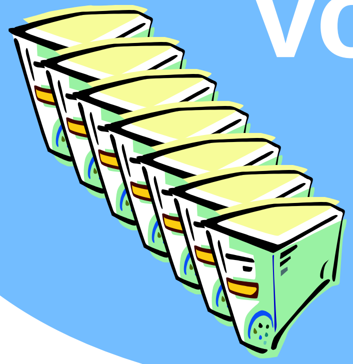
Big Data Science, Streams and Process Mining

Prof. Dr. Thomas Seidl
LMU München, Lehrstuhl für Datenbanksysteme und Data Mining

17.05.2018 – TUM Ringvorlesung „Digitalisierung“

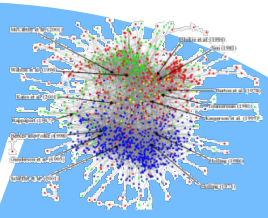
Big Data Everywhere – Many V's from Gartner 2011, IBM, BITKOM, Fraunhofer IAIS, etc.



Volume

Zettabytes
Exabytes
Petabytes
Terabytes

Variety



Video, photos, audio, texts, blogs, tables, locations.

Structured-semistructured-unstructured

Velocity

Batch
Periodic
Real-time
Anytime

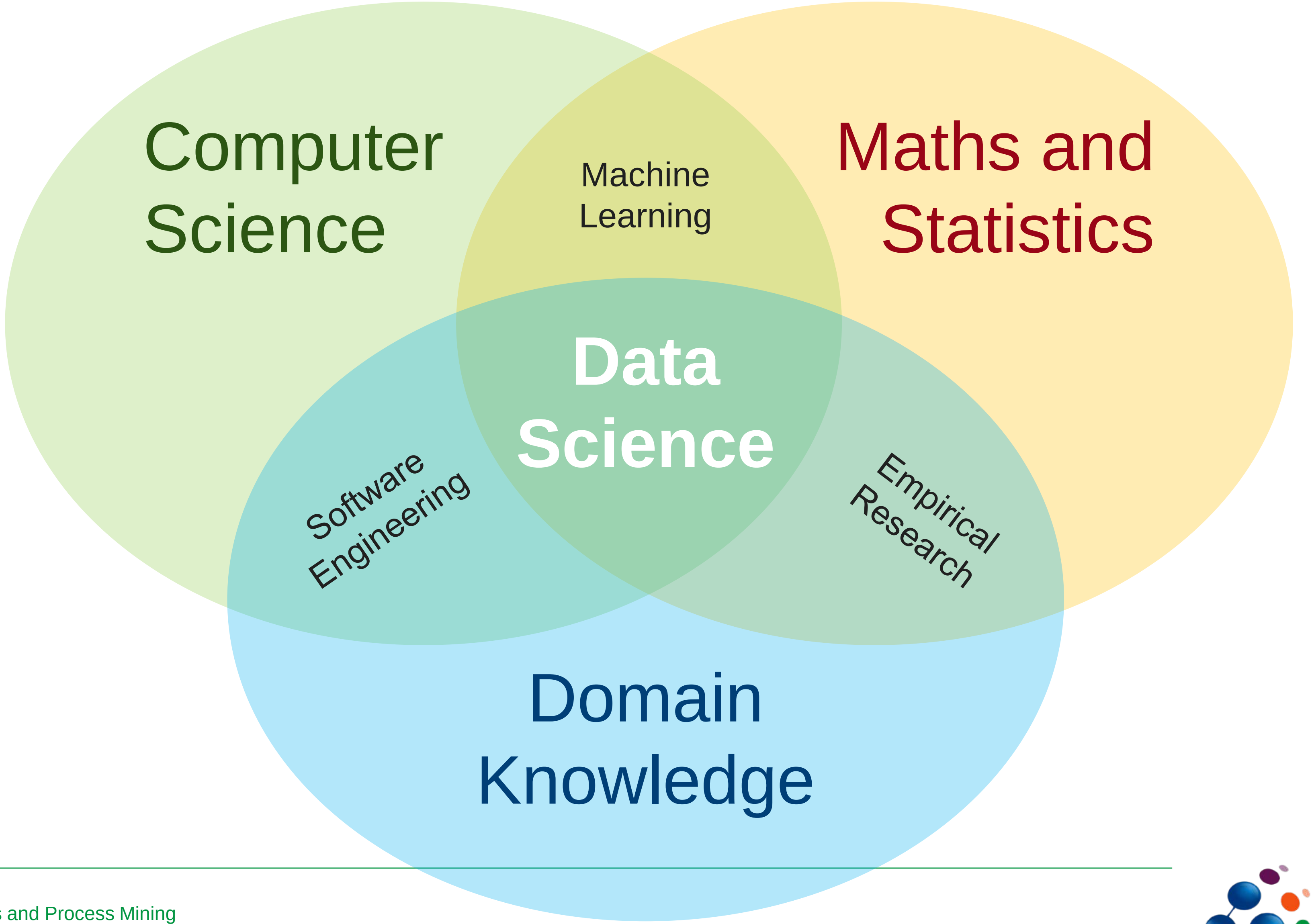
Veracity / Validity

reliability, noise, trust, provenance

Value / Visual Analytics

patterns, rules, trends, outliers
data science

Data Science Ingredients



LMU Data Science Ecosystem

Basic and Continuing Education

- BSc and MSc programs in Statistics and Informatics (19xx)
- MSc Data Science by Elite Network Bavaria (2016)
- Certified advanced training course (2018), Munich R courses (2015)

Master Data Science (est‘d 2016)

DATA
SCIENCE@LMU
STATISTICS + INFORMATICS

www.datascience-munich.de

- Funded by **Elitenetzwerk Bayern**
- Operated by **Statistics** and **Informatics** at **LMU** + TUM + U Augsburg + U Mannheim

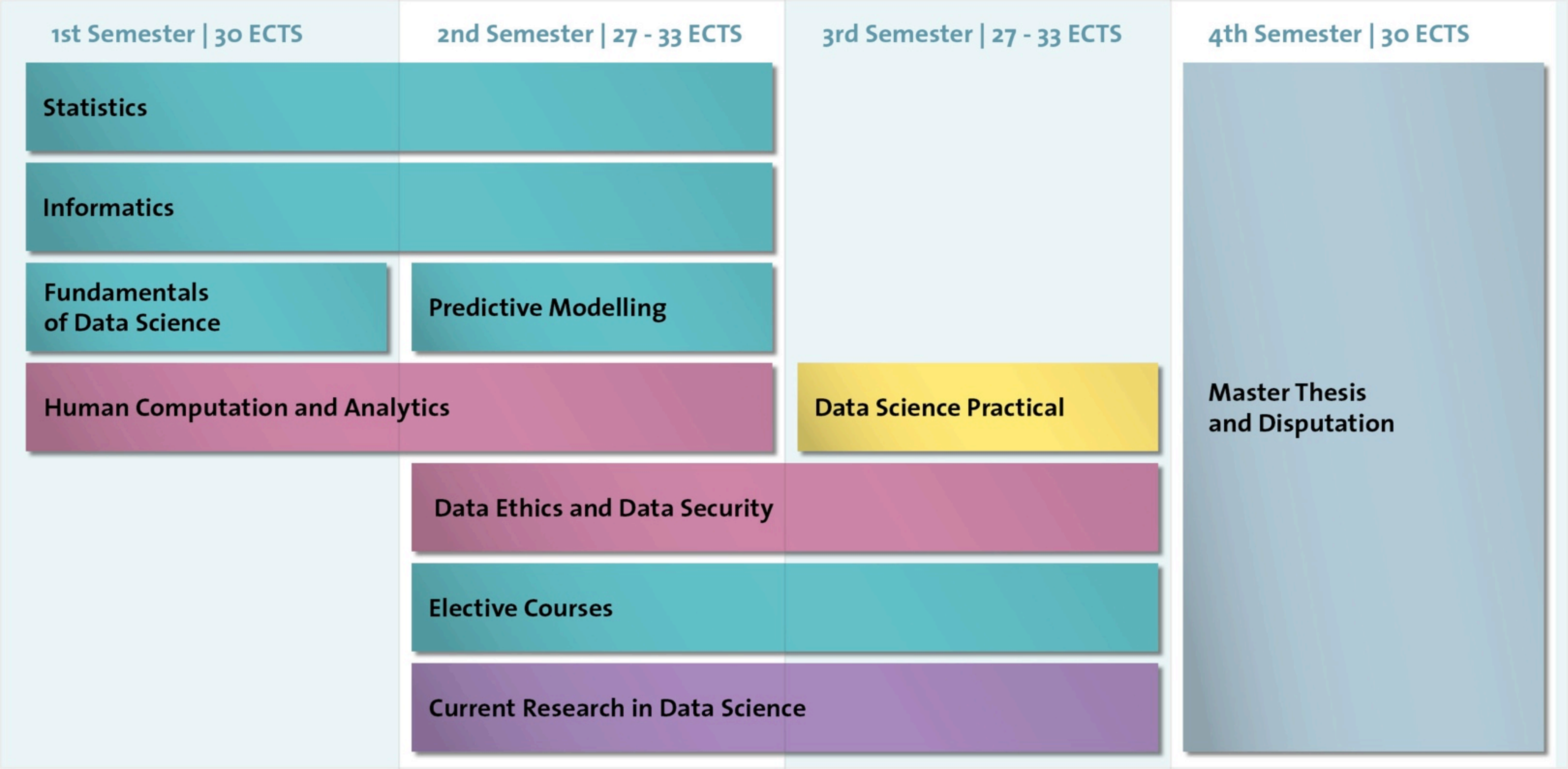
- Traditional and practical courses
  - Focused Tutorials, Summer School, Data Fest, Data Science meets Data Practice

- International scope
  - Fully English spoken, small cohorts
  - Entrance profile: excellent grades for
    $\geq$ 30 ECTS in Statistics
    $\geq$ 30 ECTS in Computer Science

- Spokespersons
  - Prof. Göran Kauermann (LMU Statistics)
  - Prof. Thomas Seidl (LMU Informatics)
  - Dr. Constanze Schmaling (coordinator)

Master Data Science: Curriculum

DATA
SCIENCE@LMU
STATISTICS + INFORMATICS



LMU Data Science Ecosystem

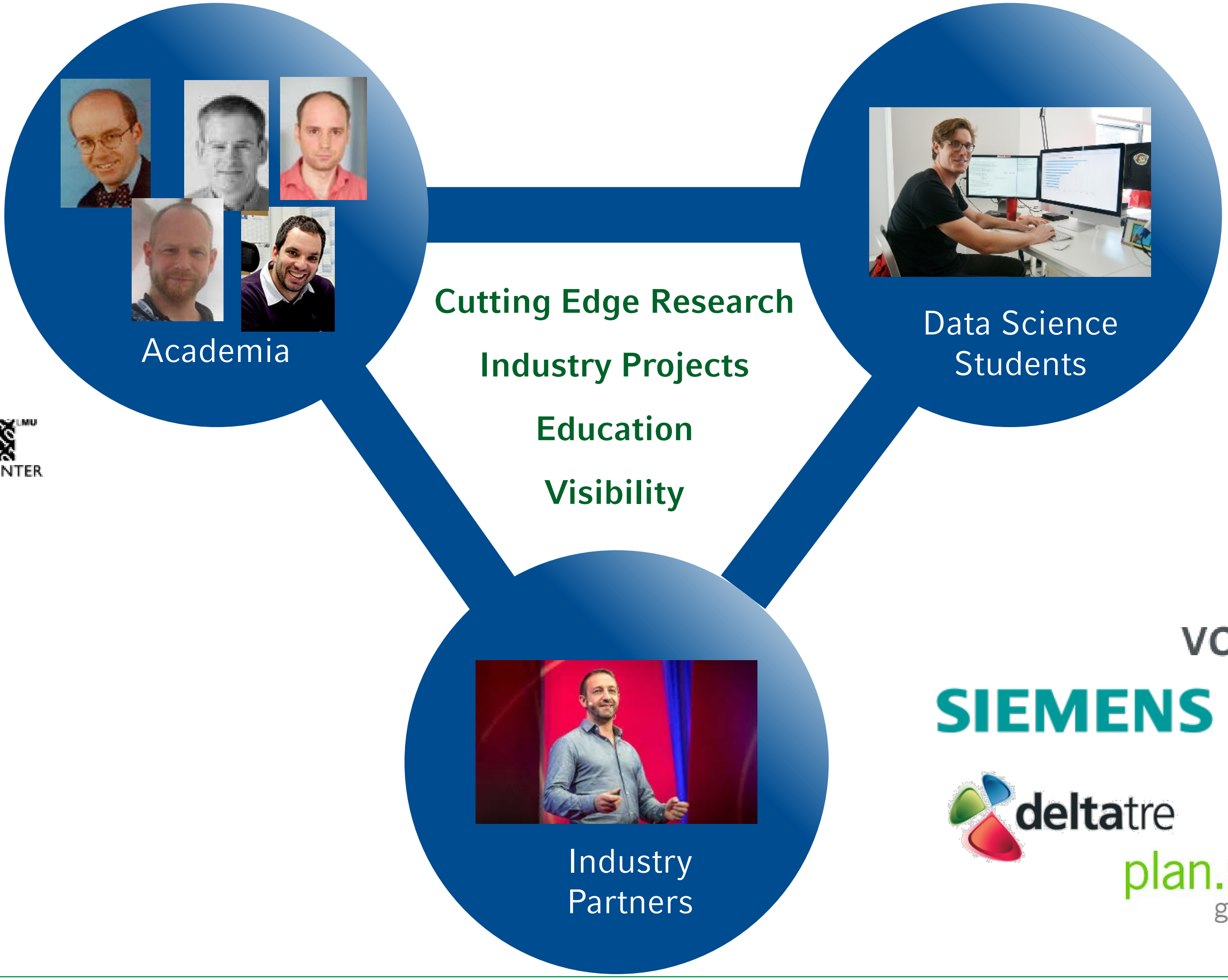
Basic and Continuing Education

- BSc and MSc programs in Statistics and Informatics (19xx)
- MSc Data Science by Elite Network Bavaria (2016)
- Certified advanced training course (2018), Munich R courses (2015)

Student Labs with Industrial Partners

- Data Science Lab @LMU (2014)
- ZD.B Innovation Lab „Big Data Science“ @LMU (2017)
- StaBLab (1997)

Data Science Lab @LMU Munich (est‘d 2014)



Data Science Lab @LMU: Working space for collaborations

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU Data Science Ecosystem

Basic and Continuing Education

- BSc and MSc programs in Statistics and Informatics (19xx)
- MSc Data Science by Elite Network Bavaria (2016)
- Certified advanced training course (2018), Munich R courses (2015)

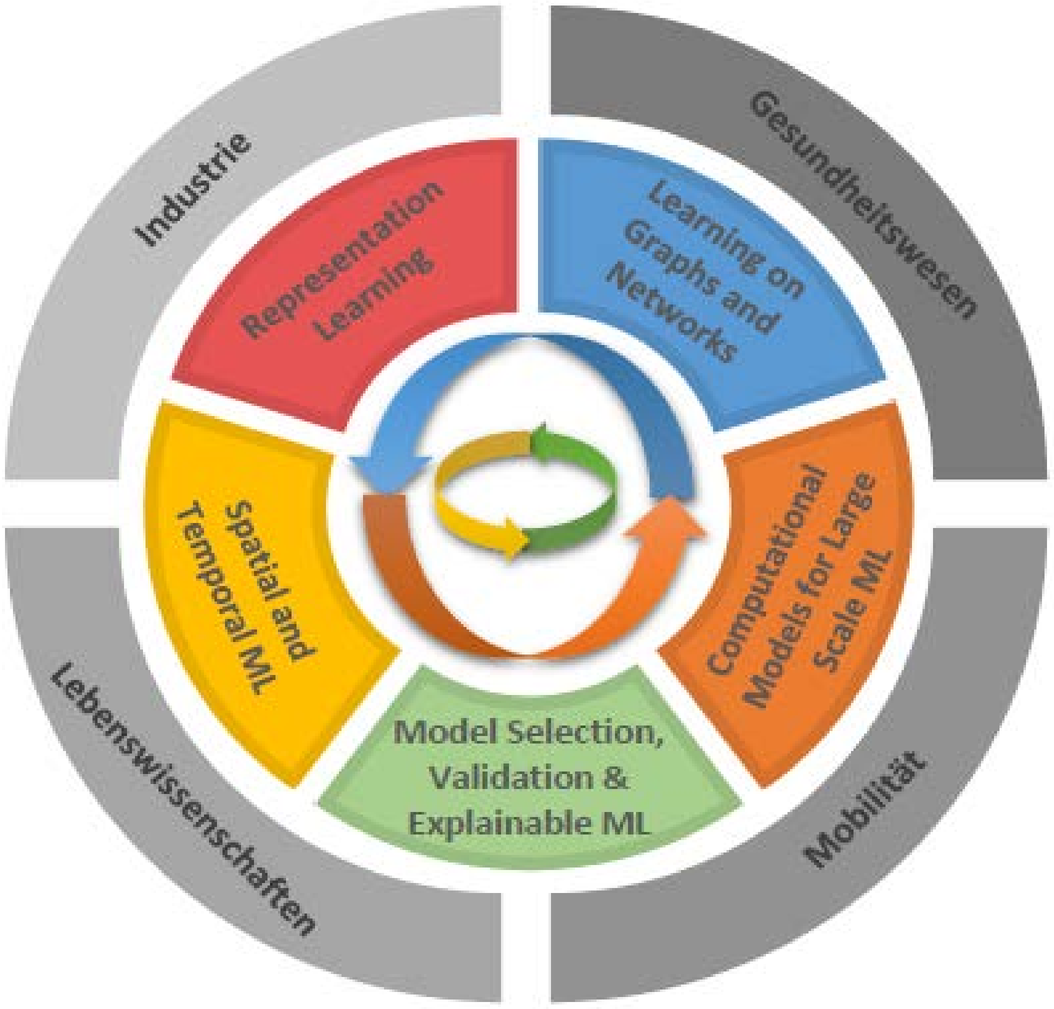
Student Labs with Industrial Partners

- Data Science Lab @LMU (2014)
- ZD.B Innovation Lab „Big Data Science“ @LMU (2017)
- StaBLab (1997)

Competence Center and Doctoral Training

- MCML – Munich Center for Machine Learning @LMU, TUM (BMBF)
- MuDS – Munich School for Data Science @Helmholtz, TUM, LMU (submitted)

Munich Center for Machine Learning (MCML)



- Funded by BMBF (2018 – 2022 – 2025)
  - Berlin, Dortmund/St. Augustin, München, Tübingen
- Partners from Informatics & Statistics (LMU & TUM)
  - Directed by Thomas Seidl (Inform.) and Bernd Bischl (Stat.)
- Four leading application areas …
  - Mobility, Life Sciences, Healthcare, Industry
- Five research areas …
  - Spatio-temporal ML, Graphs & Networks, Representation Learning, Validation & Explanation, Large Scale ML

LMU Data Science Ecosystem

Basic and Continuing Education

- BSc and MSc programs in Statistics and Informatics (19xx)
- MSc Data Science by Elite Network Bavaria (2016)
- Certified advanced training course (2018), Munich R courses (2015)

Student Labs with Industrial Partners

- Data Science Lab @LMU (2014)
- ZD.B Innovation Lab „Big Data Science“ @LMU (2017)
- StaBLab (1997)

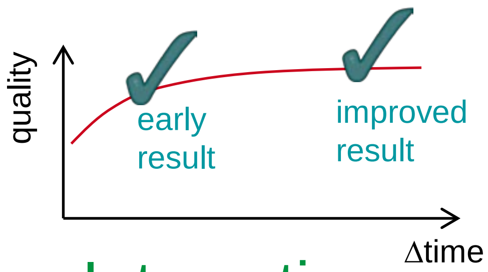
Competence Center and Doctoral Training

- MCML – Munich Center for Machine Learning @LMU, TUM (BMBF)
- MuDS – Munich School for Data Science @Helmholtz, TUM, LMU (submitted)

Solutions for Industry

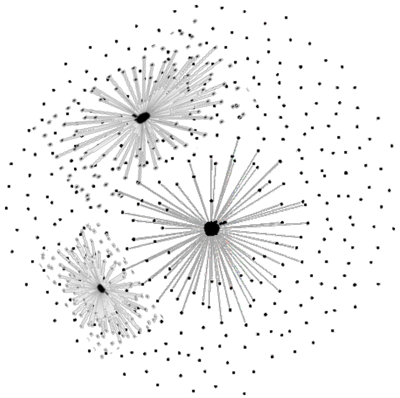
- Fraunhofer ADA: IIS, FAU, LMU (in preparation)

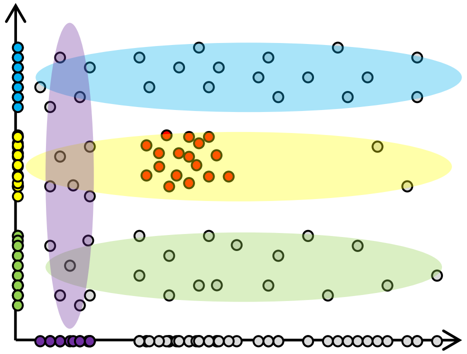
Some of our Research Areas



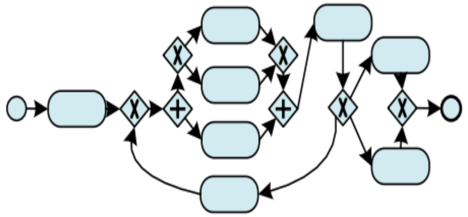
Interactive Analytics

Representation Learning

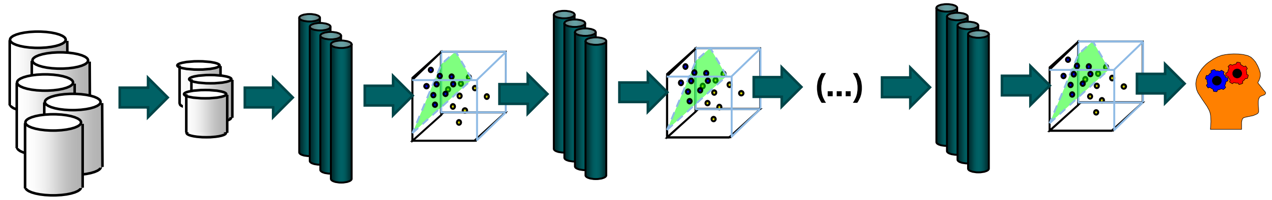




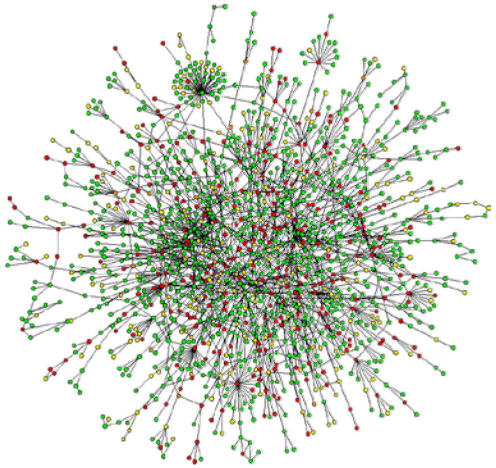
Explainable AI



Process Mining



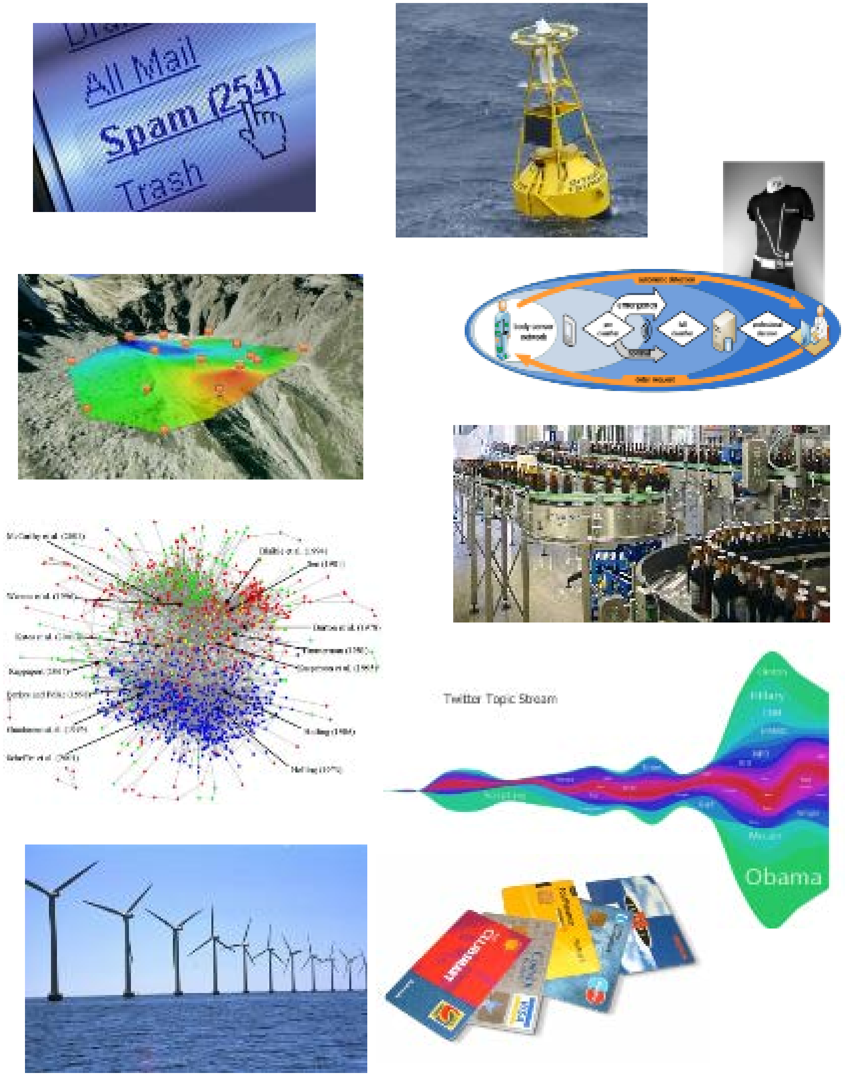
Deep Learning



Knowledge Graphs

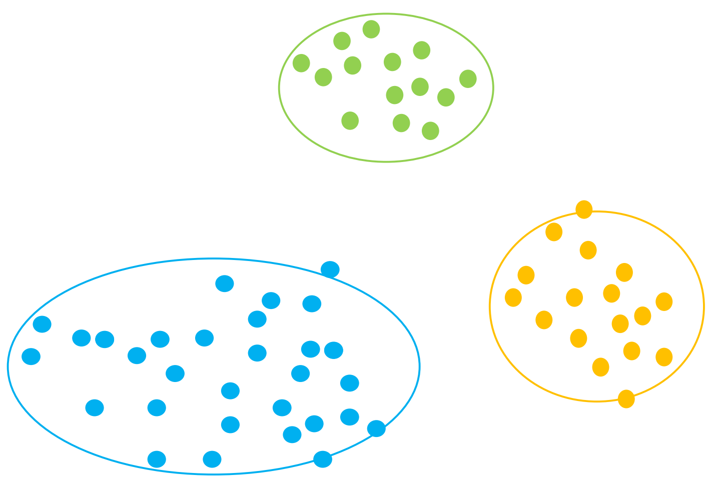
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

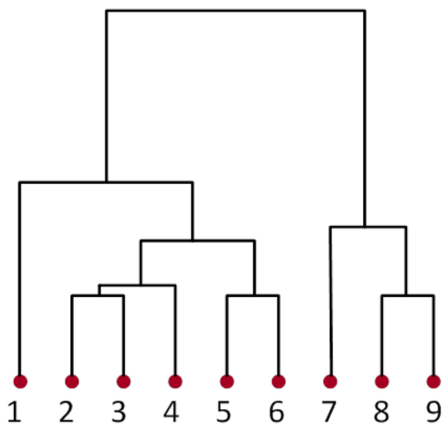
Streaming Data everywhere

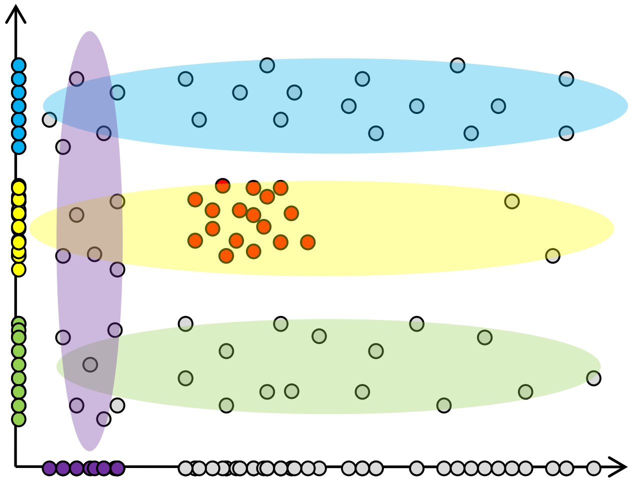


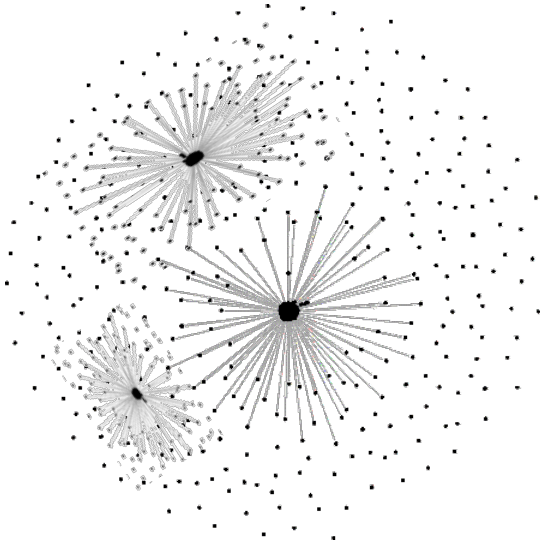
- Environment
  - Sensor data analysis, monitoring oceans or glaciers, pollution detection, lab data analysis
- Engineering
  - Power network surveillance, production monitoring, process control
- Business
  - eCommerce portals, fraud detection, product sentiment analysis, collaboration analysis
- Health
  - Detection of epidemics, monitoring of patients, personalized medicine, psychology
- Society and Humanities
  - Network analysis in Facebook, Twitter, XING, LinkedIn, DBLP; behavioral pattern analysis

Clustering / Labeling





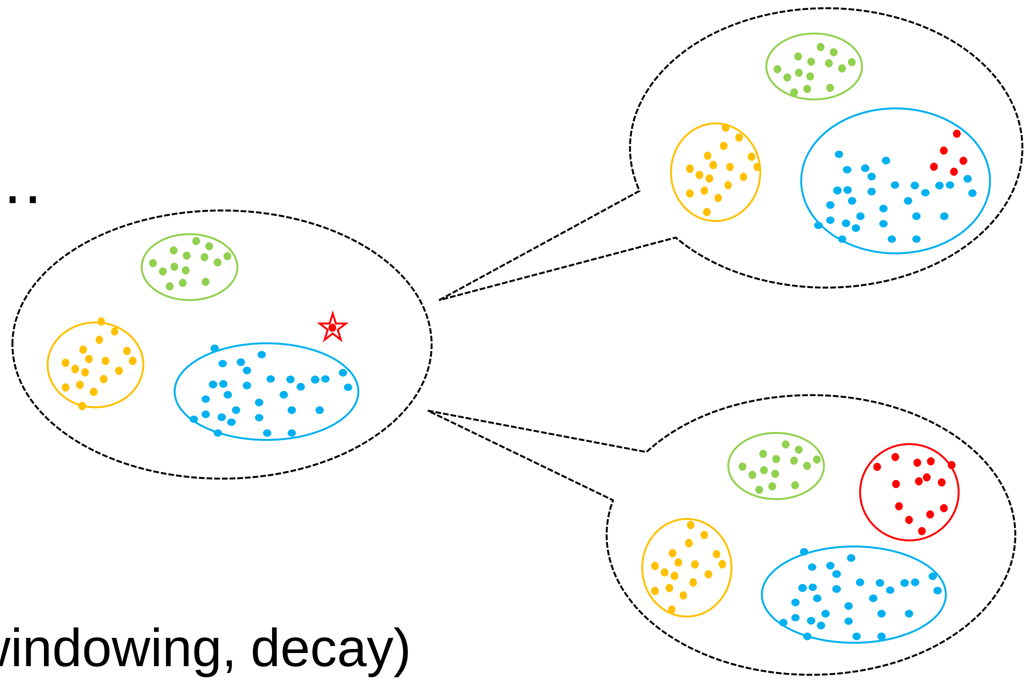




- Customer Segmentation, Labeling Products, Clique Detection, …
- Methods
  - Clustering for heterogeneous objects
  - Subspace clustering, density estimation for higher subspaces
  - Non-linear Correlation Clustering
  - Semi-supervised clustering, constraints models
  - …

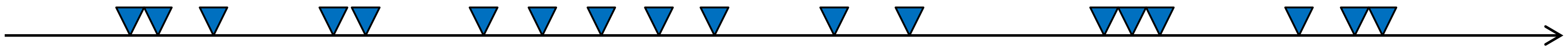
Clustering and Change Detection

- Patterns in dynamic data may evolve over time
  - Example clustering: group similar objects while separating dissimilar ones
- Does a deviating object represent …
  - An actual outlier?
  - A moving cluster?
  - A new cluster?
- Approaches
  - Aging models for objects and clusters (windowing, decay)
  - Deal with missing values

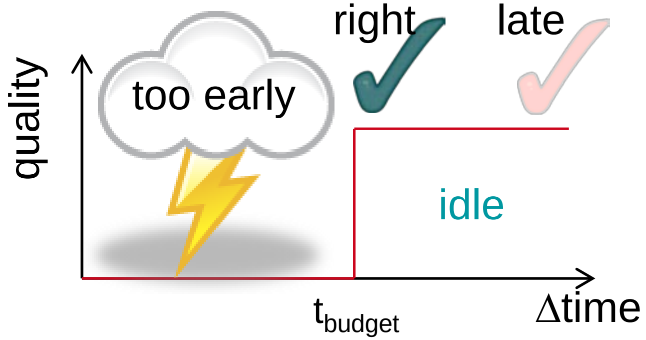


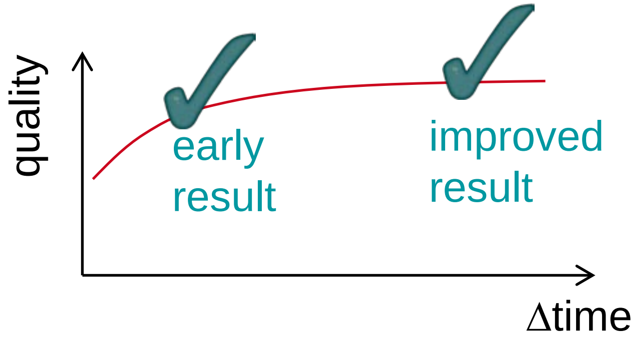
Challenge: Anytime data mining

- Scenario: Data stream with varying arrival intervals



- Classic real-time (budget) algorithm vs. Anytime algorithms

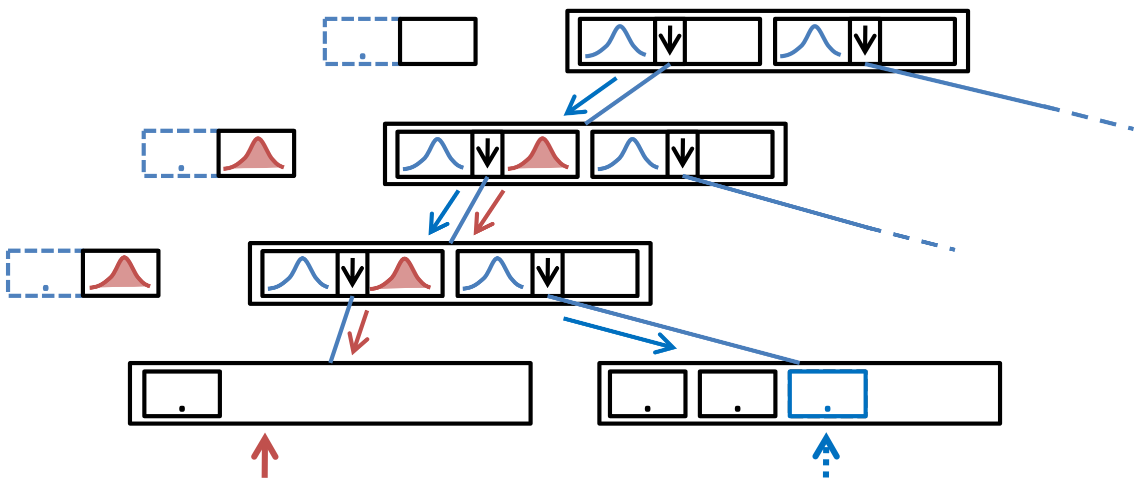




  - Idea: Provide a reasonable result after interruption at any time
  - Exploit all available time for incrementally improving the result
  - Proposed classifiers so far: SVM, nearest neighbor, boosting techniques, Bayesian networks, naïve Bayes on discrete data, Bayes on continuous data

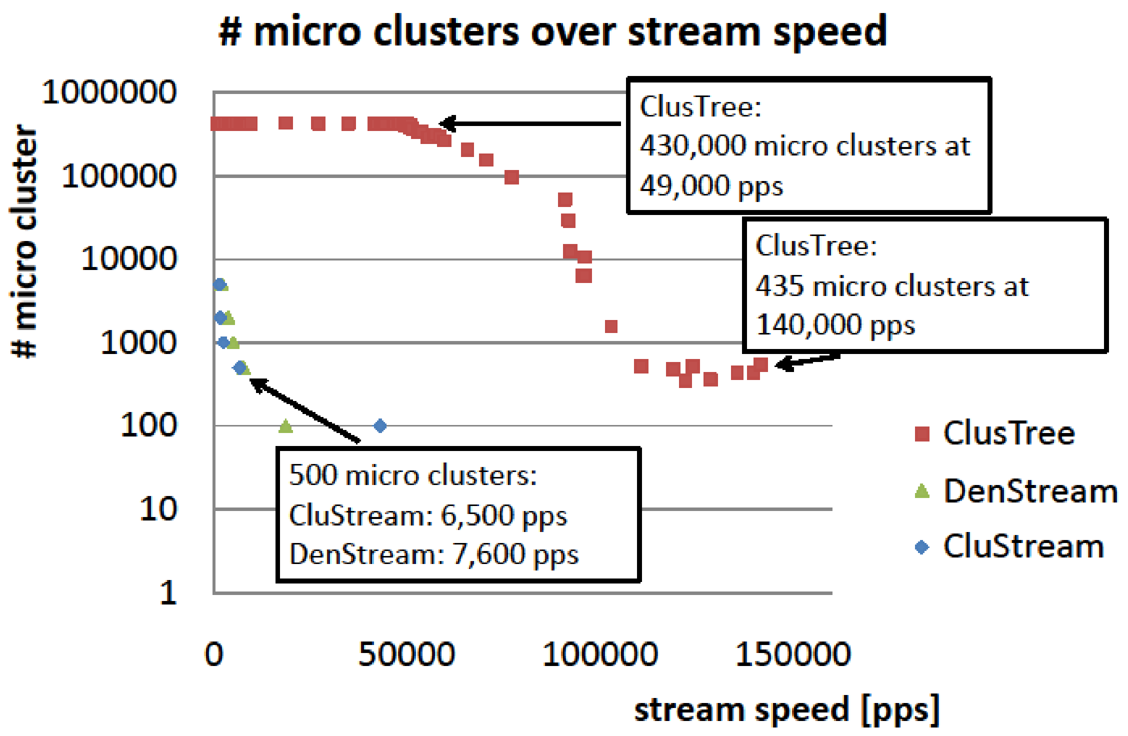
ClusTree stream clustering: buffers and hitchhikers

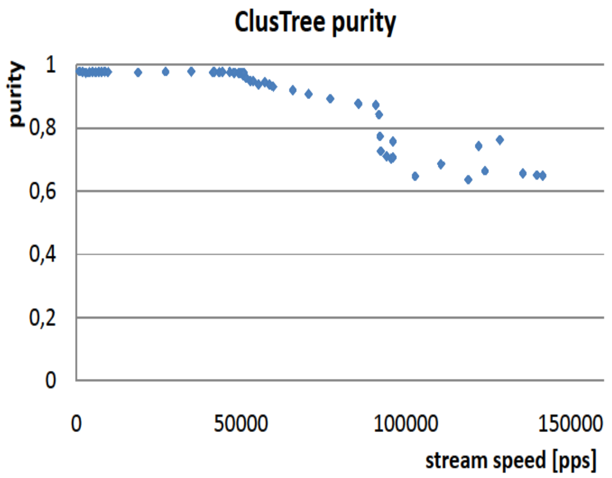
- Hierarchy of micro-clusters CF = (N, LS, SS)
- Buffer – interrupt insertion: aggregate objects on interrupt
- Hitchhiker – resume insertion: take buffer along (if on the same way)



[Kranen, Assent, Baldauf, Seidl:. KAIS 29(2), 2011]

ClusTree evaluation – adaptive clustering

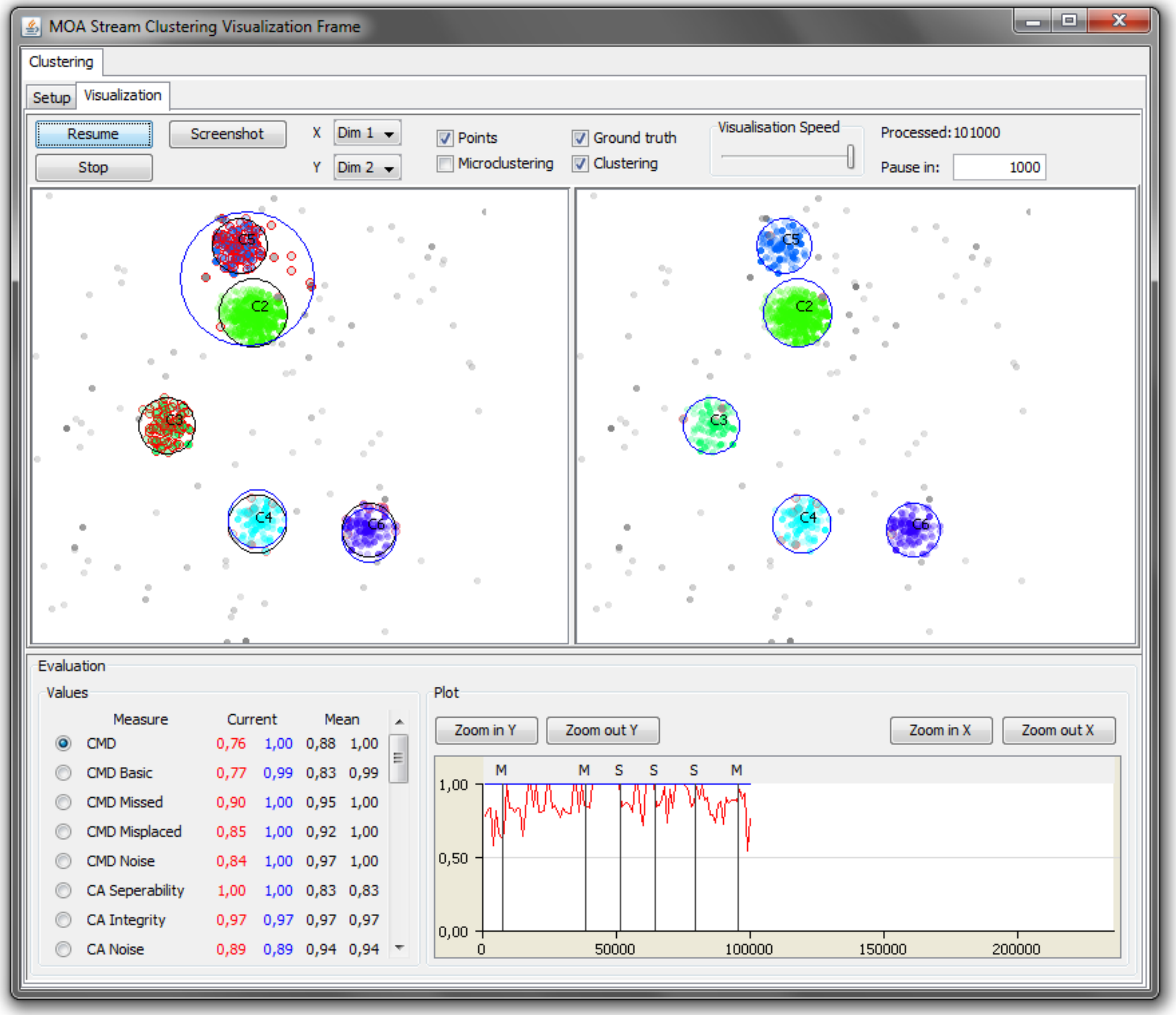




- Setup for constant streams:
  - DenStream [SDM'06] & CluStream [VLDB'03]: #MC → processable pps
  - ClusTree [ICDM'09, KAIS'11]: stream speed → maintainable #MC

[Kranen, Assent, Baldauf, Seidl: ICDM 2009, KAIS 2011]

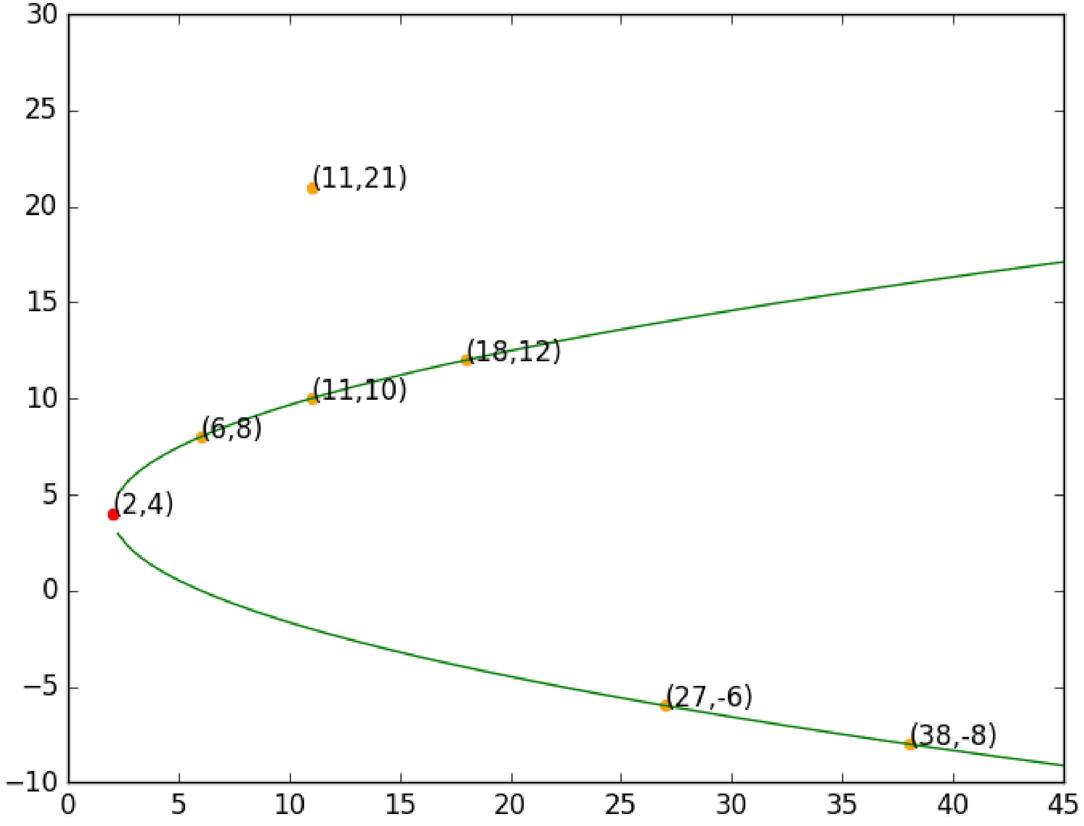
MOA: Integration into open-source framework

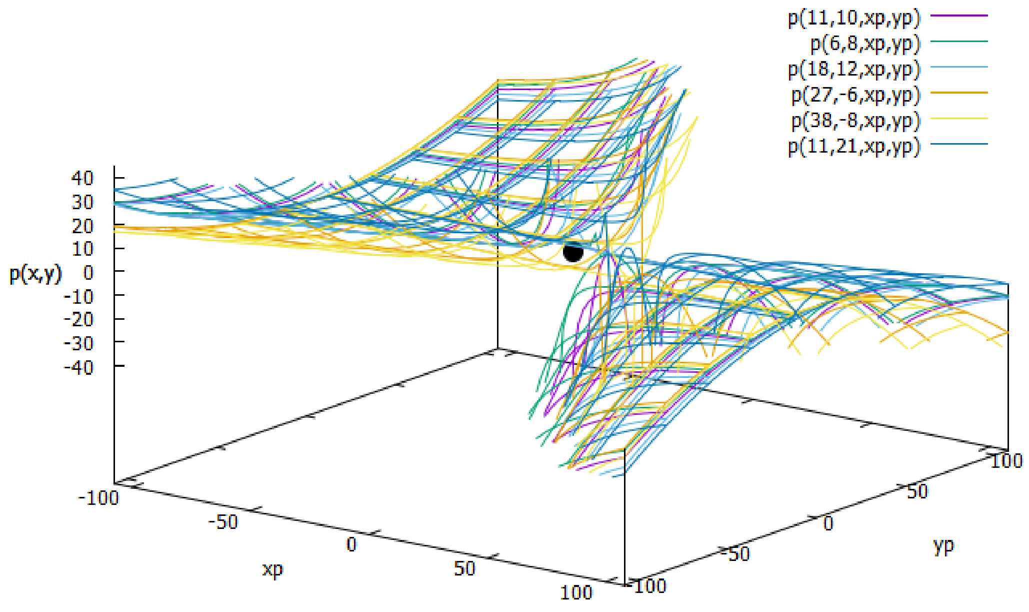


[Bifet, Holmes, Pfahringer, Kranen, Kremer, Jansen, Seidl: JMLR Proc. 11, 2010]

http://moa.cs.waikato.ac.nz/

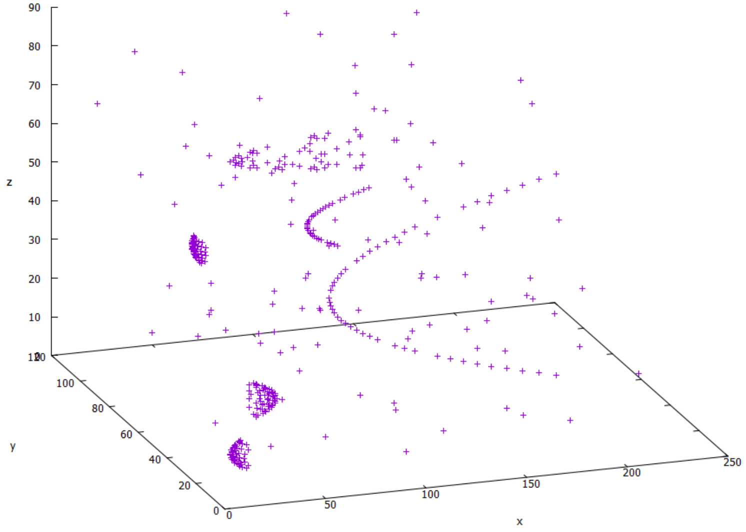
Correlation Clustering: Parameter space for parabolas



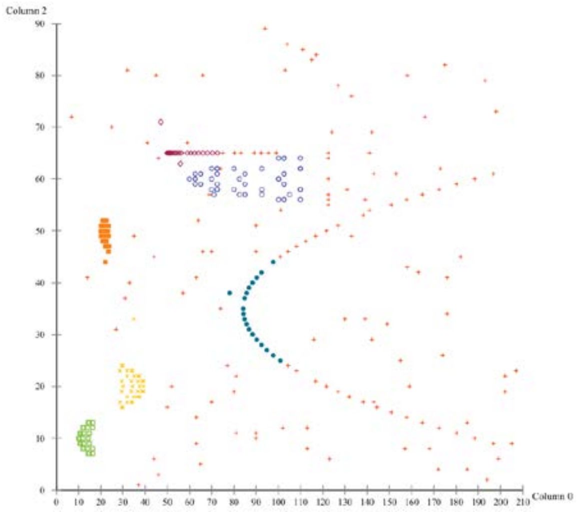


[Kazempour, Mauder, Kröger, Seidl: SSDBM 2017]

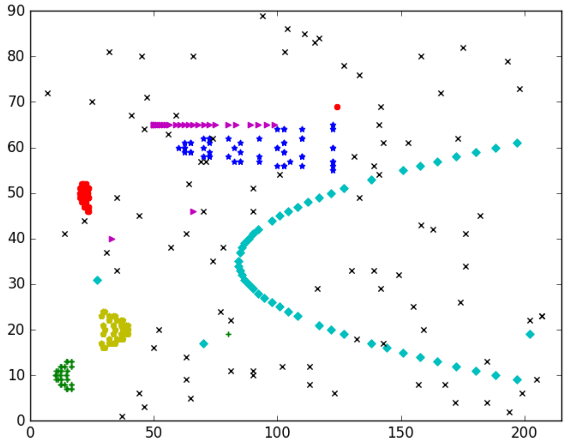
Results for Hyperparaboloid Correlation Clustering (HCC)



data set



$DBSCAN$



HCC

[Kazempour, Mauder, Kröger, Seidl: SSDBM 2017]

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Correlation Clustering on Streams

- Example for Clustering
  - Detect 2D manifolds in 3D space
  - Clusters evolve along data stream

[Borutta ~2018]

New Computation Models for Machine Learning

- Distributed Processing on Hadoop Distributed File System HDFS
  - Hadoop MapReduce
  - Apache Spark
  - Apache Flink

Similarity Self-Join $DB \bowtie_\varepsilon DB$ → Core point detection $|N_\varepsilon(p)| \geq \mu$ → Connected components $\{C_1, C_2, \dots, C_m\}$

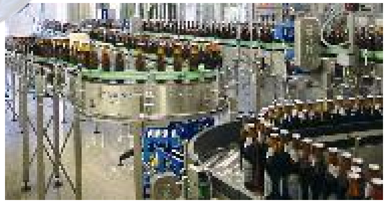
- Graph and Network Analysis
  - Pregel, Giraph, GraphX, Gelly

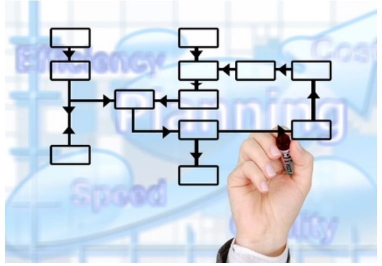
- GPU cluster computing

- New interaction models, explainable AI
  - interactive data mining, incremental algorithms
  - Visual Analytics

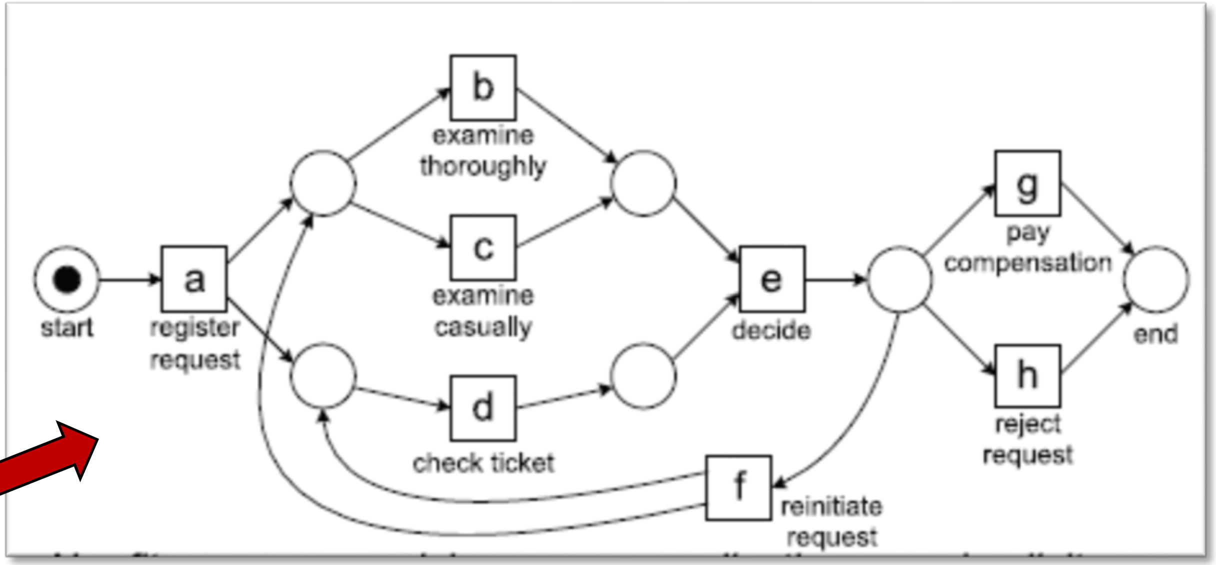
[Fries, Wels, Seidl: EDBT 2014] – [Fries, Boden, Stepien, Seidl: ICDE 2014] – [Seidl, Fries, Boden: BTW 2013] – [Seidl, Boden, Fries: ECML/PKDD 2012]

Process Discovery





#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

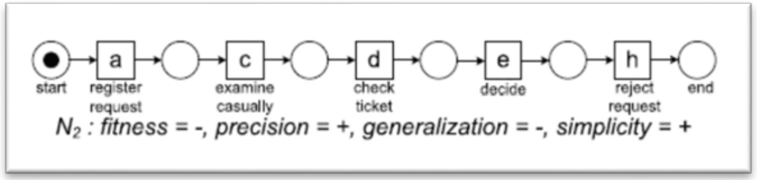


Task: Extract process model from log entries which

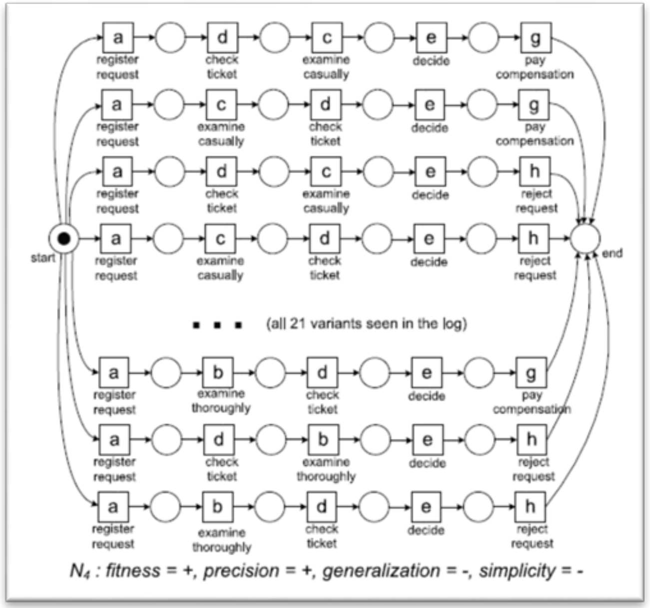
… is able to replay the log ⇒ *Fitness*
… simplifies as far as possible ⇒ *Simplicity*
… does not overfit the log ⇒ *Generalization*
… does not underfit the log ⇒ *Precision*

Process Discovery: Tune Generalization Granularity

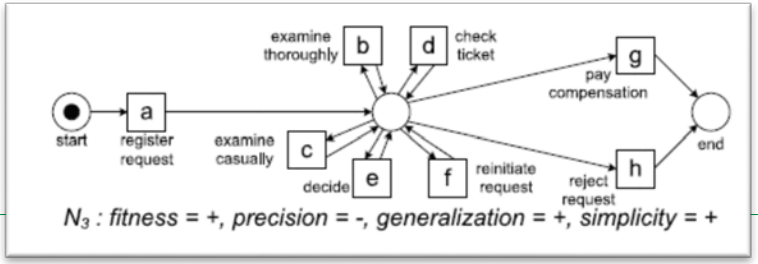
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	



N_2 : *fitness = -, precision = +, generalization = -, simplicity = +*

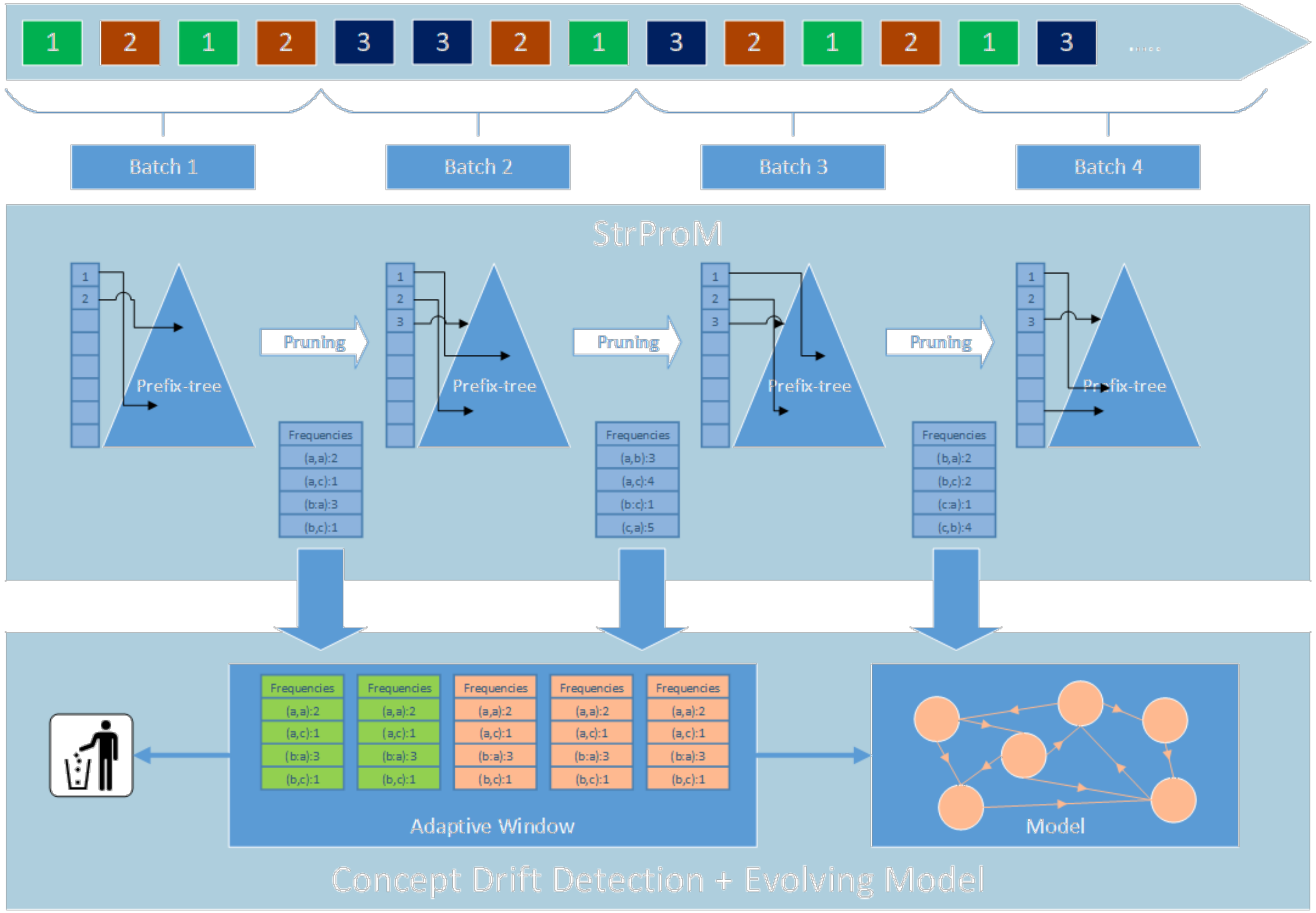


N_4 : *fitness = +, precision = +, generalization = -, simplicity = -*



N_3 : *fitness = +, precision = -, generalization = +, simplicity = +*

Stream Process Mining



Event Stream

Batched Approach

Prefix-Trees

Irregular Updates

Decaying

[Hassani, Siccha, Richter, Seidl: IEEE CI 2015]

Drifts in Processes

Human Intention

active

Technological Evolution

Changing Workflows

Pace

slow, gradual

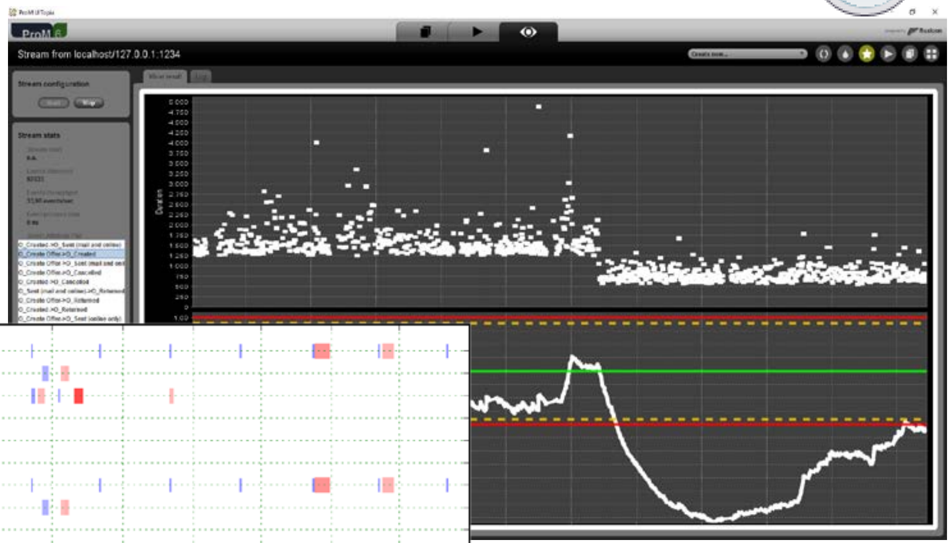
fast, sudden

Result Statistics

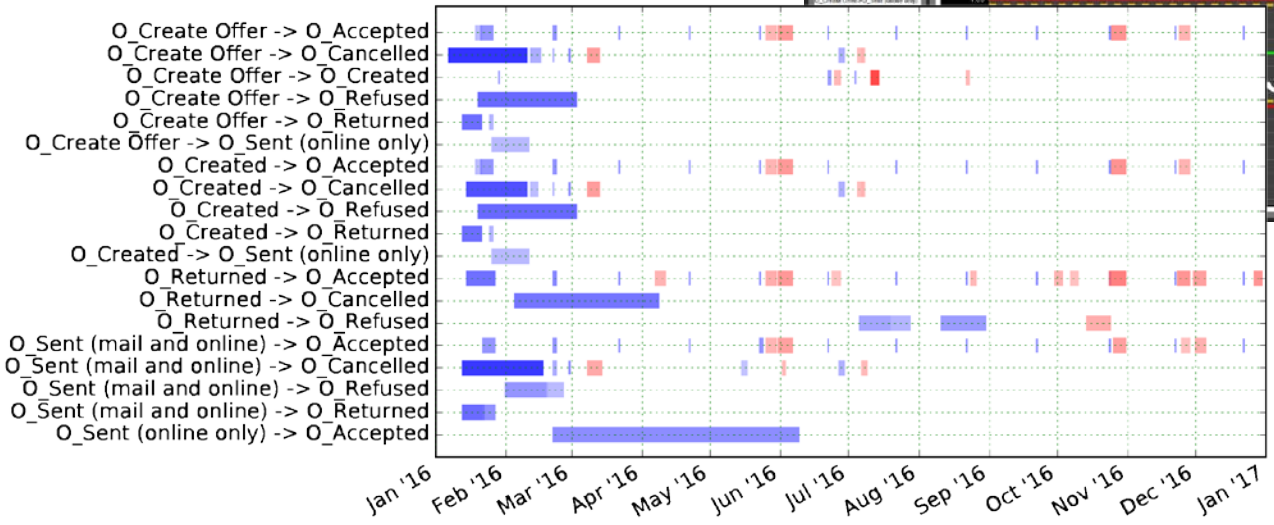
Disaster

passive, reactive

Tesseract: Temporal Drifts in Event Streams



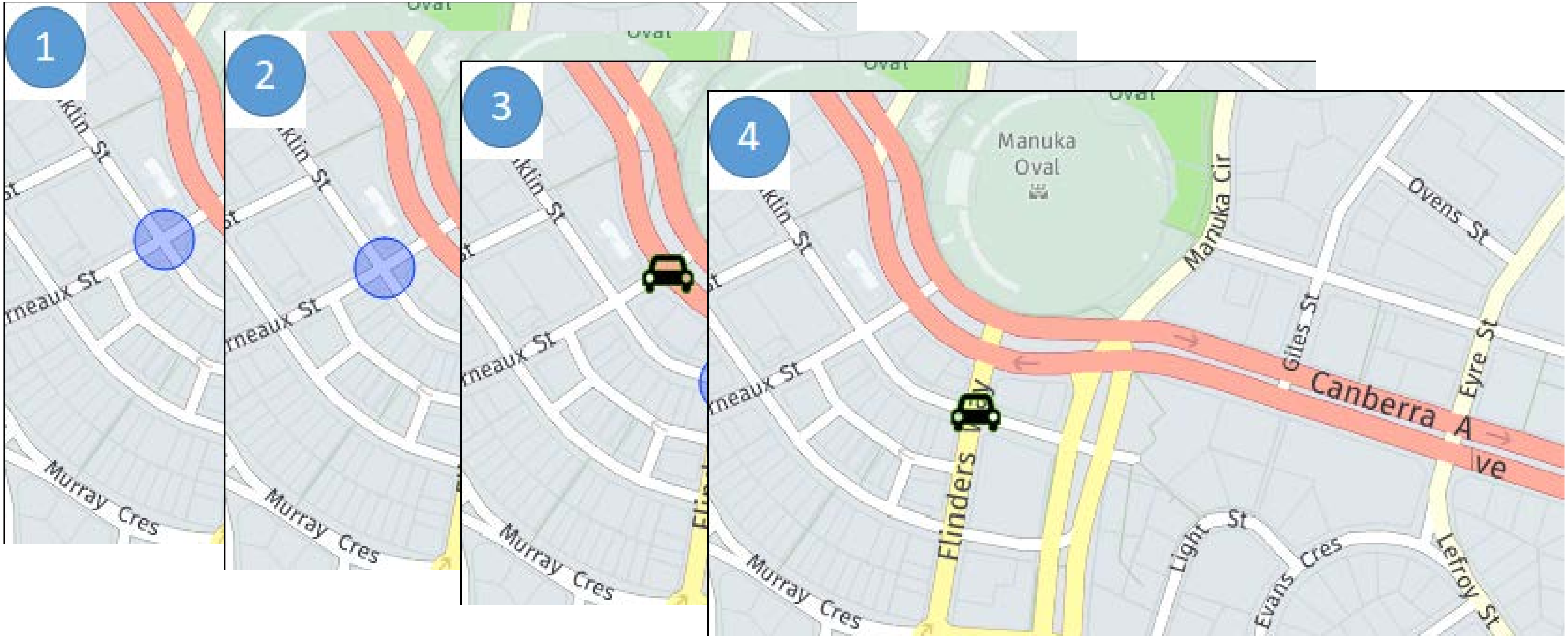
Drift Significance
and Detection



Gantt Chart
Visualization

[Richter, Seidl: BPM 2017]
best student paper award

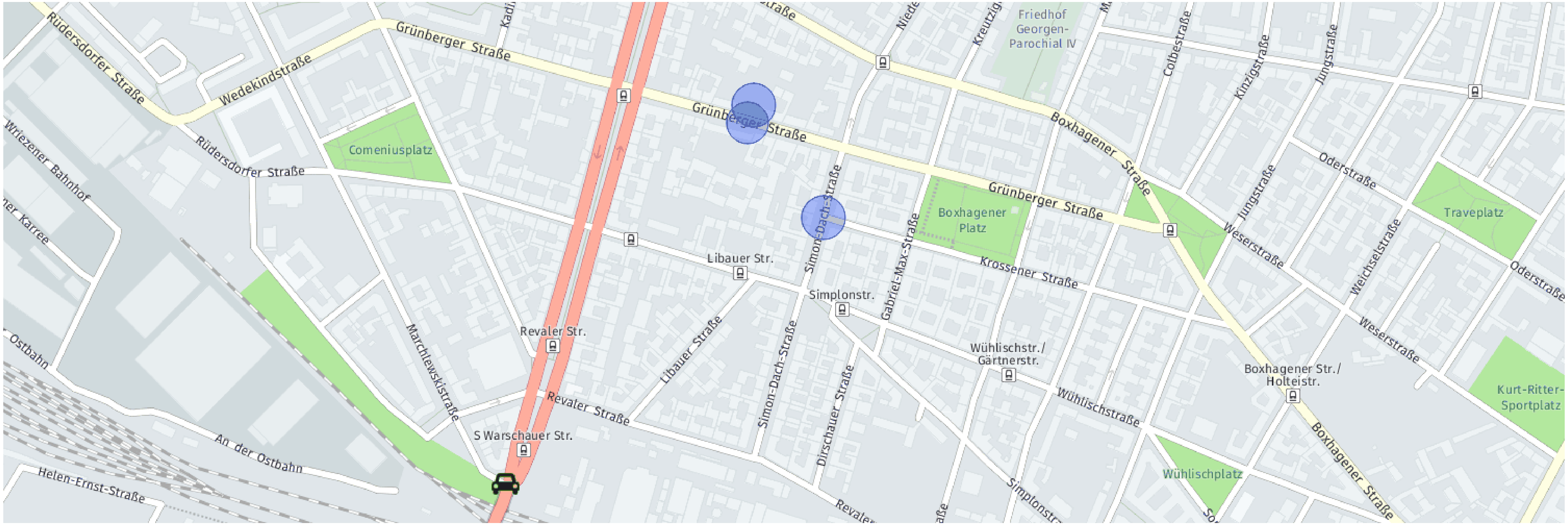
Real-Time Routing – How to Reach Free Parking Lots?



Smart cities like Melbourne installed sensors at parking spots

[Schmoll, Schubert: EDBT 2018]

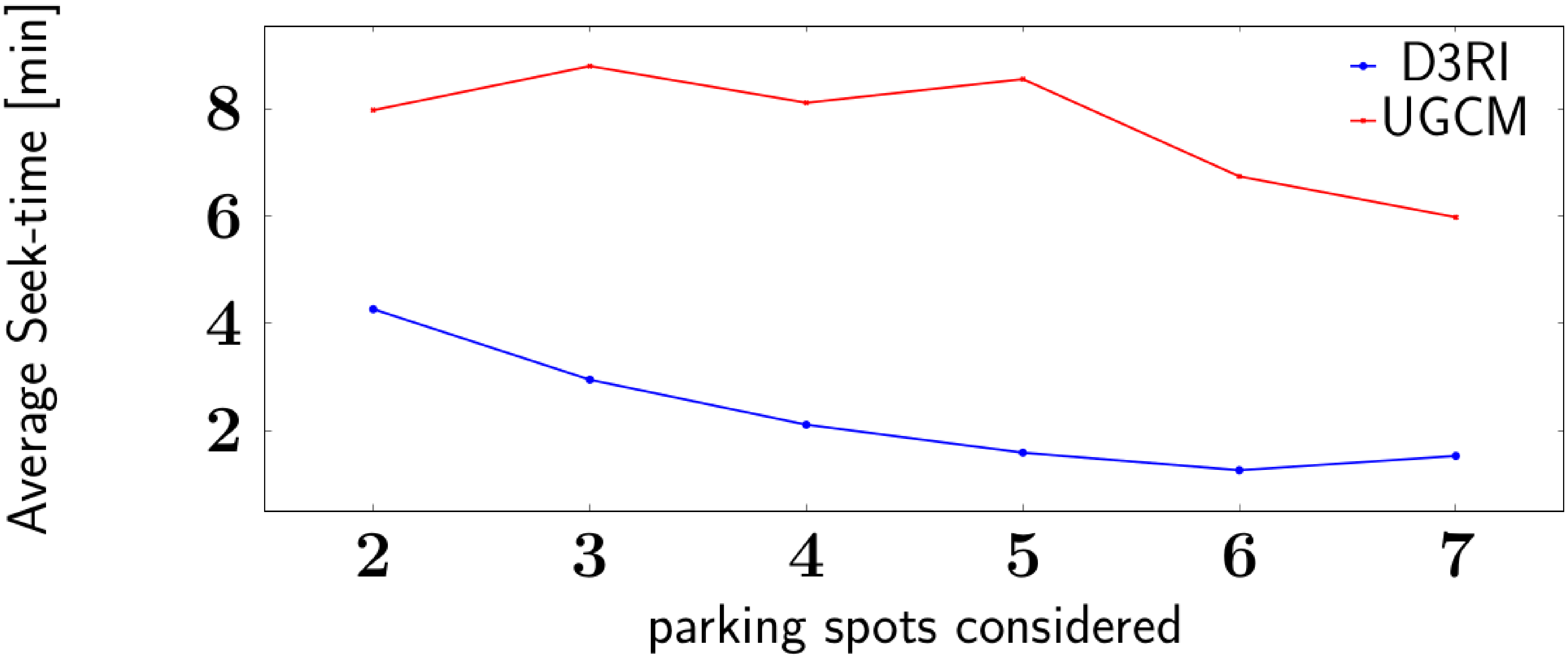
Probabilistic Approach by Markov Decision Processes



Current state is certain, but future states are not
Conventional path searching is too strict

[Schmoll, Schubert: EDBT 2018]

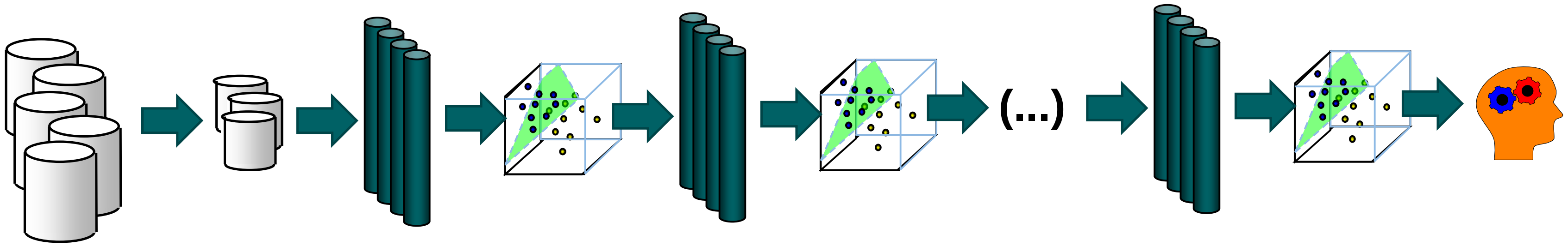
Evaluation

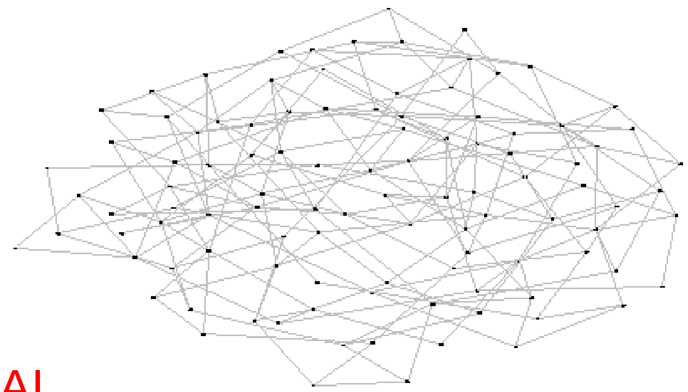


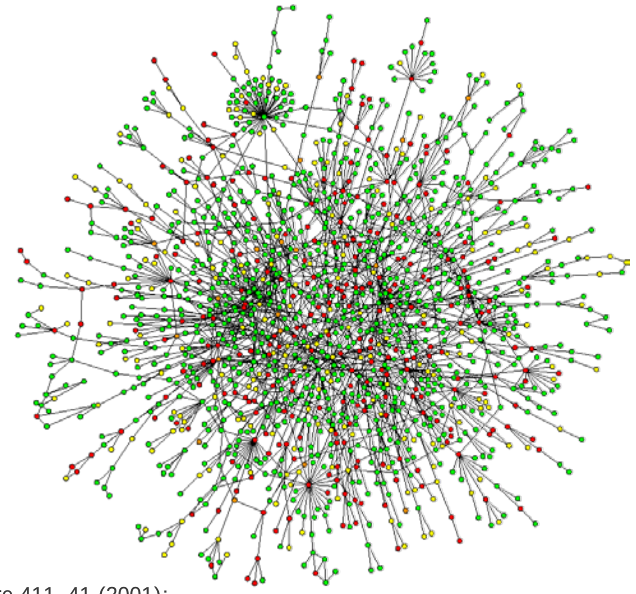
UGCM – Guo & Wolfson, GeoInformatica 2016
D3RI – Schmoll & Schubert, EDBT 2018

Representation Learning for Graph Structures

- Connection of multiple layers
- Representation learning
- How about explanations?
- How about graphs?



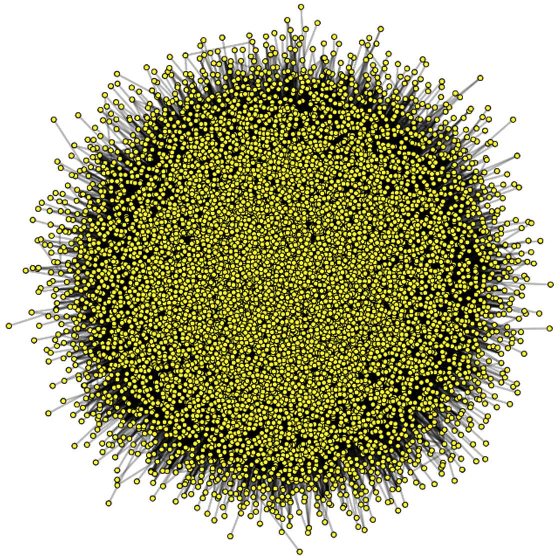


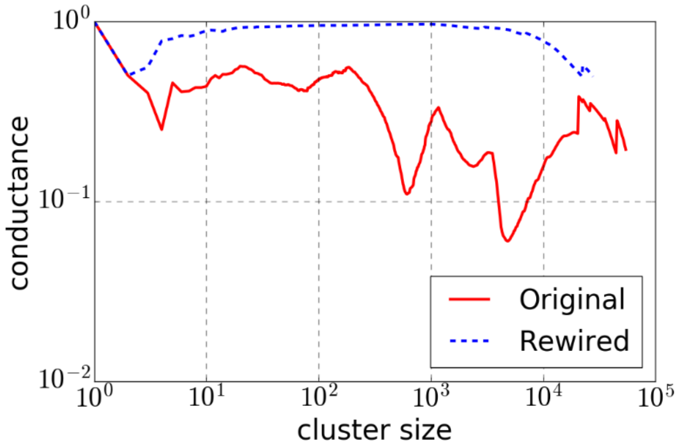


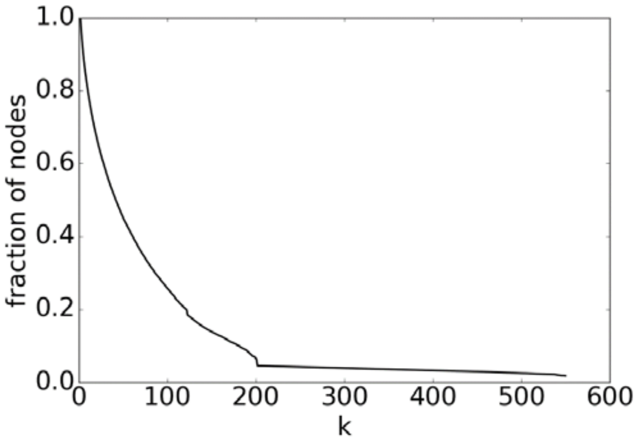
from H. Jeong et al Nature 411, 41 (2001):
Yeast protein interaction network

- New project with Siemens CT on Explainable AI
  - Analyze convolution matrices, to engineer the training data
  - Analyze decision paths in neural networks

LASAGNE: Locality And Structure Aware Graph Node Embedding





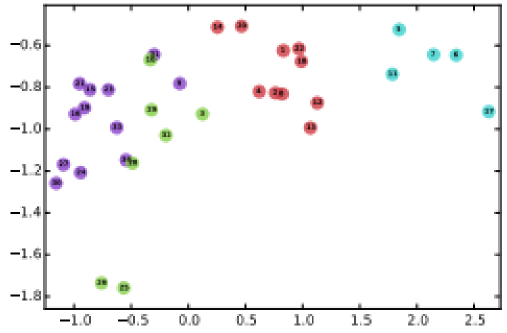


- Structure aware graph node embeddings
  - Explores neighborhood for each node individually
  - Shorter hop neighborhood for nodes in dense areas
  - Wider hop neighborhoods for nodes in sparse areas
  - Works better than previous methods for poorly structured networks

[Faerman, Borutta, Fountoulakis, Mahoney: ArXiv 2017]
Joint work with Berkeley Institute for Data Science

Representation of nodes in labeled graphs

(a) Input: Karate Graph

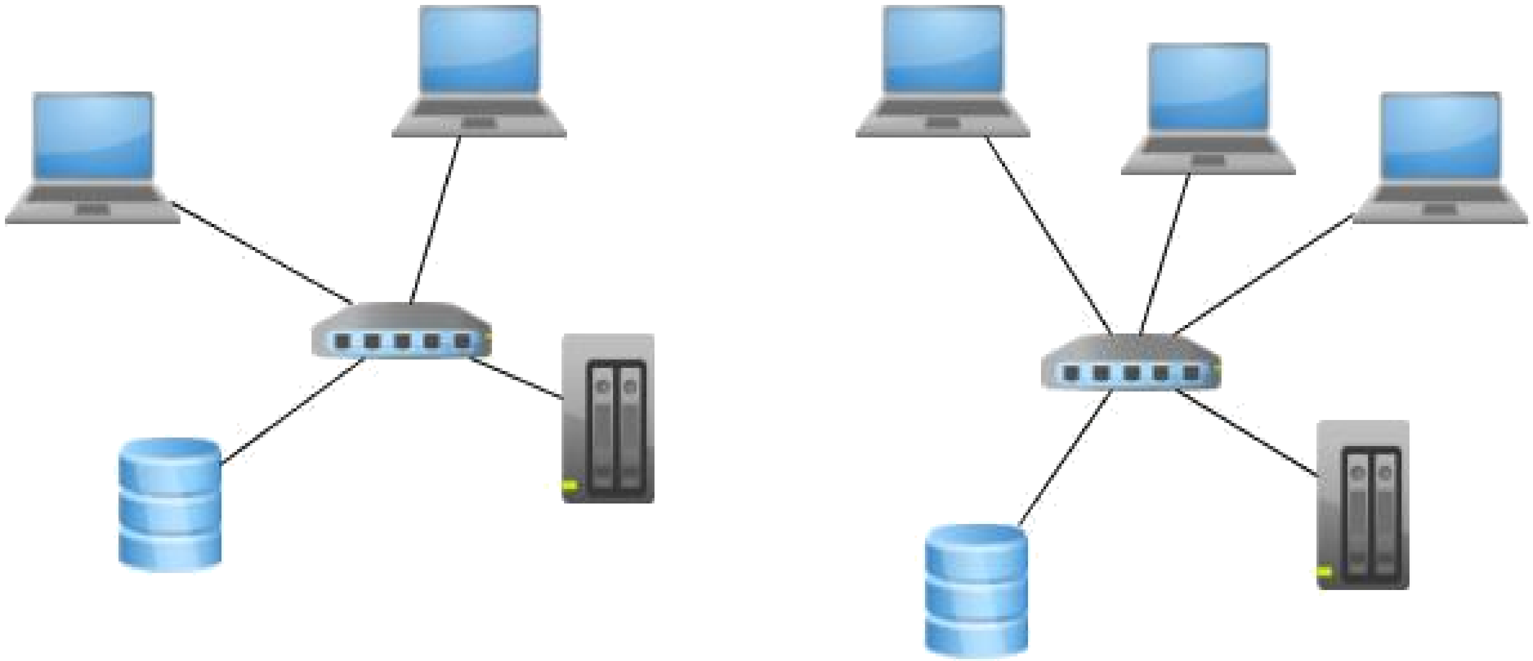


(b) Output: Representation

- Graphs naturally represent interactions
- Low dimensional vectors reflect roles of nodes
- Can be combined with other node features
- Useful for various tasks:
  - Node classification
  - Graph exploration
  - Link prediction

Perozzi, Al-Rfou,, Skiena. "Deepwalk: Online learning of social representations."

Semi-Supervised Learning on Graphs Based on Local Label Distributions



- Usually labels are used as goal of the prediction
- This work
  - Uses local label distributions from the nodes' neighborhoods to create node features
  - ML Models learn best combination of different neighborhood localities for individual nodes
  - Performs better or comparable to methods considering explicitly given node features

[Faerman, Borutta, Busch, Schubert: ~2018]

How about your data?

Prof. Dr. Thomas Seidl
LMU Munich
Data Science Lab

http://dsl.ifi.lmu.de