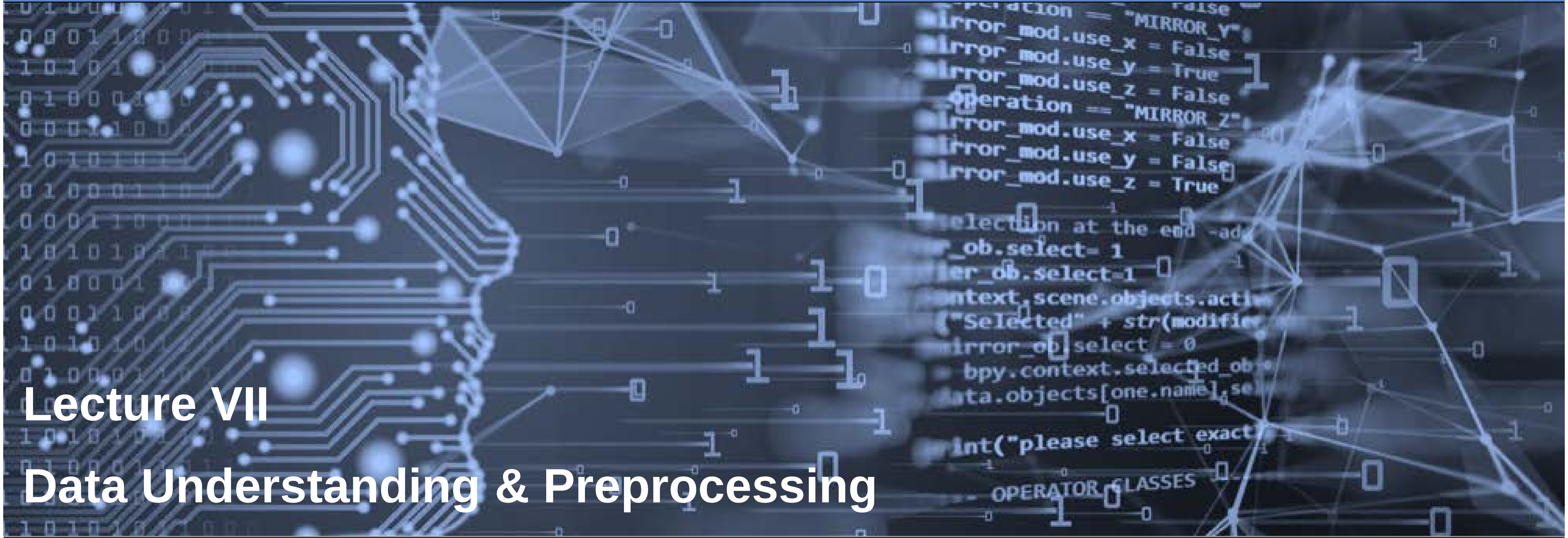
# Machine Learning Applications

TECHNISCHE UNIVERSITÄT DARMSTADT

**Winter semester 2019/2020**

**Henrik Simon & Sebastian Baumann**



## Lecture VII

## Data Understanding & Preprocessing

# What should you be able to take out of the lecture today?

TECHNISCHE UNIVERSITÄT DARMSTADT

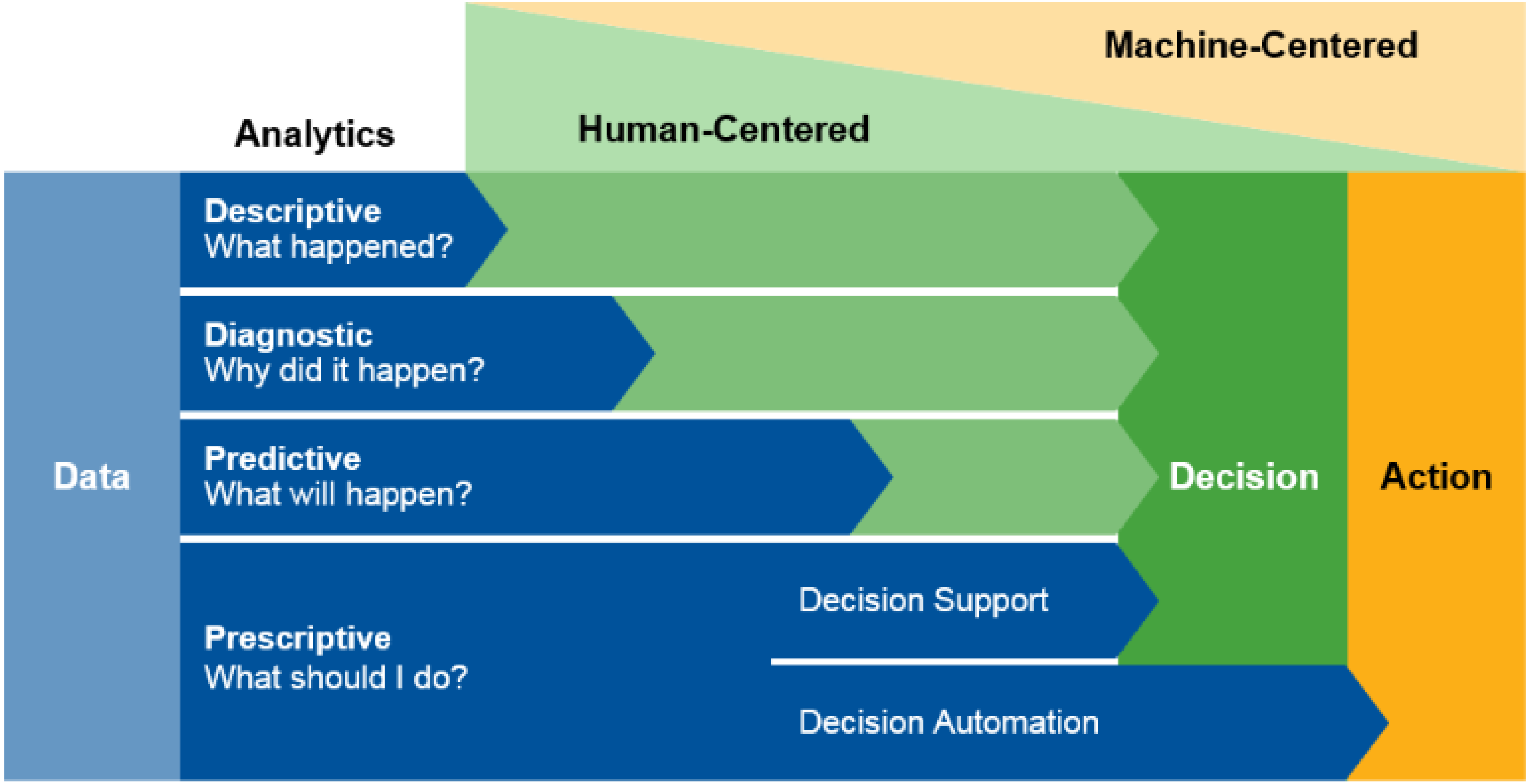
DATAbility.ai
Engineering know-how meets artificial intelligence.

- What are important approaches for machine learning projects (process models, design principles)?
- When the use of the word prediction should be avoided (diagnosis vs. prognosis)?
- Why is business and data understanding important?
- How can we assess data quality and how can we trust data?
- How to deal with outlier and measurement uncertainties? What are negative influences?
- Why correlation analyses can be useful, but should be avoided?
- How can we extract and select meaningful predictors for modeling?

| Process models | Data acquisition | Data understanding | Data preprocessing | Demo |
|---|---|---|---|---|
| Approaches | Sources | Correlations | Feature engineering | |
| Principles | Information | | | |
| comprehension | Data quality | | | |
| | Measurement errors | | | |
| | Outliers | | | |
| | Impacts | | | |

# RECAP: PROCESS MODELS AND MACHINE LEARNING APPROACH

# The four analytic capabilities of business intelligence: delivering hindsight, insight and foresight

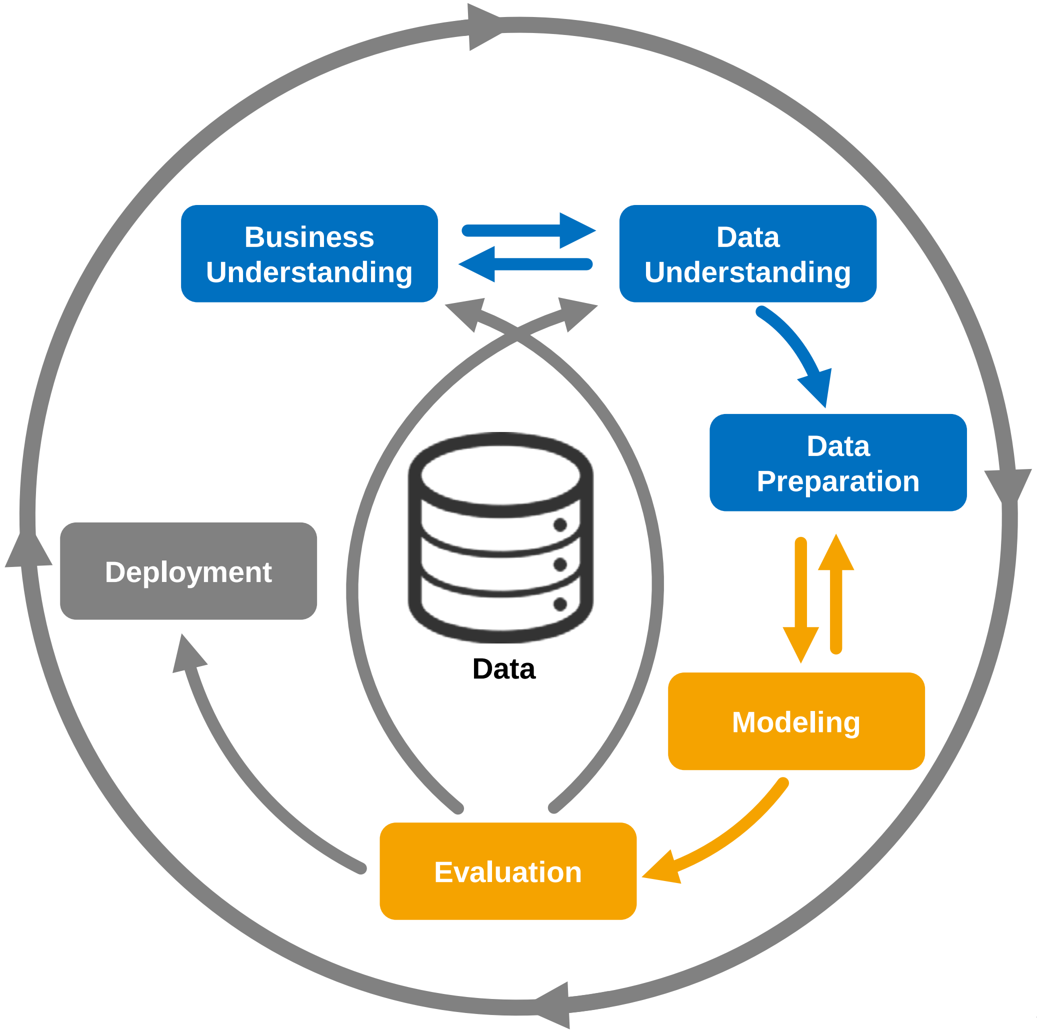



Source: Gartner Inc. [Publ.]: 2017 Planning Guide for Data and Analytics. Technical Professional Advice, G00311517 (2016)

# Business Understanding and Data Understanding form the basis of a data mining / machine learning cycle.

TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.




## Business Understanding

- initial phase that focuses on understanding the project objectives and requirements from a business perspective
- converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives
- defines a clear objective and research question. What do you want to realize?

  Clustering, Classification, Regression | Diagnosis vs. Prognosis

## Data Understanding

- based on an initial data collection; activities to get familiar with the data, to identify data quality problems, to discover first insights or to detect interesting subsets to form hypotheses for hidden information
- takes major influences on all following steps (GIGO: garbage in - garbage out)
- definition and proof of data requirements

## Data Preparation

- covers all activities to construct the final dataset from raw data, time consuming (50%-70%)
- devoting adequate energy to business understanding and data understanding can minimize this effort

Source: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm

# Deeper look into CRISP-DM process steps

TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.

### Business Understanding

**Determine Business Objectives**
*Background*
*Business Objectives*
*Business Success Criteria*

**Situation Assessment**
*Inventory of Resources*
*Requirements, Assumptions, and Constraints*
*Risks and Contingencies*
*Terminology*
*Costs and Benefits*

**Determine Data Mining Goal**
*Data Mining Goals*
*Data Mining Success Criteria*

**Produce Project Plan**
*Project Plan*
*Initial Asessment of Tools and Techniques*

### Data Understanding

**Collect Initial Data**
*Initial Data Collection Report*

**Describe Data**
*Data Description Report*

**Explore Data**
*Data Exploration Report*

**Verify Data Quality**
*Data Quality Report*

### Data Preparation

*Data Set*
*Data Set Description*

**Select Data**
*Rationale for Inclusion / Exclusion*

**Clean Data**
*Data Cleaning Report*

**Construct Data**
*Derived Attributes*
*Generated Records*

**Integrate Data**
*Merged Data*

**Format Data**
*Reformatted Data*

### Modeling

**Select Modeling Technique**
*Modeling Technique*
*Modeling Assumptions*

**Generate Test Design**
*Test Design*

**Build Model**
*Parameter Settings*
*Models*
*Model Description*

**Assess Model**
*Model Assessment*
*Revised Parameter Settings*

### Evaluation

**Evaluate Results**
*Assessment of Data Mining Results w.r.t. Business Success Criteria*
*Approved Models*

**Review Process**
*Review of Process*

**Determine Next Steps**
*List of Possible Actions*
*Decision*

### Deployment

**Plan Deployment**
*Deployment Plan*

**Plan Monitoring and Maintenance**
*Monitoring and Maintenance Plan*

**Produce Final Report**
*Final Report*
*Final Presentation*

**Review Project**
*Experience Documentation*

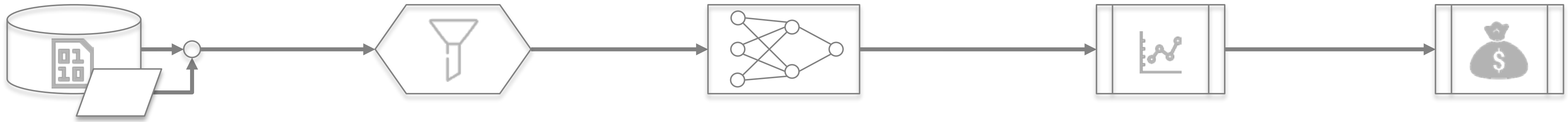
Source: Pete Chapman, *The CRISP-DM User Guide* https://s2.smu.edu/~mhd/8331f03/crisp.pdf

# Design principles

TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.

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

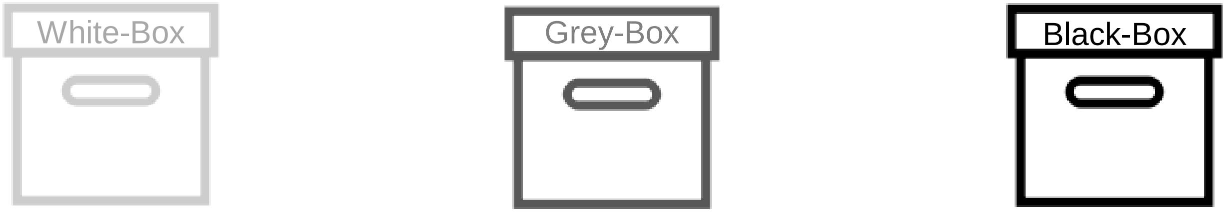


- Mathematical / physical models
- SME knowledge
- Challenging for digitization
- Expensive for complex systems

Engineering approach

**Intelligent approach**
***SMART DATA***

Computer science approach




- 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
- Derivation of explainable AI / ML frameworks

# DIAGNOSIS VS. PROGNOSIS

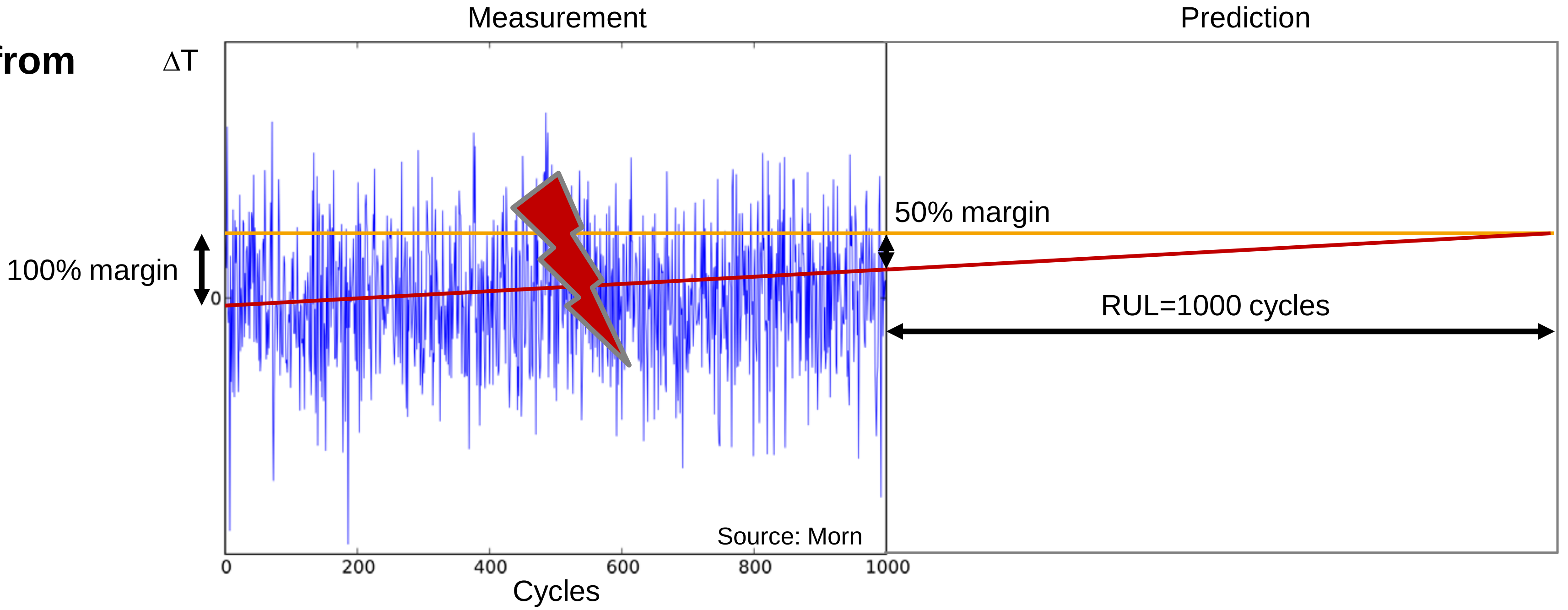
# The different capabilities provide a distinction between diagnosis and prognosis.

TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.

| 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**





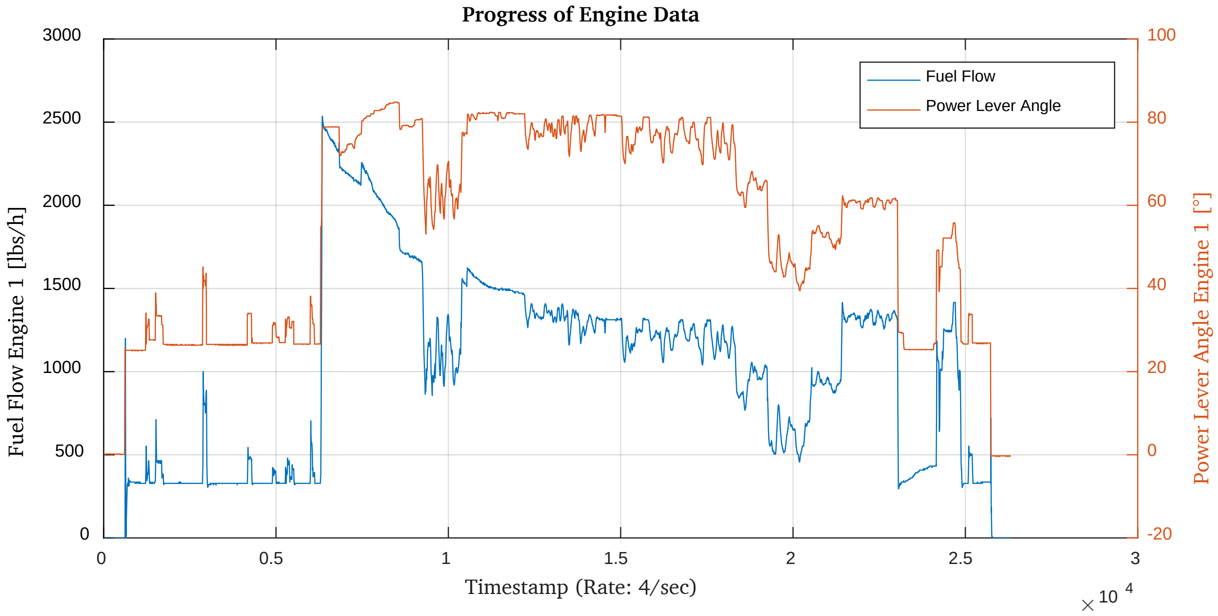

FSR

# The different capabilities provide a distinction between diagnosis and prognosis.

| 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 aviation

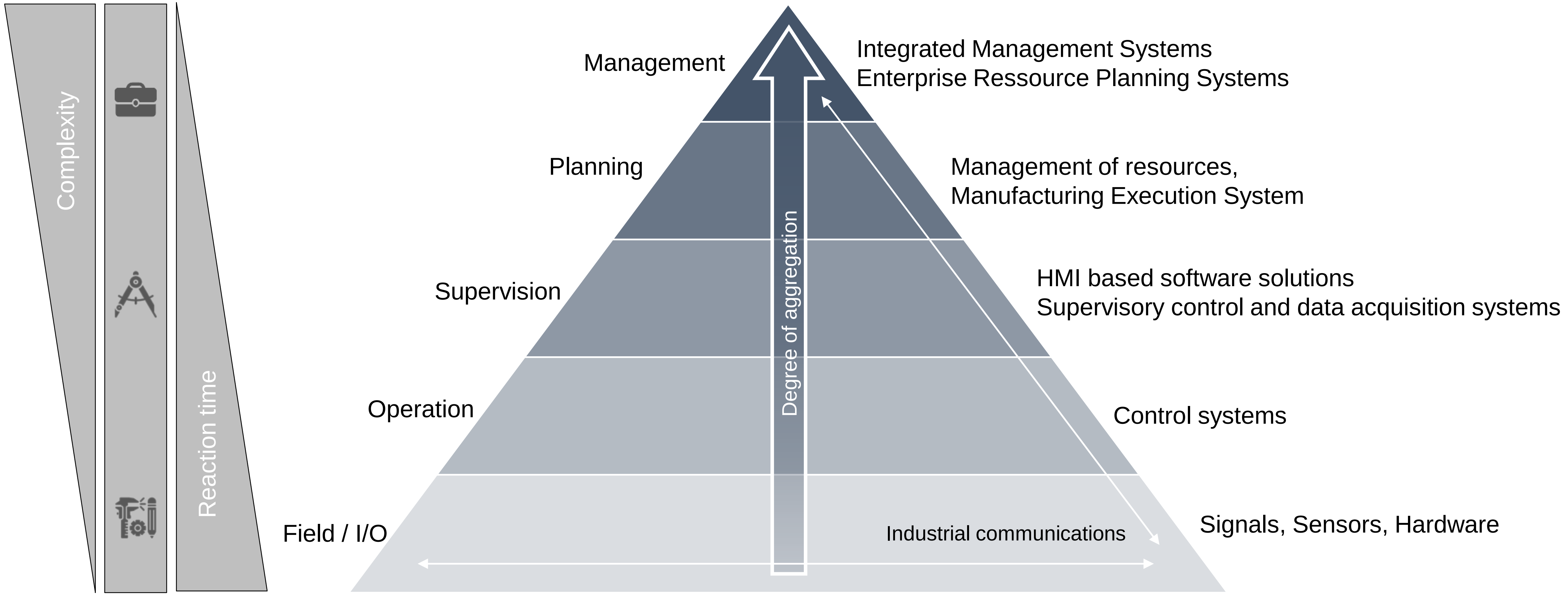
Source: karlenepetitt.blogspot.com




Predictable predictor?
Predictive capability?

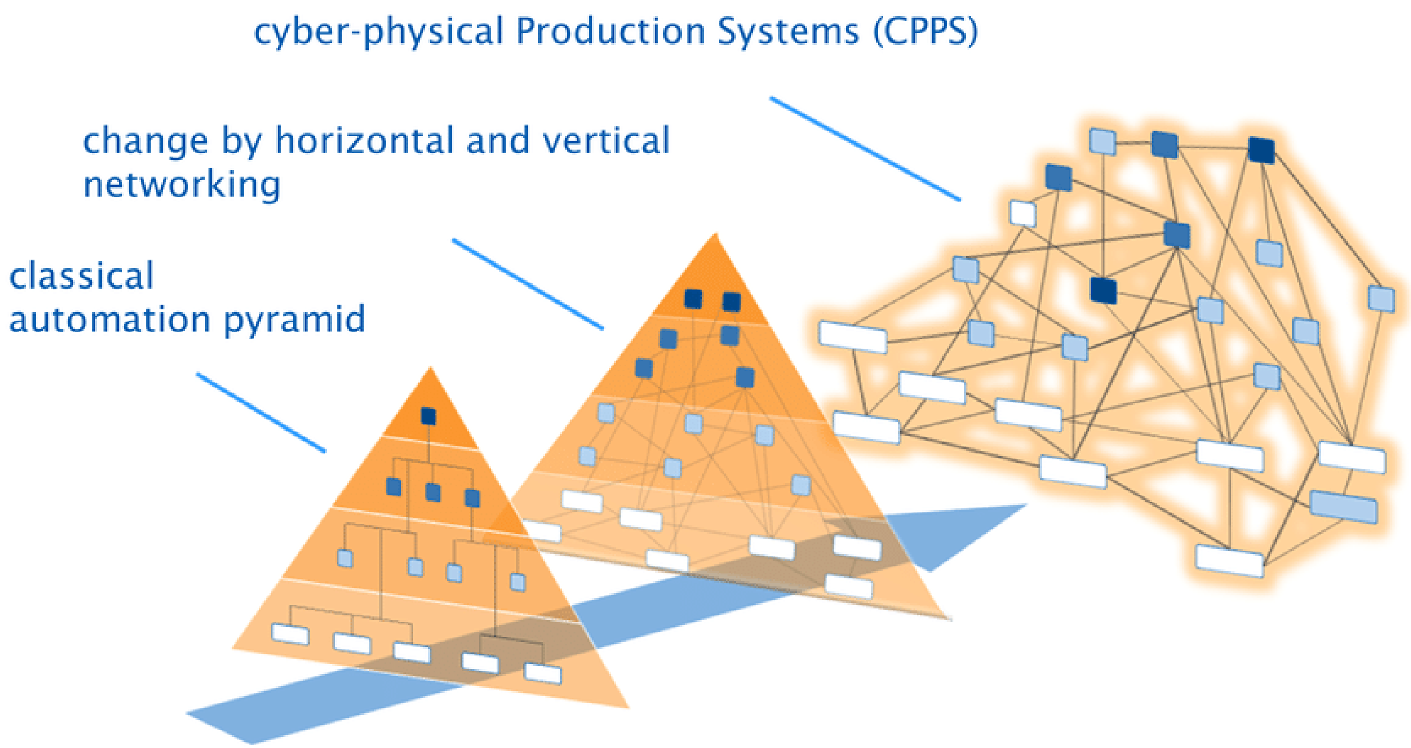
# DATA ACQUISITION

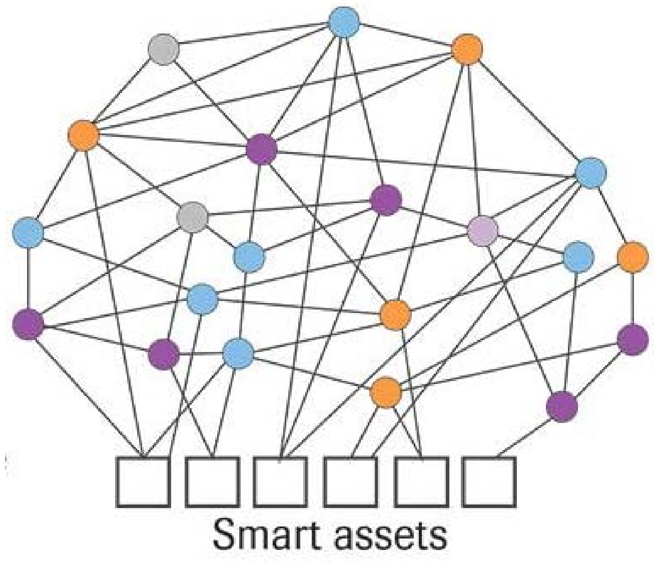
# Various automation levels lead to different information that can have the same data origin.



# Future: Cyber physical systems based automation

## Vertical & horizontal integration




Sources: Lipinski, Richter, Reiff-Stephan: Intelligent sensor systems for self-optimising production chains. In: Proceedings of the 1th Int. Conf. and Exh. on Future RFID Technologies. pp. 115–125 (2014); Labs: Industry 4.0 network architecture relies on interconnectivity, Online: https://www.foodengineeringmag.com/articles/97066-industry-40-network-architecture-relies-on-interconnectivity (2017)

# DATA QUALITY

# Assessing data quality

## How to create trust and confidence

**Definition Data quality**

Data are of high quality if they are suitable for their intended use in operations, for decision support and for the planning of those.

**Definition Meta Data**

Structured information, which describes, explains, localizes, or simplifies in another way the fetch, usage or management of an information source.

# Assessing the data quality

## Classes and dimensions for data quality assessment

TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.

### Understandability

- Expressiveness
- Unambiguousness
- Documentation
- Conformity

### System inherence

- Accuracy/variance
- Authenticity
- Konsistency
- Objektivity
- Redundancy
- Credibility

### Availability

- Access & reference
- Safety & security
- Latency
- Processing
- Reliability

### Relevance

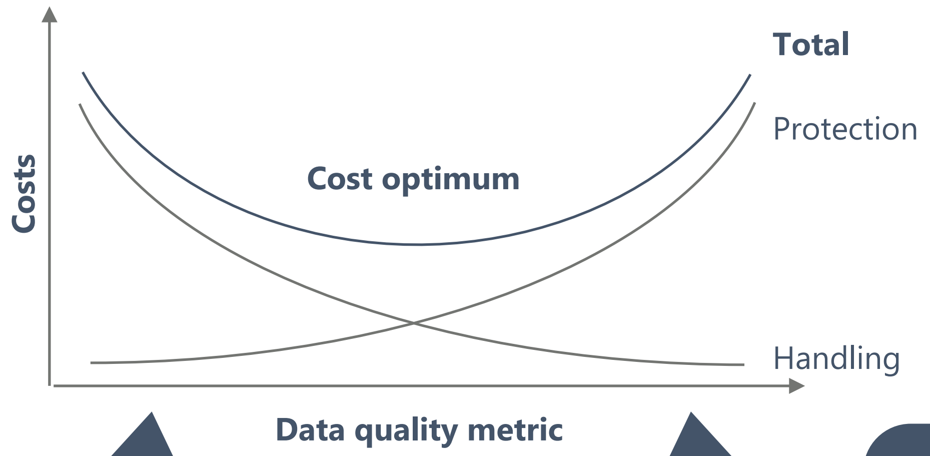
- Timeliness
- Completeness
- Coherence

Area of high data quality

# Assessing the data quality

## Classes and dimensions for data quality assessment

**Investment trade-off for data quality**




### Advancement

- Data cleaning
- System optimization
- Process optimization
- Training
- …

### Management

Define
Measure
Analyze
Improve
Control

### Protection

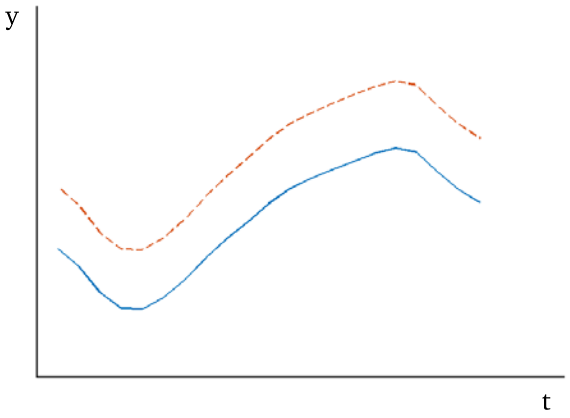
- Auditing
- Self-test mechanisms
- …

# MEASUREMENT ERRORS

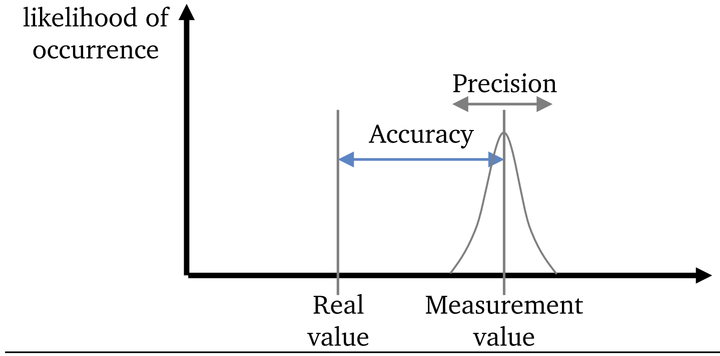
# Illustration of measurement errors

Errors can be classified into three types that can overlap.

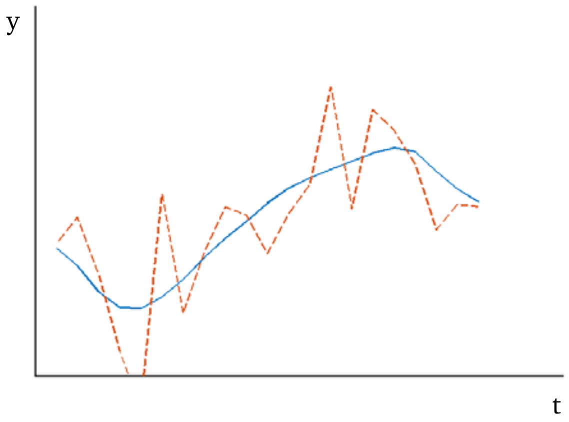
## Systematic errors




### Bias / Offset




## Random errors

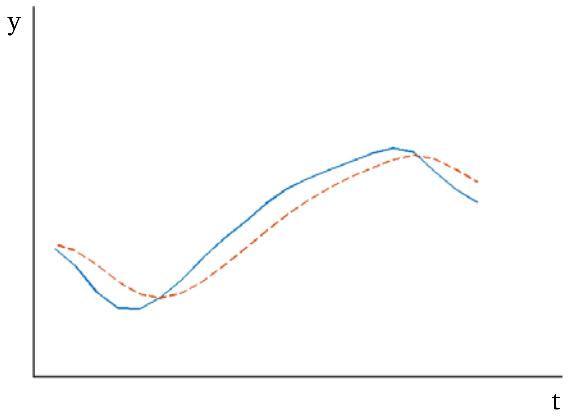



### Variance

Signal to noise ratio (SNR)

$$\frac{P_{Signal}}{P_{Noise}} = 10^{\frac{SNR}{10\,dB}}$$

## Dynamic errors




### Lag / Delay

Delay 1st order

$$G(s) = \frac{1}{1 + Ts}$$

# Measurement errors have a significant influence on the model quality of machine learning models.

TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.

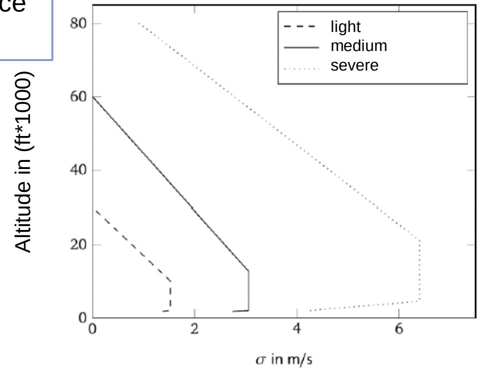
### Model result and target of an aircraft fuel flow in cruise flight

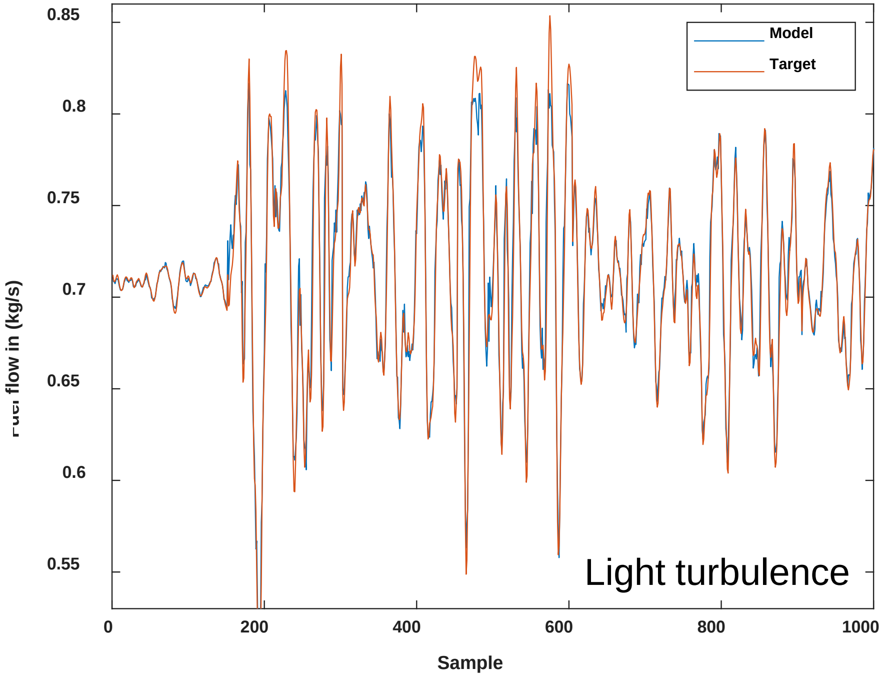
15 Predictors: TAS, ALT, AIL, AOA, TAT, SAT, YAW, ROLL, ACC

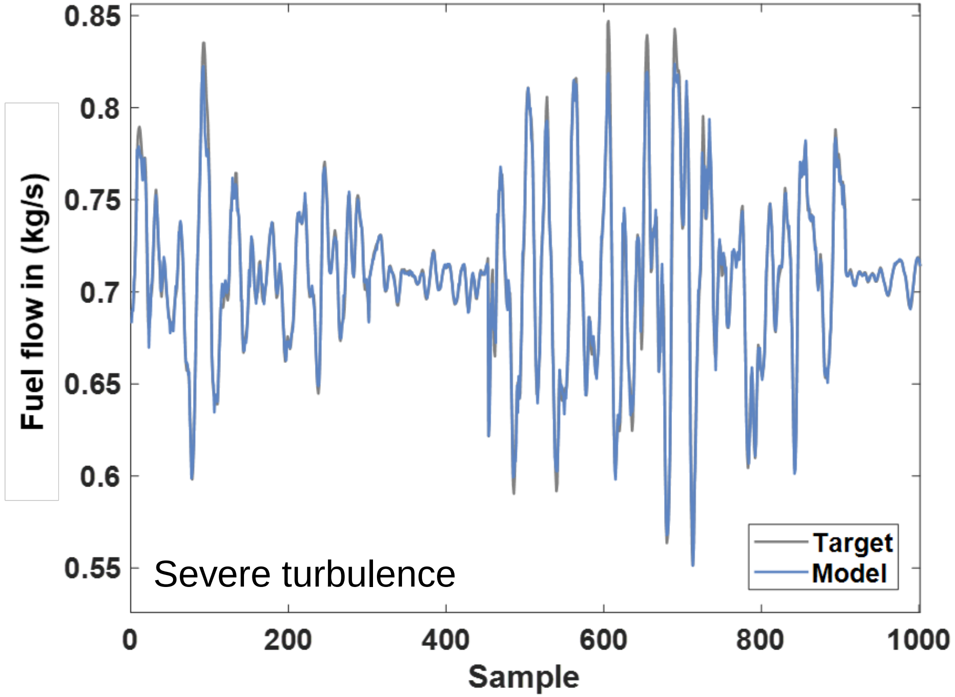
Random Forest

FF

Turbulence model



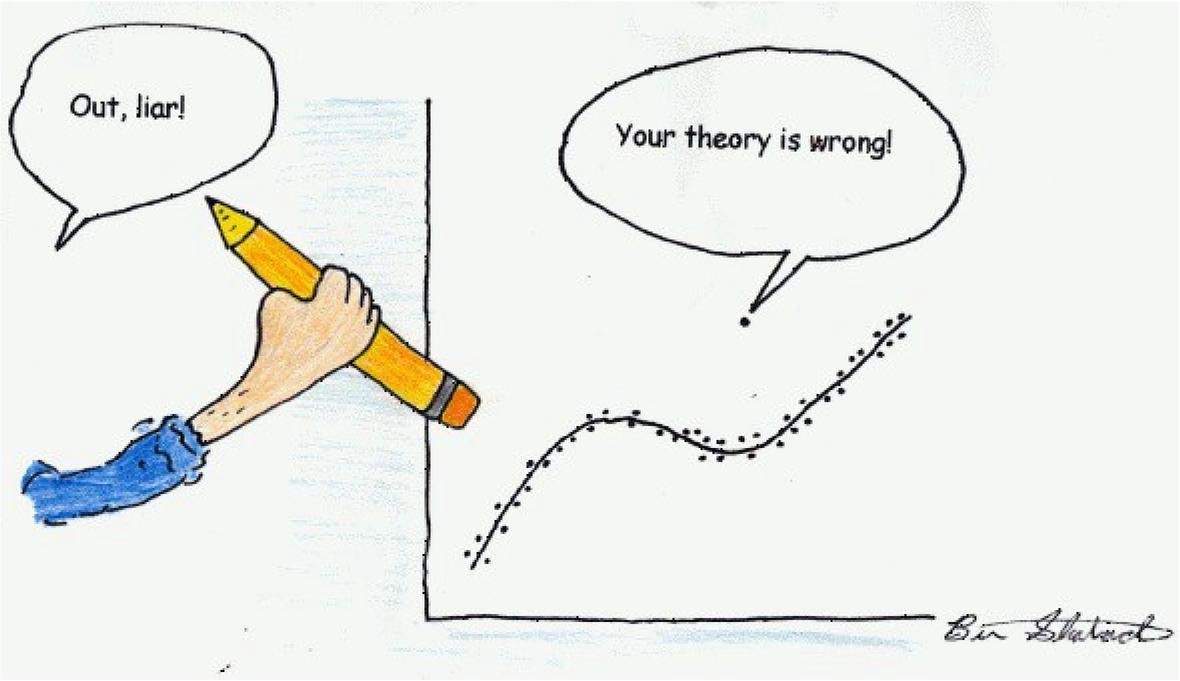

### Performance metrics for models trained with original and manipulated data

| Errors | Turbulence | Model | MAE \| RMSE [kg/s*10-3] | MRE [%] | $R^2$ [%] | p [-] \| h [-] \| r [-] |
|---|---|---|---|---|---|---|
| - | - | Original | .3 \| 1.8 | .04 | 99.01 | |
| dyn. & stoch.. | Light | Manipulation | 6.4 \| 10.3 | .9 | 95.84 | <.01 \| 1 |
| | Medium | Manipulation | 4.2 \| 6.8 | .6 | 98.31 | 0.84 |
| | Severe | Manipulation | 4.4 \| 7.4 | .62 | 97.96 | |



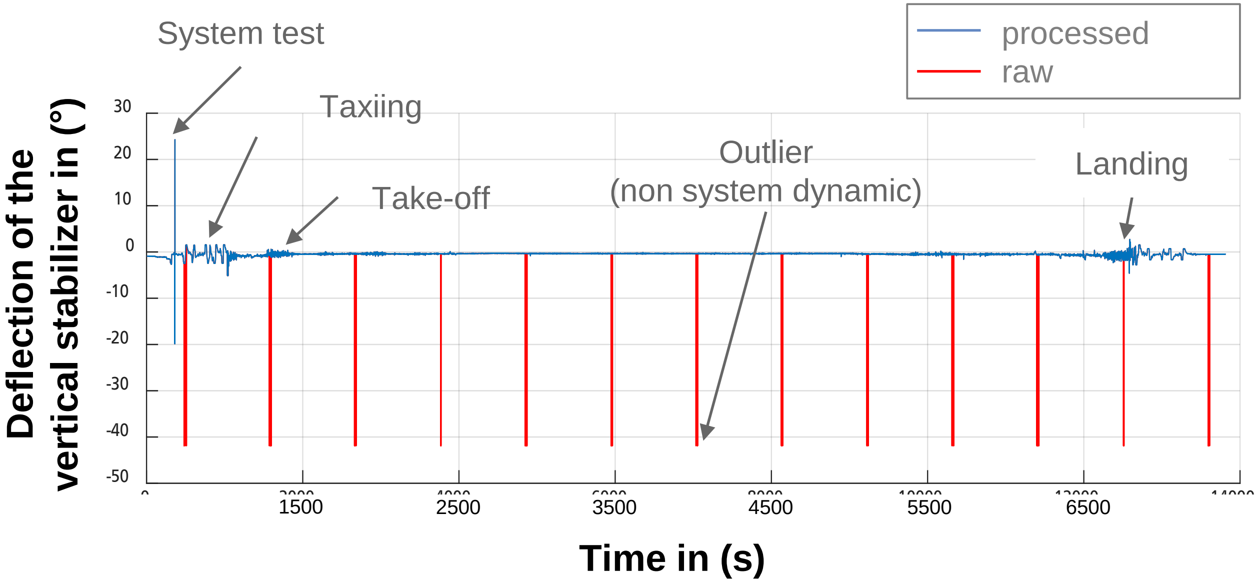
FSR

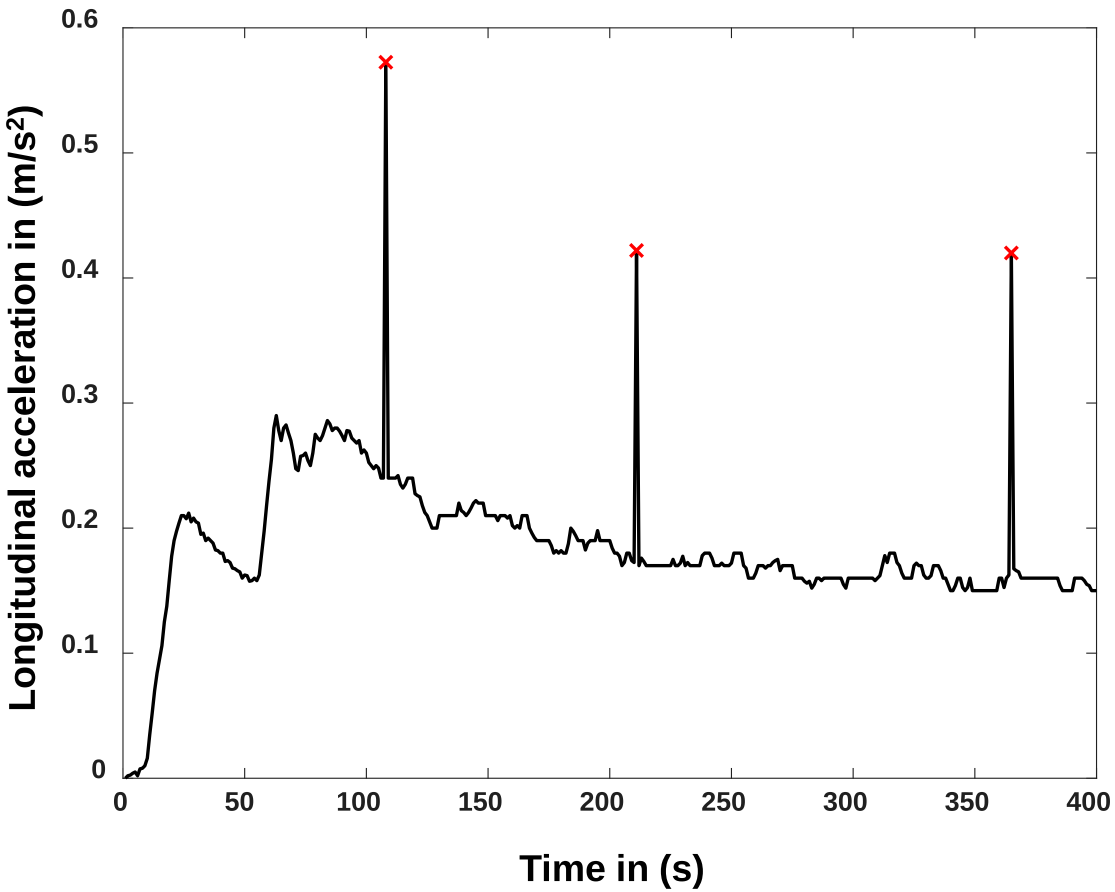
# OUTLIER



# Outliers distort the system dynamic characteristic.

- Time series can contain non-system dynamic data points.
- Assessments turn out to be difficult
- Characteristics can be distorted by outliers

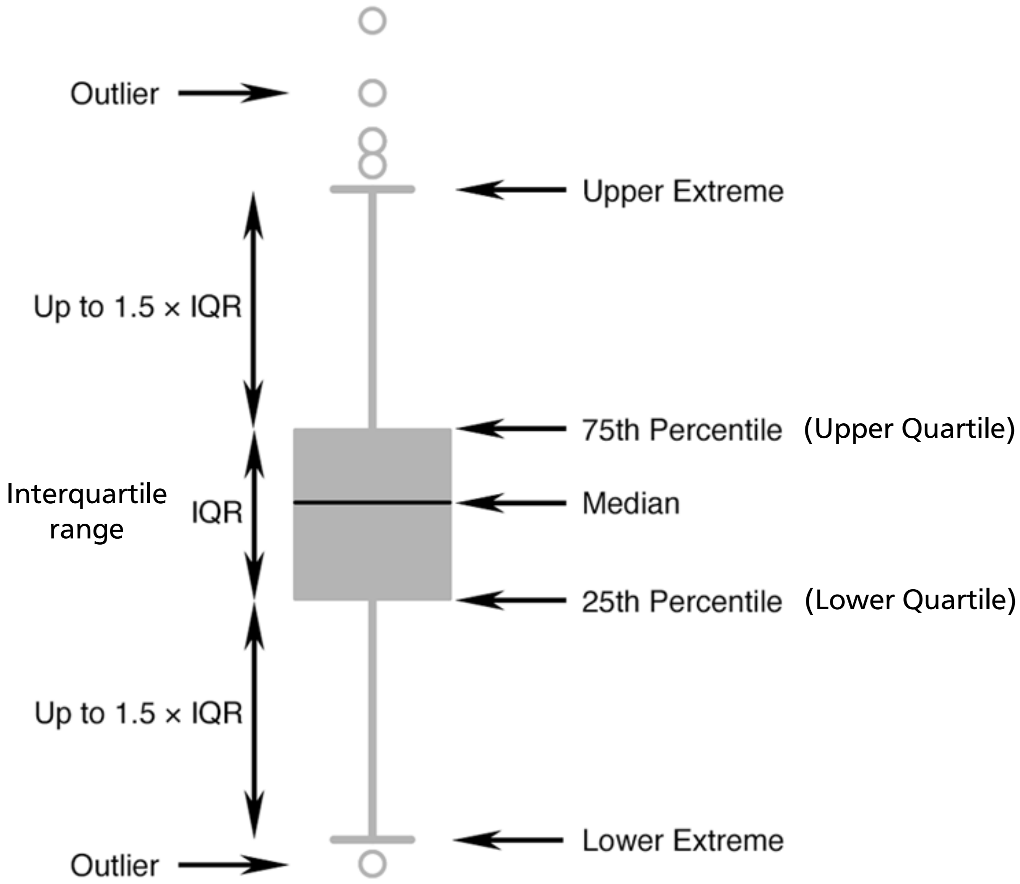


# There is no unambiguous mathematical / physical description of outliers.

- Outliers stand out clearly from the time series
- Outliers are conspicuously far away from a measure of location (e. g. mean, median, interquartile distance)
- 5-point descriptions often only serve as an informal test

**Boxplot according TUKEY**

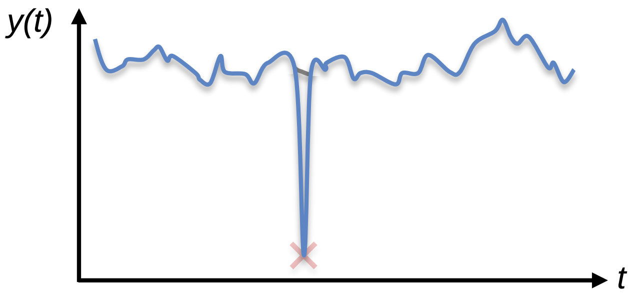



Tukey, J. W.: Exploratory Data Analysis. Pearson Publishing, Cambridge (1977) Picture Source: https://www.infragistics.com/community/blogs/b/tim_brock/posts/demystifying-box-and-whisker-plots-part-1

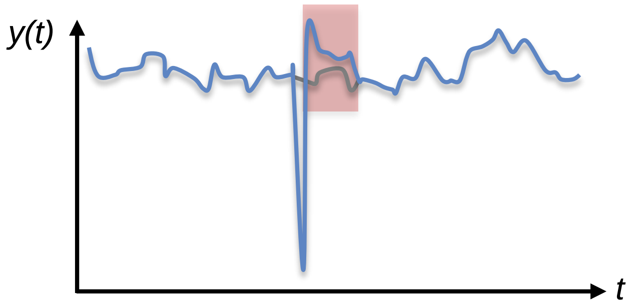
# Sorts of outliers according to Aguinis



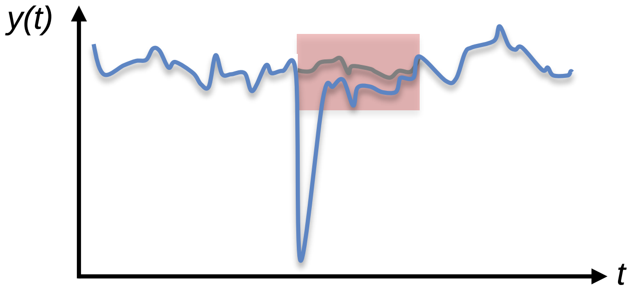
**Additive outlier**



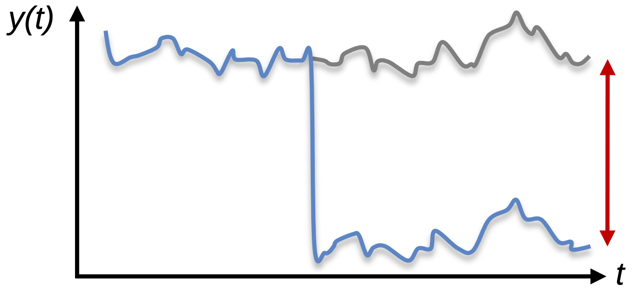

**Innovation outlier**




**Temporary outlier**




**Level shift**




*Picture according to Darné*

Source: Aguinis, H.; Gottfredson, R. K.; Joo, H.: Best-Practice Recommendations for Defining, Identifying, and Handling Outliers. In: Organizational Research Methods, Vol. 16, Iss. 2, S. 270–301 (2013)

# Examples for identification strategies

## Median / Hampel filter

- Window width and threshold (variance / standard deviation) as settings
- Distance evaluation according to

$$|x_i - \tilde{x}_i| > n_\sigma \sigma_i$$

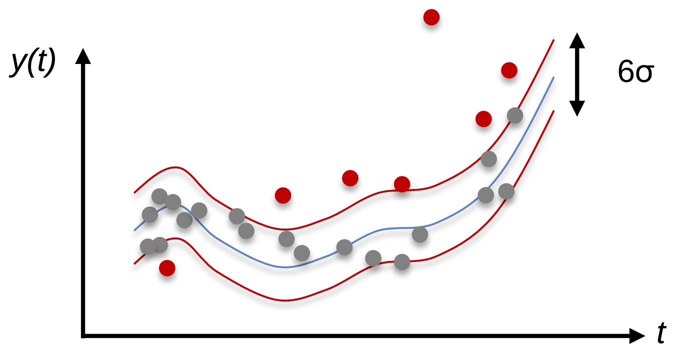
## DFFITS

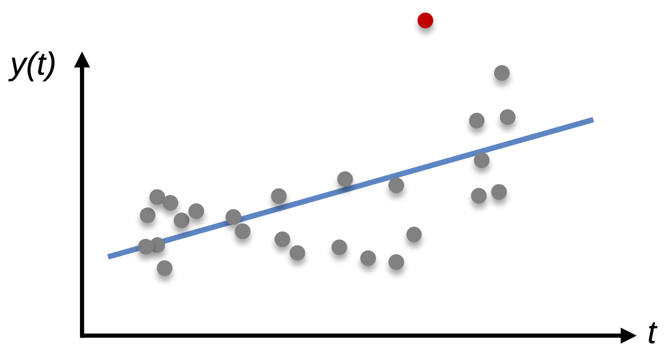
- Identification of influential data points on a regression model
- Outlier dimension with

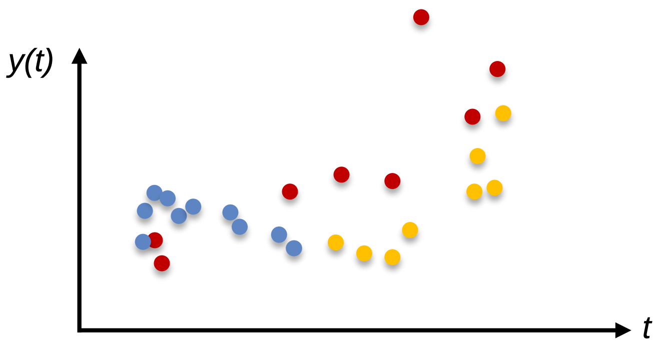
$$DFFITS_i = \frac{y - y_{(i)}}{\sigma_{(i)}\sqrt{h_{ii}}}$$

## k-means

- E.g. calculation of first-degree differences
- Initialization of k cluster centers
- Iterative assignment of data points to clusters based on distances of the first-degree differences
- Distance dimension between cluster centers as quality functional

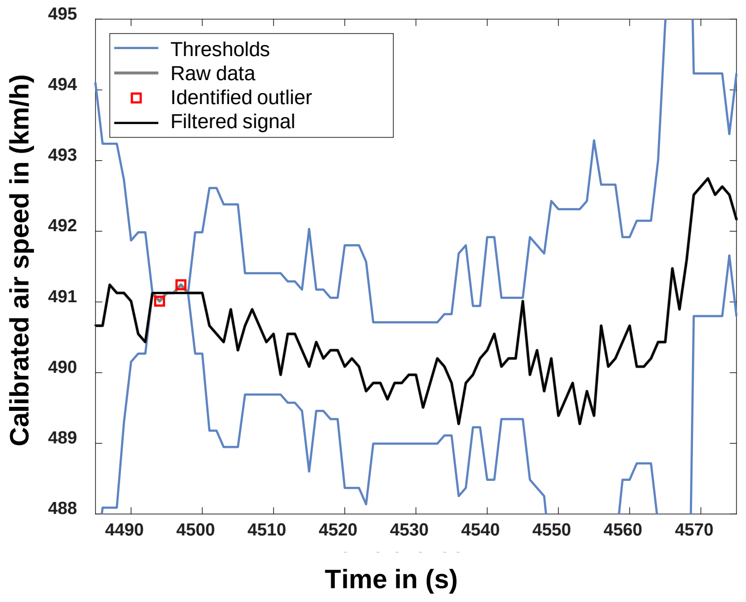


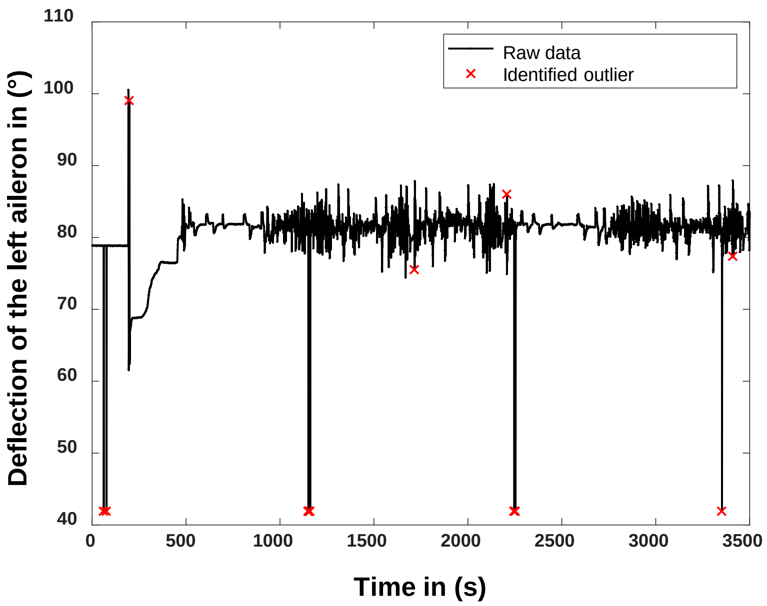



# The assessment of system dynamics and outlier identification in time series is not trivial.

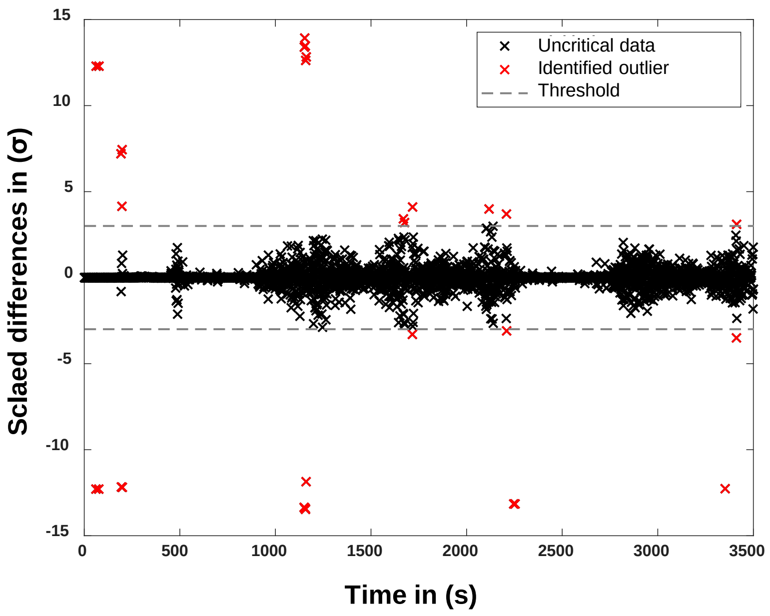
## Sensitivity of the Hampel filter



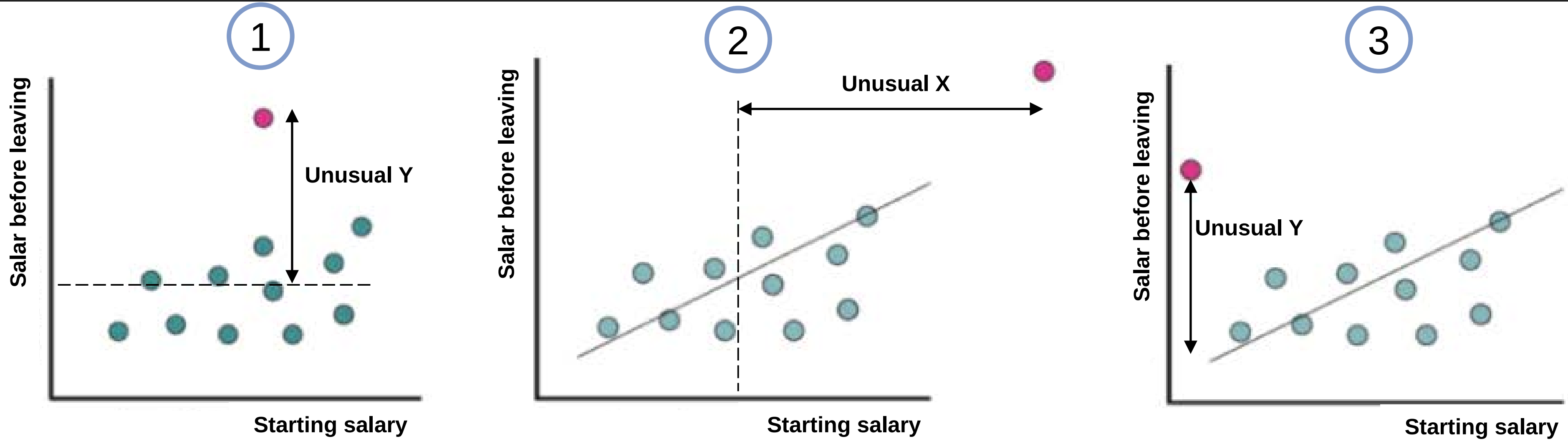

## Identified outlier of a hybrid approach




## Identified outlier in the differences



# Effects of outliers on statistical tests of more verifiable data.




**Impacts on**

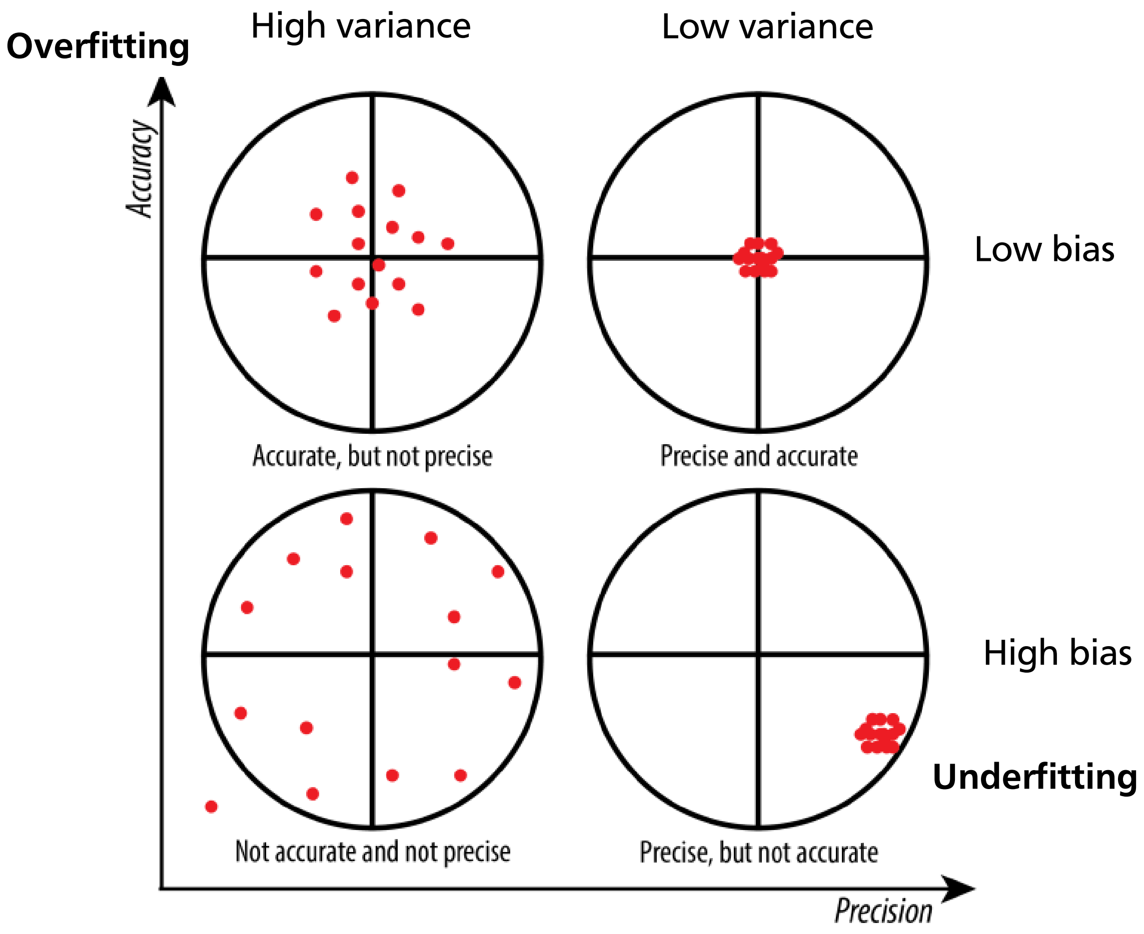
- **Slope (Effect Size):** Type 3 outliers have a large impact on the estimated effect size
- **Standard Error of Regression**: Type 1 and 3 outliers increase the standard error of regression significantly
- **P-Value**: Type 1 outliers increase the p-value, 2 reduces it in a misleading way, and 3 has a wild and unpredictable effect on the p-value.

Source: https://idatassist.com/do-you-know-what-outliers-in-your-data-really-mean/

# INFLUENCES OF INSUFFICIENT DATA QUALITY

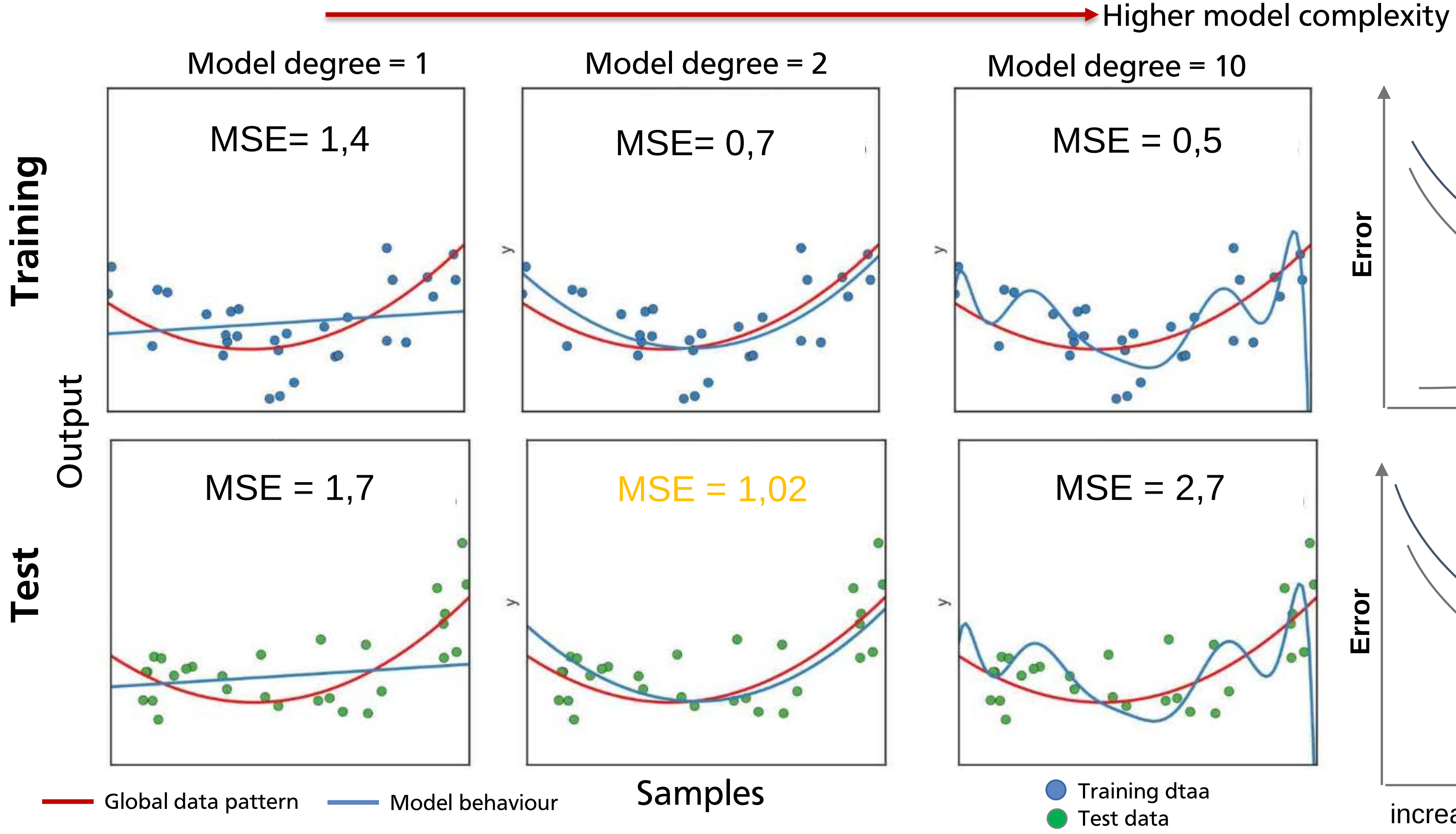
# Bias-Variance-Trade-Off

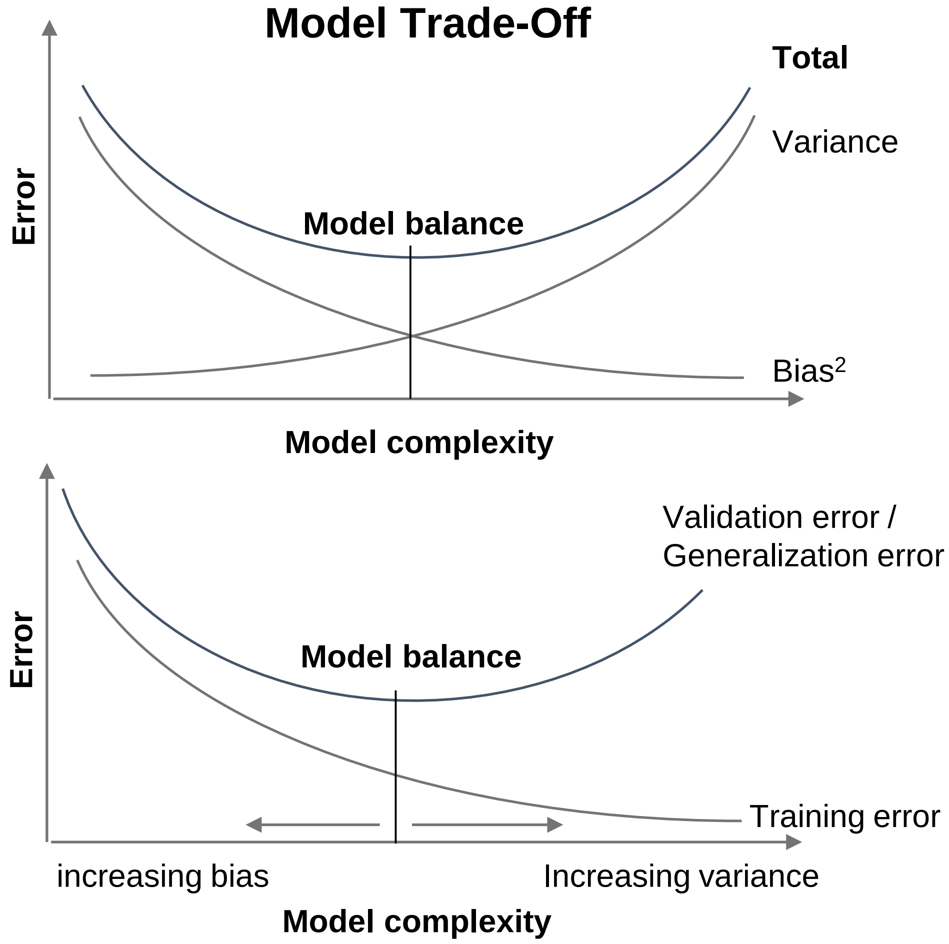
- **Underfitting**
  Model is unable to capture the global behavior or pattern of the data. Possible causes: less amount of data, linear modeling with nonlinear data.
- **Overfitting**
  Model complexity approaches the complexity of training data, e. g. captures noise in data.
- **Bias-Variance-Trade-Off**
  mean approximation of the data
  - simple models: bias to generalized data behavior
  - complex models: Variance increases on test data
- **Generalizability**
  A model is over adjusted by agent A, if agent A* describes the training data worse with a larger error but the overall distribution of the data with a smaller error better than A.




Source: according to https://wp.stolaf.edu/it/gis-precision-accuracy/

# Achievement of generalizability

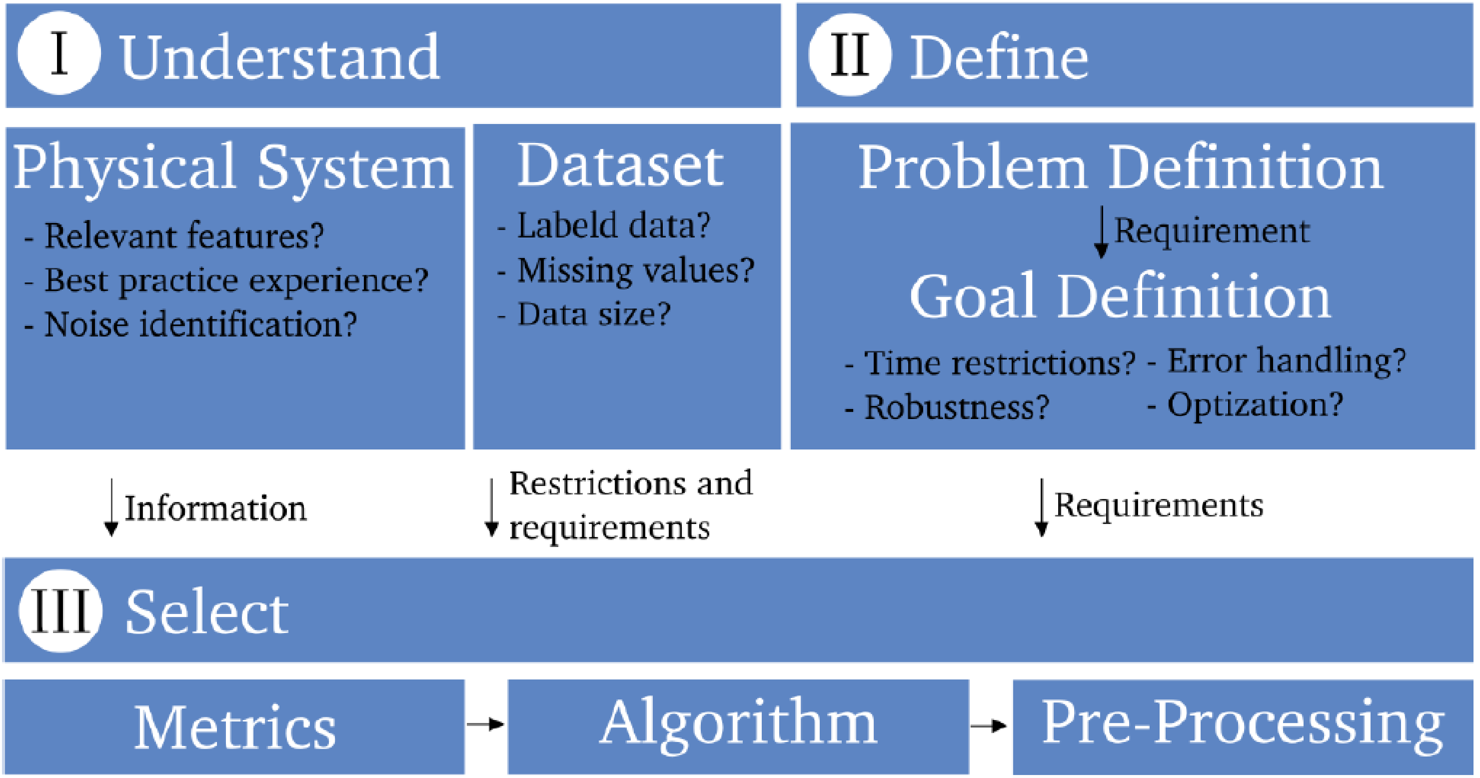



Source: according to https://neeravbasant.wordpress.com/tag/bias-variance-trade-off/

# DATA UNDERSTANDING

# Suggestion for an approach to select algorithms and strategies



# Statistical data understanding and analysis techniques

| Descriptive statistics | Explorative statistics | Inductive statistics |
|---|---|---|
| Characterisation of the data through metrics and graphics | Search for anomalies in the data and development of hypotheses | Formulation of statements that can be statistically evaluated beyond the data set |
| **Examples:**<br>▪ Locational and variance metrics<br>▪ Graphical methods | **Examples:**<br>▪ Relational metrics<br>▪ Graphical methods | **Examples:**<br>▪ Statistical models<br>▪ Significance tests |

# Statistical data understanding and analysis

Methods

TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.

## Descriptive methods

### Locational metrics

- Arithmetic mean
- Median
- Mode
- Quantile
- Minimum, maximum

### Variance metrics

- Empirical variance
- Empirical standard deviation
- Mean absolut deviation
- Range
- Inter-quartil-distance
- Variation coefficient

### Graphical methods

- Frequency table
- Frequency distribution
- Histogram
- Empirical distribution function

## Explorative methods

### Graphical methods

- Boxplot
- Q-Q-Plot
- P-P-Plot
- Violinplots
- Scatterplot
- Covariance
- Covariance coefficients

### Correlation analysis

- Bravais and Pearson
- Fechner
- Spearman
- Kendall

# DATA UNDERSTANDING

# EXAMPLE

# Example: Statistical Data Understanding
## Dataset provided by NASA (CMAPSS – turbofan simulation)

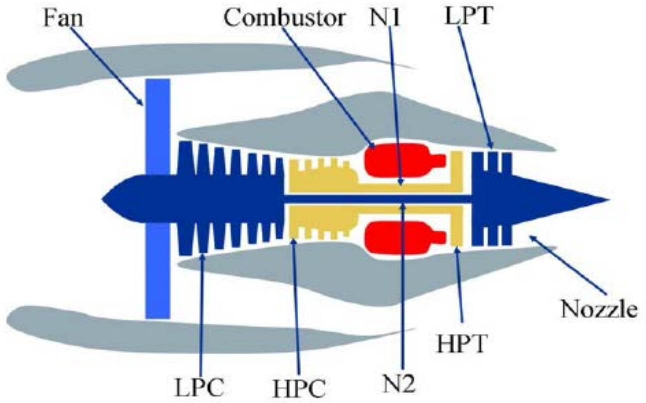
- Task:
  - Diagnosis of health index (HI)
  - Prognosis of the rest of useful lifetime (RUL)

Variables:

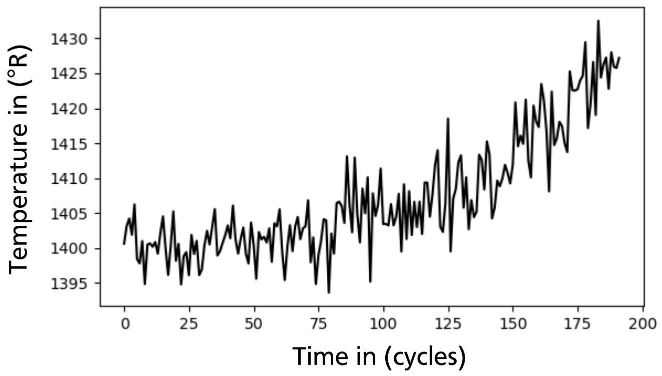
- 3 operational settings
- 21 sensor variables
- Artificial noise overlain

➢ Some features/targets are not directly measurable -> HI indicator

➢ data processing to eliminate or at least attenuate unwanted signal components



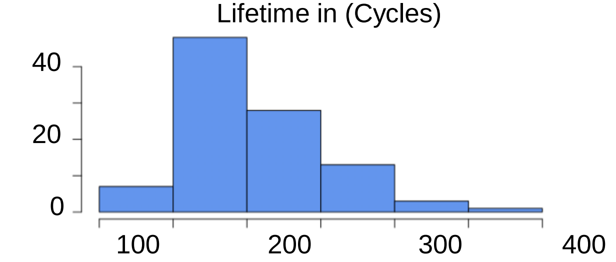

Soruce: Saxena *PHM* 2008



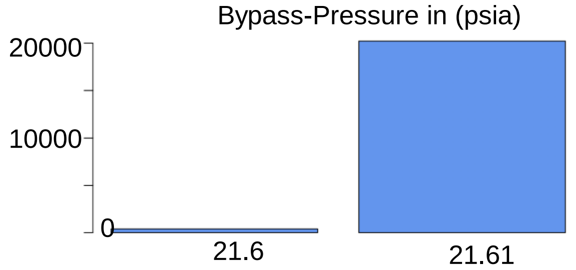

Dataset available for free under:
https://ti.arc.nasa.gov/tech/dash/groups/pcoe/prognostic-data-repository/

# Example: Statistical Data Understanding

Descriptive statistics ➡ Explorative statistics

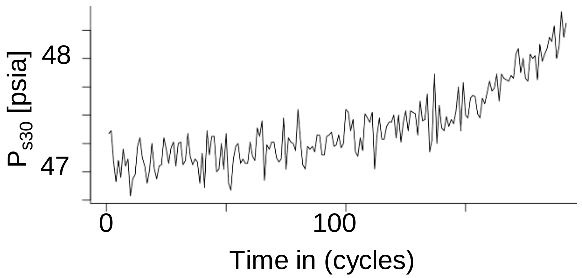



- Average Lifetime: 205 cycles (standard deviation 46,3)
- Range: 127 - 361




10 variables without relation to health status;

- constant, or discrete characteristics
- only varying through sensor noise

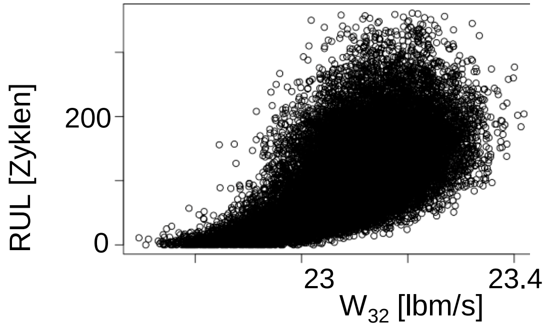


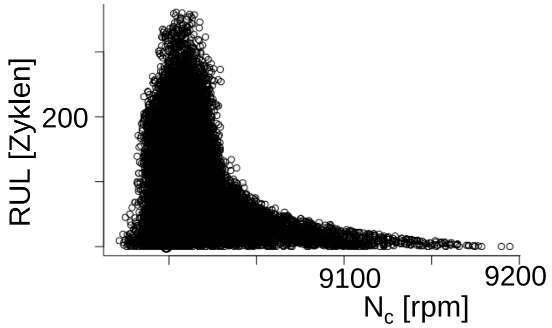

- Different locational and variance metrics at the beginning and end of operations
- Characteristic trend over time

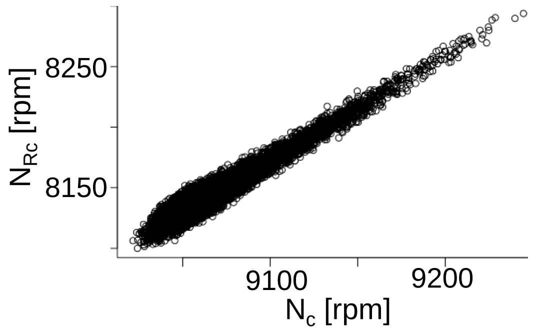
➢ **12 variables with consistent trend over time suitable for RUL prognosis**

# Example: Statistical Data Understanding

Descriptive statistics ➡ Explorative statistics




**Correlation analysis:**

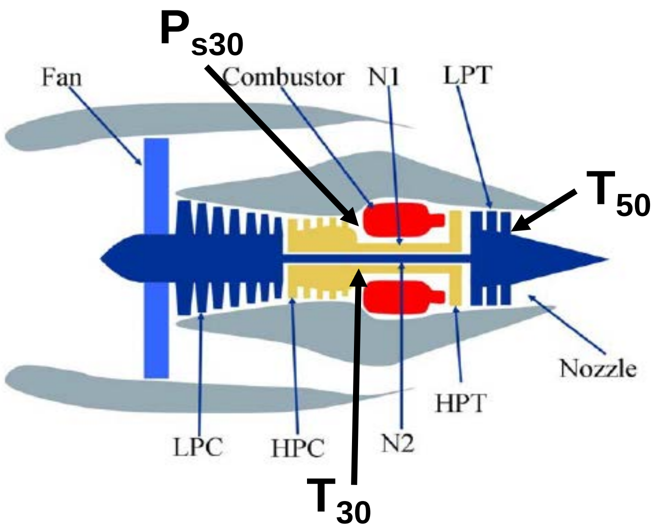
Varying correlations with RUL

- low importance of the operational variables
- medium correlation for the majority of the sensors
- low correlations for sensors which are constant or show inconsitent trend

Partly high correlations underneath the sensors → redundancies

- ➢ **Different importance and relevance for prognosis**
- ➢ **Potential of feature reduction through dependences**

# Physical interpretation (based on statistical analysis)




**Fault in high pressure compressor (HPC)**

↓

$P_{s30}$: Static pressure a HPC exit ↗

↓

Health index/status ↘

Change of other components and sensors

↓

**RUL prognosis** ↘

$T_{30}$ ↗ | $T_{50}$ ↗ | $\varphi$ ↘

Correlations can help to evaluate redundancies.

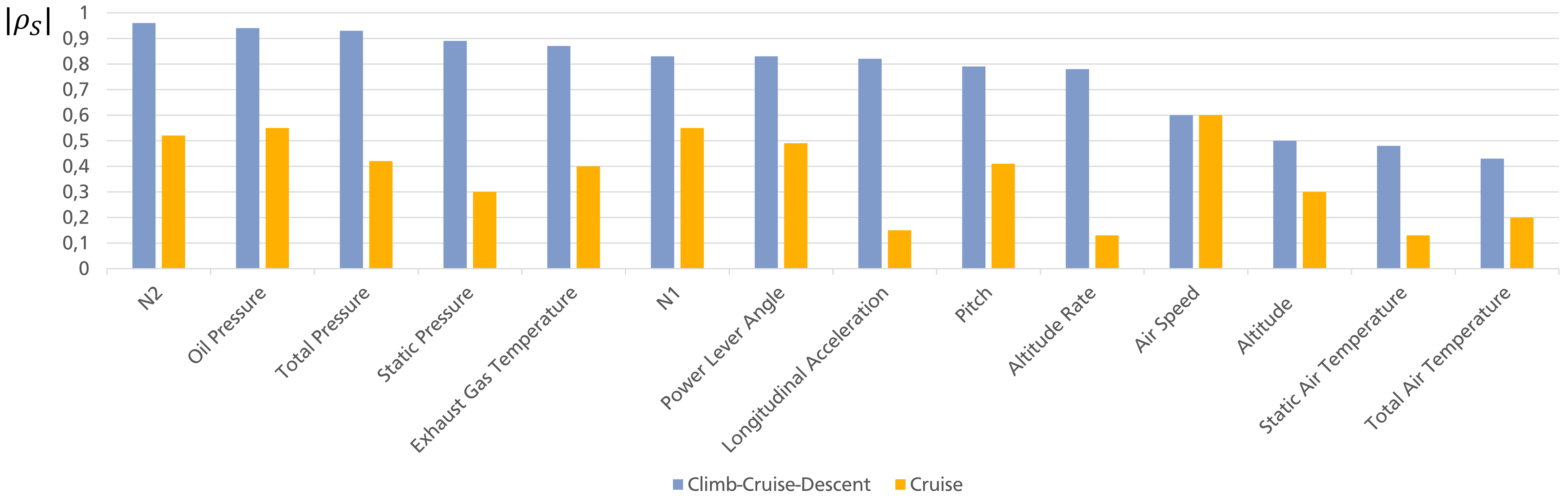
However, interpretations may differ from physical insights.

# CORRELATION

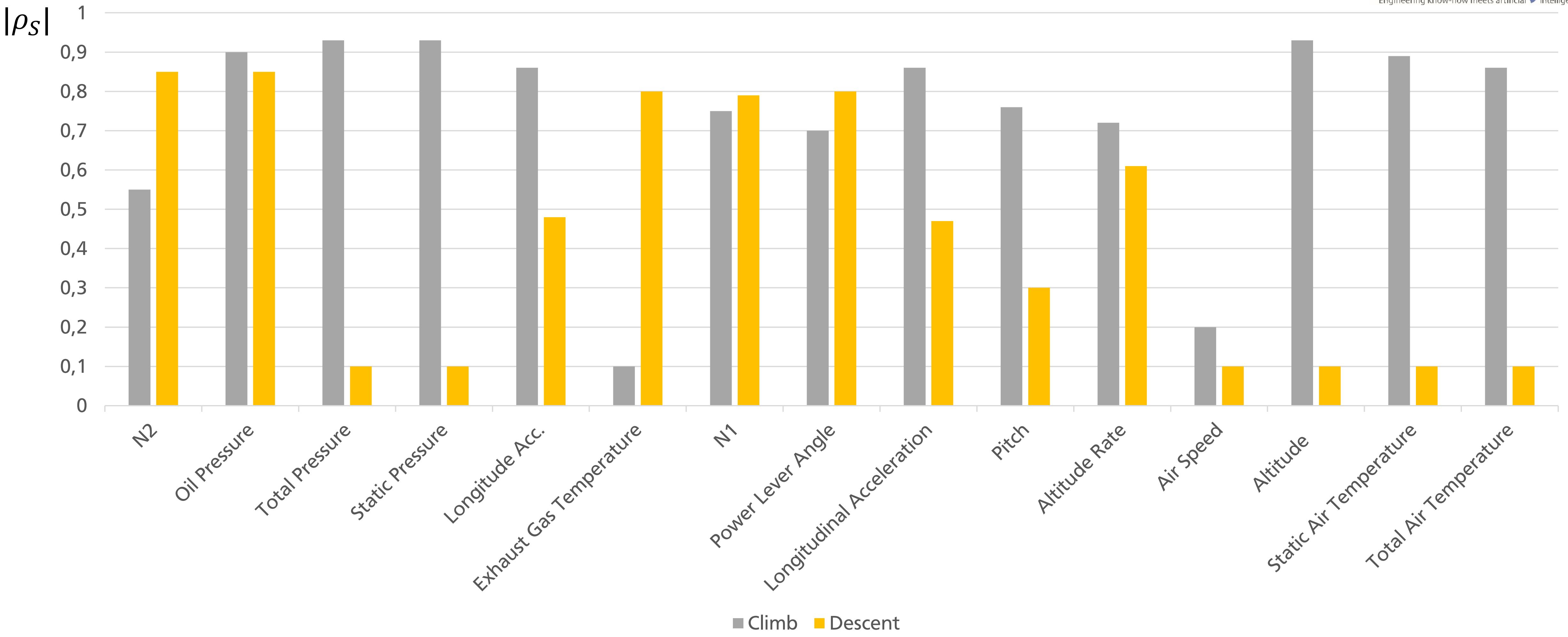
# Correlations with the fuel flow of an aircraft in different flight phases.

SPEARMAN correlation coefficient:
(Testing for monotony)

$$\rho_s = \frac{\operatorname{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_X \cdot \sigma_Y}$$

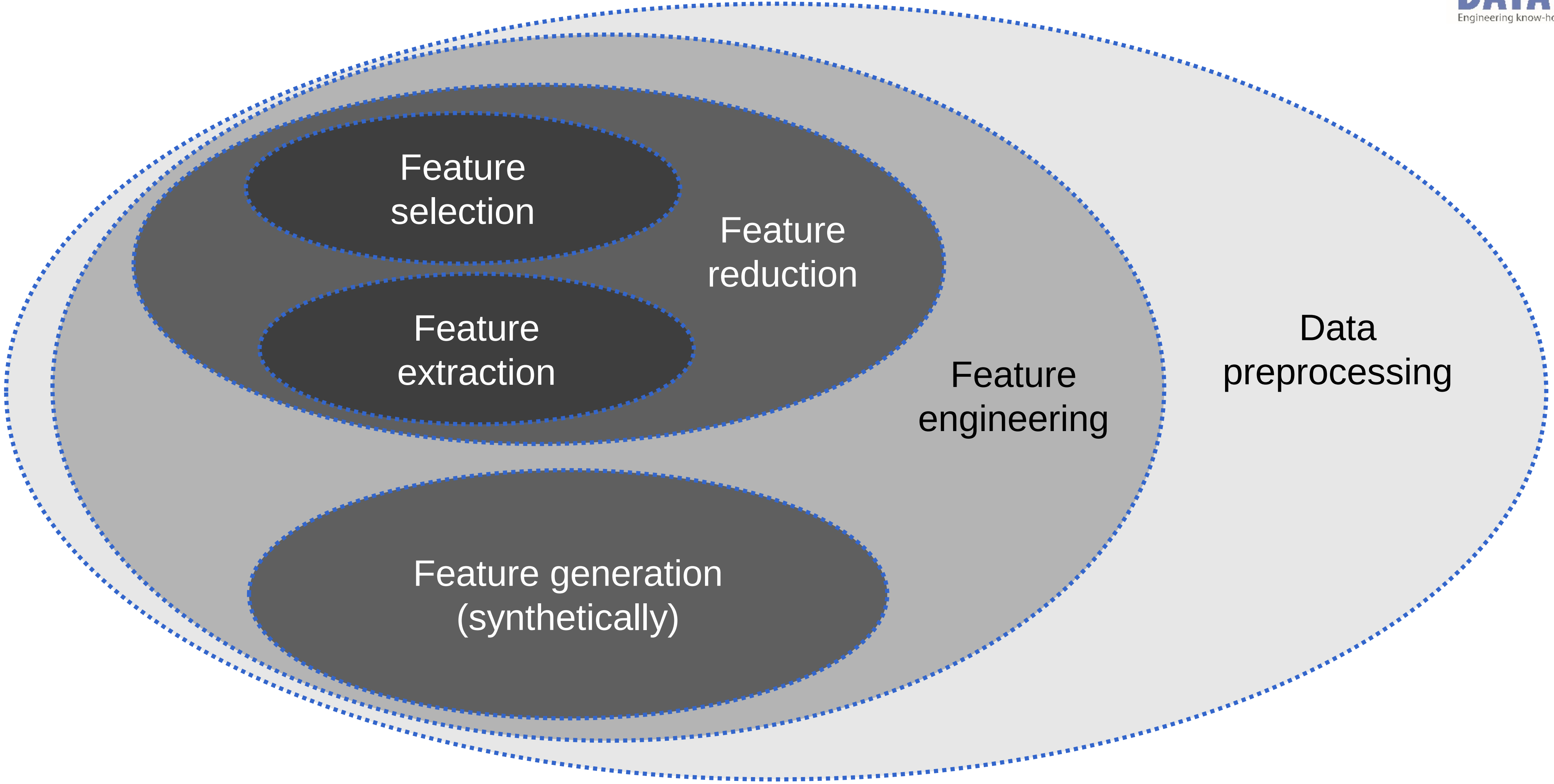


# Correlations with the fuel flow of an aircraft in different flight phases.

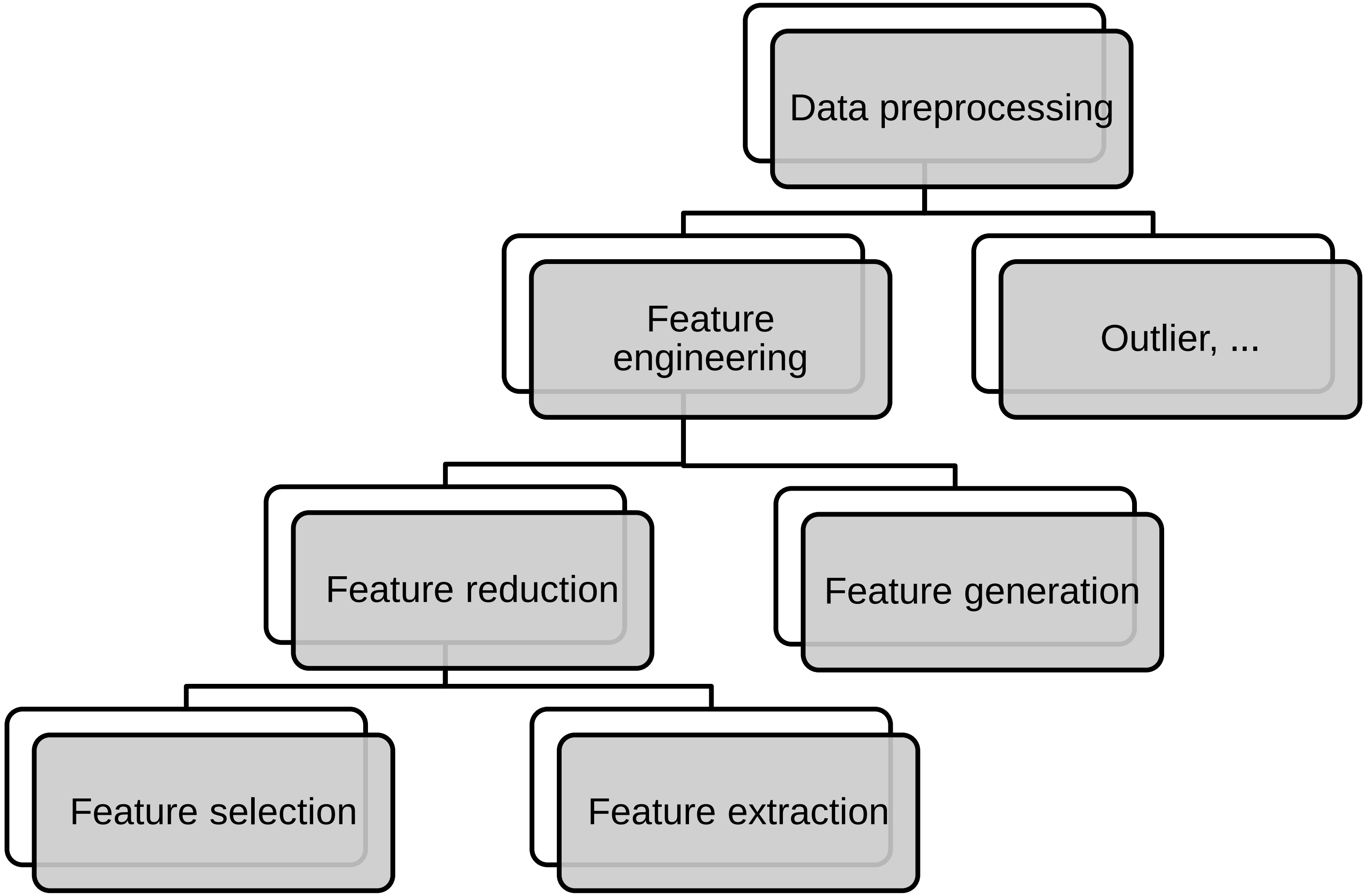


# DATA PREPROCESSING: FEATURE ENGINEERING

# Classification of feature engineering



# Dependencies in feature engineering



# Feature reduction can be devided into feature extraction and feature selection

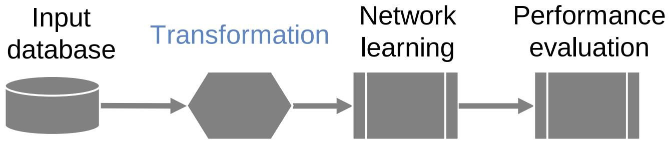
- Avoidance of multi collinearities and rudundant parameters
- Better generalizability
- Evalutaion of reduction methods through model performance/quality

**Feature reduction**

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

**Feature selection**
- Wrapper
- Filter
- Embedded

# Feature Extraction
# PCA – Principal Component Analysis




TECHNISCHE UNIVERSITÄT DARMSTADT

DATAbility.ai
Engineering know-how meets artificial intelligence.

## Definition: *Source:* Jolliffe I.T. *Principal Component Analysis* (1986)

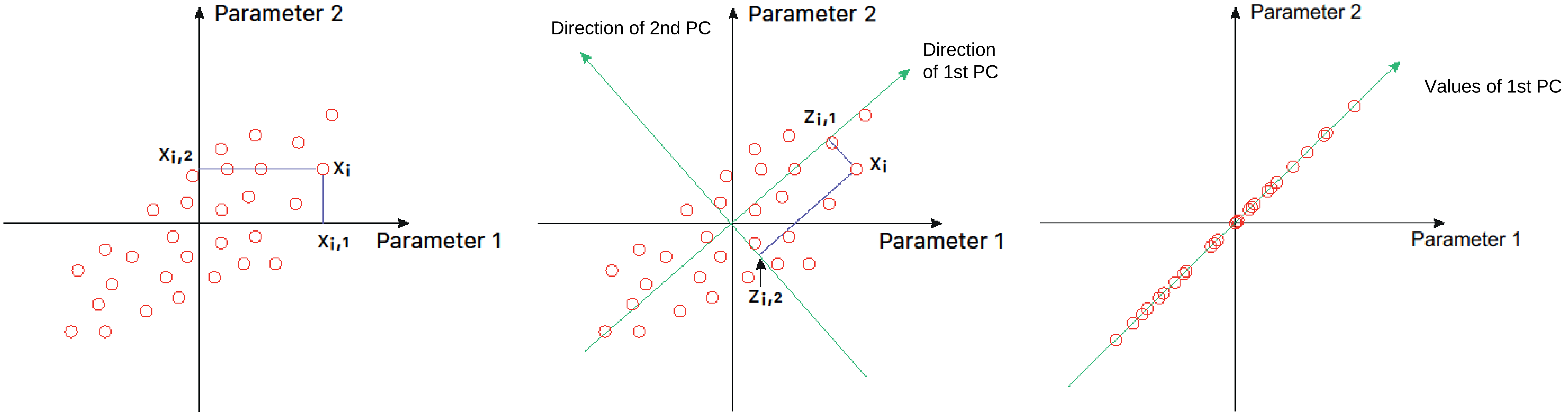
For a set of *d*-dimensional vectors $\{\mathbf{t}_\mathrm{n}\}, n \in \{1 \dots N\}$, the *q* principal axes $\mathbf{w}_\mathrm{j}, \mathrm{j} \in \{1 \dots q\}$, are those orthonormal axes onto which the retained variance under projection is maximal.
The eigenvectors $\mathbf{w}_\mathrm{j}$ are given by the *q* dominant eigenvectors of the sample covariance matrix $\mathbf{S} = \sum_\mathrm{n}(\boldsymbol{t}_n - \bar{\boldsymbol{t}})(\boldsymbol{t}_n - \bar{\boldsymbol{t}})^T/N$ such that $\mathbf{S}\mathbf{w}_\mathrm{j} = \lambda_j \boldsymbol{w}_j$ and where $\bar{\boldsymbol{t}}$ is the sample mean. The vector $\boldsymbol{x}_n = \boldsymbol{W}^T(\boldsymbol{t}_n - \bar{\boldsymbol{t}})$, where $\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_q)$, is thus a *q*-dimensional reduced representation of the observed vector $\mathbf{t}_\mathrm{n}$.

Source: http://www.analytik.ethz.ch/vorlesungen/chemometrie/2_PCA_Monitor.pdf

## Summary

- PCA is an orthogonal linear transformation into a new coordinate system representing the maximum variance
- It is a popular tool for dimensionalty reduction
- The first principal component explains the greatest variance
- The relevance of individual features within a principal component can be interpreted from transformation matrix
- Purpose: explorative data analysis for correlation discovering; modeling with the transformed data can eliminate irrelevant information such as noise and the risk of artifacts such as outliers

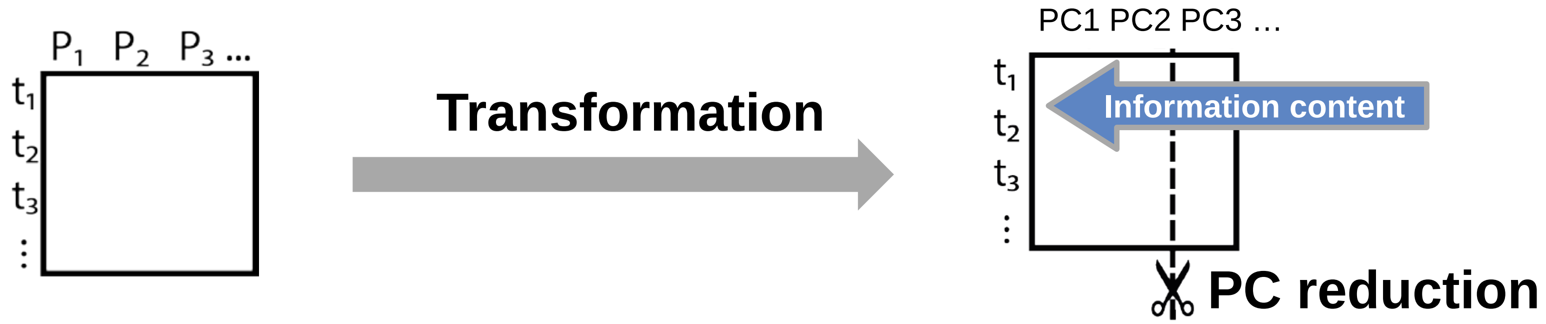
# PCA – Visual explanation (Two dimensional)

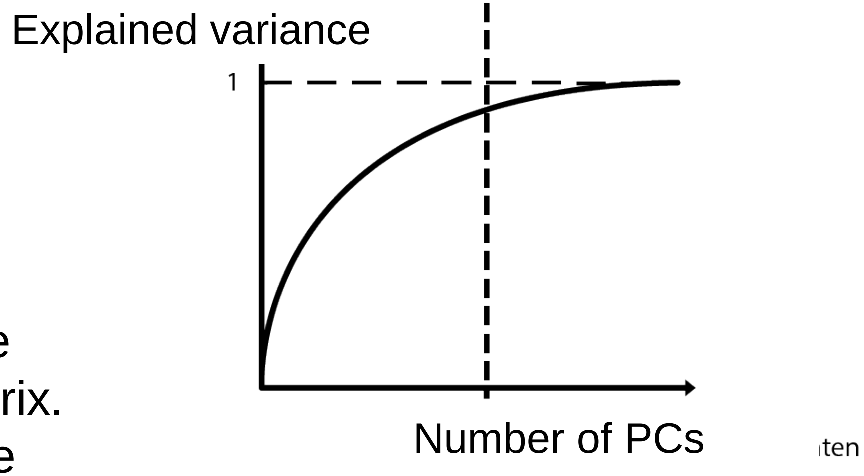


# PCA / Factor Analysis – Visual explanation

**Original data set**

**PC data set**



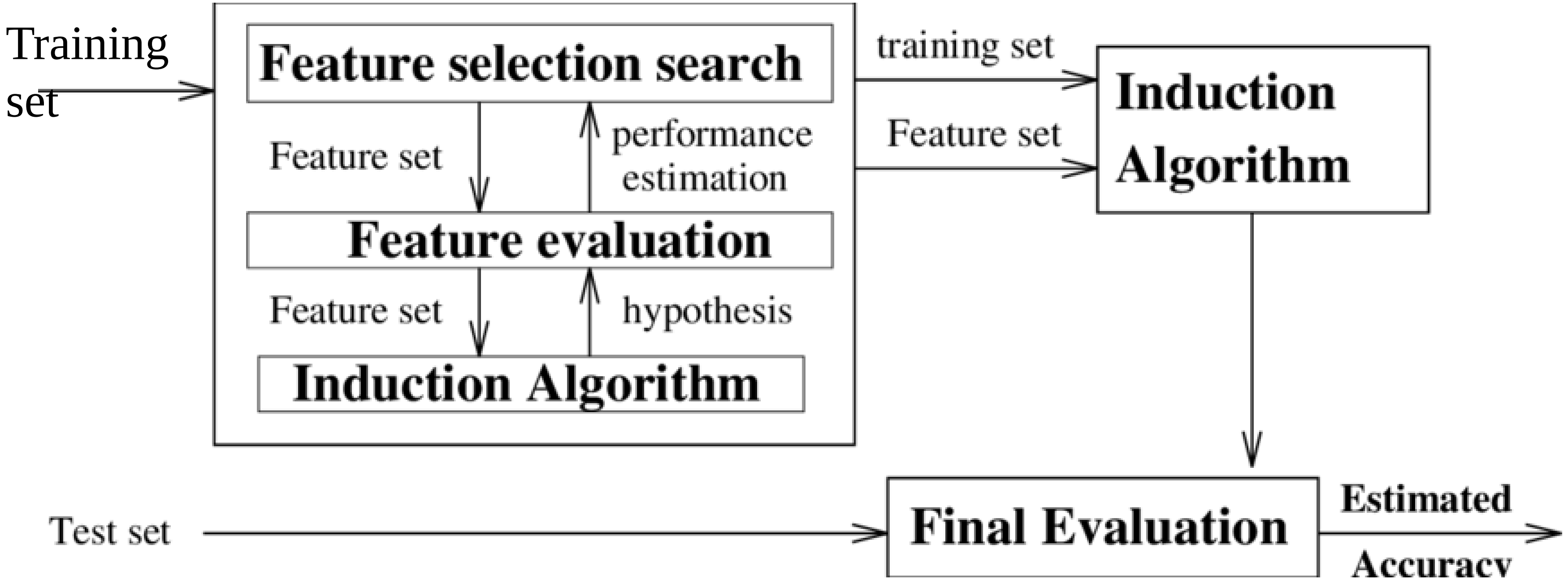




- Explained variance **α** of the total variance
- **Alternative: Kaiser-criterion** Determine the eigenvalues of the PCs from the transformation matrix. Eigenvalues greater than one are taken into account for feature (PC) selection.

**Reduction**:

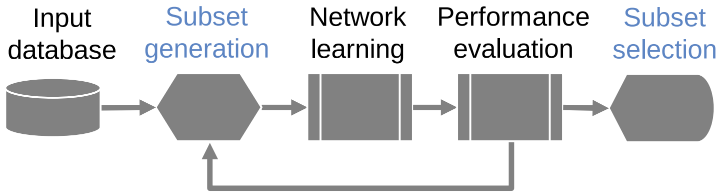
- find a feature set that describes most of the variance while using a fewer number of features
- The obtained latent variables (PC) describe the variance of the data in decreasing order
- The first PC describes most of the variance.
- In order to decide how many PCs are enough to describe the data characteristics, a threshold between 70% to 90% of the described variance needs to be reached by the PCs

# Feature Selection approach

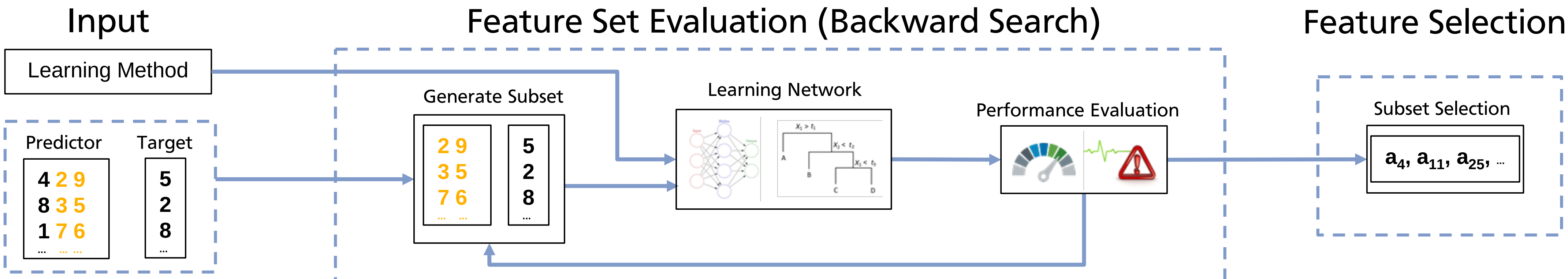



Source: Ron Kohavi: Wrappers for Performance Enhancement and Oblivious Decision Graphs. Stanford University (2015)

# Wrapper using measures for data characterization to select features.

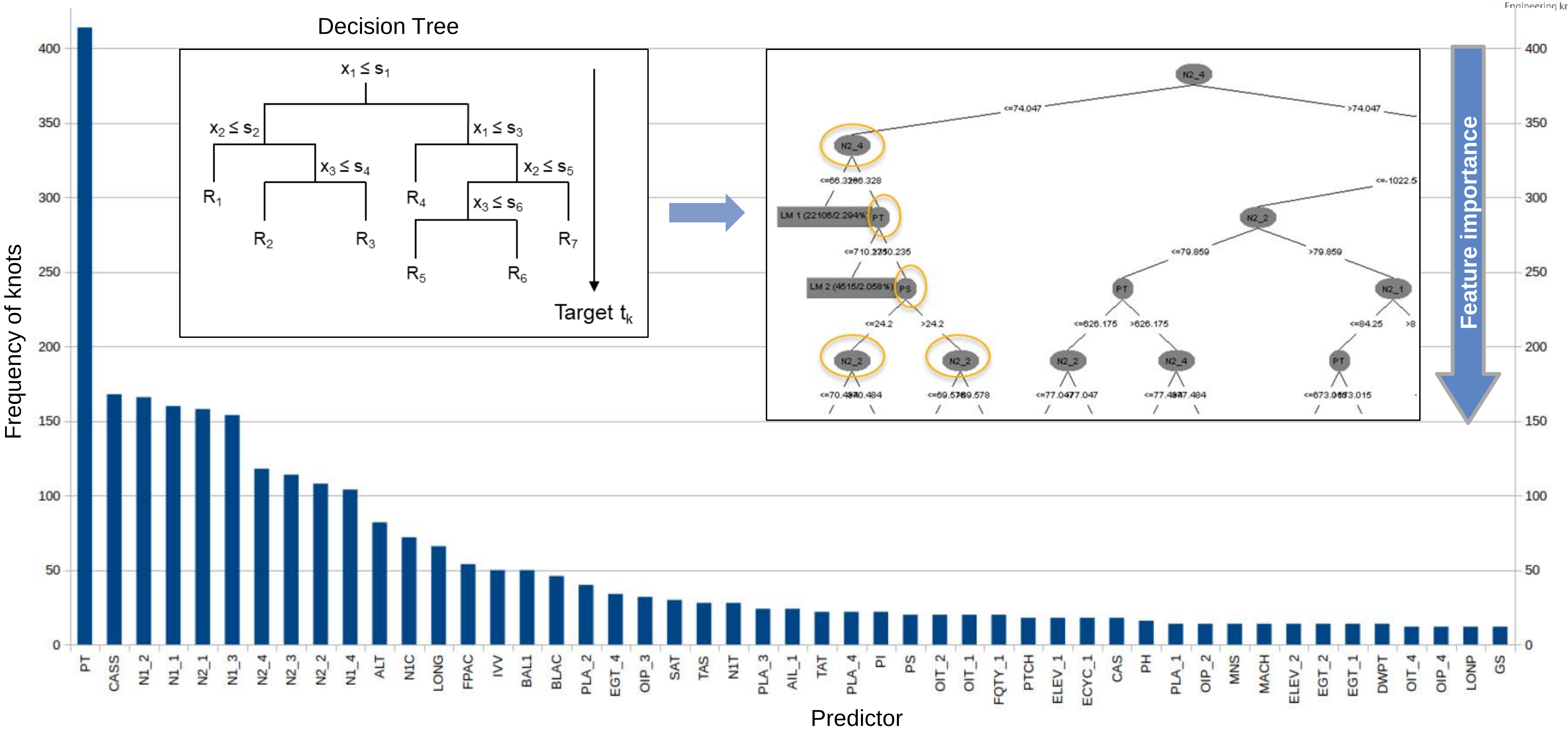



- **Forward** Search starts with an empty feature subset. The algorithm tries to add every possible attribute to its set and evaluates changes with some performance estimate and, thus, the optimal attribute will be added. The algorithm has also a parameter that allows to make some suboptimal steps, when no attributes can be added with a positive estimation. This is done with back-propagation, therefore if there are no positive tendencies further, last most optimal feature subset is recovered.
- **Backward** Search starts with a subset of all possible attributes included. Algorithm tries to remove some of them and evaluates the results iteratively.




➢ More expensive than a filter approach such as variance threshold

# Embedded: Importance of individual predictors for machine learning models via decision trees.



Data Acquisition | Data Understanding | Data Preprocessing

# **APPLICATION: HANDLING DEVICE** (see Lecture 1)

# Setup of Automatic Handling Device

- 4-DoF robotic arm that moves boxes
- No integrated sensors → no process information

→ **Does the robotic arm move a heavy or a light container?**

- Custom retrofit of robotic arm
- Low-cost sensor and microcontroller (~20 €)
- Raw data send via Wi-Fi

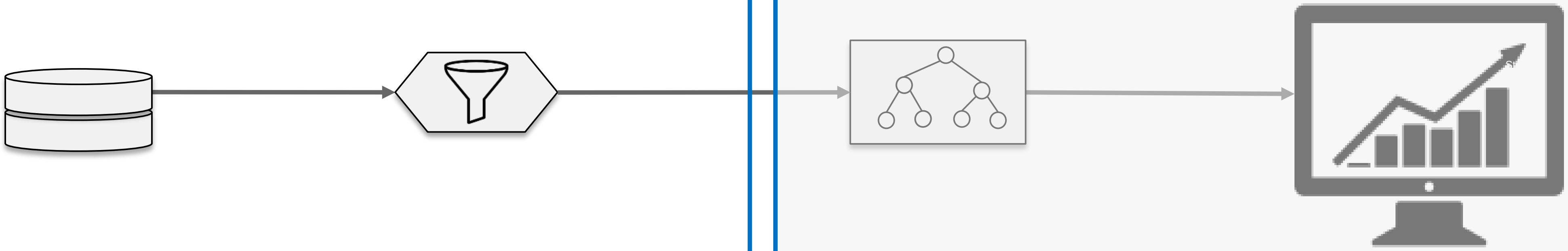
→ **Automatic data-driven classification in MATLAB**

**ESP 32**
240 MHz
520 KB RAM
12 bit ADC
GPIO pins
Wi-Fi, BT

**MPU-9250**
3-axis Gyro
3-axis Accelerometer
3-axis Magnetometer

# Application Framework

**Acquisition** | **Preprocessing** | **Modeling** | **Evaluation**



Collect and categorize
- Statistical Description
- Outlier?

→ Knowledge discovery

- Filtering
- Normalising
- Feature Extraction: FFT/PCA
- Feature Selection: Wrapper

Classification (Diagnosis)
- Methods:
  - Artificial Neural Network
  - Bayesian Network
  - Decision Tree
  - Random Forest

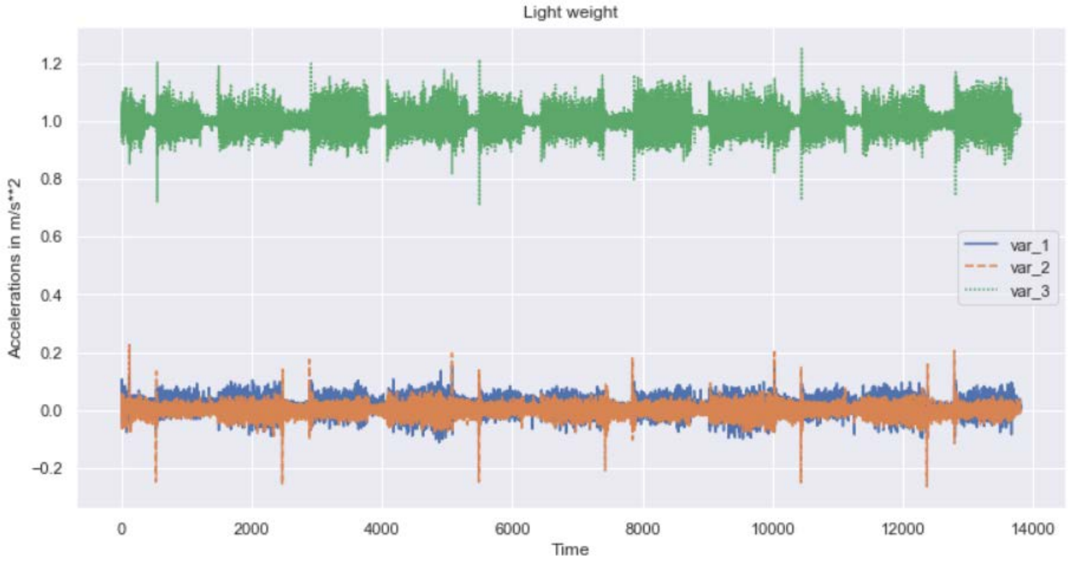
Performance metrics:
- Accuracy
- Precision/Recall

# Results: Data Understanding

## Descriptive statistics

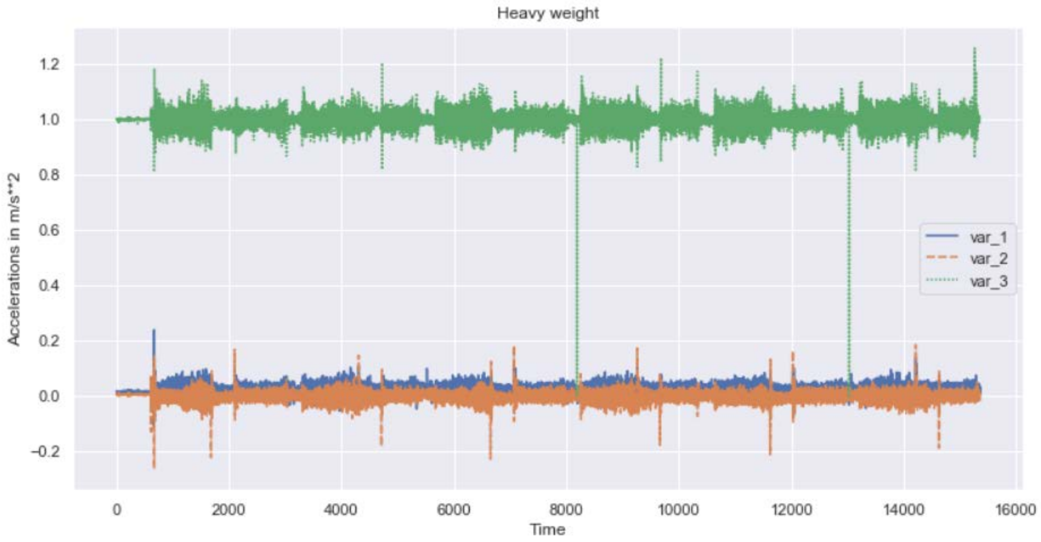
```
df_light_full.describe()
```

| | t | var_1 | var_2 | var_3 |
|---|---|---|---|---|
| count | 13806.000000 | 13806.000000 | 13806.000000 | 13806.000000 |
| mean | 400889.231711 | 0.013772 | 0.000039 | 0.999970 |
| std | 39877.035927 | 0.025820 | 0.024085 | 0.041814 |
| min | 331765.000000 | -0.112000 | -0.264000 | 0.708000 |
| 25% | 366364.500000 | -0.000000 | -0.012000 | 0.976000 |
| 50% | 400897.000000 | 0.015000 | 0.001000 | 0.999000 |
| 75% | 435409.500000 | 0.029000 | 0.013000 | 1.021000 |
| max | 469952.000000 | 0.154000 | 0.226000 | 1.250000 |



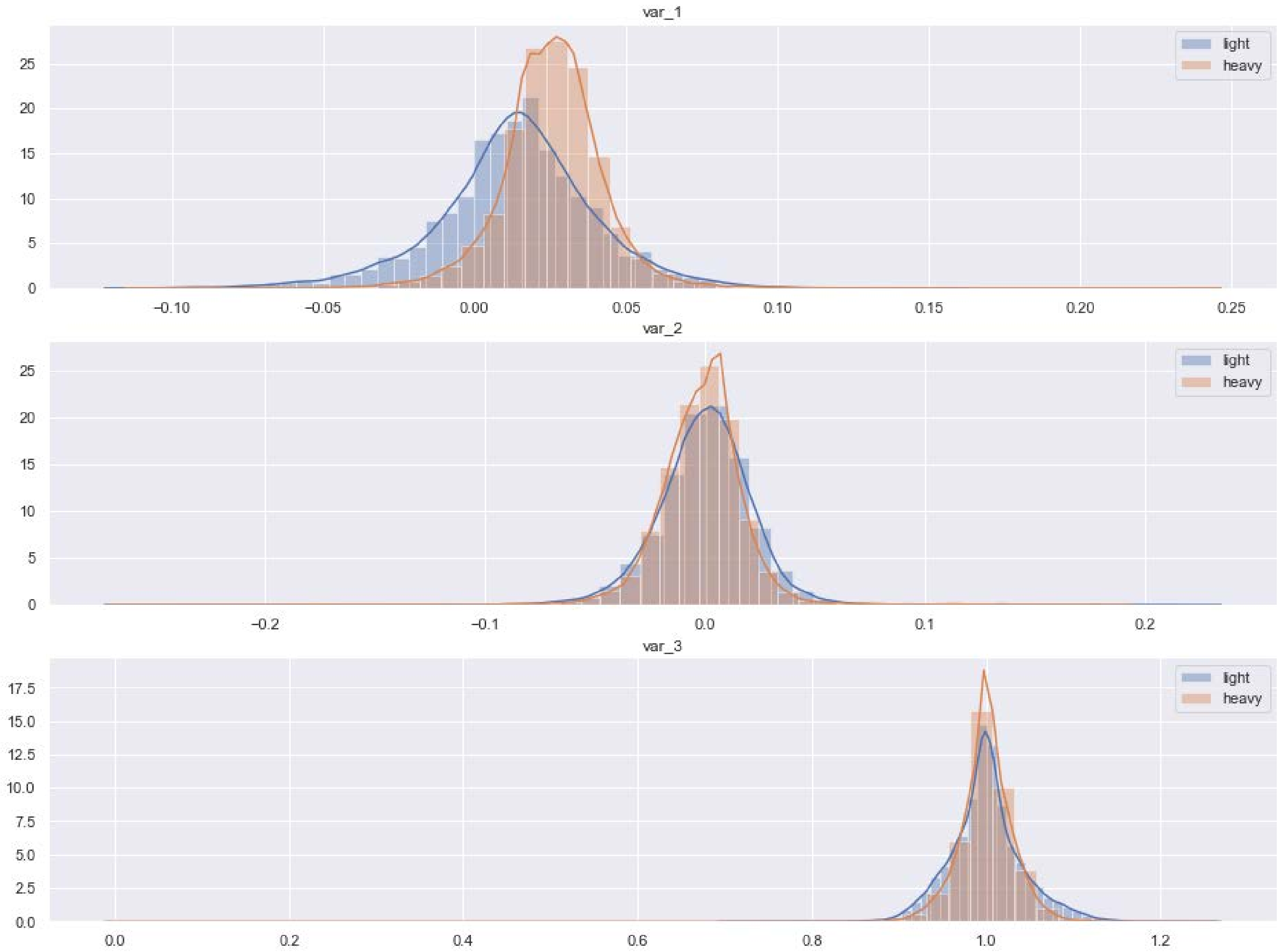

```
df_heavy_full.describe()
```

| | t | var_1 | var_2 | var_3 |
|---|---|---|---|---|
| count | 15352.000000 | 15352.000000 | 15352.000000 | 15352.000000 |
| mean | 217449.840607 | 0.025510 | -0.001491 | 1.000084 |
| std | 44658.663301 | 0.016597 | 0.020733 | 0.032306 |
| min | 140254.000000 | -0.109000 | -0.263000 | -0.000000 |
| 25% | 178677.500000 | 0.016000 | -0.012000 | 0.984000 |
| 50% | 217437.000000 | 0.026000 | -0.000000 | 1.000000 |
| 75% | 256122.500000 | 0.035000 | 0.009000 | 1.017000 |
| max | 294706.000000 | 0.240000 | 0.186000 | 1.258000 |



# Results: Data Understanding

## Explorative statistics



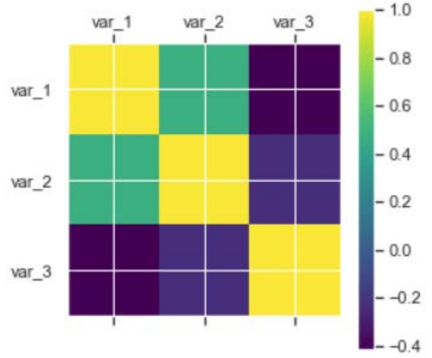

Correlations between accelartions

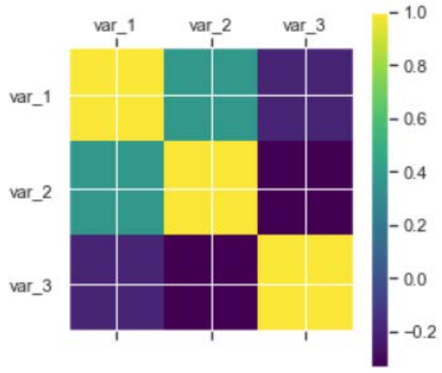
`df_light.corr()`

| | var_1 | var_2 | var_3 |
|---|---|---|---|
| var_1 | 1.000000 | 0.476959 | -0.415627 |
| var_2 | 0.476959 | 1.000000 | -0.224667 |
| var_3 | -0.415627 | -0.224667 | 1.000000 |

`df_heavy.corr()`

| | var_1 | var_2 | var_3 |
|---|---|---|---|
| var_1 | 1.000000 | 0.381186 | -0.194681 |
| var_2 | 0.381186 | 1.000000 | -0.330959 |
| var_3 | -0.194681 | -0.330959 | 1.000000 |

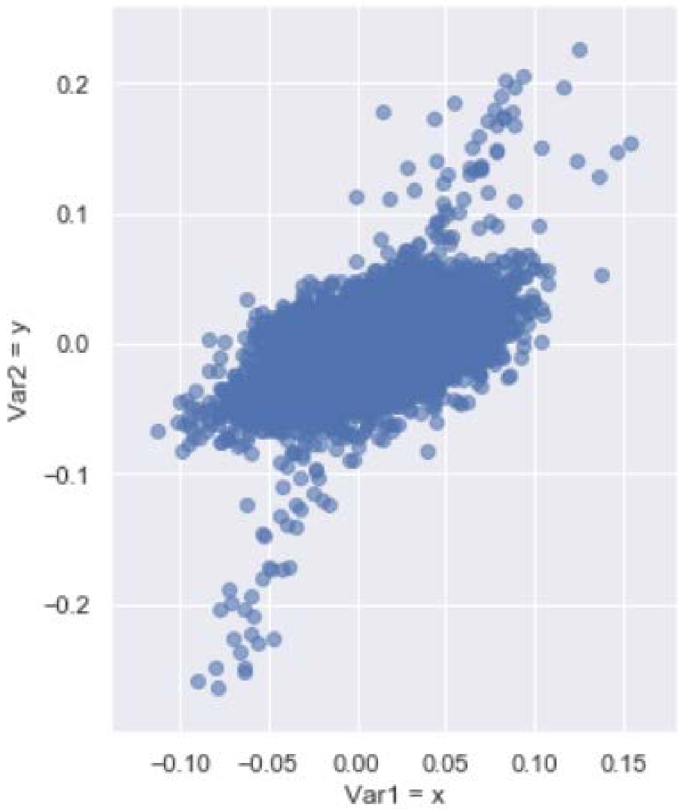


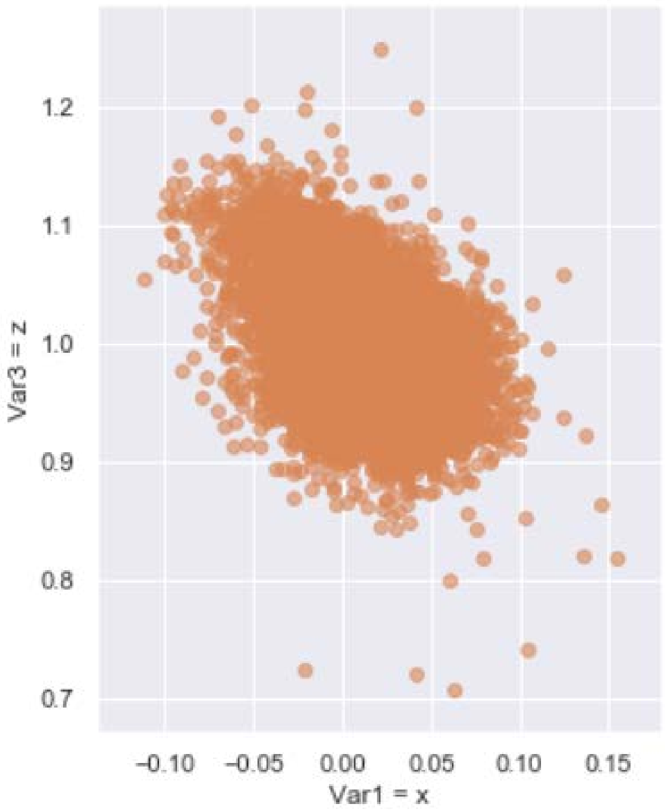



# Results: Data Understanding

## Explorative statistics

### Plotting accelerations against each other

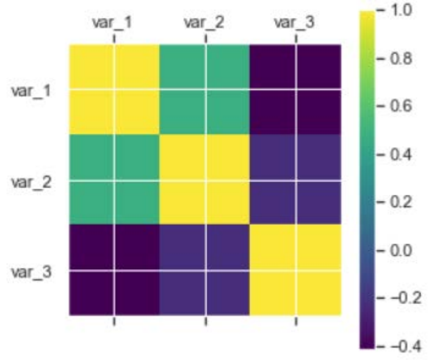



### Correlations between accelartions

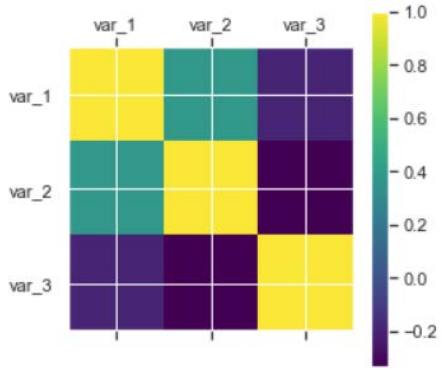
```
df_light.corr()
```

| | var_1 | var_2 | var_3 |
|---|---|---|---|
| var_1 | 1.000000 | 0.476959 | -0.415627 |
| var_2 | 0.476959 | 1.000000 | -0.224667 |
| var_3 | -0.415627 | -0.224667 | 1.000000 |



```
df_heavy.corr()
```

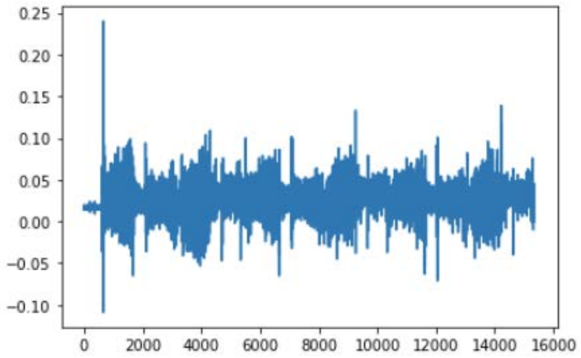
| | var_1 | var_2 | var_3 |
|---|---|---|---|
| var_1 | 1.000000 | 0.381186 | -0.194681 |
| var_2 | 0.381186 | 1.000000 | -0.330959 |
| var_3 | -0.194681 | -0.330959 | 1.000000 |

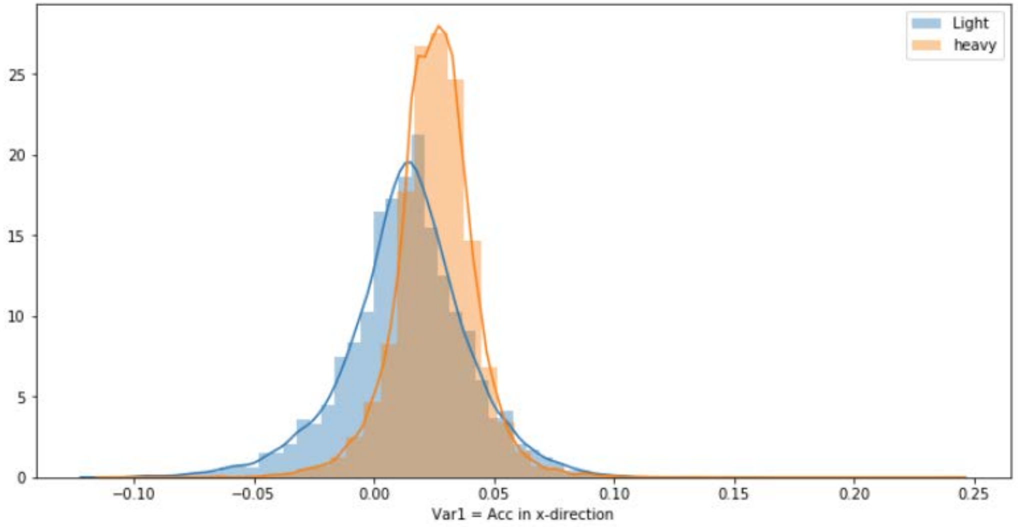


# Results: Data Preprocessing

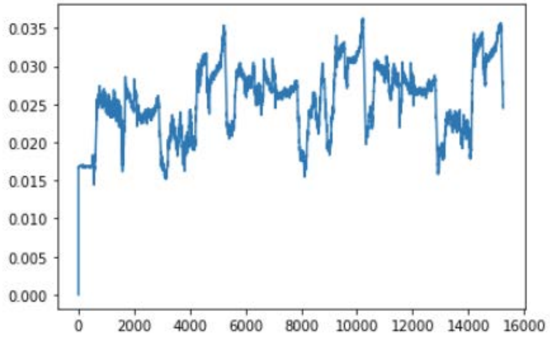
## Moving mean (window size = 100)

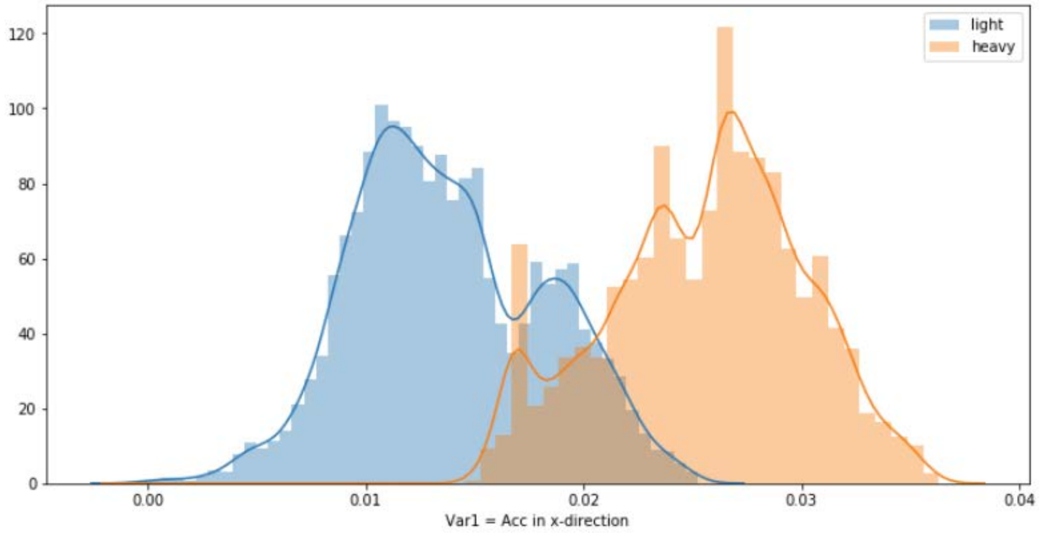
Not preprocessed




Moving mean preprocessed

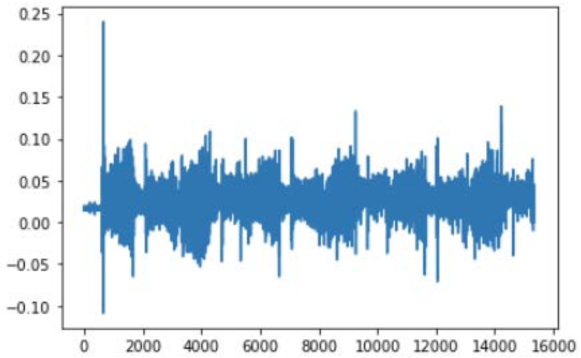


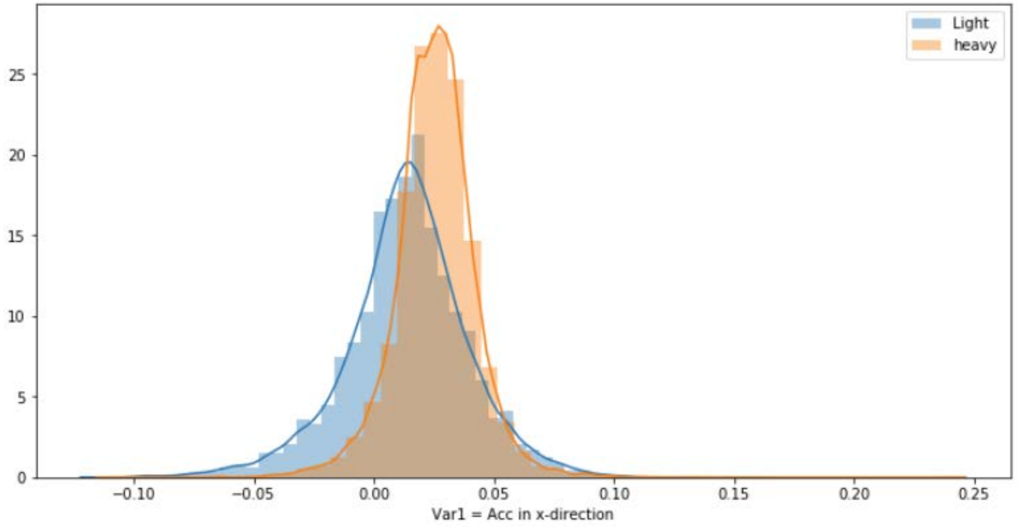



# Results: Data Preprocessing

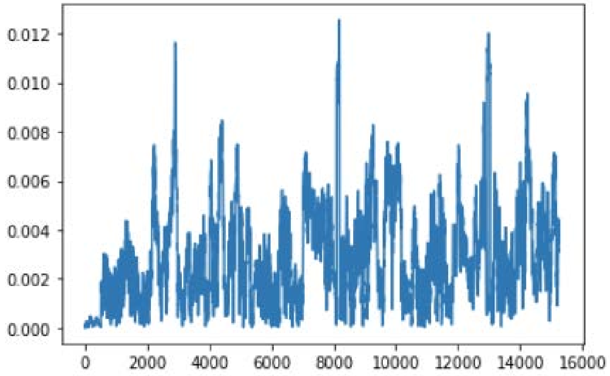
## Wrapper based on FFT (150 features!)

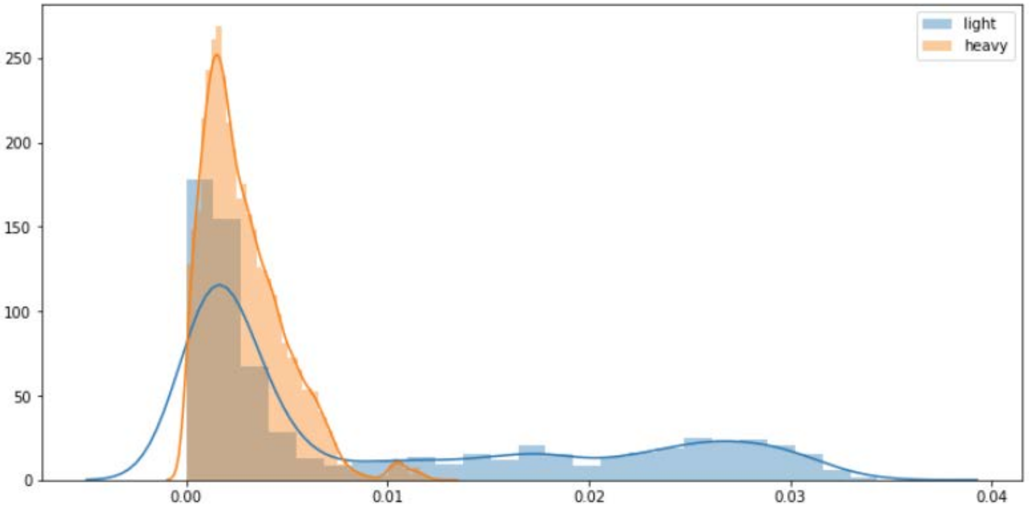
Not preprocessed




Wrapper: Feature No. 147

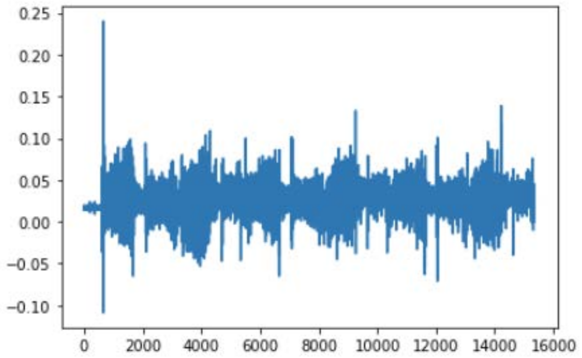


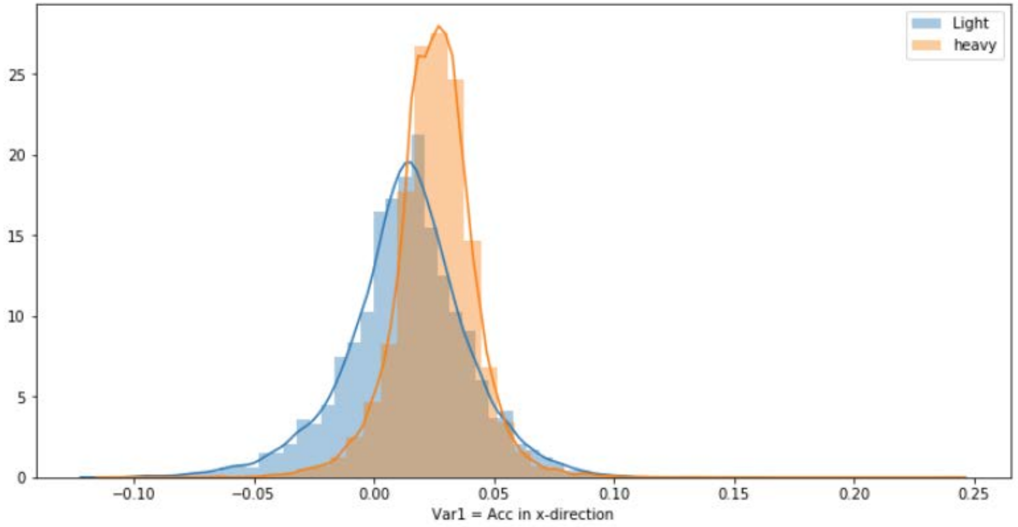



# Results: Data Preprocessing

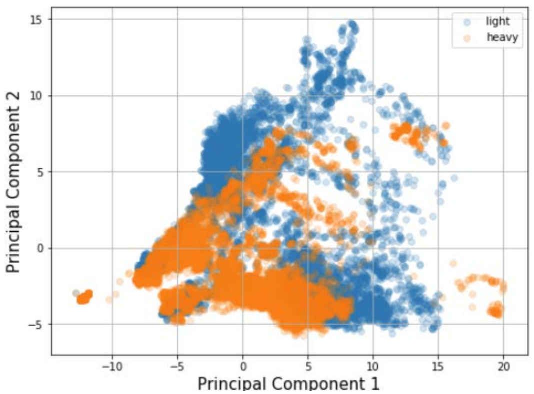
## PCA with 3 principle components, based on FFT

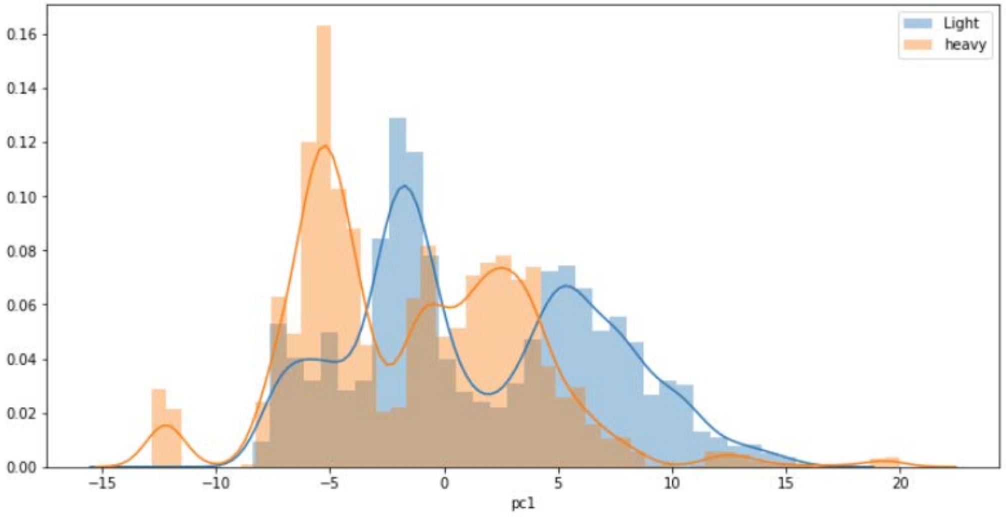
Not preprocessed




PCA on FFT features

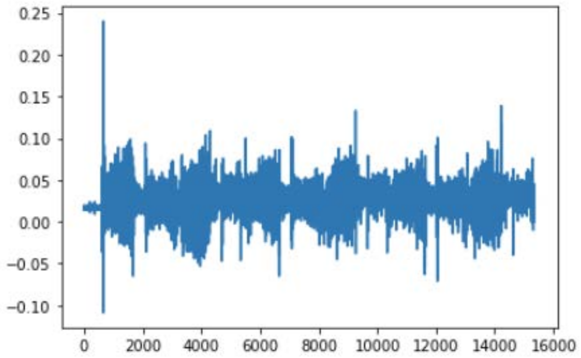


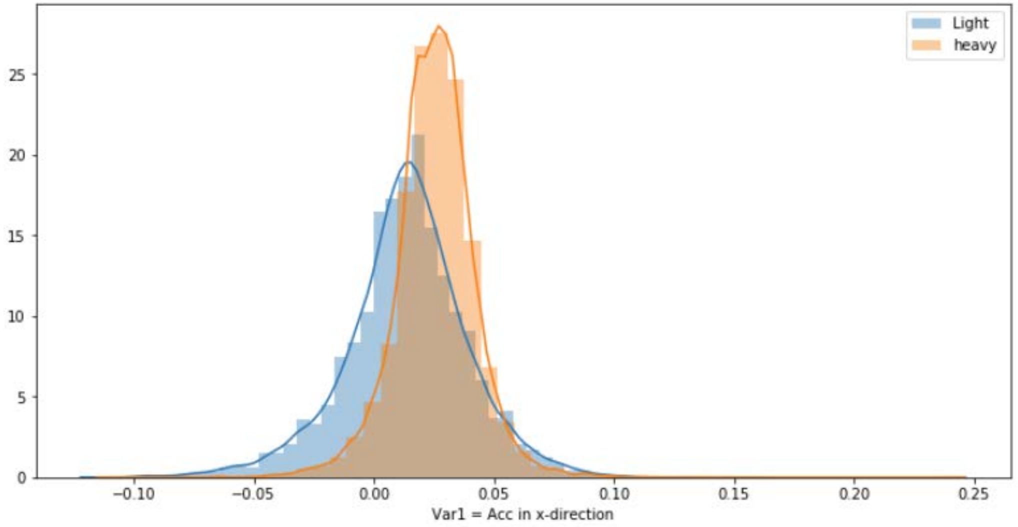



# Results: Data Preprocessing

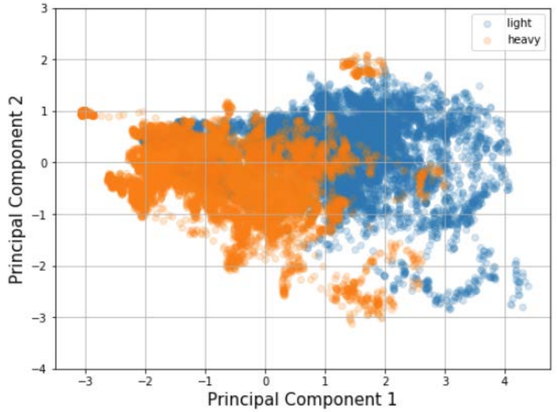
## PCA with 3 principle components, based on moving mean

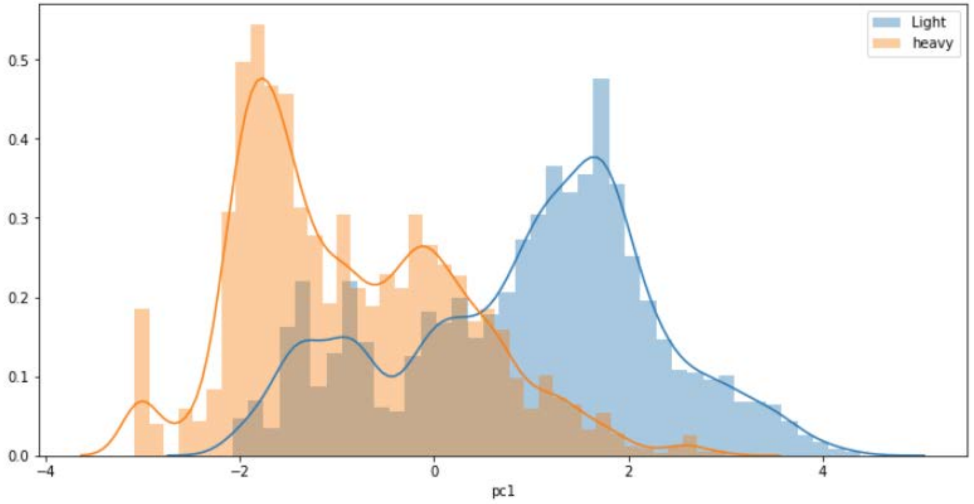
Not preprocessed





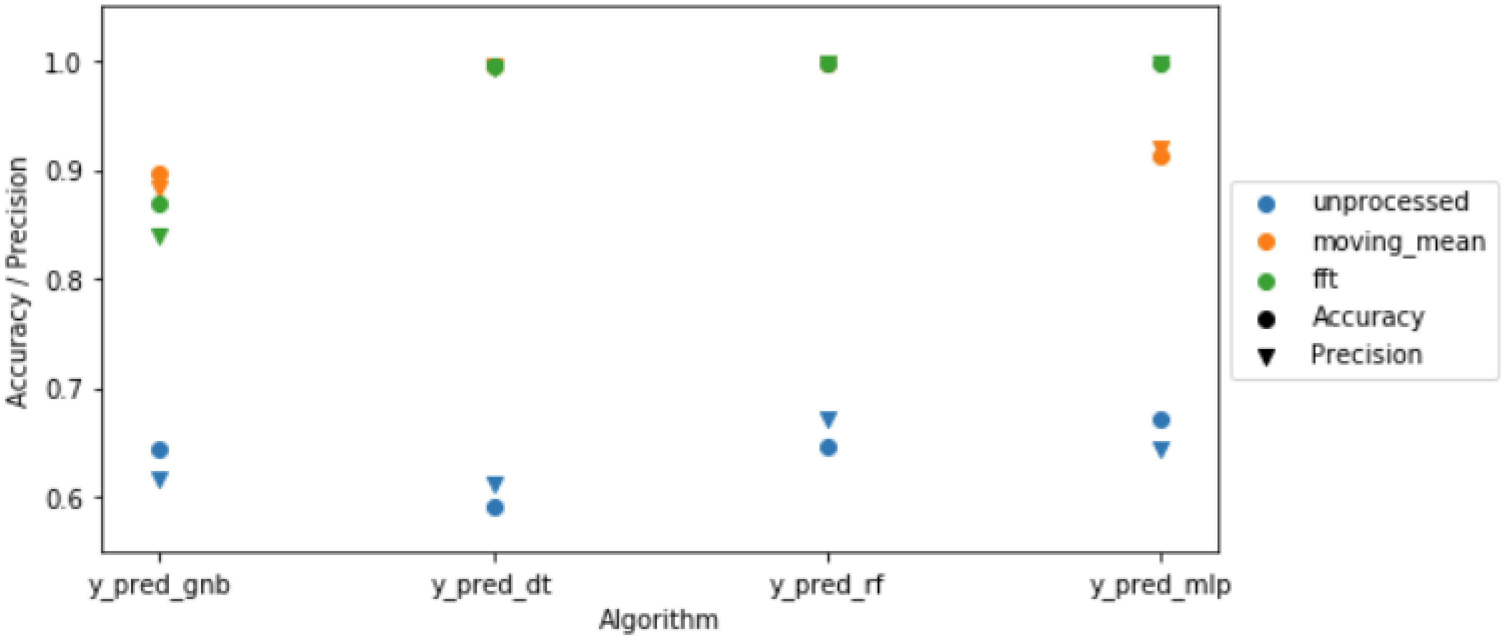

PCA on moving mean features



# Results: Data Preprocessing
## Final results

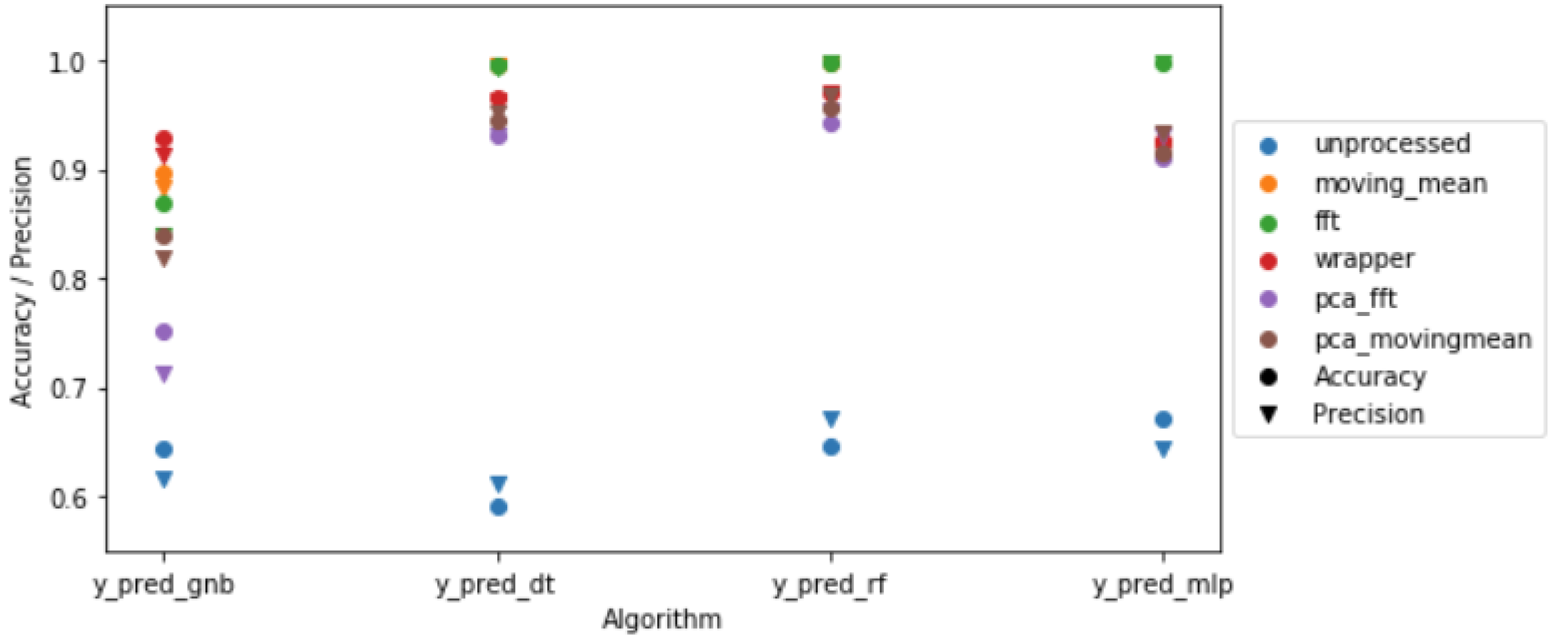
- Different algorithms
- Different preprocessing methods
- Showing accuracies and precision



# Results: Data Preprocessing

## Final results

- Different algorithms
- Different preprocessing methods
- Showing accuracies and precision



What to take with you?

# LEARNING OUTCOMES

# Key Findings

- Business and data undestanding built the foundation for your model
- Various methods are available for data preprocessing, an appropriate selection depends on many influences, mostly on the analysis question
- Many of them can be combined in creative ways
- Outlier detection is an important step to increase data quality
- Data quality is the key to success – prevent *garbage in, garbage out*
- Feature selection methods are used to focus on meaningful features
- Feature reduction methods reduce model complexity
- Grey box modeling is the key for technical domains, reviews are obligatory

# References


- Jason Bell: *Machine Learning: Hands-On for Developers and Technical Professionals.* John Wiley & Sons, Inc. (2015)
- Addison, D.; Wermter, S.; Arevian, G.: *A comparison of feature extraction and selection techniques*. In: International Conference on Artificial Neural Networks, S. 212–215 (2003)
- Guyon, I.; Elisseeff, A.: *An introduction to variable and feature selection*. Journal of machine learning research, 3(3): S. 1157–1182 (2003)
- Guyon, I. et al.: *Feature Extraction. Foundations and Applications*. Springer (2006)
- Hildebrand, K. et al.: Daten und Informationsqualität. Springer (2018)
- L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot, und G. Tutz, *Statistik: der Weg zur Datenanalyse*. Berlin Heidelberg: Springer Spektrum (2016)
- R. Kosfeld, H.-F. Eckey, und M. Türck, *Deskriptive Statistik: Grundlagen - Methoden - Beispiele - Aufgaben*. Wiesbaden: Springer Gabler (2016)
- T. Becker, R. Herrmann, V. Sandor, D. Schäfer, und U. Wellisch, *Deskriptive Statistik und explorative Datenanalyse.* In: Stochastische Risikomodellierung und statistische Methoden, Berlin, Heidelberg: Springer (2016) S. 27–91
- PCA: T. Runkler: *Data Mining, Modelle und Algorithmen intelligenter Datenanalyse*. Springer Verlag (2015) pp. 39-43
- Ansgar Steland: *Basiswissen Statistik*. Springer, Heidelberg (2016)
- Thomas Cleff: *Deskriptive Statistik und explorative Datenanalyse*. Springer, Heidelberg (2015)


THANK U