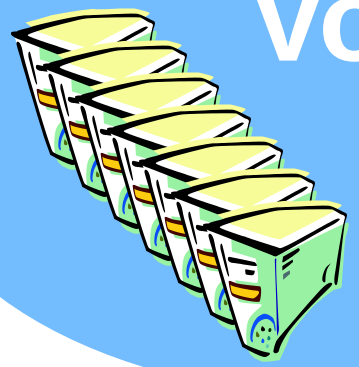
Big Data and Machine Learning

Prof. Dr. Thomas Seidl
LMU Munich, Chair of Database Systems and Data Mining

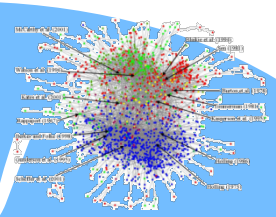
Nov. 22nd, 2018 | LRZ Symposium SuperMUC-NG | Garching

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

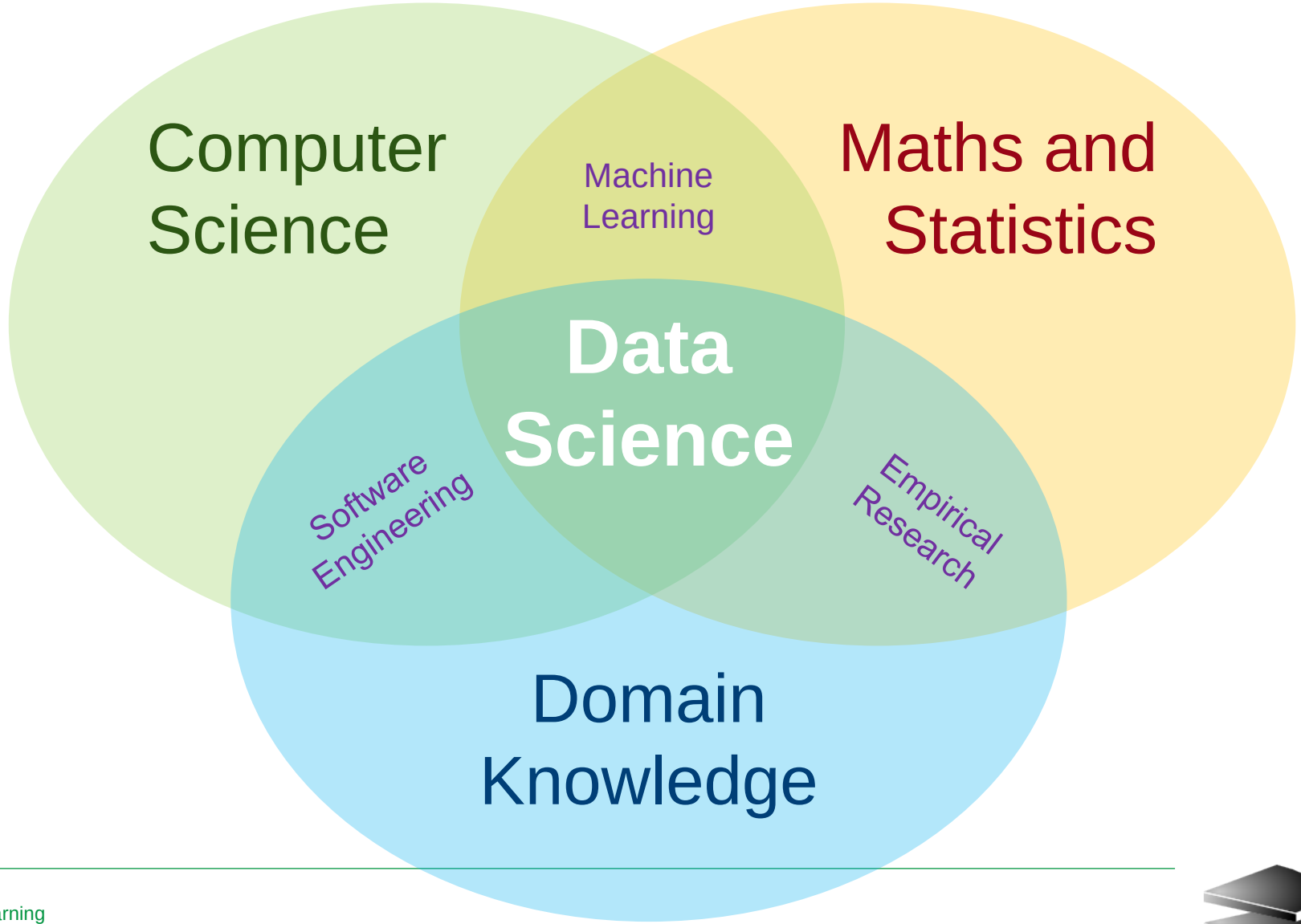
reliability, noise, trust, provenance

Value / Visual Analytics

patterns, rules, trends, outliers
data science

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Data Science Ingredients



Machine Learning – Tasks

Descriptive Learning

Better understanding

Pattern recognition
Clustering
Outlier detection

Data Mining

Predictive Learning

Better forecasts

Traffic prediction
Labeling
Fraud detection

Regression

Prescriptive Learning

Better actions

Predictive maintenance
Autonomous driving
Medical therapies

Artificial Intelligence

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

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

Competence Center and Doctoral Training

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

Solutions for Application Domains

- LRZ Competence Center on Big Data (2018, StMWK)
- Fraunhofer ADA-Center: IIS, FAU, LMU (2018, StMWi)

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Master Data Science

www.datascience-munich.de

- Funded by **Elite Network of Bavaria**
- 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

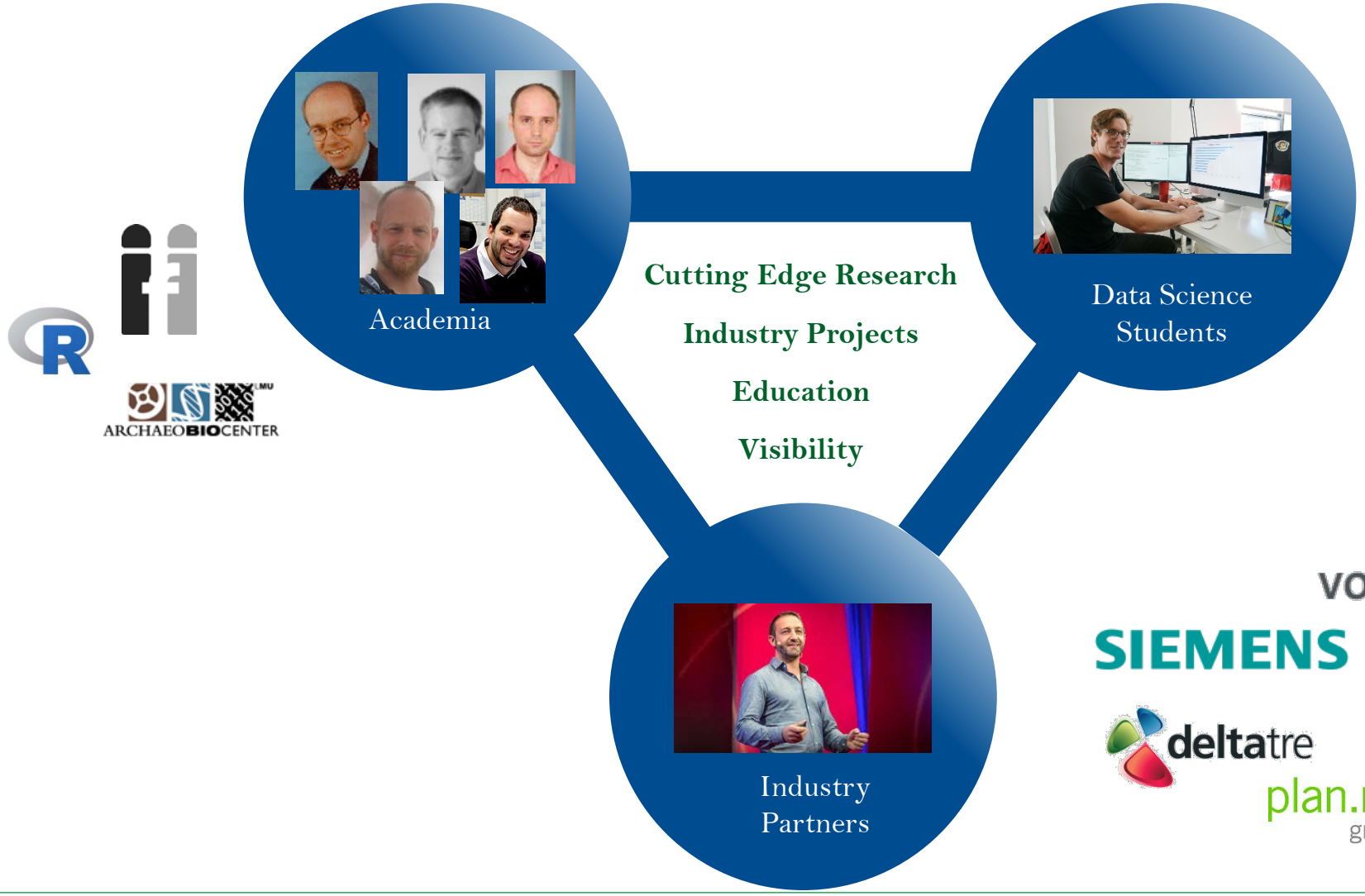
- International scope
  - Fully English spoken, small cohorts
  - Entrance profile: excellent grades for
    - ≥ 30 ECTS in Statistics
    - ≥ 30 ECTS in Computer Science

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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Data Science Lab @LMU: Working Space for Collaborations

Data Science Lab @LMU



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Munich Center for Machine Learning (MCML)

Bundesministerium für Bildung und Forschung

- Funded by BMBF (2018 – 2022 – 2025)
  - Berlin, Dortmund/St. Augustin, München, Tübingen
- Joint Initiative of Informatics and Statistics
  - 15 principal investigators from LMU and TUM
  - Directed by Thomas Seidl, Bernd Bischl, Daniel Cremers
- 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 LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Helmholtz Data Science Initiative

Munich School for
Data Science
@ Helmholtz, TUM & LMU

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LRZ Big Data Competence Center

Leibniz-Rechenzentrum
der Bayerischen Akademie der Wissenschaften

Customized Concepts	Consulting	User Support
Innovative Technologies	Hardware Resources	Big Data Infrastructure
Open Data	Continuing Education	Training

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Fraunhofer ADA Center

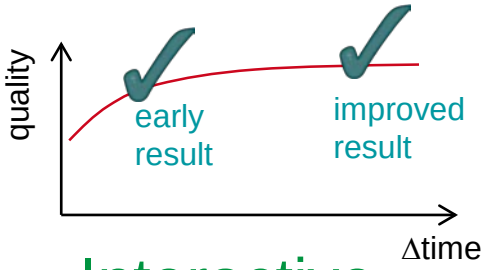
Fraunhofer IIS

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

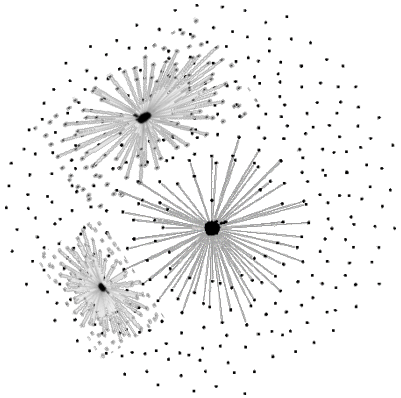
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

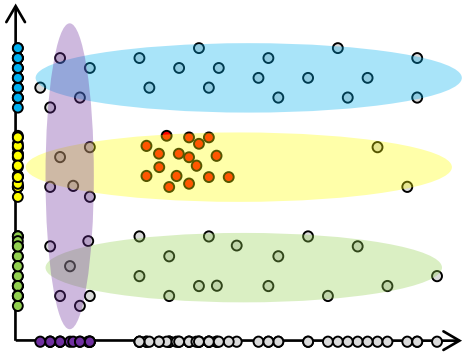
Some of Our Research Areas



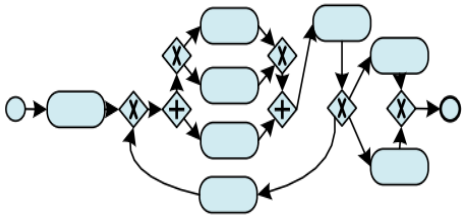
Interactive Analytics

Representation Learning

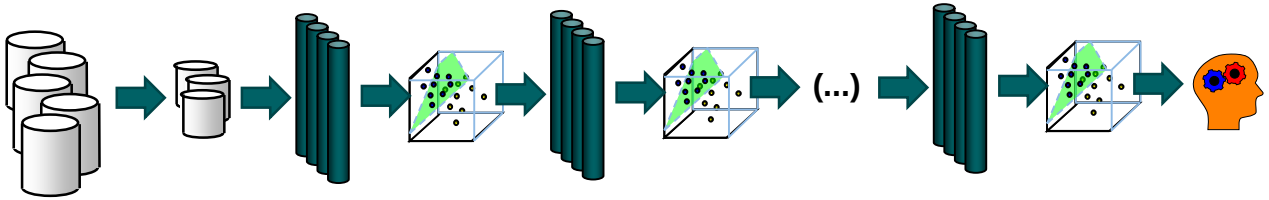




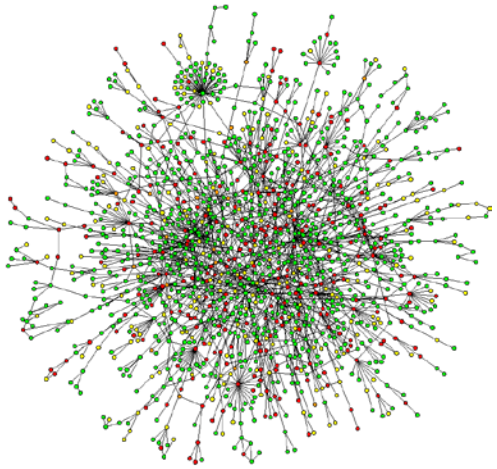
Explainable AI



Process Mining

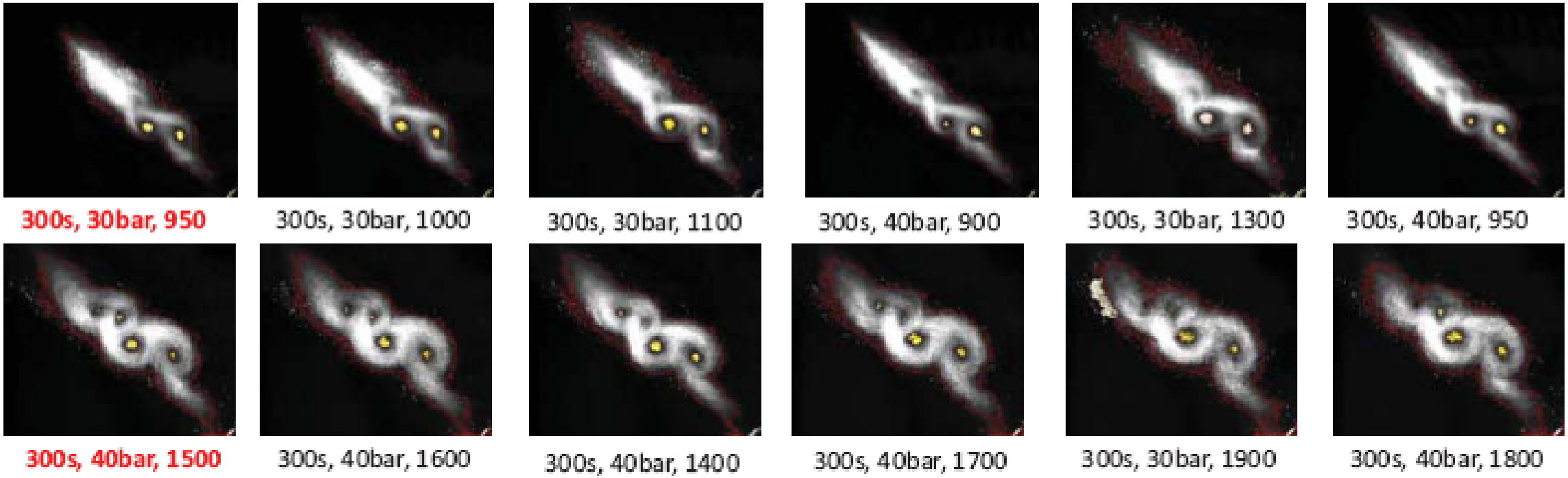


Deep Learning



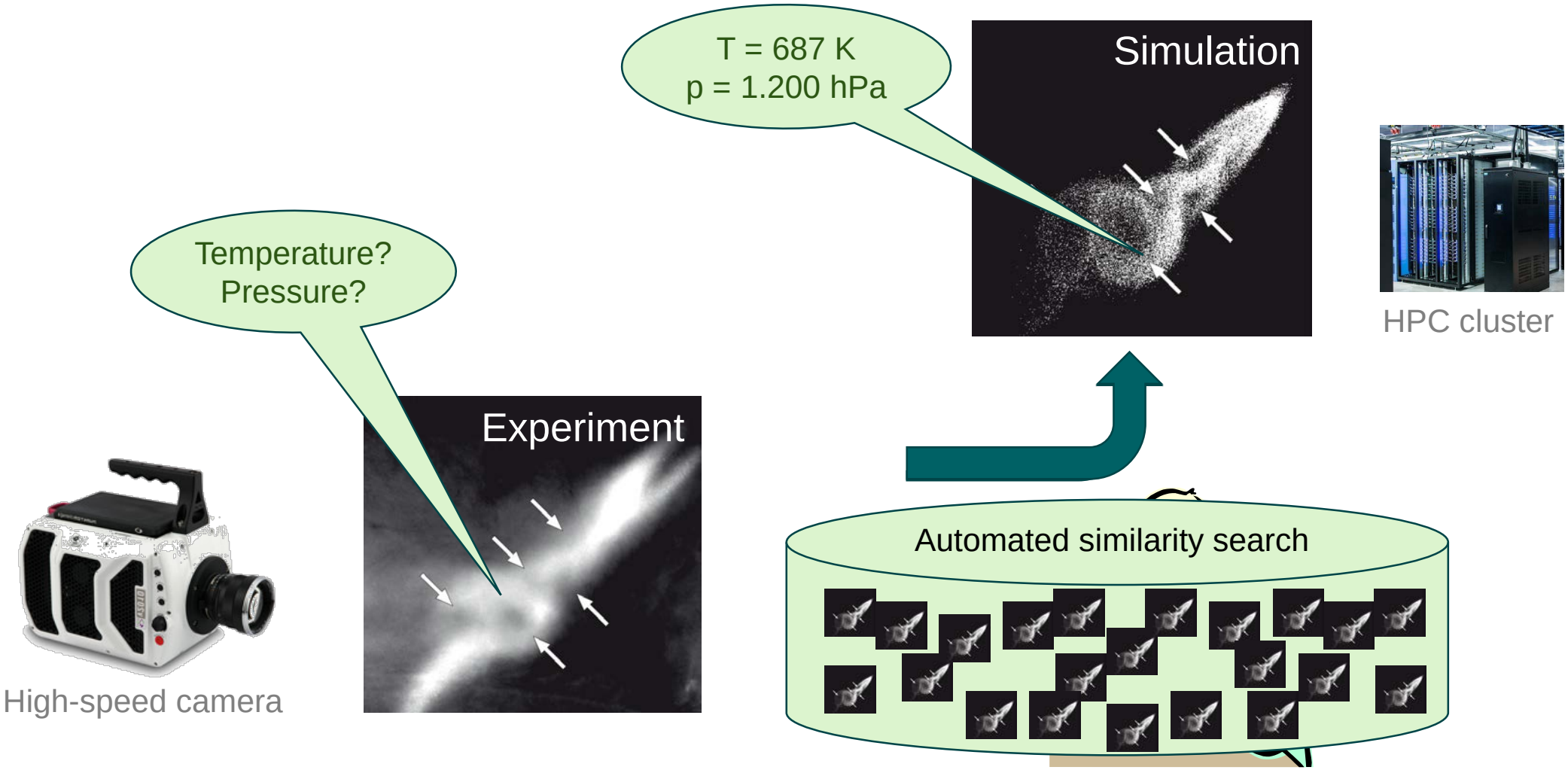
Knowledge Graphs

Similarity Search for Fuel Injection (SFB 686)



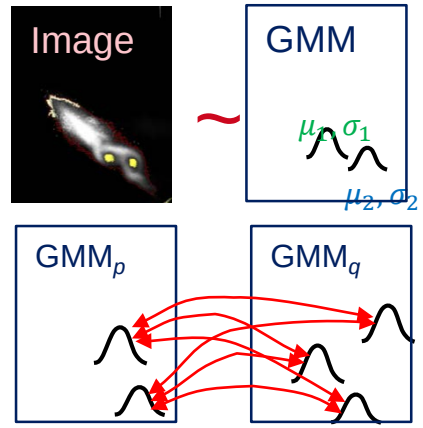
Spray vortex analysis in combustion engineering: compare experiments with simulations

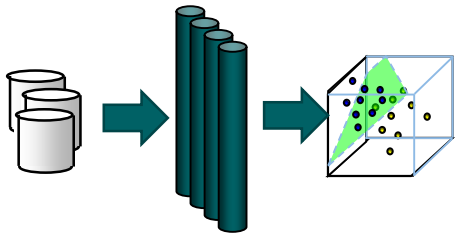
Virtual Sensors for Fuel Injection (SFB 686)



Beecks, Zimmer, Seidl, Martin, Pischke, Kneer: **SISAP** 2011
Beecks, Zimmer, Kirchhoff Seidl: **ICCV** 2011

Similarity Modeling





Classic methods

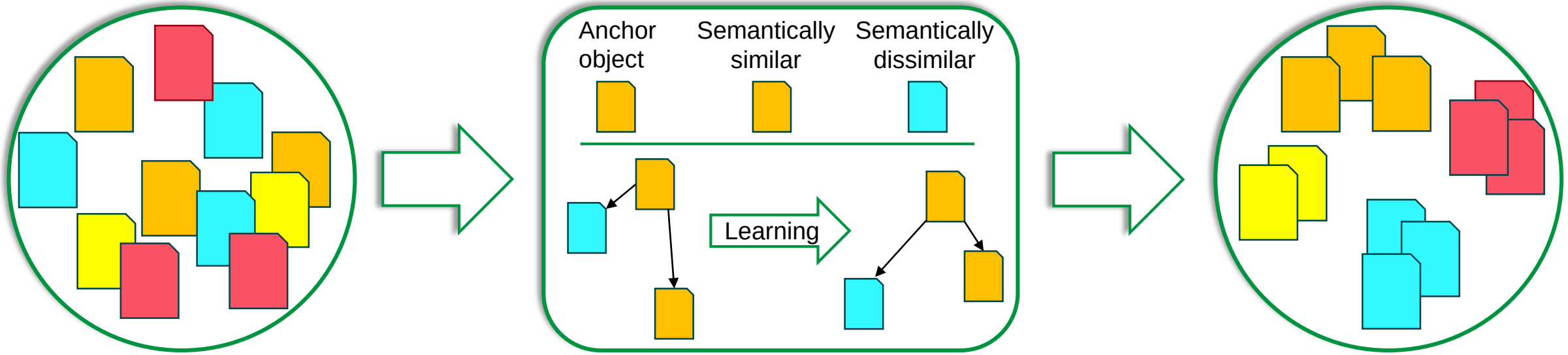
- Feature engineering: bag of words, term frequency, feature signatures, …
- Similarity functions, distance functions

Neural learning

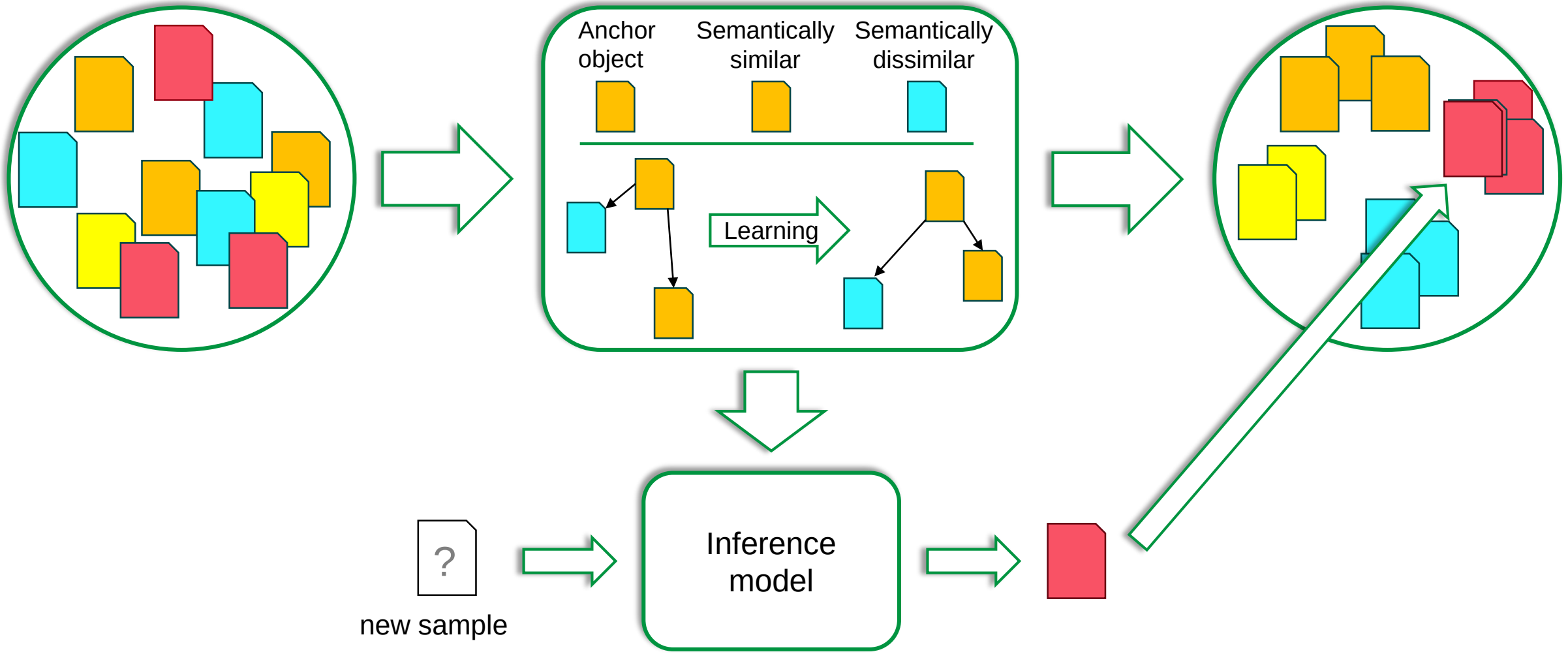
- Representation learning, metric learning
- Learn analogies of similar objects, e.g. by Siamese networks [AT&T 1993] [LMU 2015]

[Bromley, Guyon, Le Cun, Sackinger, Shah: NIPS 1993]
[Yin, Schütze, Xiang, Zhou: arXiv 2015]

Similarity Learning Through Embedding



Similarity Learning Through Embedding



Similarity Learning – Are There Labels Available?

Many

- Focus on given concepts
- Supervised learning

Some few

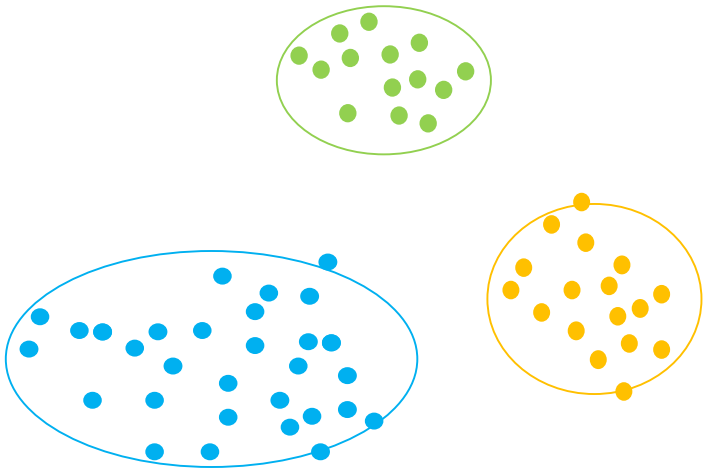
- Focus on few hidden concepts
- Semi-supervised learning

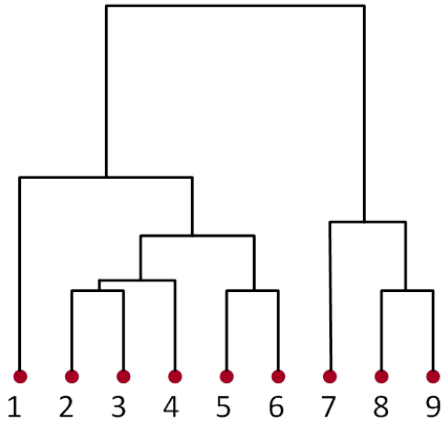
None

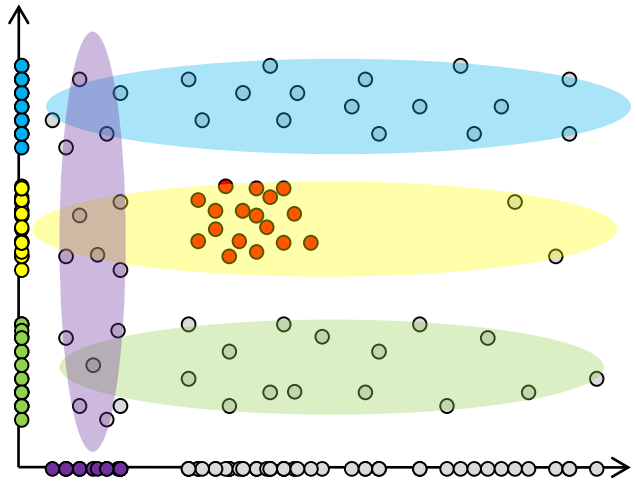
- Many hidden concepts
- Unsupervised learning

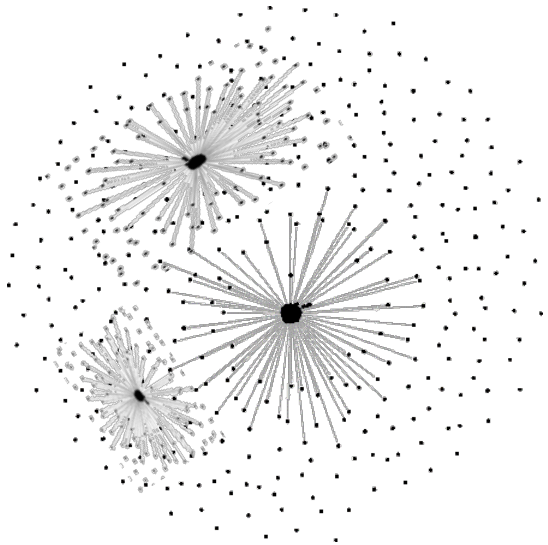
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Clustering



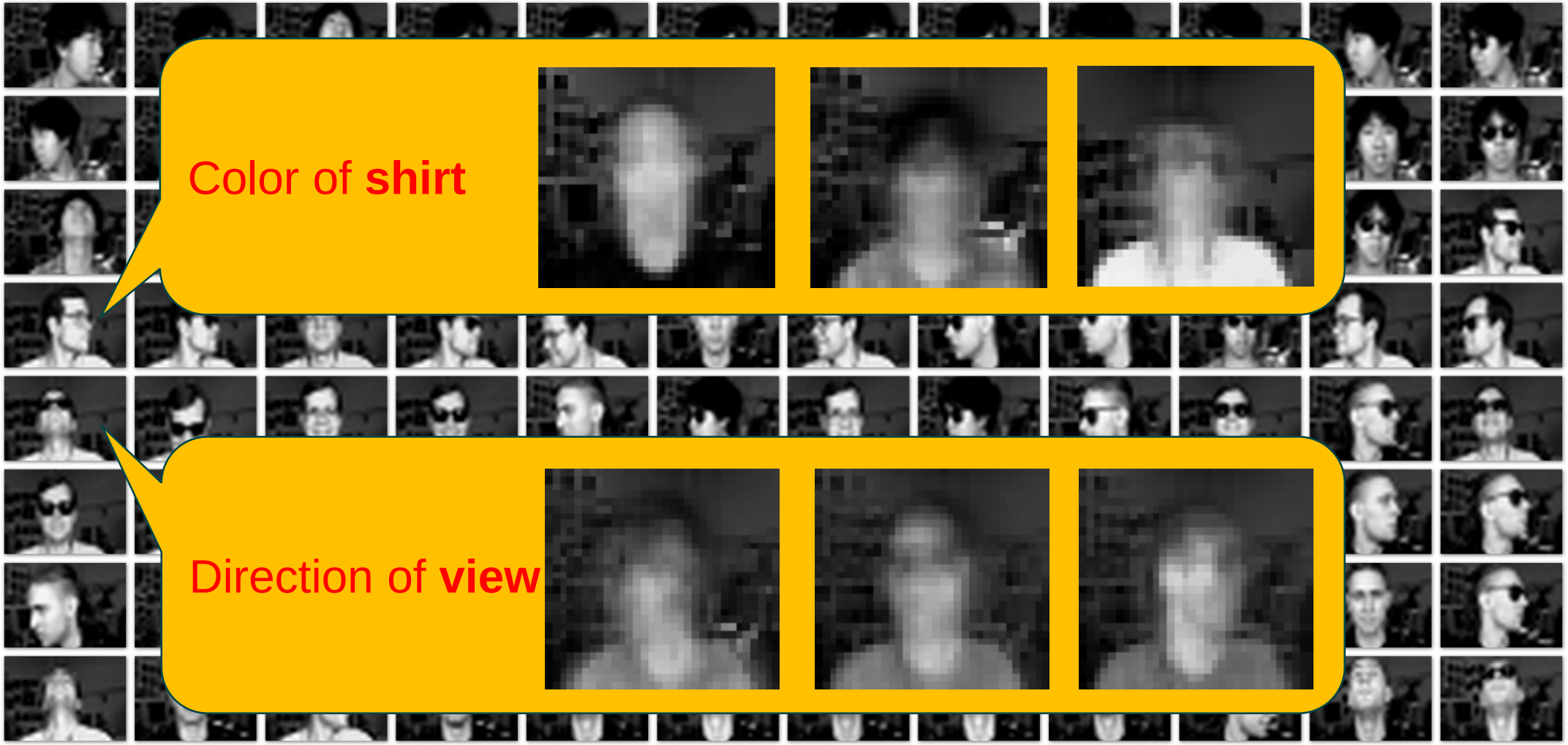






- Customer Segmentation, Labeling Products, Clique Detection, …
  - Clustering for heterogeneous objects
  - Subspace clustering, density estimation for higher subspaces
  - Non-linear Correlation Clustering
  - Semi-supervised clustering, constraints models
  - …

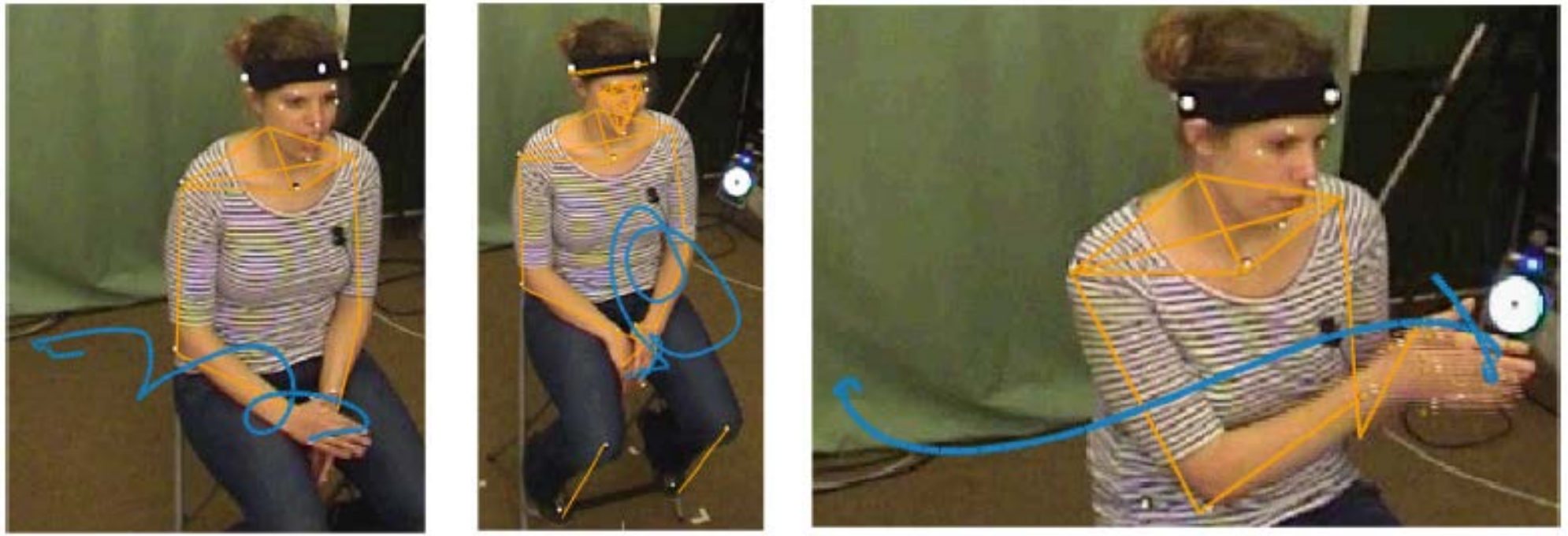
Alternative Clustering, Multi-view Labeling



[Günnemann, Färber, Seidl: **KDD** 2012]
[Günnemann, Färber, Rüdiger, Seidl: **KDD** 2014]

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

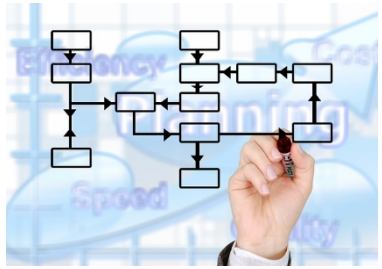
Labeling of Co-speach Gestures



[Schüller, Beecks, Hassani, Hinnell, Brenger, Seidl, Mittelberg: **Göttingen Dialog in Digital Humanities** 2015] – *best paper award*
[Beecks, Hassani, Hinnell, Schüller, Brenger, Mittelberg, Seidl: **SSTD** 2015]

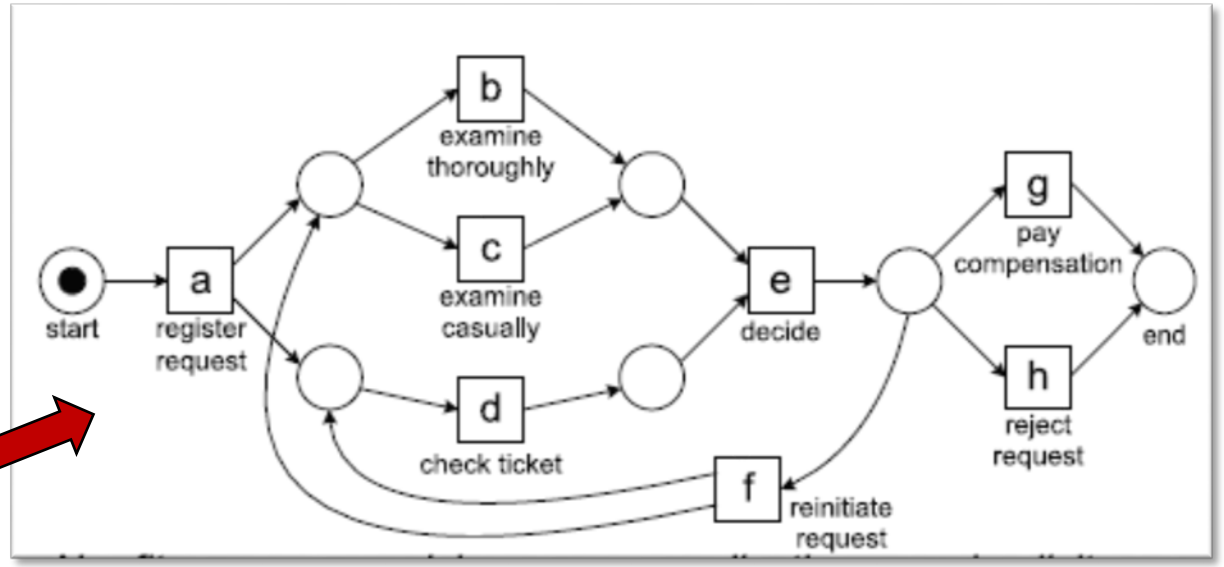
- Gestures are a non-verbal modality of human expression
- Supervised: Recognition of known gestures from dictionary
- Unsupervised: Extraction of frequent patterns to hypothesize gestures

Process Mining



time	case	event
2018-6-6-6:29	732	a
2018-6-6-6:32	744	a
2018-6-6-6:33	732	b
2018-6-6-6:34	728	a
2018-6-6-6:35	732	d
2018-6-6-6:37	744	b
2018-6-6-6:38	728	c
2018-6-6-6:39	751	a
2018-6-6-6:42	744	d
2018-6-6-6:43	732	d
2018-6-6-6:44	744	e
2018-6-6-6:45	751	c
2018-6-6-6:47	732	e
2018-6-6-6:48	744	g
2018-6-6-6:59	751	d
2018-6-6-7:02	751	e
2018-6-6-7:03	728	e
2018-6-6-7:04	768	a
2018-6-6-7:05	751	h
2018-6-6-7:07	768	c
2018-6-6-7:08	728	h
2018-6-6-7:09	732	g
2018-6-6-7:12	768	d
2018-6-6-7:13	779	a
2018-6-6-7:14	768	e
2018-6-6-7:15	779	b
2018-6-6-7:17	768	h
2018-6-6-7:18	779	d

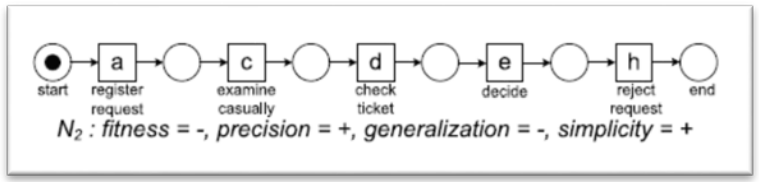
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

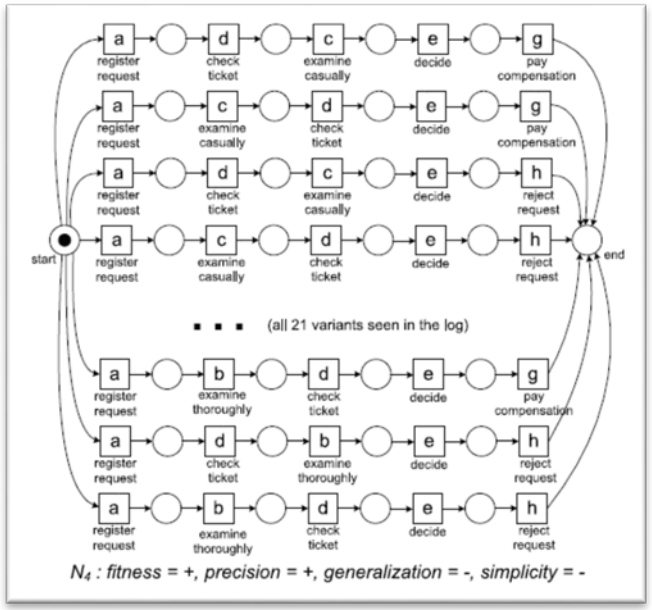


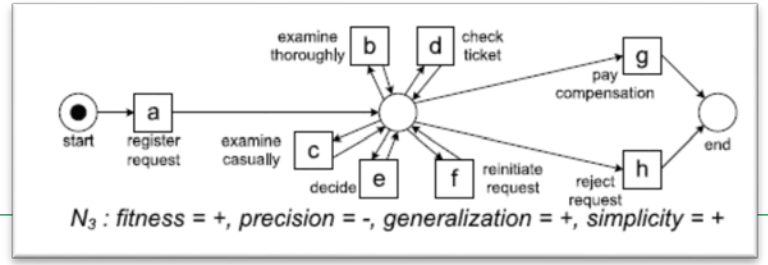
- Task: Extract process model from log entries which
  - … is able to replay the log ⇒ $Fitness$
  - … simplifies as far as possible ⇒ $Simplicity$
  - … does not overfit the log ⇒ $Generalization$
  - … does not underfit the log ⇒ $Precision$

Process Discovery: Tune Generalization Granularity

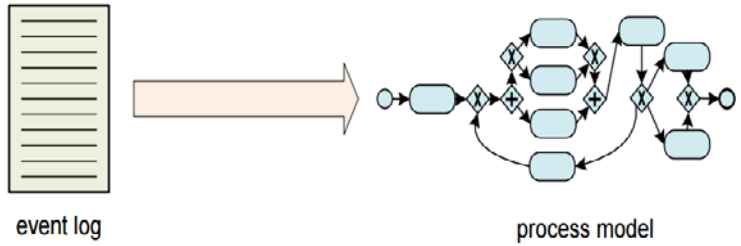
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	





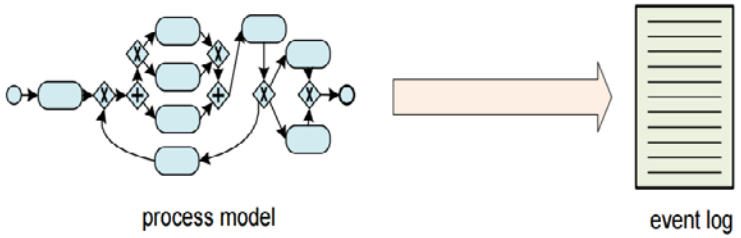


Process Mining: Towards Holistic Analytics in Industry 4.0 Environments



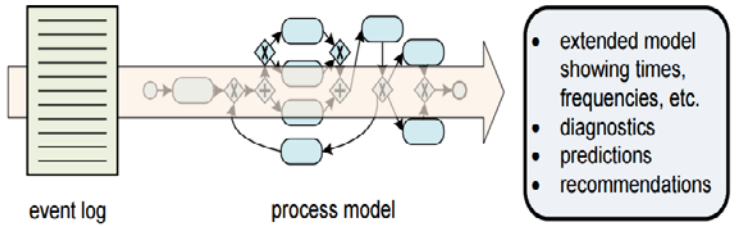
Process Discovery

„Play-In“



Conformance Check

„Play-Out“



Process Enhancement

„Replay“

Applications

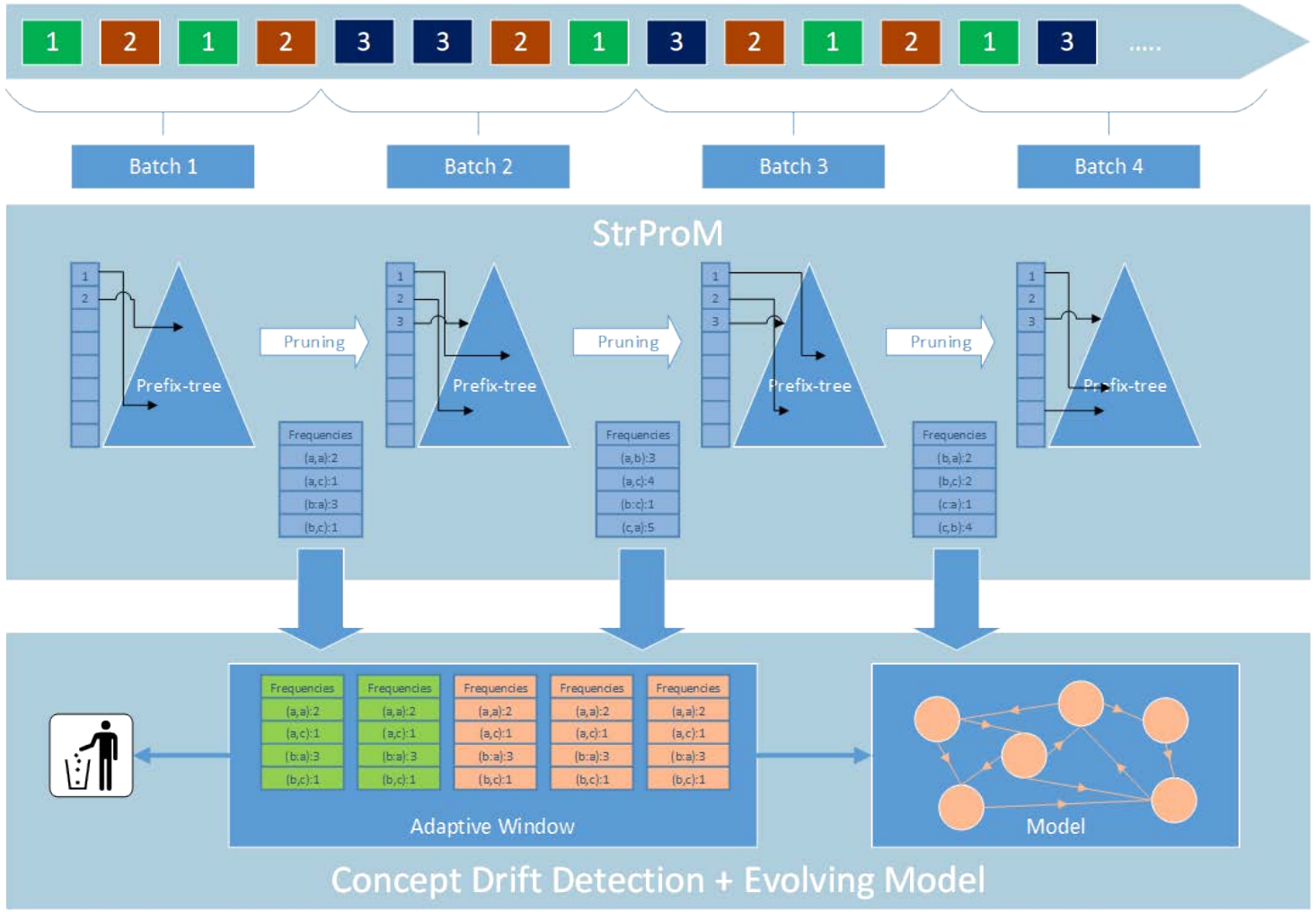
- Fleet management
- Monitoring of train schedules
- Predictive maintenance for mechanical parts in use
- Monitoring of production processes

Challenges for Process Mining on Complex Events and Cases

- Multi-source data descriptions
- Multimodal and heterogeneous data
- Spatio-temporal contexts
- Uncertainty in object representations
- Evolution of models over time

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Stream Process Mining



Event Stream

Batched Approach

Prefix-Trees

Irregular Updates

Decaying

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

Big Data Technology for Machine Learning

- Distributed Processing on Hadoop Distributed File System HDFS
  - Hadoop MapReduce
  - Apache Spark
  - Apache Flink
- Graph and Network Analysis
  - Pregel, Giraph, GraphX, Gelly
- GPU cluster computing
- New interaction models, explainable AI
  - interactive data mining, incremental algorithms
  - Visual Analytics

https://www.lrz.de/presse/fotos/

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

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

How About Your Data?