Machine Learning Methods for Anomaly Detection:

Examples from Dam and Levee Condition Monitoring

Valeria Krzhizhanovskaya

University of Amsterdam, The Netherlands
ITMO University, Russia

V.Krzhizhanovskaya@uva.nl

ITMO UNIVERSITY

UNIVERSITY OF AMSTERDAM

Outline

- IoT, Big Data, Data Science
- Data-driven modelling
- Machine learning
- Anomaly detection

- Example:
  dam and levee condition monitoring

BIG
DATA
USE
ONE
SINCE
SOFTWARE
UBIQUITOUS
REQUIRING
RADIO-FREQUENCY
COMPLEXITY
SOLID
WIRELESS
SAN
TOLERABLE
SIZE
NEEDED
QUALITIES
PETABYTES
INTERNET
TECHNOLOGIES
PROCESSING
LOGS
BURIED
ANALYTICS
PARALLEL
BUSINESS
MOVING
ORGANIZATIONS
DISTRIBUTED
USED
CAPTURE
MAY
MANAGEMENT
MANAGE
GROW
SOCIAL
LARGE
STORE
DEFINING
EVERY
CASE
LARGER
STORAGE
RESEARCH
DATABASES
MPP
TERABYTES
CONTINUES
COST
CITATION
COMPLEX
INDEXING
DIFFICULTY
TARGET
ABILITY
SETS
EXAMPLES
APPLIED
AMOUNT
DESCRIBING
ELAPSED
CURRENT
THOUGHT
SENSOR
ARCHIVES
WORKING
TYPES
USING
GARTNER
MASSIVELY
BIOLOGICAL
RELATED
ALSO
HUNDREDS
CREATED
PERFORMANCE
DISK
RELATIONAL
DESKTOP
CURRENTLY
TENS
CAPACITY
FORMS
PRESENTATIONS
WORLD'S
PRACTITIONERS
NOW
FC
ZETTABYTES
GENOMICS
SET
COMPUTING
TOOLS
SHARED
TIME
COMBAT
SIGNIFICANT
WITHIN
PROCESS
INCLUDE
SYSTEMS
DEFINITION
BIOGEOCHEMICAL
NETWORKS
SEARCH
RECONSIDER
OPPORTUNITIES
INFORMATION
CONNECTOMICS
RECORDS

Big Nonsense:
THE END OF SCIENTIFIC THINKING IS NEAR…

Big Nonsense: the end of scientific thinking is near…

Peter M.A. Sloot

'The real purpose of the scientific method is to discover that Nature hasn't misled you into thinking you know something you don't actually know.'

Robert M. Pirsig,1974

Astronomers of the Maya civilisation and astronomers of the Babylonian civilisation were brilliant in predicting astronomical events. For instance, from meticulous observations of the Sun, Moon, Venus and Jupiter they were able to predict with astonishing accuracy the 584-day cycle of Venus or the details of the celestial track of Jupiter [1]. Yet they had no clue about our heliocentric solar system, they believed that the earth was flat and they were completely ignorant of the real movement of stars and planets while being convinced that the sky was supported by four jaguars, each holding up a corner of the sky. If we would be sent back in time and speak to them about the planets orbiting the sun they would laugh at us and challenge us to come with the accurate predictions that they are able to make. With all our knowledge, but without thousands of years of technological development, we would not be able to come close to any of their predictions. So being laughed at would be a small punishment, more likely we would be ritually slaughtered…

This is how I feel, time and again when one or another data fetishist tries to convince me of the importance of big data over proper scientific inference and methods.

We are on the edge of a fundamental clash of scientific methods. If we do not solve this soon we risk being thrown back into pre-medieval scientific methods. This warning should reverberate strongly for us complexity- and system scientists, as we are slowly being drawn into data obscurantism.

Indeed, our era of data intensive science raises fundamental questions at the basis of modern science [2]. In the history of science, two opposing philosophies have been advocated, inductivism and deductivism, due to two knights of the realm, Sir Francis Bacon (1561–1626) and Sir Karl Popper (1902–1994). Most students of the physical sciences and engineering accept that their disciplines adhere to the precepts of Popperian falsifiability where one proceeds through development and use of deductive models, whose predictions are taken to furnish an accurate account of the objective nature of the world. On the other hand, Baconian science, long considered dead in those sciences, has reared its head with a vengeance in the past quarter of a century, based essentially on the compilation of all sorts of 'look up tables' summarizing the results of a myriad observations and measurements. Within Baconian inductivism there is no model to be falsified, just ever more data to be collected, and conclusions to be drawn on the basis of some more or less sophisticated form of statistical inference.

As we push forward the frontiers of knowledge and attempt to understand ever more complex systems (life, climate, health, economy), we should recognize the need for a constructive interplay between these two philosophies of science. Complexity science is *the* discipline in which both these approaches merge as one.

This merger is crucial for the sake of scientific progress such that we once again can 'stop and think about the complexity, the inconceivable nature of Nature' [3]

[1] M. Ossendrijver, Science: Vol. 351, Issue 6272, pp. 482-484. (2016)
[2] P.M.A. Sloot, P. Coveney and J. Dongarra: Journal of Computational Science Vol. 1 (2010) 3–4;
[3] 'The Fantastic Mr Feynman', interview with R. Feynman on YouTube: https://youtu.be/LvqlelxXTpw

P.M.A. Sloot.
"Big Nonsense: the end of scientific thinking is near…"
43 Visions for Complexity, Ed. S. Thurner, p. 65-66.
World Scientific Publishing Co. Pte. Ltd.
ISBN 978-981-3206-84-7, 2017

Big Nonsense: the end of scientific thinking is near…

Peter M.A. Sloot

'The real purpose of the scientific method is to discover that Nature hasn't misled you into thinking you know something you don't actually know.'

Robert M. Pirsig,1974

Astronomers of the Maya civilisation and astronomers of the Babylonian civilisation were brilliant in predicting astronomical events. For instance, from meticulous observations of the Sun, Moon, Venus and Jupiter they were able to predict with astonishing accuracy the 584-day cycle of Venus or the details of the celestial track of Jupiter [1]. Yet they had no clue about our heliocentric solar system, they believed that the earth was flat and they were completely ignorant of the real movement of stars and planets while being convinced that the sky was supported by four jaguars, each holding up a corner of the sky. If we would be sent back in time and speak to them about the planets orbiting the sun they would laugh at us and challenge us to come with the accurate predictions that they are able to make. With all our knowledge, but without thousands of years of technological development, we would not be able to come close to any of their predictions. So being laughed at would be a small punishment, more likely we would be ritually slaughtered…

This is how I feel, time and again when one or another data fetishist tries to convince me of the importance of big data over proper scientific inference and methods.

We are on the edge of a fundamental clash of scientific methods. If we do not solve this soon we risk being thrown back into pre-medieval scientific methods. This warning should reverberate strongly for us complexity- and system scientists, as we are slowly being drawn into data obscurantism.

How big is the Big Data?

Google "internet of things"

All News Images Videos

ut 34,000,000 results (0.75 seconds)

Google "big data"

All News Images Videos

About 69,300,000 results (0.56 seconds)

Google "gl

All News Books Images

About 424,000 results (0.73 seconds)

oud computing"

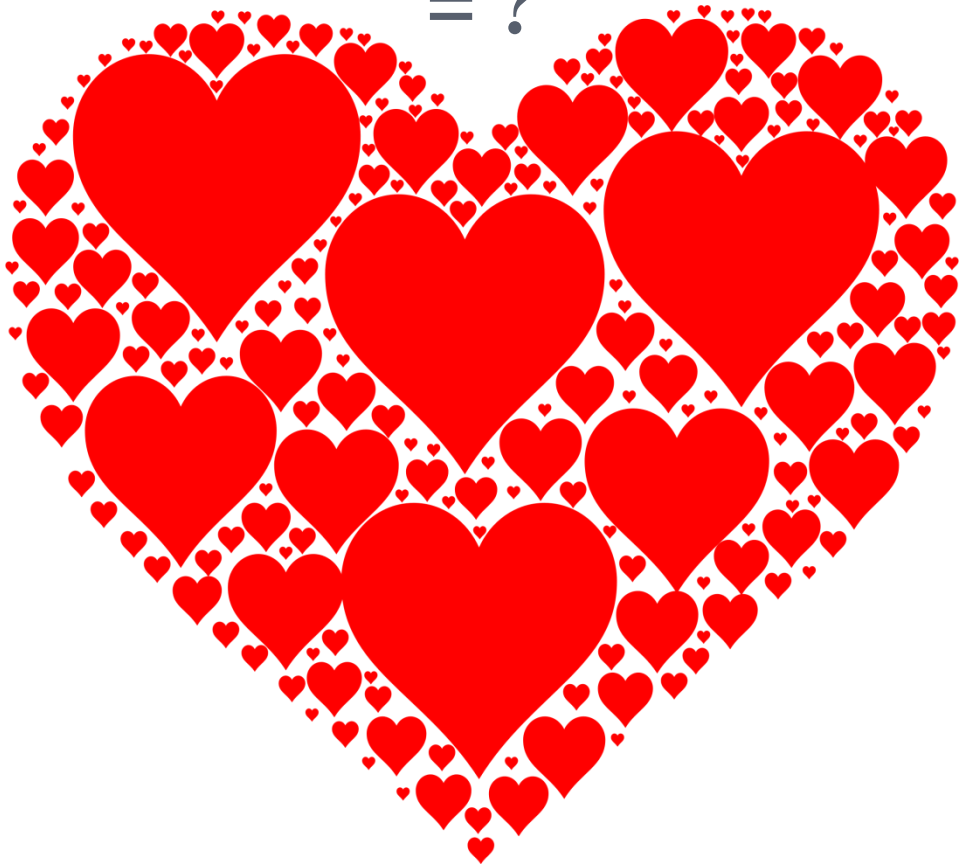
All News Images Books

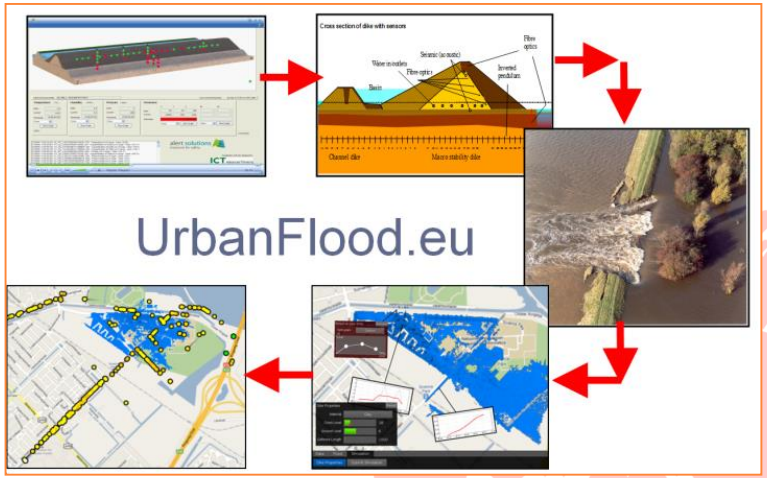
About 79,200,000 results (0.82 seconds)

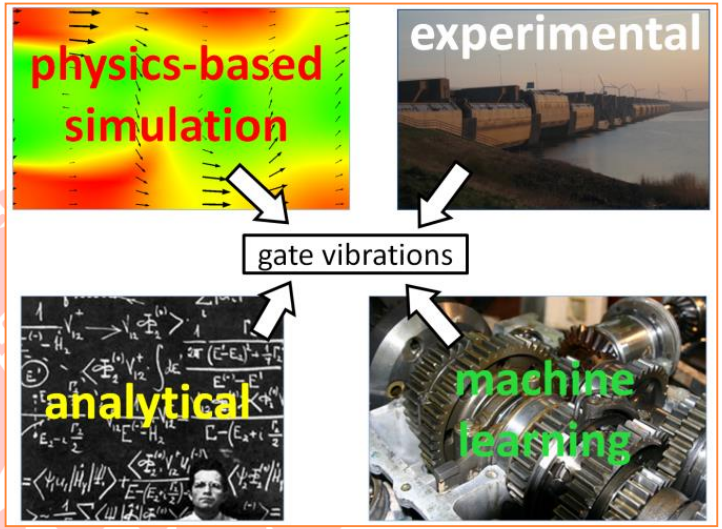
Big data science is dangerous but it is all around us!

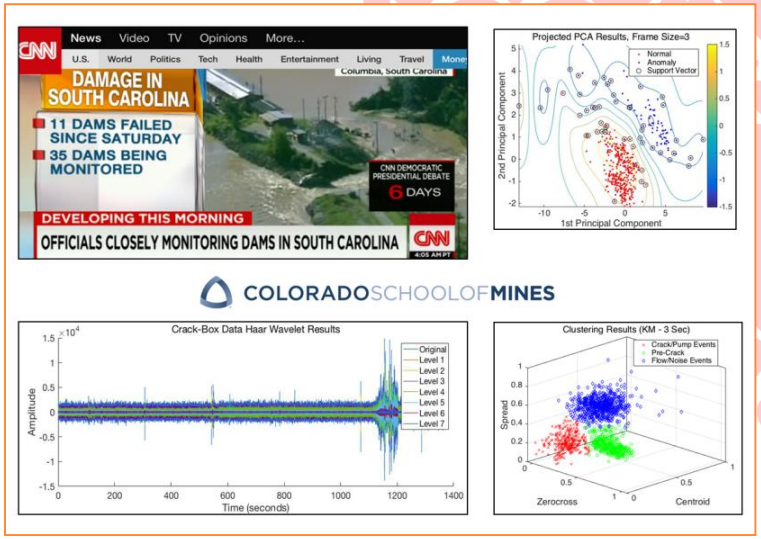
- Data science ➔
  - new methods and tools
  - to derive mathematical models from real-life observations
  - to detect anomalies
  - to predict the future

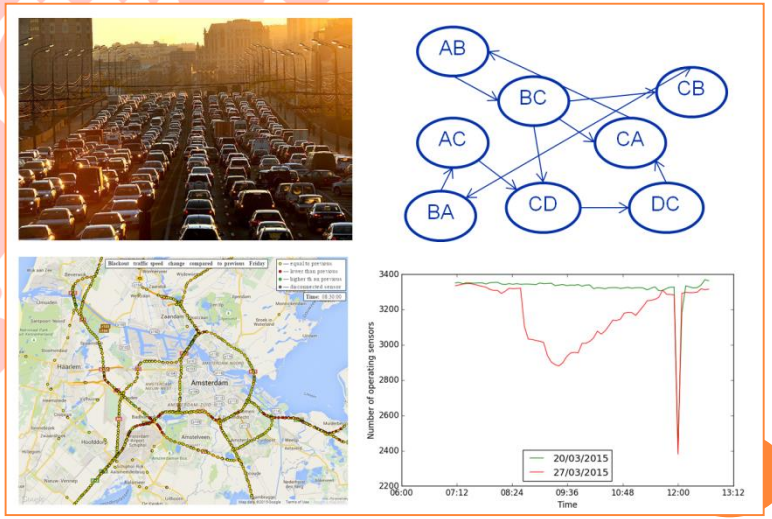
DATA SCIENCE + COMPUTATIONAL SCIENCE = ?



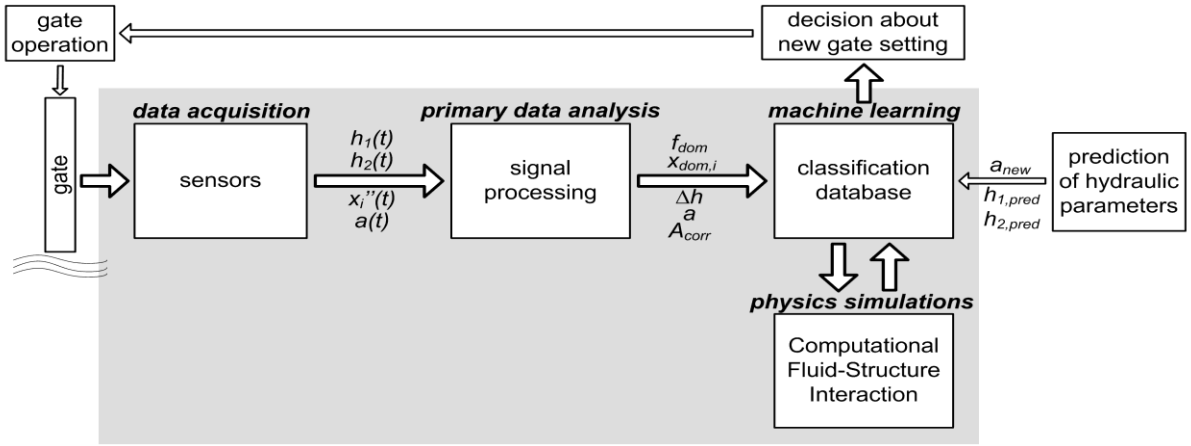

Cross section of dike with sensors
Fiber optics
Seismic (acoustic)
Water in cracks
Fiber optics
Inverted piezometers
Basin
Channel dike
Macro stability dike
UrbanFlood.eu

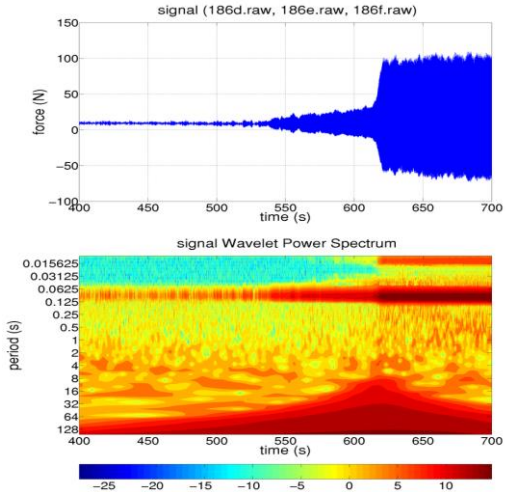

physics-based simulation
experimental
gate vibrations
analytical
machine learning


CNN
News Video TV Opinions More...
U.S. World Politics Tech Health Entertainment Living Travel
DAMAGE IN SOUTH CAROLINA
11 DAMS FAILED SINCE SATURDAY
35 DAMS BEING MONITORED
CNN DEMOCRATIC PRESIDENTIAL DEBATE
6 DAYS
DEVELOPING THIS MORNING
OFFICIALS CLOSELY MONITORING DAMS IN SOUTH CAROLINA
Projected PCA Results, Frame Size=3
Normal
Anomaly
Support Vector
2nd Principal Component
1st Principal Component
COLORADOSCHOOLOFMINES
Crack-Box Data Haar Wavelet Results
Amplitude
Time (seconds)
Original
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Level 7
Clustering Results (KM - 3 Sec)
Crack/Pump Events
Pre-Crack
Flow/Noise Events
Spread
Zerocross
Centroid

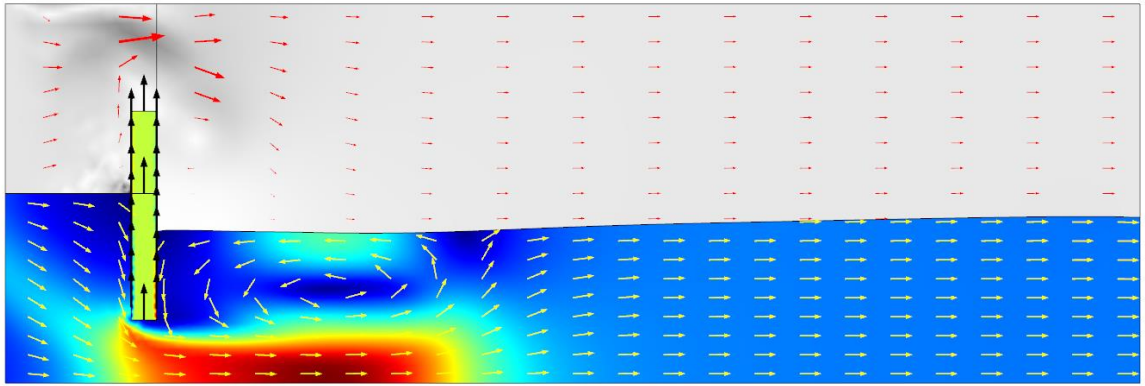

AB
BC
CB
AC
CA
BA
CD
DC
Amsterdam
Number of operating sensors
Time
20/03/2015
27/03/2015

Flood barrier gate vibrations

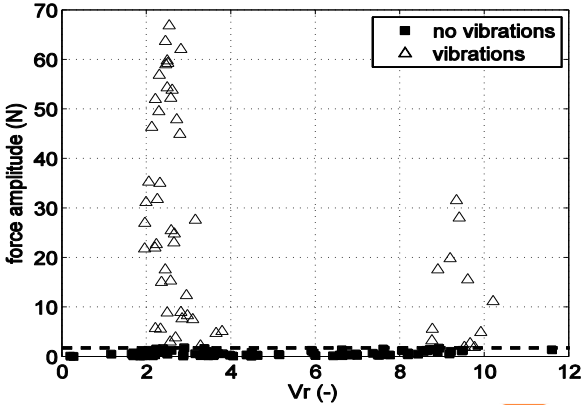




Simulation



Experiment



- Christiaan Erdbrink

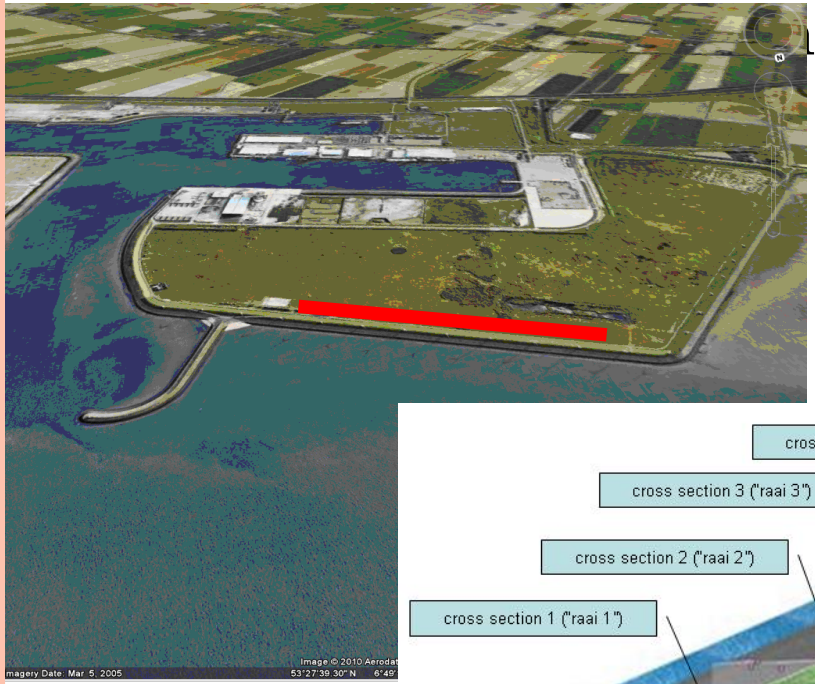
ITMO, UvA, Deltares

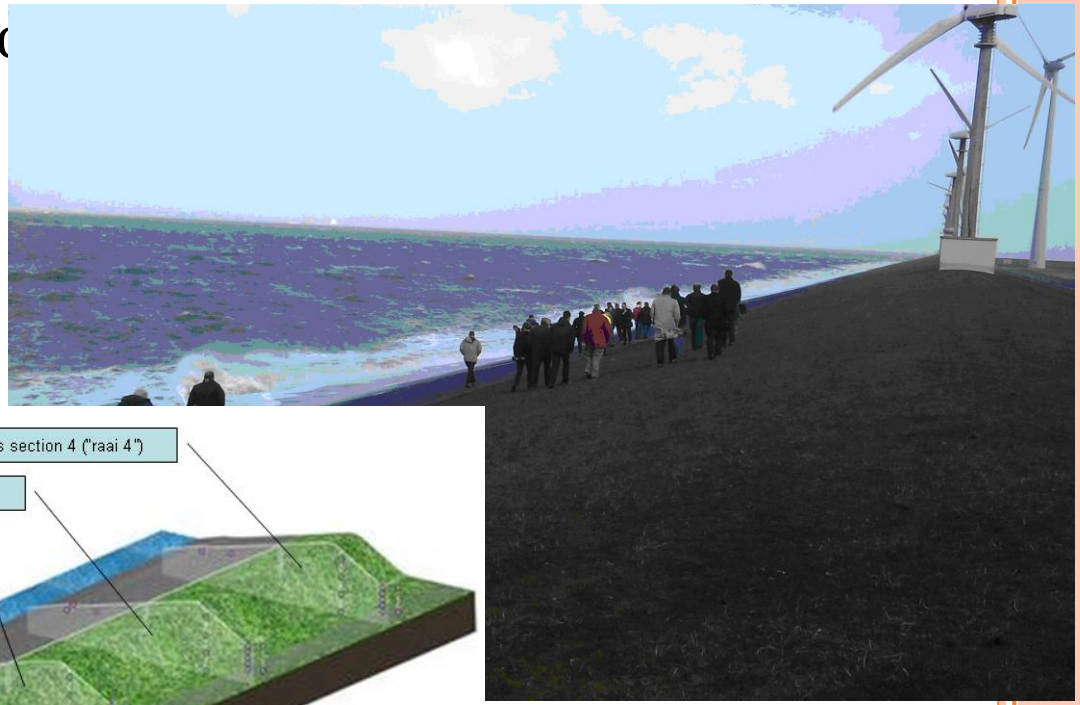
C.D. Erdbrink, V.V. Krzhizhanovskaya, P.M.A. Sloot, "Reducing cross-flow vibrations of underflow gates: Experiments and numerical studies", Journal of Fluids and Structures, V. 50, 2014, pp. 25-48, doi: 10.1016/j.jfluidstructs.2014.06.010

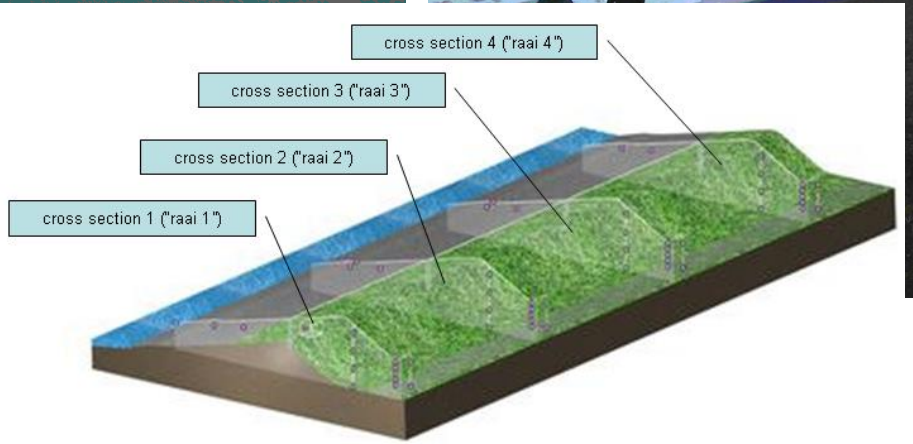
C.D. Erdbrink, V.V. Krzhizhanovskaya, "Differential evolution for system identification of self-excited vibrations", Journal of Computational Science, V. 10, 2015, pp. 360-369. doi: 10.1016/j.jocs.2015.03.004

Data-driven levee modeling and anomaly detection

- 2D full-scale simulation of porous flow & structural dynamics

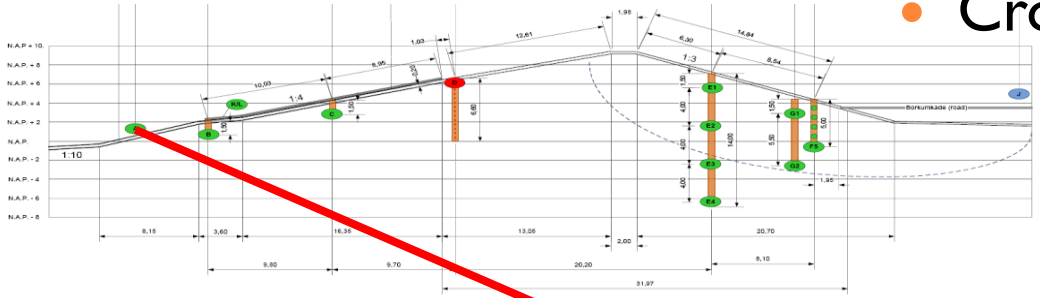




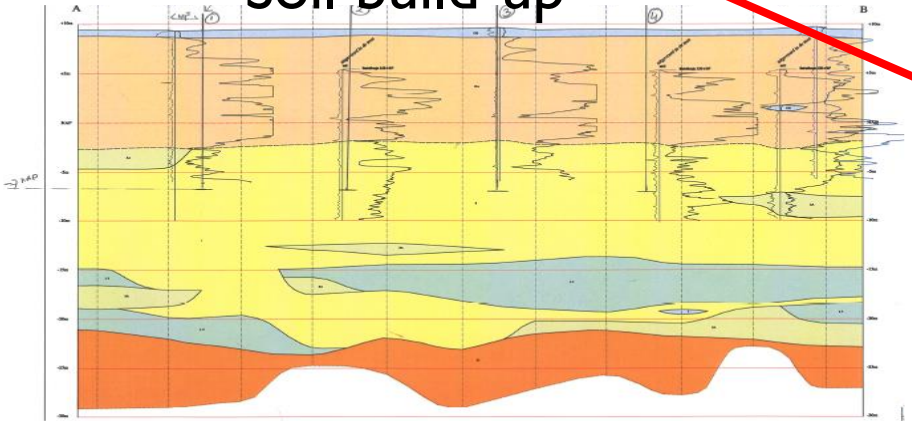


INPUT DATA

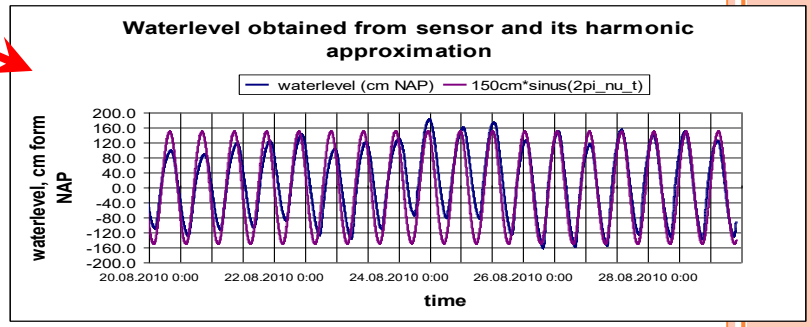
- Cross section geometry



- Soil build-up

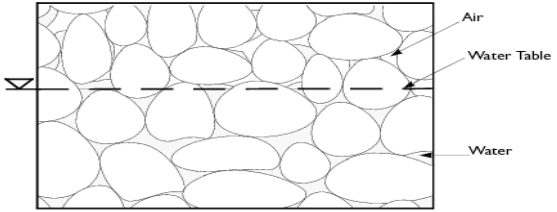


- Water level sensor signal

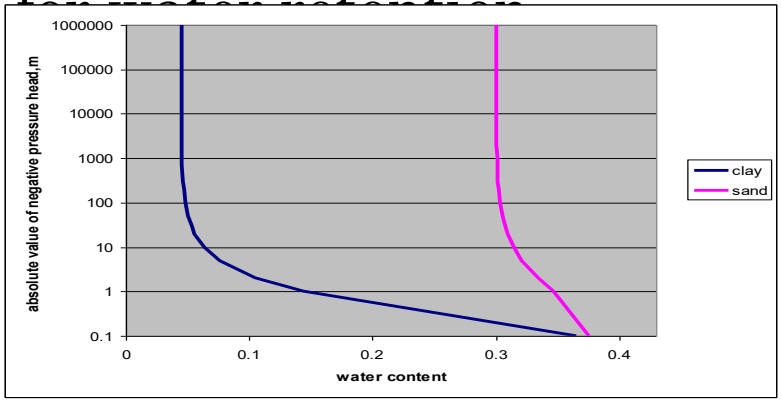


Simulation components (1)

- Porous flow:
  - Richards' eq. for water flow in partially saturated soils
  - Van Genuchten model for water retention



partially saturated soil



$$(C+\theta_e S)\frac{\partial p}{\partial t}+\nabla\cdot[-K_S k_r \nabla(p+\rho g z)]=\frac{\partial \varepsilon}{\partial t}$$

FSI coupling terms

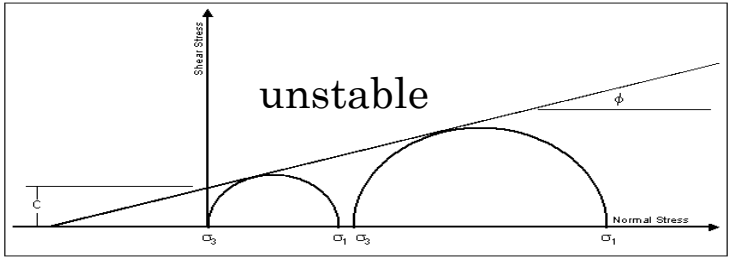
SIMULATION COMPONENTS (2)

- Structural stability:
  - Plane stress
  - Linear elastic soil
  - Mohr-Coulomb failure criterion

$$\tau_f = c + \sigma_f \tan\phi$$

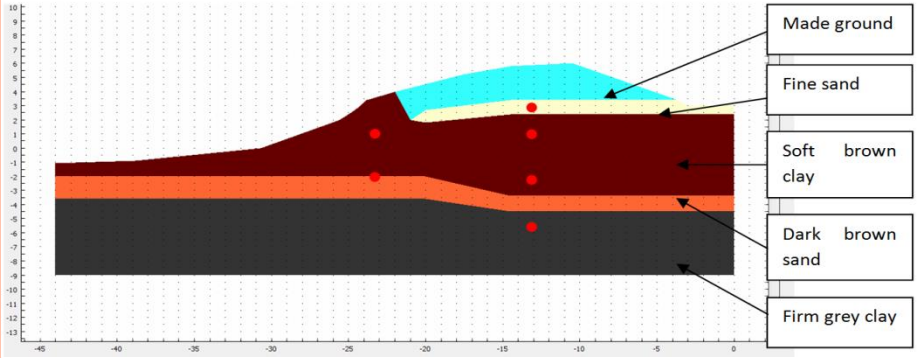
FSI coupling terms

$$\begin{cases} \nabla \cdot \underline{\sigma} - \nabla p + \rho_s g = 0 \\ \underline{\sigma} = \lambda \varepsilon \underline{E} + 2\mu \underline{\varepsilon} \end{cases}$$

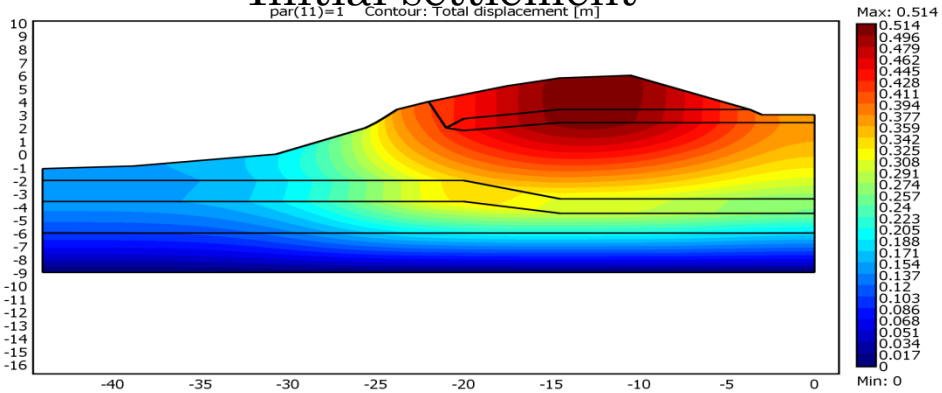


Virtual Dike: Boston

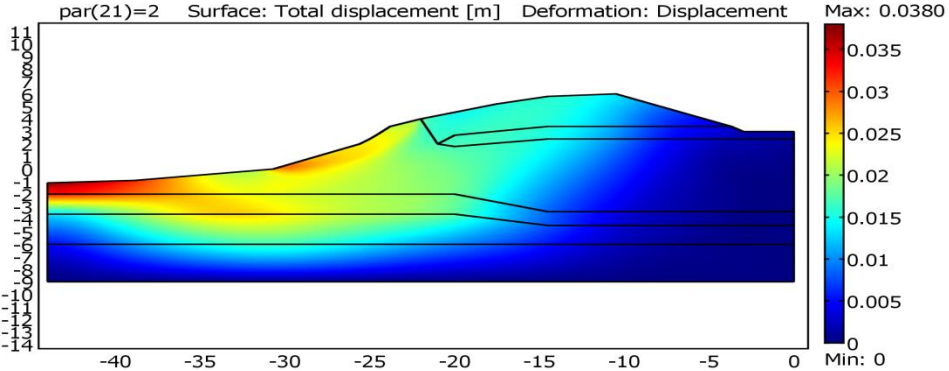
Real layers



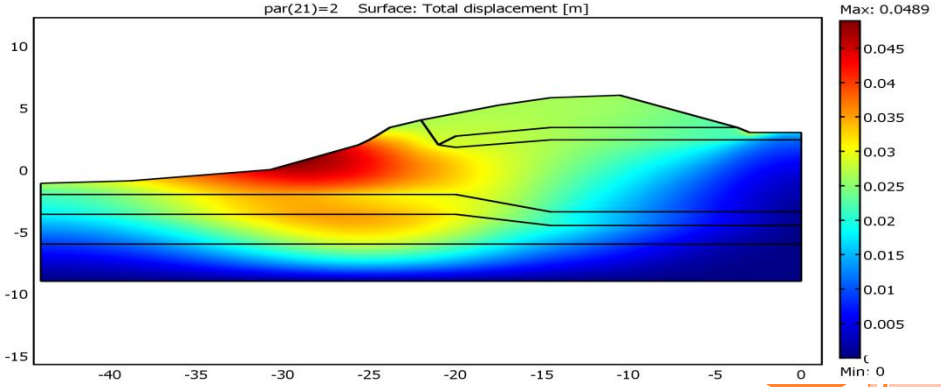
Initial settlement



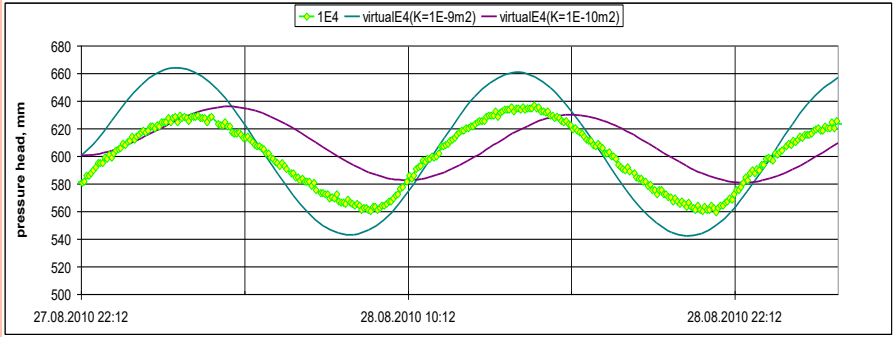
Low tide

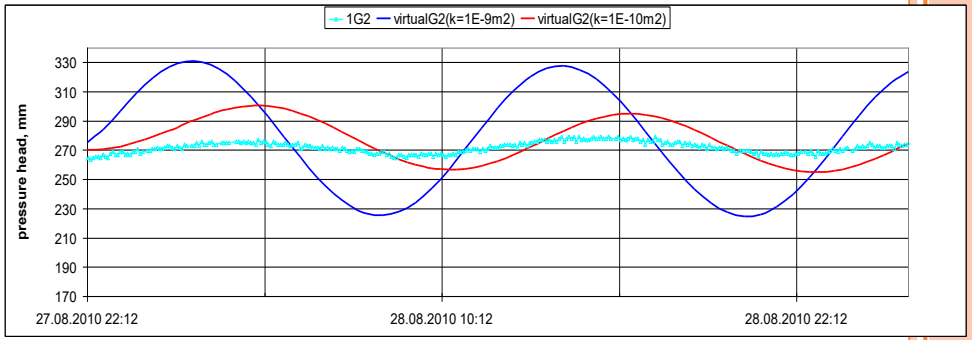


High tide

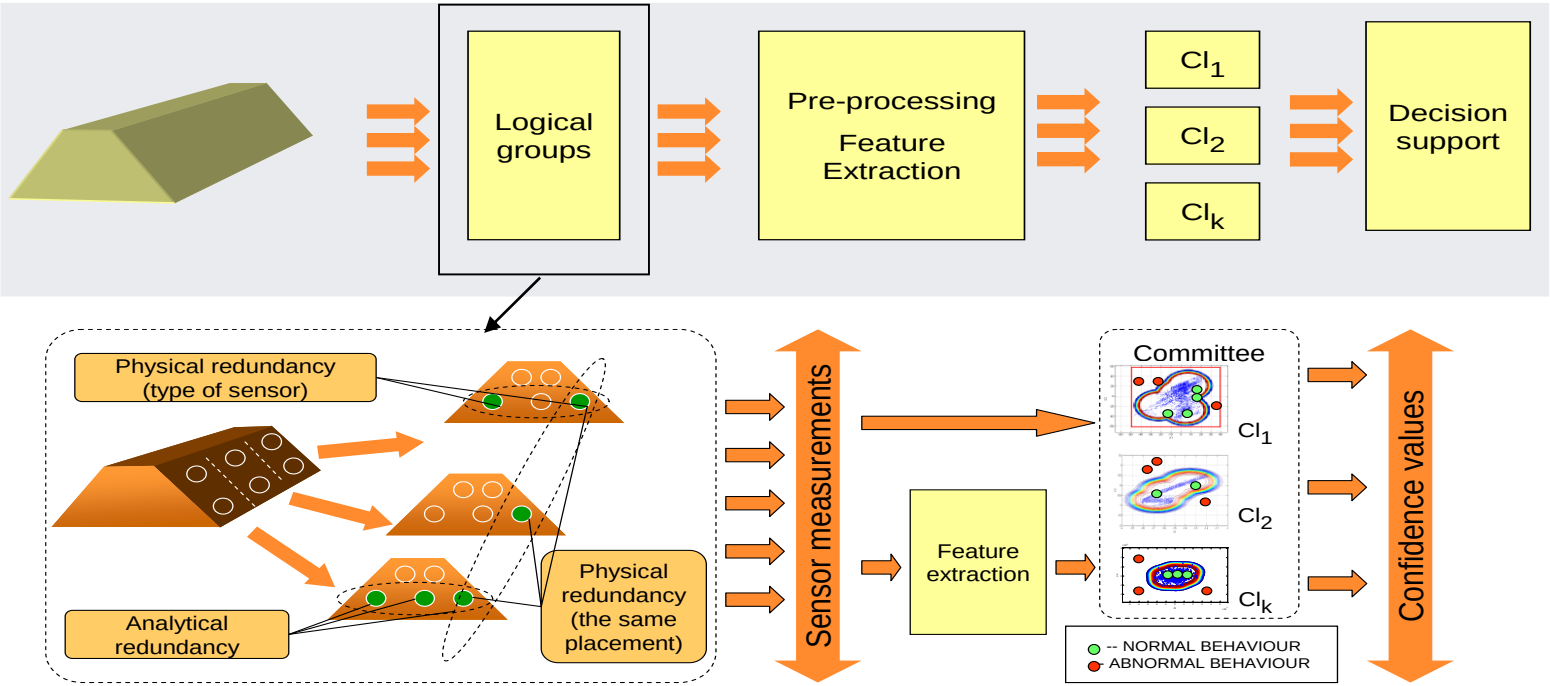


Model calibration based on data

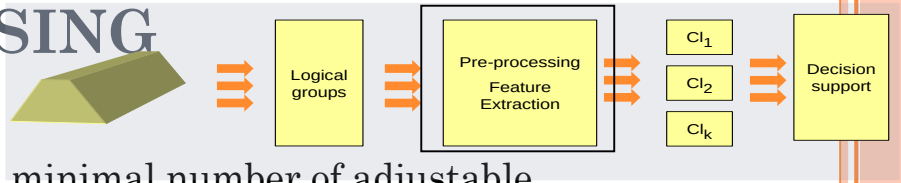




Anomaly Detection Approach



PRE-PROCESSING



The main idea: to apply methods, which have minimal number of adjustable parameters and require minimal information about signal

1) Wavelet denoising

$y \longrightarrow$ DWT $\xrightarrow{\text{Coefficients}}$ Hard thresholding $\longrightarrow$ iDWT $\longrightarrow x$

2) Spectrum Singular Analysis (SSA)

$y \longrightarrow$ SSA $\longrightarrow y = \sum_{n=1}^{N} c_n \longrightarrow$ Take only c which corresponds to maximal eigenvalues $\longrightarrow x$

3) Hodrick-Prescott filter

$$\min \sum_{i=1}^{T} (y_i - x_i)^2 + \lambda \sum_{t=2}^{T-1} [(x_{i+1} - x_i) - (x_i - x_{i-1})]^2, \ \lambda - smoothing\ parameter$$

4) L1 trend filtering

$$\min \sum_{i=1}^{T} (y_i - x_i)^2 + \lambda \sum_{i=2}^{T} |x_i - x_{i-1}|, \ \lambda - smoothing\ parameter$$

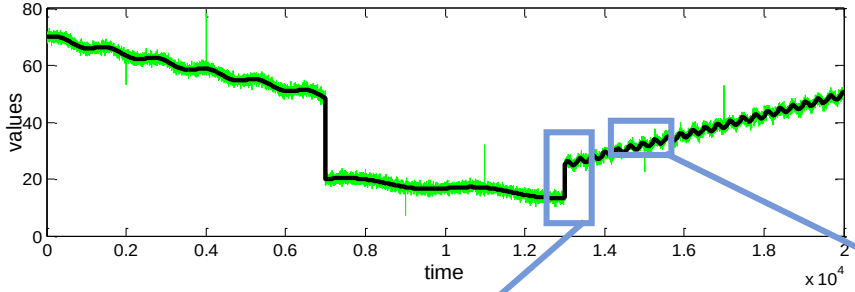
5) Moving average

$$x_i = \frac{1}{N} \sum_{l=0}^{N-1} y_{i-l}$$

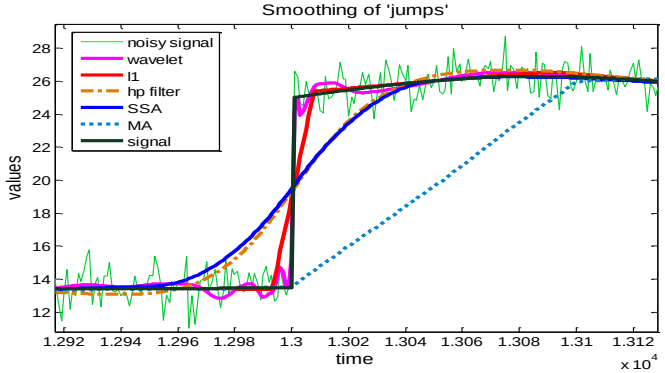
y – measurements
x – estimation of a signal

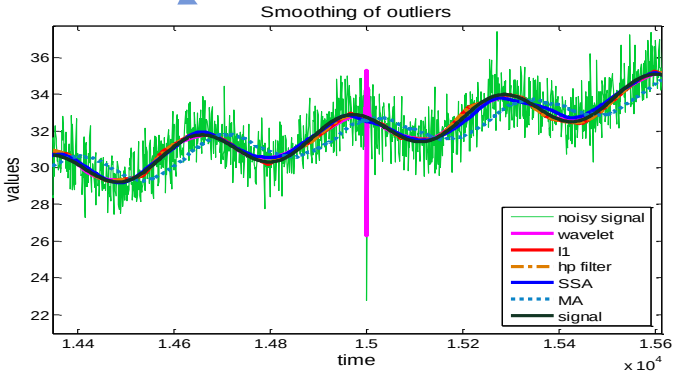
DATA PRE-PROCESSING

Synthesized data

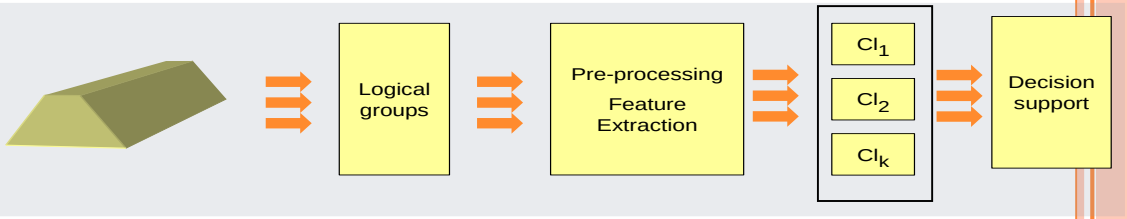


	Wavelet	*l1*	*hp*	*SSA*	*MA*
RMSE	**0.23**	0.24	0.49	0.50	1.18

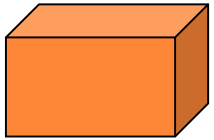




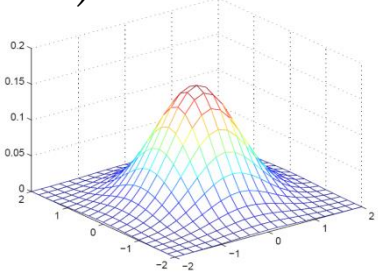
ONE-SIDE CLASSIFICATION



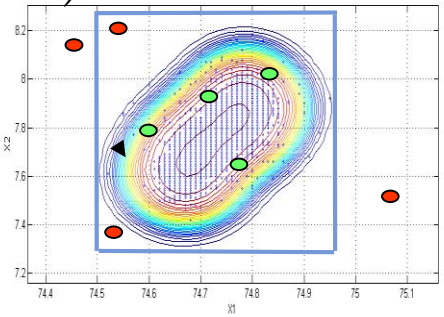
1) Hypercube

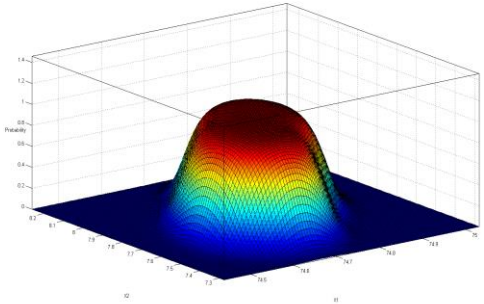


2) Parzen window



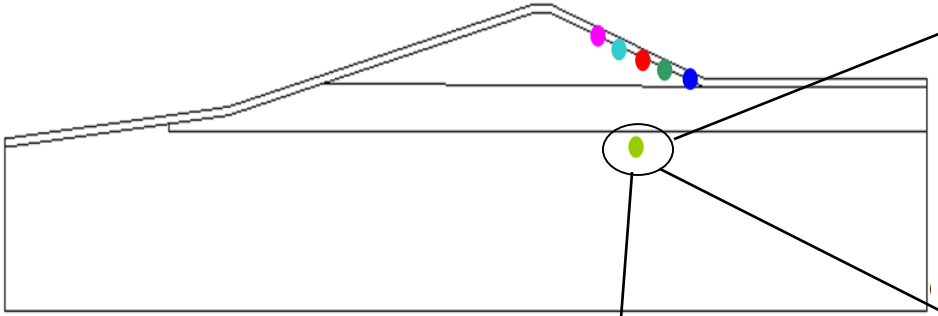
3) Neural Clouds

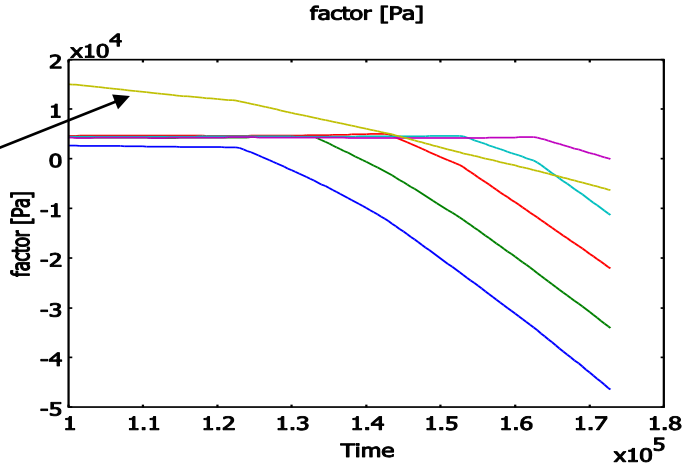


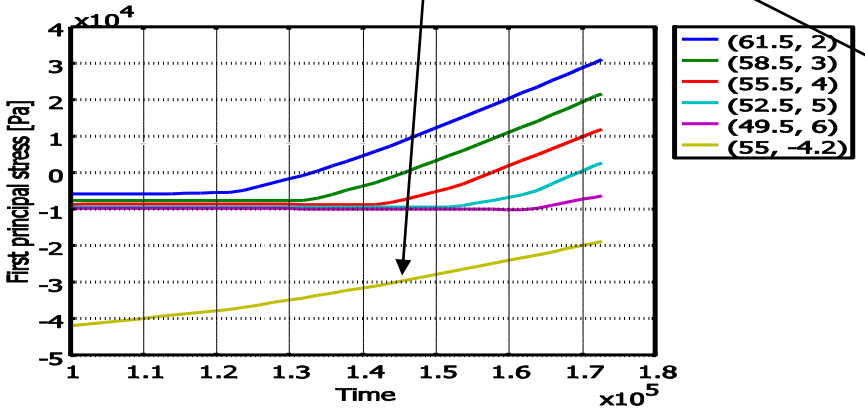


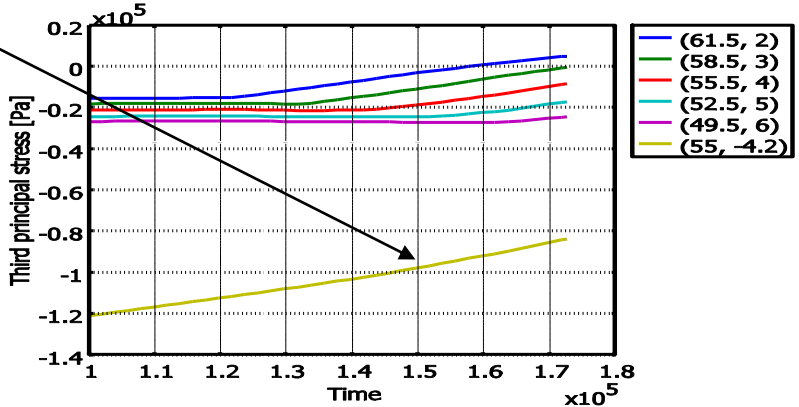
--ABNORMAL BEHAVIOUR
-- NORMAL BEHAVIOUR

Simulated sensor data

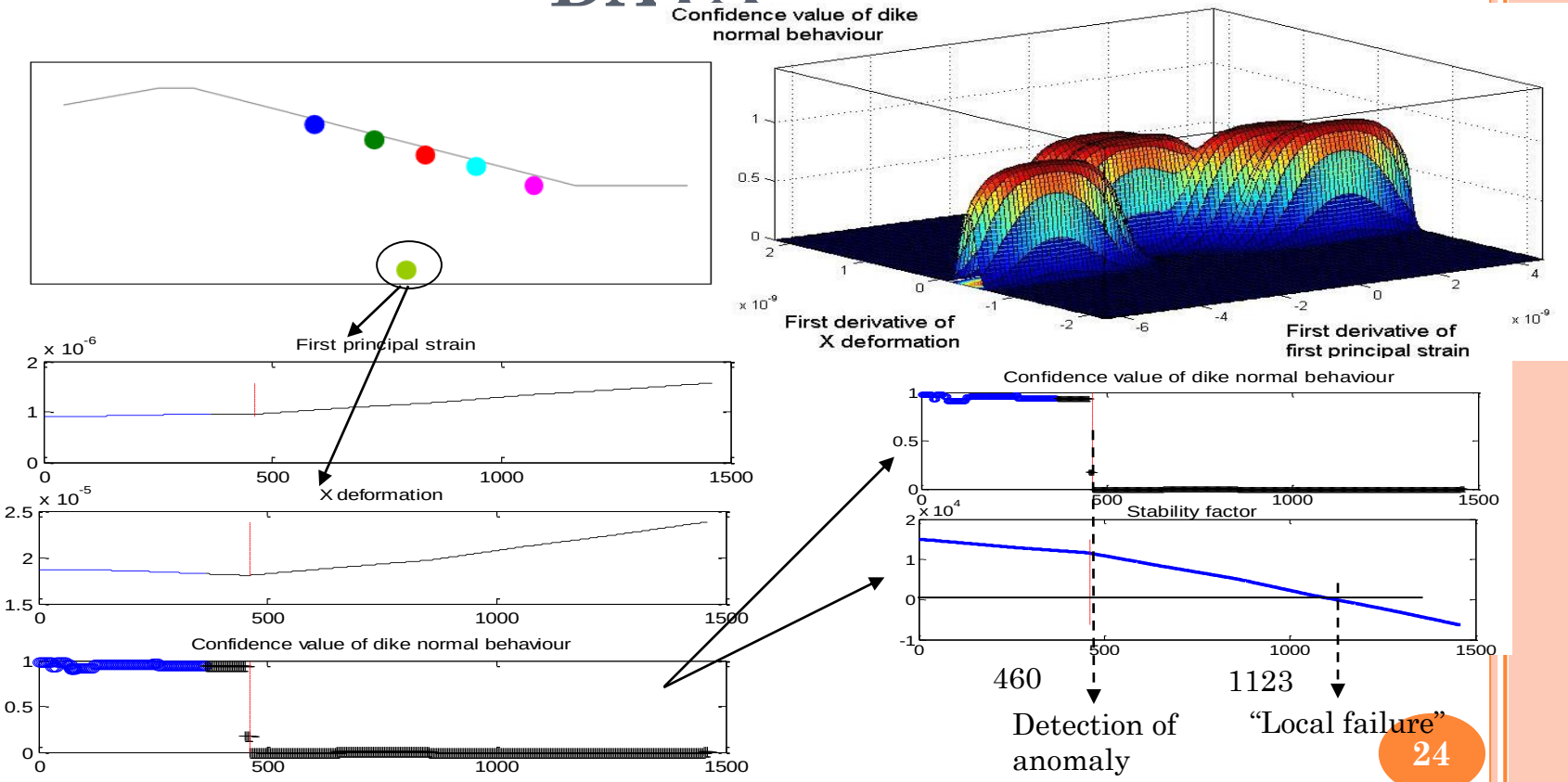




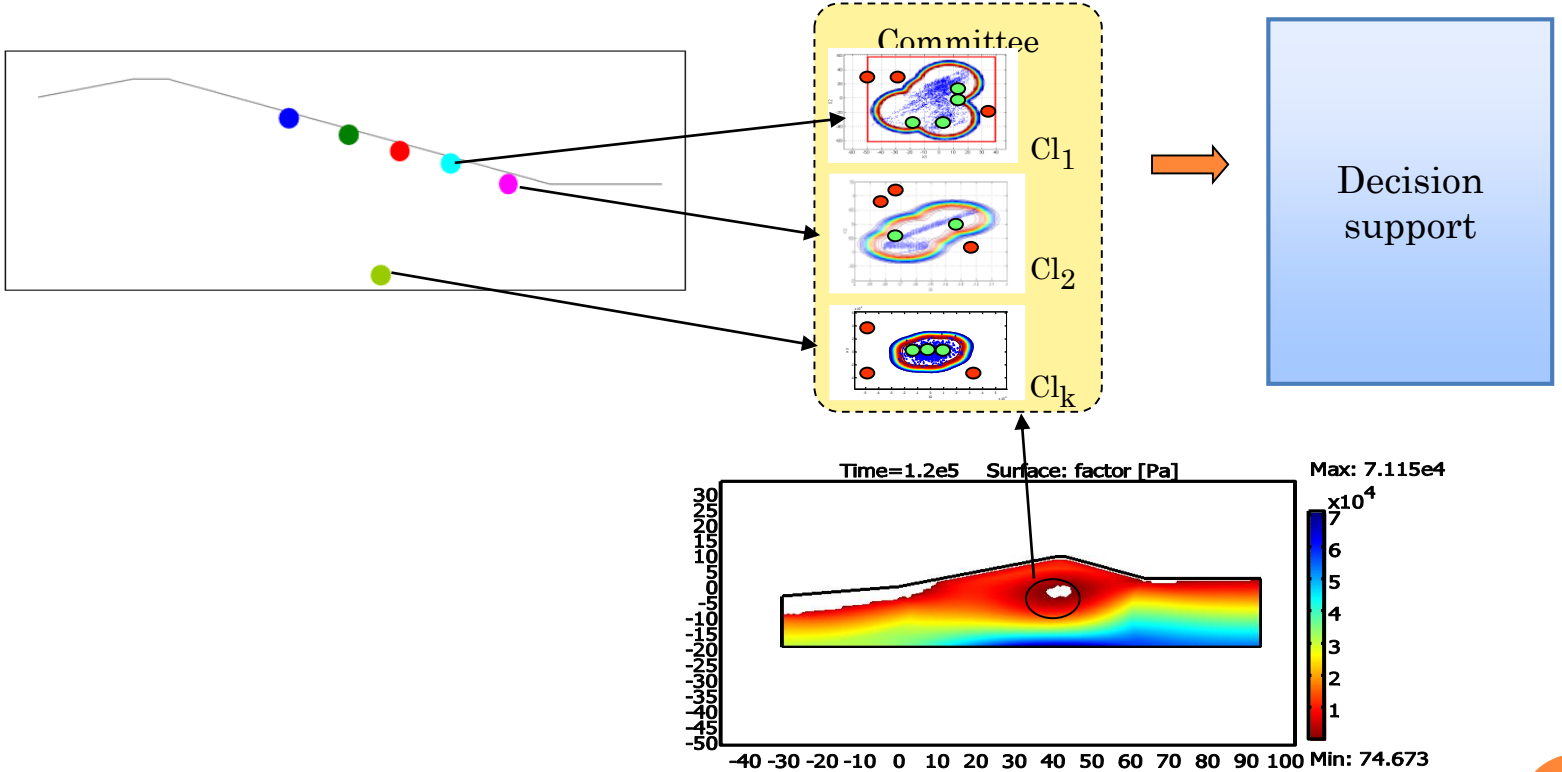




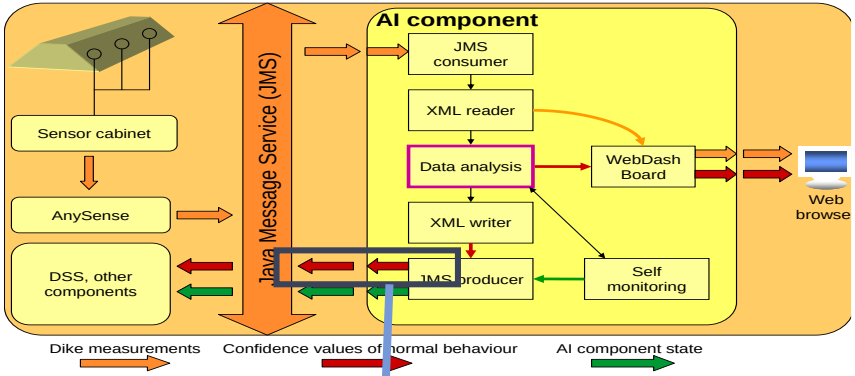
Anomaly detection in artificial data

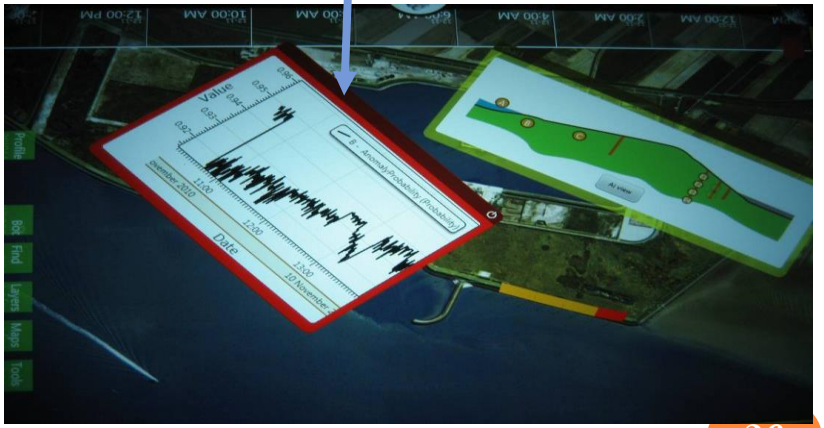


Combining AI with Virtual Dike



AI component as part of the UrbanFlood Early Warning System





Krzhizhanovskaya et al. 2011 " Flood early warning system: design, implementation and computational modules", Procedia Computer Science, V. 4, pp.106-115, doi: 10.1016/j.procs.2011.04.012

- Alexander Pyayt

- Natalia Melnikova

Pyayt et al. 2015 "Combining Data-Driven Methods with Finite Element Analysis for Flood Early Warning Systems", Procedia Computer Science, V. 51, 2015, pp. 2347-2356, doi: 10.1016/j.procs.2015.05.404

Melnikova et al. 2015 “Experience of using FEM for real-time flood early warning systems: Monitoring and modeling Boston levee instability”, Journal of Computational Science, V. 10, 2015, pp. 13-25, doi: 10.1016/j.jocs.2015.04.033.

Melnikova et al. 2015 "Numerical prediction of the IJkDijk trial embankment failure", Proceedings of the Institution of Civil Engineers - Geotechnical Engineering, V. 168, Issue 2, 2015, pp. 158-171, doi: 10.1680/geng.14.00040

LATEST WORK ON MONITORING LEVEES

Photo from www.cnn.com, South Carolina flooding: Dams breached, more trouble ahead, Oct 7, 2015 Update

Earth Dams and Levees

- Internal Erosion
- Worldwide Problem

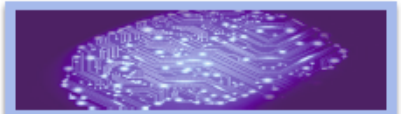
- Approach: Non-intrusive(!) monitoring

Tunbridge Dam, Tasmânia, Australia, 11/28/2008, Source: Jeffery Farrar (2008)

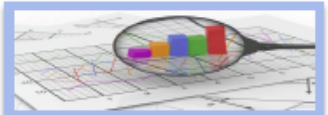
Our approach

Data Collection
- Experimental Levees
  Crack-Box Testbed
  Piping Experiment
- Passive Seismic Data

Preprocessing
- Wavelet Denoising
- Spectral Frames
- Feature Extraction
- Standardization
- Automatic Feature Selection



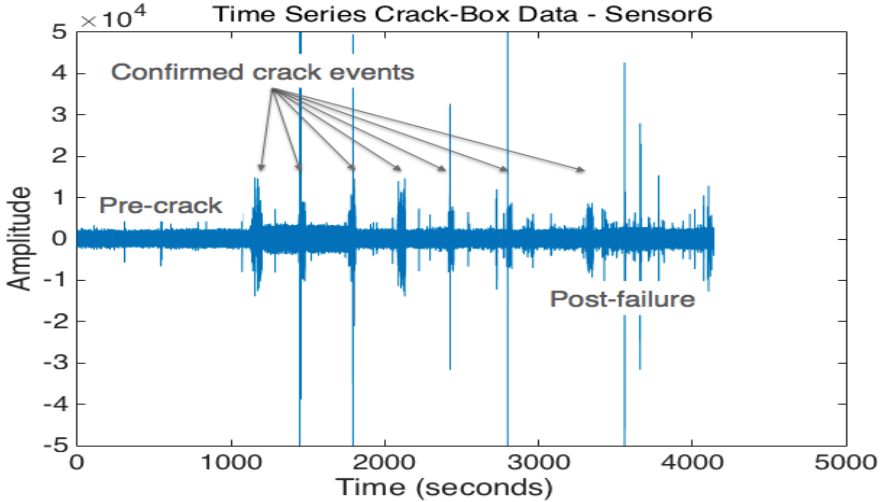
Machine Learning
- Classification
- Anomaly Detection

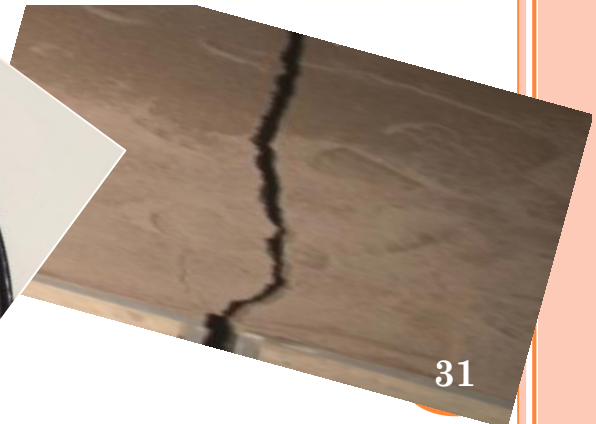


Results
- Visualization
- Validation

Passive Seismic Data

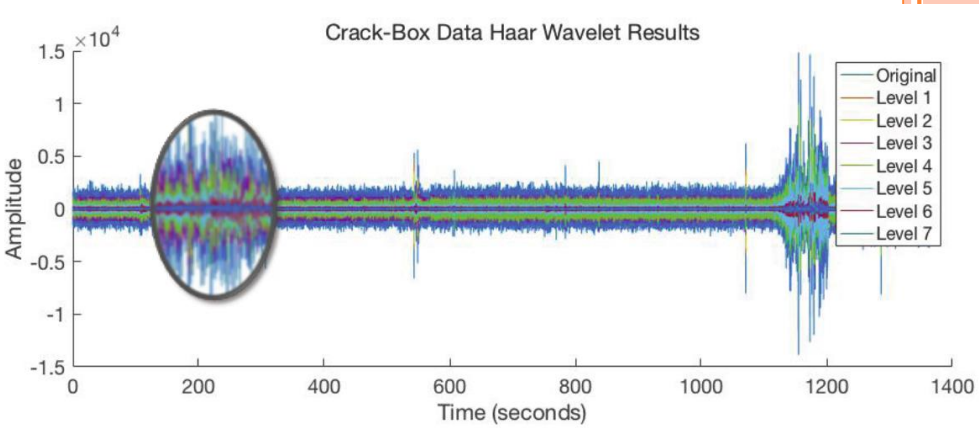
- Crack-Box & Piping Experiments
  - Internal Erosion and Cracking
  - Geophones
  - 4,140 Seconds of Data



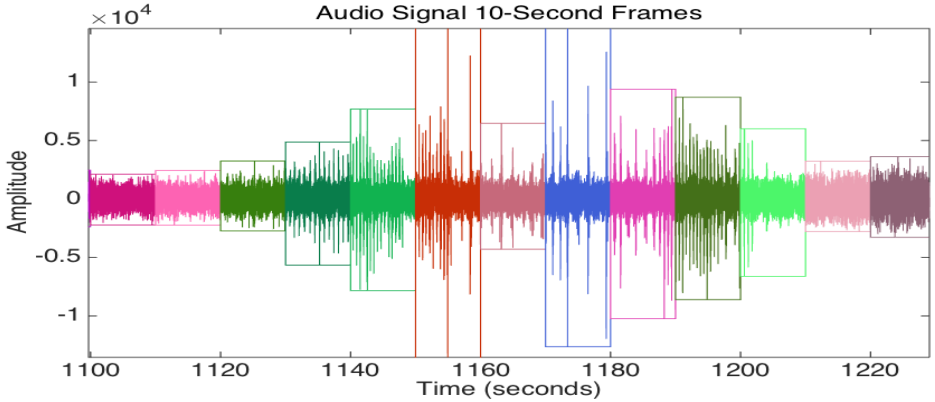


Data Preprocessing

1. **Wavelet de-noising: MODWT, Haar level 3**

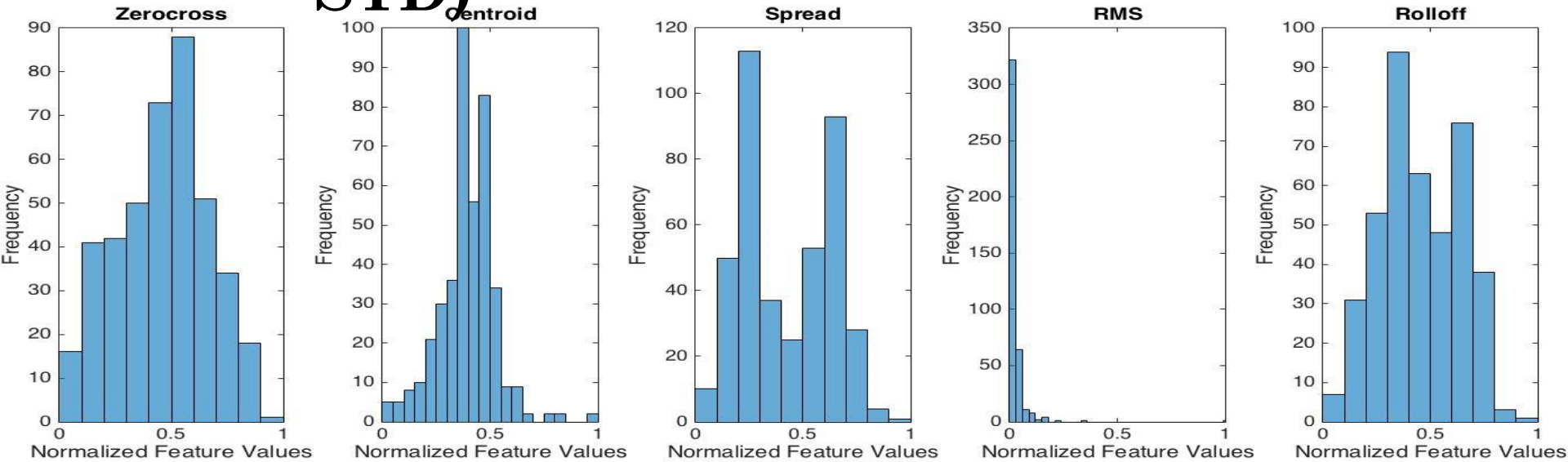


2. **Spectral frames (1, 2, 3, 5, and 10-second frames)**



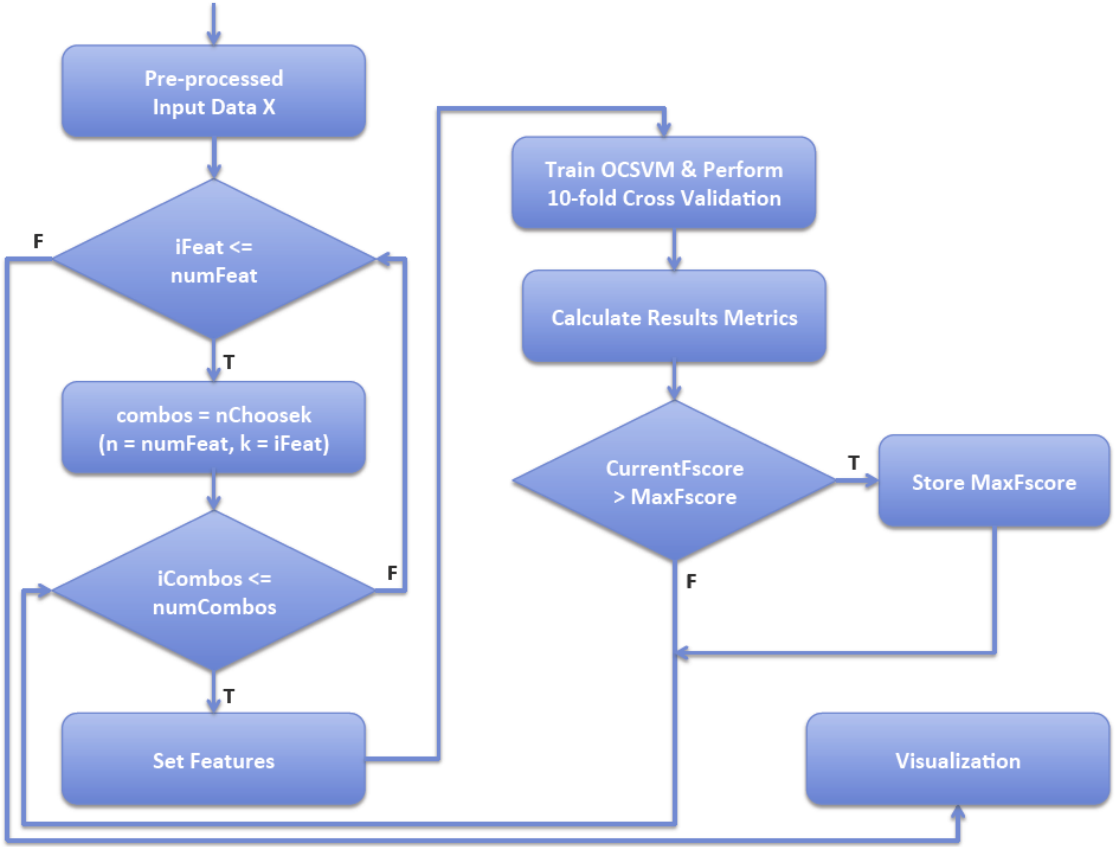
3. **Feature extraction**

3. **Feature extraction**
4. **Standardization (zero mean, unit STD)**

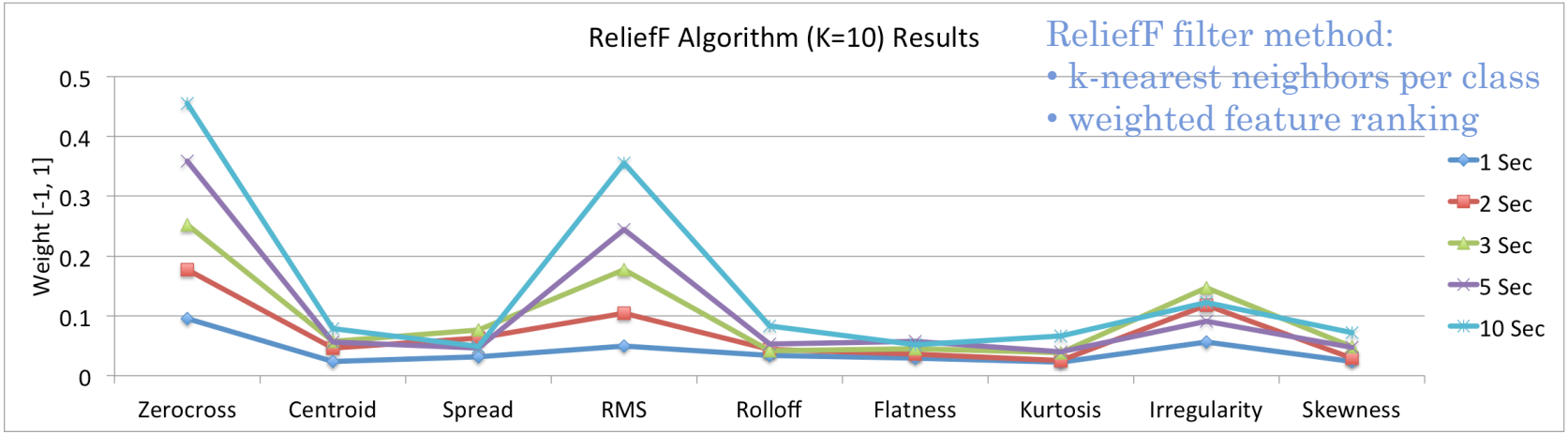


+ Flatness, Kurtosis, Irregularity, Skewness

5. Automatic feature selection



SELECTED FEATURES



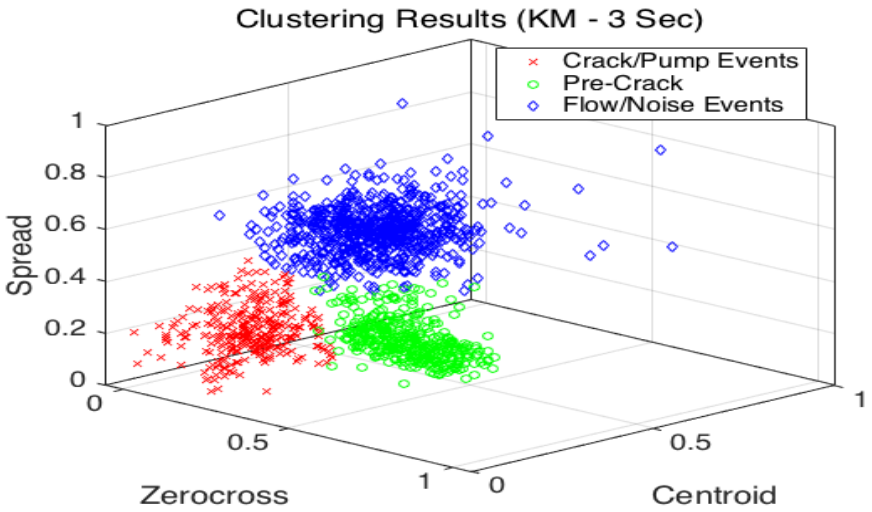
ReliefF filter method:
- k-nearest neighbors per class
- weighted feature ranking

Crack	1 Second	2 Second	3 Second	5 Second	10 Second
Accuracy	0.957	0.961	0.965	0.949	0.986
F1-score	0.973	0.976	0.978	0.968	0.991
Best # of features	2	2	2	2	2
Selected features	RMS, SP	RMS, KU	RMS, KU	RMS, RG	RMS, FL

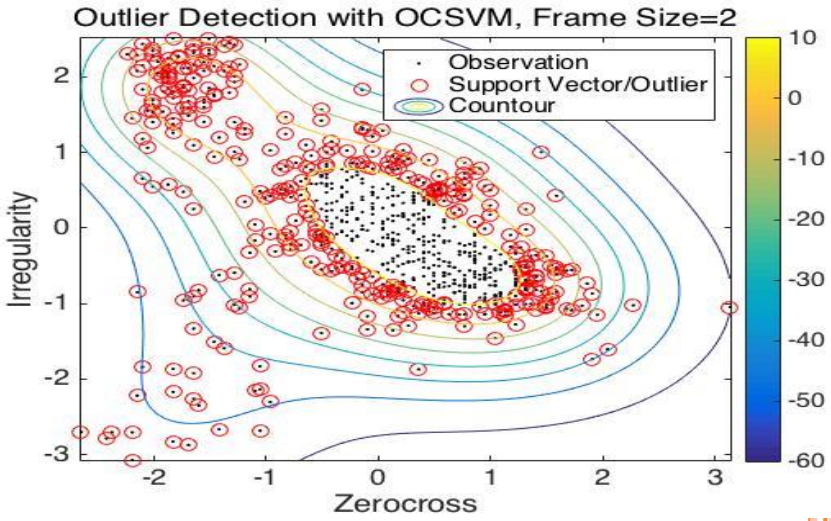
Piping	1 Second	2 Second	3 Second	5 Second	10 Second
Accuracy	0.940	0.950	0.935	0.919	0.951
F1-score	0.963	0.969	0.960	0.950	0.970
Best # of features	2	2	2	3	3
Selected features	ZC, RMS	ZC, RMS	ZC, RMS	ZC, RMS, RG	ZC, RMS, RG

Machine Learning

- Clustering



- Anomaly Detection

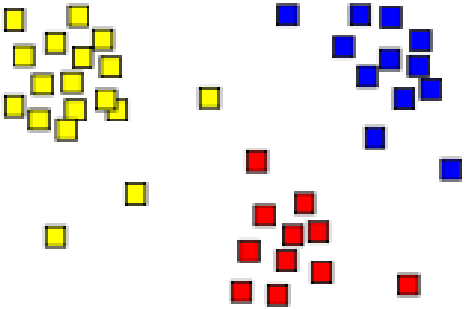


Stages of Clustering

- Pattern Representation
- Distance Measure
- Grouping Membership
- Cluster Representation
- Validation Techniques

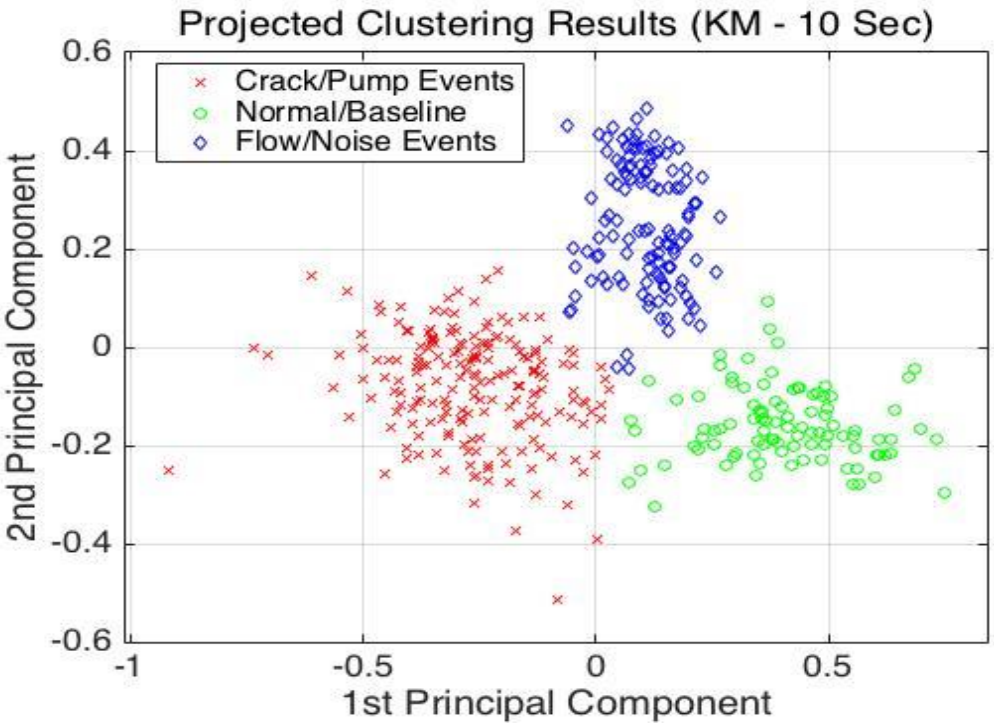
Clustering Algorithms

- K-Means Clustering
- Partitioning Around Medoids
- Fuzzy C-Means Clustering
- Gaussian Mixture Model
- Agglomerative Hierarchical Clustering



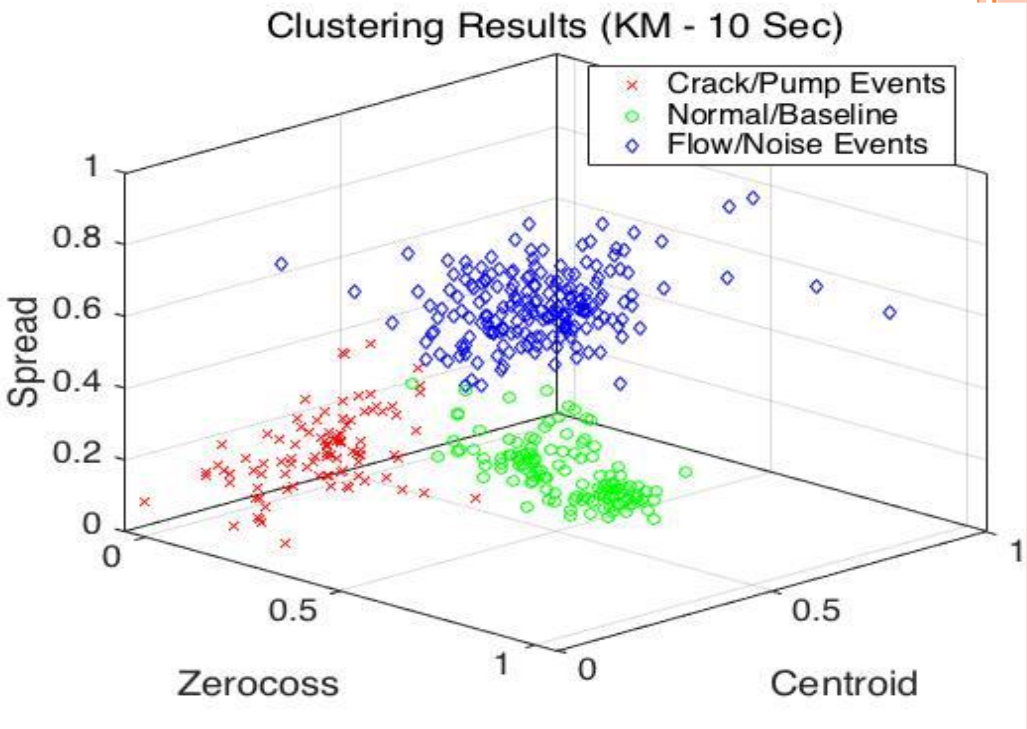
Cluster Representation

- Principal Component Analysis
  - K-means
  - K=3
  - 10-Second Frames



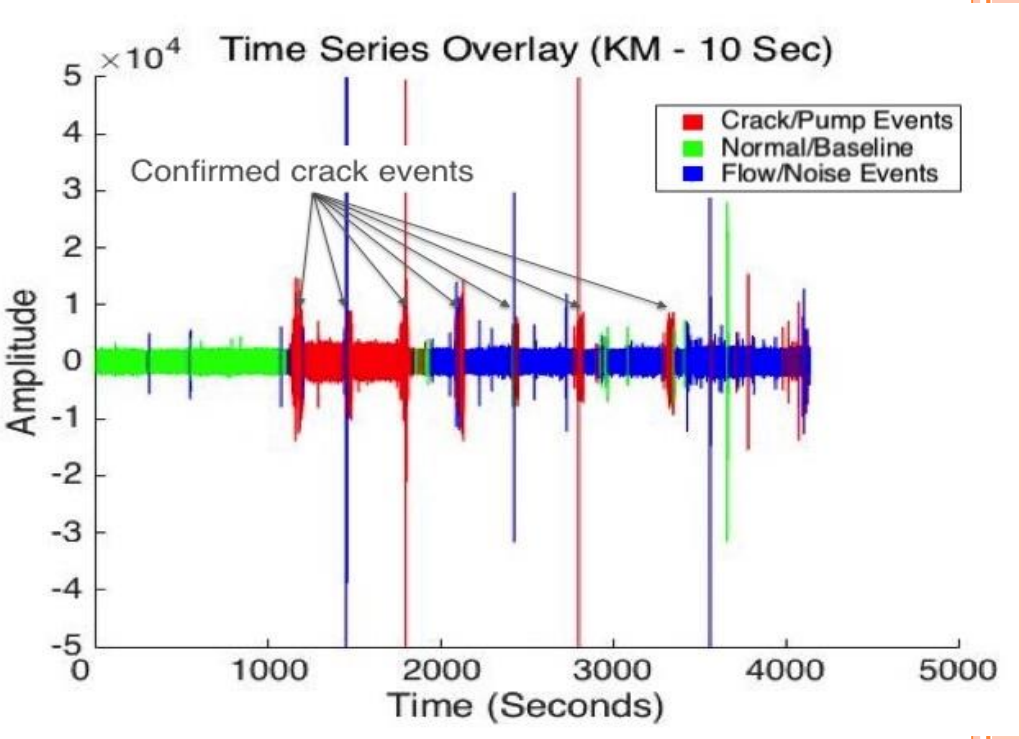
Cluster Representation

- 3D Plot of Selected Features
  - K-means
  - K=3
  - 10-Second Frames



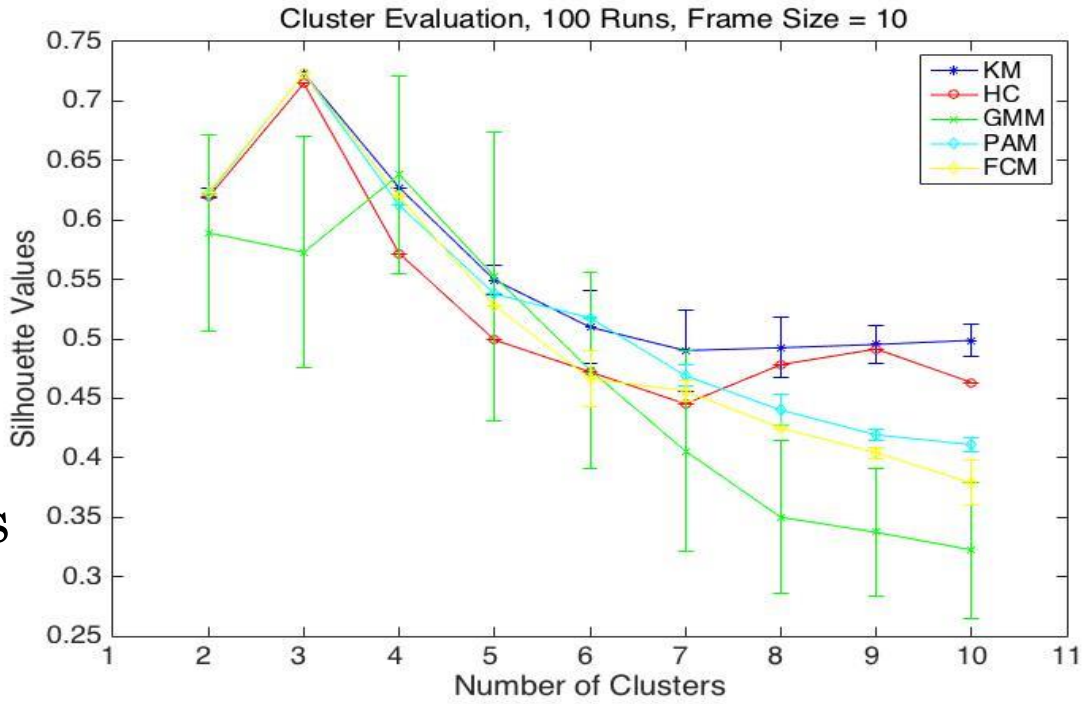
Cluster Representation

- Time Series Cluster Color Overlay
  - K-means
  - K=3
  - 10-Second Frames



Internal Validation

- Silhouette Width
  - Range: [-1, 1]
  - Averaged over 100 Runs
  - 95% CI
  - 10-Second Frames



External Validation

- Purity
  - Range: [0, 1]
  - Averaged over 100 Runs
  - Comparing all 5 Algorithms

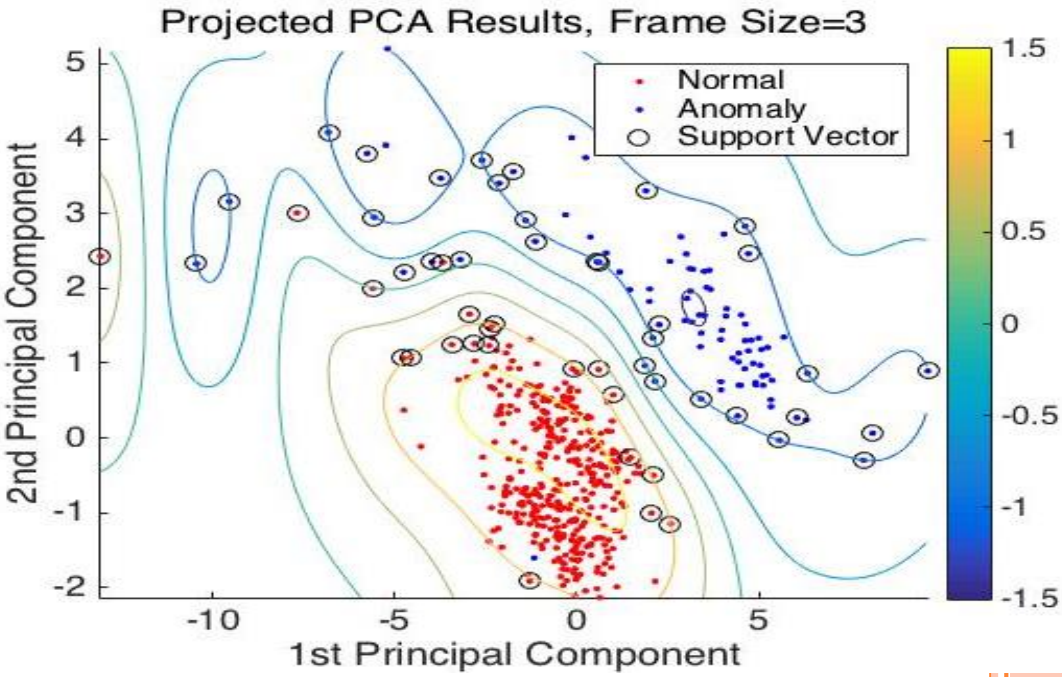
4 out of 5 Algorithms Performed well (> 82%)

Frame Size	Obs.	KM	HC	GMM	PAM	FCM
3 Seconds	1380	0.830	0.827	0.715	0.825	0.826
5 Seconds	828	0.825	0.829	0.728	0.823	0.826
10 Seconds	414	0.831	0.838	0.765	0.829	0.831

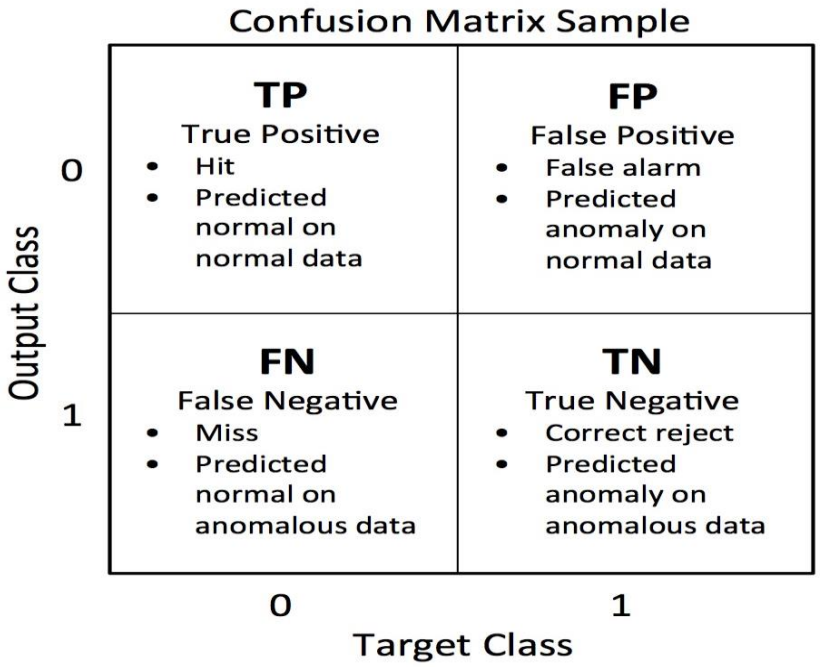
GMM is less accurate

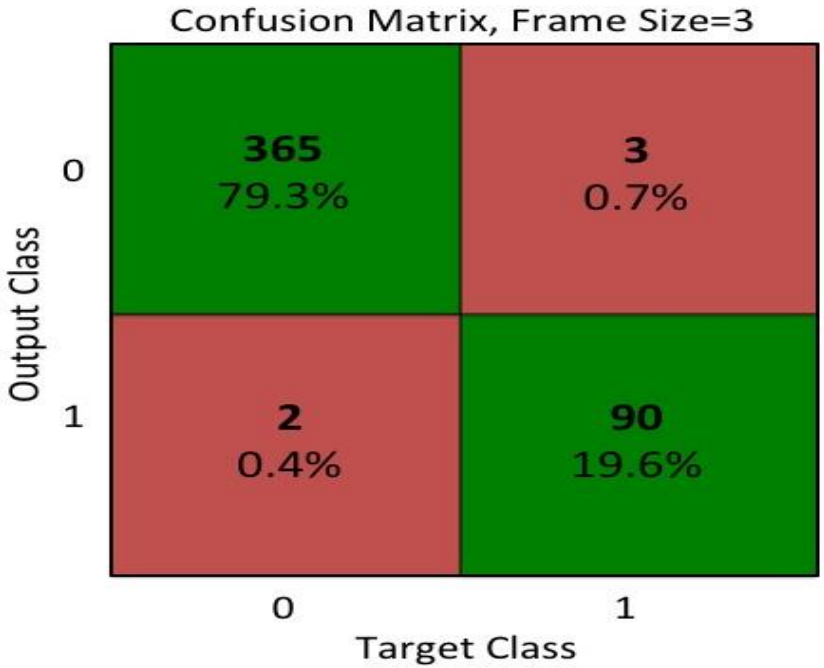
Two-Class Support Vector Machine

- Nine Features
- RBF Kernel
- Cross Validation
- 97%+ Accuracy



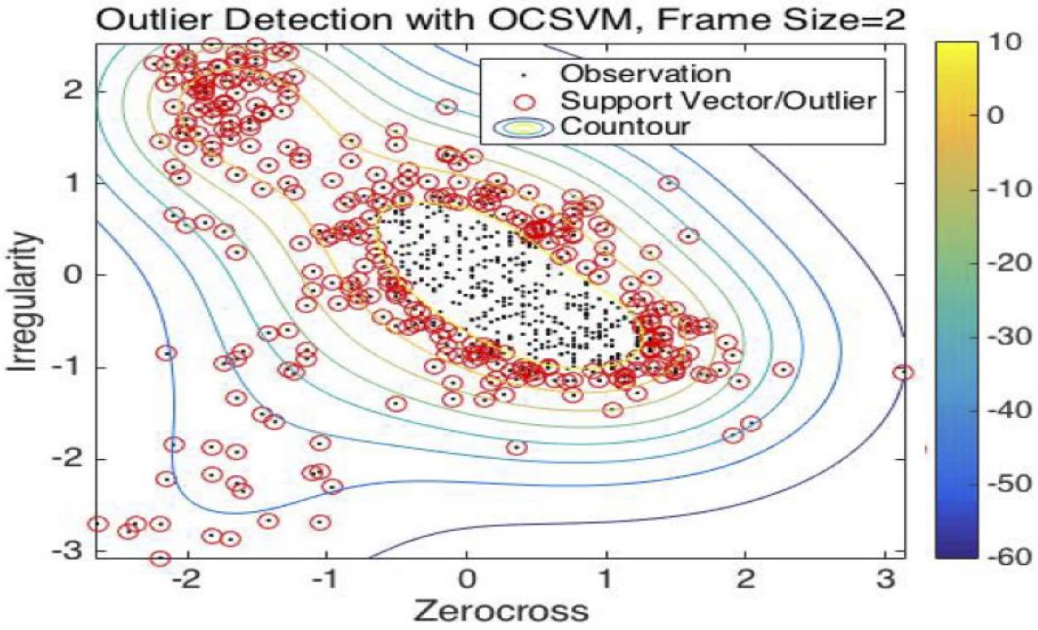
Prediction Result: 97%+ Accuracy



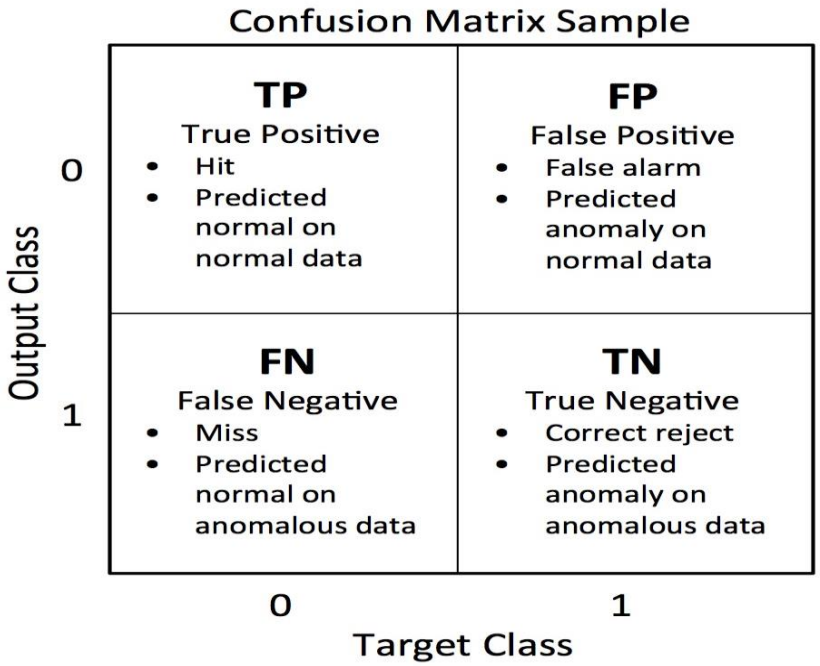


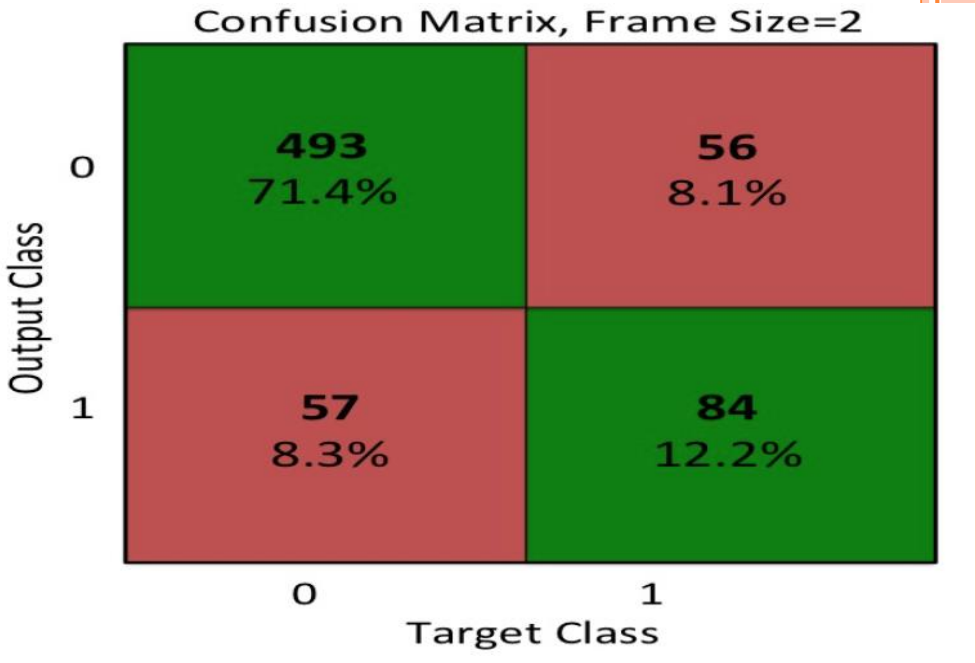
One-Class Support Vector Machine

- Two Features
- RBF Kernel
- Cross Validation
- 83%+ Accuracy



PREDICTION RESULT: 83%+ ACCURACY





Automatic feature selection

- Improved accuracy

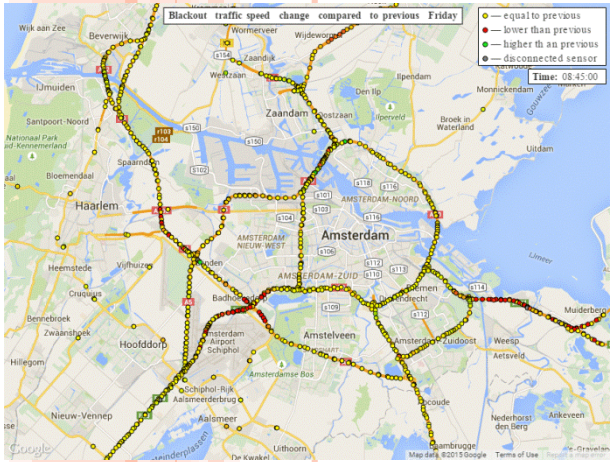
  from 83% ➔ to over 91%

- Wendy Fisher

- Tracy Camp

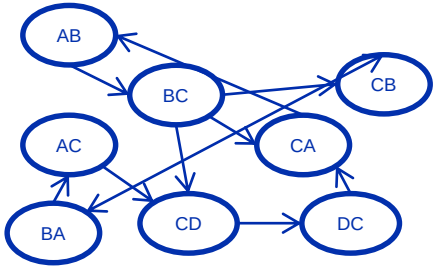
W. Belcher, T. Camp, V. V. Krzhizhanovskaya, "Detecting Erosion Events in Earth Dam and Levee Passive Seismic Data with Clustering", 2015, doi: 10.1109/ICMLA.2015.9

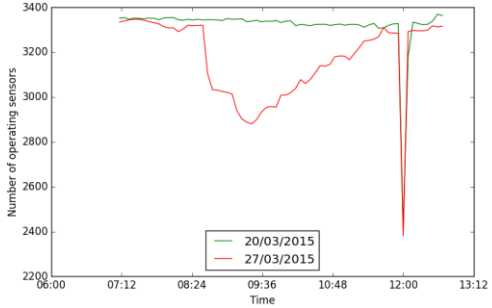
W. Fisher, T. Camp, V. Krzhizhanovskaya, "Crack Detection in Earth Dam and Levee Passive Seismic Data Using Support Vector Machines", 2016, doi: 10.1016/j.procs.2016.05.339



Skip?

Data-driven traffic simulation





DATA

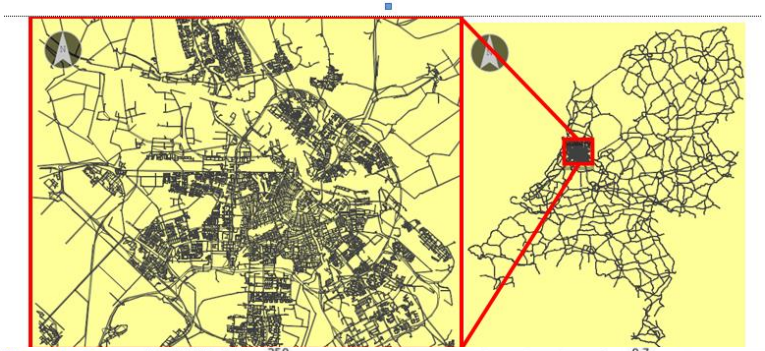
National Data Warehouse for Traffic Information

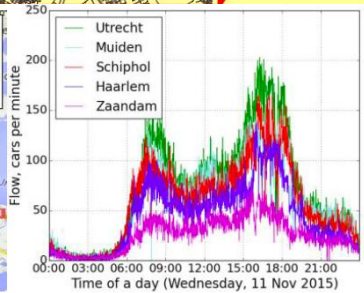
- NDW = National Data Warehouse for traffic information
- 40,000 sensors in the Netherlands

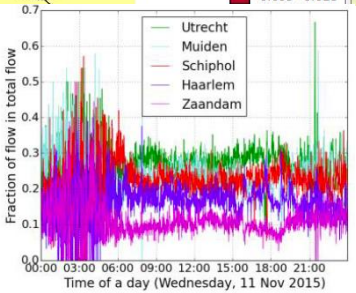
Vehicle length categories, meters		
Set 1		**Set 2**
1,85 – 2,4		… – 5,6
2,4 – 5,6		
5,6 – 11,5		5,6 – 12,2
11,5 – 12,2		
12,2 – …		12,2 – …

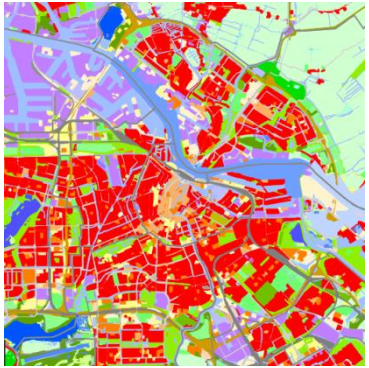
Data types	
Traffic data	**Status data**
Traffic flow (intensity)	Road works
Speed	Incidents
Travel time	Bridge/dynamic lanes status

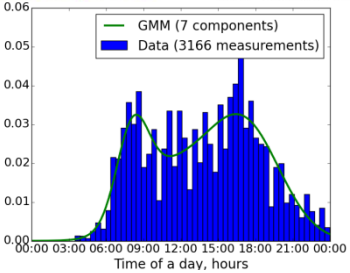
Other sources of data

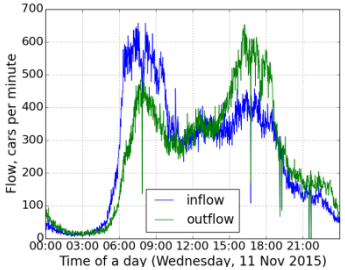






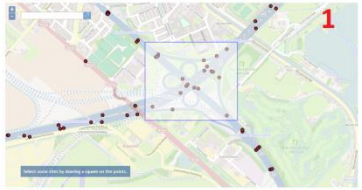






Blackout Friday traffic

[1] Airport Schiphol



[2] Amsterdam Slotervaart

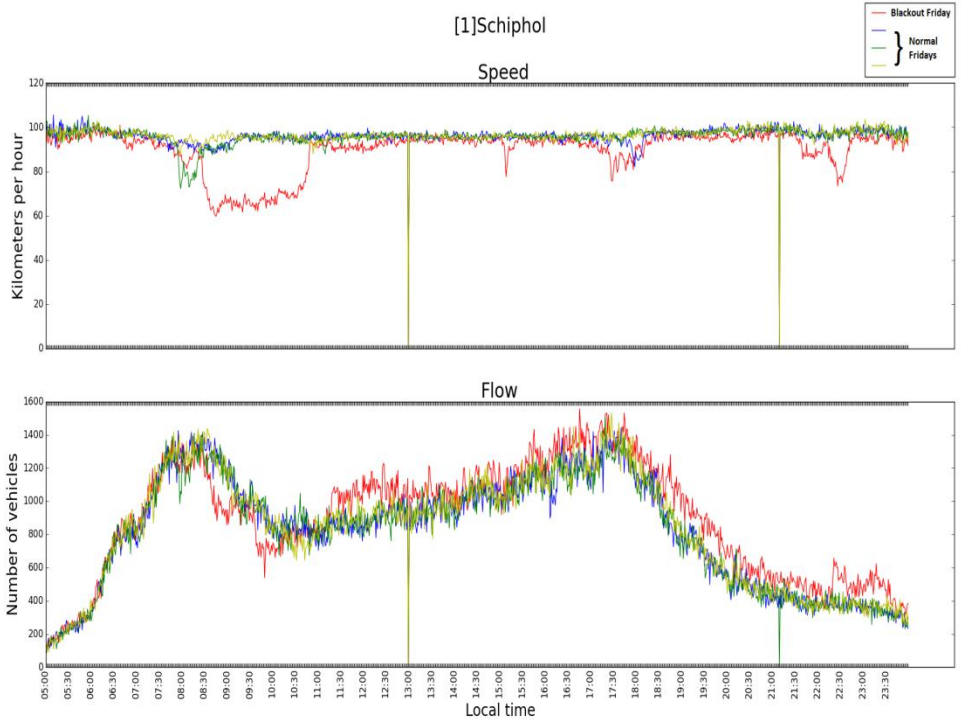
[3] Amsterdam Sloterdijk

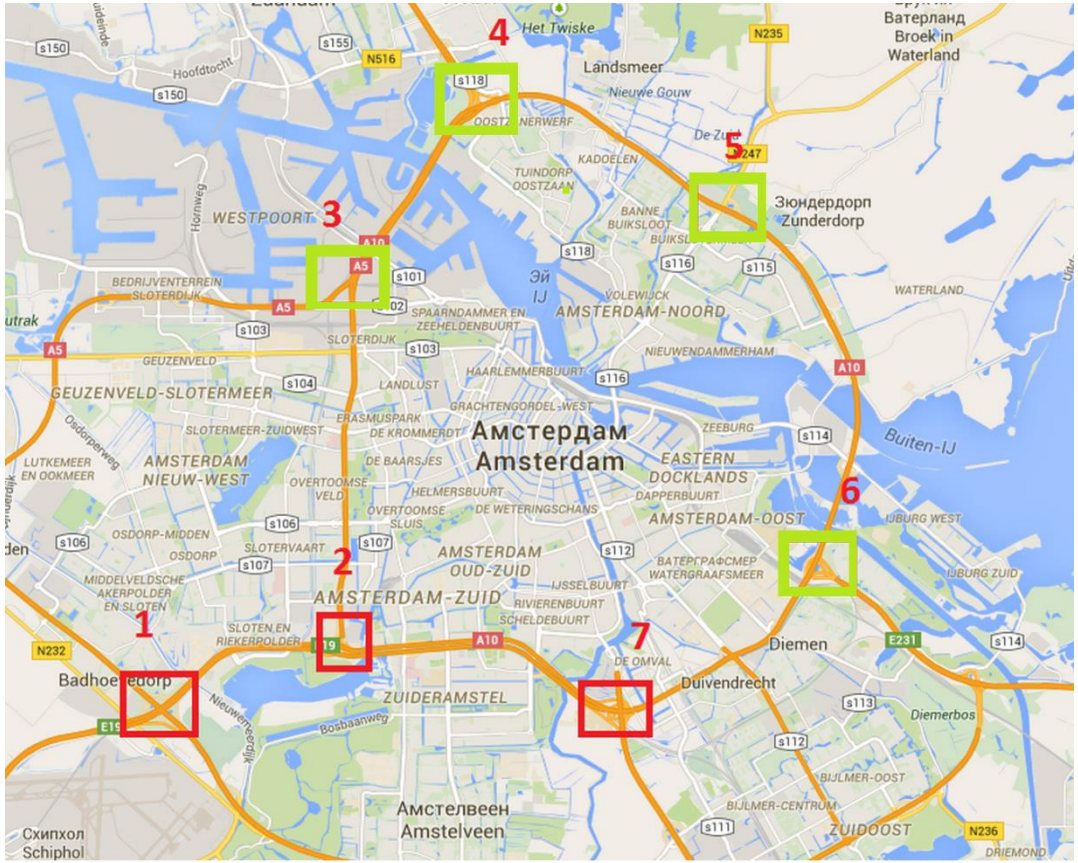
[4] Amsterdam Tuindorp Oostzaan

[5] Amsterdam Noord

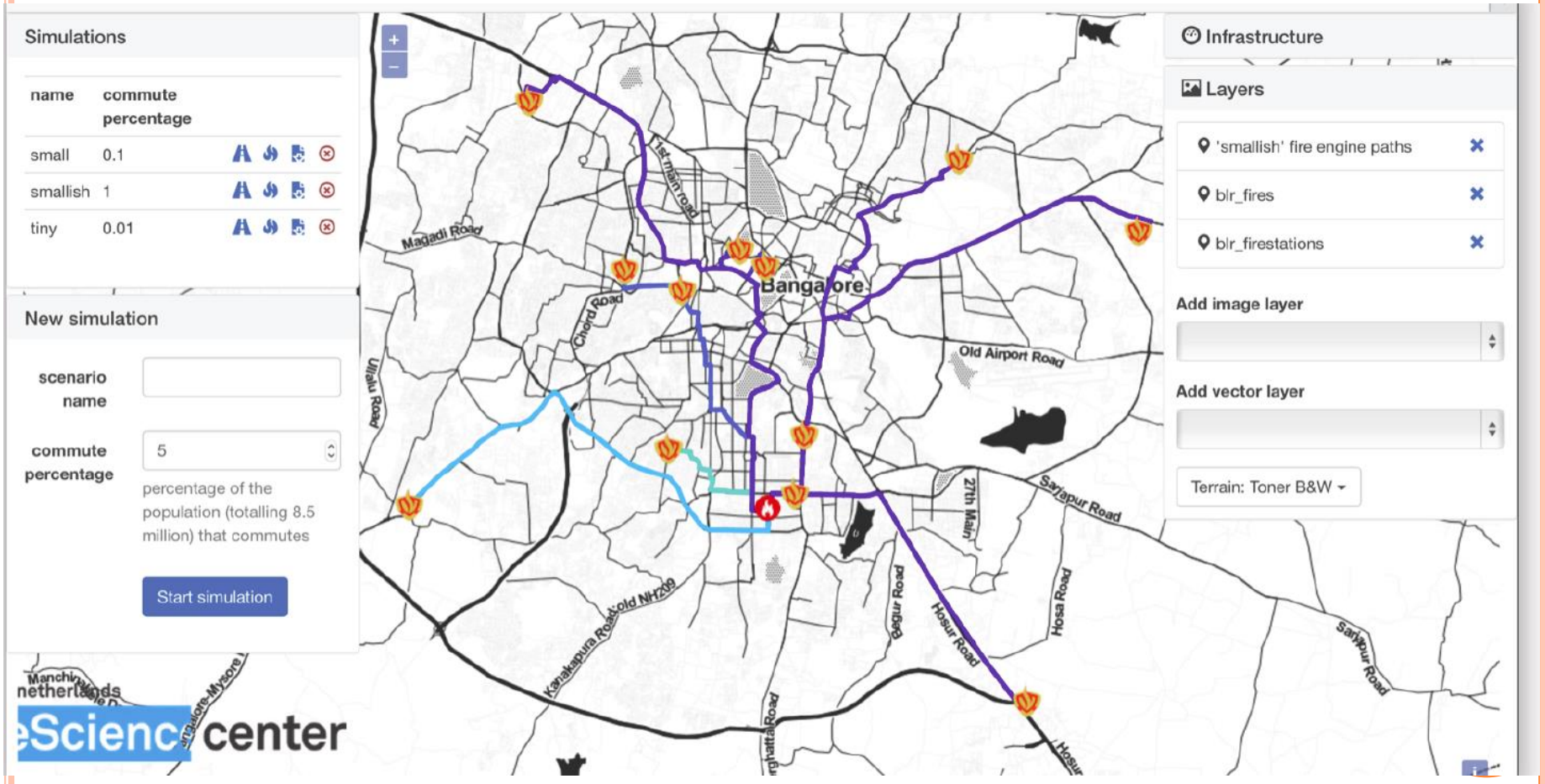
[6] Amsterdam Watergraafsmeer

[7] Amsterdam Rivierenbuurt

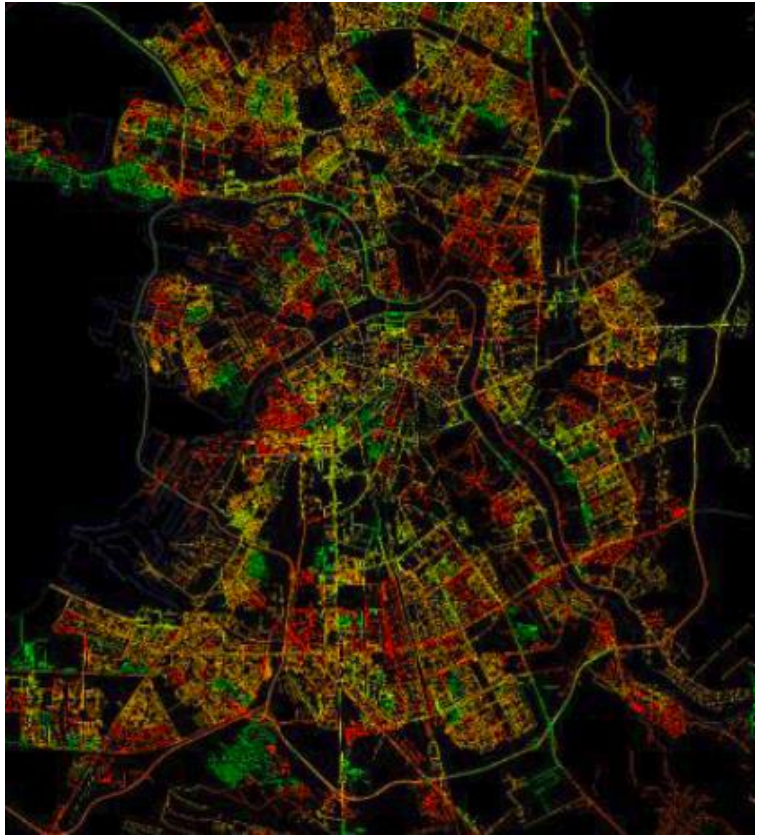


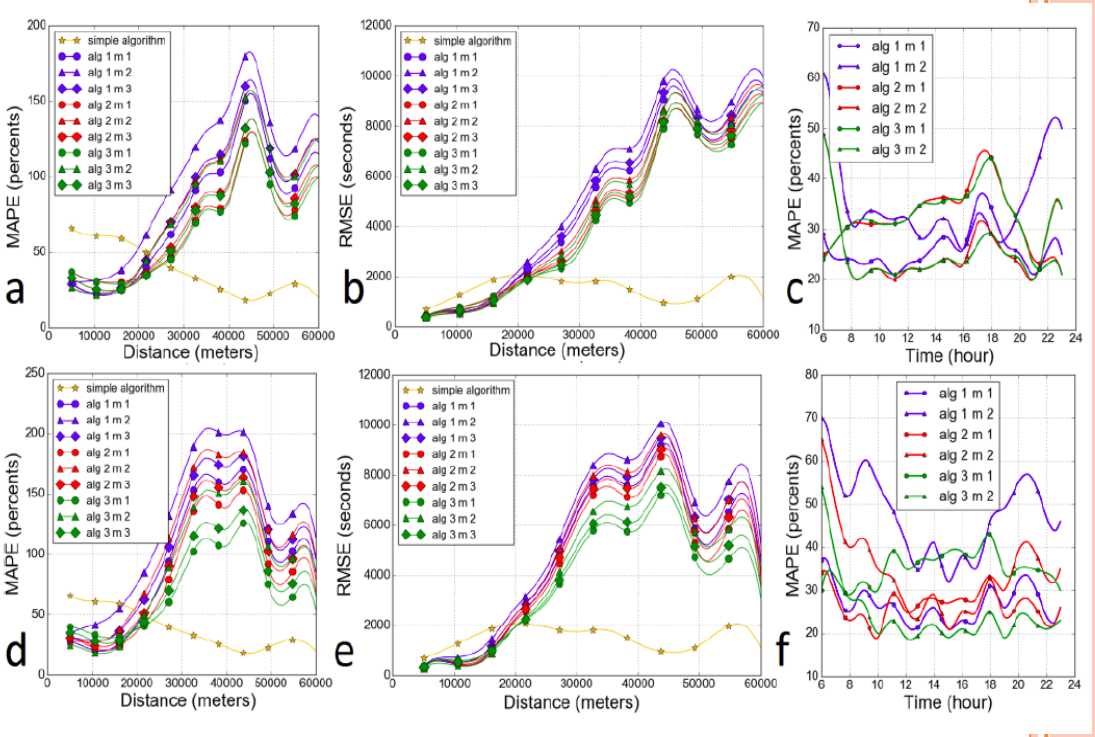

1
2
3
4
5
6
7
Het Twiske
Landsmeer
Nieuwe Gouw
Ватерланд
Broek in
Waterland
Зюндердорп
Zunderdorp
WESTPOORT
Эй
IJ
AMSTERDAM-NOORD
WATERLAND
GEUZENVELD-SLOTERMEER
AMSTERDAM
NIEUW-WEST
Амстердам
Amsterdam
EASTERN
DOCKLANDS
AMSTERDAM-OOST
Buiten-IJ
AMSTERDAM
OUD-ZUID
AMSTERDAM-ZUID
ZUIDERAMSTEL
Diemen
Duivendrecht
Diemerbos
Badhoevedorp
Амстелвеен
Amstelveen
Схипхол
Schiphol
ZUIDOOST

Fire (Bangalore)



Traffic model development based on partial data (St. Petersburg)





- Valentin Melnikov

V.R. Melnikov, V.V. Krzhizhanovskaya, A.V. Boukhanovsky, P.M.A. Sloot, "Data-driven Modeling of Transportation Systems and Traffic Data Analysis During a Major Power Outage in the Netherlands", Procedia Computer Science, V. 66, 2015, pp. 336-345.
doi: 10.1016/j.procs.2015.11.039

V.R. Melnikov, V.V. Krzhizhanovskaya, M.H. Lees, A.V. Boukhanovsky, "Data-driven Travel Demand Modelling and Agent-based Traffic Simulation in Amsterdam Urban Area", Procedia Computer Science, V. 80, 2016, pp. 2030-2041.
doi: 10.1016/j.procs.2016.05.523

Conclusions

- Data Science
  + Computational Science
  + Computer Science
- Machine learning is great for
  - Model calibration
  - Anomaly detection
- Even 1-class SVM can give 91% accuracy

- More challenges and exciting time ahead!

Brace yourselves for an interesting

Many thanks for your attention!

Happily blending into Krakow flowers ☺

Valeria Krzhizhanovskaya

University of Amsterdam, The Netherlands
ITMO University, Russia

V.Krzhizhanovskaya@uva.nl

ITMO UNIVERSITY

UNIVERSITY OF AMSTERDAM