Big Data Mining

What ? Why ? How ? Where ? Who ?

Prof. Dr. Bart De Moor

March 2018

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

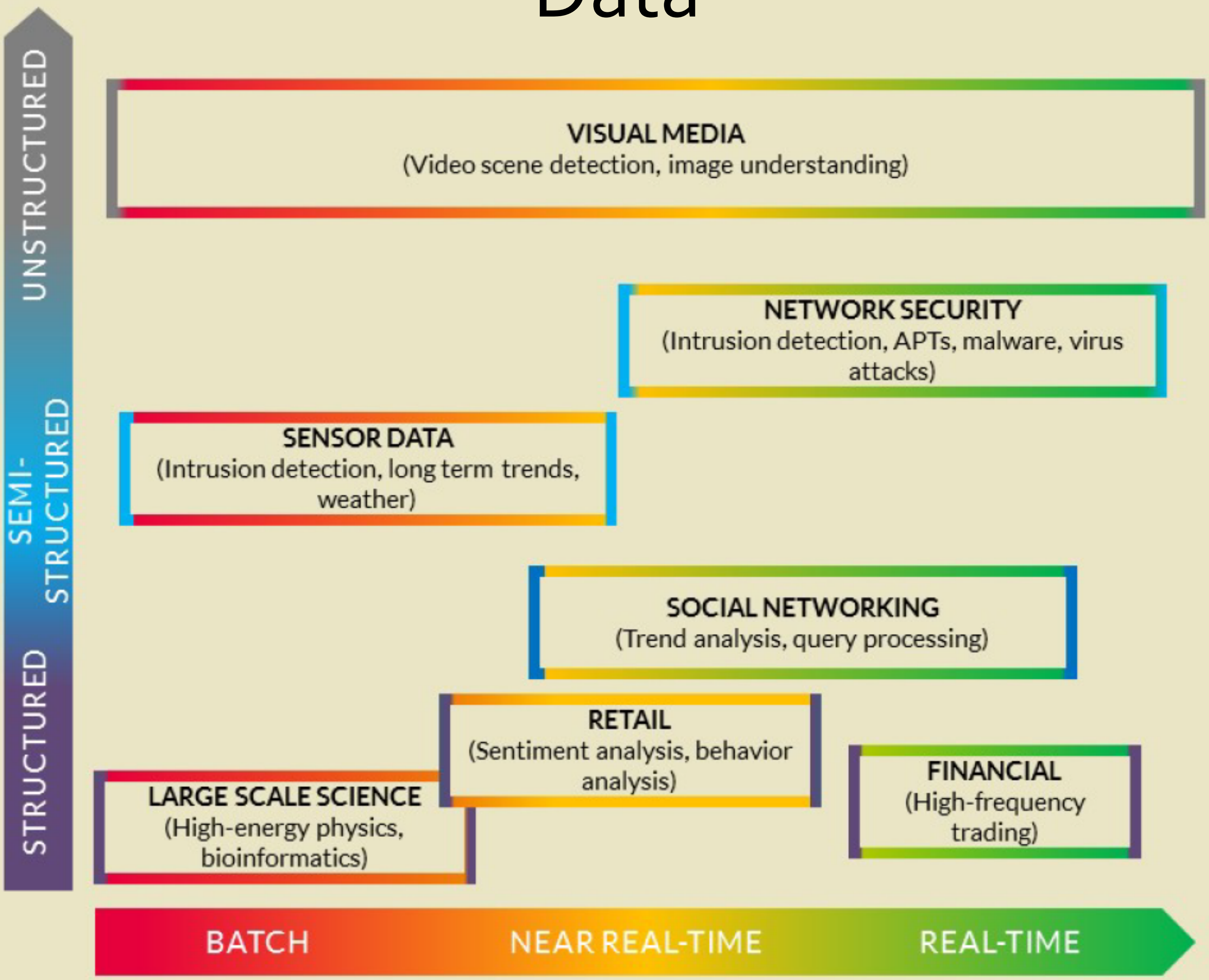
I have data

Which hypotheses can I check?

Data first!


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

Big Data Mining

What ? Why ? How ? Where ? Who ?

Prof. Dr. Bart De Moor

March 2018

Main tasks

Prediction

Regression

Segmentation

Clustering

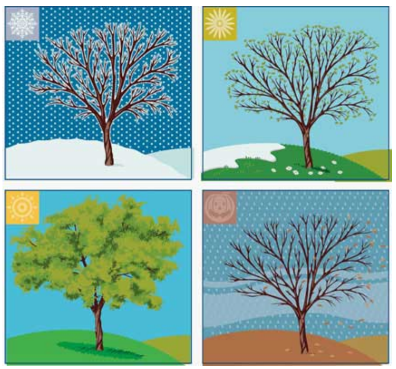
Classification

Anomalies

Detect outliers

Main tasks

Filtering effects



⇩

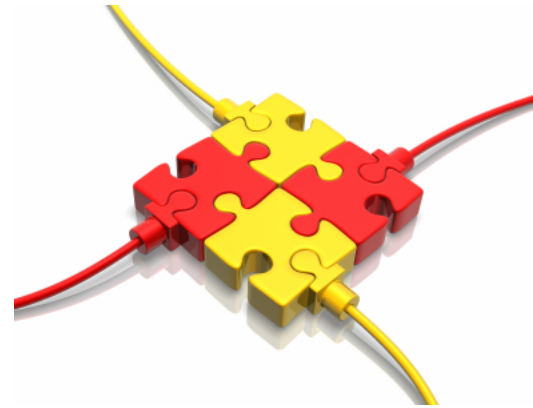
Normalization

Assess relevance

⇩

Ranking

Combining info



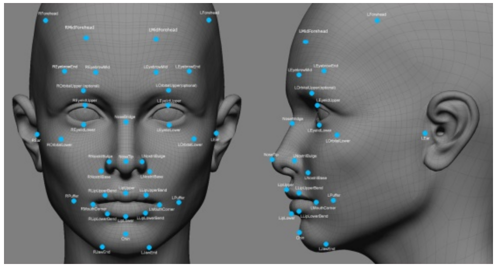
⇩

Data fusion

Objectives - ICT

Communication networks

Facial recognition



Home automation

Digital signing

Data center optimization

Objectives - Finance

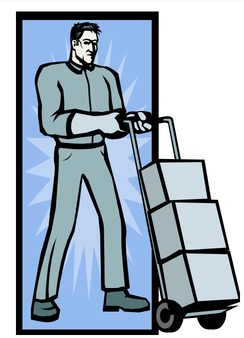
Fraud detection

Credit worthiness

Risk assessment

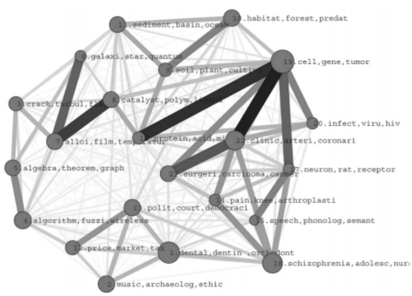
Portfolio management

Just-in-time production



Objectives - Education

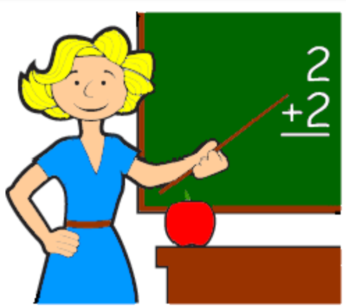
Scientometrics



Detecting plagiarism

Grading

Teacher performance



Student performance

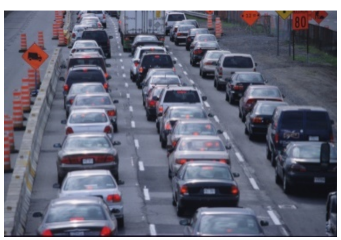
Objectives – Smart Cities

Predictive maintenance

Flood prediction

Smart lighting

Traffic management



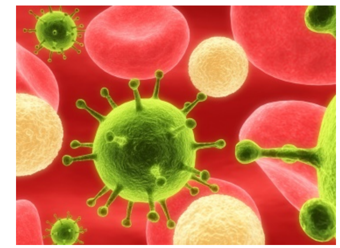
Electricity Demand

Objectives – Health

Diagnostics

Genome sequencing

Disease spreading



Tumour detection

Medical fraud detection

Big Data Mining

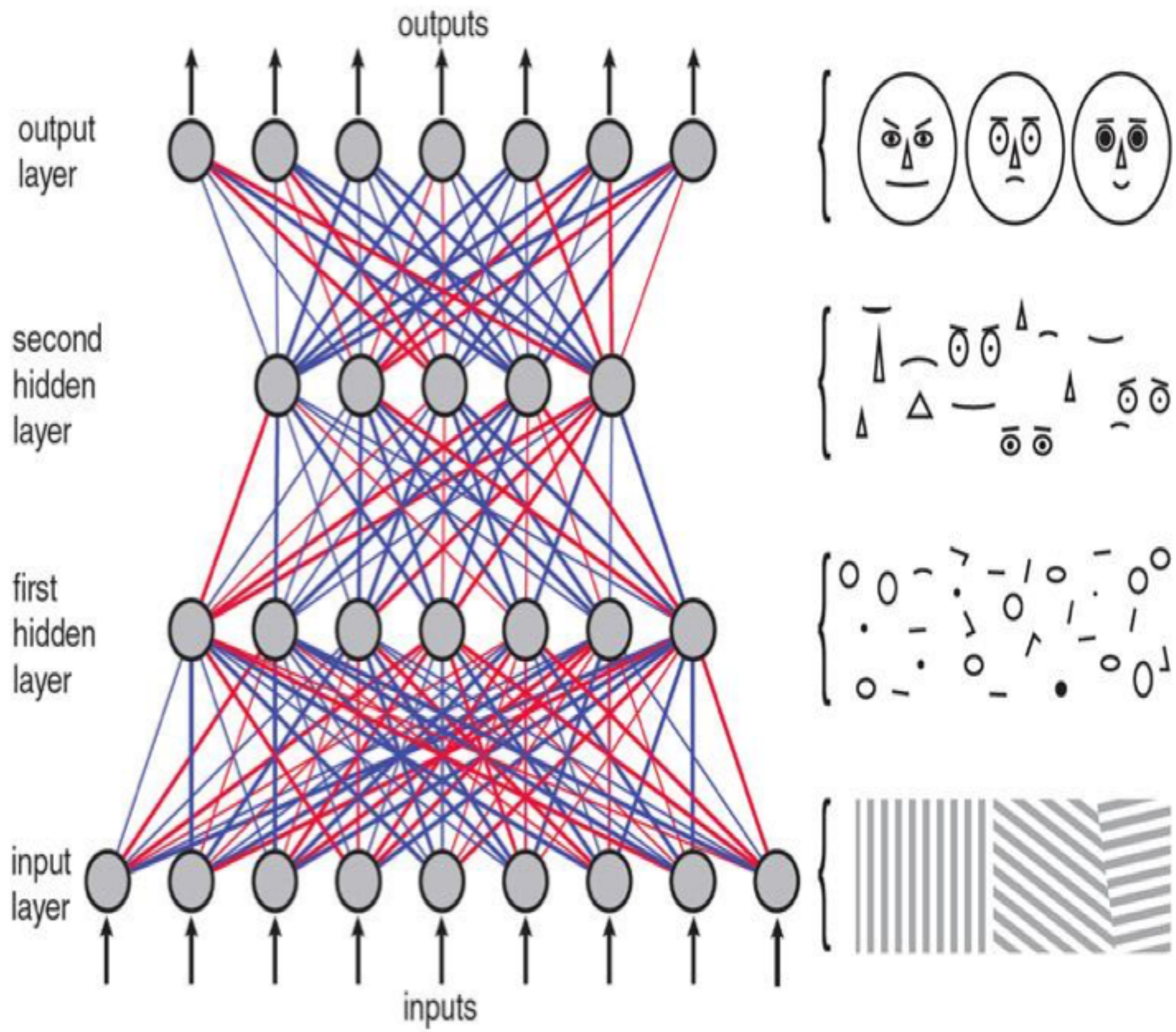
What ? Why ? How ? Where ? Who ?

Prof. Dr. Bart De Moor

March 2018

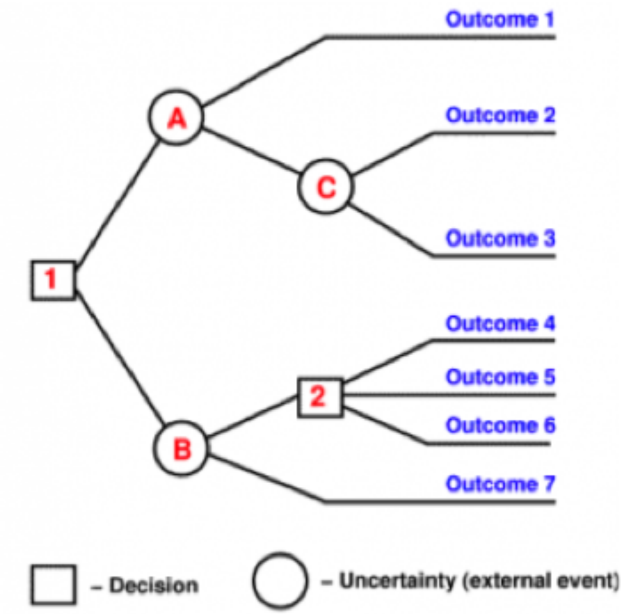
Deep Learning & Neural Networks

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



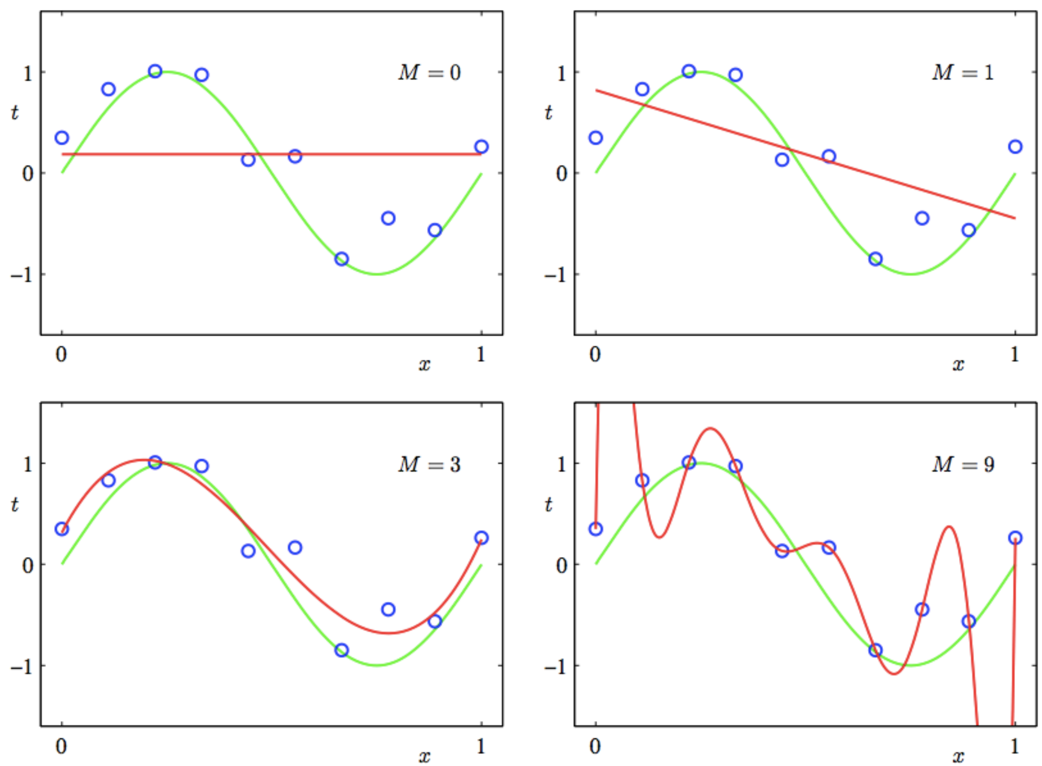
Decision trees

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.



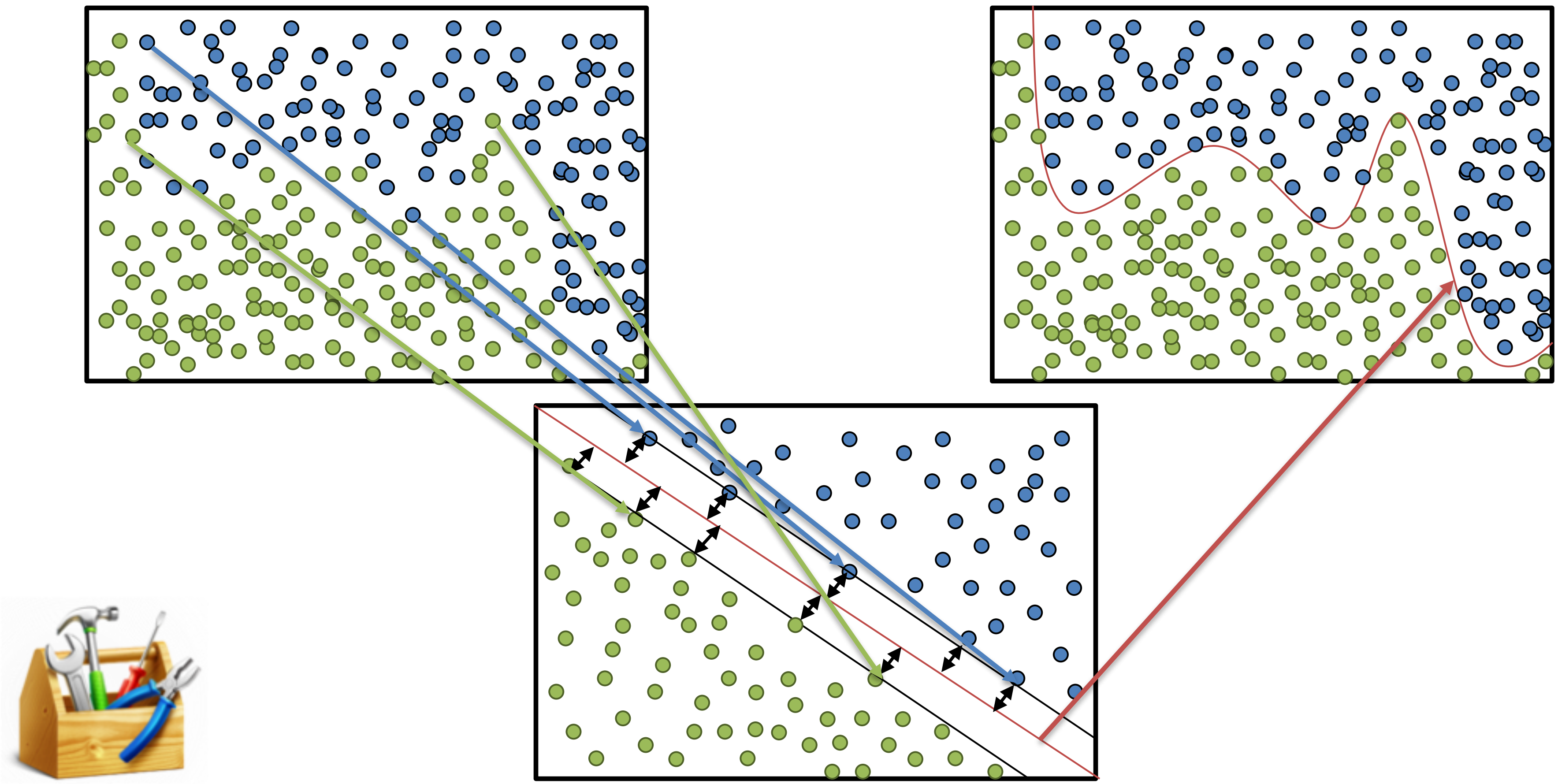
Regularized Regression

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.

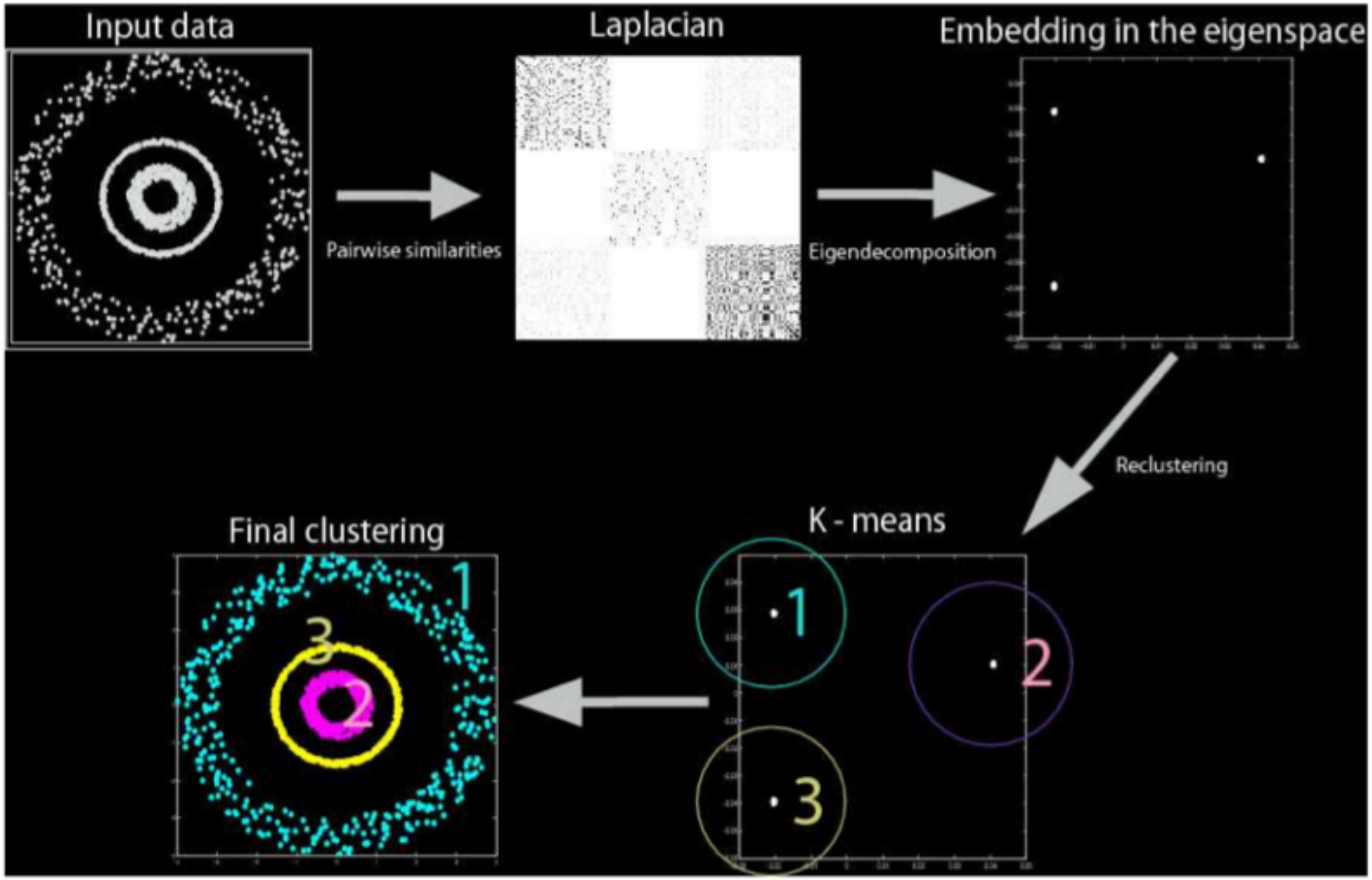


Support Vector Machine

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.

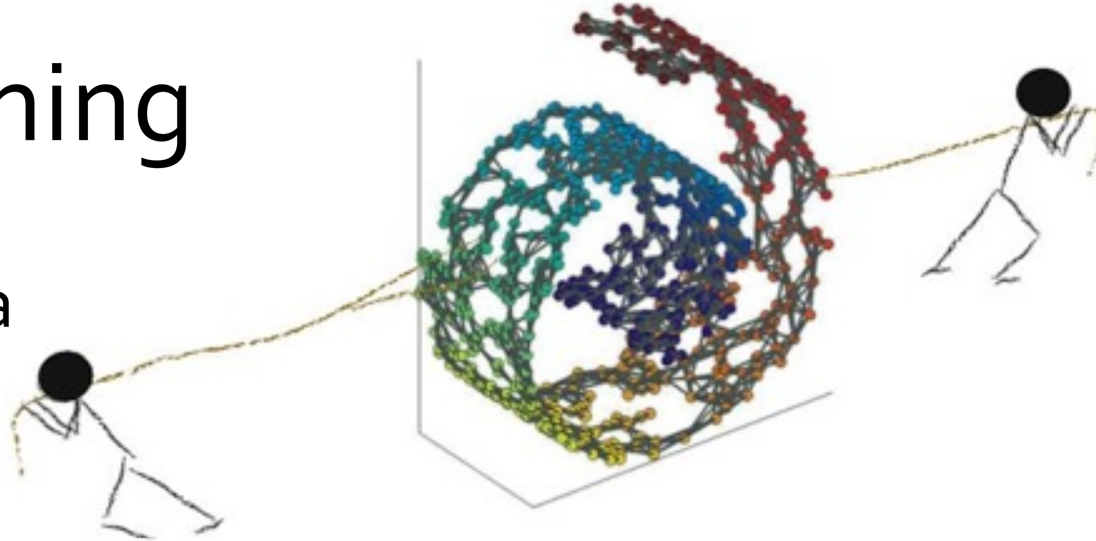


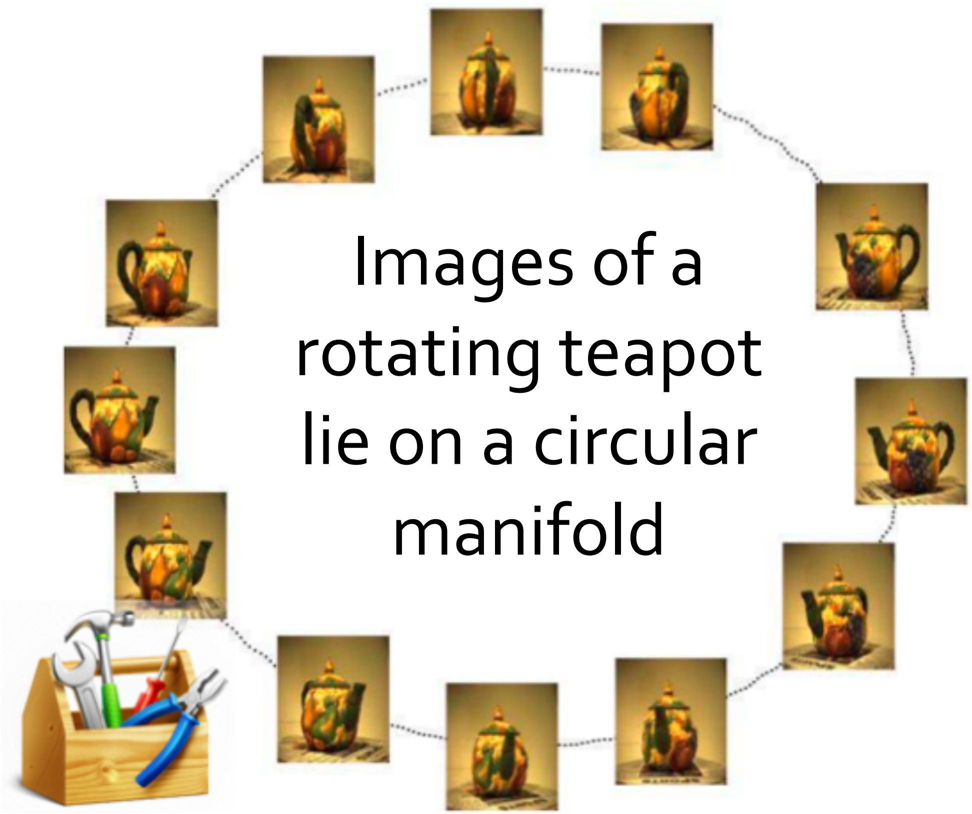
Spectral clustering



Manifold learning

A lot of datasets live on a low dimensional manifold.



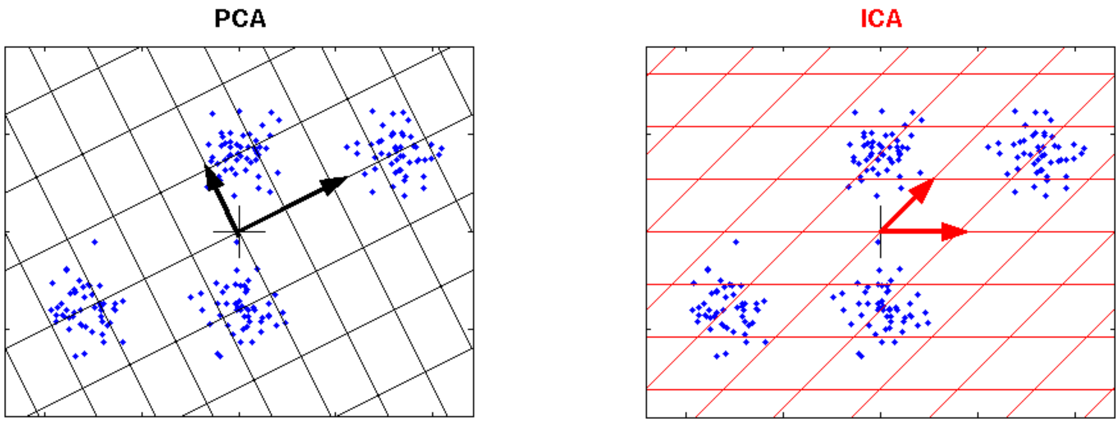


Images of a rotating teapot lie on a circular manifold

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

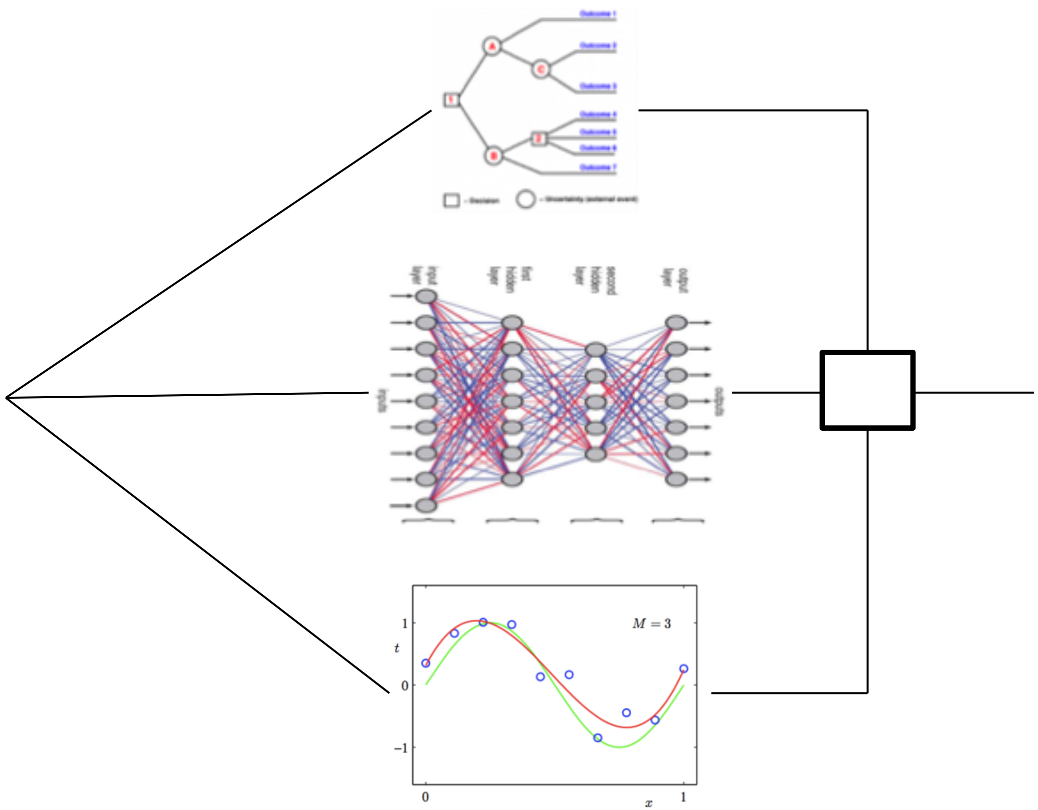
Component analysis

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).



Ensemble methods

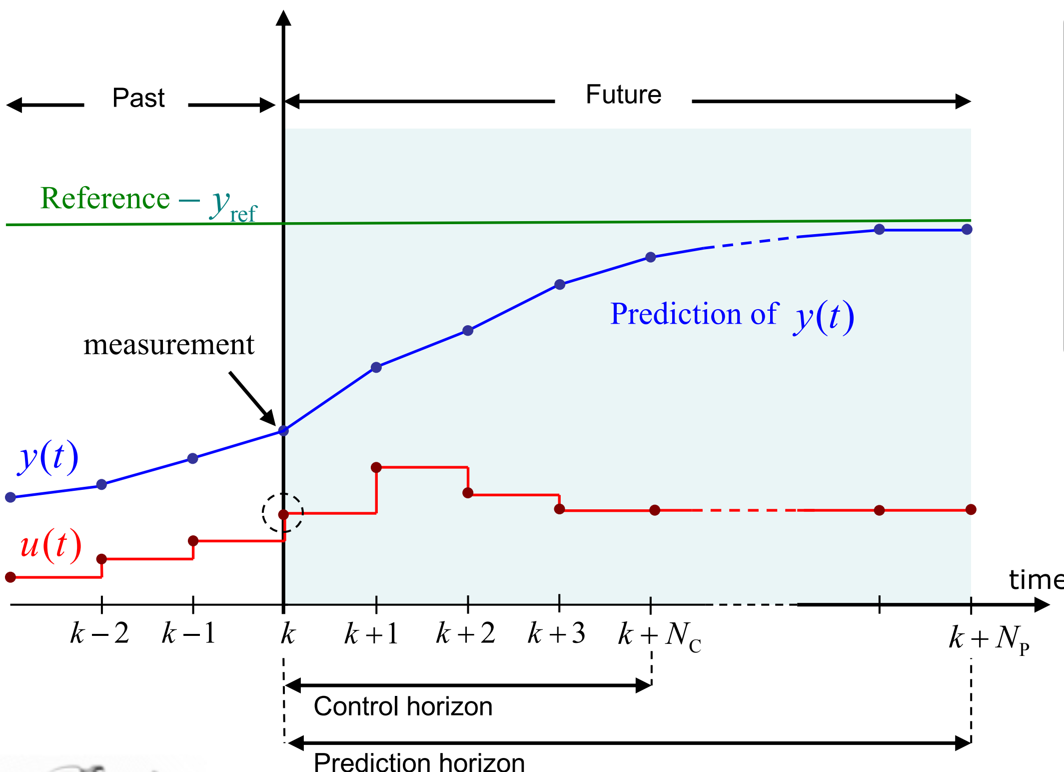
Several machine learning algorithms are implemented in parallel to each other. A decision on the outcome is then made, based on some decision rule (e.g., majority voting).



Model Predictive control (MPC)

Control method for handling input and state constraints within an optimal control setting.

Principle of predictive control



$$\min_{u(k),\ldots,u(k+N_{\text{c}}-1)} \sum_{i=1}^{N_{\text{p}}} \left(y_{\text{ref}} - y(k+i)\right)^2$$

subject to

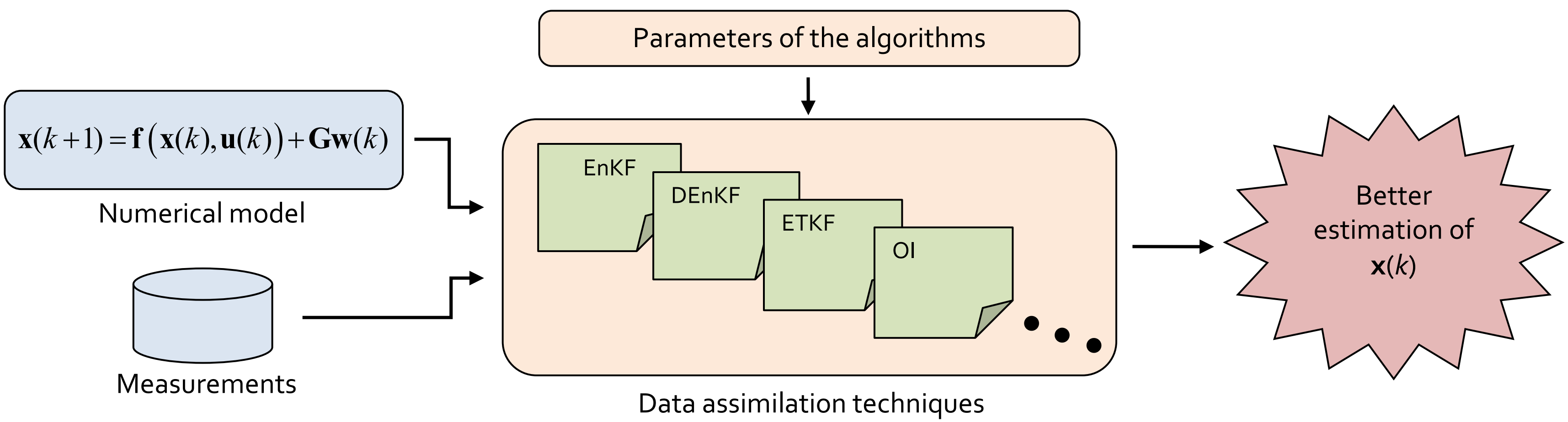
- model of the process
- input constraints
- output / state constraints

Why MPC ?

- It handles multivariable interactions
- It handles input and state constraints
- It can push the plants to their limits of performance.

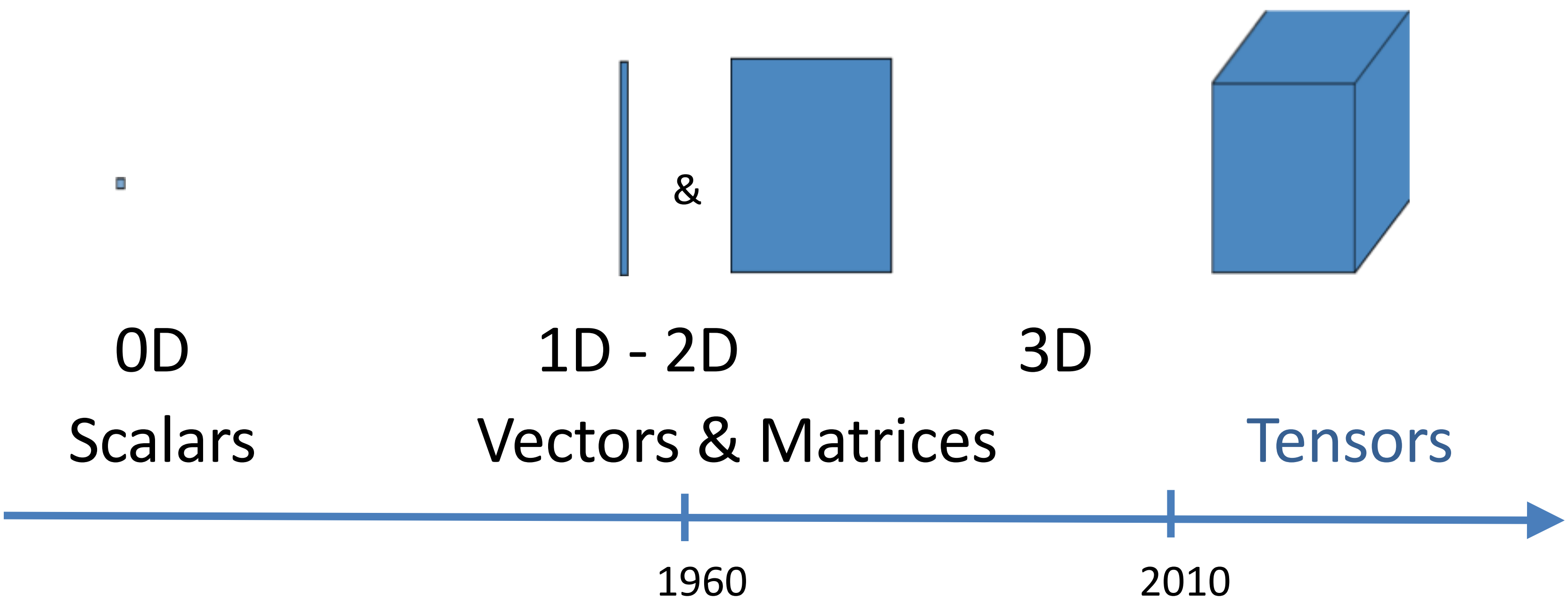
Data Assimilation

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.

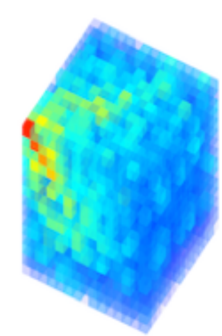


Some data assimilation techniques: 3DVAR, 4DVAR, Ensemble Kalman Filter (EnKF) and its variants, Optimal Interpolation (OI), particle filters, etc.

From matrices to tensors



- Exciting new possibilities in tensor framework
- Shift of paradigm



Big Data Mining

What ? Why ? How ? Where ? Who ?

Prof. Dr. Bart De Moor

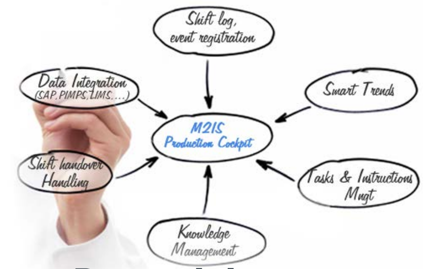
March 2018

STADIUS - SPIN-OFFS "Going beyond research"
www. esat.kuleuven.be/stadius/spinoffs.php

Transport & Mobility research

www.tmleuven.be

d square



Data mining industry solutions

www.dsquare.be

Data handling & mining for clinical genetics

www.cartagenia.com

Automated PCR analysis

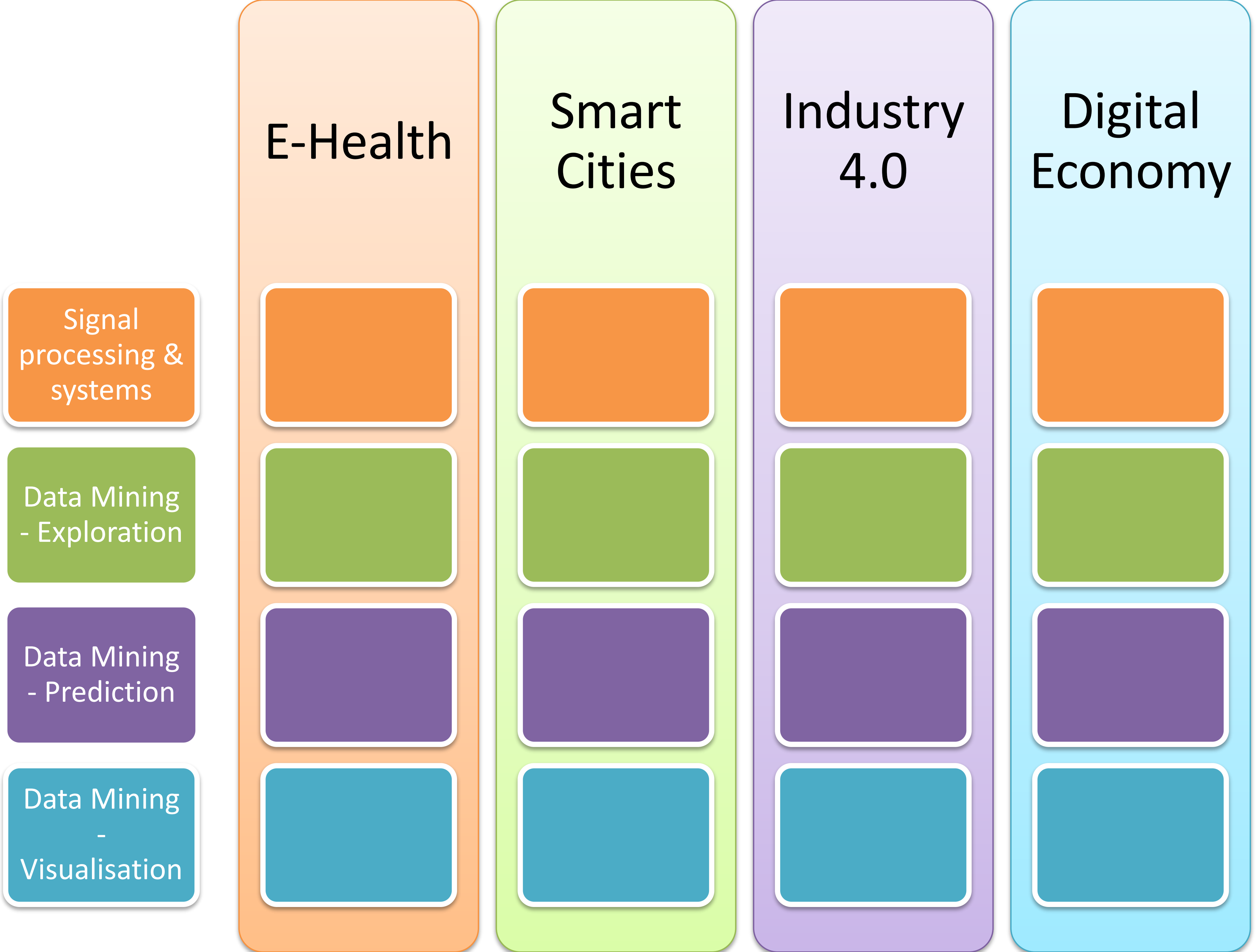
www.ugentec.com

Financial compliance

www.baesystems.com

Automation & Optimization

www.ipcos.be


E-Health
Smart Cities
Industry 4.0
Digital Economy
Signal processing & systems
Data Mining - Exploration
Data Mining - Prediction
Data Mining - Visualisation

Tsunami of medical data

sequencing all newborns by 2020 (125k births / year)

125 PetaByte / year

index of 20 million Biomedical PubMed records

23 GigaByte

raw NGS data of 1 full genome

1 TeraByte

PACS
UZ Leuven

1,6 PetaByte

Genomics core HiSeq 2000 full speed exome sequencing

1 TeraByte / week

1 CD-ROM

750 MegaByte

1 small animal image

1 GigaByte

1 slice mouse brain MSI at 10 µm resolution

81 GigaByte

Solid tradition of working with doctors

International Clinical Research Consortia e.g. IETA International Endometrial Tumor Analysis Group

Jessa Ziekenhuis

UMC
Universitair Medisch Centrum Utrecht

Erasmus MC
Universitair Medisch Centrum Rotterdam

academisch ziekenhuis Maastricht
azM

Intensive Care Unit

Radiology

Urology

Cardiology

Pneumology

Orthopaedics

Neonatology

Dentistry

Forensics

Rehabilitation

Gynaecology

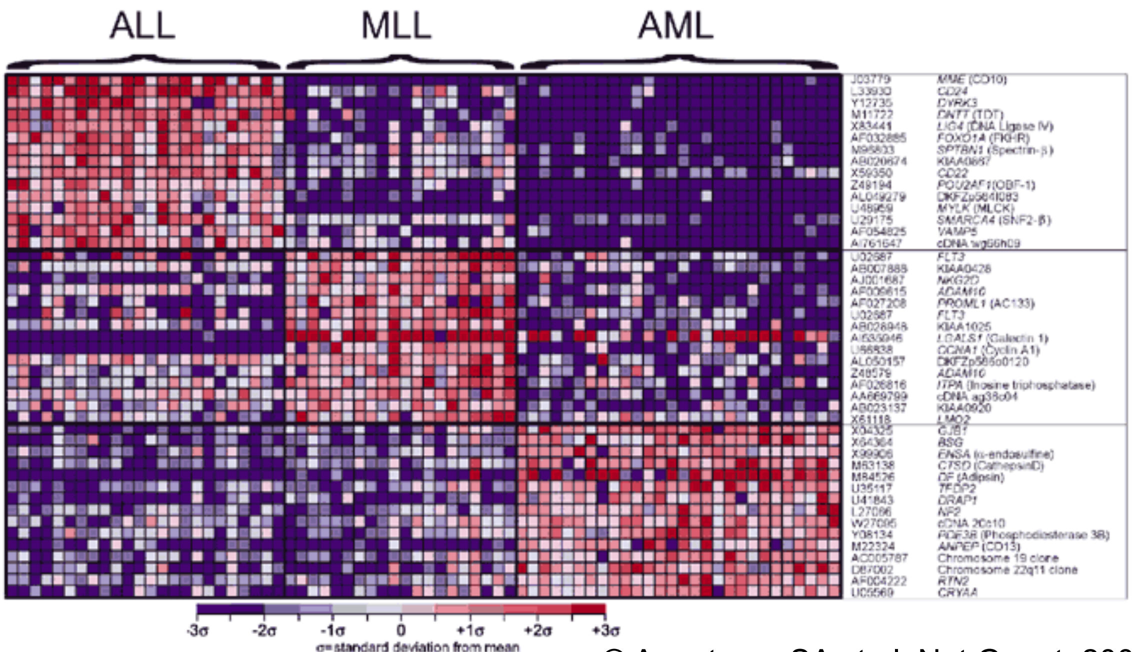
Radiotherapy

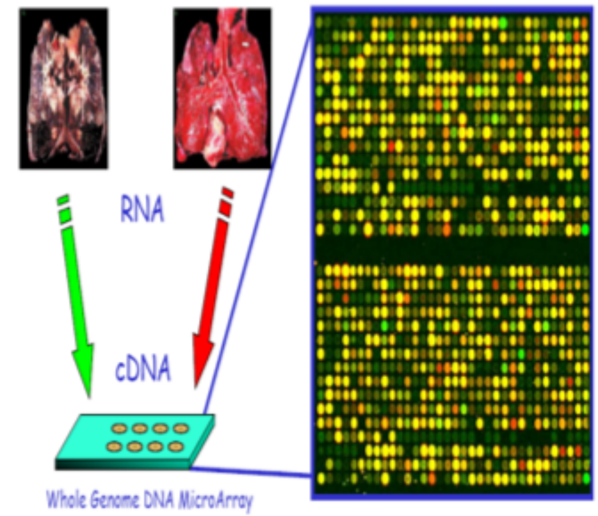
Neurology

Human Genetics

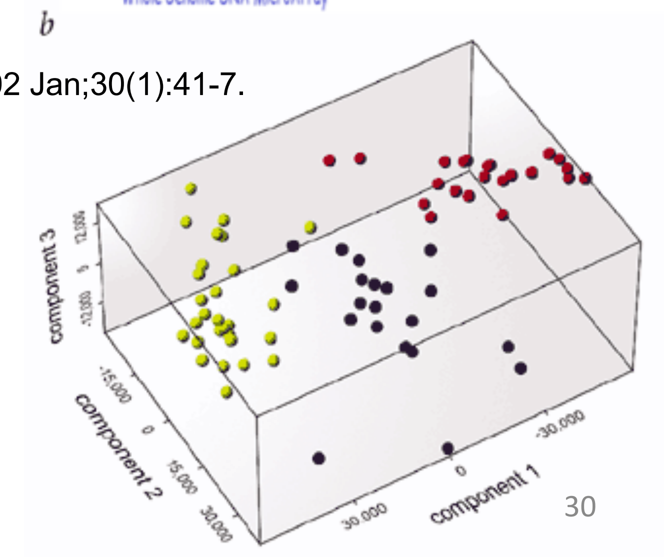
Oncology

Genomic markers for Leukemia





© Armstrong SA et al. Nat Genet. 2002 Jan;30(1):41-7.

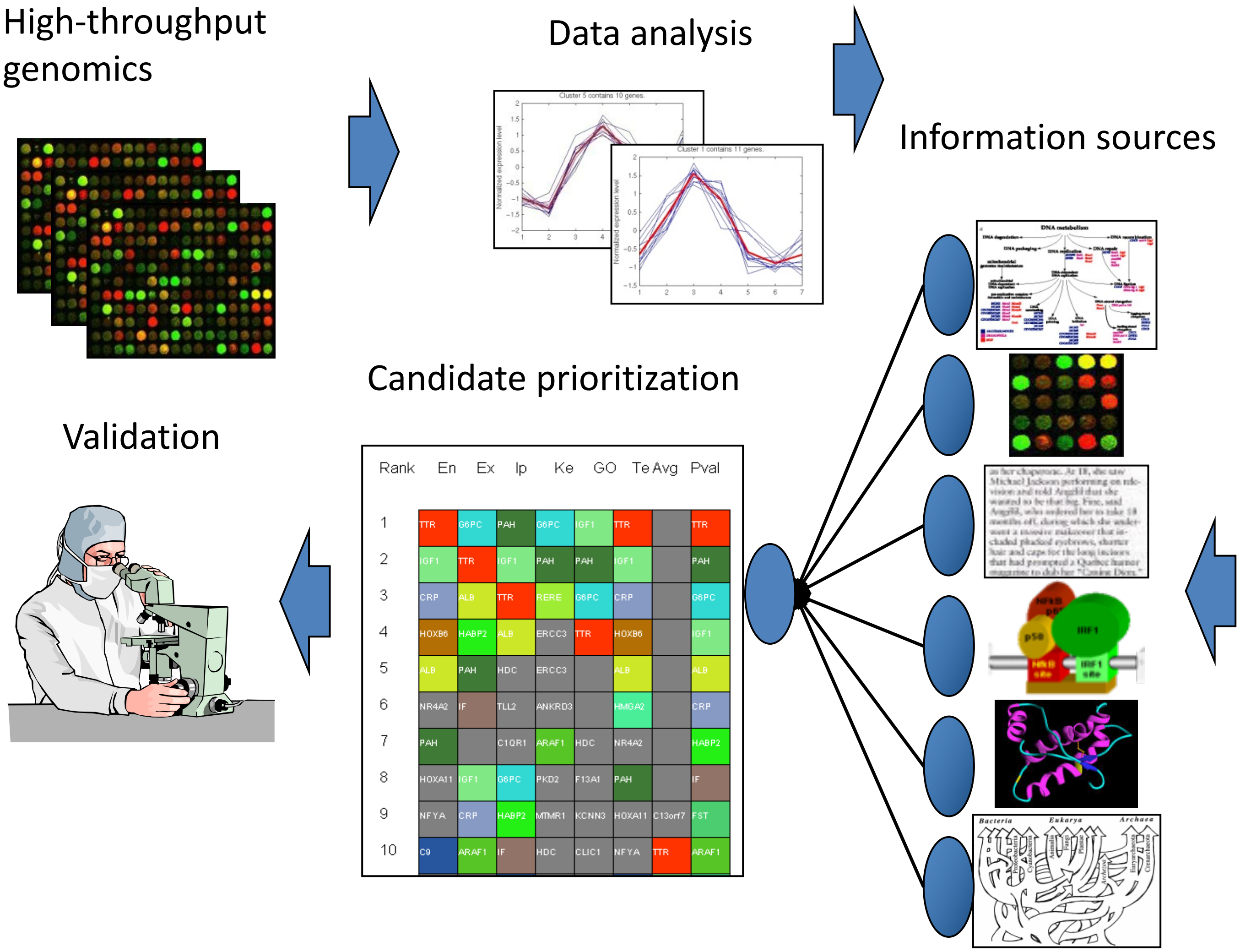


12 600 genes

72 patients

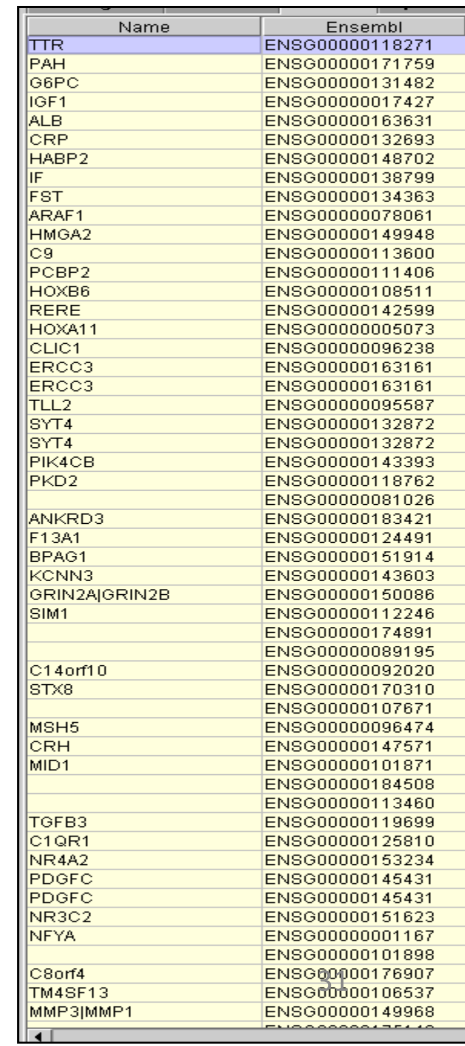
- 28 Acute Lymphoblastic Leukemia (ALL)
- 24 Acute Myeloid Leukemia (AML)
- 20 Mixed Linkage Leukemia (MLL)

Genomic Data Fusion



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

Candidate genes

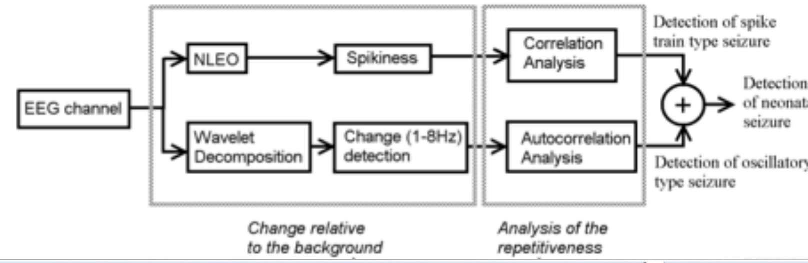


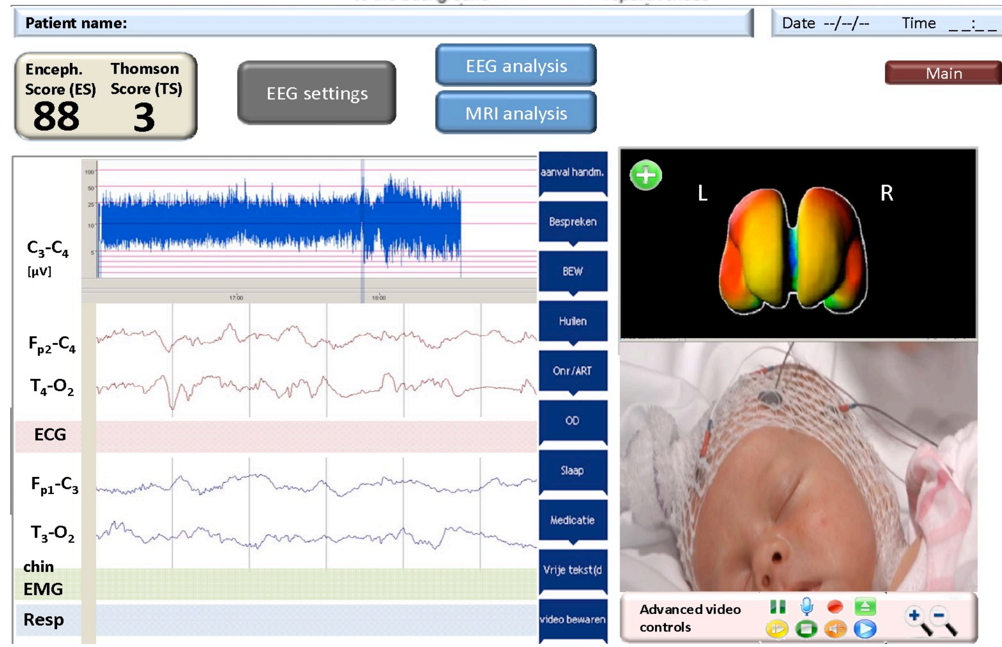
Name	Ensembl
TTR	ENSG00000118271
PAH	ENSG00000171759
G6PC	ENSG00000131482
IGF1	ENSG00000017427
ALB	ENSG00000163631
CRP	ENSG00000132693
HABP2	ENSG00000148702
IF	ENSG00000138799
FST	ENSG00000134363
ARAF1	ENSG00000078061
HMGA2	ENSG00000149948
C9	ENSG00000113600
PCBP2	ENSG00000111406
HOXB6	ENSG00000108511
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

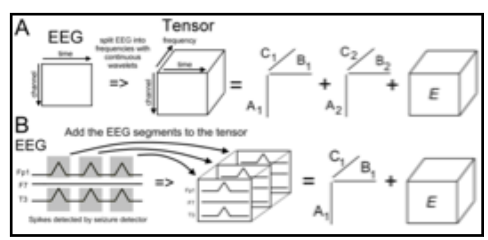
Endeavour: Aerts et al., Nature Biotechnology, 2006

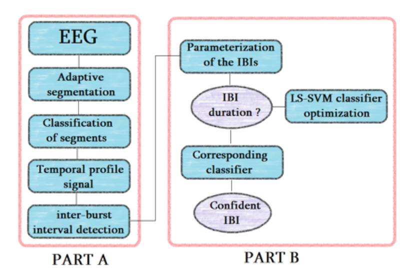
NeoGuard : decision support

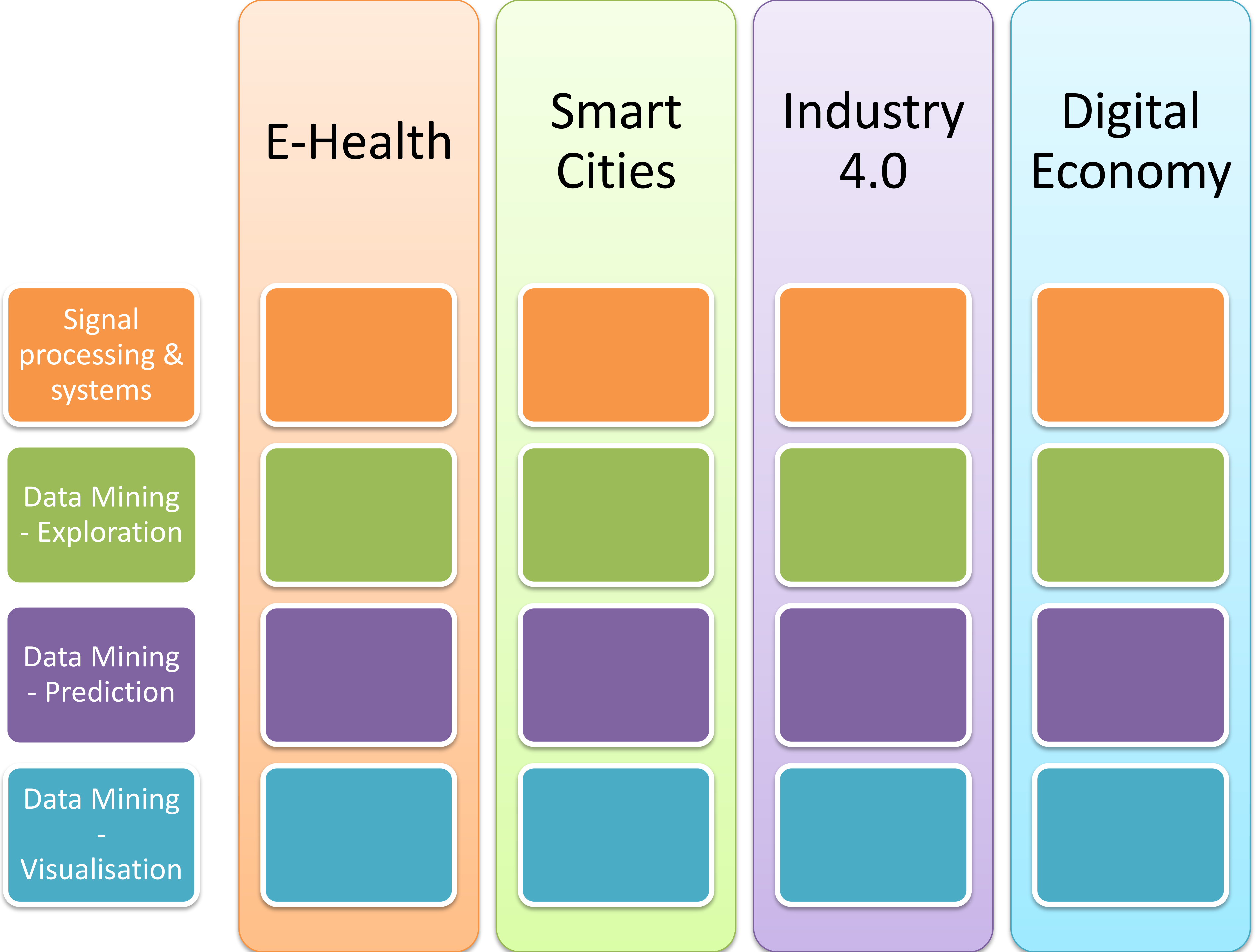
- Brain injury estimate
  - Detection of neonatal epileptic seizures
  - Seizures localization
  - Inter-burst intervals
- Incorporated expertise
  - Knowledge of neurophysiologists are incorporated into algorithms
- Monitoring
  - evolution rate of the background EEG
  - Maturity in premature
- Outcome prediction
  - Good
  - Poor







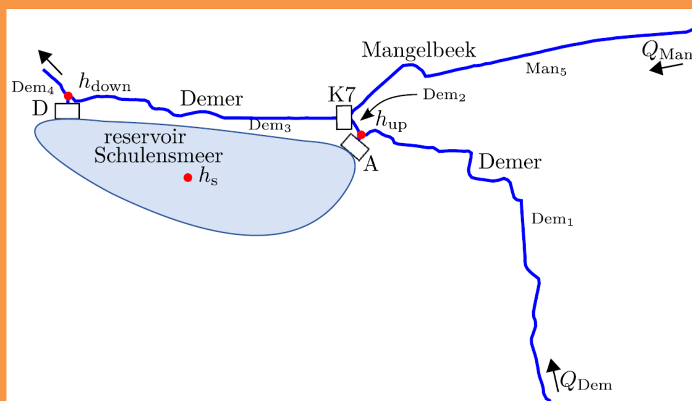



E-Health
Smart Cities
Industry 4.0
Digital Economy
Signal processing & systems
Data Mining - Exploration
Data Mining - Prediction
Data Mining - Visualisation

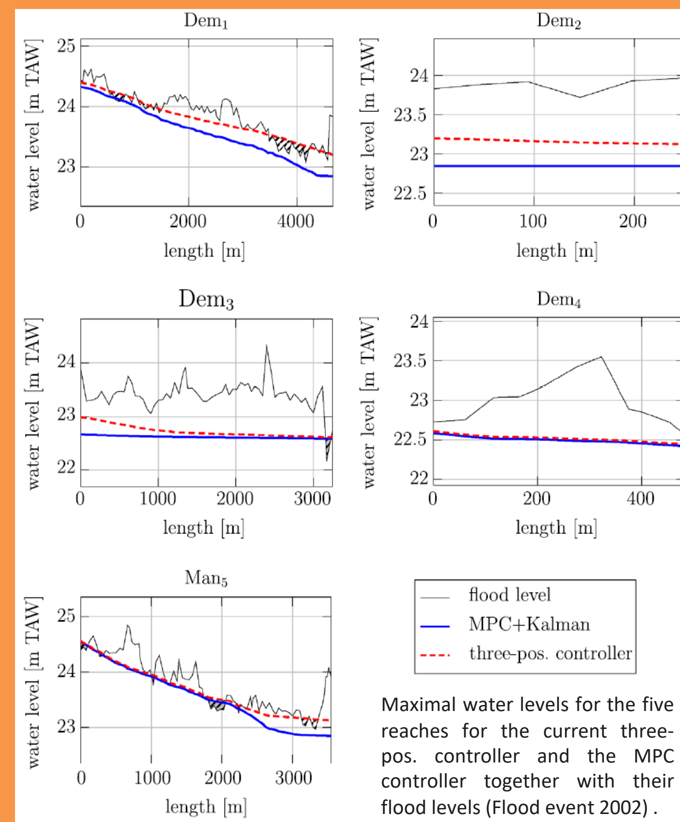
Smart Cities – Water monitoring

Implementation of a Nonlinear Model Predictive controller (NMPC) for the Demer

This project focuses on the development and implementation of an advanced control strategy for avoiding future floodings of the Demer river in Belgium.



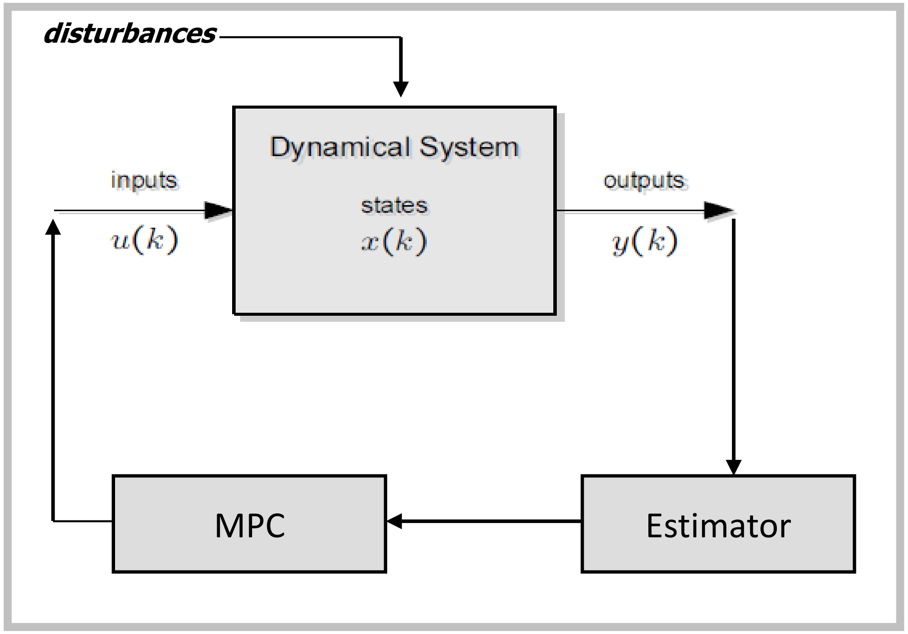
Upstream part of the Demer that is modelled and controlled in the preliminary study done by KU Leuven/ESAT/Stadius

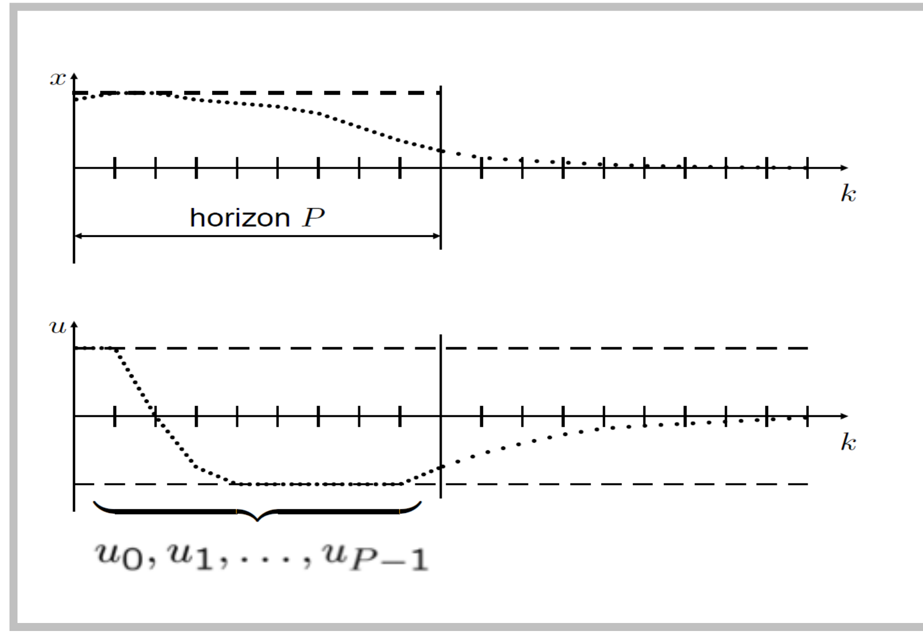


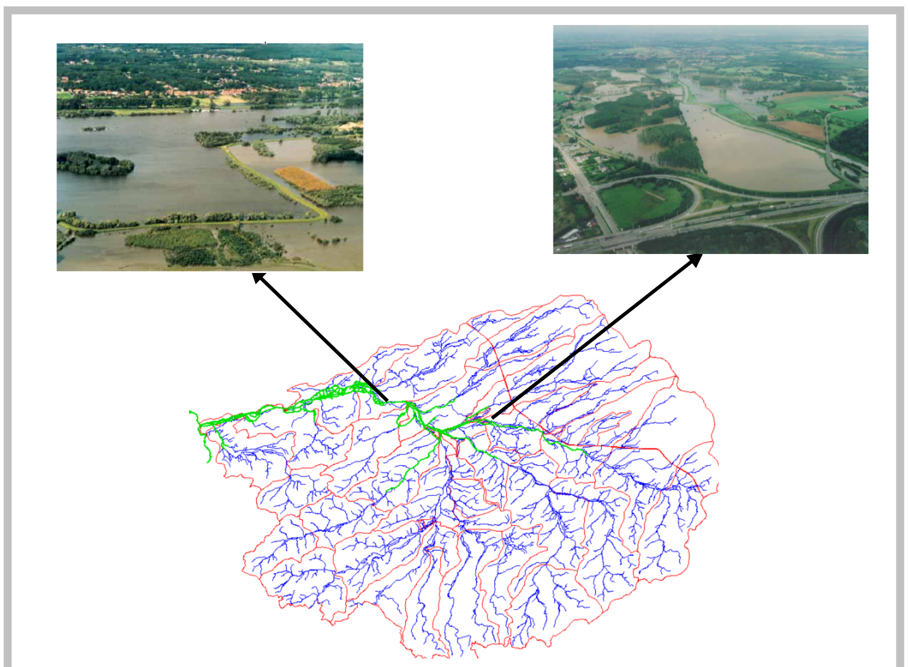
Maximal water levels for the five reaches for the current three-pos. controller and the MPC controller together with their flood levels (Flood event 2002).

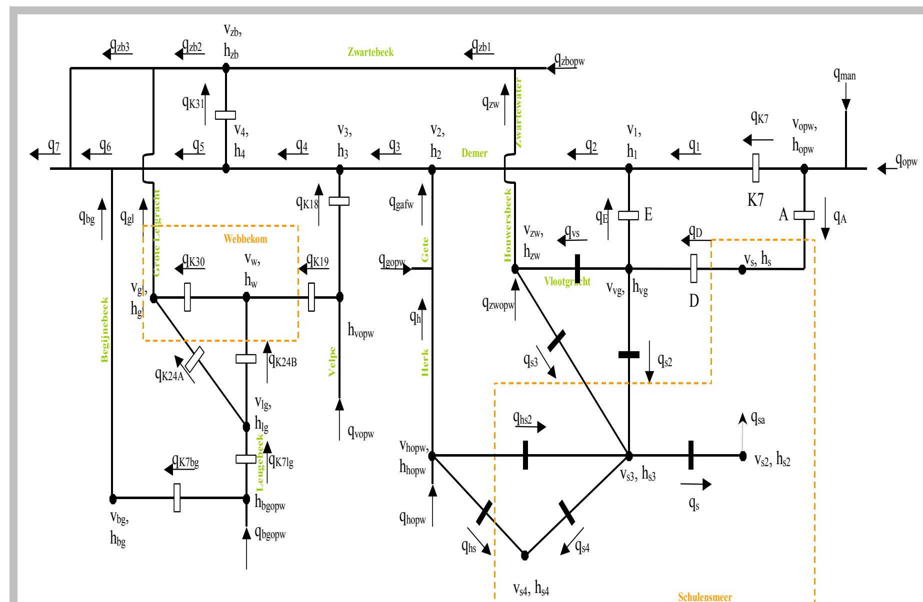
IPCOS
IMDC International Marine & Dredging Consultants
COFELY FABRICOM GDF SUEZ
anteagroup

Model Based Predictive Control for Flood Regulation: Demer







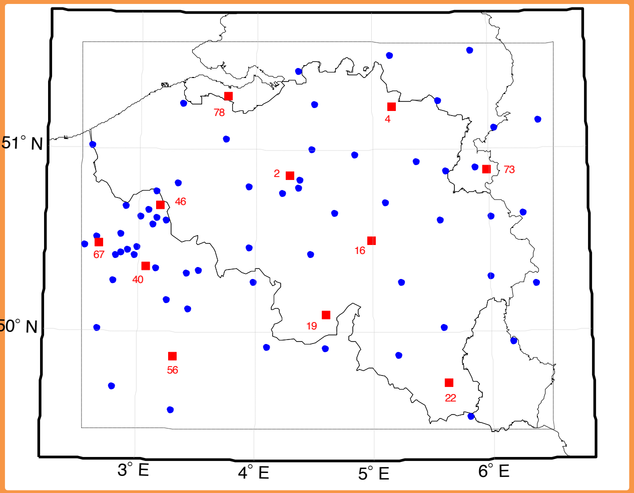


Smart Cities – Air quality

Data assimilation in the Air-quality model Aurora

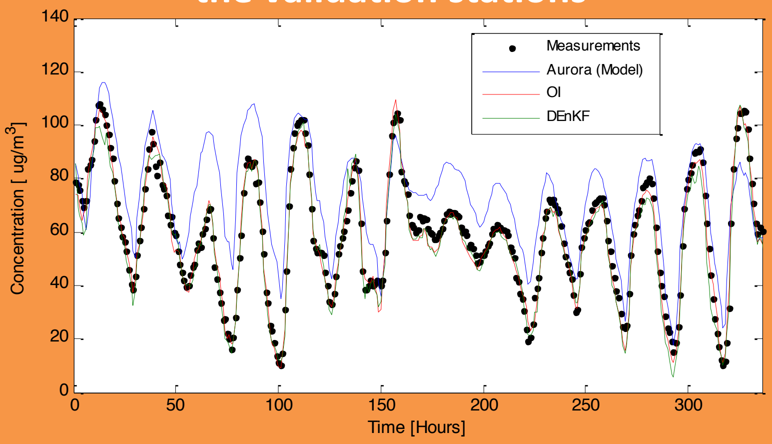
The objective of this research was to improve the concentration estimates of the air-quality model Aurora by using data assimilation techniques (e.g., Optimal Interpolation (OI), Deterministic Ensemble Kalman Filter (DEnKF, etc.)

O_3 air-quality stations



- Assimilation stations
- Validation stations

Average of the O_3 concentration over the validation stations



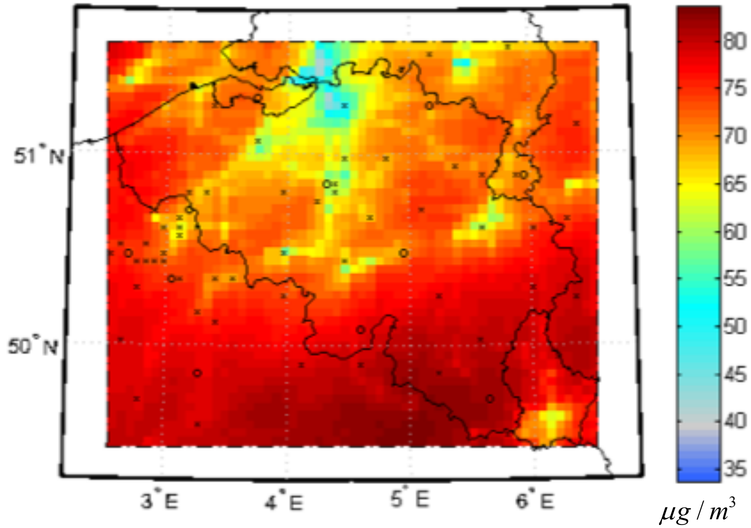
Starting date: May 28th, 2005 at midnight

This study was carried out within the framework of the IWT project CLIMAQS, "Climate and Air Quality Modeling for Policy Support".

vito
vision on technology

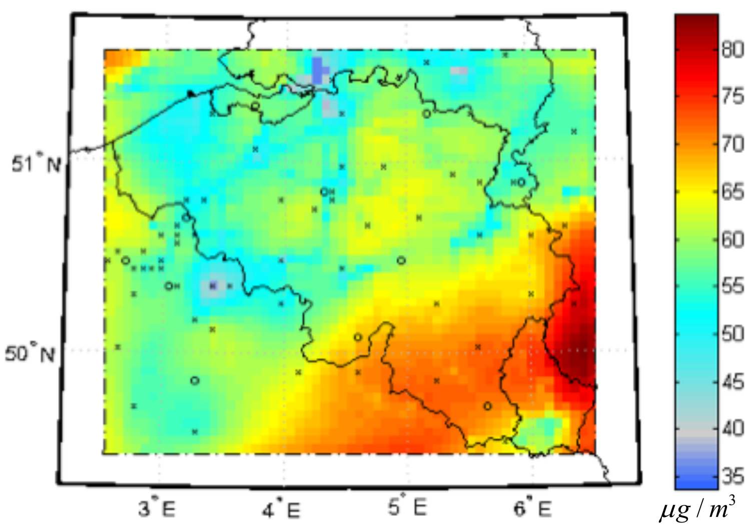
Average of the O_3 concentration field over the 14 day period

Free-run of Aurora

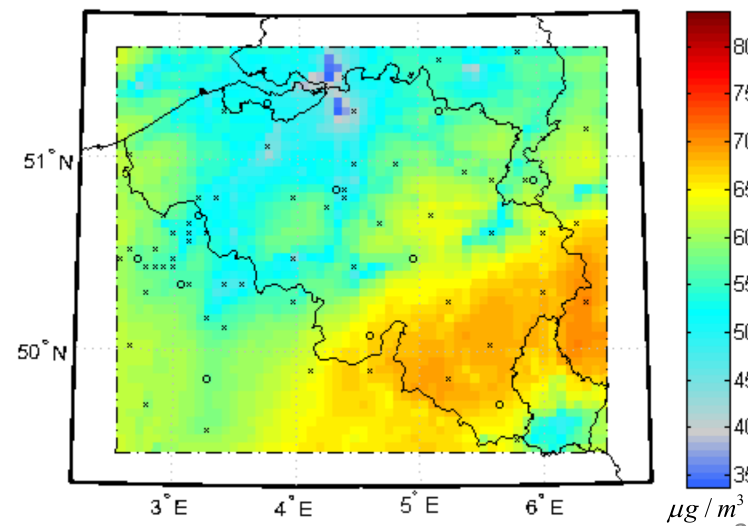


x - Assimilation station
o - Validation station

Optimal Interpolation



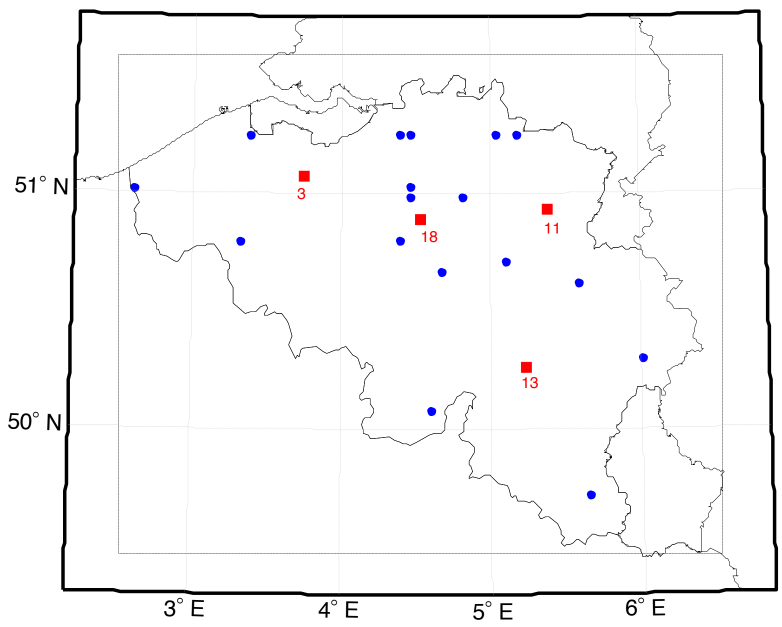
DEnKF



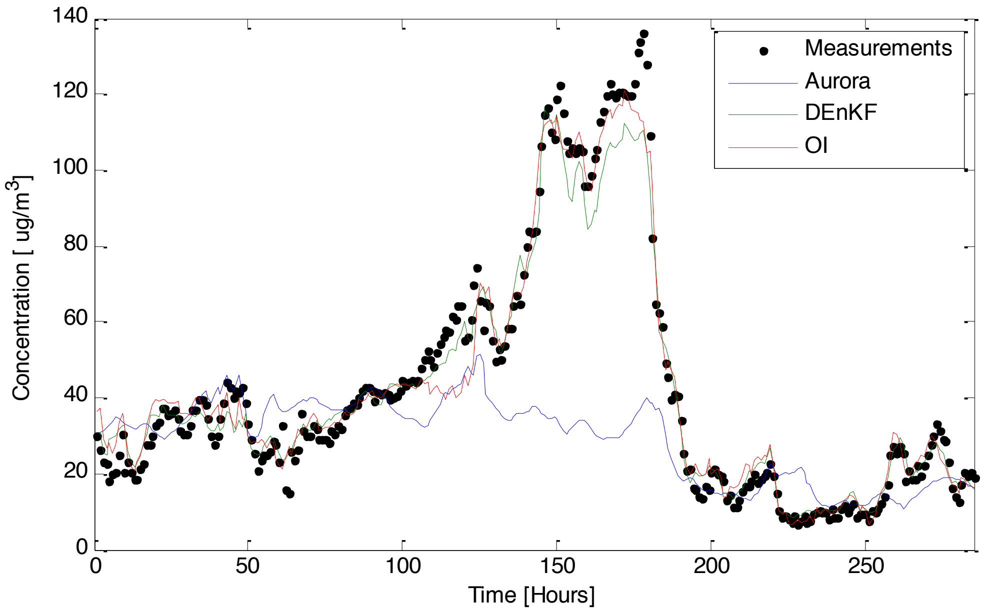
Data Assimilation

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



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

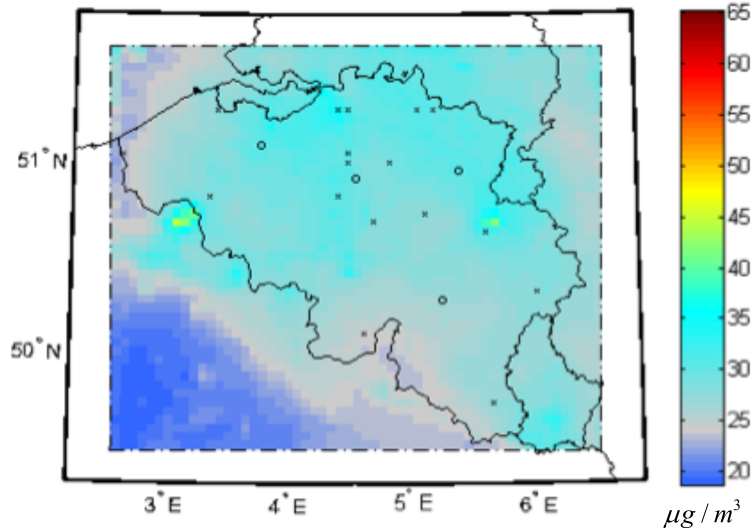


- Assimilation stations
- Validation stations

Starting date: January 20th, 2010 at midnight

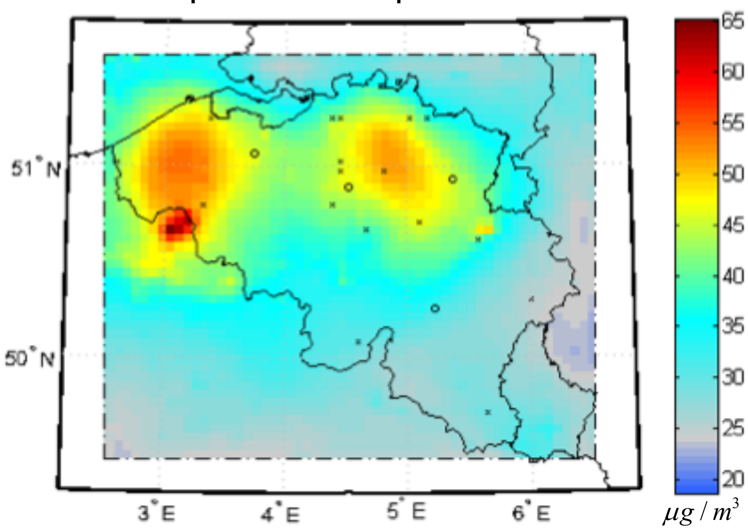
Average of the PM_{10} concentration field

Free-run of Aurora

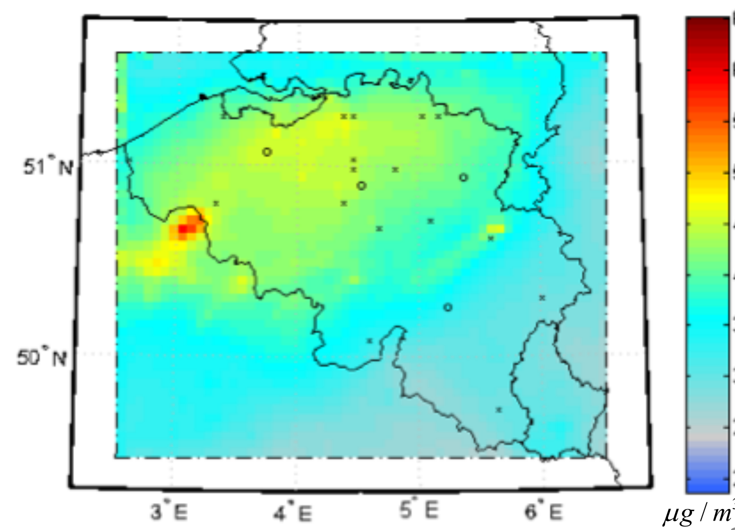


x - Assimilation station
o - Validation station

Optimal Interpolation

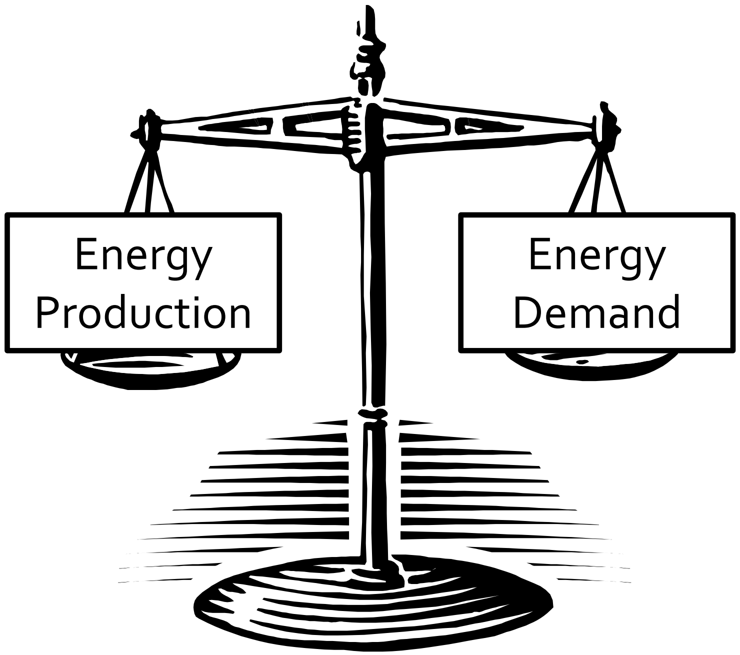


DEnKF



Electric load forecasting

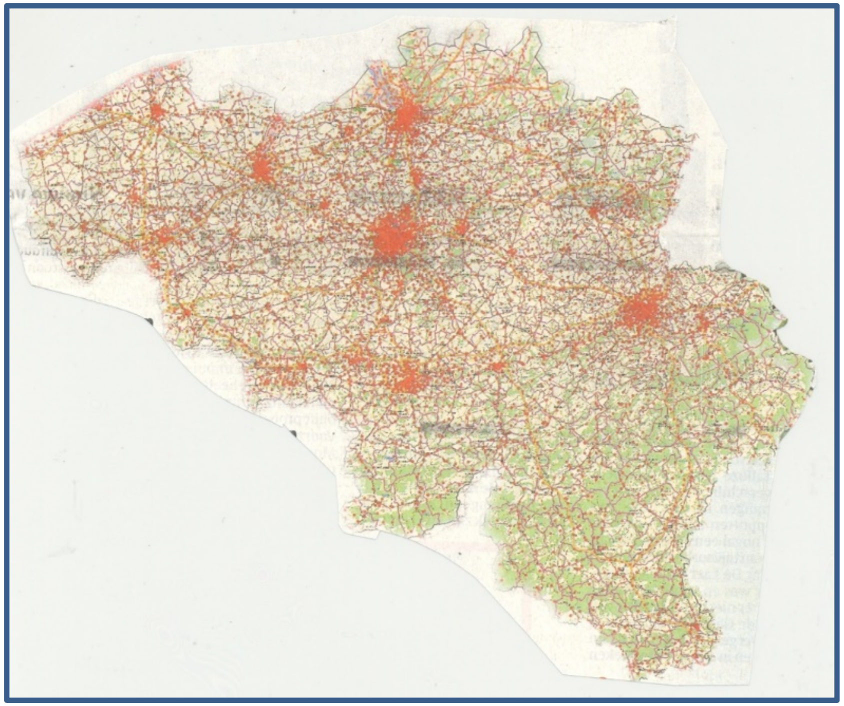
Problem

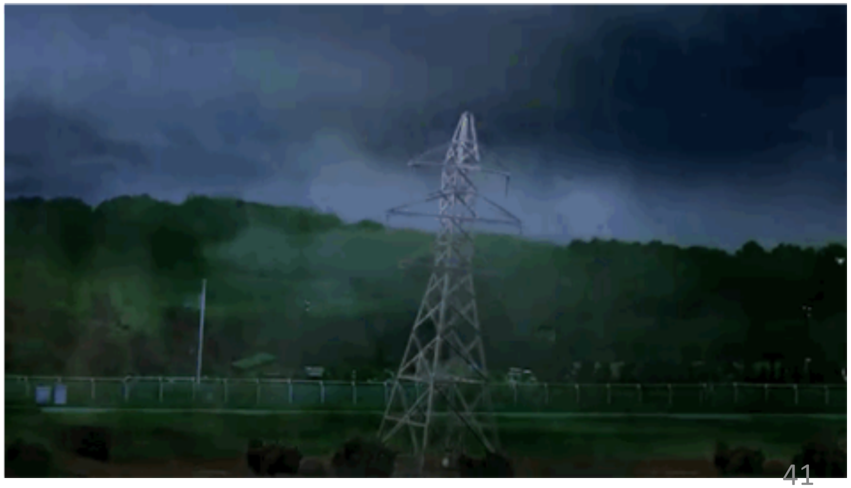


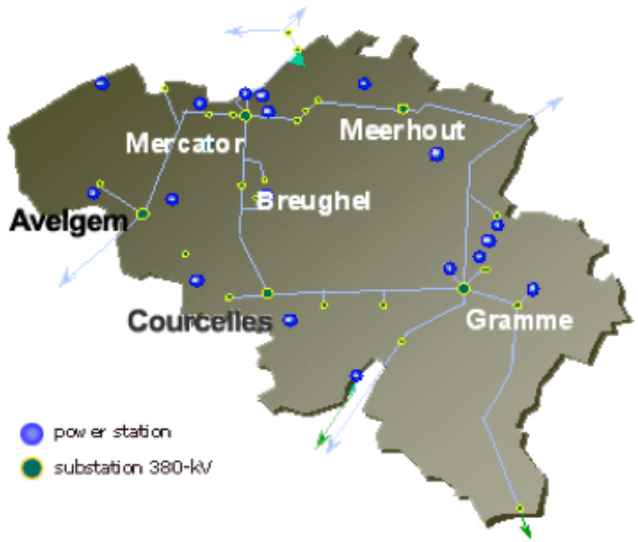
How to forecast the demand?

Power grid

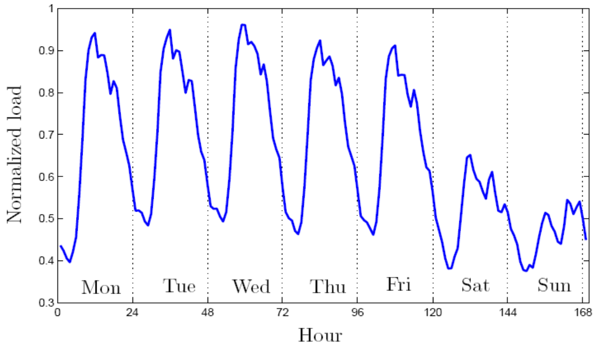

België en Europa
Het Elia-net:
knooppunt van elektriciteitverkeer in Europa
380 kV interconnectienet met hoogspanningsstations

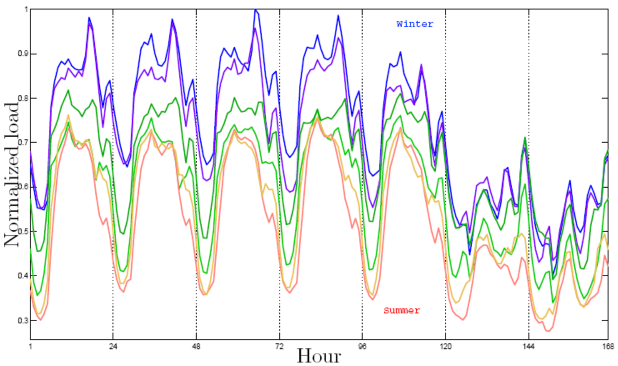






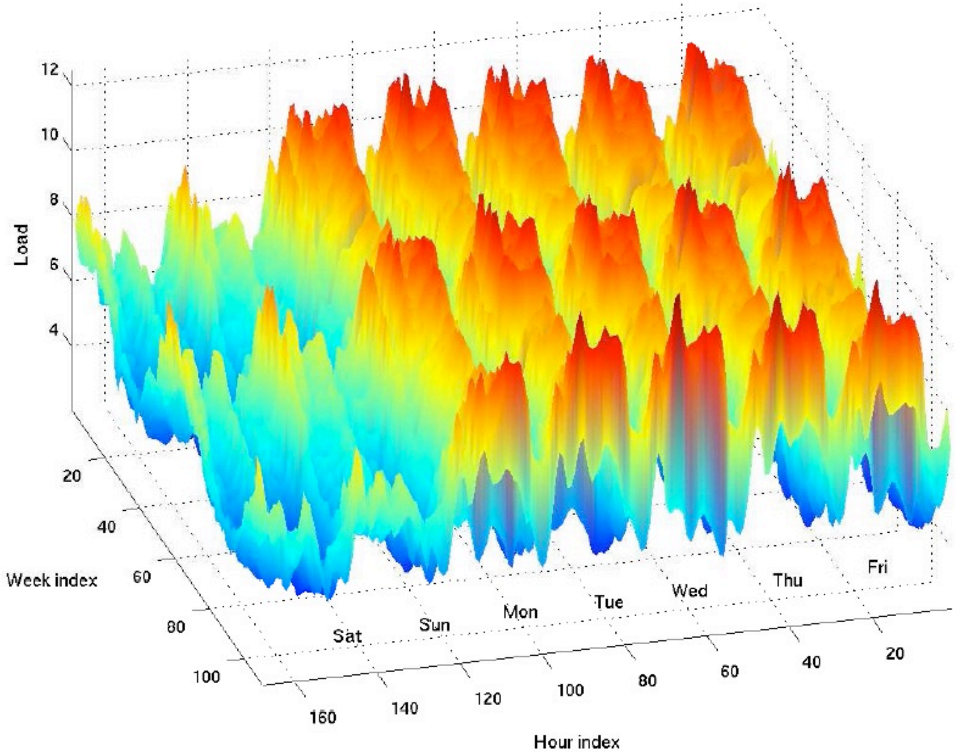
250 transformer substations
Every 15 min, 5 years



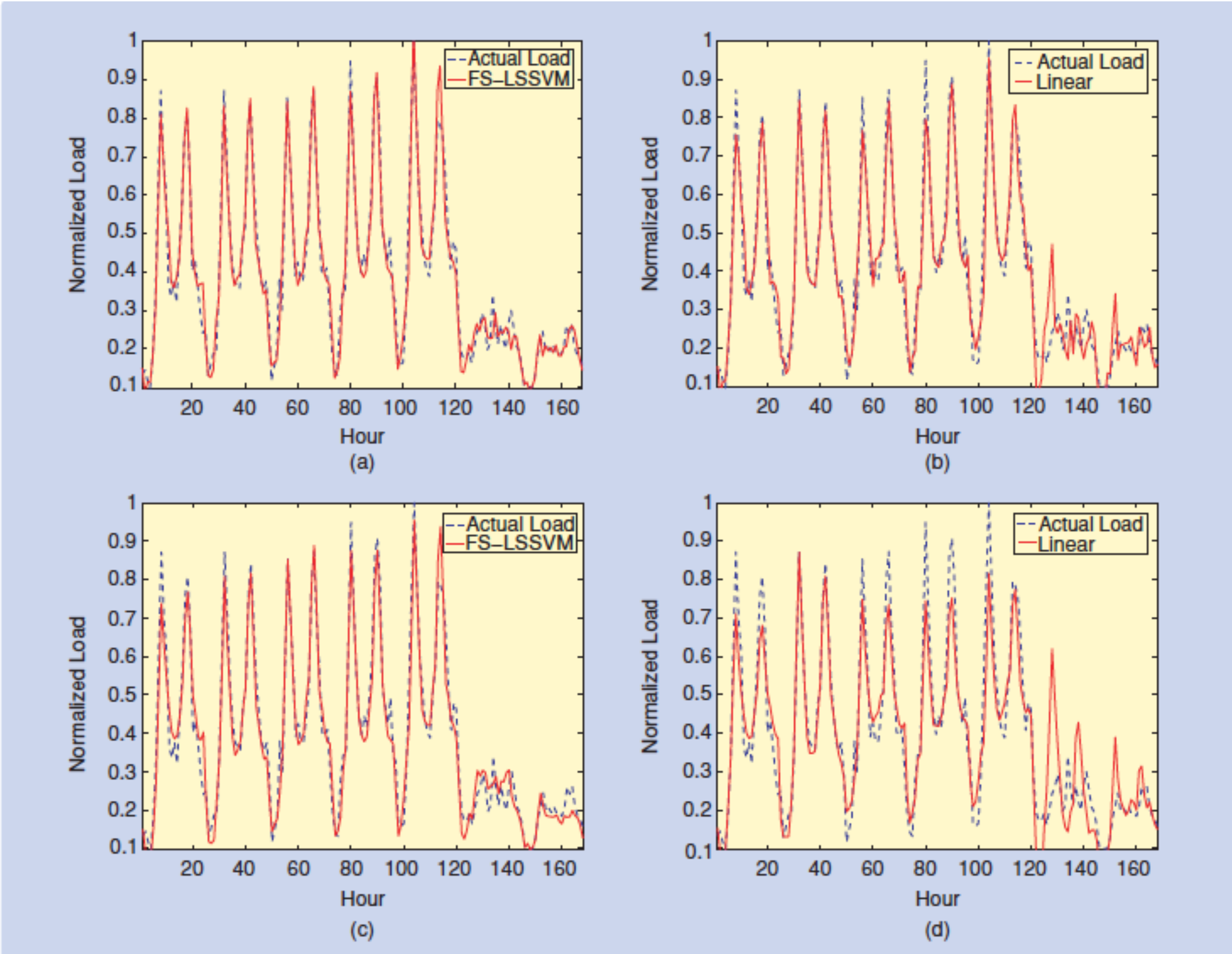


1 post, 1 week

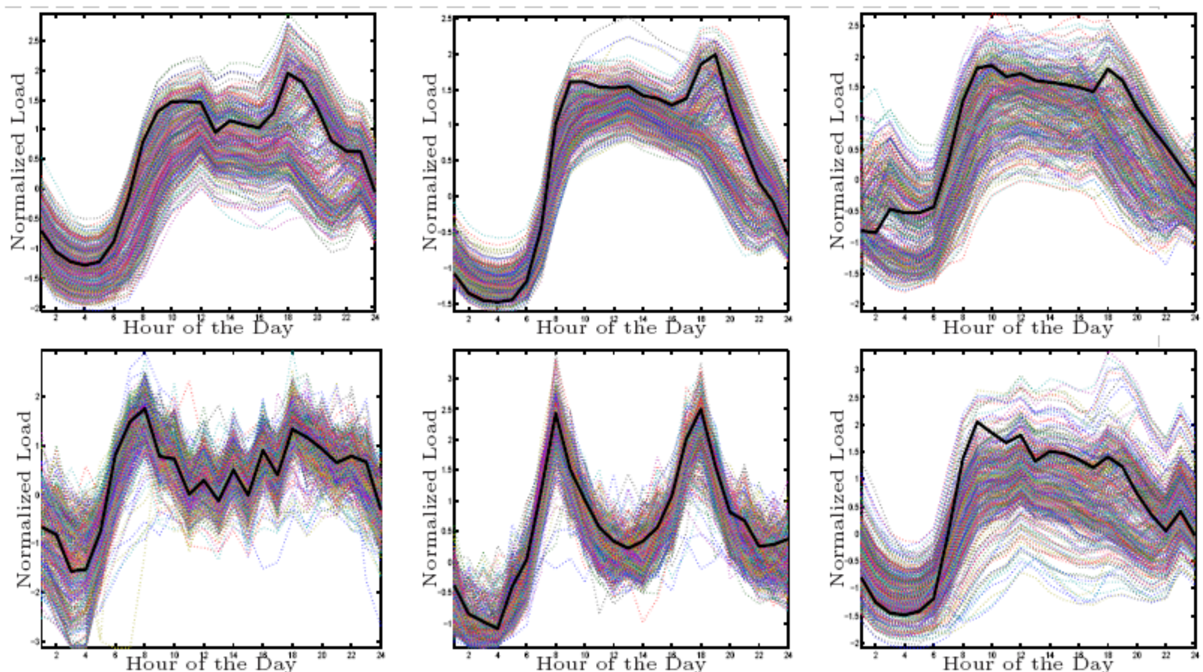
1 post, four seasons



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



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

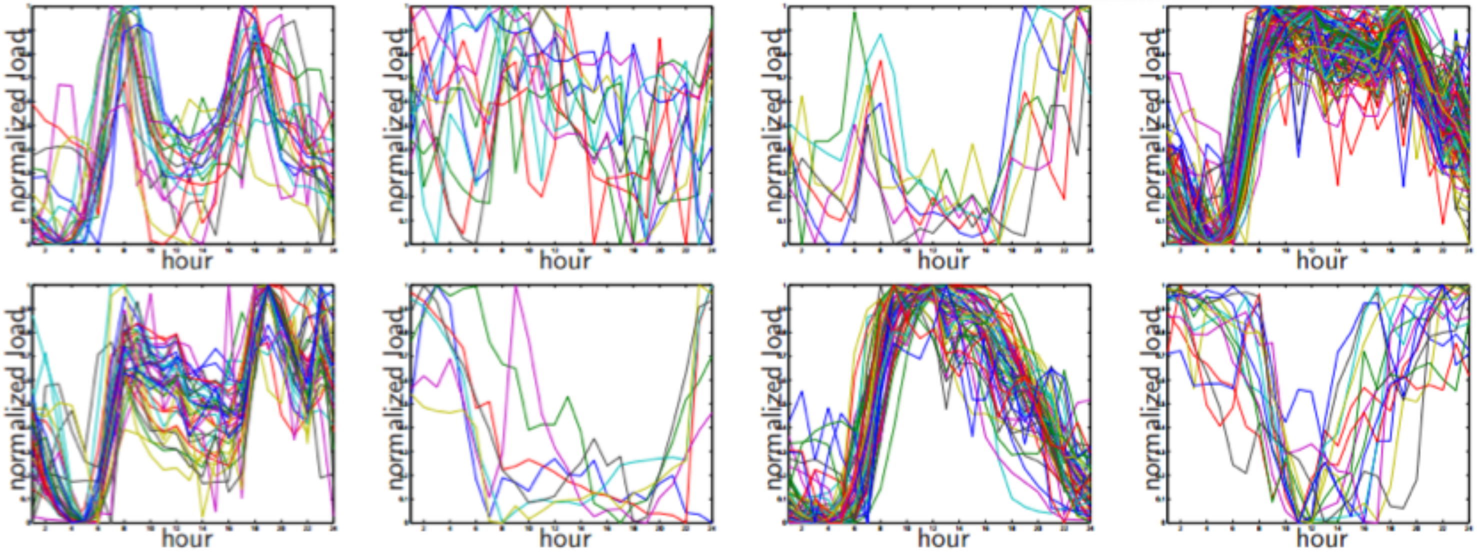


6 posts, 1 year
Seasonalities, calender holidays !

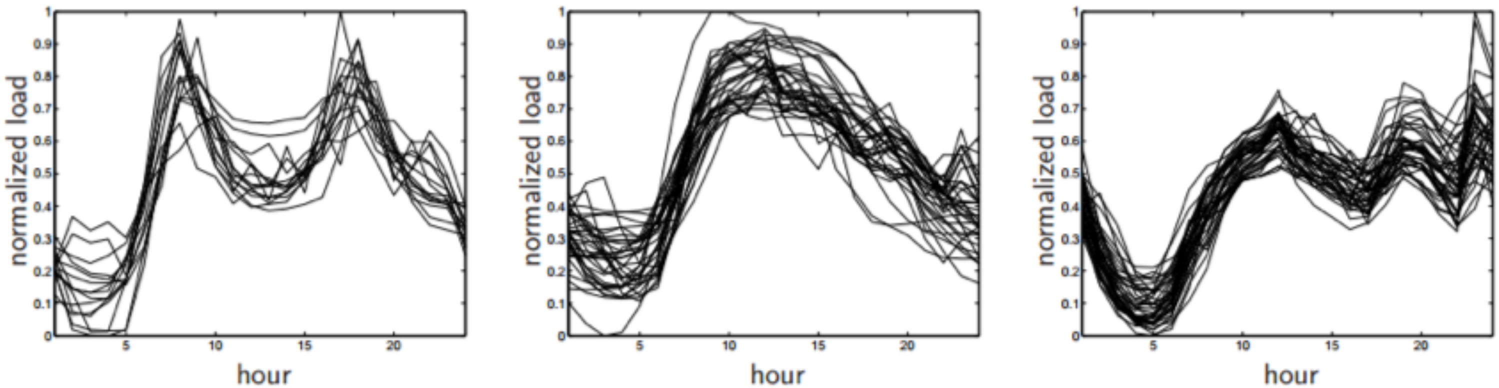
Electric Market Segmentation

Data

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

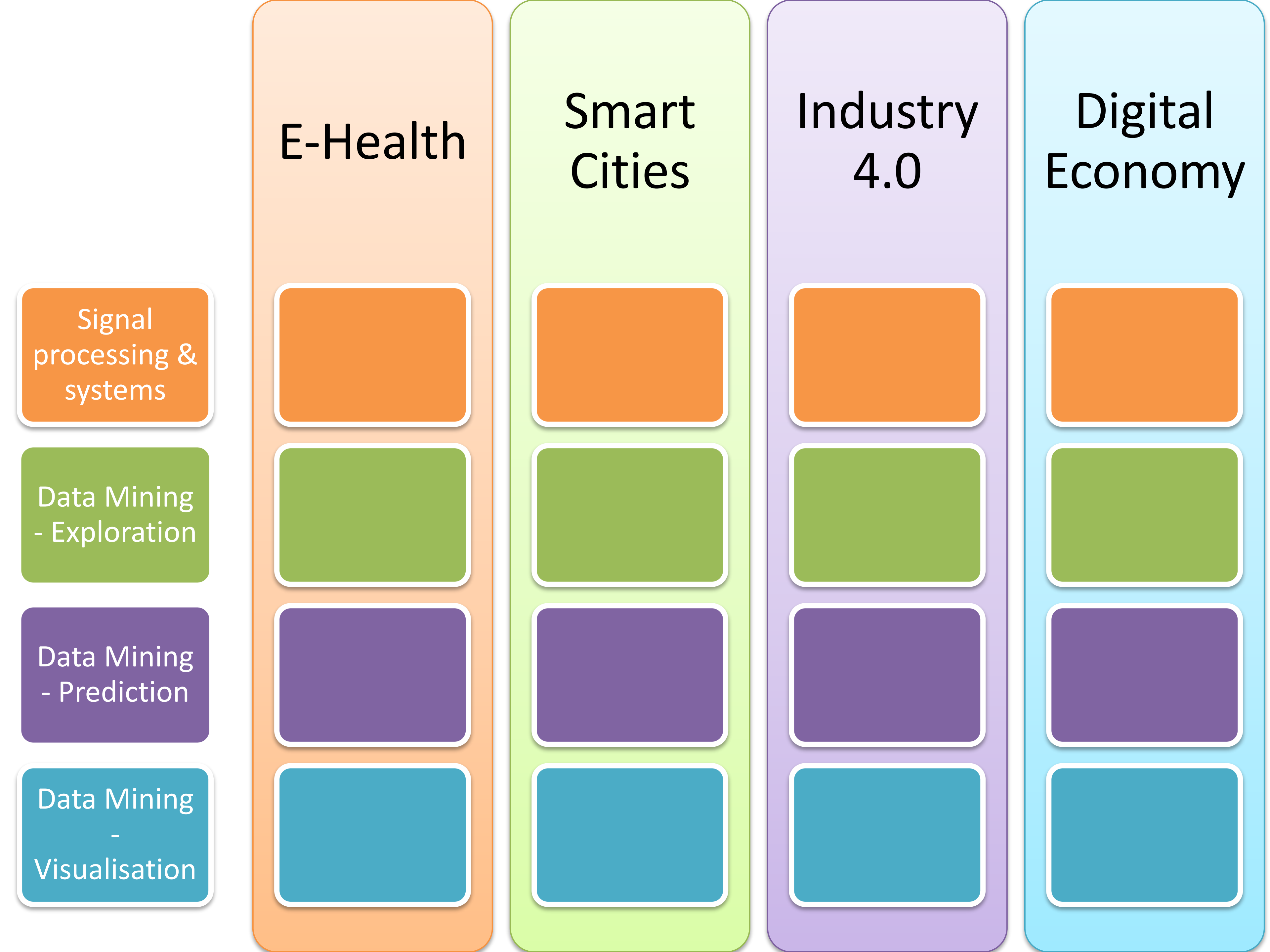


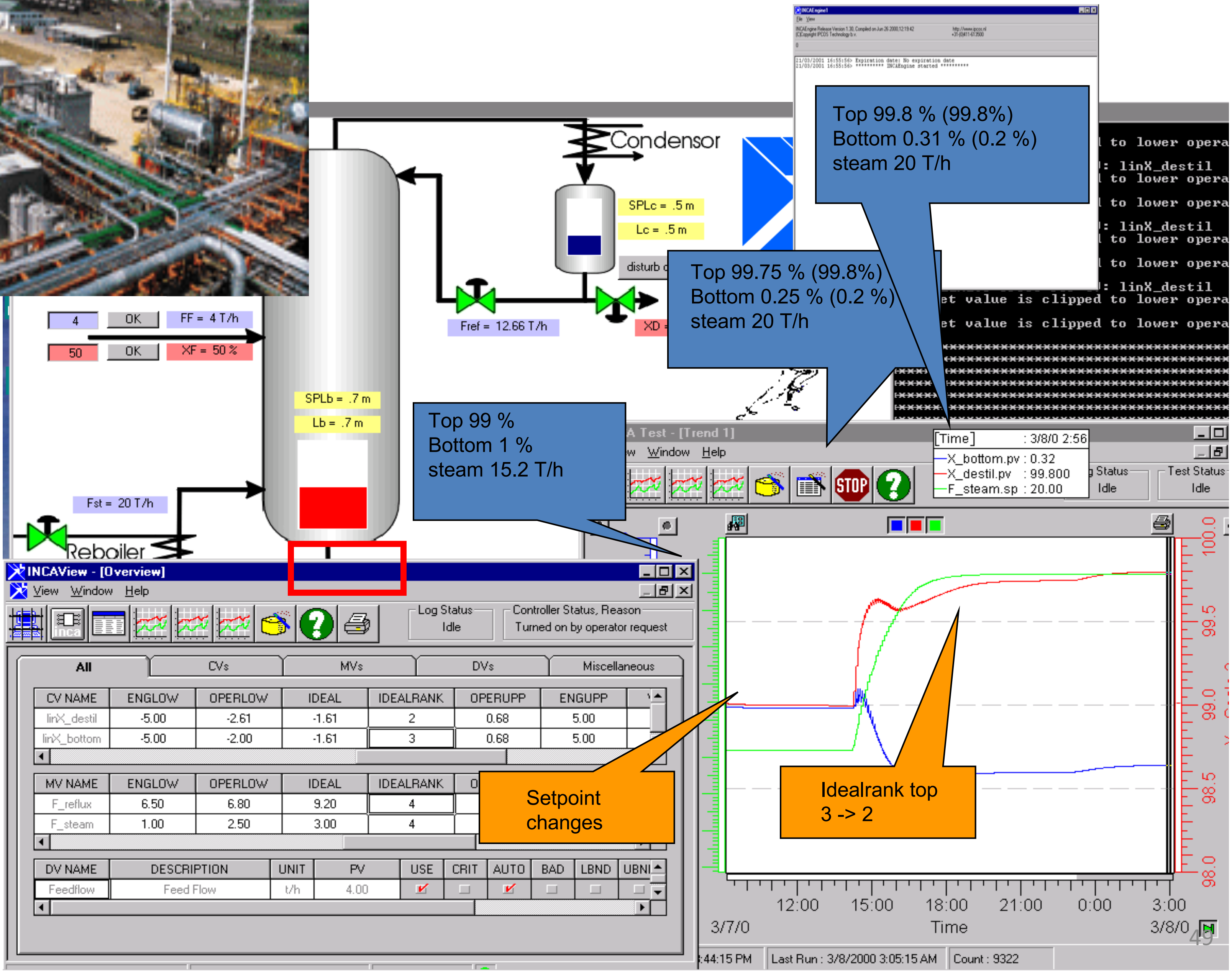
Electric Market Segmentation



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*


E-Health
Smart Cities
Industry 4.0
Digital Economy
Signal processing & systems
Data Mining - Exploration
Data Mining - Prediction
Data Mining - Visualisation

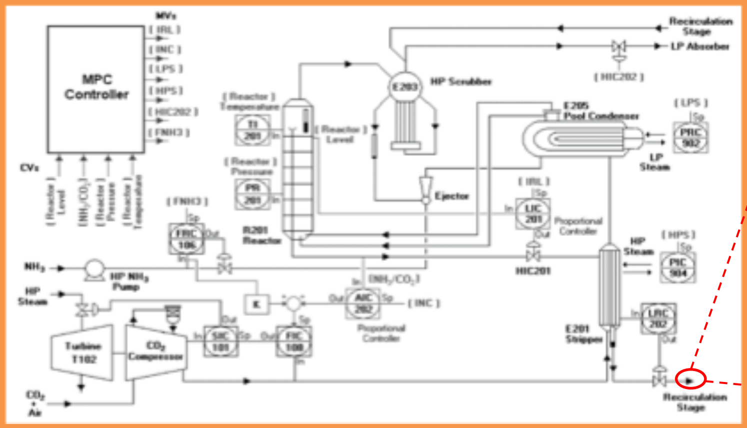


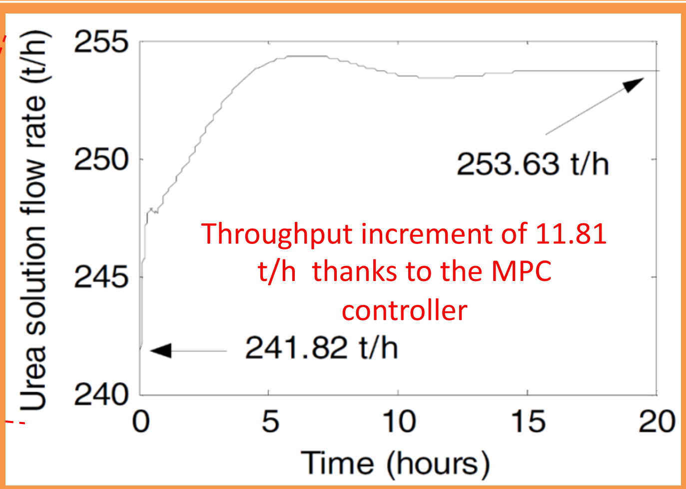
Condensor
Reboiler
SPLc = .5 m
Lc = .5 m
Fref = 12.66 T/h
FF = 4 T/h
XF = 50 %
SPLb = .7 m
Lb = .7 m
Fst = 20 T/h
Top 99.8 % (99.8%)
Bottom 0.31 % (0.2 %)
steam 20 T/h
Top 99.75 % (99.8%)
Bottom 0.25 % (0.2 %)
steam 20 T/h
Top 99 %
Bottom 1 %
steam 15.2 T/h
Setpoint changes
Idealrank top
3 -> 2
INCAView - [Overview]
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
Feedflow Feed Flow t/h 4.00
[Time] : 3/8/0 2:56
X_bottom.pv : 0.32
X_destil.pv : 99.800
F_steam.sp : 20.00
Test Status Idle
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

Industry 4.0 – Control

Control of the synthesis section of a Urea plant using MPC control techniques

In this project, Model Predictive Control (MPC) strategies were used for stabilizing and maximizing the throughput of the synthesis section of a urea plant, while satisfying all the process constraints.



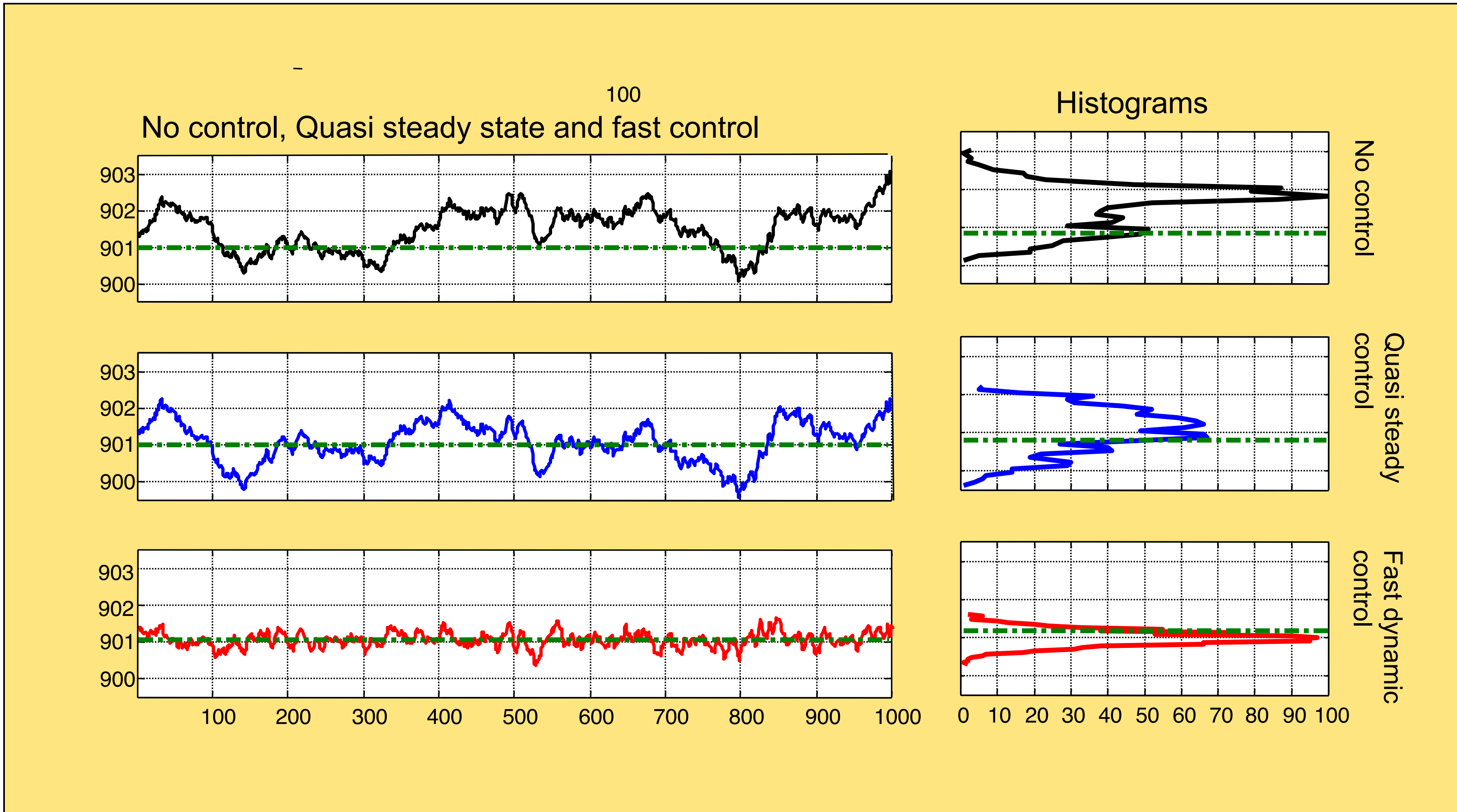


ROI:
- IPCOS successfully implemented an MPC control system on the urea plant of Yara Brunsbüttel.
- IPCOS developed an advanced control solution for urea plants, "urea@Max"
- …

IPCOS

Protomation

Modelling for control



HELPING MANUFACTURING COMPANIES RAPIDLY IMPROVE PLANT AVAILABILITY AND ASSET EFFECTIVENESS THROUGH BIG DATA IOT DISCOVERY ANALYTICS

Google FOR PROCESS INDUSTRY: BECOME THE NEXT BIG THING IN MANUFACTURING IT

TrendMiner

MEET PETER

Plant Manager and under pressure to maximize **Plant Availability** and **Asset Effectivness**

MEET ALICE

Process Engineer. Reports to Peter. Needs to optimize the process and minimize the number of **Abnormal Situations**. Currently not satisfied with the level of **Operational Intelligence**.

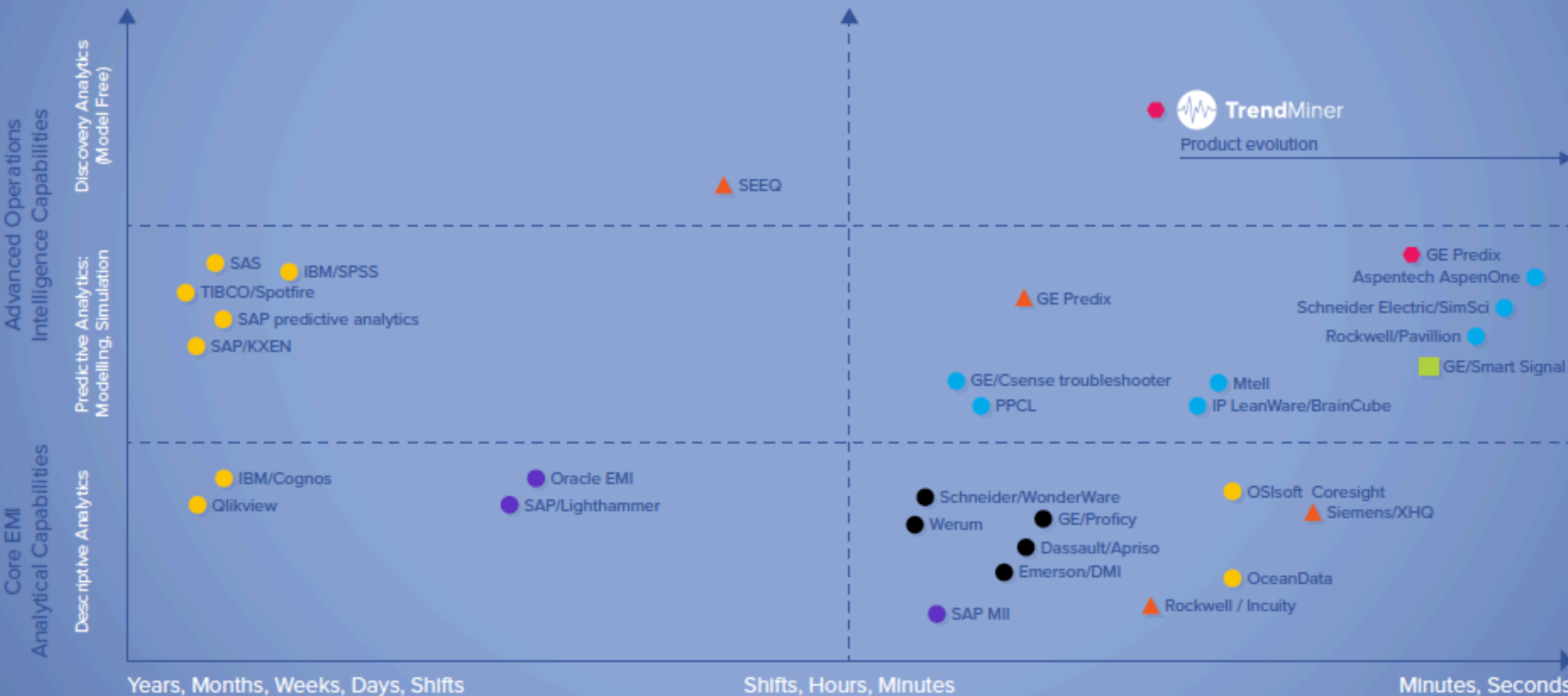
umicore
materials for a better life

TrendMiner

COMPETITIVE POSITIONING

ERP-based EMI · BI/Analytics · Adv Analytics · MES · Asset Monitoring · EMI-framework · APC/RTO/RPO/APS

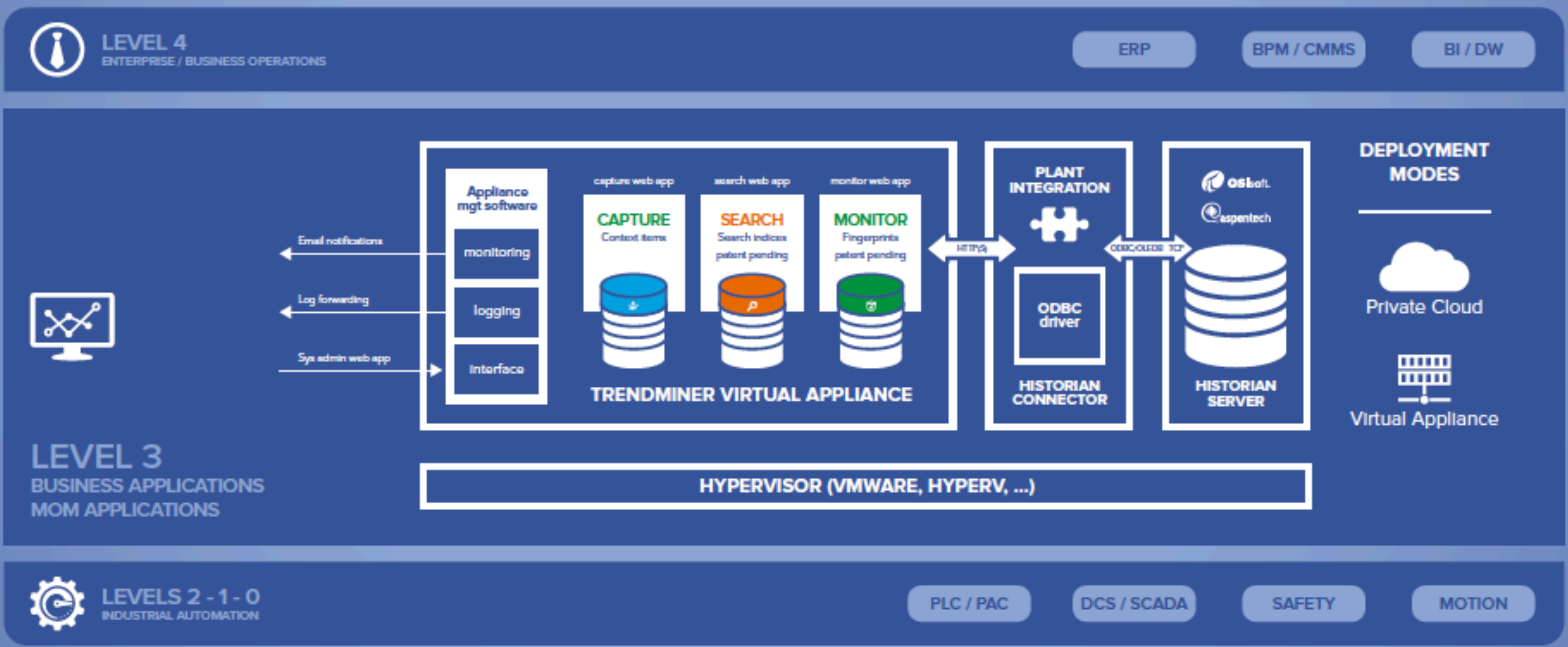
OUR UNIQUE POSITIONING ALLOWS US TO PARTNER WITH DOMINANT LEGACY PLAYERS



TrendMiner

TECHNOLOGY

PLUG N' PLAY INTEGRATION WITH EXISTING STANDARDS AND SYSTEMS



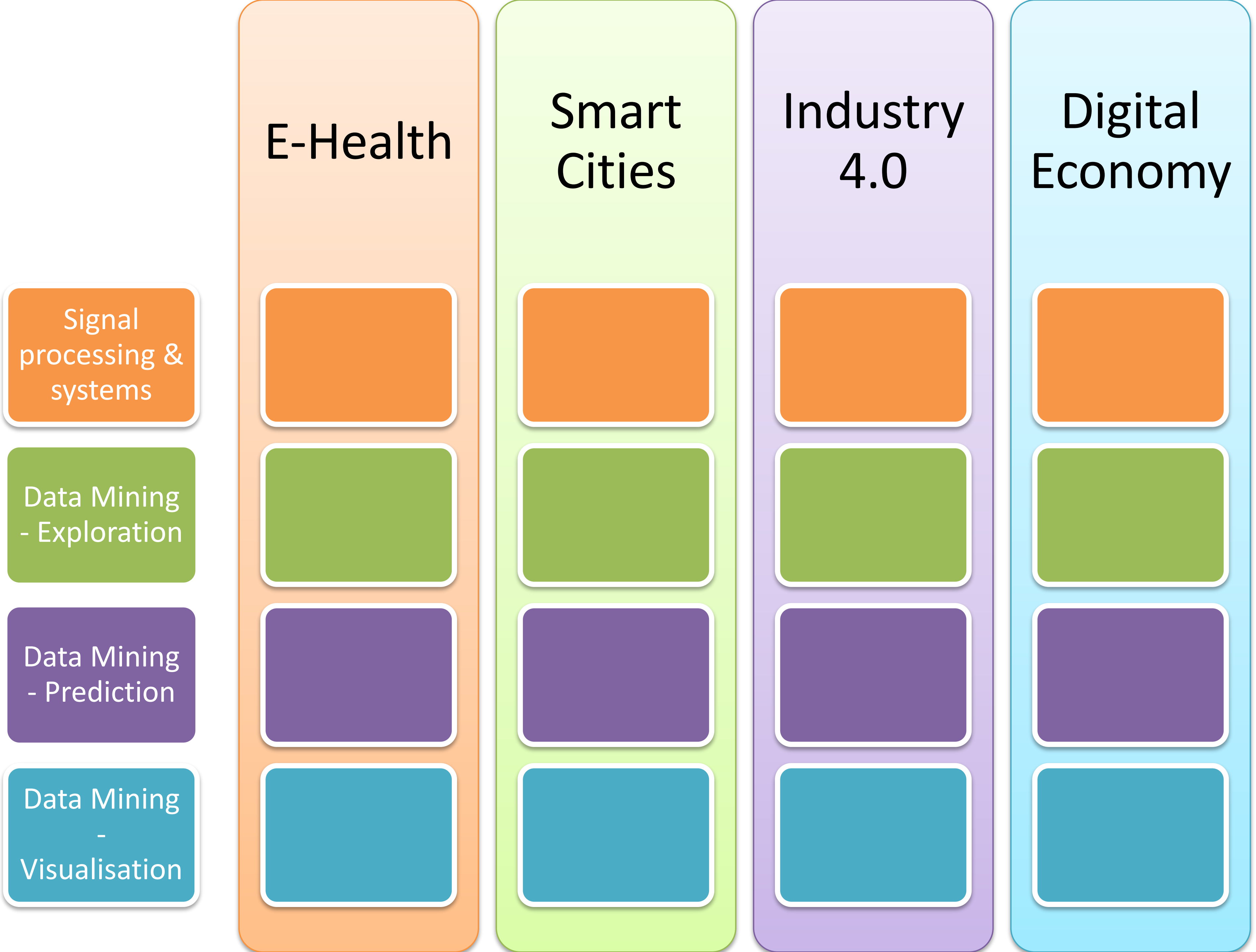
TRACTION

100% CLOSE RATE SO FAR WITH **19** GLOBAL COMPANIES WITH **SHORT SALES CYCLE**

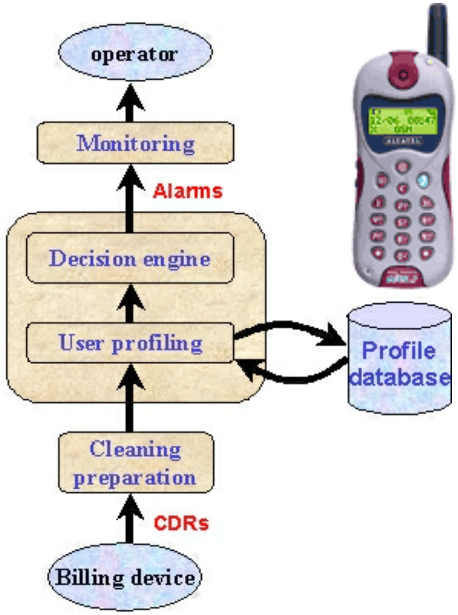
64+
customers

100+
manufacturing plants

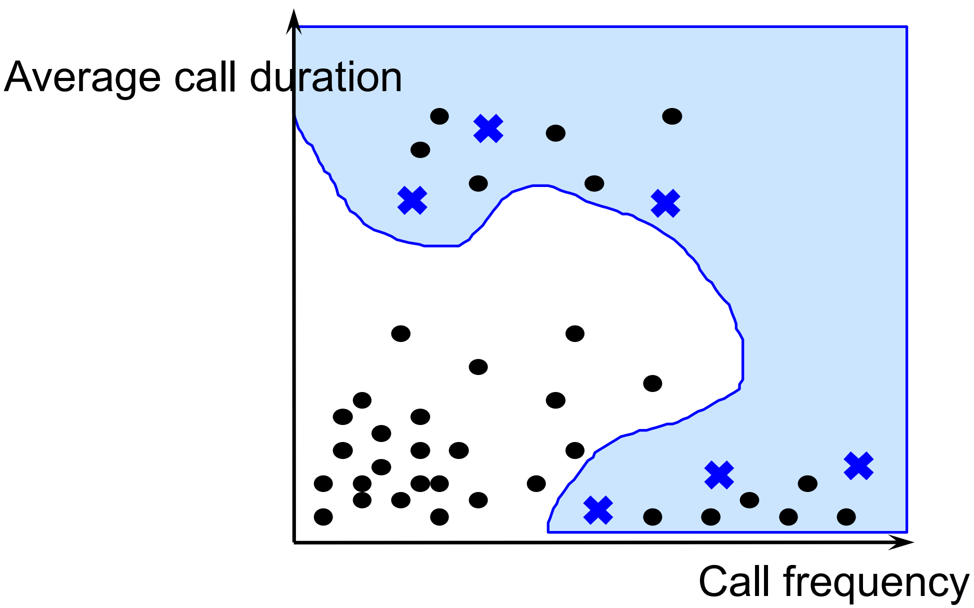
TrendMiner


E-Health
Smart Cities
Industry 4.0
Digital Economy
Signal processing & systems
Data Mining - Exploration
Data Mining - Prediction
Data Mining - Visualisation

Fraud detection on mobile phone network



	Short Duration	Long Duration	High Frequency	International	Same Destination	Off Peak	Call Forwarding	Behaviour Change
Direct call selling		X	X	X			X	
PABX fraud	X		X		X	X		X
Freephone fraud	X		X		X			X
Premium rate fraud		X	X		X			X
Subscription fraud			X					
Handset theft		X	X	X	X			X



Big Data Mining

What ? Why ? How ? Where ? Who ?

Prof. Dr. Bart De Moor

March 2018