Machine Learning Applications

Winter semester 2019/2020

Simon Mehringskötter

Consultation Hour for Written Exam:
12.02.2020 13.00 – 15.00 L1|01-595

Lecture XIV

Recap on Lecture for Written Exam

What is the objective of the lecture today?

- Brief recap of all previous MLA lectures
- Information for Written Exam
- Preview for next week (last lecture before written exam)

Recap

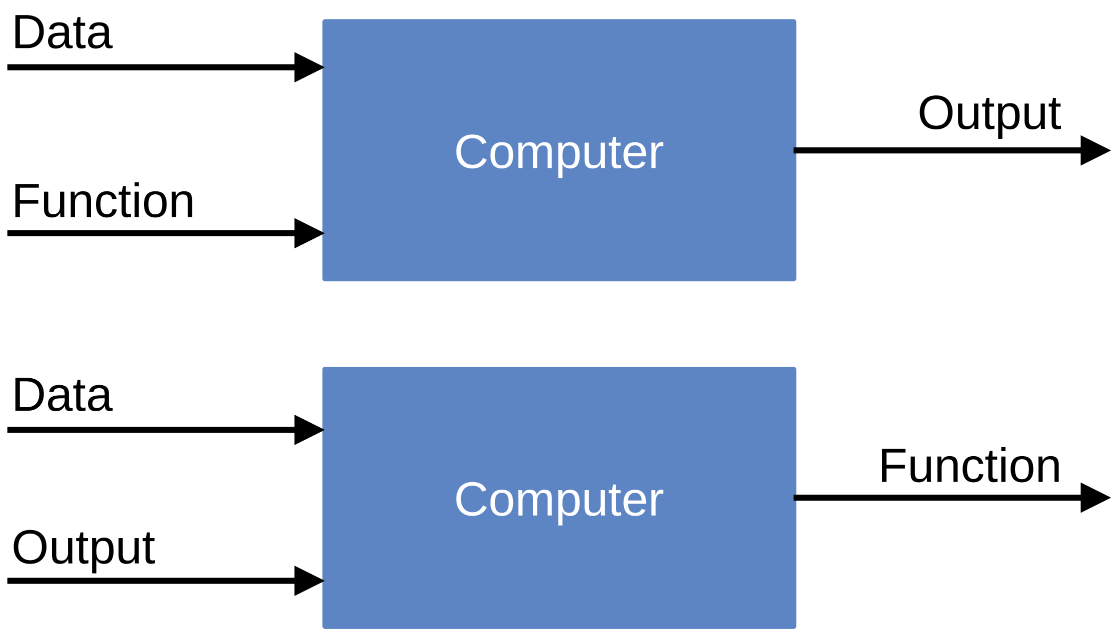
WHAT IS MACHINE LEARNING?

What is AI, what is machine learning and what is deep learning?

Artificial Intelligence is the science and engineering of making intelligent machines, especially intelligent computer programs.

Traditional programming and data analysis

Machine Learning



Deep learning is a form of machine learning that uses artificial neural networks.

Which are the forms/techniques of machine learning?

- **Supervised learning:**

  Training data also include the desired outputs

- **Unsupervised learning:**

  Trainng data does not contain the desired outputs

- **Semi-supervised learning:**

  Traning data does contain some of the desired outputs

- **Reinforcement learning:**

  Reward from sequences of actions

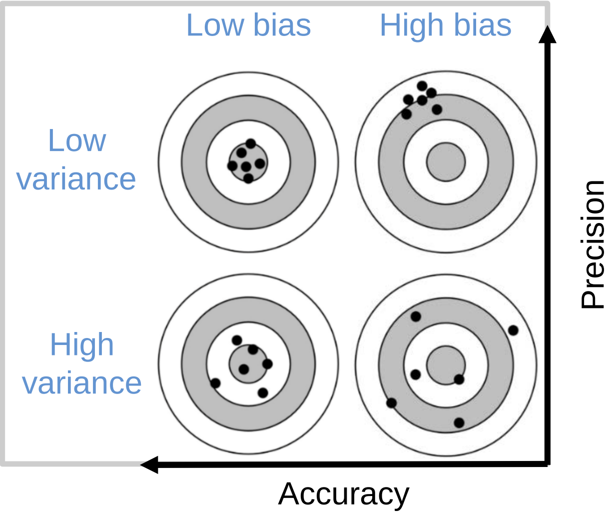
Bias-variance tradeoff

Decomposing errors in machine learning models

Reducible error:

- **Bias** error: occurring by erroneous assumptions in the underlying model
- **Variance** error: sensitivity to small fluctuations in the training set

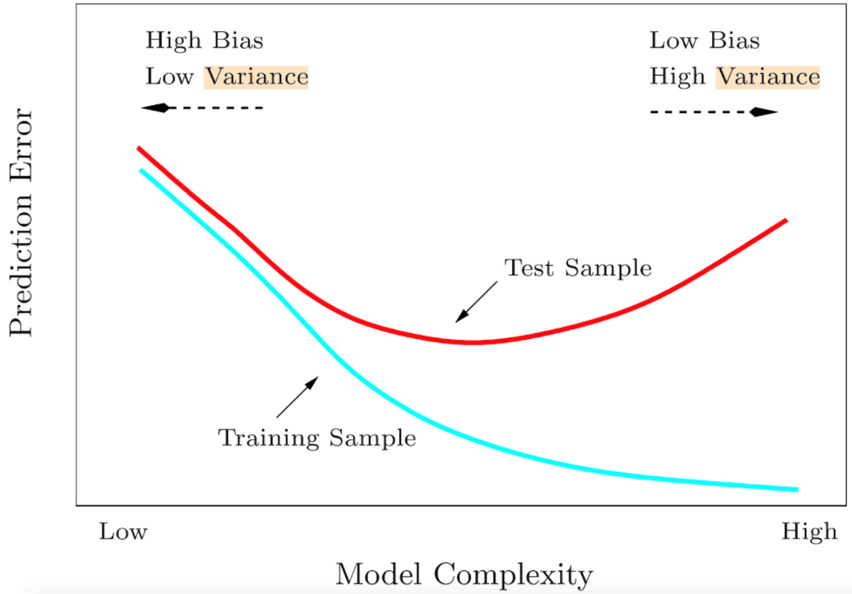
If you try to reduce one error, the other might go up!



Source: http://snoek.ddns.net/~oliver/mysite/the-bias-variance-tradeoff.html

Irreducible error:

Natural variability in a system caused by unknown/unpredictable factors



Source: https://techpolicyinstitute.org/wp-content/uploads/2017/12/Woloszko_Forecasting-GDP-growth-with-adaptive-trees-002.pdf

Bias-variance tradeoff

Overfitting and Underfitting and how to tackle these phenomenon.

- **Overfitting problem:** → Too close to training data; does not generalize

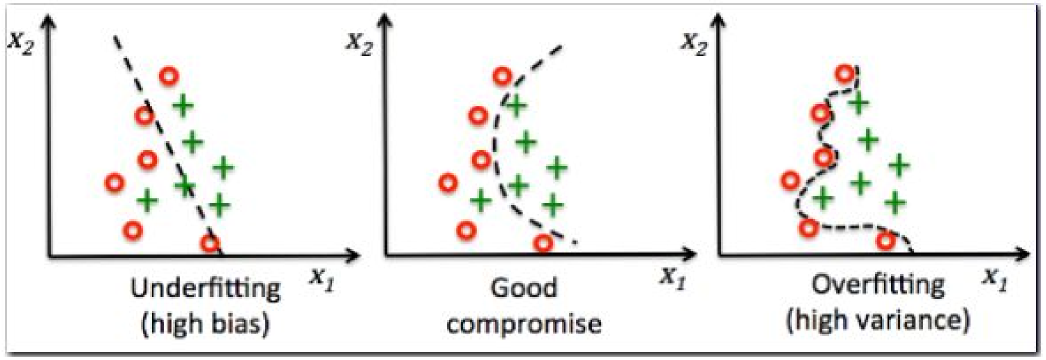
  Starting position high bias: Reducing the bias causes the variance to go up which leads to an overfitting problem

- **Underfitting problem:** → Too much generalized; training data not covered

  Starting position high variance: Reducing the variance causes the bias to go up which leads to an underfitting problem

- How to tackle these phenomenon?

  Build a more complex model, Cross Validation, Dropout method, etc.



Source: https://www.sigs-datacom.de/trendletter/2019-11/2-ki-und-testen.html

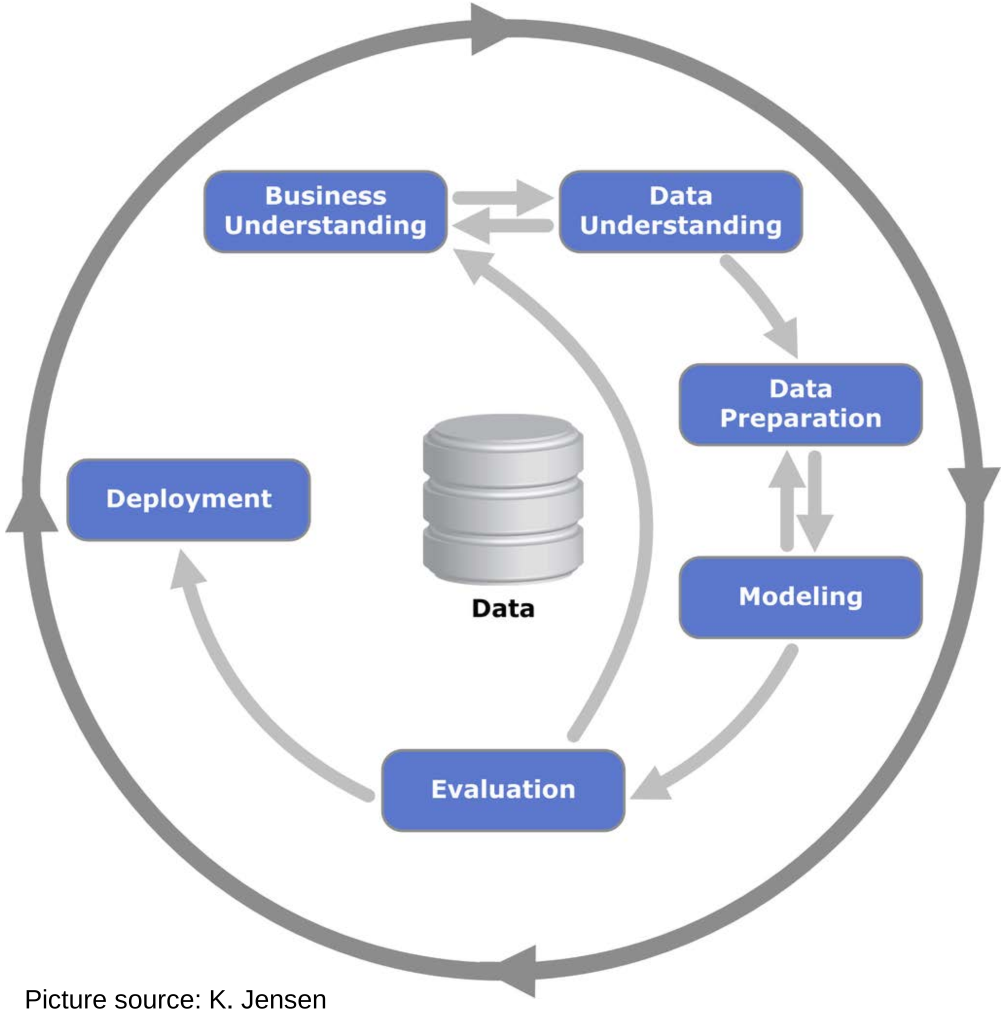
Recap

PROCESS MODELS

Data Mining Life Cycle: CRISP-DM

Cross-Industry Standard Process for Data Mining

TECHNISCHE UNIVERSITÄT DARMSTADT



Picture source: K. Jensen

- Iterative open standard process model to gain knowledge from data related to business goals
- Feedback loops allow iteration of goals and models

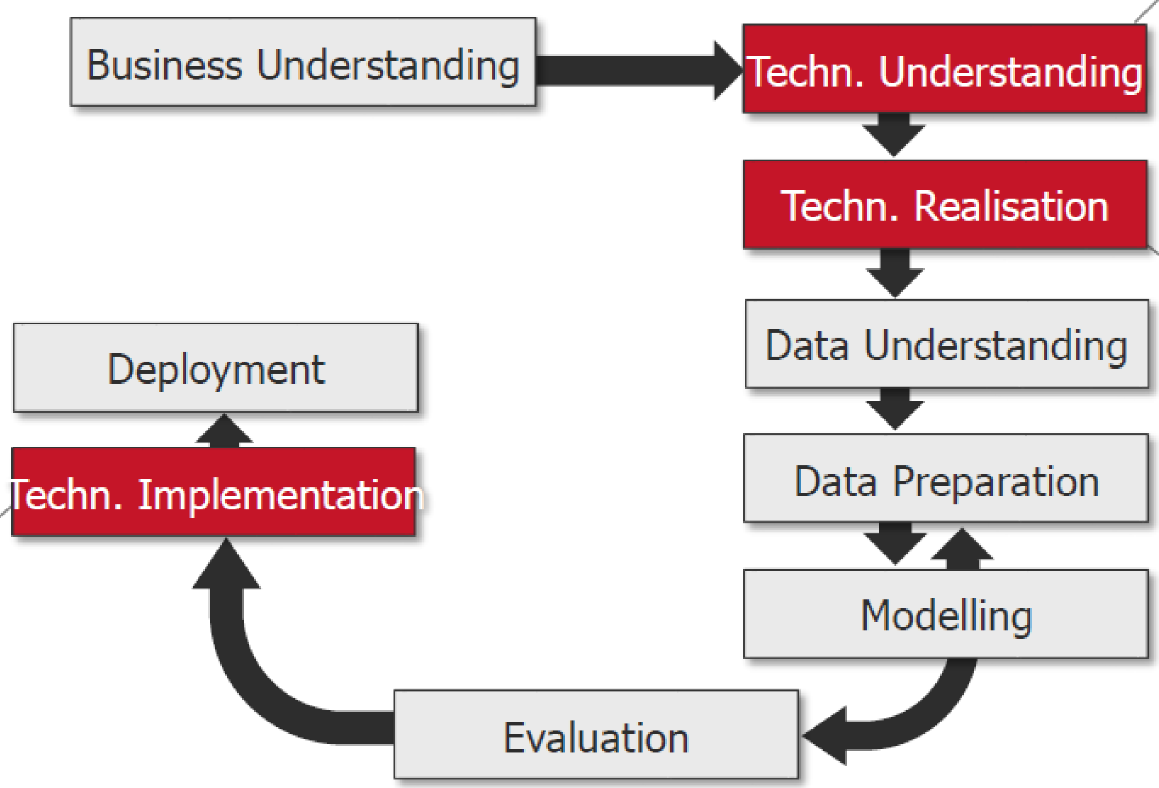
1. Define project goals and business objectives
2. Understand the available data and their quality
3. Filter and select useful and relevant data
4. Create data-models that might meet the defined goals
5. Evaluate models' performance related to the goals
6. Set the best model into operation

Source: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf

Data Mining Methodology for Engineering Applications

DMME Process

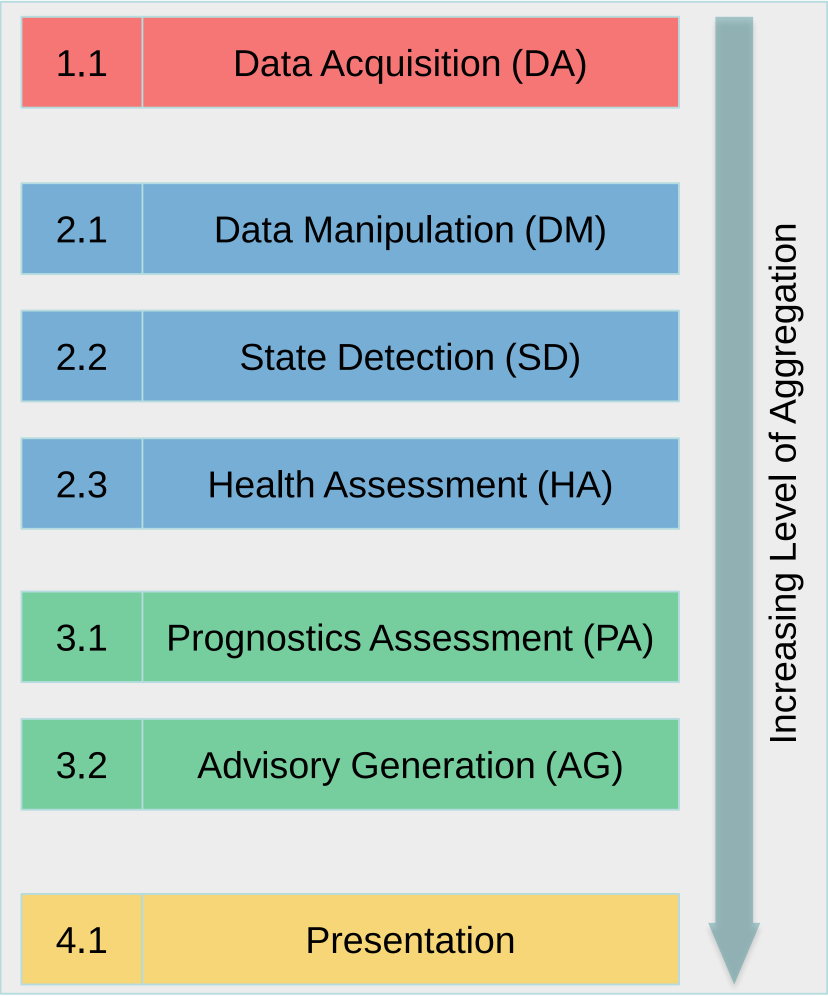
- Measurable technical targets
- Domain knowledge
- Experimental design



- Measuring concept
- Conduct Experiments

- Adapt sensor concept to production environment
- Software integration

Open System Architecture for Condition-Based Maintenance (OSA-CBM)



- OSA-CBM is a pure technical representation and does not value financial benefits
- Transformation of raw data into simple usable information to optimally plan maintenance operations

- 1.1 sensor/parameter selection & data acquisition
- 2.1 filter sensor data, preprocess data
- 2.2 feature extraction/selection/generation that describe state
- 2.3 quantify health (e.g. health index)
- 3.1 predict health degradation/remaining useful life
- 3.2 estimate advisories on system knowledge focused on goals
- 4.1 role-based human-machine-interface to inform user

Recap

LINEAR MODELS AND EVALUATION

Classification vs. Regression

- In **Supervised learning** in addition to each observation $\vec{x}$ there is a label (class) y given, i.e. we have observations $(\vec{x}, y) \in X \times Y$

- y can be a **qualitative** as well as a **quantitative** description of $\vec{x}$

- For the quantitative case e.g. $Y \in \mathbb{R}$ and we try to predict for an unknown $\vec{x}$ the value y (**Regression**)
  - E.g. linear Regression $\hat{y} = \sum_{i=1}^{p} \beta_i x_i + \beta_0$ where $y, \beta_0 \in \mathbb{R}$; $\vec{x}, \vec{\beta} \in \mathbb{R}^p$

- In the case of qualitative descriptions, Y is a discrete quantity and we use a function f for a **Classification**

Curse of the high dimension in linear models

- Underlying objective: Minimizing the training error and estimating the error probability
  - Simple and commonly used error function: Residual Sum of Squares (RSS): $RSS = \sum_{i=1}^{N}(\hat{y}_i - y_i)^2$

- The overall x_i averaged training error of linear models consists of noise, variance and bias.

$$\frac{1}{N}\sum_{i=1}^{N} Err(x_i) = \sigma_\epsilon^2 + \frac{p}{N}\sigma_\epsilon^2 + \frac{1}{N}\sum_{i=1}^{N}\left[f(\vec{x}_i) - E\hat{f}(\vec{x}_i)\right]^2$$ (*no need to memorise for exam*)

- Therefore, model complexity (p, N) and variance of estimates for different training datasets are directly related in linear models

距离计算困难

sparse 数据样本

→ **Curse of the high dimension** in linear models

Evaluation of learned models

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Validation through experts**
  - A domain expert evaluates the plausibility of a learned model 合理性

    + but often the only option (e.g., clustering)

    - subjective, time-intensive, costly
- **Validation on data**
  - Evaluate the accuracy of the model on a separate dataset drawn from the same distribution as the training data

    + fast and simple, off-line, no domain knowledge needed, methods for re-using training data exist (e.g., cross-validation)

    - labeled data are scarce, could be better used for training
- **On-line Validation**
  - Test the learned model in a fielded application

    + gives the best estimate for the overall utility

    - bad models may be costly

Confusion Matrix
(Classification)

TECHNISCHE UNIVERSITÄT DARMSTADT

	Classified as +	Classified as -	
Is +	true positive (tp)	false negative (fn)	tp + fn = P
Is -	false positive (fp)	true negative (tn)	fp + tn = N
	tp + fp	fn + tn	|E| = P + N

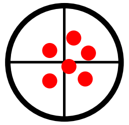
- The confusion matrix summarizes all important information
- How often is class *i* confused with class *j*
- Most evaluation measures can be computed from the confusion matrix
- Accuracy, Precision, Recall, Specificity, False Negative Rate, False Positive Rate

Frequently used are **Accuracy**, **Precision, Recall and Specificity**

$\text{Accuracy} = \frac{tp+tn}{tp+tn+fn+fp}$ $\text{Precision} = \frac{tp}{tp+fp}$ $\text{Recall} = \frac{tp}{tp+fn}$ $\text{Specificity} = \frac{tn}{tn+fp}$

精度 真阳性 真阴性率

Different performance metrics exist



Accuracy based metrics

- **Error** $\Delta(i) = r_*(i) - r(i)$

  $r(i)$: RUL estimate at time t_i

  $r_*(i)$: True RUL at time t_i

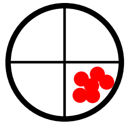
  → represents the deviation

- **Mean absolute percentage error**

  $$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{100\Delta(\mathrm{i})}{r_*(i)}\right|$$

  → unit free metric

  → weighs small errors differently to big errors



Precision based metrics

- **Sample standard deviation**

  $$SSD = \sqrt{\frac{\sum_{i=1}^{N}(\Delta(i)-\mu_\Delta)}{N-1}}$$ with μ_Δ: mean of errors

  → measures dispersion/spread of the error

  → normal distribution is assumed

- **Mean absolute deviation**

  $$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|\Delta(i) - \mathrm{median}\big(\Delta(i)\big)\right|$$

  → estimator of dispersion/spread of the error

  → can be used for small number of data

Mean squared error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\Delta(i)^2$$

Cross-Validation

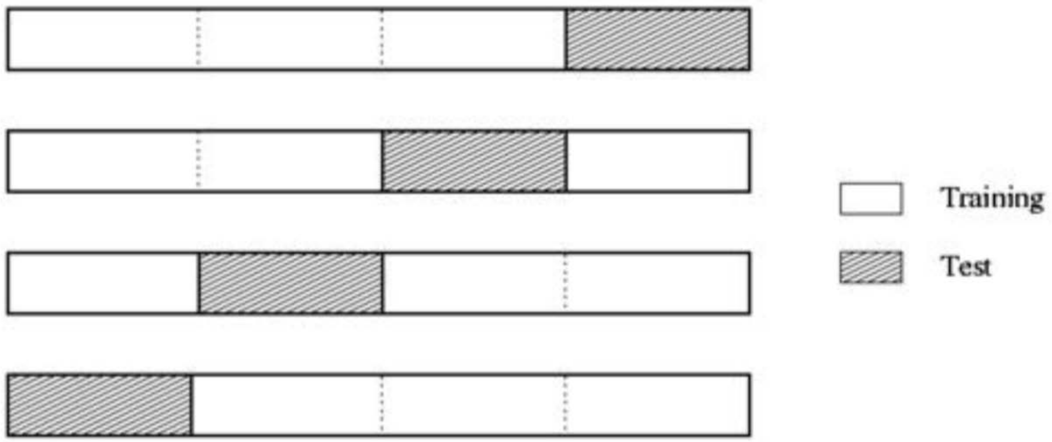
- **Algorithm**:
  - split dataset into n (usually 10) partitions
  - for every partition n
    - use other n-1 partitions for learning and 1 partition for testing
  - average the results
- **Properties:**

  \+ makes best use of available data
    - only one example not used for testing

  \+ no influence of random sampling
    - training/test splits are determined deterministically

  \- typically very expensive

  \- bias
    - e.g., majority classifier in a perfectly balanced problem



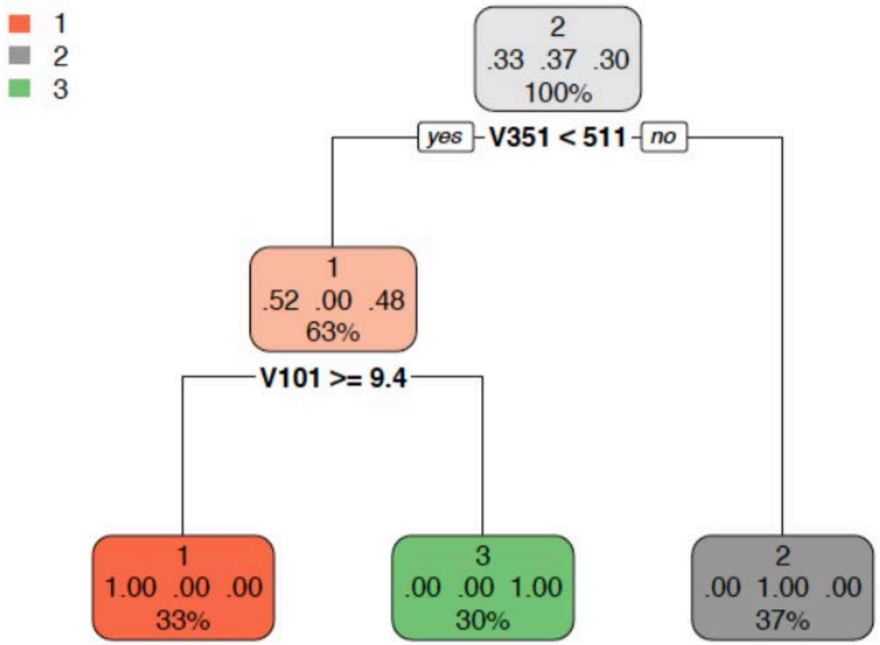
Example: 4-fold cross-validation

Recap

TREE BASED METHODS & ENSEMBLES

What are Decision trees?

- **Decision tree**: (*n*-ary) decision trees consist of a sequence of *n* decisions (test), which can be represented as a tree.
- In binary trees, a “Yes/No" (binary) decision is made in each inner node. If "Yes", the decision in the next node follows on the left side, if "No" on the right side.
- The decisions are made in such a way that the nodes represent a "pure" class if possible.
- In each terminal node (leaf) an assignment is made to the class that occurs there most often.
- There are decision trees for classification and regression as well as in combination with linear models



Pruning of Decision trees

entropy / Entrupie

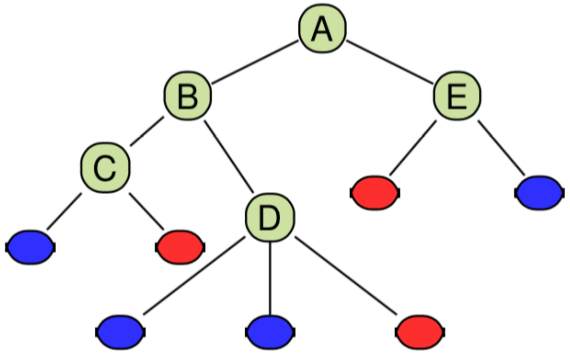
TECHNISCHE UNIVERSITÄT DARMSTADT

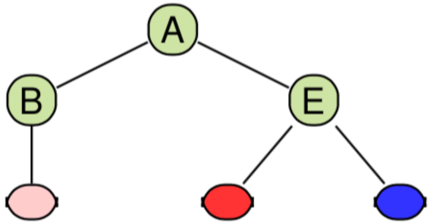
- The **goals** of pruning:
  - Reducing Overfitting of the tree to the training data
  - Increase intelligibility!
- **Operations** of pruning:

  a) Set nodes in place of a subtree

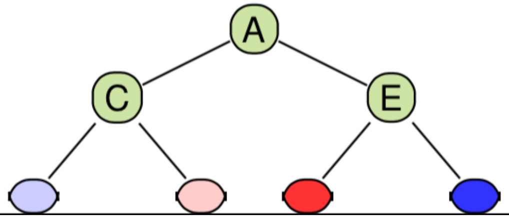
  b) Move a subtree one level higher
- **Estimating** how the real error in trimming develops.

  If the error of a node is smaller than the sum of the errors of its subnodes, the subnodes can be pruned away. To do this we have to estimate (bottom-up) the errors at all nodes.





a) Set nodes in place of a subtree



b) Move a subtree one level higher

What are Random Forrests?

- Random Forrest:

  A random forest consists of several uncorrelated decision trees. All decision trees have grown under a certain type of randomisation during the learning process.

- Prediction Random Forrest:

  A new object is classified by the Random Forest by classifying it once from each of the calculated trees and then assigning it to the class that most trees prefer.

Input

⇩

Every tree makes its predictions

⇩

Formal majority decision

\+ Any variable that contributes to class segregation is also used at some point in the classification

– Comprehensibility is lost, because the classification rule is no longer easy to read.

Bagging and Boosting

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Bagging**:
  Bagging (Bootstrap aggregation) refers to the collection of many similar learners and their common decision rule, where the learners are generated from bootstrap samples of both observations and variables of a data set. Typically, the learners are of a rather simple structure, e.g. random forests as bagging procedures that use individual trees as learners.

- **Boosting**:
  Boosting is a procedure that establishes an efficient decision rule for a classification problem by combining several simple rules. These rules are called weak learner classifiers or basic classifiers (e.g. naive Bayes, logistic regression, decisions stumps or flat decision trees)
  The result of boosting, on the other hand, is called a strong classifier (strong learner)

Recap

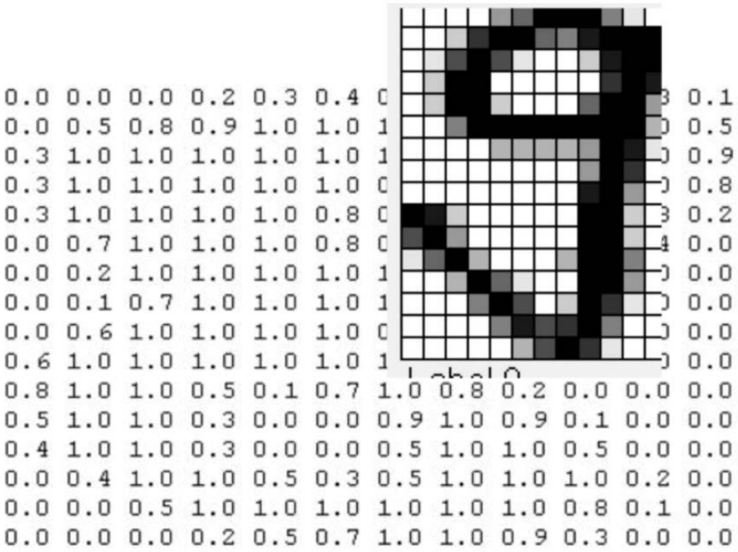
FURTHER METHODS OF MACHINE LEARNING AND DEEP LEARNING

k-Nearest-Neighbors (kNN)

Supervised

On the example of handwritten digit recognition

- Two images represent the same digit if the images are similar
- Similar = similar gray value distribution
- We present images as a matrix of gray values
- The similarity is described by the distance. The more similar, the smaller the distance. They are identical if their distance is 0. The most important distance measure for real features/vectors is the „Euclidean distance“
- k defines the amount of neighbors considered for a majority decision



Digit = 12 x 16 Matrix of gray values in [0,1]

Vector of gray values of length 192

Clustering

Two types of clustering

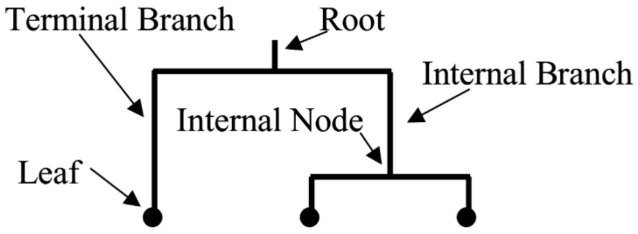
Unsupervised

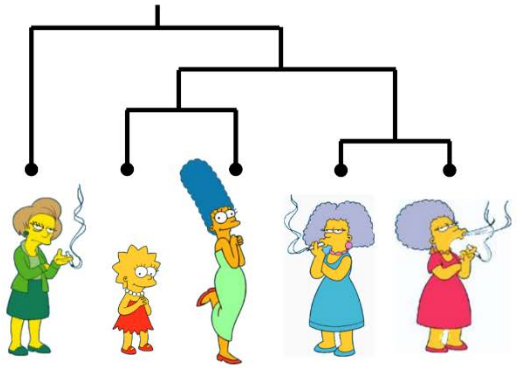
- **Partitioning approaches**: Construct partitionings (divisions) of the data and evaluate them using an evaluation function
- **Hierarchical approaches**: Construct a hierarchical division of the data based on a criterion

层次聚类

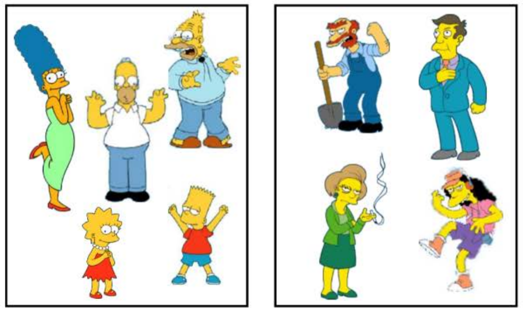
树状图

- Dendrograms: Representing similarities by a tree. The similarity of two objects is expressed in a dendrogram by the height (as seen from the leaves) of the lowest internal node that both objects have in common





Hierarchical



Partition

k-Means Algorithm (kMA)

Clustering by means of partitioning

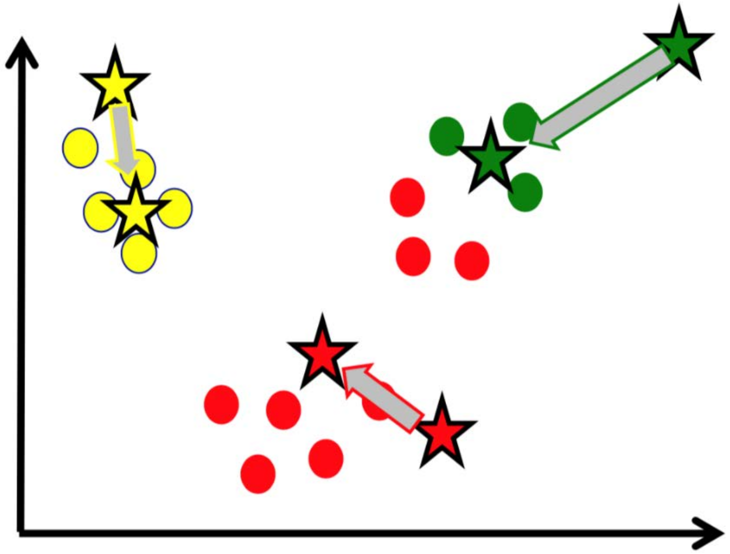
k = Number of clusters (you specify)

One „mean" per cluster

1. Initialize the mean value (e.g. by randomly selecting k data points).

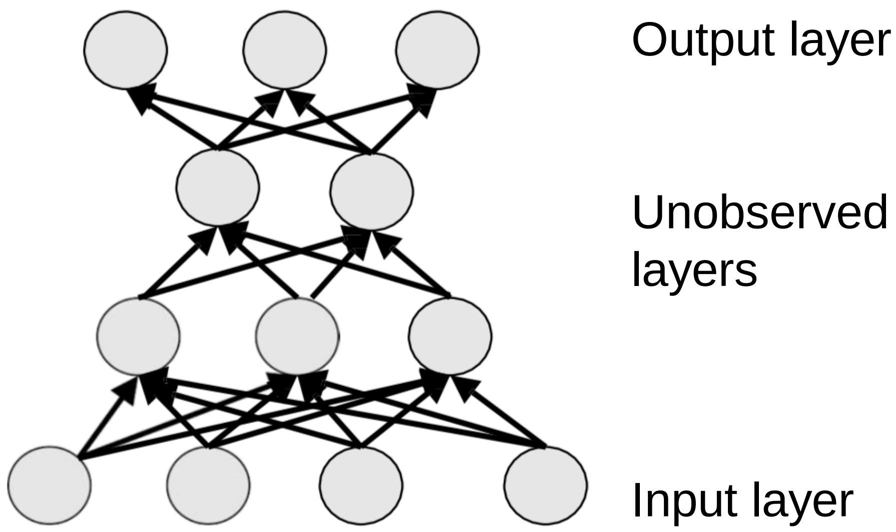
   Now, we repeat the following two steps until convergence:

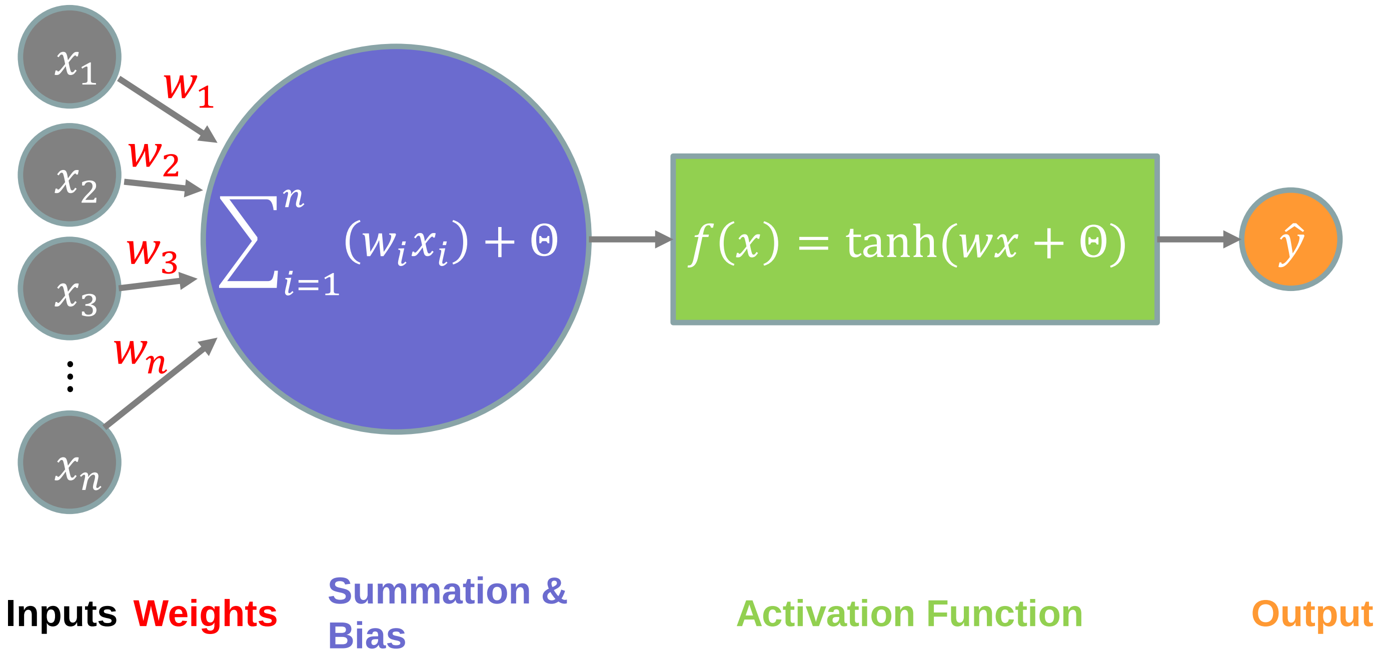
2. Assign each data point to the cluster of its next average value
3. Compute the centroids for the clusters by taking the mean of all data points that belong to each cluster.



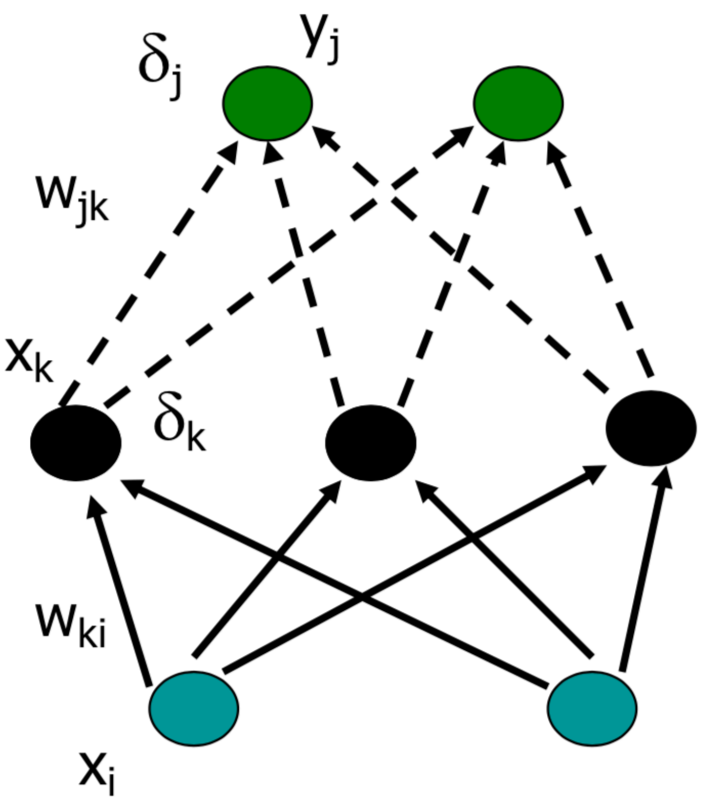
Deep Architectures and Neuronal Networks

- Deep architectures consist of several layers of non-linear computation, such as neuronal networks with several unobserved layers





Backpropagation



传播

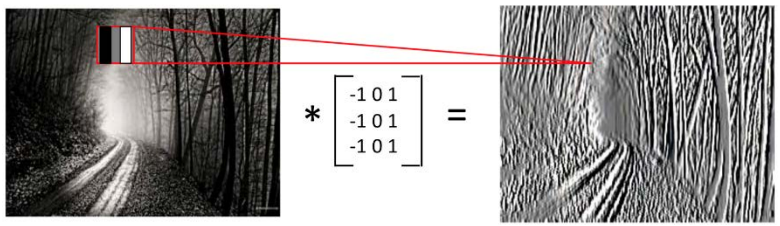
Backward step / feedback:
Reports the error from the output layer (successively) to the unobserved layers

Forward step / forward message:
Propagates activations from the input layer to the output layer

Deep Convolutional Neuronal Network (DCN/CNN)

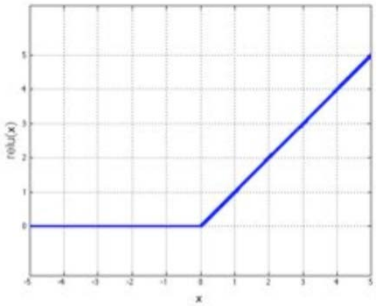
Fully interconnected multi-layer network

- **Convolutional Layer**: Filters detect local patterns such as color values, edges, ...



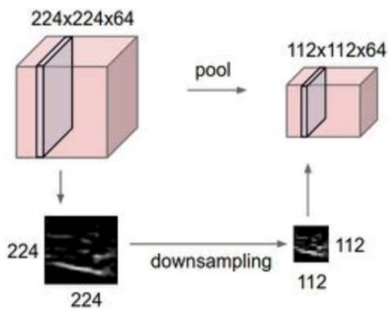
Conv. Layer

- **Rectified Linear Unit** (ReLU): Non-linear activation functions are applied per element



max(0,x) ReLU

- **Pooling Layer**: Compress the representation (downsampling/sub-sampling). They are applied to each checkbox independently and are intended to make the network invariant to smaller transformations



Pooling

- Output Layer with **Activation Function** Soft-Max

Filter of Convolutional Network

3D

Convolutional

Pooling

Convolutional

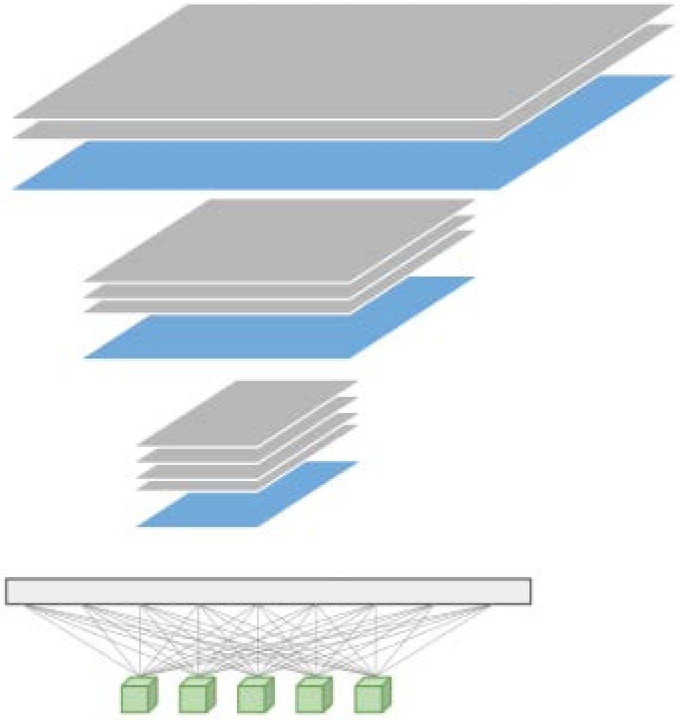
Pooling

Convolutional

Pooling

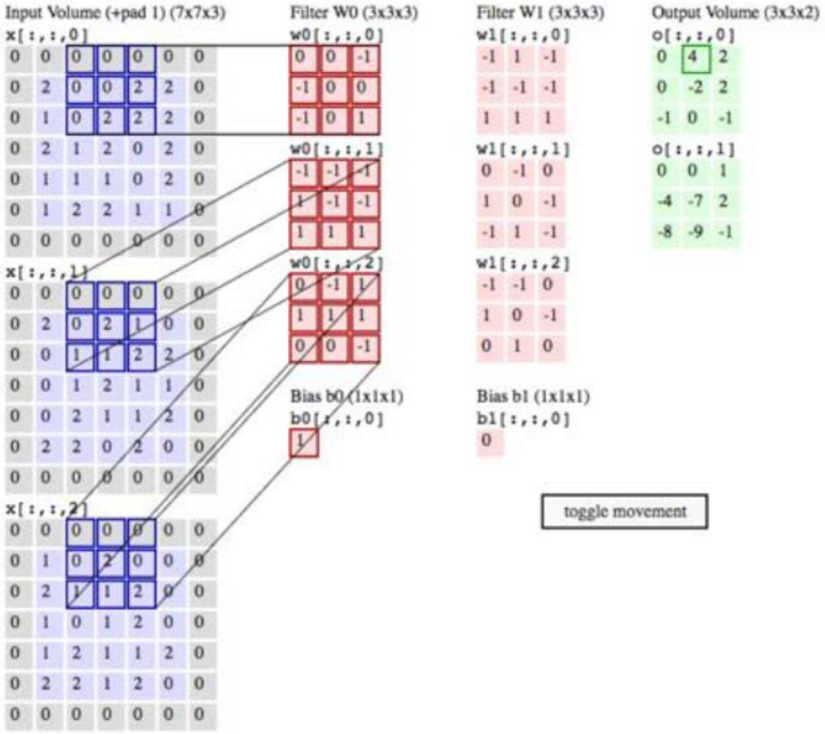
Flatten

Dense



Source: towardsdatascience.com

2D



Convolution: "slide the filter over the picture and calculate at each set the scalar product”

Self Organizing Maps / Best Matching Unit

SOM / BMU

TECHNISCHE UNIVERSITÄT DARMSTADT

General purposes of SOM:

- Clustering of data (i.e. understand how many clusters might be in a specific data set)
- Dimensionality reduction (i.e. find out which features are truly relevant or can be neglected)

Algorithm:

对齐

- Neurons assigned to a weight vector are aligned within a grid (e.g. rectangular)
- Each node‘s weights are randomly initialized and the input vector is chosen at random from the set of training data
- Every node (weight vector) is examined to calculate which one‘s weights are most like the input vector (via Euclidean distance) → Winning node: **Best Matching Unit** (BMU)
- Winning node and neighborhood (distance related) are rewarded with becoming more like the sample vector
- Repeat for a number of iterations and determine distances between neurons

Self Organizing Maps / Best Matching Unit

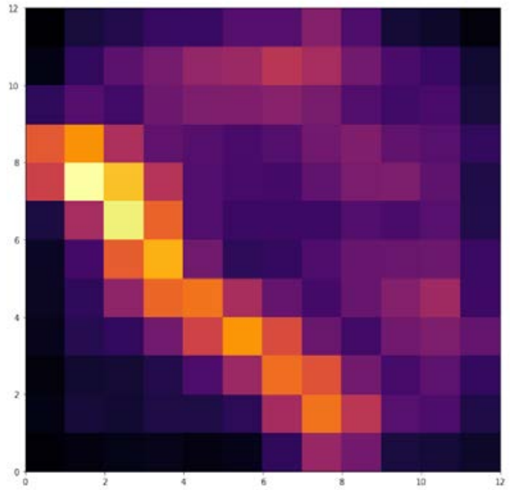
SOM / BMU

Typical decisions to make when using SOM:

- Learning rate: How close shall the BMU and its neighbors be pulled towards the data point
- Neighbors: How many neighbors shall be affected from the BMU. (Usually in the beginning the neighborhood should be large and decrease monotonically over time) 单调

Interpretation of results:

- If the average distance is high, then the surrounding weights are very different and a light color is assigned to the location of the weight.
- If the average distance is low, a darker color is assigned.

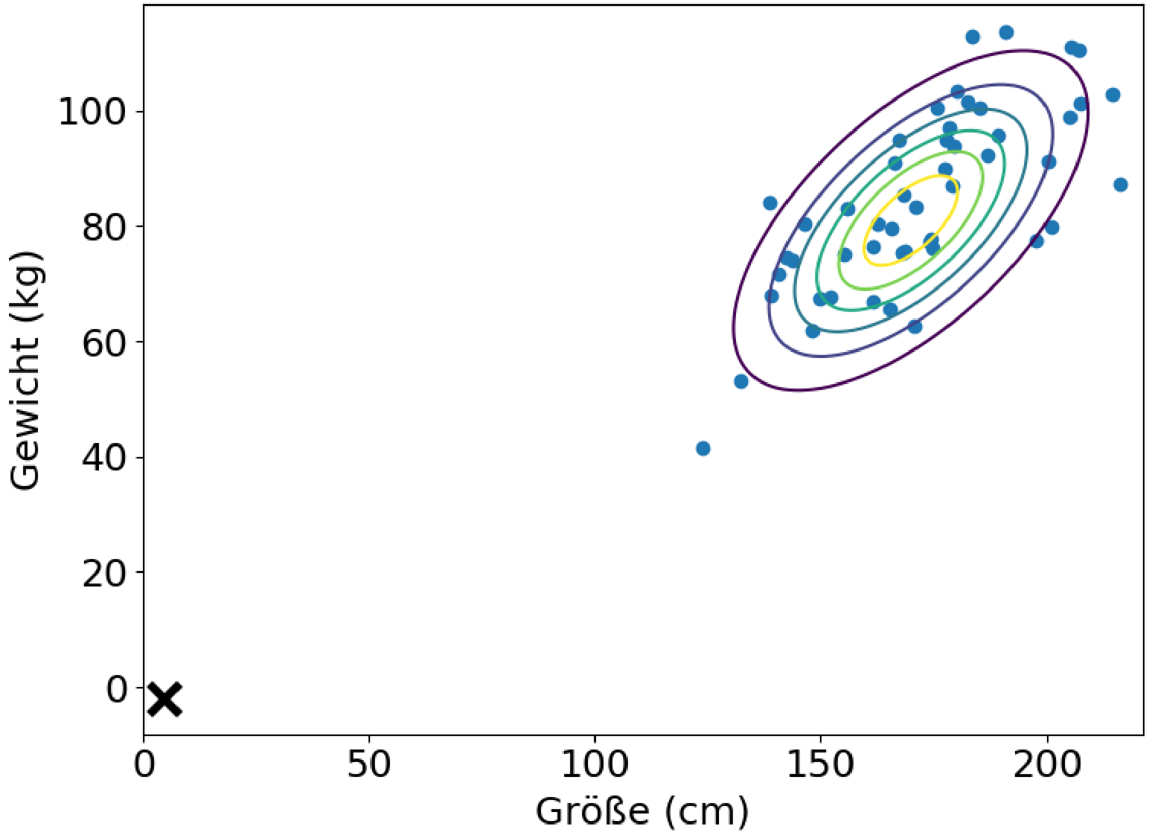


→ SOM forms a semantic map where similar samples are mapped close together and dissimilar ones apart.

Generative Models

- Generative models aim to model the representation over the entire data, i.e. $p(x, y)$ instead of $p(y \mid x)$
- This allows answering a variety of additional queries
- For instance, we can evaluate the input likelihood to detect outliers

$$p(x) = \int p(y, x) dy$$



Typical methods: Sum-Produkt Networks, Normalizing Flows, Autoregressive Density Estimators, Variational Autoencoders → see Lecture 5

Recap

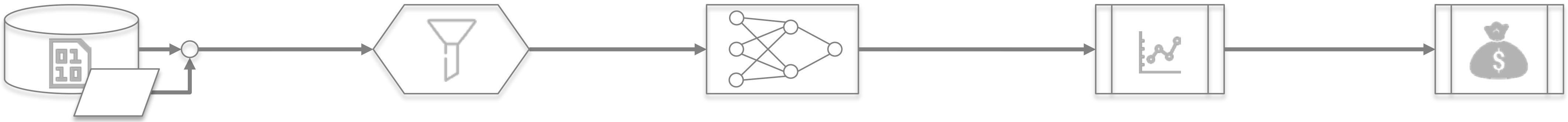
DATA UNDERSTANDING & PREPROCESSING

Design principles

(for machine learning projects)

TECHNISCHE UNIVERSITÄT DARMSTADT

Data base | **Preprocessing** | **Modeling** | **Surplus value / Benchmarking** | **Deployment**

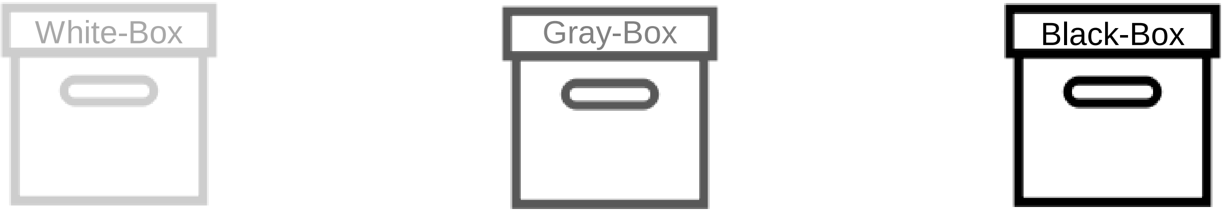


Engineering approach

Intelligent approach
SMART DATA

Computer science approach

- Mathematical / physical models
- SME knowledge
- Challenging for digitization
- Expensive for complex systems



- Input-output analyses
- Mastering system complexity
- Hidden functionalities

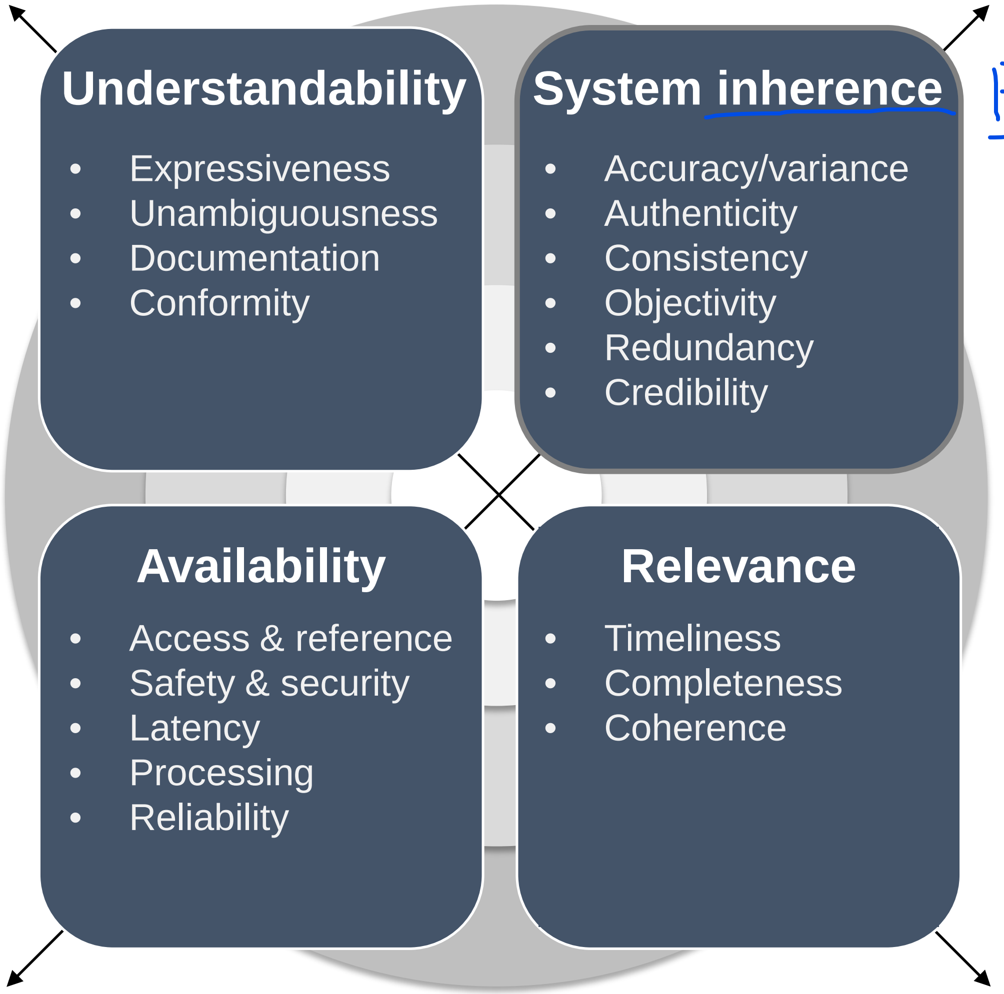
Domain knowledge ←

→ Complexity

- Objection: optimization of model efforts (time, money, resources, accuracy) – benchmarks are obligatory
- Correlation ≠ Causality 因果
- Derivation of explainable AI / ML frameworks

Assessing the Data Quality

Classes and dimensions for data quality assessment



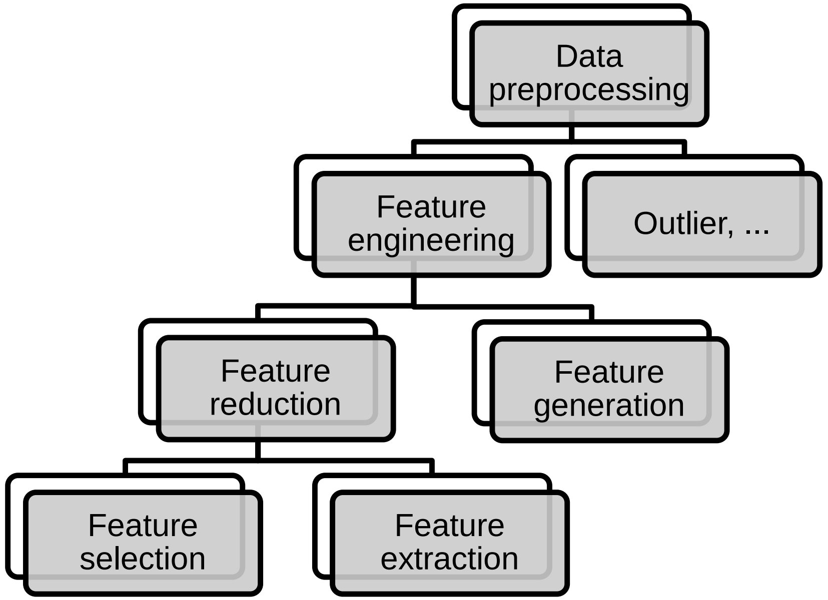
固有

Area of high data quality

Feature Engineering

Feature reduction can be divided into feature extraction and feature selection

- Avoidance of multi collinearities and redundant parameters
- Better generalizability
- Evalutaion of reduction methods through model performance/quality



Feature reduction

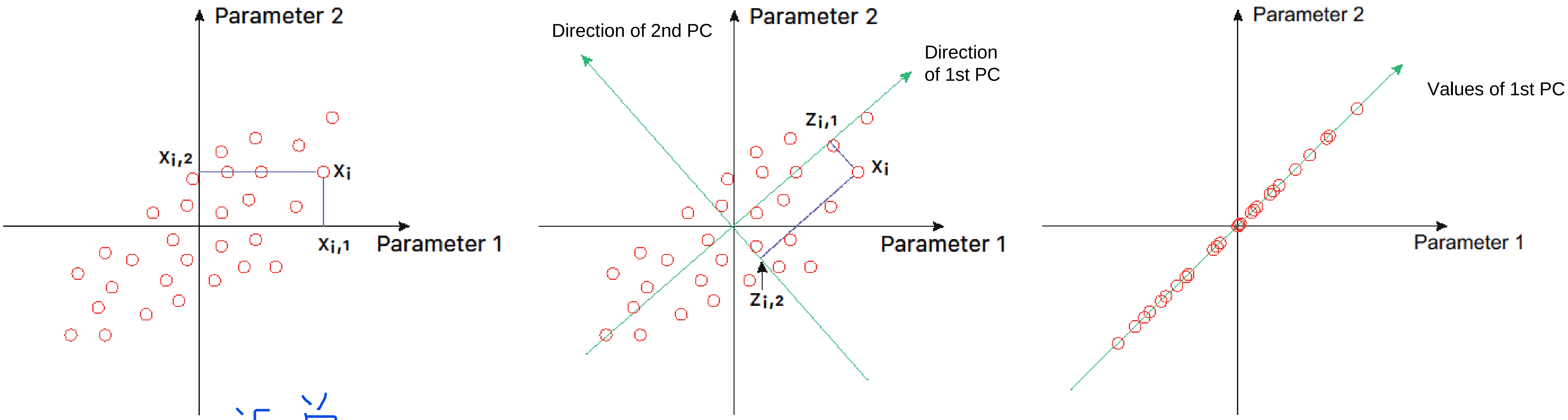
Feature extraction 提取
- Principal component analysis (PCA)
- Factor analysis

Feature selection
- Wrapper
- Filter
- Embedded

Principal Component Analysis (PCA)

Visual explanation (Two dimensional)

TECHNISCHE UNIVERSITÄT DARMSTADT



汇总

➔ Aggregate data in new space (orthogonal transformation) in order to represent most of the variance and reduce the dimensionality

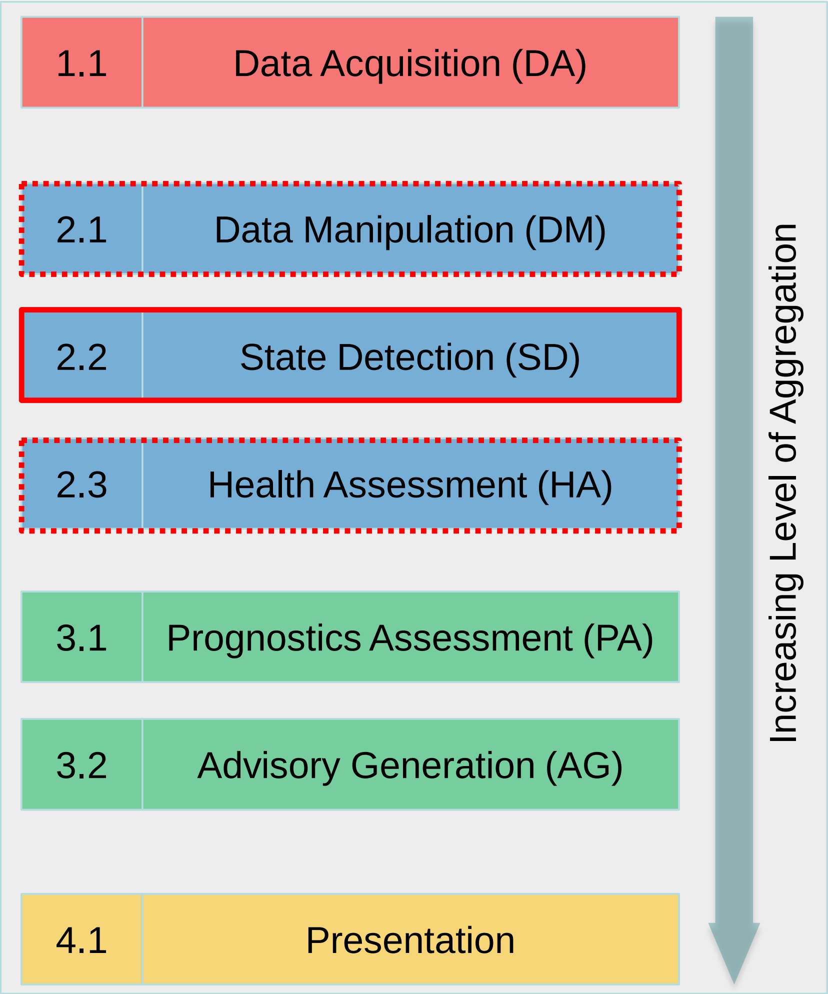
➔ Specify value for variance that should be explained or use elbow plot

Recap

DIAGNOSIS MODELS VS. PROGNOSIS

Diagnosis is the beginning in OSA-CBM

Open System Architecture for Condition Based Maintenance



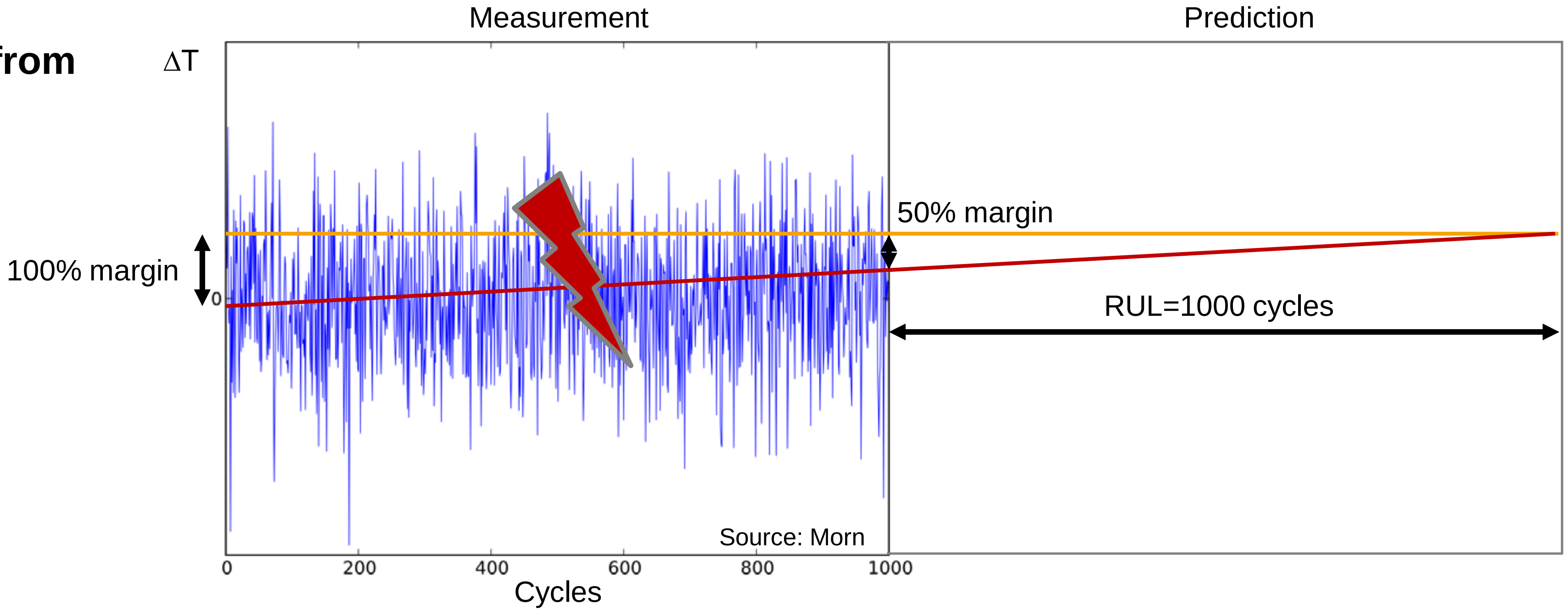
- Diagnosis focusses on part 2 in OSA-CBM
- Given dataset is manipulated (filtered) in order to find or create features for state detection
- Health assessment by comparing the state with run to failure data, threshold values or similar
- The Health indicator can then be used for prognosis of the remaining useful lifetime (RUL)

Diagnosis vs Prognosis

Different capabilities provide a distinction

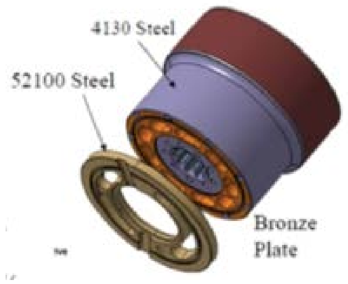
Ops	Description	Diagnosis	Prognosis	Prescription
Acquisition	Understanding & wrangling	Use of descriptive findings and models to describe a system state	Behavioral estimation on the basis of the current state and future influences	Derivation of decision support and recommendations / automation

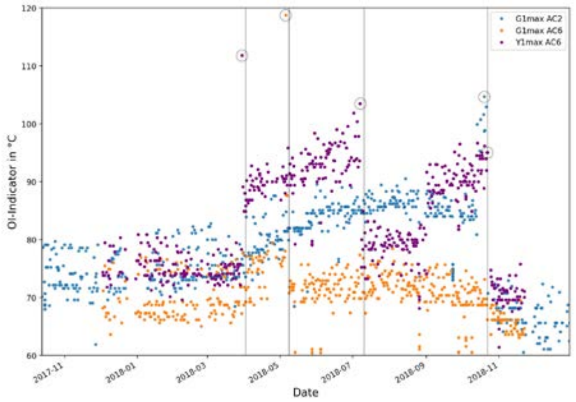
An example from rail transport

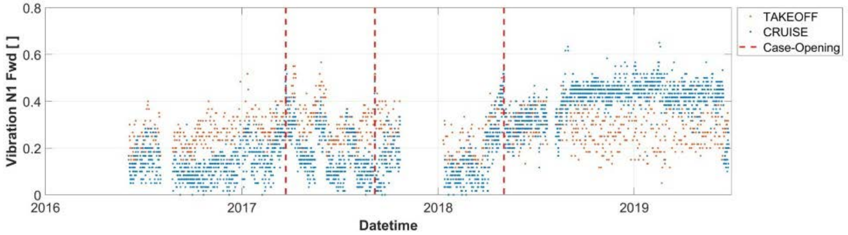


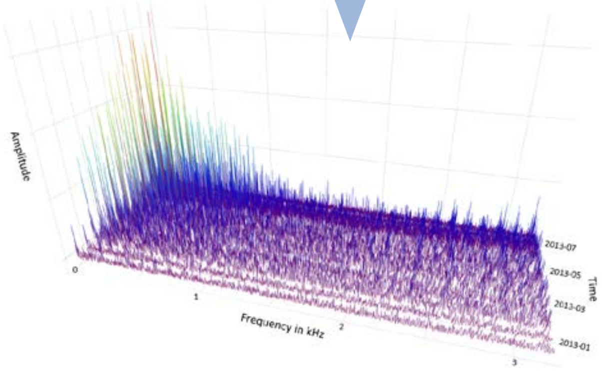
Real world data!

Challenges? Not always the best feature









- Only temperature (-ratio) as feature
- Vibration/temperature/speed signals, but filtered down to singular data points
- Vibration sampled with 40kHz at multiple positions, constant intervals, almost constant measurement conditions

Challenges when handling real world data

- Data is not available at all
- Data might be restricted to specific owner (data ownership)
- Data does not represent underlying behavior
- High class inbalance (more than 95 % „healthy“ samples)
- Unexpected/unknown changes of components
- Varying environment conditions
- Varying operational conditions and settings
- Changes of components before failure (no run-to-failure)

The prognosis is one of the last steps in OSA-CBM

TECHNISCHE UNIVERSITÄT DARMSTADT

1.1	Data Acquisition (DA)
2.1	Data Manipulation (DM)
2.2	State Detection (SD)
2.3	Health Assessment (HA)
3.1	Prognostics Assessment (PA)
3.2	Advisory Generation (AG)
4.1	Presentation

Increasing Level of Aggregation

- PA (typically) completely relies on previous steps 1.1 till 2.3
- PA aims to predict future behavior of investigated component
  How will the health of my component develop in the future?
- HA describes the health of a component (e.g. Health Index)
  What is the health status of my component?

咨询

- AG combines result of PA with system/expert knowledge
  Which actions and when should I take for the component?

There are four assumptions that are ideally considered for prognosis

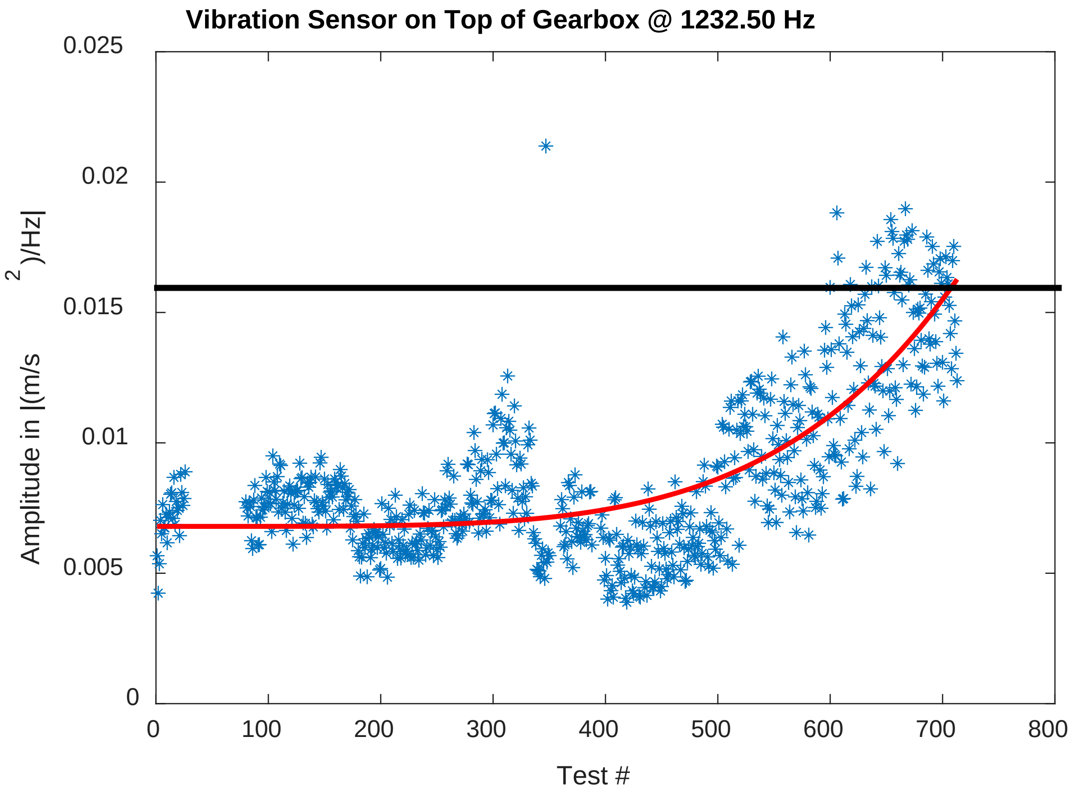
TECHNISCHE UNIVERSITÄT DARMSTADT

1. The monitored system **degrades** as a function of **use**, **time** and **environmental conditions**

2. The **aging** and **damage accumulation** is a **monotonic process** 单调

3. Signs of **aging** are **visible before** the **failure** of the system occurs

4. Signs of **aging** can be **fitted** to a **model** to estimate the remaining useful life



Uncertainty in Prediction

A look into the future never gives a certain answer

- Phenomena of any prognosis like weather forecasts, stock forecasts, etc.
- Input uncertainty
  - Material properties → the reason to test more than one component
  - Initial or boundary conditions → the environment has an influence
  - Sensor uncertainty → the reason for sensor calibration
- Discretization uncertainty
  - Time steps (sample rate) → real world is continuous – information between samples is lost
  - Floating-point number precision → conversion of analog values to discretized values
- Model uncertainty
  - Representation of the real world problem → algorithm output vs. real world output

Each prognosis needs to be expressed as a probability

- Each prognosed degradation at time t_k is a probability distribution for $x(t_k)$
- The probability of failure is expressed as a 累积 cumulative distribution function (CDF)

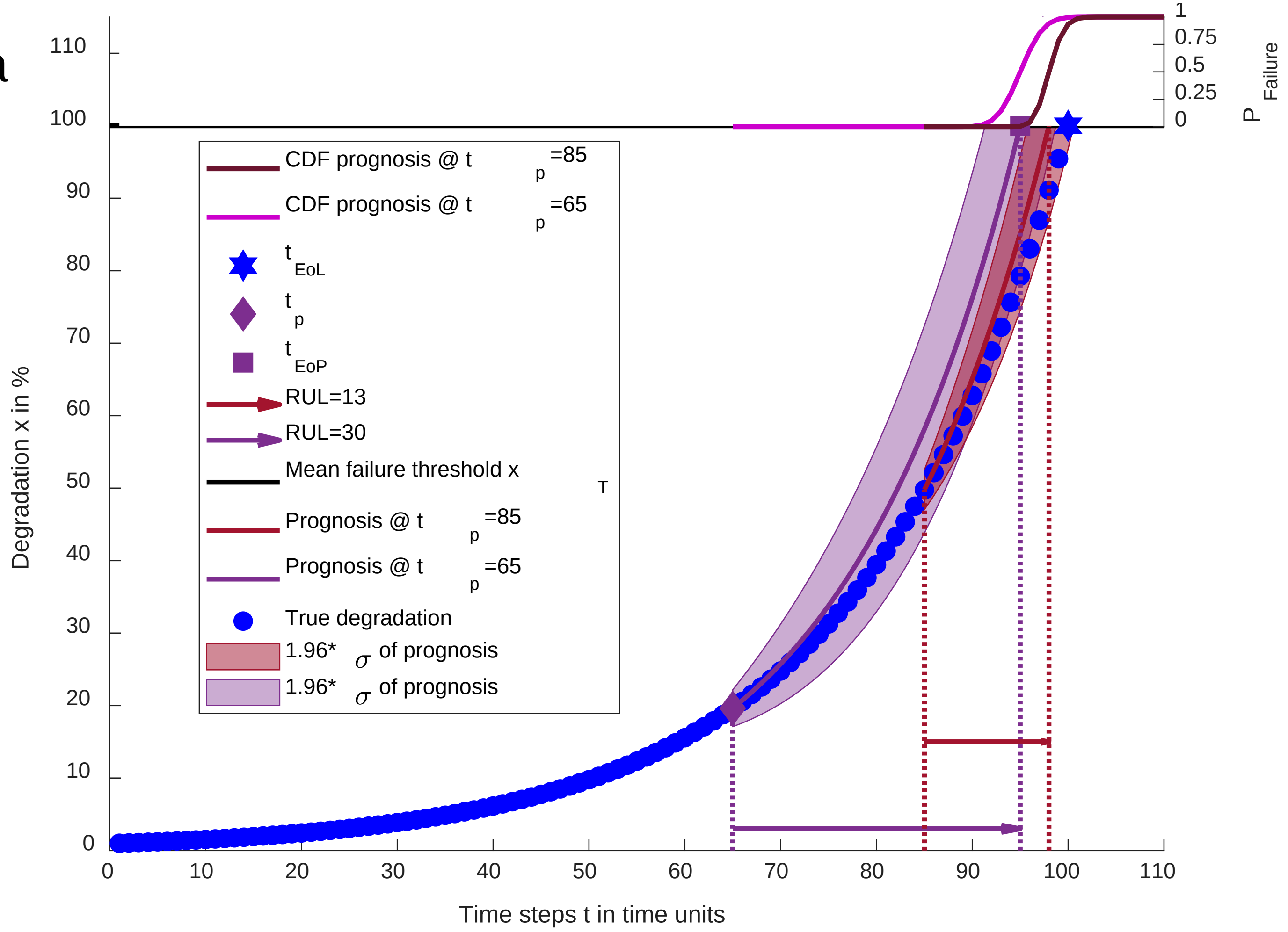
→ Reaching the threshold at time t_k is given with a probability

→ How to calculate the RUL?

  → Mean of CDF reflects the expectation value of discrete distribution

  → Median of CDF reflects a probability of 50 % that a component will have failed until that time

  → Specify a distinct probability value for the CDF



Recap

PHM FOR COMPLEX SYSTEMS

Complex Systems

Performability = Performance + Dependability

Performance
ability of a system to accomplish its intended services within given non-functional constraints (e.g. time)

Timeliness
ability of the system to provide a service according to given time requirements

Precision
ability of the system to provide the same results under unchanged conditions

Accuracy
ability of the system to provide exact results

Capacity
ability of the system to hold a certain amount of data

Throughput
ability to handle a certain amount of operations

Dependability
ability of a system to provide its intended services in a justifiable way

Availability
readiness for correct service

Reliability
continuity of correct service

Safety
absence of catastrophic consequences

Integrity
absence of improper system state alterations

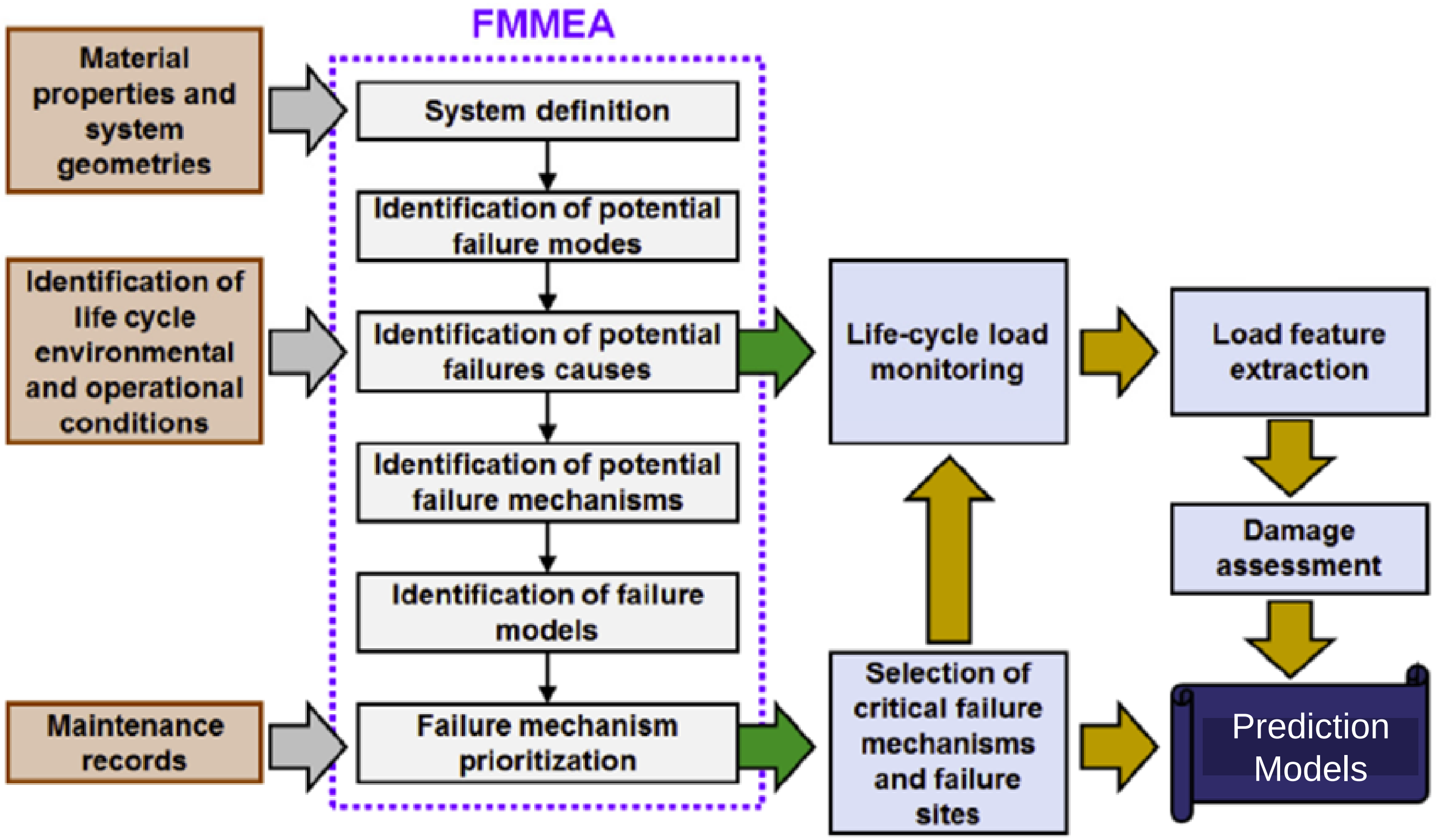
Maintainability
ability to undergo modifications and repairs

[Bertolino, 2011]

PHM & Reliability
General approach

How to identify suitable PHM monitoring strategy

- Run failure modes, mechanisms and effects analysis
- Identification of relevant (physical) failure precursors 先导
- Select feasible diagnosis/prognosis scheme 可行

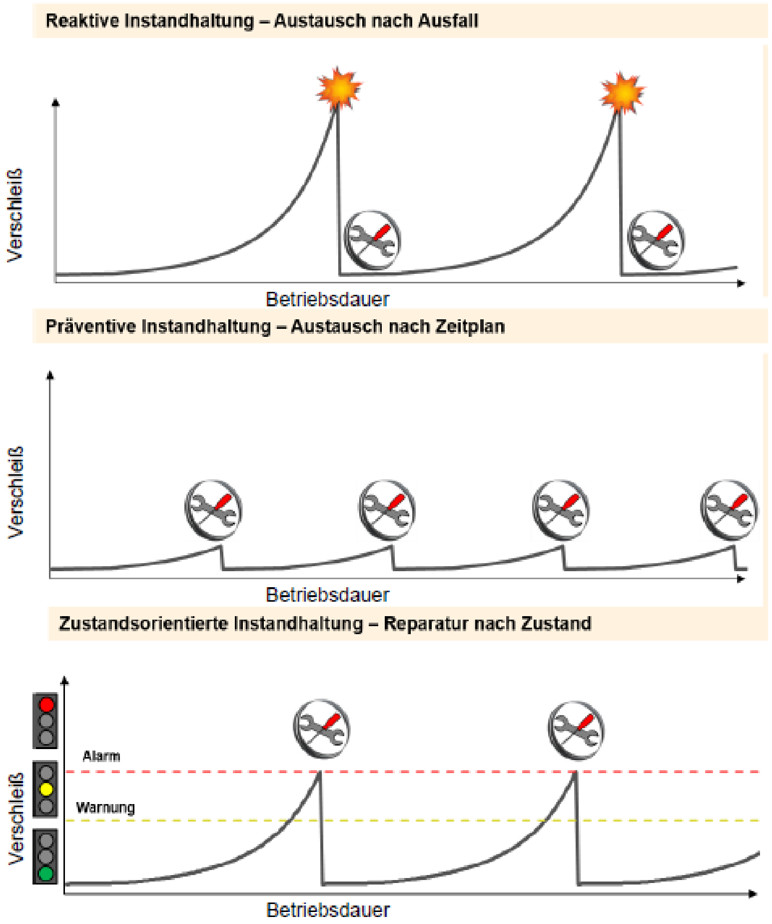


[M. Pecht, 2018]

Recap

PREDICTIVE MAINTENANCE AND PREDICTIVE QUALITY

Different Models for Maintenance and Necessity for Action



Reactive maintenance:
- Maintenance after machine failure
- Unplanned, after shutdown
- Best possible utilisation of the lifetime

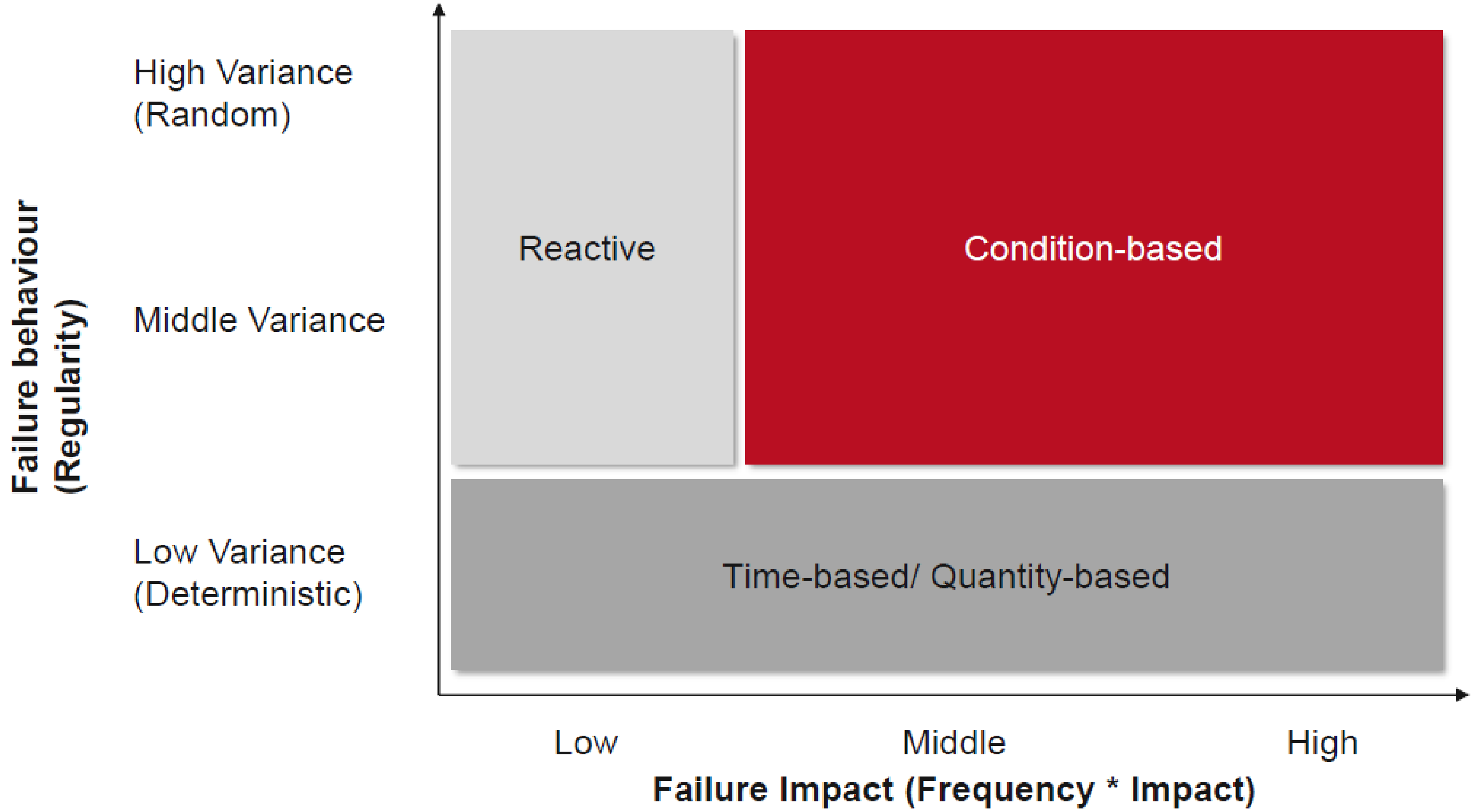
Time-based, Quantity-based, preventive maintenance:
- Maintenance according to fixed intervals
- True machine condition remains unnoticed: Exchange often not necessary

Condition-based, predictive maintenance:
- Indication of imminent failures
- Best possible utilization of the "machine life"

即将来

利用率

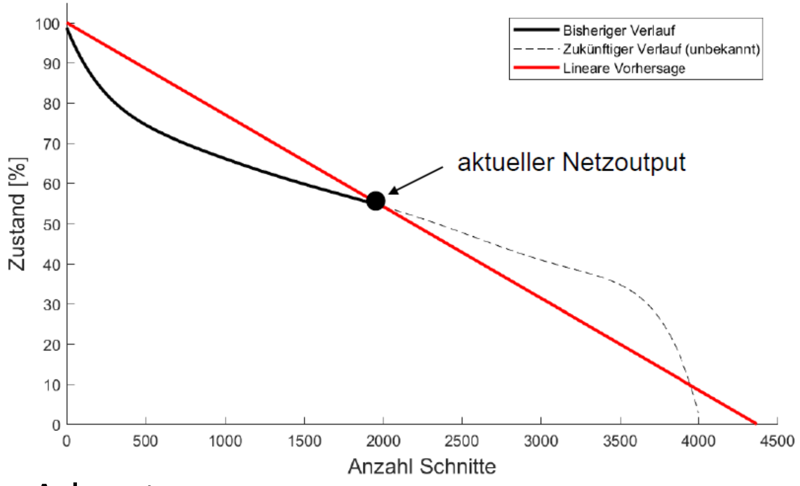
Simple Decision Rules for Maintenance Strategies



Prediction of the Future

(1) linear, (2) approximation, (3) time series prediction with neural network

(1) Linear lifetime prediction



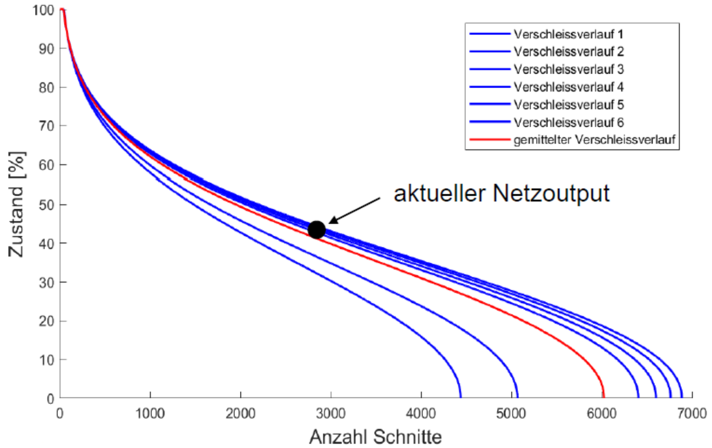
Advantage:

- Ease of implementation

Disadvantage:

- Assumption of a linear wear behavior

(2) Randomly generated and average wear curves



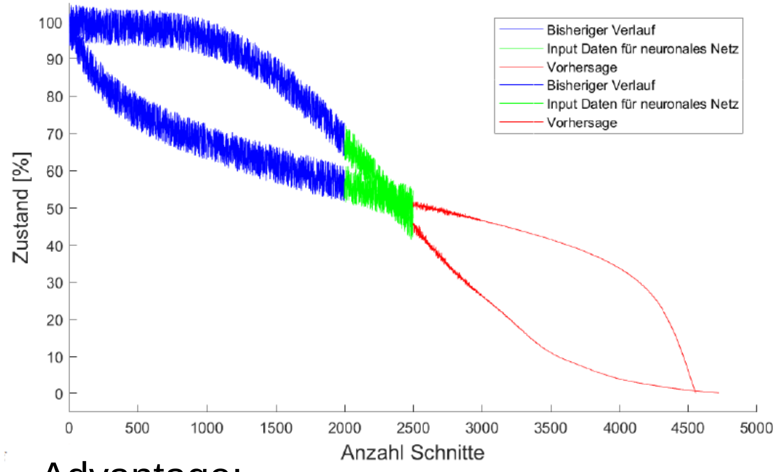
Advantage:

- Prediction of the real wear behavior
- Valid, even if boundary conditions have been changed before the estimation

Disadvantage:

- Only valid if current and future boundary conditions correspond to the curve

(3) Lifetime prediction based on the last steps



Advantage:

- Prediction of the real wear behavior
- Can recognize and distinguish different wear behavior

Disadvantage:

- At least one continuous wear curve must be available for each case
- Data of the last cuts must be available

Predictive Quality

Approach

predictive quality

Determination of quality indicators on the basis of data without a dedicated measuring operation.

conventional

machining → measurement of produced part to determine product quality

- ⊖ time consuming
- ⊖ equipment and staff necessary (investment and maintenance)
- ⊕ conventional documentation possible

predictive quality

machining

analysis of machining data to determine product quality

- ⊕ trained models allow fast analysis and feedback
- ⊕ server maintenance cost < measuring equipment maintenance cost
- ⊖ conventional documentation not possible

DMME Process

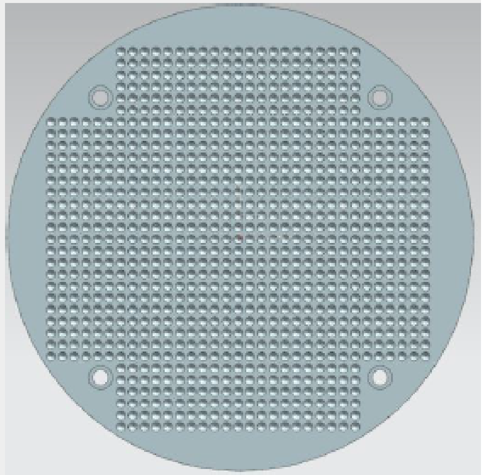
Technical Understanding

TECHNISCHE UNIVERSITÄT DARMSTADT

procedure

How can the quality characteristics be produced?

- determine framework conditions (e.g. machine tool selection)
- perform technical system analysis
- identify relevant parameters
- create measurement concept
- create experimental design



use case

Scope:

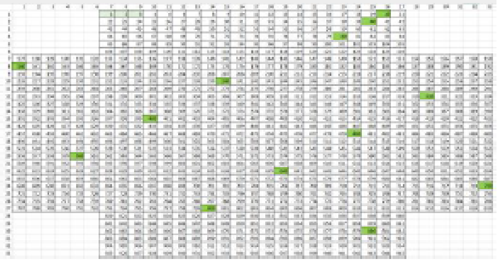
Machine tool, measuring machine, tool cycle

Analysis of the relevant parameters using domain knowledge about the machining process (spindle current, axis position, rpm)

In this step process knowledge is indispensable so far!

Experimental design:

- G code
- material & tools
- No. and arrangement of bores



Result | **Clarity about the experimental design achieved.**

DMME Process

Technical Realisation

procedure	use case

▪ set up experimental design
▪ conduct an experiment
▪ documentation
	▪ machine tool: Hermle
	  ▪ machine tool setup and conduction
	  ▪ problem of machine data acquisition
	▪ coordinate measuring machine
	  ▪ setup measuring programme

Open Research:

Required frequency

- required frequency strongly depends on underlying business case

Required parameter

- ML aims for finding correlations that go beyond expert knowledge ->the more parameters the better

Approach: Use max. option and reduce in deployment

Problem: Technical restrictions result in a trade-off between frequency and number of recorded parameters

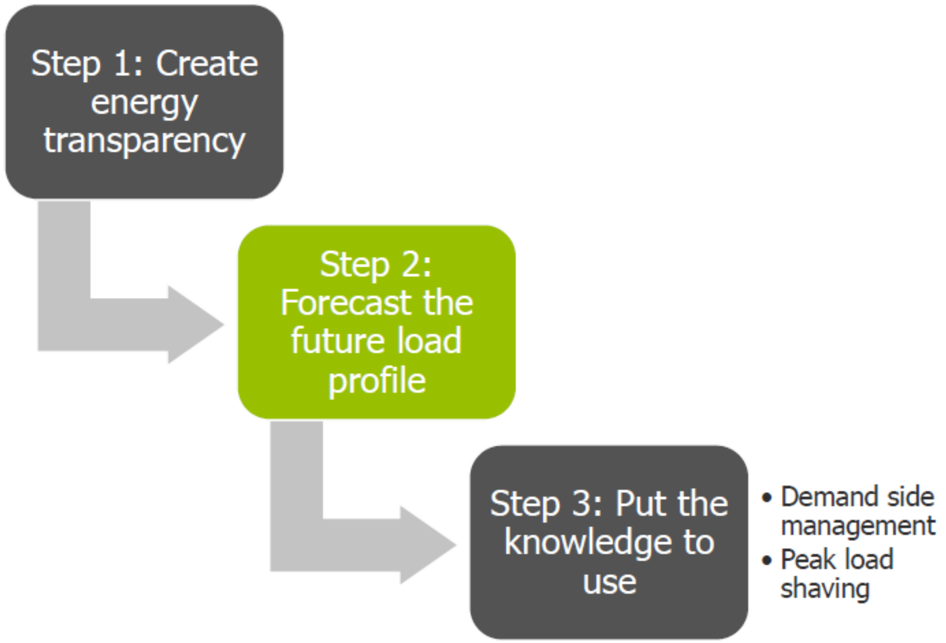
-> Generalised findings are still subject of research

Recap

ENERGY FORECASTING

Energy forcasting in industry – why?

- Pricing structures and price fluctuations at the electricity market **force the industry to adapt their electric load profile** to the electricity supply
- The load profile of the factory is strongly influenced by the load profiles of the production machines inside the factory.
- Knowing the future load profile of the production machines enables us to control the load profile of the factory.



Forecasting vs. prediction

Definition of terms

TECHNISCHE UNIVERSITÄT DARMSTADT

Prediction:

- General term → predict an unknown value from known inputs
- Example: Prediction of the net income of households from house location, house size, number of rooms, etc…

Forecasting:

- Time related → forecast the future values of a time series
- Example: Weather forecast of tomorrow from current and past weather conditions, time of year, …
- Challenge of forecasting in Machine Learning:
  - Feature engineering gets a second dimension: Time
  1. Model the exogenous, non-temporal features (the feature model)
  2. Model the historical, temporal features (the temporal model)
  - Therefore, the feature set often becomes much larger than in regular Machine Learning problems, which can evoke the “curse of dimensionality”

Data understanding in the course of CRISP-DM

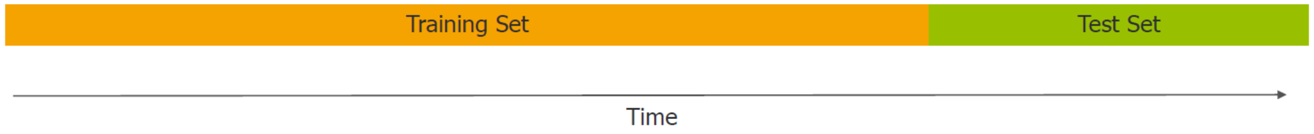
Peculiarities of time series

- The values have a strong time dependency (autocorrelation)
  - ➢ Different splitting into training, validation and test data required
    - Test data must always be future values to ensure generalization capability of the model
    - No shuffling to prevent mixing future and past values
  - ➢ Preparation for supervised learning needed
    - Target must be time shifted so that the model learns the future behavior from the current/past inputs
  - ➢ Different feature engineering required
    1. Model the exogenous, non-temporal features (the feature model)
    2. Model the historical, temporal features (the temporal model)

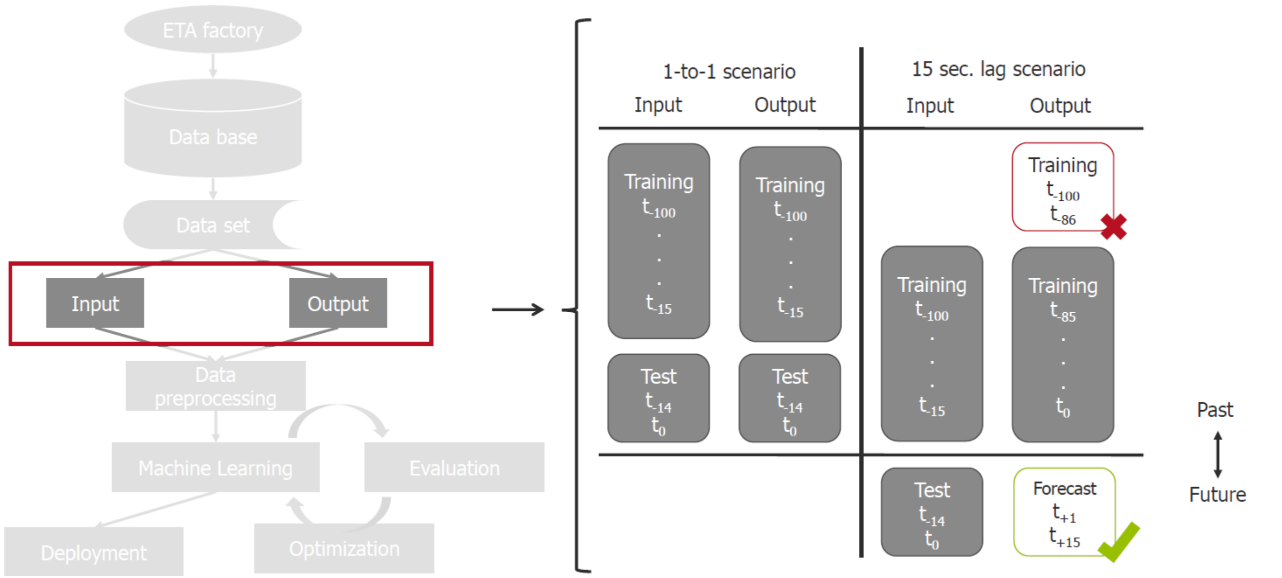
Data Preparation on the course of CRISP-DM

Splitting of the data and time shift method

- Time series values are strongly dependent on values that are close in time
- Therefore, the test/validation set should always be a set of future values
- No shuffling to prevent mixing future and past values



→ **Time shift method** for forecasting target preparation:



Data Preparation on the course of CRISP-DM

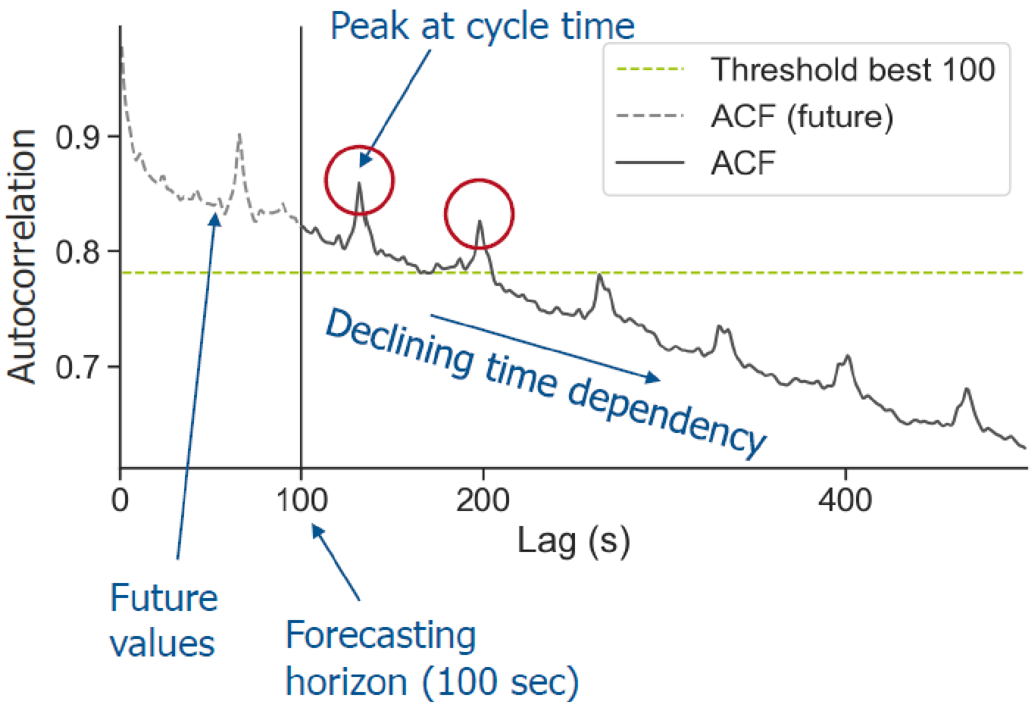
Feature engineering for the temporal feature set

Which time step(s) are significant?

Time	Param 1	Engineered feature 1 (time lag ?? steps)	...
t-0	$x_1^{(0)}$	$x_{e1}^{(0)} = x_1^{(??)}$	
t-1	$x_1^{(1)}$	$x_{e1}^{(1)} = x_1^{(??)}$	
t-2	$x_1^{(2)}$	$x_{e1}^{(2)} = x_1^{(??)}$	
t-3	$x_1^{(3)}$	$x_{e1}^{(3)} = x_1^{(??)}$	
t-4	$x_1^{(4)}$	$x_{e1}^{(4)} = x_1^{(??)}$	
...	...	...	

The autocorrelation function (ACF)

Goal: Identify promising time lags for Feature Engineering



TECHNISCHE UNIVERSITÄT DARMSTADT

Information Related to the Written Exam

INFORMATION ON EXAM

Exam information

TECHNISCHE UNIVERSITÄT DARMSTADT

Organisational information

- Date: February 17th, 2020
- Time: 10.00 – 11.00 (60 minutes)
- Room: L4|02 – 1 & 2

Allowed aids

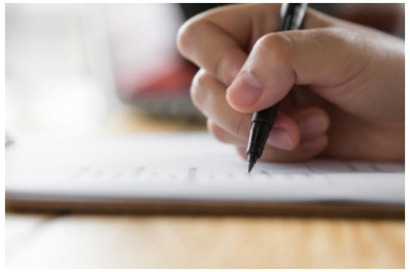
- Calculator (not programmable)
- Pen & Ruler

Question types

- Terms & Definitions
- Understanding of Methods/Algorithms (What? How? Pro/Con? Application?)
- Tasks with graphical or calculation solution

No need to learn

- Formulas; except something easy like
  - *Mean* $\frac{1}{N}\sum_{n=1}^{N} x_n$
  - *Error* $\Delta = y_{real} - y_{predict}$
  - *Linear functions (also in* $\mathbb{R}^p$*)* $y = m \cdot x + b$
  - *Accuracy, Precision, MAE, MSE,…*
  - *…*
- Any programming commands



Picture source: freepik.com, unsplash.com

Last lecture before Written Exam

PREVIEW FOR NEXT WEEK

Operational Control

Optimized Control of Cross-Linked Energy Systems by Means of Reinforcement Learning

- Deep Reinforcement Learning (DRL)
- How DRL can be used to control industrial utilities (and other interactive systems)
- Students will be given tools to model their own complex systems with DRL

OH AND ONE LAST THING…

Oh and one last thing… Evaluation of MLA lecture!

TECHNISCHE UNIVERSITÄT DARMSTADT

- Evaluation will be available from February, 24th (12 p.m.) to March, 09th (12 p.m.).
- Link will be provided in the MLA moodle course

Why a second evaluation?

- We want to improve the lecture in your interest
- We want to have your feedback on the structure of the MLA lecture, the Hackathon and the structure of the written examination

What we took from previous evaluation?

- Lots of work for 4 CP
- It's not clear what will be part of written exam
- Redundancy in the lecture's content

Time for your questions and suggestions…