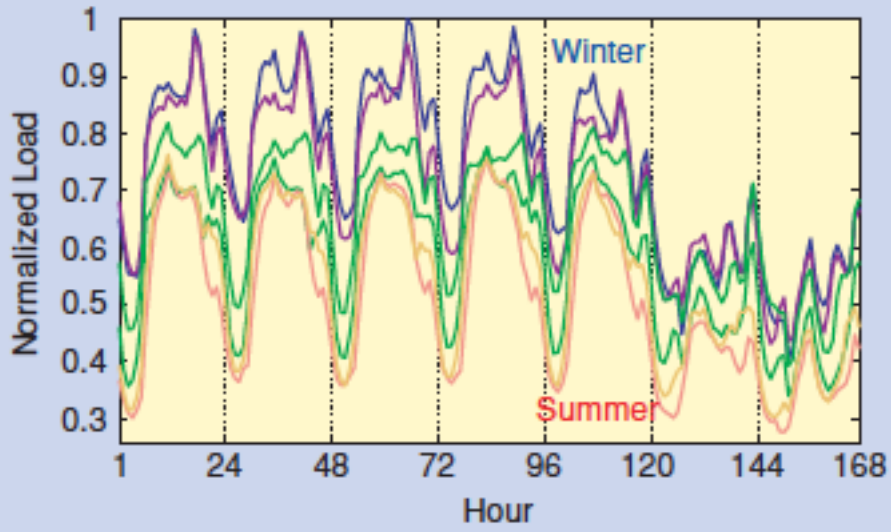


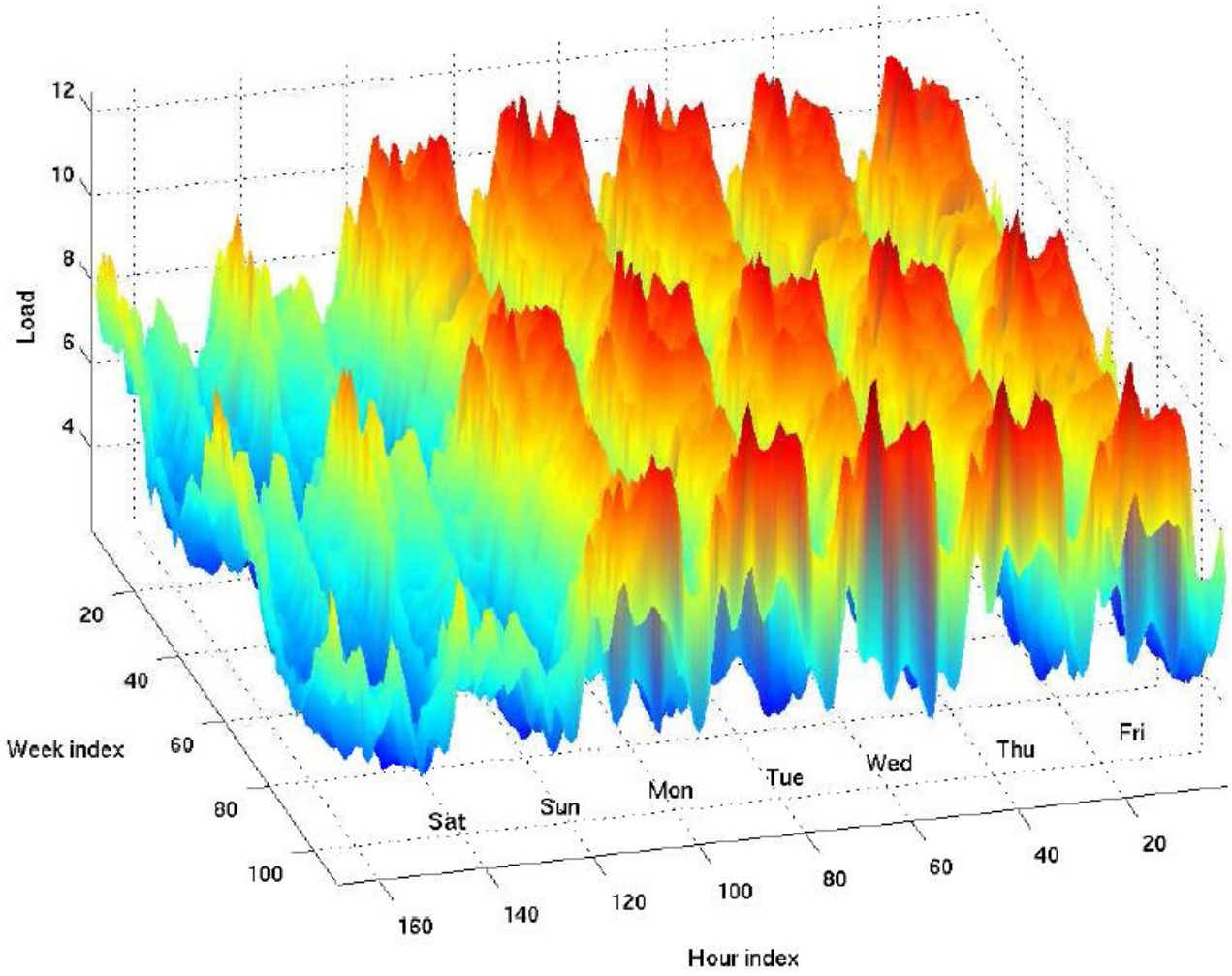
1 post, 1 week

1 post, four seasons

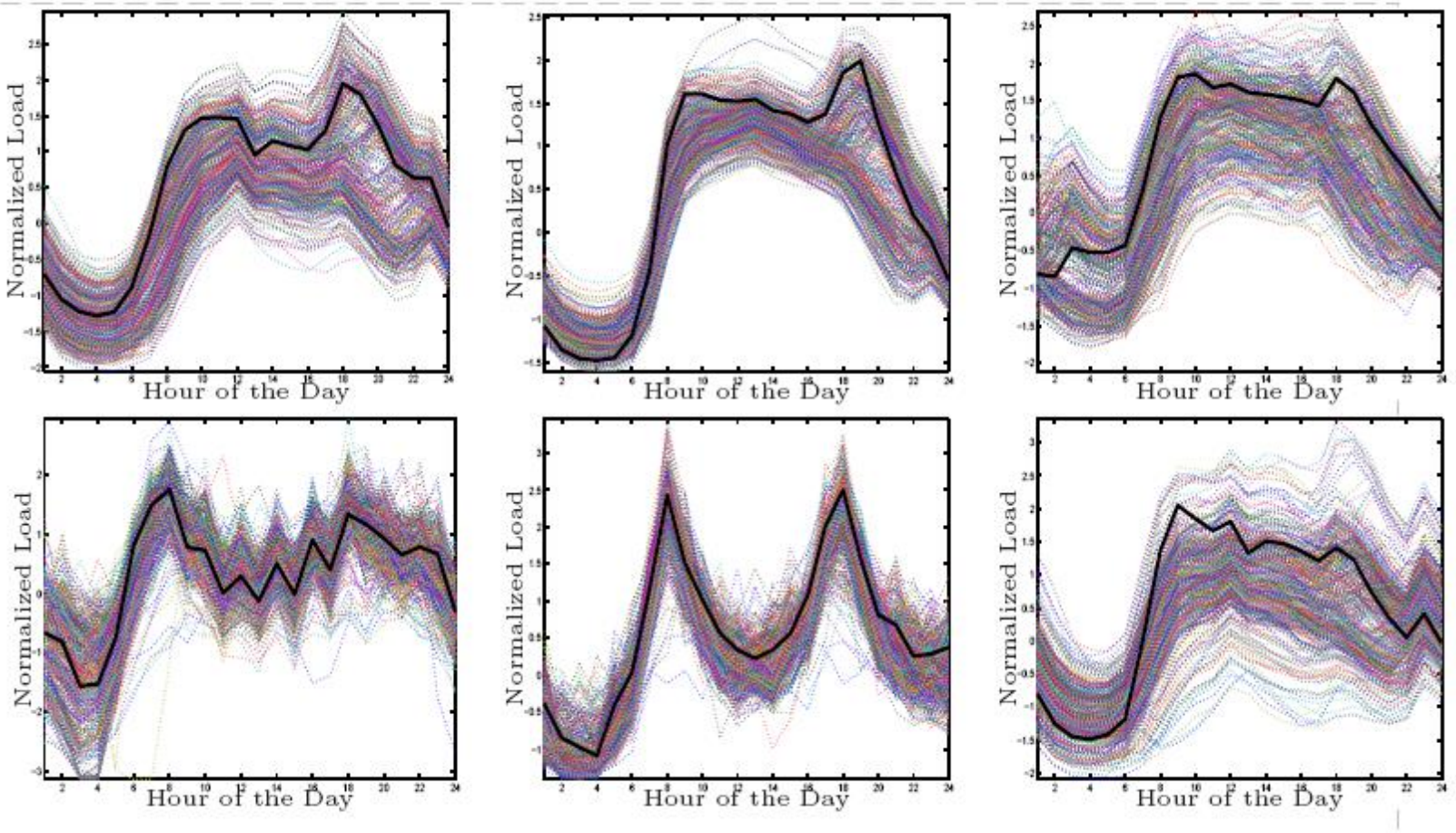
Electric load forecasting



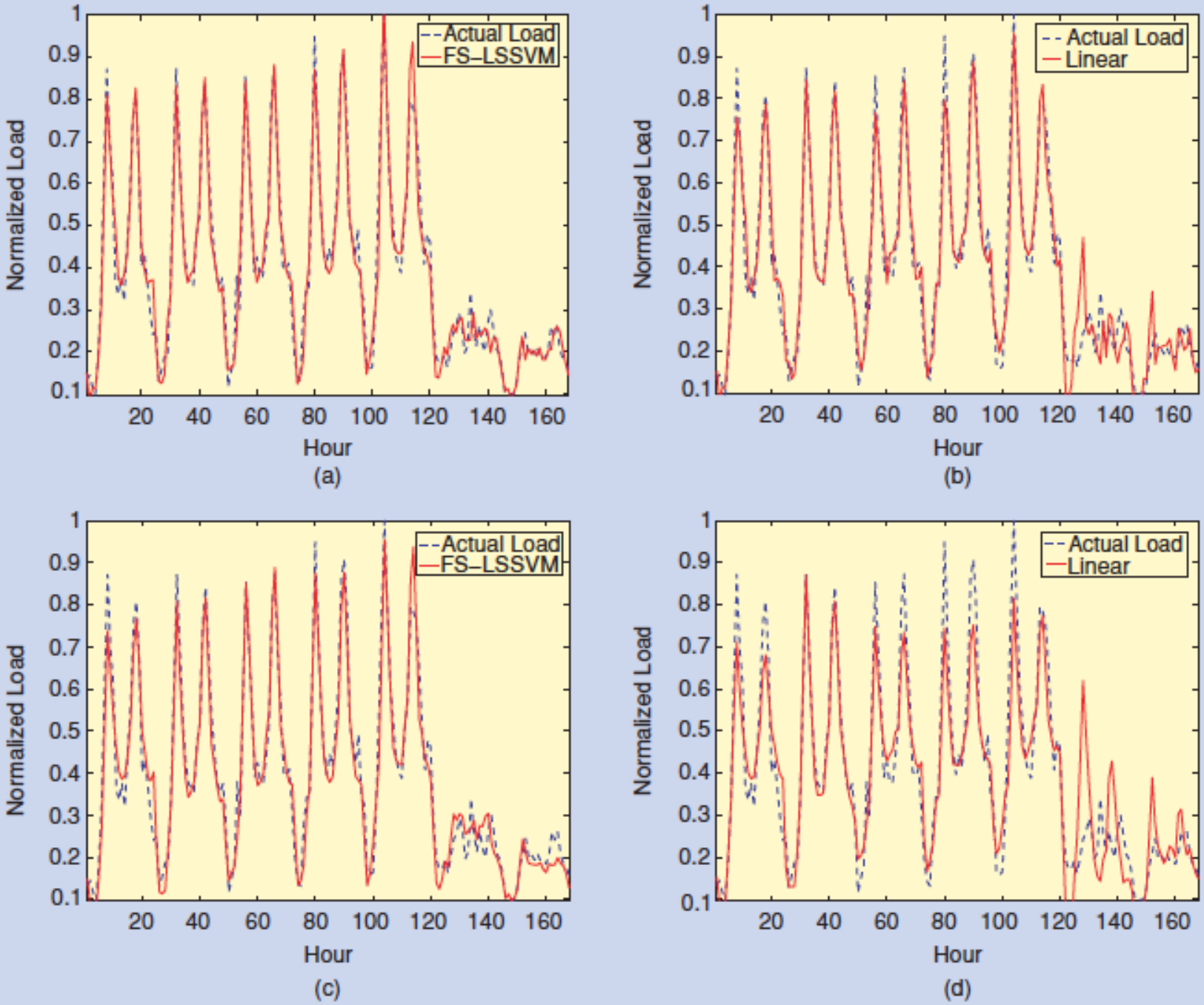




Seasonalities in the load: day, week, year, holidays

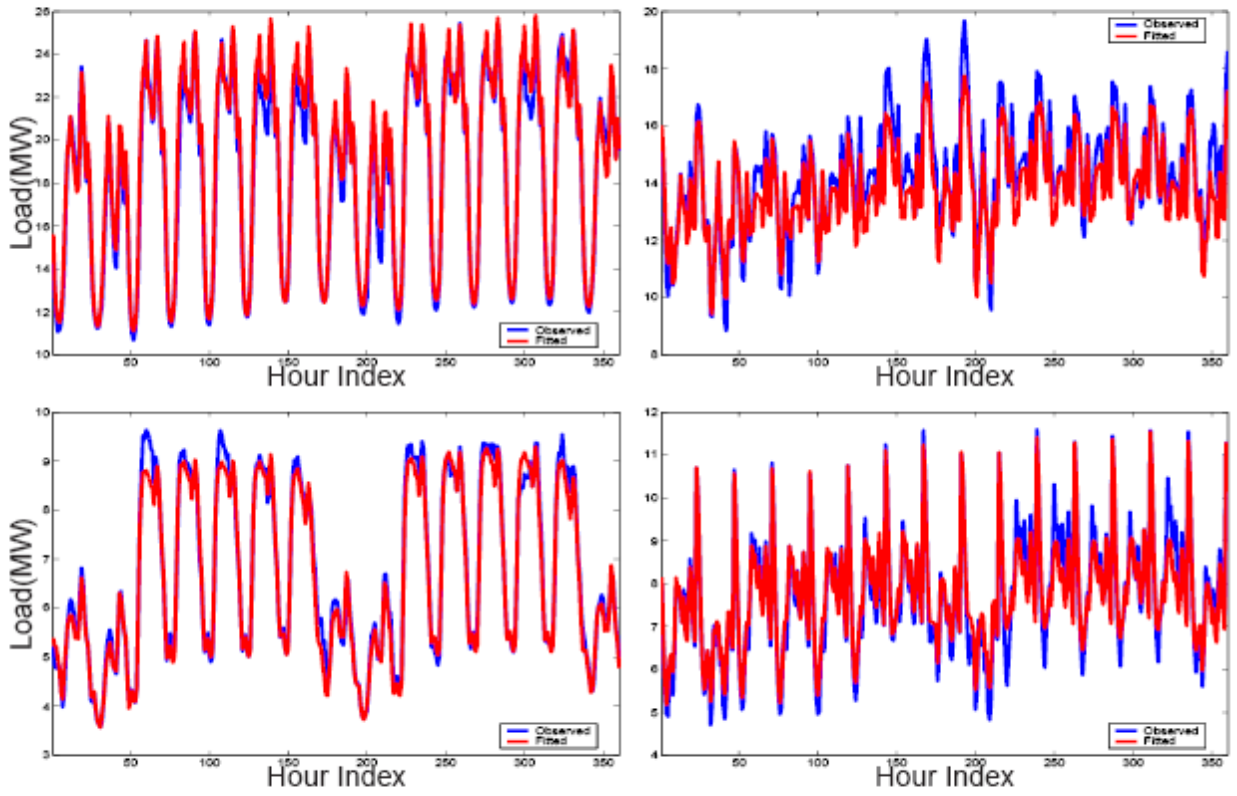


6 posts, 1 year
Seasonalities, calender holidays !


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)

- **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:**



Energy

Industry

Environment

Social networks

Fraud and predictive analysis

Health

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

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

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

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

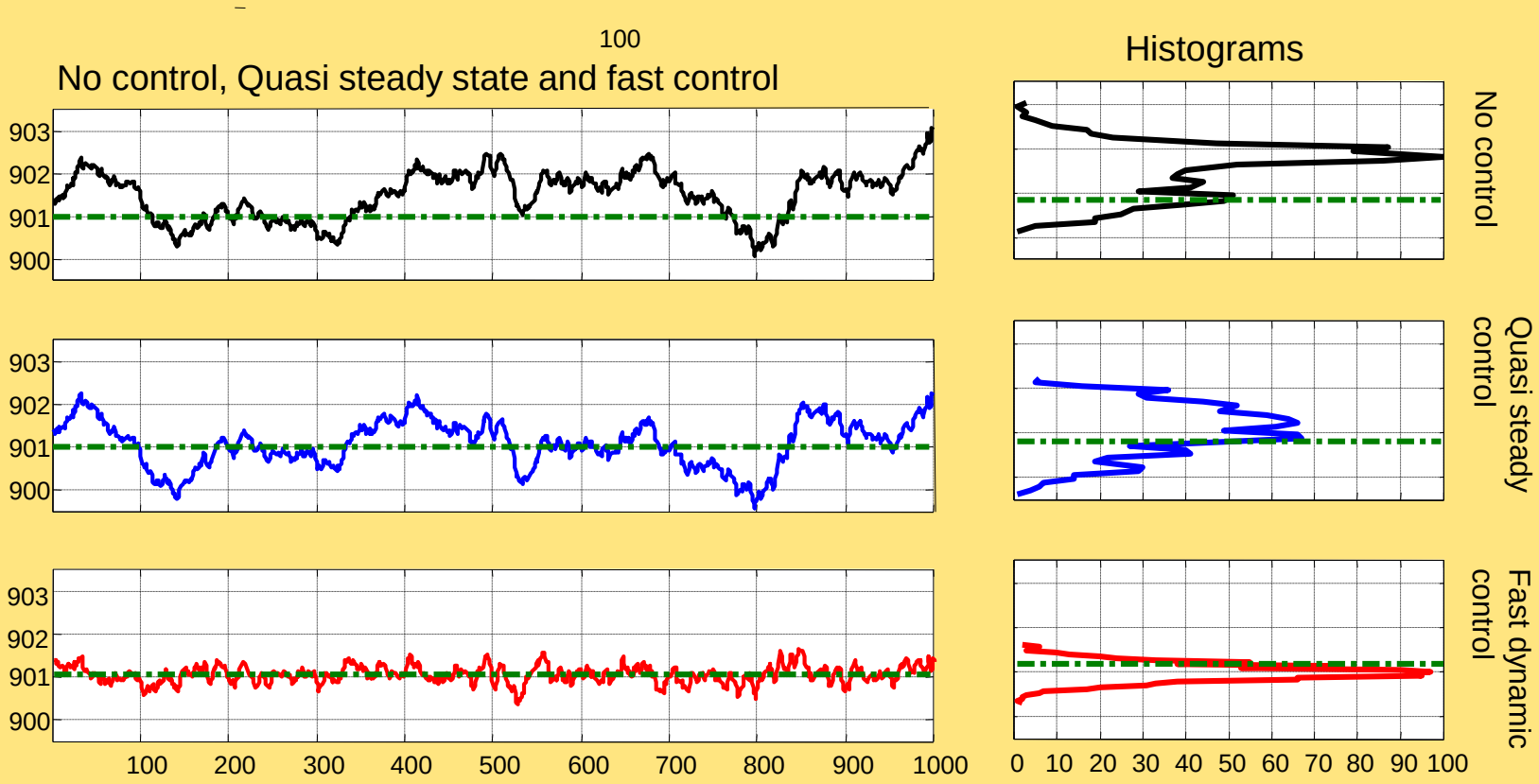
100.0 99.5 99.0 98.5 98.0

12:00 15:00 18:00 21:00 0:00 3:00

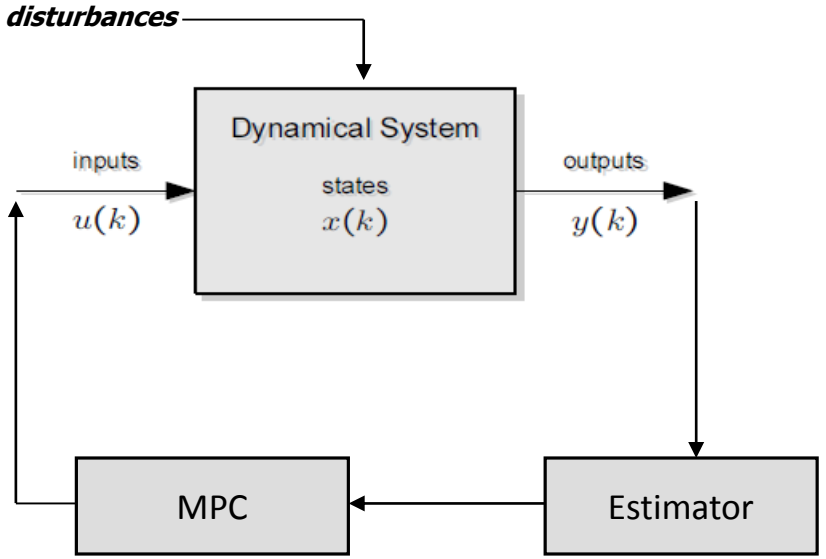
3/7/0 Time 3/8/0

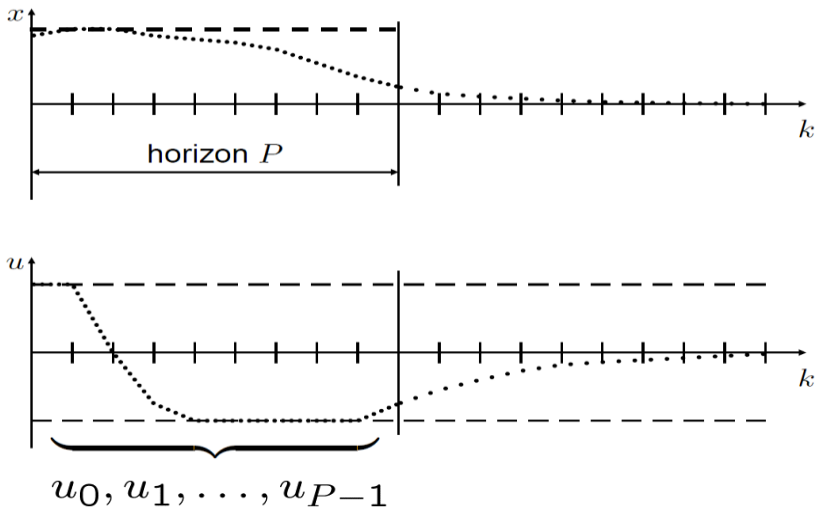
Last Run : 3/8/2000 3:05:15 AM Count : 9322

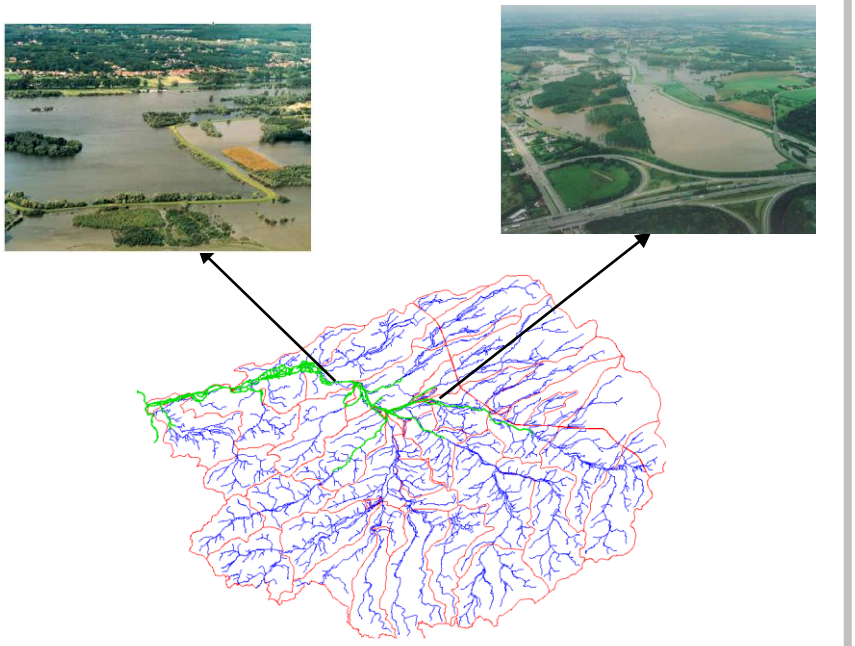
Modelling for control

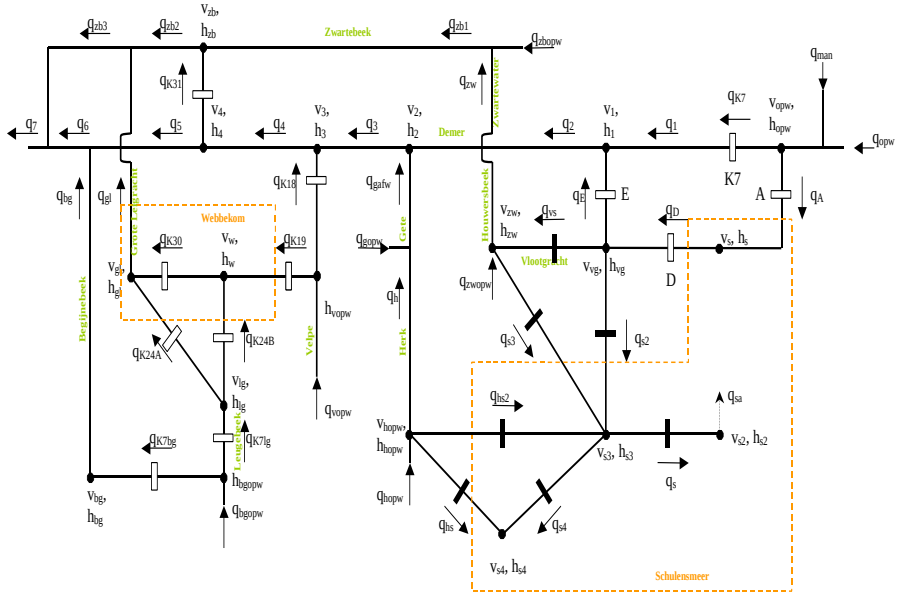


Model Based Predictive Control for Flood Regulation: Demer









d square

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)

I²

APPLICATION ARCHITECTURE

OS SOFTWARE

HARDWARE

inversys INFUSION

I²

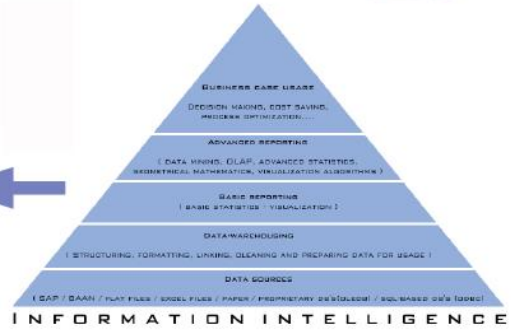
BUSINESS INTELLIGENCE USING PREDICTIVE TECHNOLOGIES
ROI OF 145%

BUSINESS INTELLIGENCE WITHOUT PREDICTIVE TECHNOLOGIES:
ROI OF 89%

PRODUCTION IMPROVEMENT
ROI OF 237%

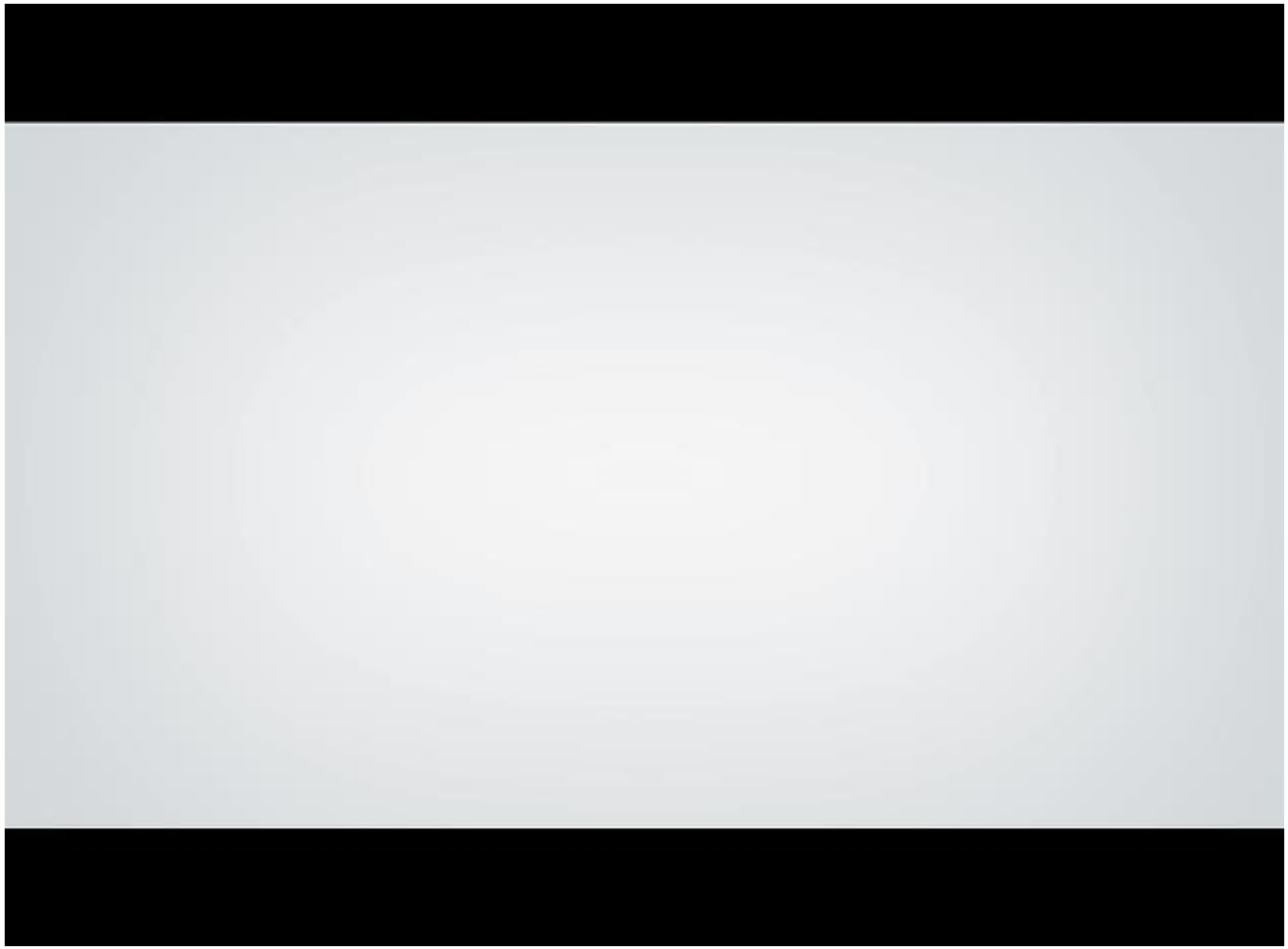
IDC Analyze the Future

BUSINESS ADDED VALUE, COMPETITIVE ADVANTAGE VS. DEGREE OF INTELLIGENCE



INFORMATION INTELLIGENCE

Dsquare website



Energy

Industry

Environment

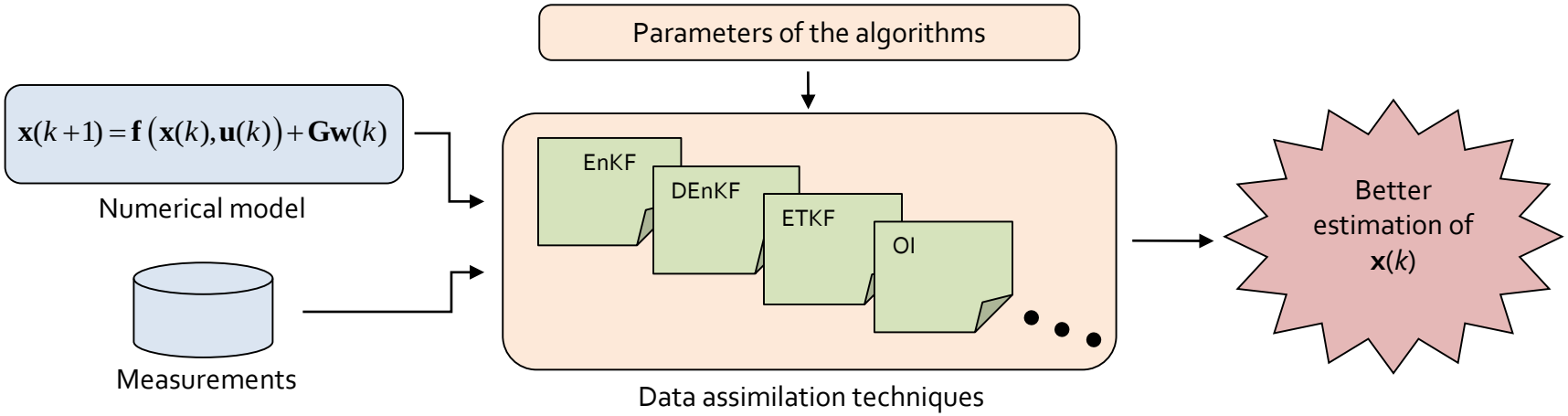
Social networks

Fraud and predictive analysis

Health

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.

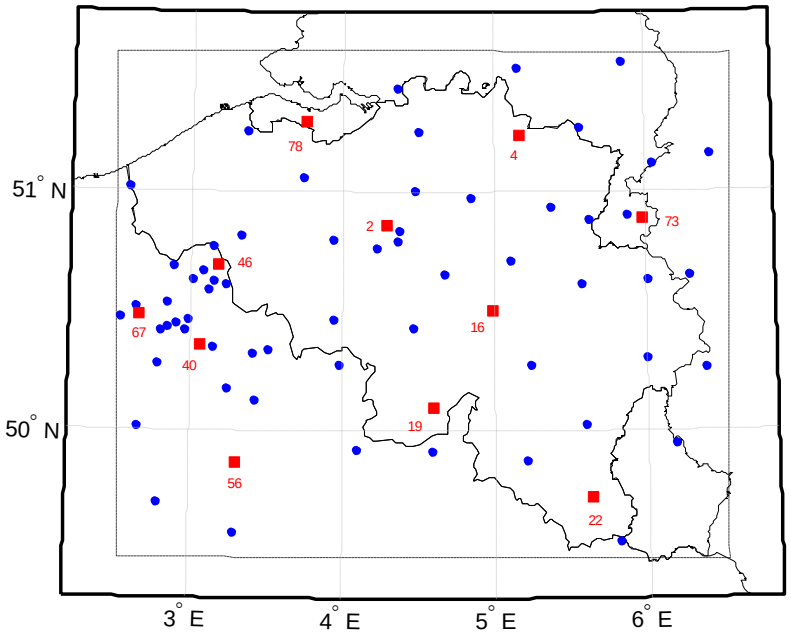


Some data assimilation techniques: 3DVAR, 4DVAR, Ensemble Kalman Filter (EnKF) and its variants, Optimal Interpolation (OI), particle filters, etc.

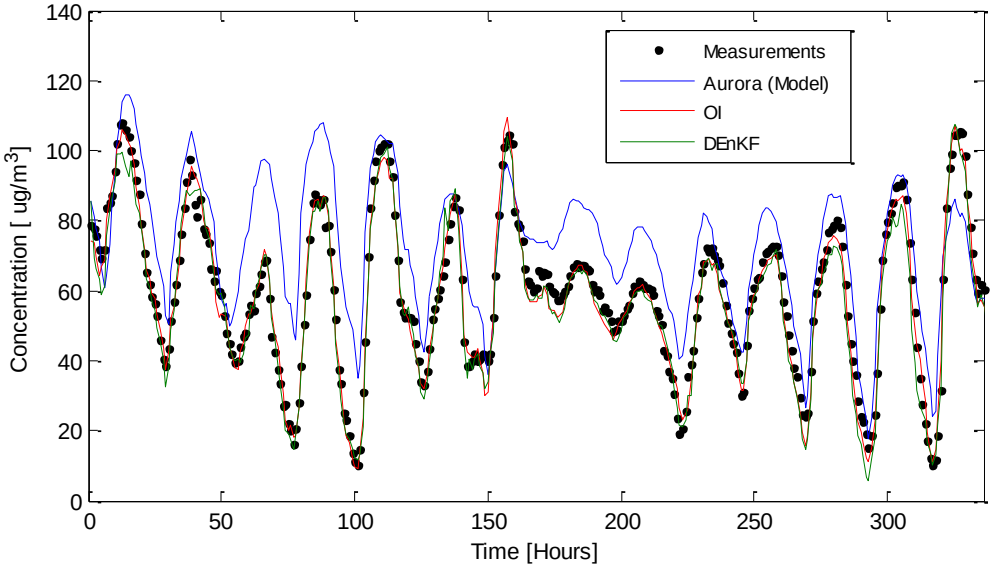
Data Assimilation

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

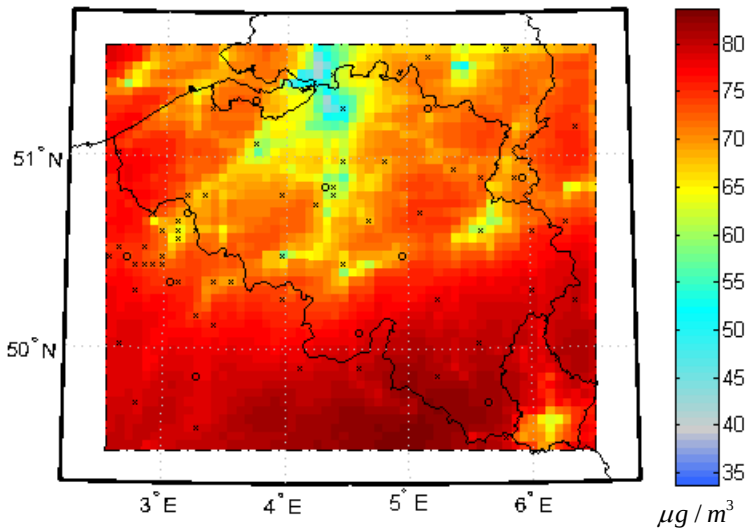


- Assimilation stations
- Validation stations

Starting date: May 28th, 2005 at midnight

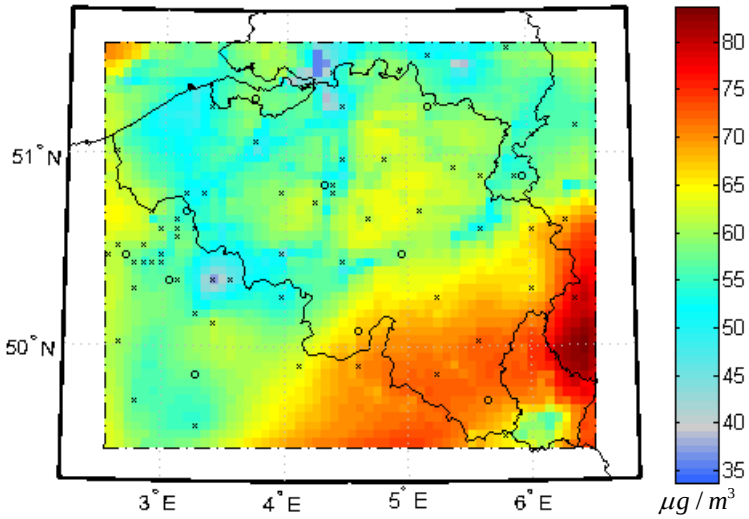
Average of the O_3 concentration field over the 14 day period

Free-run of Aurora

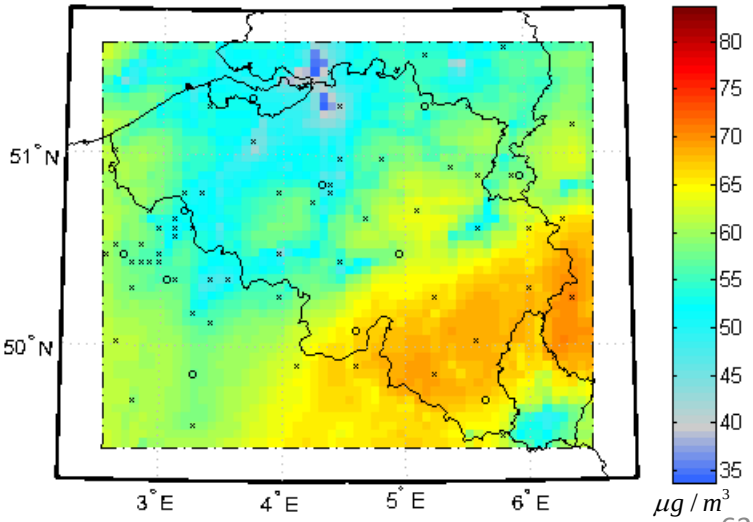


x - Assimilation station
o - Validation station

Optimal Interpolation



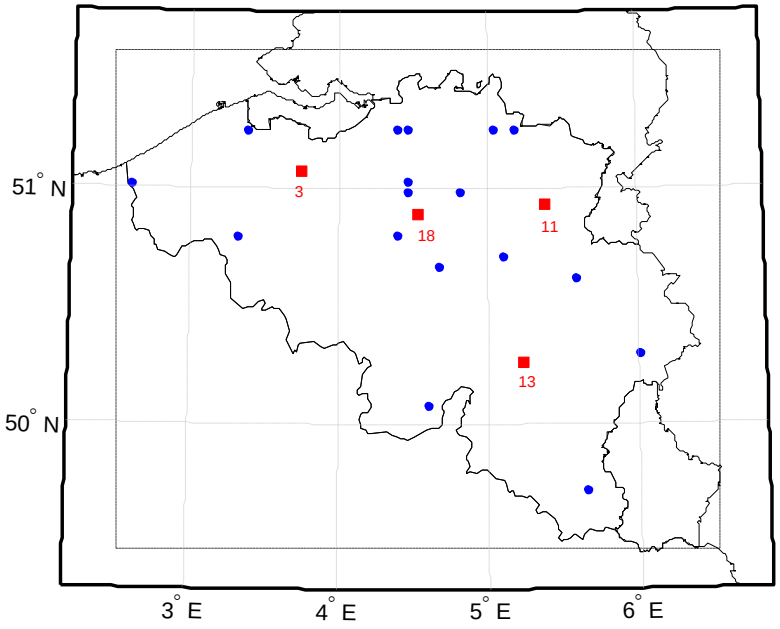
DEnKF



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.

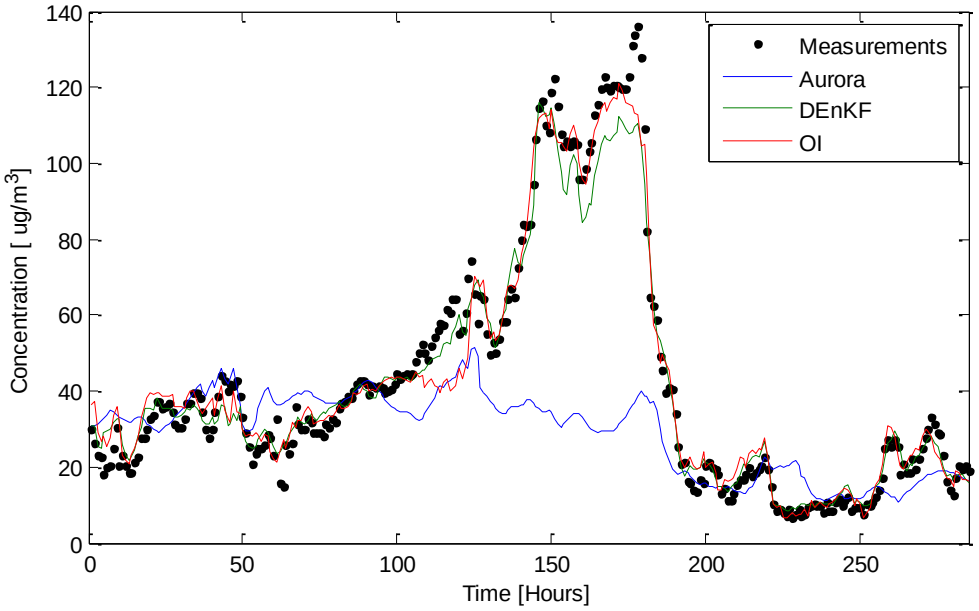
PM_{10} air-quality stations



● - Assimilation stations

■ - Validation stations

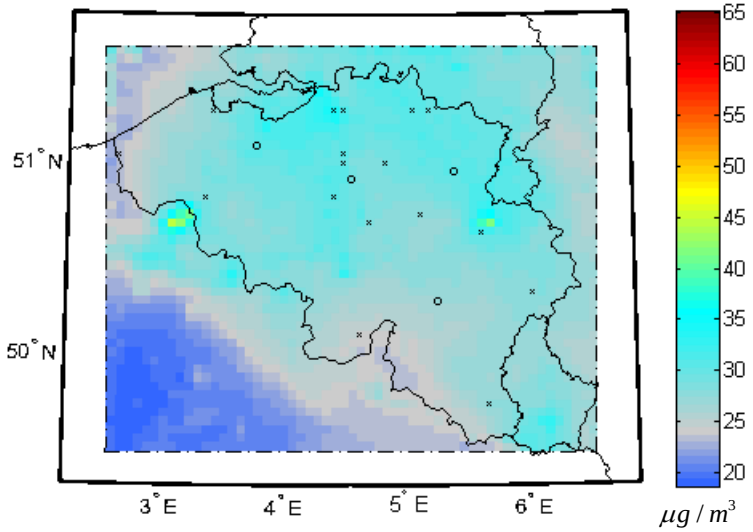
Average of the PM_{10} concentration over the validation stations



Starting date: January 20th, 2010 at midnight

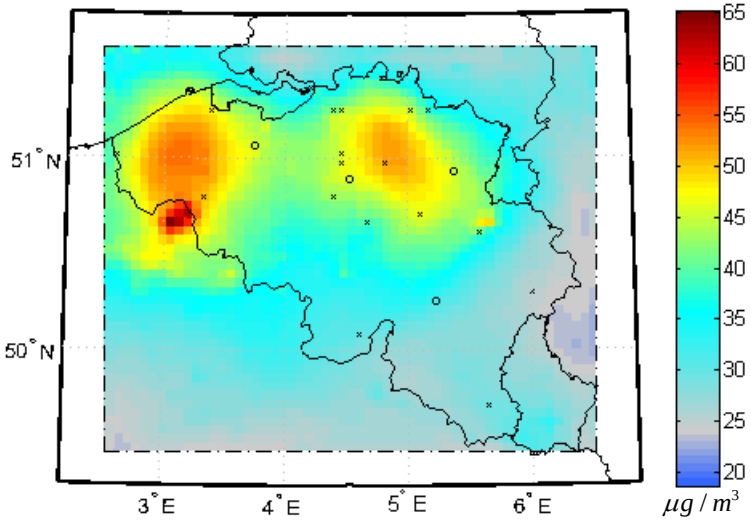
Average of the PM_{10} concentration field

Free-run of Aurora

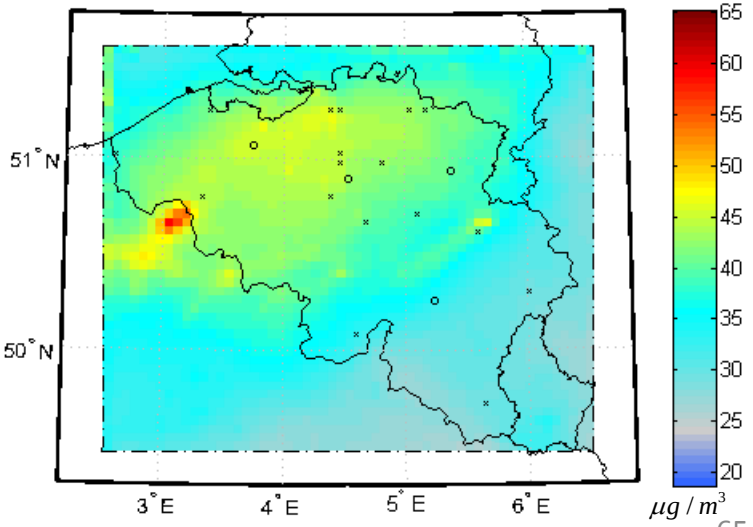


x - Assimilation station
o - Validation station

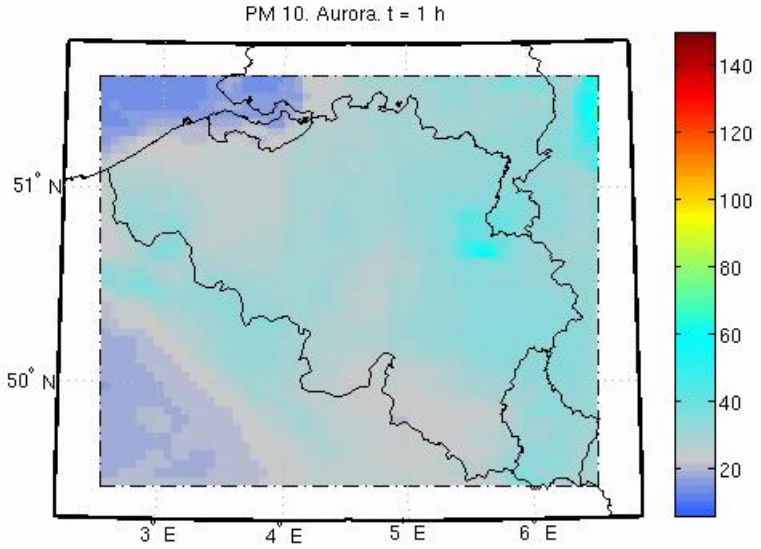
Optimal Interpolation

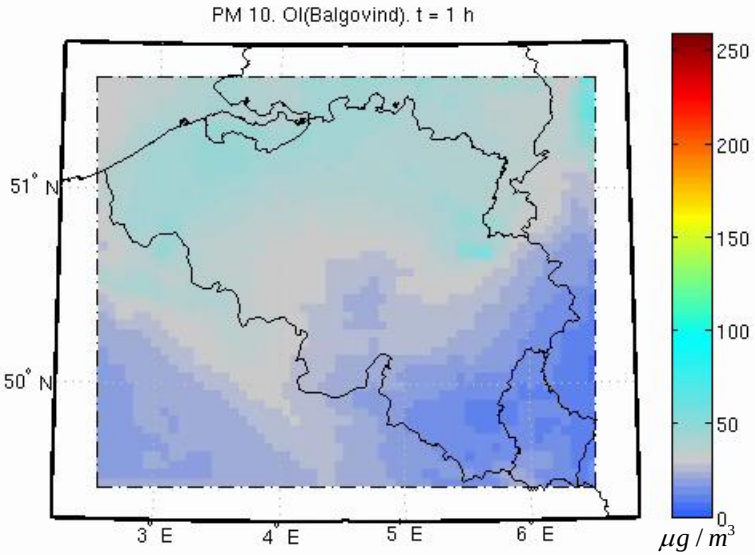


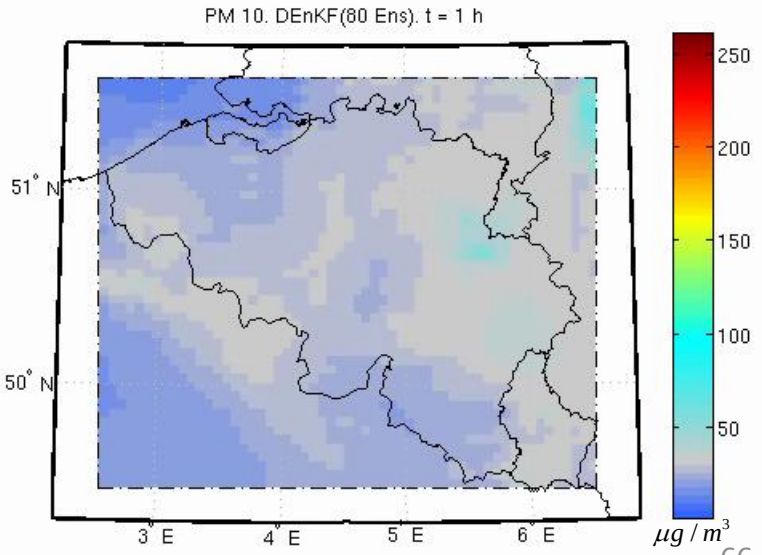
DEnKF



Average of the PM_{10} concentration field







Energy

Industry

Environment

Social networks

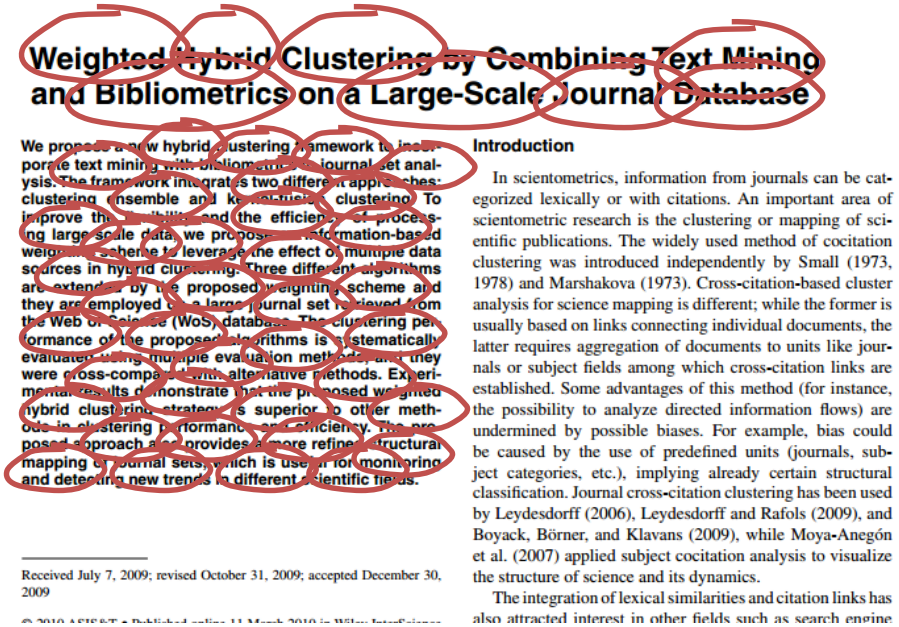
Fraud and predictive analysis

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.

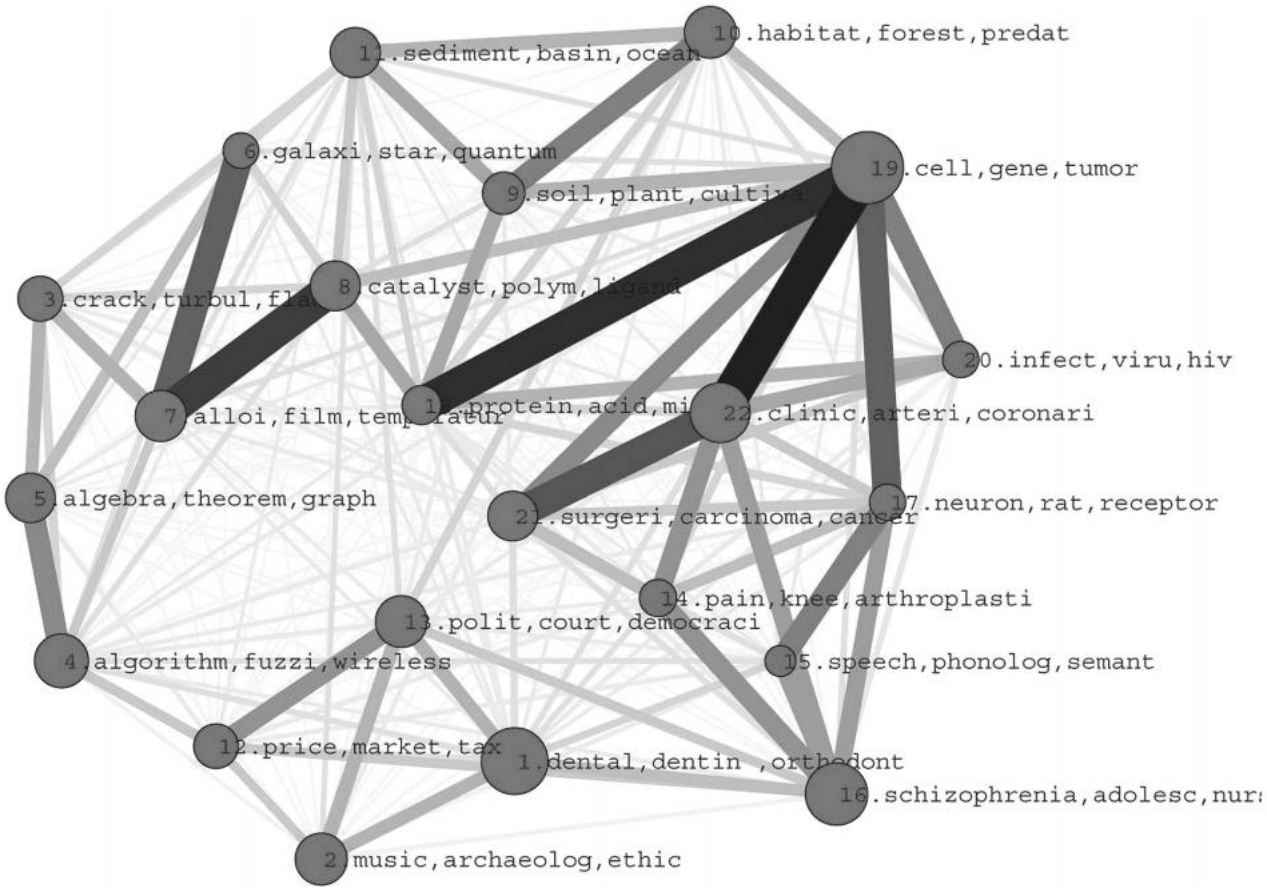
Received July 7, 2009; revised October 31, 2009; accepted December 30, 2009

Introduction

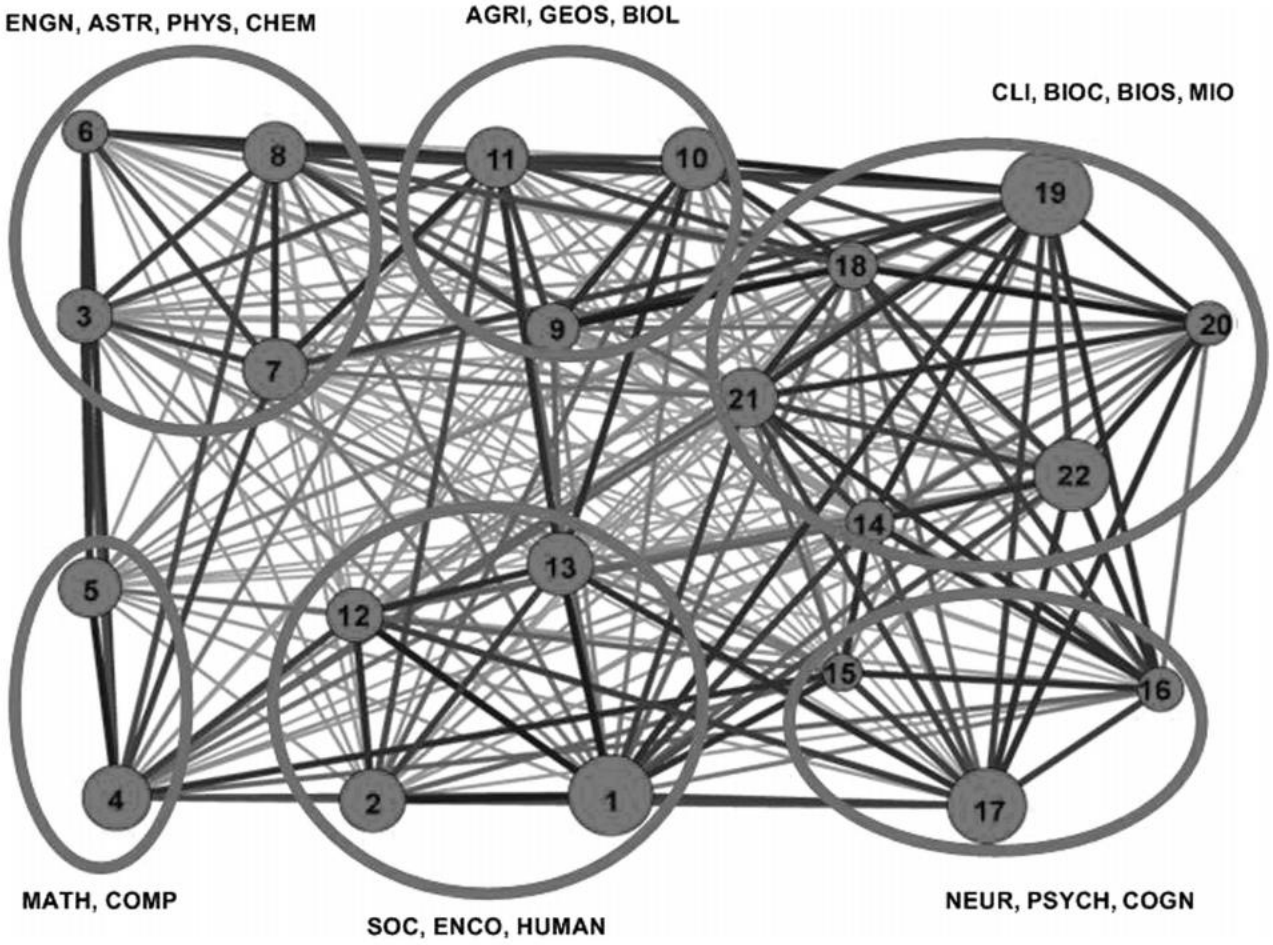
In scientometrics, information from journals can be categorized lexically or with citations. An important area of scientometric research is the clustering or mapping of scientific publications. The widely used method of cocitation clustering was introduced independently by Small (1973, 1978) and Marshakova (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

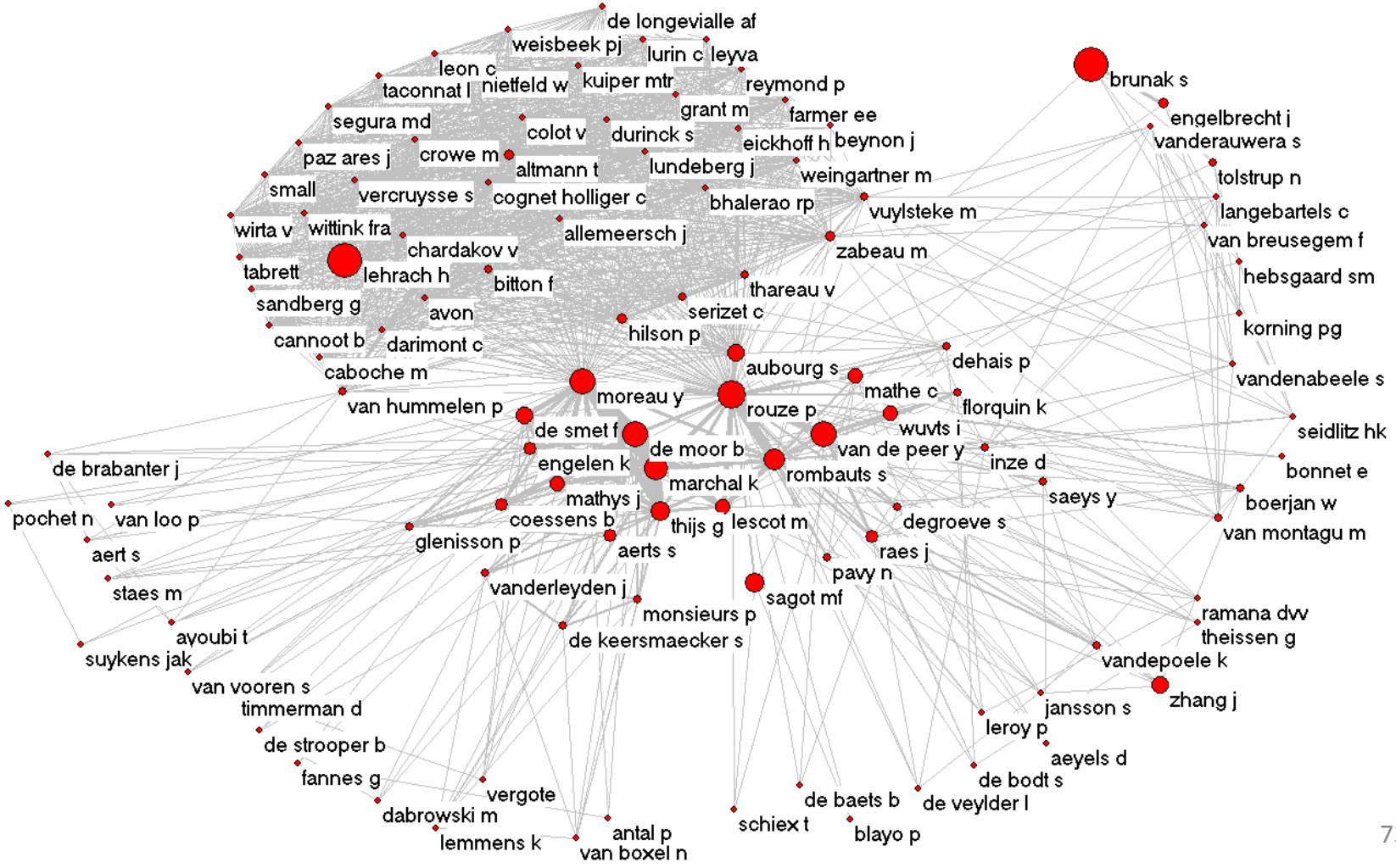


Journal Clustering

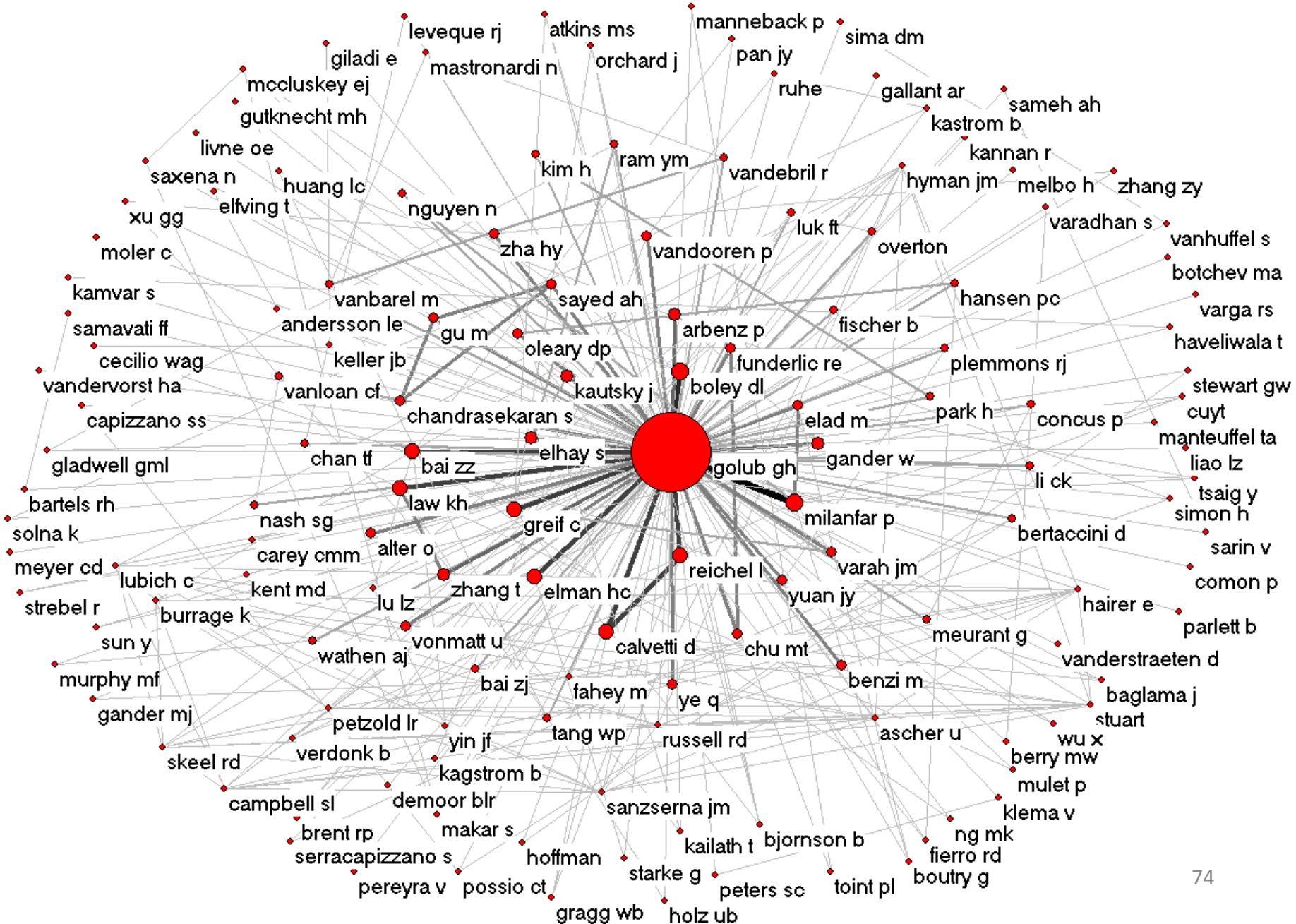


Author Collaboration Clustering

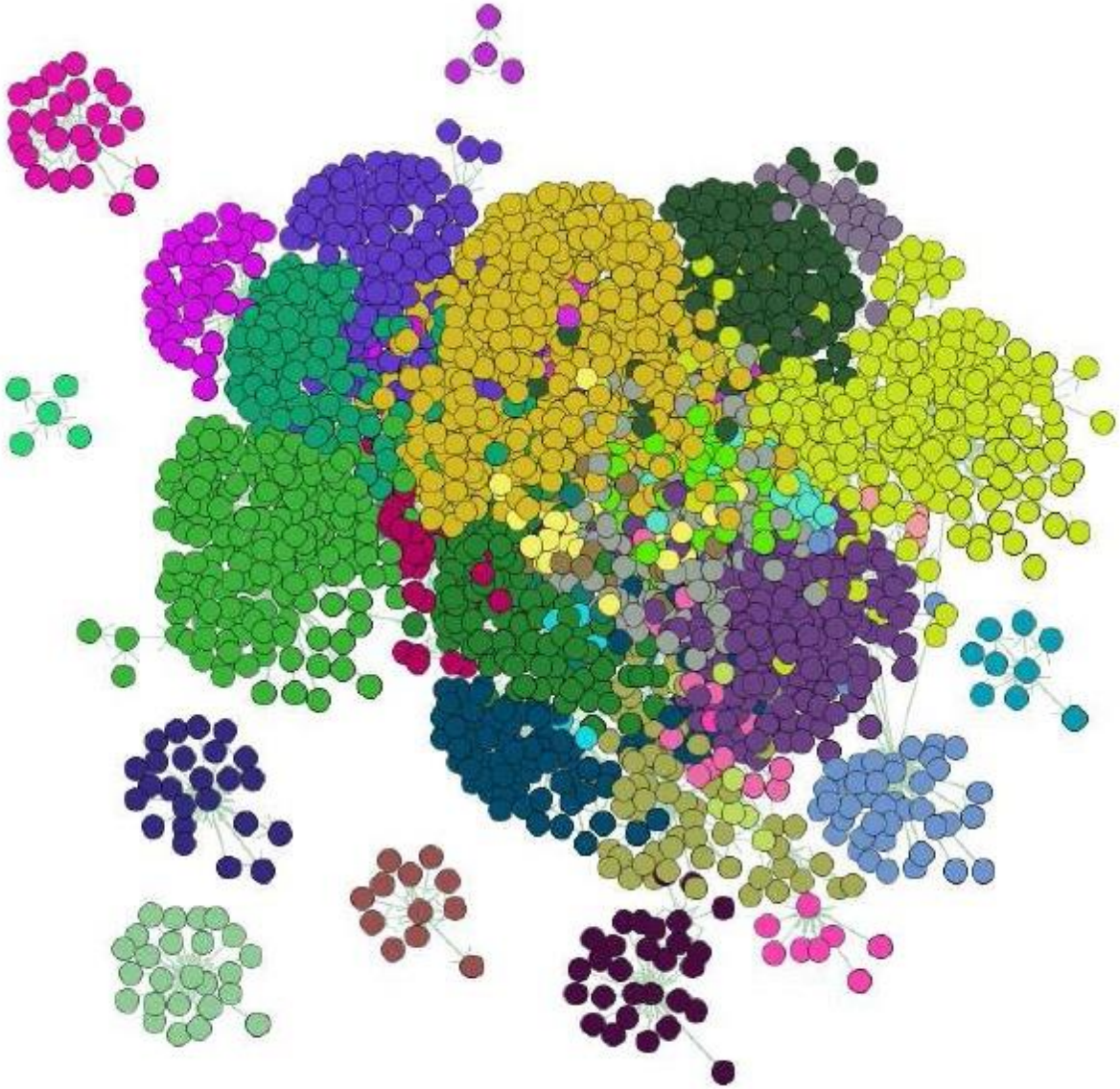
Author Collaboration Clustering



Golub around the world commemoration February 29 2008



Web of Science based literature network for Lennart Ljung



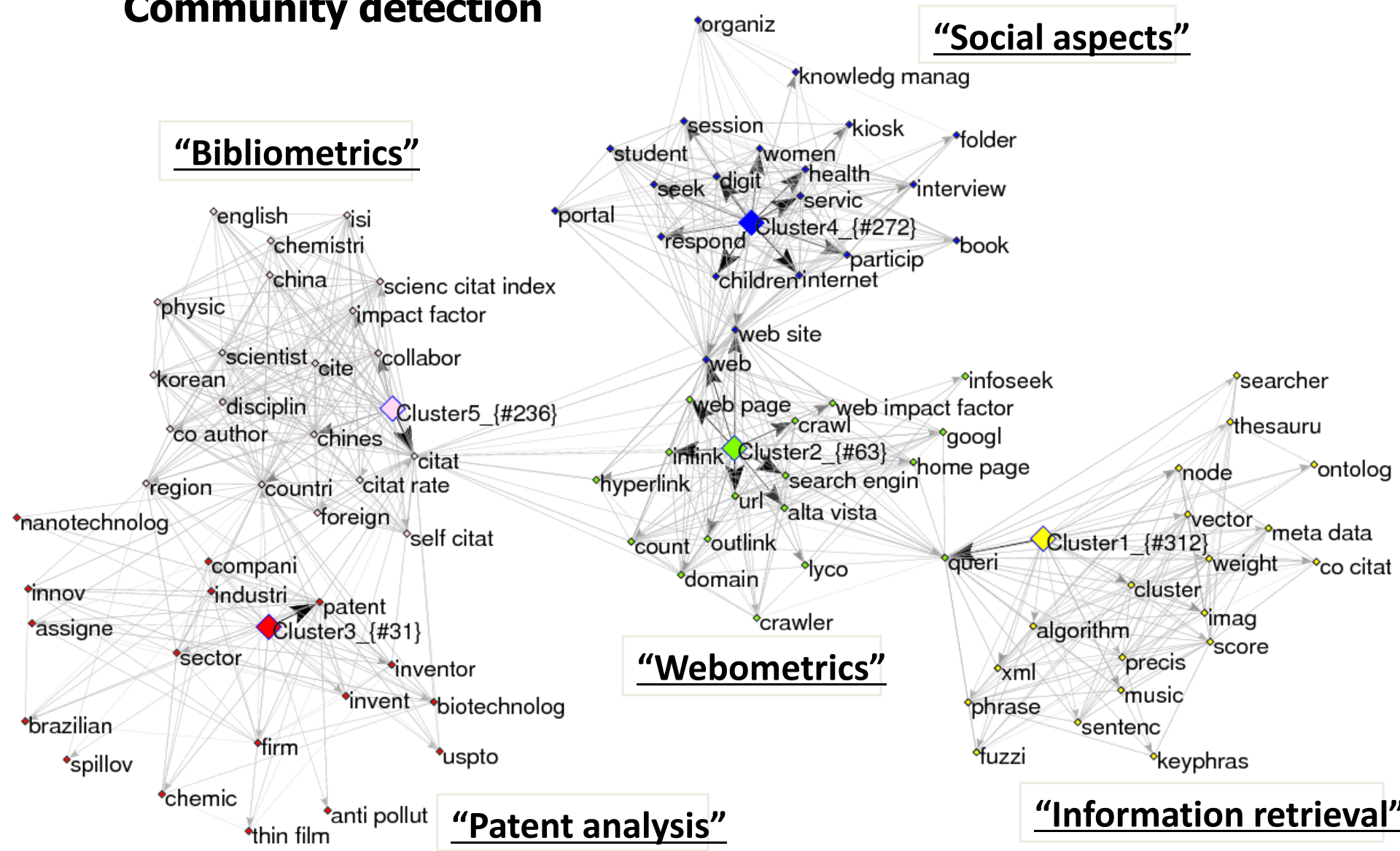
138 seed papers
+ all cited and citing publications

Result: 4943 nodes, 6216 edges

Link based clustering identifies topically homogeneous clusters.

13 papers are written by another L. Ljung.

Community detection



Energy

Industry

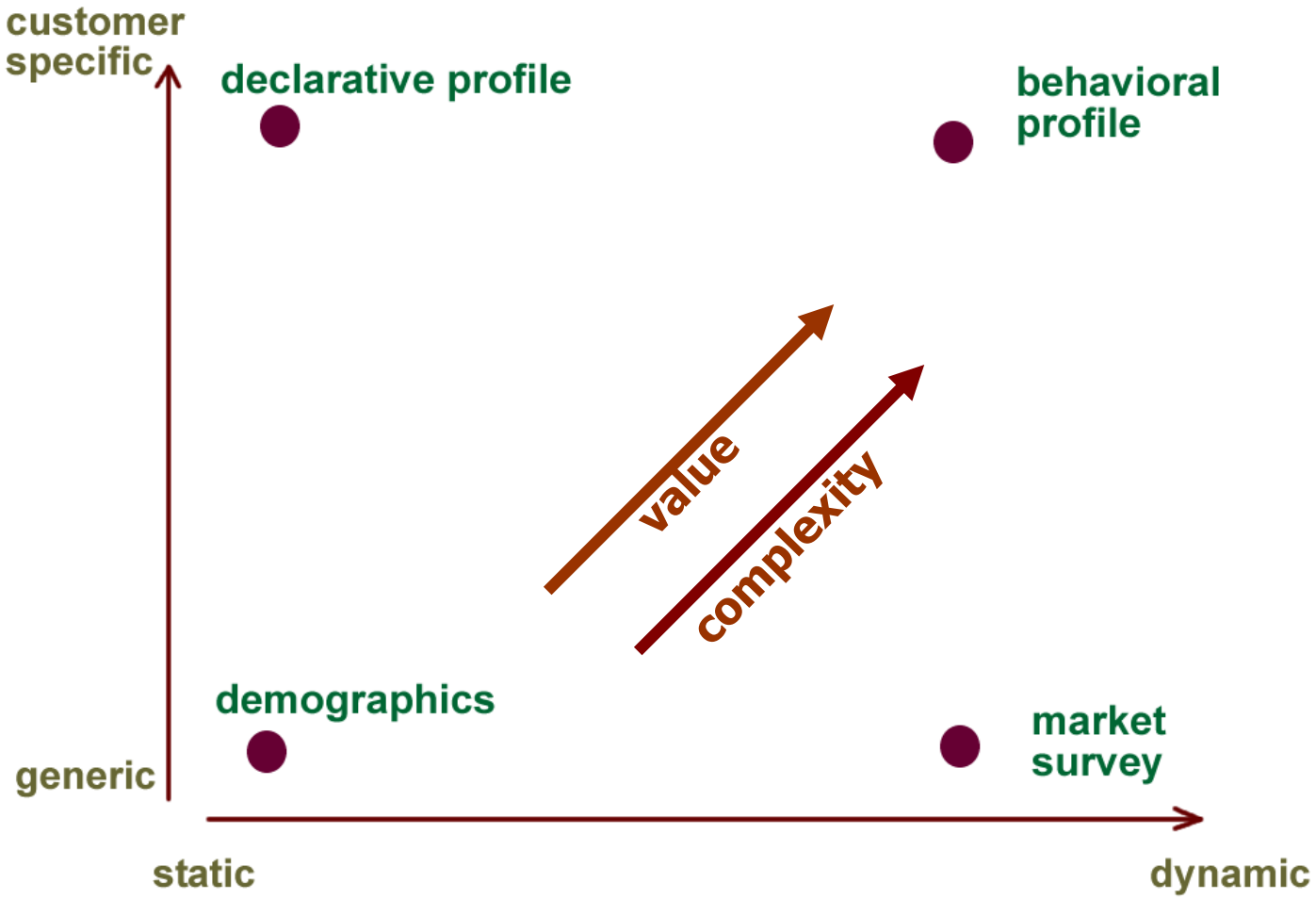
Environment

Social networks

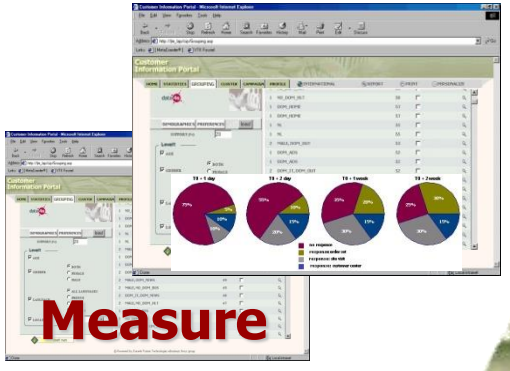
Fraud and predictive analysis

Health

Customer Intelligence

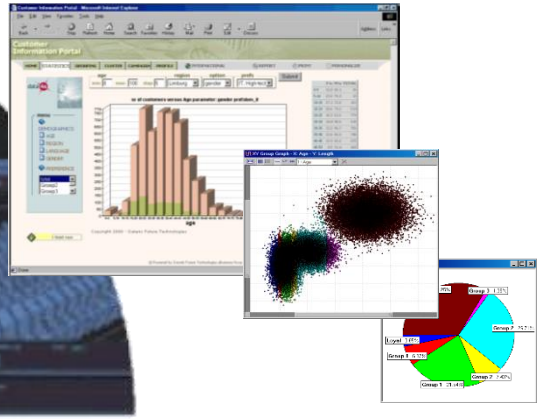


Customer Intelligence

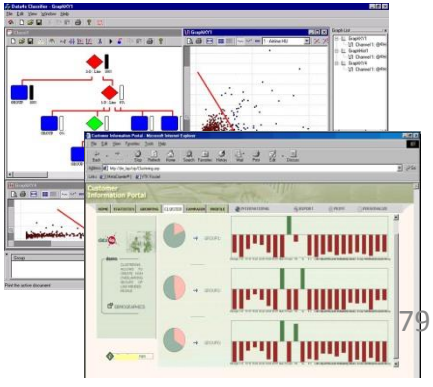


Measure response

Analyze & Model

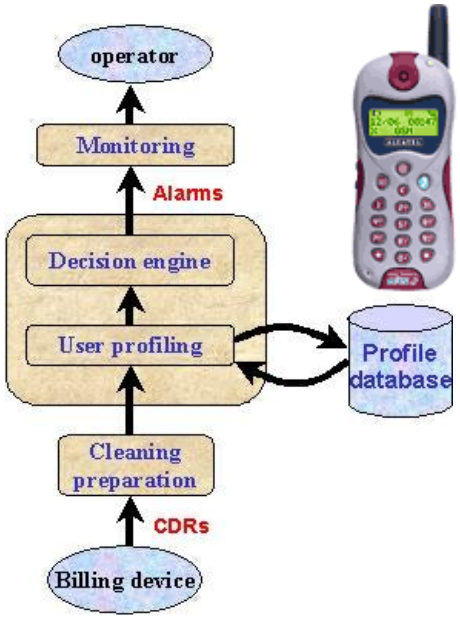


Score & Target

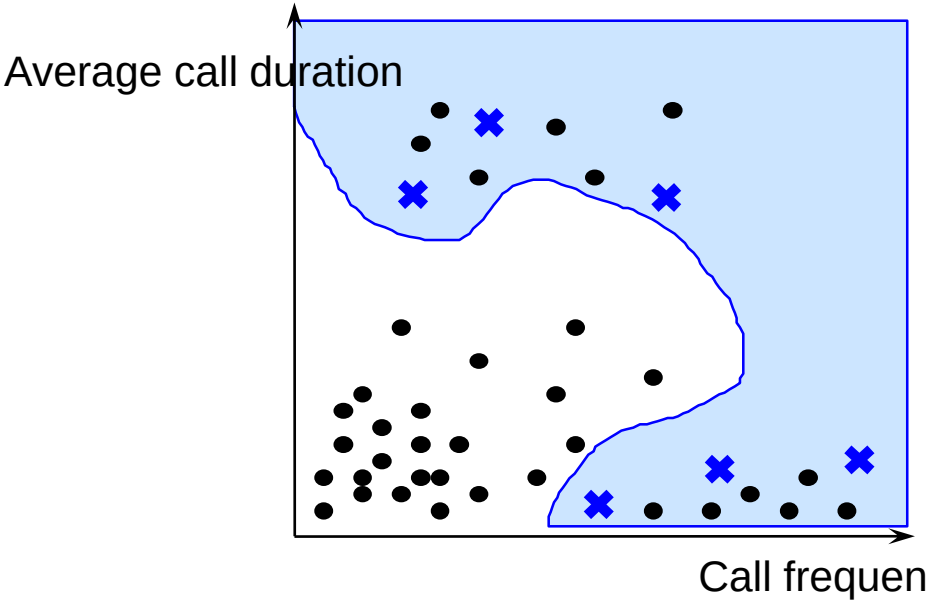


Contact

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



Energy

Industry

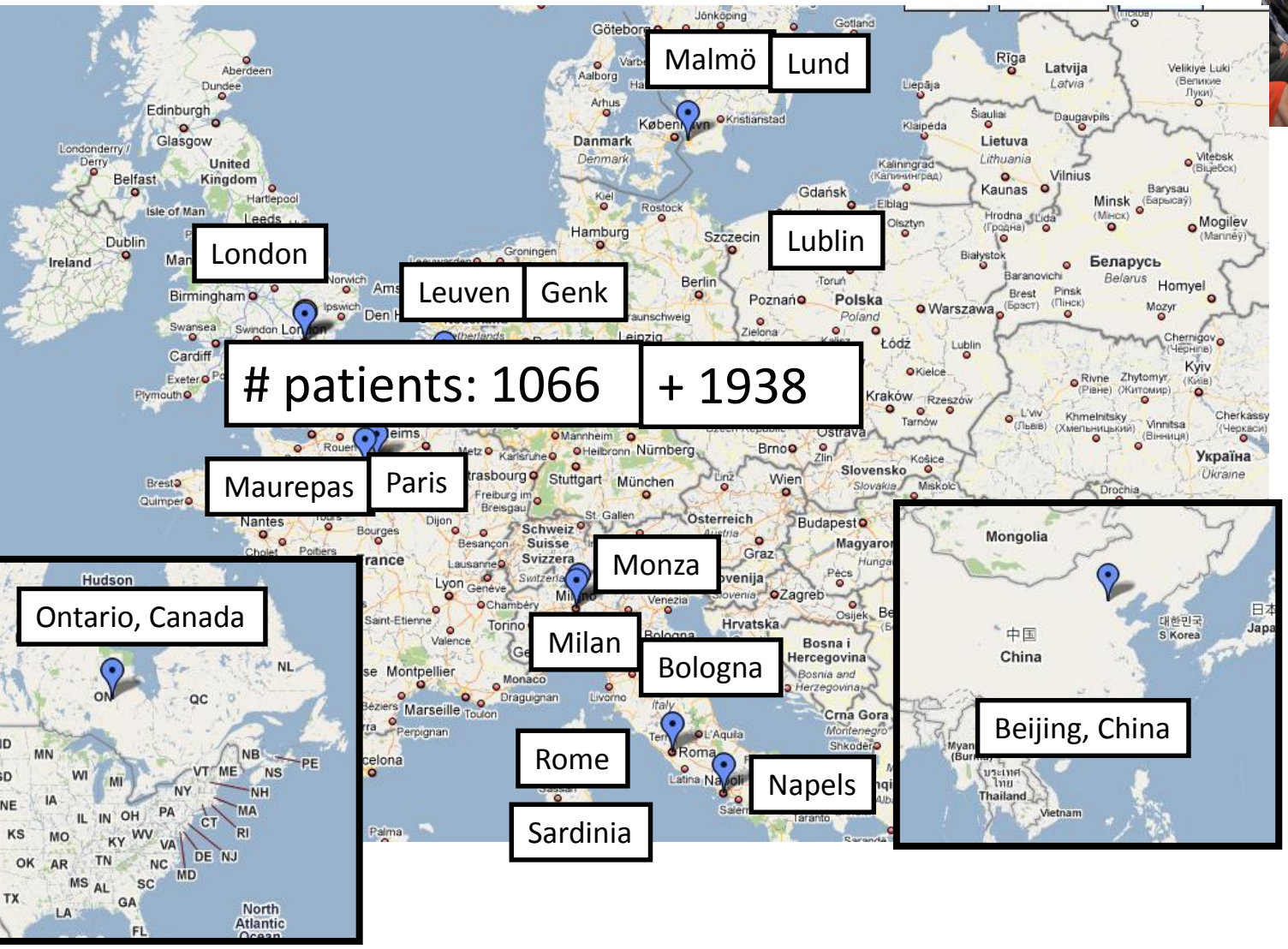
Environment

Social networks

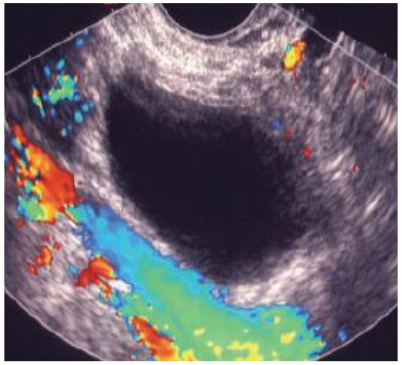
Fraud and predictive analysis

Health

Participatory



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



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

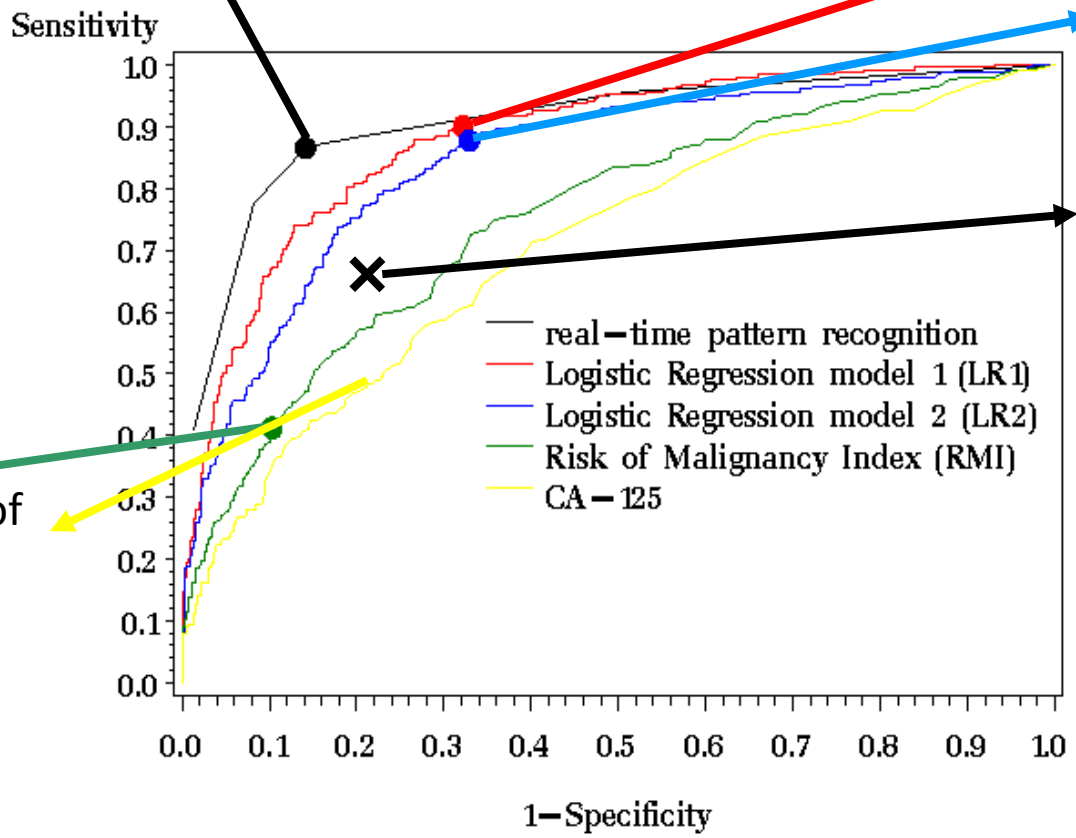
Performance

Performance of an expert

Performance of IOTA models

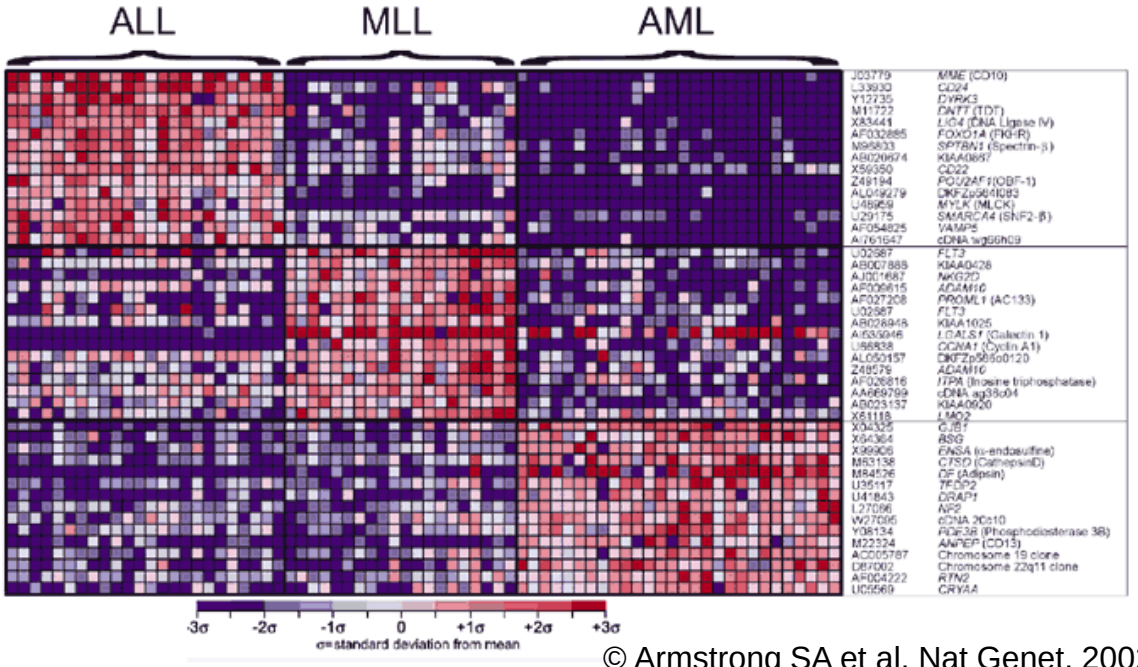
Performance of non-experts

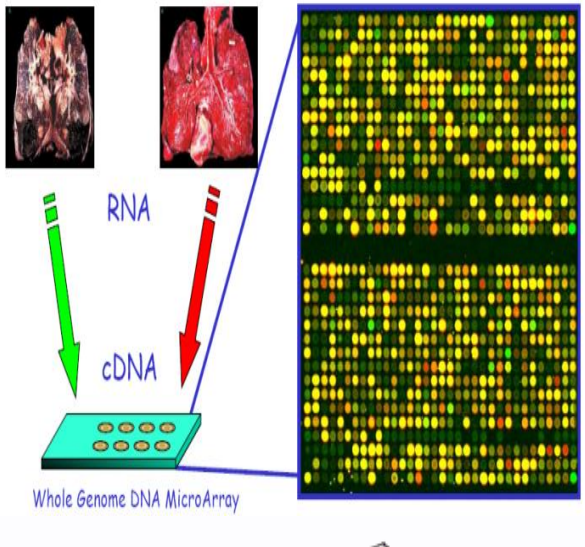
Performance of old models



You share, we care !

Genomic markers for Leukemia

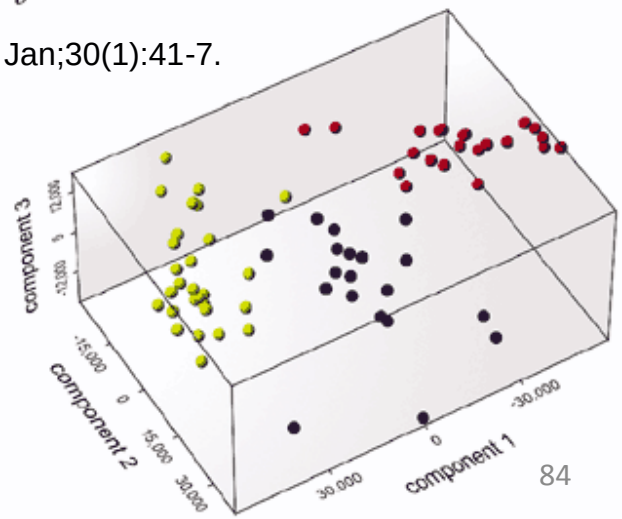




12 600 genes

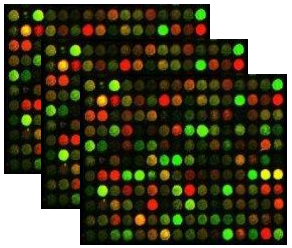
72 patients

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

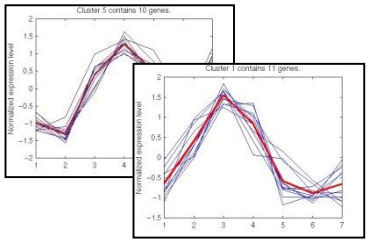


Genomic Data Fusion

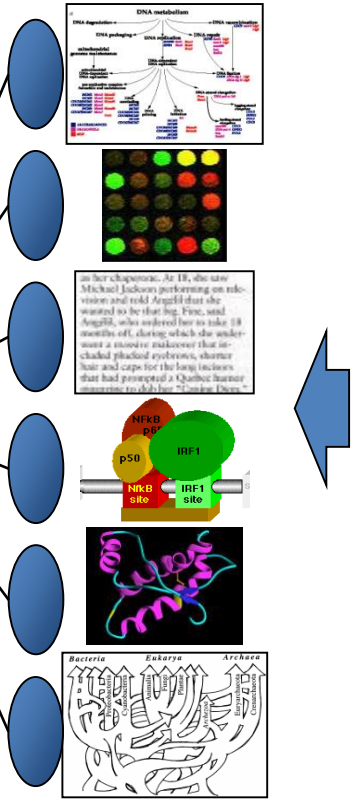
High-throughput genomics



Data analysis



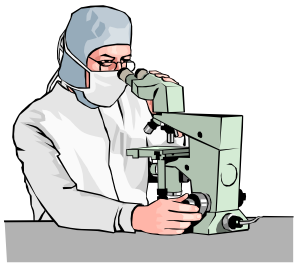
Information sources



Candidate 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



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

Endeavour: Aerts et al., Nature Biotechnology, 2006

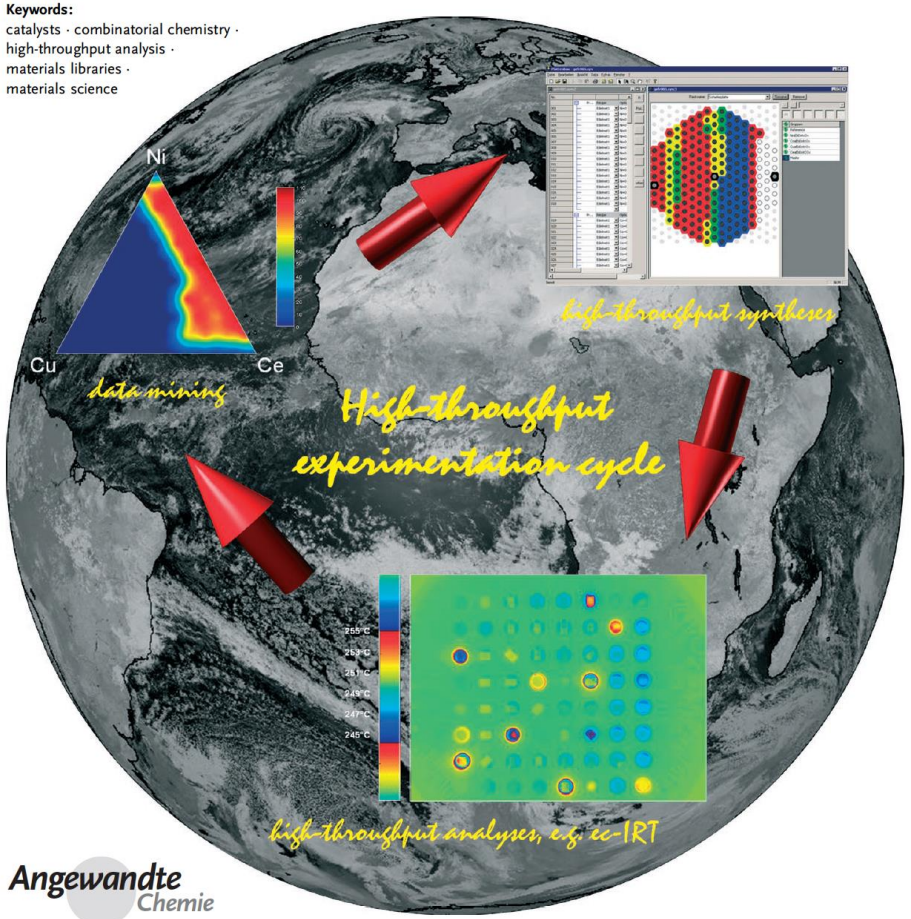
Materials Research

DOI: 10.1002/anie.200603675

Combinatorial and High-Throughput Materials Science

Wilhelm F. Maier, Klaus Stöwe, and Simone Sieg*

Keywords:
catalysts · combinatorial chemistry · high-throughput analysis · materials libraries · materials science



Electrochimica Acta 56 (2011) 9679–9699

Contents lists available at SciVerse ScienceDirect

Electrochimica Acta

journal homepage: www.elsevier.com/locate/electacta

Review

A review of high throughput and combinatorial electrochemistry

T.H. Muster[a,*,1], A. Trinchi[a], T.A. Markley[a], D. Lau[a], P. Martin[b], A. Bradbury[a], A. Bendavid[b], S. Dligatch[b]

[a] CSIRO Division of Materials Science and Engineering, Private Bag 33, Clayton South, Victoria 3169, Australia
[b] CSIRO Division of Materials Science and Engineering, PO Box 218, Lindfield, NSW, 2070, Australia

ARTICLE INFO

Article history:
Received 20 June 2011
Accepted 2 September 2011
Available online 14 September 2011

Keywords:
Electrochemistry
Combinatorial
High-throughput
Multielectrode
Review

ABSTRACT

Many 21st century technological solutions are reliant on the development of new materials with improved properties, and increasingly on materials that can be optimised to perform more than one function. High-throughput and combinatorial methodologies are being used more frequently to discover and design improved materials in a time efficient manner for a variety of applications. A number of technological challenges involve the field of electrochemistry, such as battery development, electrocatalysis, photocatalysis, corrosion protection, sensor development, photovoltaics and light-emitting materials. This review focuses on the utilisation of high-throughput and combinatorial methods that have incorporated, or are associated with, electrochemical methods. In many cases electrochemical determinations are well-suited for high-throughput methodologies, enabling direct quantitative analysis of properties. However, in other circumstances electrochemical measurements are complicated by additional factors. Hence the limitations of high-throughput and combinatorial electrochemistry are also discussed within.

Top Curr Chem (2014) 345: 139–180
DOI: 10.1007/128_2013_486

Published online: 28 November 2013

Data Mining Approaches to High-Throughput Crystal Structure and Compound Prediction

Geoffroy Hautier

Abstract Predicting unknown inorganic compounds and their crystal structure is a critical step of high-throughput computational materials design and discovery. One way to achieve efficient compound prediction is to use data mining or machine learning methods. In this chapter we present a few algorithms for data mining compound prediction and their applications to different materials discovery problems. In particular, the patterns or correlations governing phase stability for experimental or computational inorganic compound databases are statistically learned and used to build probabilistic or regression models to identify novel compounds and their crystal structures. The stability of those compound candidates is then assessed using ab initio techniques. Finally, we report a few cases where data mining driven computational predictions were experimentally confirmed through inorganic synthesis.

Keywords Ab initio computations · Crystal structure prediction · Data mining · High-throughput computing