Master-Thesis

Data Mining in Industrial Processes:
Evaluation of different machine learning models for product quality prediction

ANGEWANDTE INFORMATIK IV
UNIVERSITÄT BAYREUTH

Andreas Braun 1200197
Master Applied Computer Science
INF301: Master-Thesis (30LP)

Overview

- Motivation
  - Introduction and Quality Goal
  - Advantages of Machine Learning and Requirements
  - Manufacturing Process – Data Understanding and Data Partitioning
- Data Processing and Dataset Description
- Analytical Processing: Prediction of Quality
  - Selection of Discipline and Algorithms
  - Exploration of Algorithm types
  - Evaluation of best Algorithms
  - Combination of Algorithms

Introduction

Process Industry Data Usage

Passive Use

- Technical checks
- Log requirements

Active Use

- Knowledge extraction
  → competitive advantage

Past → Present

Goal:
Further improvement of the quality control mechanism

- Customers expect products free from error
- Less (financial) compensation in case of defect

Quality in Manufacturing (Quality Goal)

Product developement stages:

- Manufacturing process directly adds value → quality analysis here

- Associated quality task:

Goal:
Prediction of quality

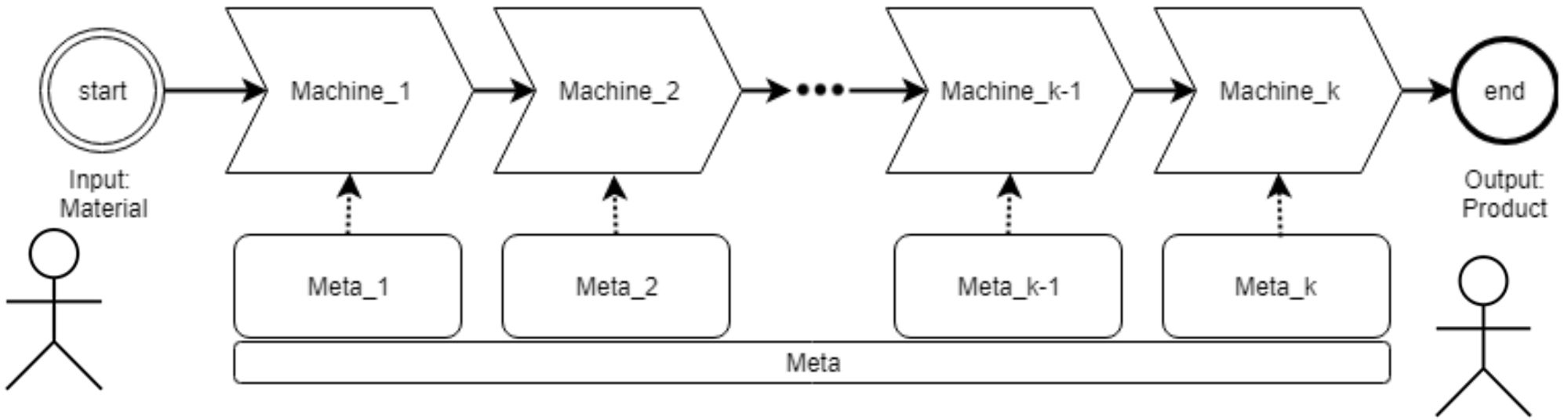
- Core question: Final product quality, good or poor?

Advantages of Machine Learning (ML) and Requirements

- Advantages of ML over traditional methods:
  - High dimensionality of the feature space manageable
  - Automatic identification and removal of failed products (vs batch wise)

- Requirements:
  - General approach --> useable for all articles
  - Assistance system for the workers
  - Data driven knowledge creation
  - Complement exisiting techniques with a ML approach

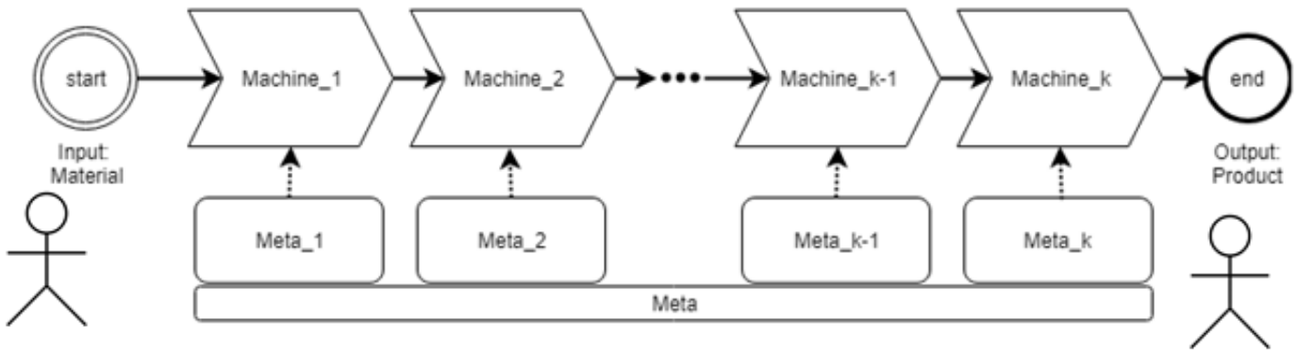
Basic Data Understanding – Manufacturing Process

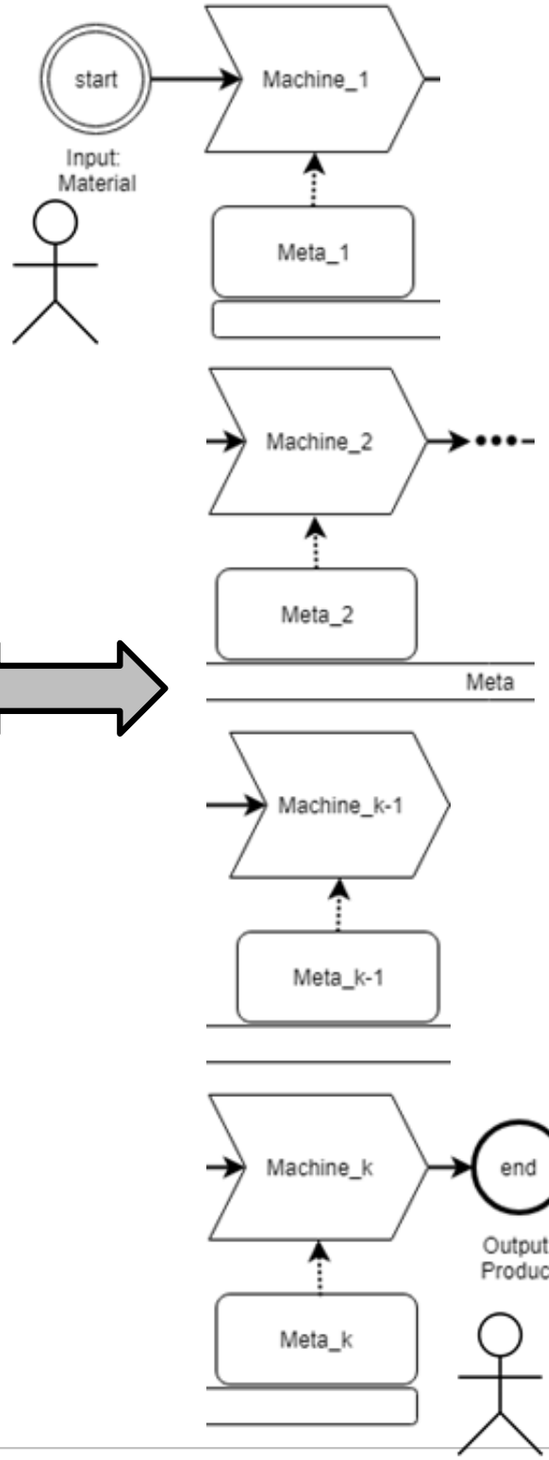


Properties of the manufacturing process:

- Plastic profile extrusion of window profiles
- Linear, continuous production process
- Product quality is inherently connected to the production line itself

Data Partitioning – Manufacturing Process





- Multiple Time Series stand for one Machine with Meta Information
- Relevant field: Time Series Data Mining (TSDM)

$$Y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_j(t) \\ \vdots \\ y_d(t) \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1i} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2i} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{j1} & y_{j2} & \cdots & y_{ji} & \cdots & y_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{d1} & y_{d2} & \cdots & y_{di} & \cdots & y_{dn} \\ t_1 & t_2 & \cdots & t_i & \cdots & t_n \end{bmatrix}$$

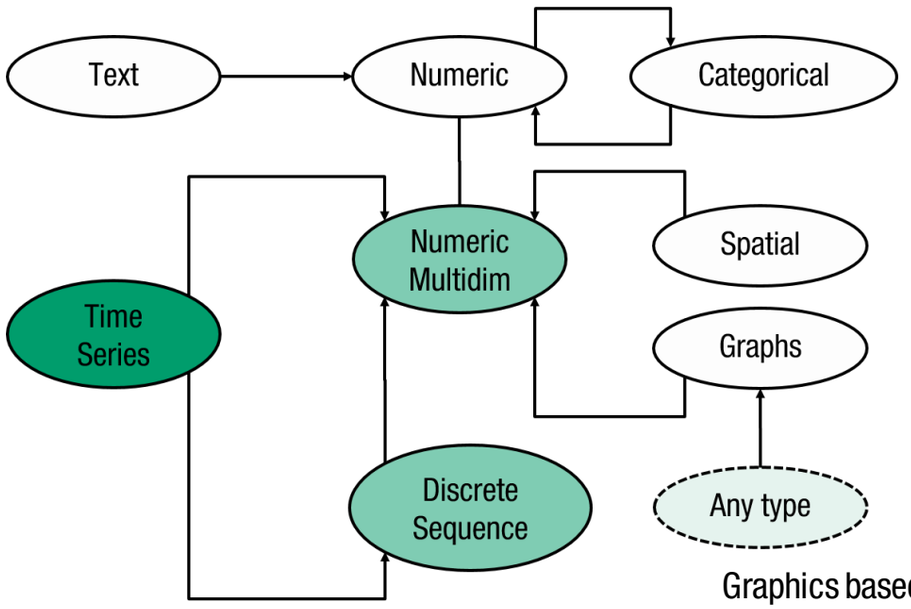
Horizontal & <u>vertical</u> approach

==

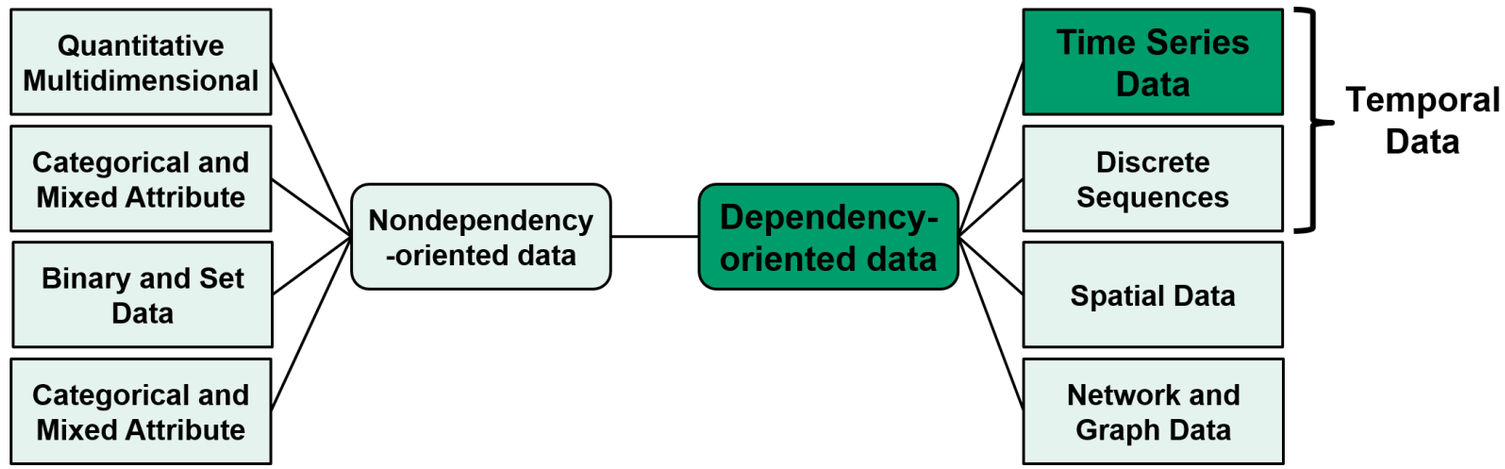
Time series based & <u>state based</u>

Data Types of the vertical and horizontal partitioning

Possible Data Type Conversions



Examples (Non)Dependency Data



Graphics based on textual description of
[C. Aggarwal. Data Mining The Textbook. Springer, 2015.]

- Vertical state based = nondepedency oriented + multidimensional data type

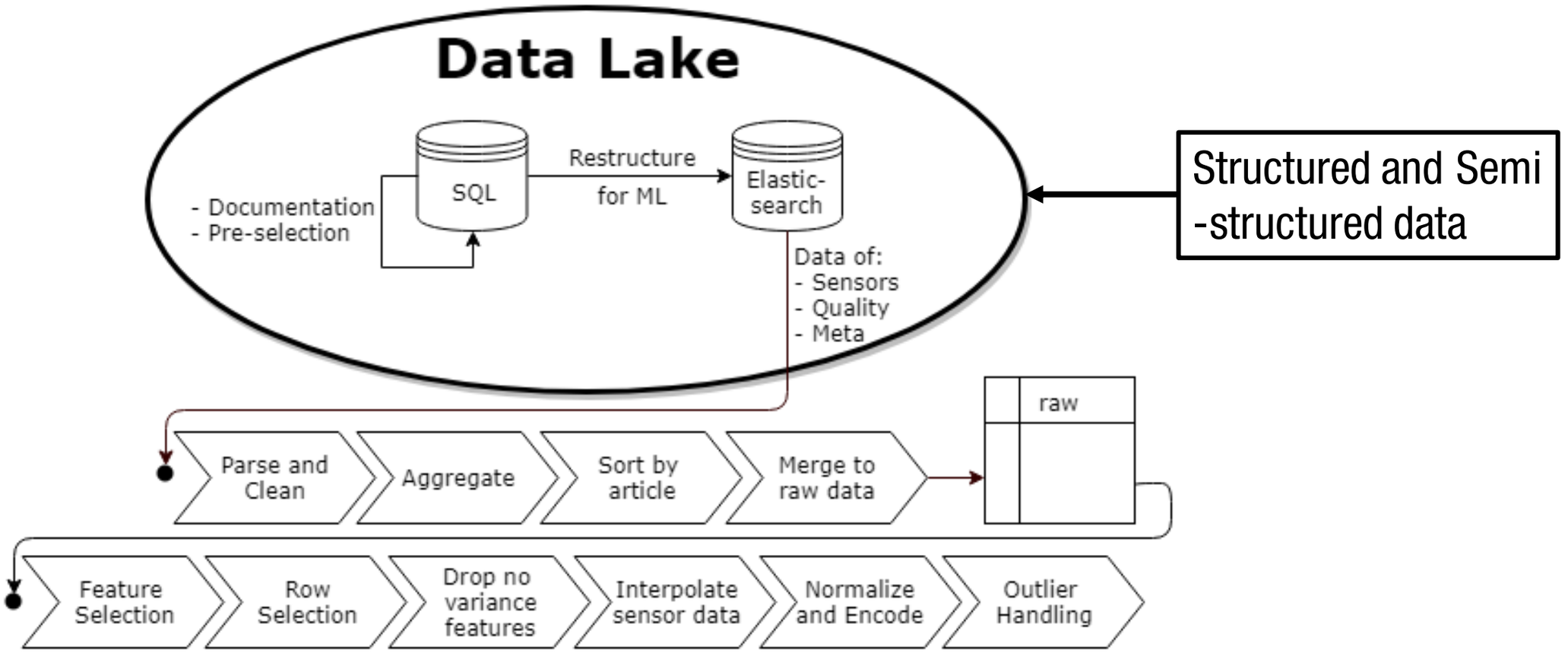
$$Y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_j(t) \\ \vdots \\ y_d(t) \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1i} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2i} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{j1} & y_{j2} & \cdots & y_{ji} & \cdots & y_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{d1} & y_{d2} & \cdots & y_{di} & \cdots & y_{dn} \\ t_1 & t_2 & \cdots & t_i & \cdots & t_n \end{bmatrix}$$

- Horizontal time series based = dependency oriented + time series data type

Overview

- Motivation
  - Introduction and Quality Goal
  - Advantages of Machine Learning and Requirements
  - Manufacturing Process – Data Understanding and Data Partitioning
- Data Processing and Dataset Description
- Analytical Processing: Prediction of Quality
  - Selection of Discipline and Algorithms
  - Exploration of Algorithm types
  - Evaluation of best Algorithms
  - Combination of Algorithms

Data Processing



A. Extract, Transfer, Load (ETL): From Data Lake to flat ML-raw-file

B. Preprocessing: Feature selection, Interpolation, Normalization, Encoding, Outlier handling, …

C. Analytical Processing

Dataset Description: Representative Article

~100 cols
400k rows

	M996_S99	M996_S100	M996_S101	T0	T1	T2_1.0	T2_2.0	T2_3.0	T2_4.0	T3_1615.0
2018-04-07 08:55:18	1.04	-2.23	0.24	11.20	1.04	0	0	1	0	0
2018-04-07 08:55:19	1.01	-2.23	0.24	11.20	1.04	0	0	1	0	0
2018-04-07 08:55:20	0.98	-2.23	0.24	11.20	1.04	0	0	1	0	0

- ETL and Preprocessing is finished
- Dataset is specific for the prodcution line and the produced article --> representative article
- Sensor values S_i: temperature, pressure, weight, speed, effective power,…
- Meta information T_i: material number, variant number, garbage production, recycled material, …

Overview

- Motivation
  - Introduction and Quality Goal
  - Advantages of Machine Learning and Requirements
  - Manufacturing Process – Data Understanding and Data Partitioning
- Data Processing and Dataset Description
- Analytical Processing: Prediction of Quality
  - Selection of Discipline and Algorithms
  - Exploration of Algorithm types
  - Evaluation of best Algorithms
  - Combination of Algorithms

Analytical Processing: Prediction of Quality

- Most basic information about quality available (error time, production time)

--> *Discipline of Classification (Binary)*

  - Performance determination by scoring function
    - Assistance system: No unnecessary Alarms
    - Complement exisiting methods: Be sure about bad quality detection
    - Class imbalance present in the dataset

--> *Primary score of Precision and secondary score of Recall*

- Precision = TP / (TP+FP)
- Recall = TP / (TP+FN)

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

Selection of Algorithms for Exploration

- Overview about popular ML algorithms in the process industry
- Data Science expert knowledge of supervisor at Rehau AG
- Availability of stable implementations in major frameworks
- Explainability of models

Optional step

- principal component analysis (PCA)
- independent component analysis (ICA)

- gaussian naive bayes (GNB)
- k nearest neighbour (KNN)
- logistic regression (LR)
- linear support vector machine (SVM)
- random forest (RF)

Overview

- Motivation
  - Introduction and Quality Goal
  - Advantages of Machine Learning and Requirements
  - Manufacturing Process – Data Understanding and Data Partitioning
- Data Processing and Dataset Description
- Analytical Processing: Prediction of Quality
  - Selection of Discipline and Algorithms
  - Exploration of Algorithm types
  - Evaluation of best Algorithms
  - Combination of Algorithms

Analytical Processing

- Goal: Find good working ML pipeline
  - Challenge: No Free Lunch Theorem --> What model to use?

$$Y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_j(t) \\ \vdots \\ y_d(t) \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1i} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2i} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{j1} & y_{j2} & \cdots & y_{ji} & \cdots & y_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{d1} & y_{d2} & \cdots & y_{di} & \cdots & y_{dn} \\ t_1 & t_2 & \cdots & t_i & \cdots & t_n \end{bmatrix}$$

Analytical Processing: Prediction of Quality

1. Exploration of algorithm types: Use a subset (10% of the data) --> best algorithm type
2. Evaluation of best Algorithms: Best algorithms on the entire dataset
3. Combination of Algorithms: Time series based + state based algorithm

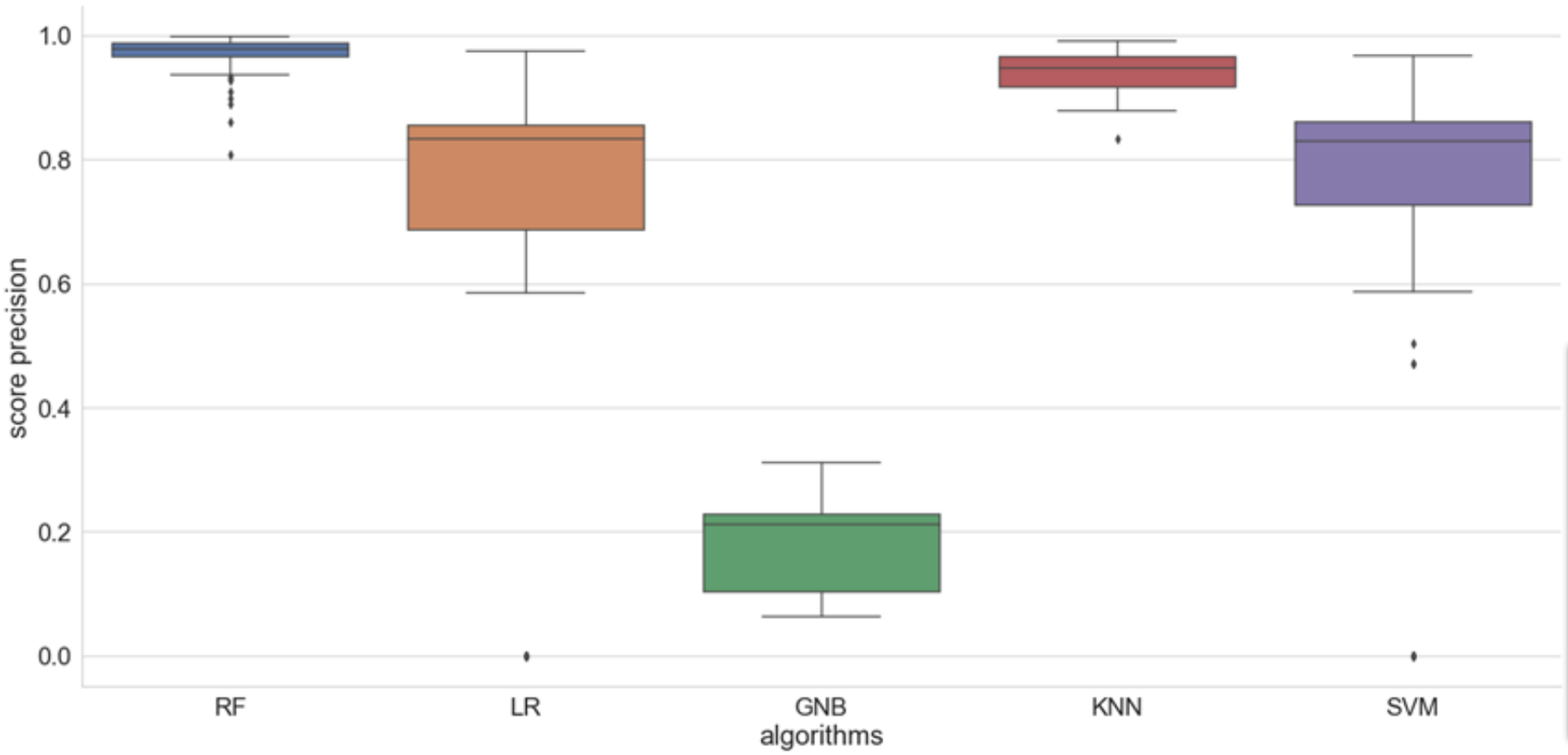
Tab. 4.1: Cases for algorithm exploration.

Case	Approach	Scoring Function
1	state based	precision
2	state based	recall
3	time series based	precision
4	time series based	recall

Tab. 4.6: Evaluation cases for the two best algorithms.

Case	Approach	Scoring Function
5	Best of state based	precision
6	Best of state based	recall
7	Best of time series based	precision
8	Best of time series based	recall

State-based: 1. Exploration on Precision (Case 1)



100 algorithms each with
5 times cross validation which
Results in 2500 times fitting
Onto the exploration dataset

Total population	Condition positive	Condition negative
Predicted condition positive	**True positive**, Power	**False positive**, Type I error
Predicted condition negative	**False negative**, Type II error	**True negative**

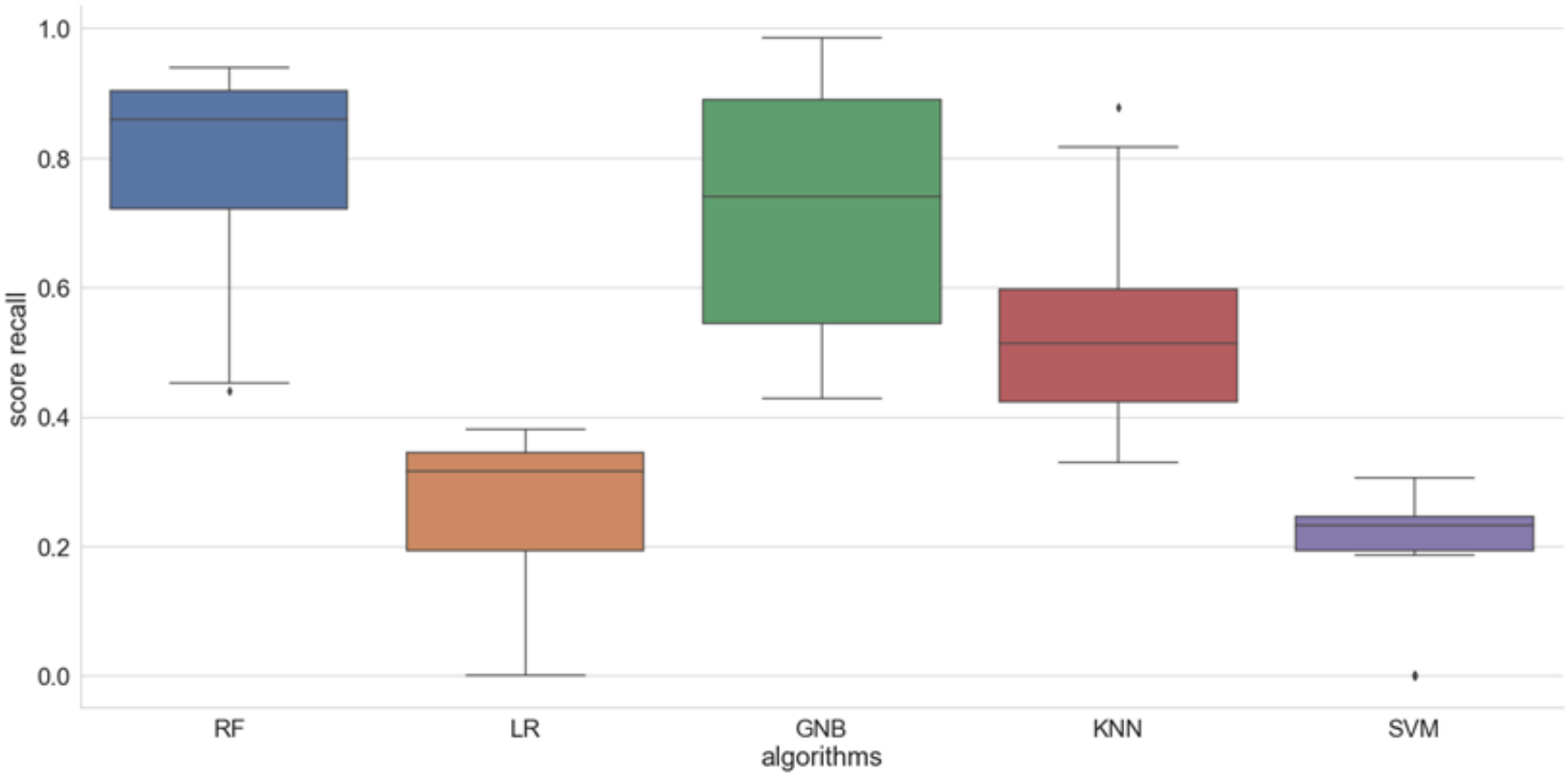
Optional use of PCA / ICA

Top 3 precision (=TP/(TP+FP)):

- RF
- KNN
- LR

	RF	LR	GNB	KNN	SVM
mean	0.9705	0.6917	0.1736	0.9398	0.7156
min	0.8076	0.0000	0.0632	0.8340	0.0000
25%	0.9662	0.6860	0.1021	0.9161	0.7263
50%	0.9790	0.8326	0.2112	0.9478	0.8287
75%	0.9870	0.8554	0.2278	0.9661	0.8608
max	0.9977	0.9742	0.3110	0.9916	0.9664

State-based: 1. Exploration on Recall (Case 2)



Total population	Condition positive	Condition negative
Predicted condition positive	**True positive**, Power	**False positive**, Type I error
Predicted condition negative	**False negative**, Type II error	**True negative**

Optional use of PCA / ICA

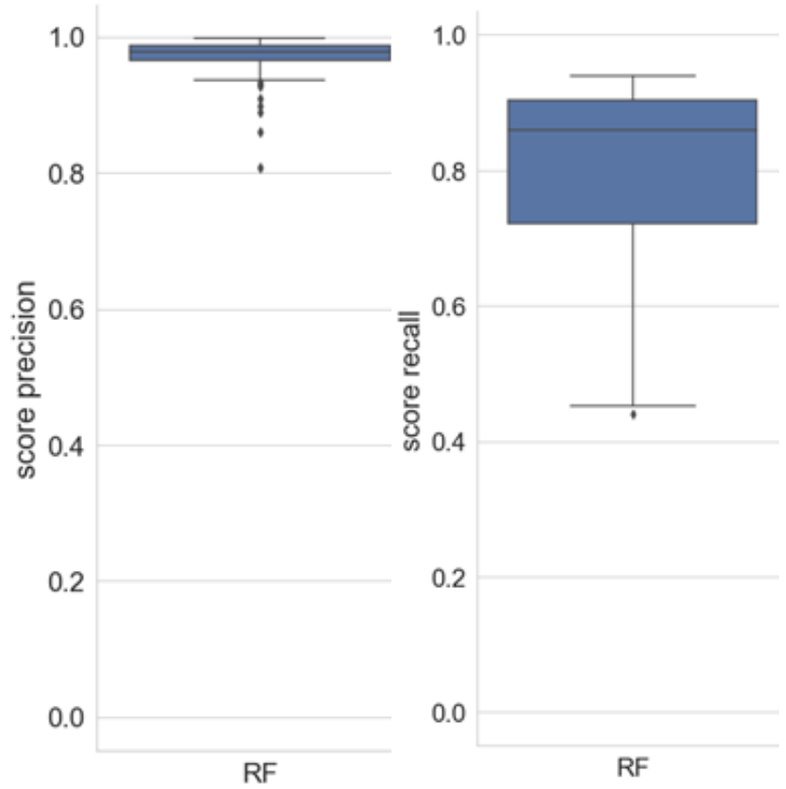
Top 3 recall (=TP/(TP+FN)):

- RF
- GNB
- KNN

	RF	LR	GNB	KNN	SVM
mean	0.8019	0.2497	0.7353	0.5221	0.2053
min	0.4397	0.0000	0.4276	0.3300	0.0000
25%	0.7204	0.1924	0.5439	0.4232	0.1924
50%	0.8586	0.3152	0.7404	0.5132	0.2316
75%	0.9032	0.3443	0.8897	0.5965	0.2462
max	0.9397	0.3805	0.9846	0.8772	0.3059

State-based: 2. Best Algorithm (Cases 5+6)

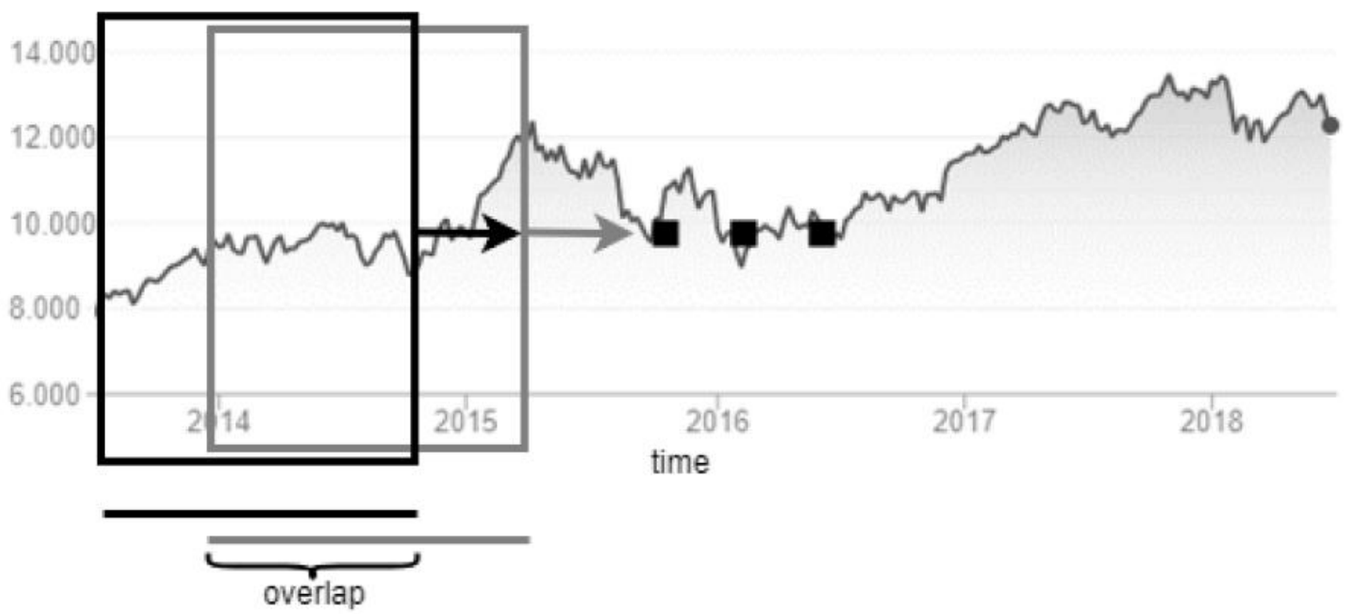
	RF_precision	RF_recall			
RF0	0.9977	0.4397	**RF5**	0.9918	0.8722
RF1	0.9951	0.5482	**RF6**	0.9911	0.4885
RF2	0.9949	0.7494	**RF7**	0.9910	0.9024
RF3	0.9944	0.6749	**RF8**	0.9899	0.8657
RF4	0.9920	0.8838	**RF9**	0.9895	0.6201



- Entire dataset with cross validation cv=10:

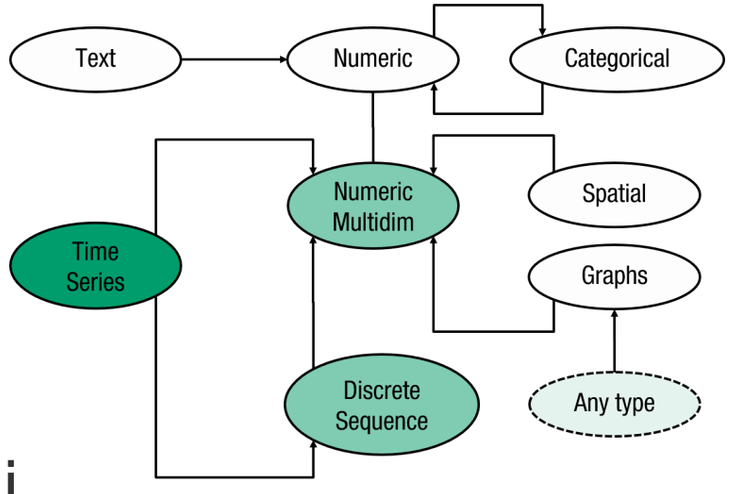
RF2 (state b.)	Precision (Case 5)	Recall (Case 6)
CV Score	99.83% (+/-00.17)% [99.66, 100]%	98.83% (+/-00.63)% [98.20, 99.46]%
Test Score	99.97%	98.99%

Time Series based: Windowing



$$Y(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \\ \vdots \\ y_j(t) \\ \vdots \\ y_d(t) \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1i} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2i} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{j1} & y_{j2} & \cdots & y_{ji} & \cdots & y_{jn} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{d1} & y_{d2} & \cdots & y_{di} & \cdots & y_{dn} \\ t_1 & t_2 & \cdots & t_i & \cdots & t_n \end{bmatrix}$$

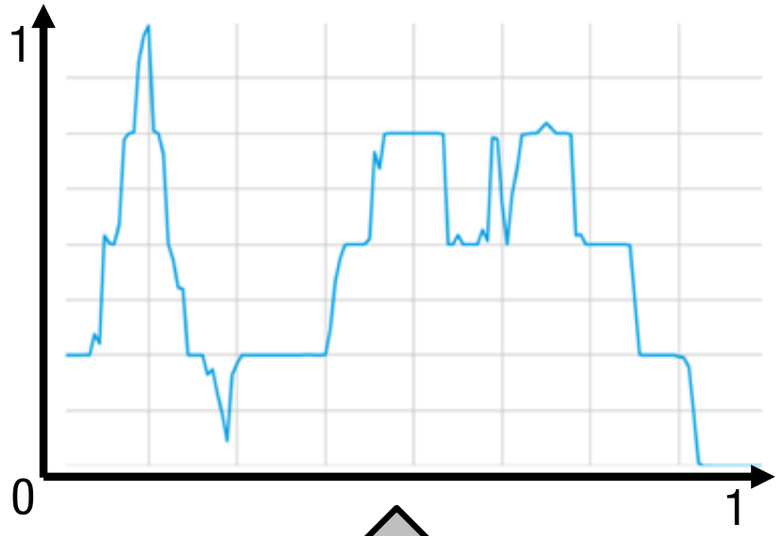
Different data types need different processing.



- Dataset of d-many very long time series y(t)
  - --> Windowing retrieves sub time series
  - One window represents the information for a timestamp t_i
- Heuristics influence parameters of windowing (manageable size)
  - --> window_size=5min, window_shift=1min -- > 40k windows

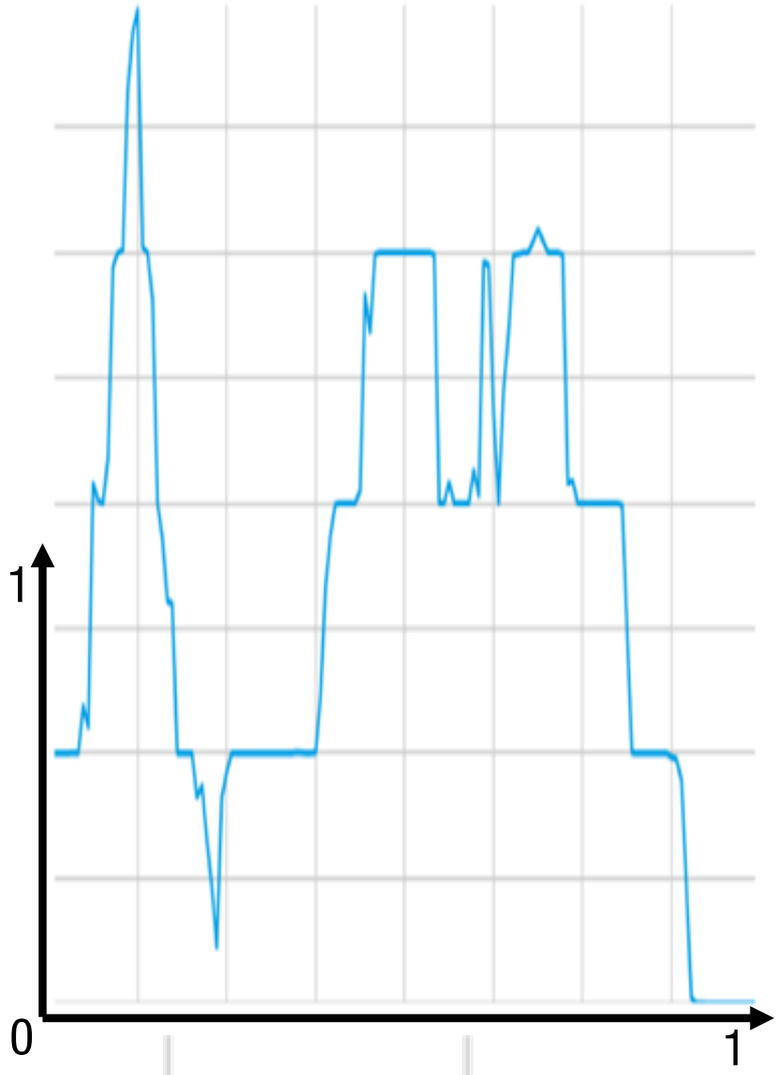
TSDM: Representation and Distance Measure

- Classification needs a distance measure
- Examples:



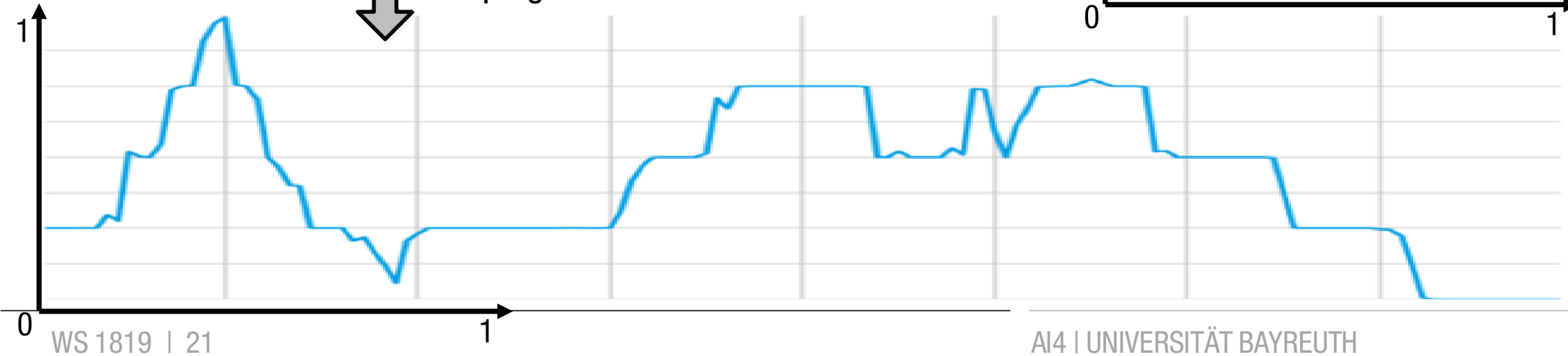
vs.

Amplitude Scaling

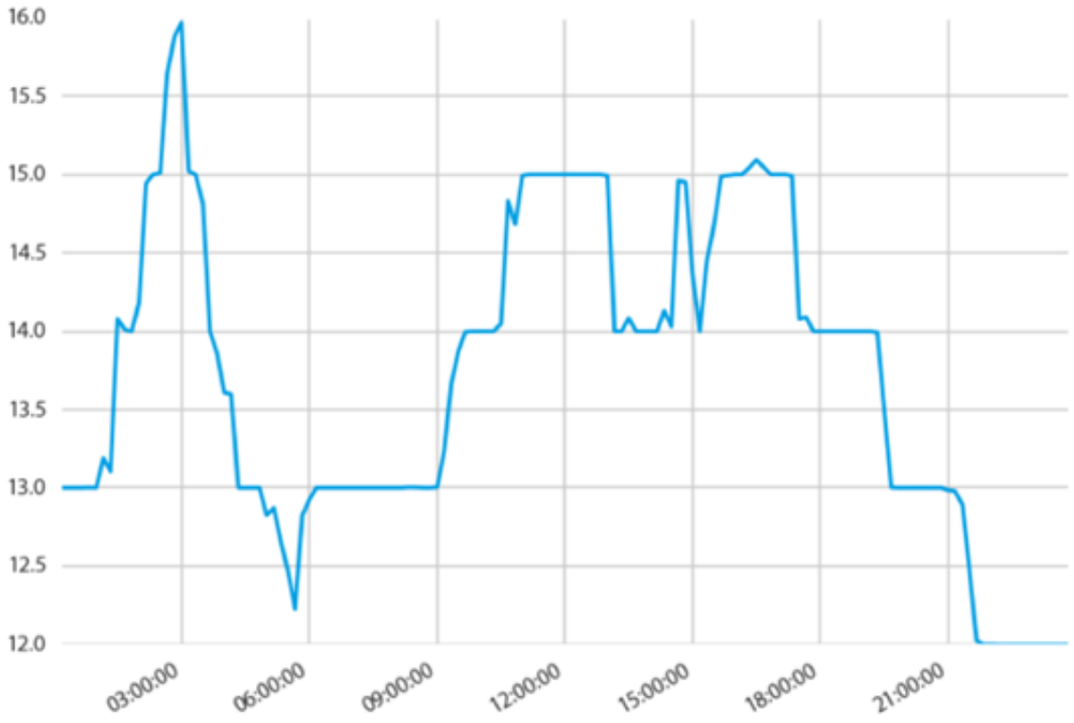


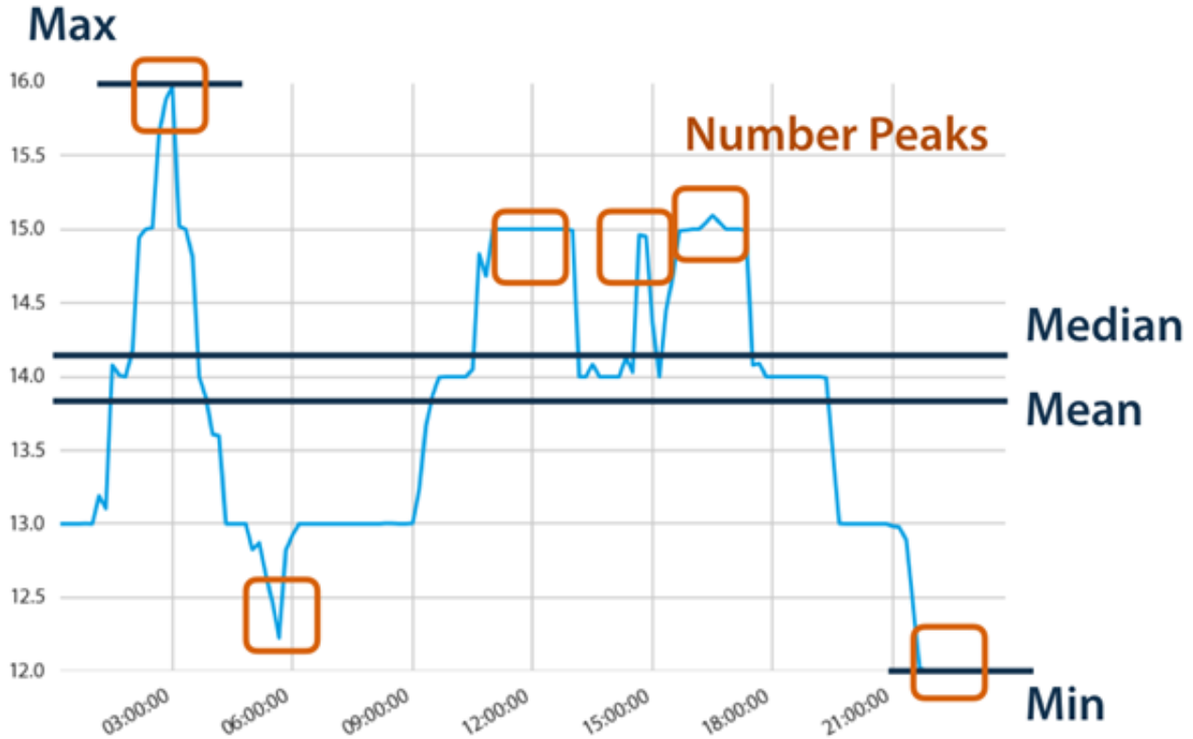
vs.

Time Warping



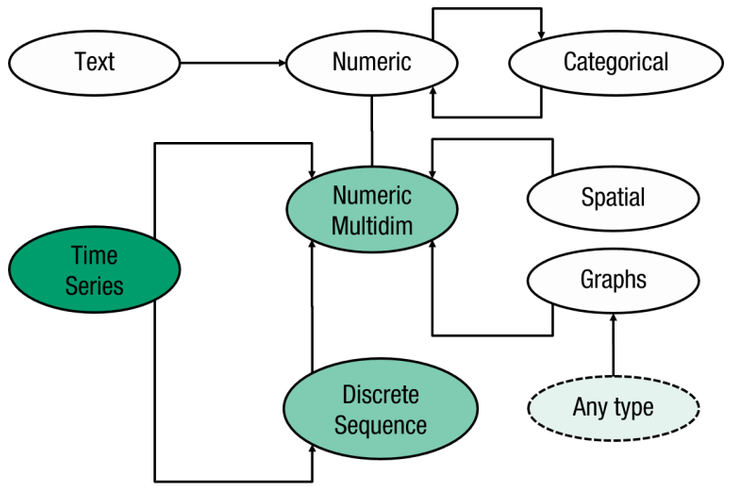
Time Series feature based distance approach



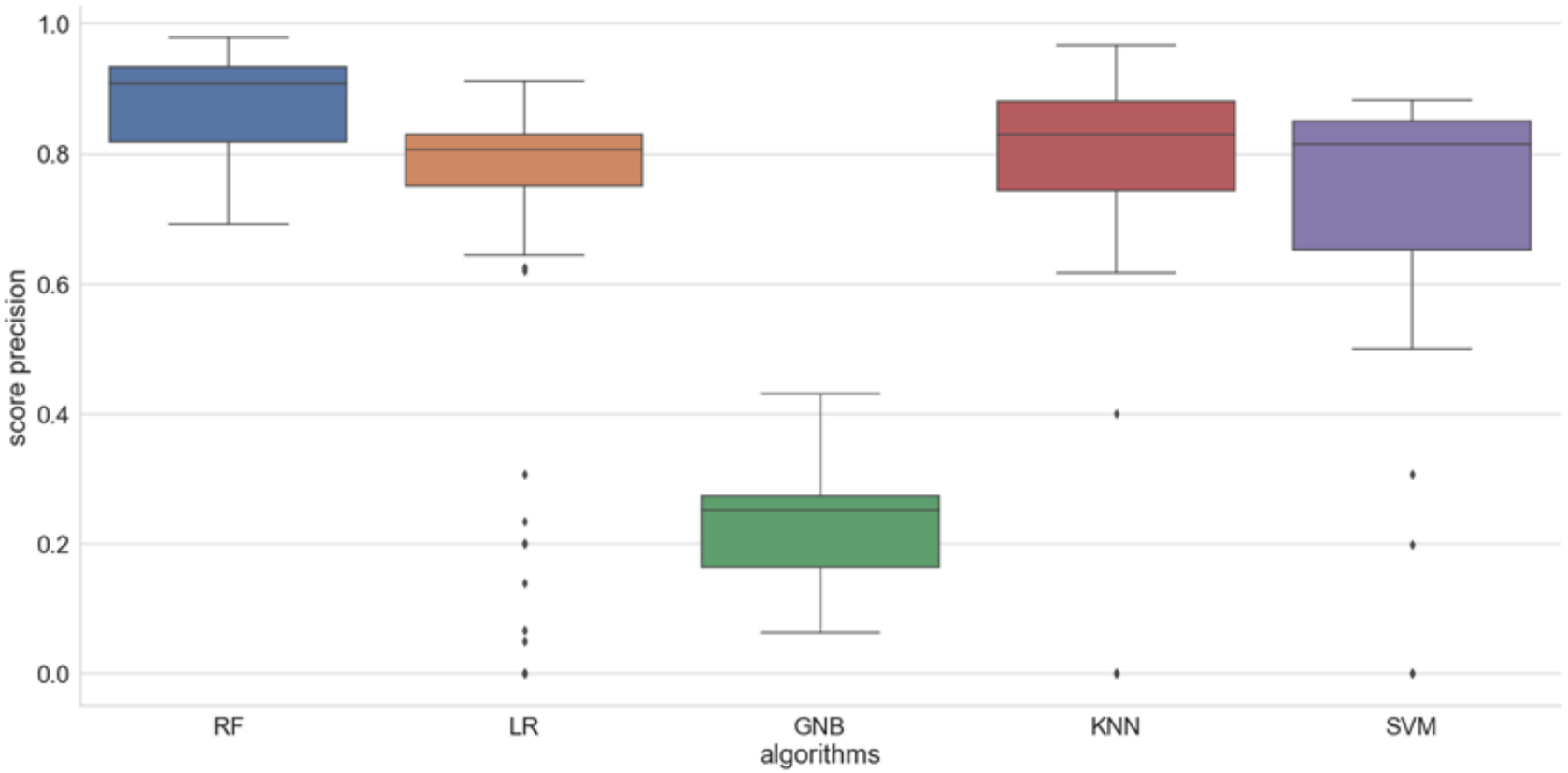


Raw time series representation with feature based distance approach

- Data size reduction
- Numeric multidimensional data type which is input in many standard ML-frameworks



TS-based: 1. Exploration on Precision (Case 3)



Total population	Condition positive	Condition negative
Predicted condition positive	**True positive**, Power	**False positive**, Type I error
Predicted condition negative	**False negative**, Type II error	**True negative**

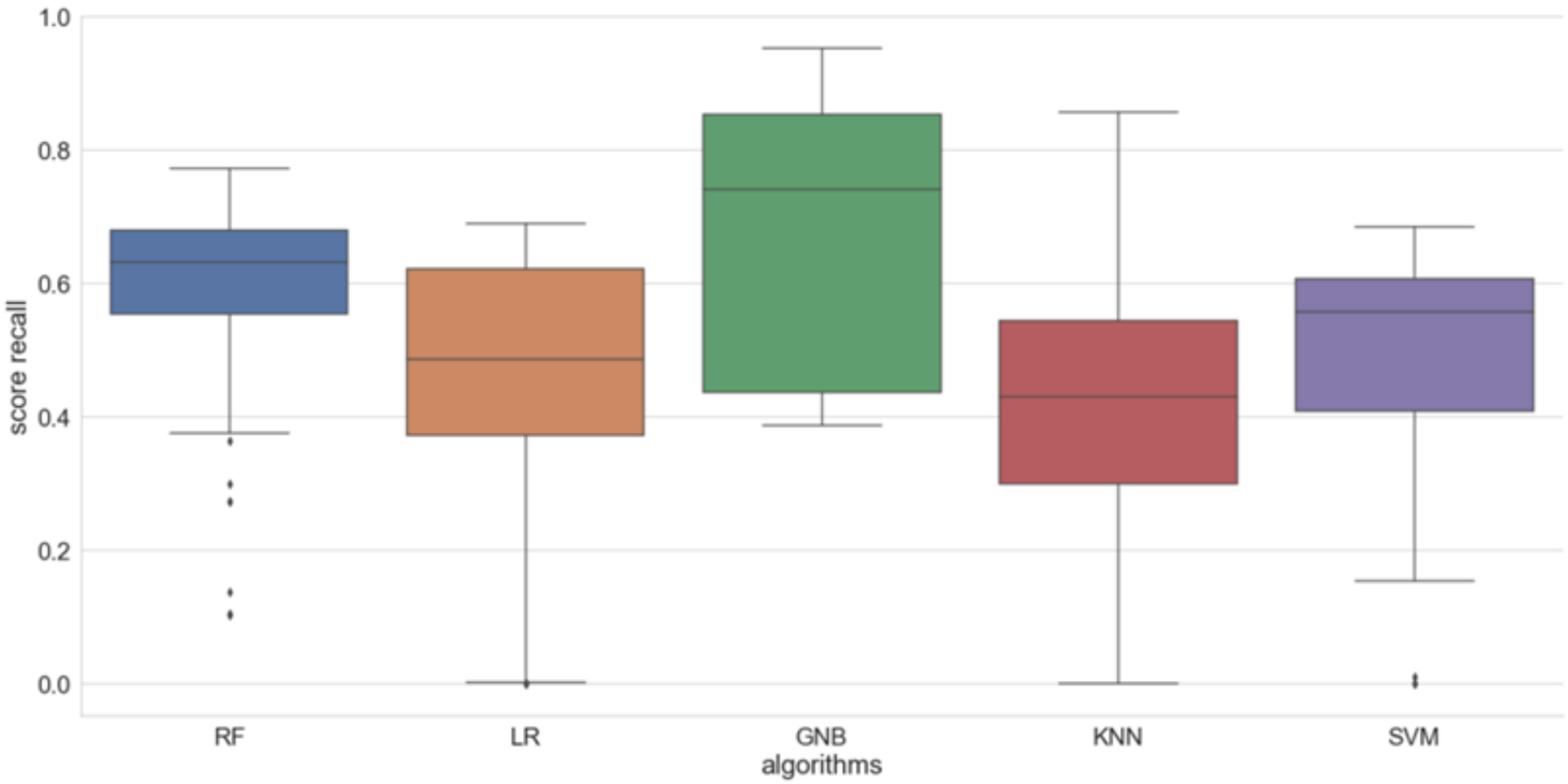
Optional use of PCA / ICA

Top 3 precision (=TP/(TP+FP)):

- RF
- KNN
- SVM

	RF	LR	GNB	KNN	SVM
mean	0.8753	0.7006	0.2277	0.7705	0.7284
min	0.6901	0.0000	0.0640	0.0000	0.0000
25%	0.8179	0.7503	0.1631	0.7440	0.6515
50%	0.9078	0.8060	0.2517	0.8300	0.8146
75%	0.9322	0.8301	0.2739	0.8797	0.8501
max	0.9780	0.9106	0.4297	0.9667	0.8816

TS-based: 1. Exploration on Recall (Case 4)



Total population	Condition positive	Condition negative
Predicted condition positive	**True positive**, Power	**False positive**, Type I error
Predicted condition negative	**False negative**, Type II error	**True negative**

Optional use of PCA / ICA

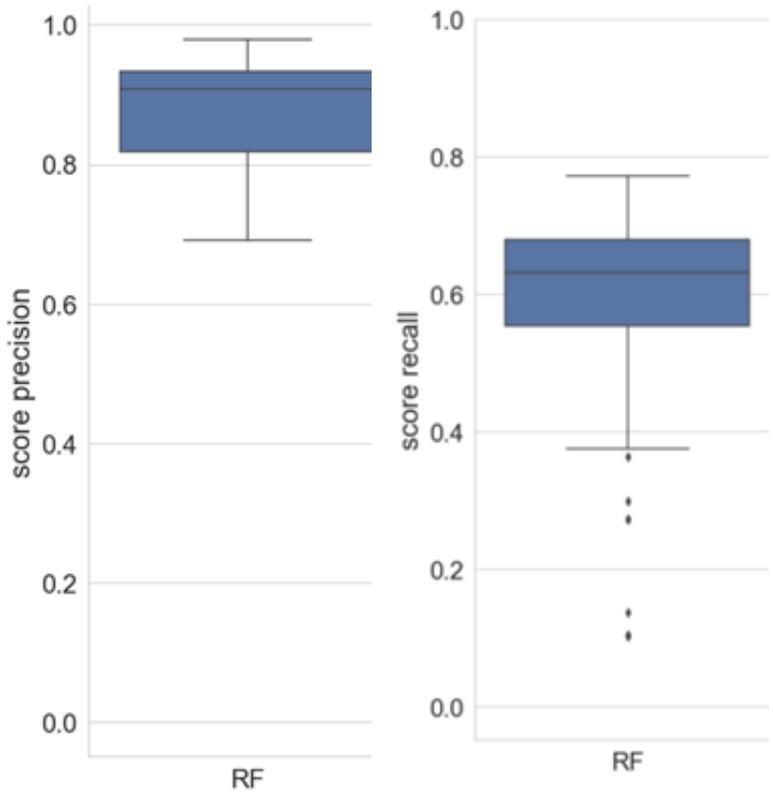
Top 3 recall (=TP/(TP+FN)):

- GNB
- RF
- SVM

	RF	LR	GNB	KNN	SVM
mean	0.5905	0.4619	0.6618	0.4186	0.4919
min	0.1025	0.0000	0.3871	0.0000	0.0000
25%	0.5551	0.3734	0.4368	0.2999	0.4087
median	0.6326	0.4860	0.7404	0.4310	0.5578
75%	0.6808	0.6217	0.8537	0.5443	0.6077
max	0.7733	0.6906	0.9530	0.8565	0.6850

Time Series based: 2. Best Algorithm (Cases 7+8)

	RF_precision	RF_recall			
RF0	0.9780	0.4419	**RF5**	0.9579	0.5523
RF1	0.9776	0.6353	**RF6**	0.9568	0.6712
RF2	0.9757	0.4060	**RF7**	0.9553	0.7264
RF3	0.9651	0.4556	**RF8**	0.9546	0.6849
RF4	0.9609	0.4446	**RF9**	0.9528	0.6711



- Entire dataset with cross validation cv=10:

RF1 (time series b.)	Precision (Case 7)	Recall (Case 8)
CV Score	99.20% (+/-01.74)% [97.46, 100]%	86.45% (+/-04.19)% [82.26, 90.64]%
Test Score	99.35%	84.53%

3. Combination of algorithms

- Locigal OR combines the algorithms‘ prediction
- Idealized Idea: Perfect precision with non overlapping recalled sets
- Aim: Further performance enhancement with combination of most precise algorithms

Scoring Function	RF0 state based [%]	RF0 TS based [%]	Combined [%]
Precision (subset)	99.77	97.80	-- / --
Recall (subset)	43.97	44.19	-- / --
Precision	99.85 +/- 00.46	97.87 +/- 01.50	97.94 +/- 01.52
Recall	98.62 +/- 00.84	89.11 +/- 02.23	99.00 +/- 00.94

--> No improvement achieved in this case

Requirements and Result Summary

- General approach --> useable for all articles
- Data driven knowledge creation
- Assistance system for the workers
- Complement exisiting techniques with a ML approach

Goal:
Further improvement of the quality control mechanism
Prediction of quality

Random Forest	Precision [%]	Recall [%]
State based	99.83 +/-00.17	98.83 +/-00.63
Time Series based	99.20 +/-01.74	86.45 +/-04.19
Combination	97.94 +/-01.52	99.00 +/-00.94

Thank you for your attention.

Now Q & A.

Additional Information

- Random forests are the best working algorithm types
  - Robust to outliers, interpretability by variable importance
  - Ensemble method vs non ensemble methods
- Paritioning: state based > time series based
  - Select more and invariant (shift, drift) features
- PCA/ICA implies information loss --> less achieveable performance
- More refined preprocessing steps
- Choose other discipline of TSDM to solve quality question (clustering, motiv discovery, anomaly detection)

TSDM: Representation and Distance Measure

Definition 2.3.4. *Time Series Representation.* Given a univariate time series $T \in \mathbb{R}^n$ of length $n \in \mathbb{N}$, a representation of T is a model T' of reduced dimensionality $n' \in \mathbb{N}$ with $(n' << n)$ such that T' closely approximates T [EA12].

- Distance depends on the representation
  - Representation: raw, (non)-data adaptive, model based
  - Distance: shape, edit, feature or structure based
- Distance and representation amplify the visibility of information
- Chosen approach:

Raw time series with feature based distance approach