Three Ways to make your Industrial Data Science Projects a Success

Prof. Dr.-Ing. Jochen Deuse

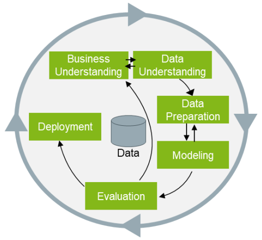
IDS 2019

Institute of Production Systems

Institute of Production Systems, Leonhard-Euler-Str. 5, D-44227 Dortmund

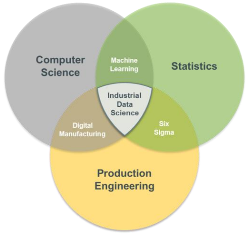
Defining the Process

- **Defining the Process**



- Dealing with Data Immaturity

- Combining Domain Knowledge with Data Science

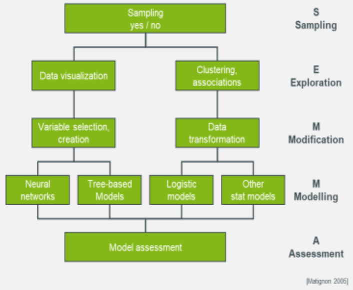


- Conclusion

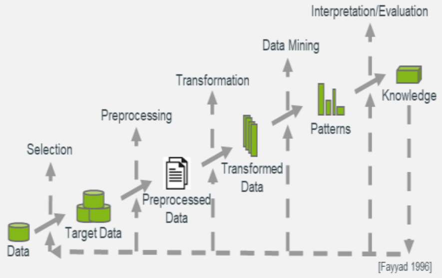
There is a Variety of Knowledge Discovery Processes

Institute of Production Systems

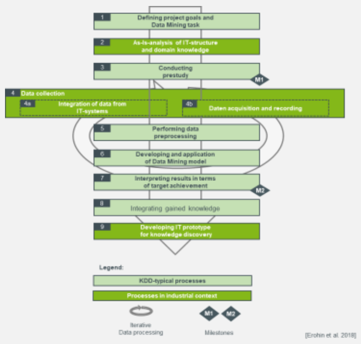
SEMMA of SAS



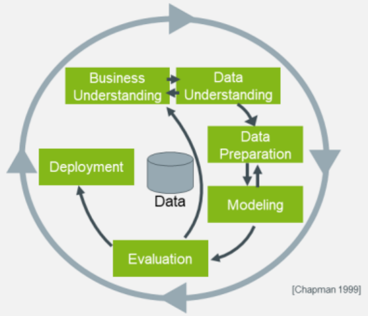
Knowledge Discovery in Databases (KDD)



Knowledge Discovery in Industrial Databases (KDID)

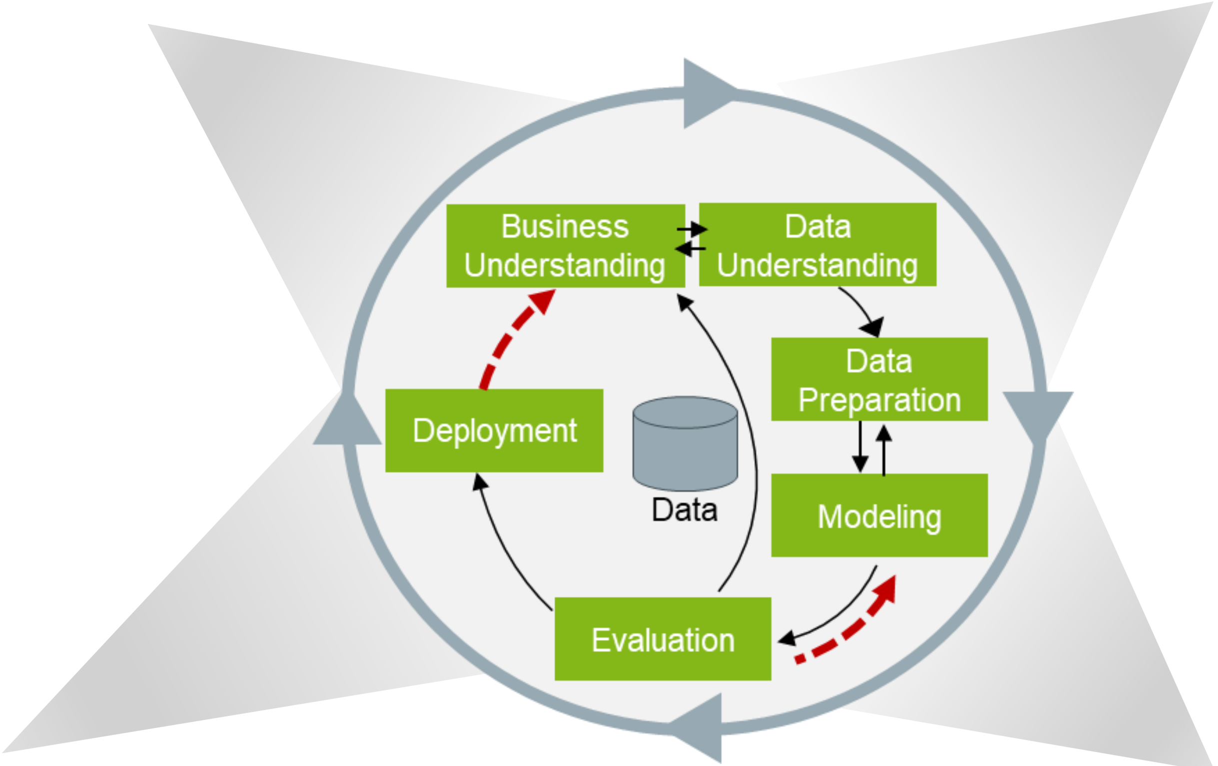


Cross-Industry Standard Process for Data Mining (CRISP-DM)



So why do we follow CRISP-DM?

- From a domain expert‘s perspective, the process is very intuitive
- It resembles a PDCA- respectively a DMAIC-circle



- It provides a well defined project structure
- It can easily be adapted across different industries

CRISP-DM provides a well defined Project Structure

SMD-Value Stream: Shortening quality control loops and reducing the need for X-Ray inspection

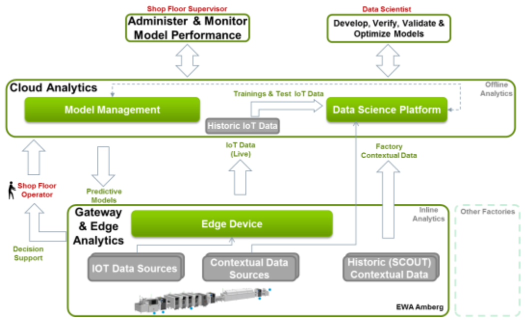
ATTRIBUTE

SPI Daten
- CSV-Dateien
- Eine Datei pro Nutzenseite
- Kopfzeile + Rumpfdaten
- 37 Attribute (numerisch & nominal)

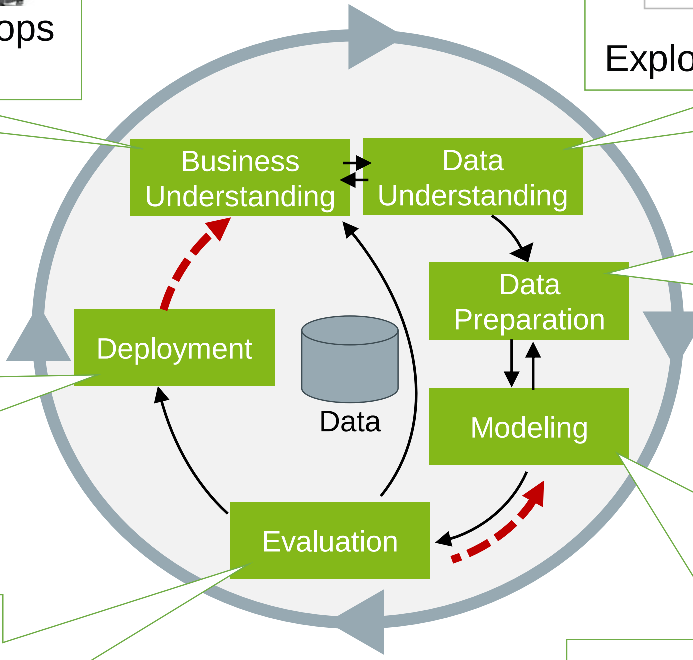
LABEL

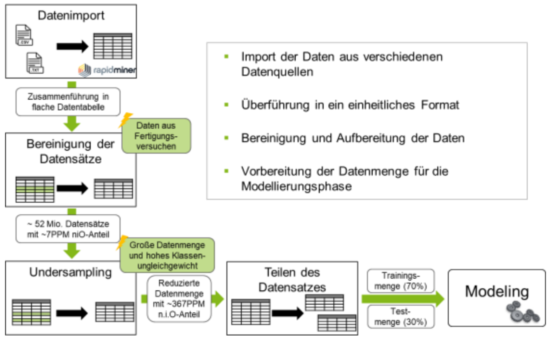
Röntgen Daten
- TXT-Dateien
- Eine Datei pro Nutzen
- Leer wenn Nutzen i.O., sonst fehlerhafte Board_Component_Pin

Exploring process and inspection data



Deploying based on IoT-Architecture



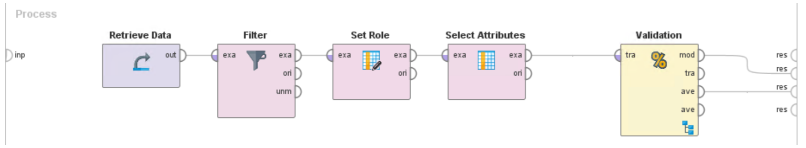


- Import der Daten aus verschiedenen Datenquellen
- Überführung in ein einheitliches Format
- Bereinigung und Aufbereitung der Daten
- Vorbereitung der Datenmenge für die Modellierungsphase

Aggregating and cleansing of data

300.103 Datensätze Gesamt		Wahres Ergebnis: n.i.O (Positive)	Wahres Ergebnis: i.O (Negative)		
Prognose	n.i.O (Positive)	107	110.664	Precision 0,10 %	Pseudofehlerrate 99,90 %
Prognose	i.O (Negative)	3	189.329	False Omission Rate 0,00 %	Negative Predictive Value 100 %
		Sensitivity / Recall 97,27 %	False Positive Rate 36,89 %	Accuracy 63,12 %	
		Schlupfrate 2,73 %	Specificity 63,11 %		

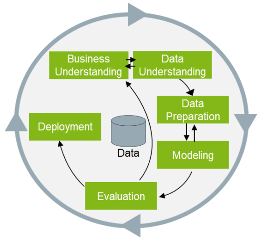
Optimising slack rate and pseudo faults



Selecting and configuring suitable prediction models

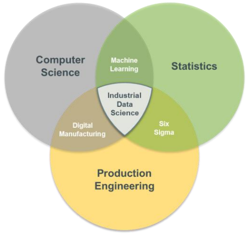
Institute of Production Systems

- Defining the Process



- **Dealing with Data Immaturity**

- Combining Domain Knowledge with Data Science



- Conclusion

Data Maturity can be assessed by applying a defined Set of Criteria

- Data Acquisition – How is data collected along the value stream?
- Sample Size – Are there enough representatives of each class and are they evenly distributed?
- Reference Level – Is the data available in a high and uniform granularity?
- Consistency – Does the relevant data set contain logical contradictions?
- Traceability – Can label and feature value characteristics be joined unambiguously?
- …

We have specified ten Criteria and four Levels of Maturity each

Institute of Production Systems

Criteria	Maturity level			
	1	2	3	4
Data collection	manual entry	electronical, must be triggered manually	data acquisition is carried out automatically in most cases	fully automated data collection
Completeness of data collection	unilateral and incomplete recording of relevant characteristics	recording of the essential characteristics	recording of a large part of the relevant characteristics	recording of all relevant, (un)influenceable characteristics
Sample size	no historic data	small sample per object group	large sample per object group, but unbalanced data	large sample with large number per object group and class
Data sources	paperbased records	decentralised data storage with simple software (e.g. Excel)	different data management systems with central data storage	comprehensive Data Warehouse
Data format	formats that are difficult to process (e.g. scans, photos)	formats with limited processability (e.g. PDF)	different, directly processable formats (e.g. CSV, XML)	comprehensive standard format
Data structure	unstructured text or images	semi-structured data (e.g. XML, JSON)	structured, mixed-scaled data	structured, metrically scaled data and standardized codes
Feature type	only set points	highly aggregated actual values	aggregated actual values or raw data with low sampling rate	raw data in real time
Reference level	value characteristics at the highest reference level	value characteristics at the upper reference level	value characteristics at the next higher level	value characteristics at individual element level
Consistency of data	no consistency/integrity	massive amount of logical differences	few logical differences	full integrity/consistency
Traceability	no ID/ time stamp	different ID/ timestamp	comprehensive ID/ time stamp	comprehensive ID/ timestamp on same reference level

Reference: Eickelmann et al. (2019): Bewertungsmodell zur Analyse der Datenreife.
In: ZWF Jg. 114, 1-2, S. 29-33

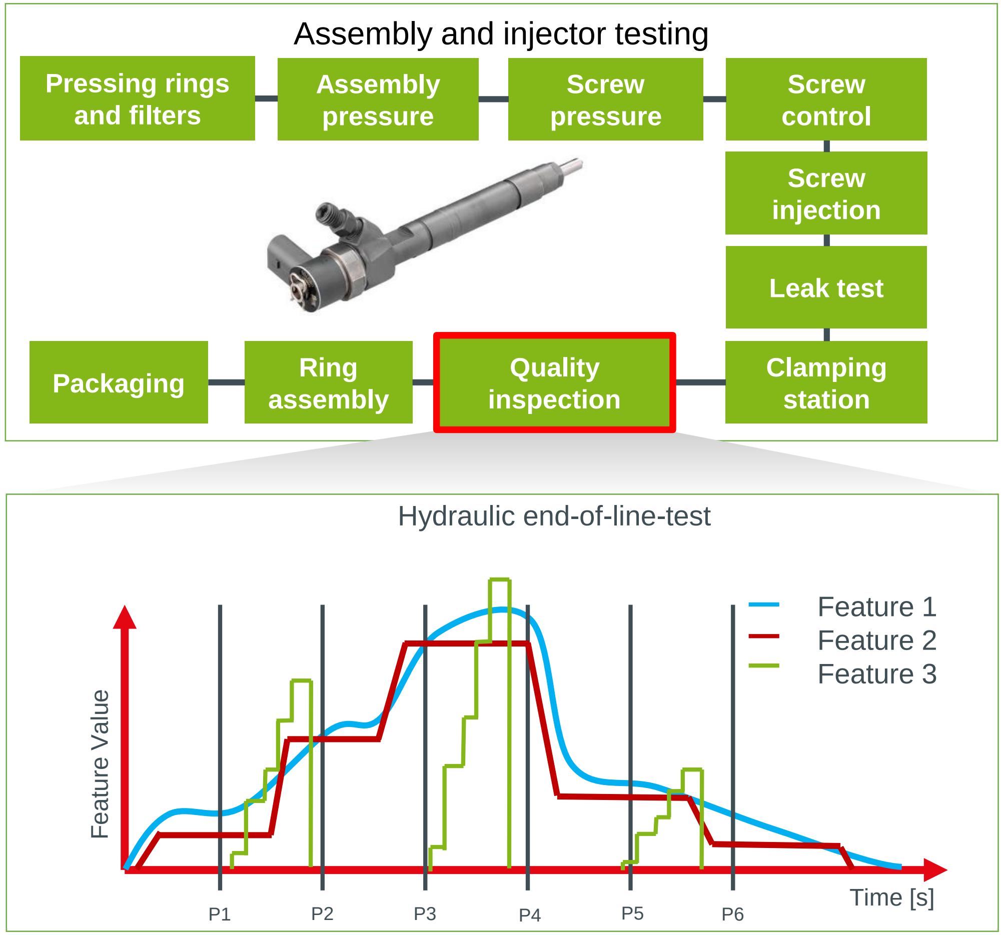
Non uniform Reference Levels prohibit Supervised Learning

Criteria	Maturity level			
	1	2	3	4
Data collection	manual entry	electronical, must be triggered manually	data acquisition is carried out automatically in most cases	fully automated data collection
Completeness of data collection	unilateral and incomplete recording of relevant characteristics	recording of the essential characteristics	recording of a large part of the relevant characteristics	recording of all relevant, (un)influenceable characteristics
Sample size	no historic data	small sample per object group	large sample per object group, but unbalanced data	large sample with large number per object group and class
Data sources	paperbased records	decentralised data storage with simple software (e.g. Excel)	different data management systems with central data storage	comprehensive Data Warehouse
Data format	formats that are difficult to process (e.g. scans, photos)	formats with limited processability (e.g. PDF)	different, directly processable formats (e.g. CSV, XML)	comprehensive standard format
Data structure	unstructured text or images	semi-structured data (e.g. XML, JSON)	structured, mixed-scaled data	structured, metrically scaled data and standardized codes
Feature type	only set points	highly aggregated actual values	aggregated actual values or raw data with low sampling rate	raw data in real time
Reference level	value characteristics at the highest reference level	value characteristics at the upper reference level	value characteristics at the next higher level	value characteristics at individual element level
Consistency of data	no consistency/integrity	massive amount of logical differences	few logical differences	full integrity/consistency
Traceability	no ID/ time stamp	different ID/ timestamp	comprehensive ID/ time stamp	comprehensive ID/ timestamp on same reference level

Supervised Learning of Quality Labels from End-of-Line-Test Data

Institute of Production Systems

- Diesel Injector Nozzle Manufacturing Value Stream



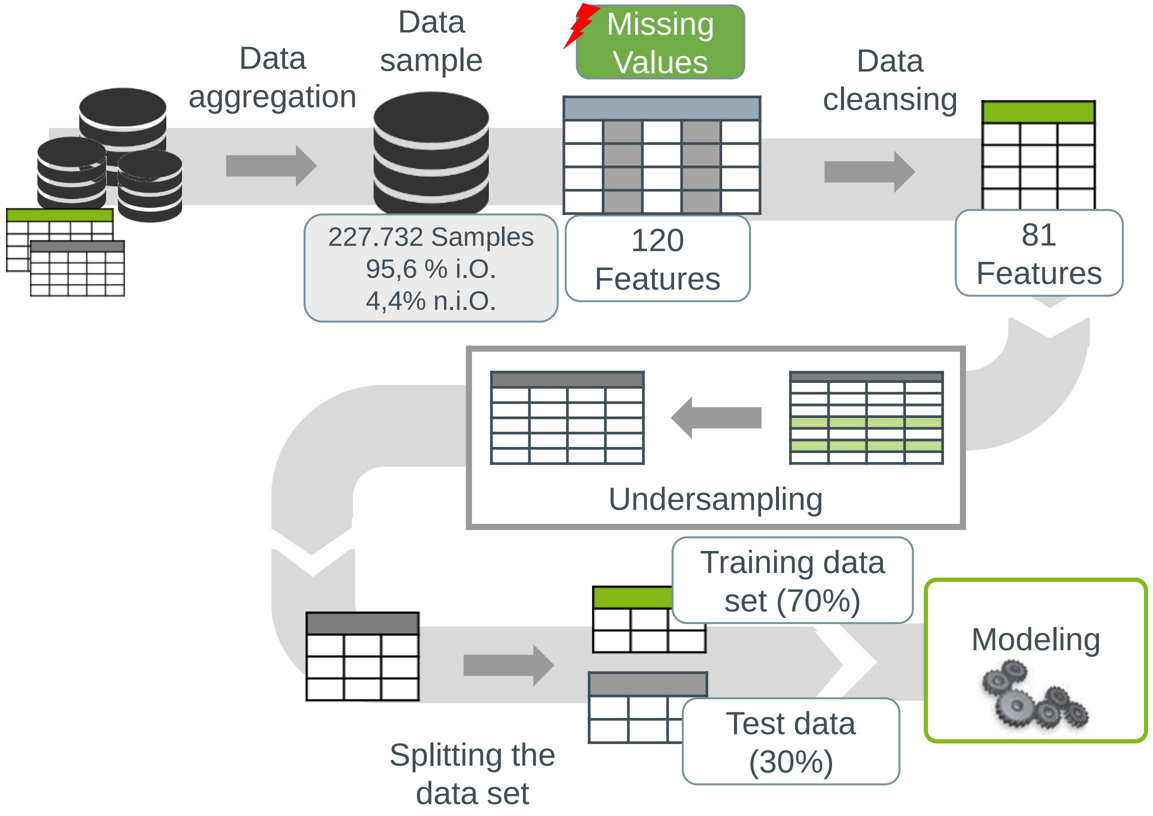
		True result			
		NOK 3.018 (4,42 %)	**OK** 65.302 (95,58 %)		
Forecast	**NOK** 1.656 (2,42 %)	1.021	635 Pseudo faults	Precision 61,65 %	Pseudo fault rate 38,35 %
Forecast	**OK** 66.664 (97,58 %)	1.997 Slack	64.667	False omission rate 3,00 %	Negative predictive value 97,00 %
		Sensitivity / Recall 33,83 %	False Positive Rate 0,97 %	Accuracy 96,15 %	
		Slack rate 66,17 %	Specificity 99,03 %		

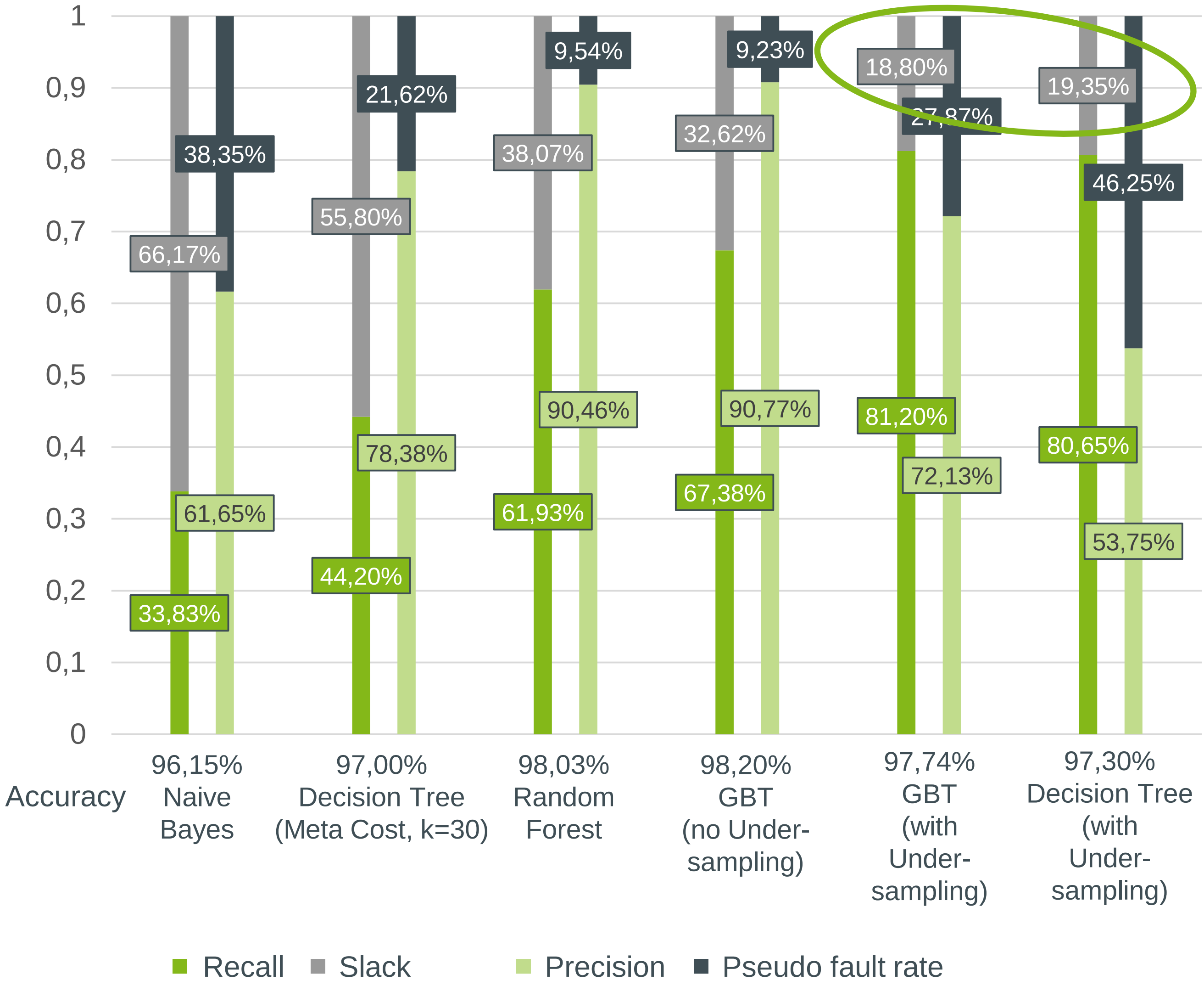
Unbalanced Label Proportions result in high Recall Rates

Institute of Production Systems

Criteria	Maturity level			
	1	2	3	4
Data collection	manual entry	electronical, must be triggered manually	data acquisition is carried out automatically in most cases	fully automated data collection
Completeness of data collection	unilateral and incomplete recording of relevant characteristics	recording of the essential characteristics	recording of a large part of the relevant characteristics	recording of all relevant, (un)influenceable characteristics
Sample size	no historic data	small sample per object group	large sample per object group, but unbalanced data	large sample with large number per object group and class
Data sources	paperbased records	decentralised data storage with simple software (e.g. Excel)	different data management systems with central data storage	comprehensive Data Warehouse
Data format	formats that are difficult to process (e.g. scans, photos)	formats with limited processability (e.g. PDF)	different, directly processable formats (e.g. CSV, XML)	comprehensive standard format
Data structure	unstructured text or images	semi-structured data (e.g. XML, JSON)	structured, mixed-scaled data	structured, metrically scaled data and standardized codes
Feature type	only set points	highly aggregated actual values	aggregated actual values or raw data with low sampling rate	raw data in real time
Reference level	value characteristics at the highest reference level	value characteristics at the upper reference level	value characteristics at the next higher level	value characteristics at individual element level
Consistency of data	no consistency/integrity	massive amount of logical differences	few logical differences	full integrity/consistency
Traceability	no ID/ time stamp	different ID/ timestamp	comprehensive ID/ time stamp	comprehensive ID/ timestamp on same reference level

Undersampling reduces the Effect of unbalanced Label Proportions





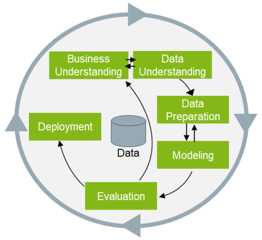
A Lack of Tracebility prohibits Supervised Learning

Institute of Production Systems

Criteria	Maturity level			
	1	2	3	4
Data collection	manual entry	electronical, must be triggered manually	data acquisition is carried out automatically in most cases	fully automated data collection
Completeness of data collection	unilateral and incomplete recording of relevant characteristics	recording of the essential characteristics	recording of a large part of the relevant characteristics	recording of all relevant, (un)influenceable characteristics
Sample size	no historic data	small sample per object group	large sample per object group, but unbalanced data	large sample with large number per object group and class
Data sources	paperbased records	decentralised data storage with simple software (e.g. Excel)	different data management systems with central data storage	comprehensive Data Warehouse
Data format	formats that are difficult to process (e.g. scans, photos)	formats with limited processability (e.g. PDF)	different, directly processable formats (e.g. CSV, XML)	comprehensive standard format
Data structure	unstructured text or images	semi-structured data (e.g. XML, JSON)	structured, mixed-scaled data	structured, metrically scaled data and standardized codes
Feature type	only set points	highly aggregated actual values	aggregated actual values or raw data with low sampling rate	raw data in real time
Reference level	value characteristics at the highest reference level	value characteristics at the upper reference level	value characteristics at the next higher level	value characteristics at individual element level
Consistency of data	no consistency/integrity	massive amount of logical differences	few logical differences	full integrity/consistency
Traceability	no ID/ time stamp	different ID/ timestamp	comprehensive ID/ time stamp	comprehensive ID/ timestamp on same reference level

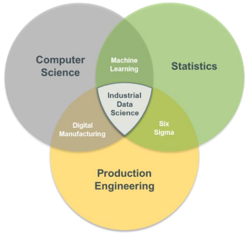
Combining Domain Knowledge with Data Science

- Defining the Process



- Dealing with Data Immaturity

- **Combining Domain Knowledge with Data Science**



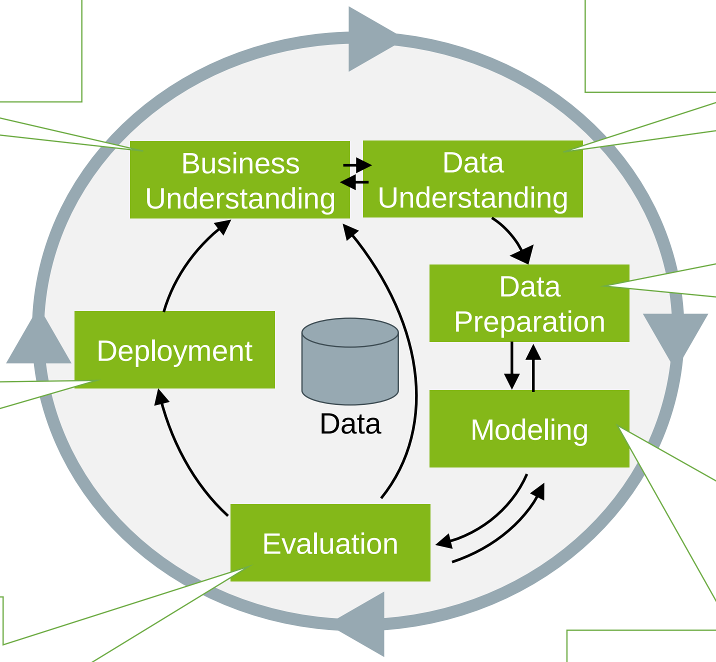
- Conclusion

Domain Knowledge is required in every Stage of CRISP-DM

Institute of Production Systems

Business-related objectives are derived and suitable success criteria are defined

Faster and better selection of critical features and labels by applying domain knowledge



Implementation and integration of prediction models in manufacturing and business processes

Domain knowledge enables the identification and elimination of unrealistic/wrong data

The project team is able to interpret and validate the model results in an application-oriented manner

Existing domain knowledge can be applied to avoid overfitting

Data Scientists and Domain Experts have been learning from each other

Institute of Production Systems

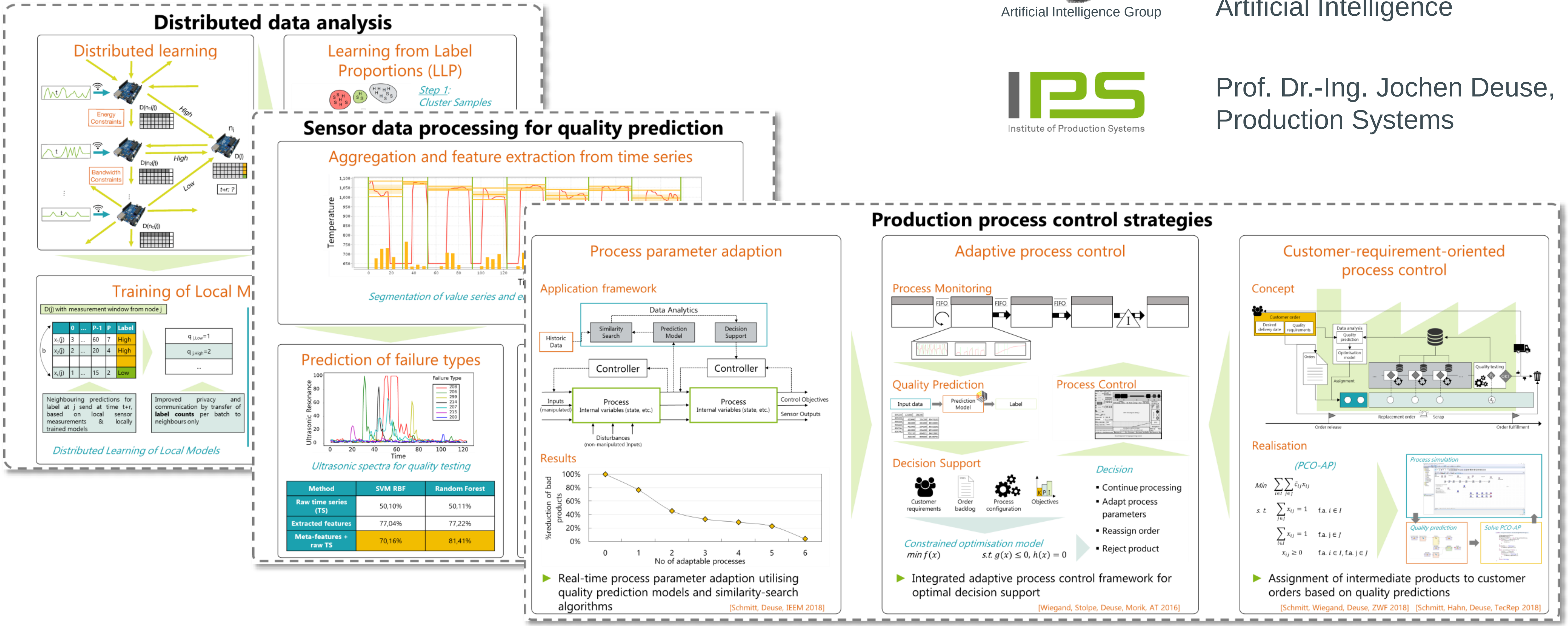
SFB 876 - B3: Data Mining on Sensor Data of Automated Processes

Artificial Intelligence Group

Prof. Dr. Katharina Morik,
Artificial Intelligence

Institute of Production Systems

Prof. Dr.-Ing. Jochen Deuse,
Production Systems



Citizen Data Scientists blending Domain Knowledge and Data Science

SYSKRON Citizen data scientists RIF Institut für Forschung und Transfer

BITBURGER BRAUGRUPPE STARKE MARKEN
Augustiner Bräu München gegründet 1328
Domain experts

rapidminer
Data scientists

Prediction of…

…malt processability

…lautering duration

…yeast processing yield

Interdisciplinary Training overcomes Faculty Boundaries

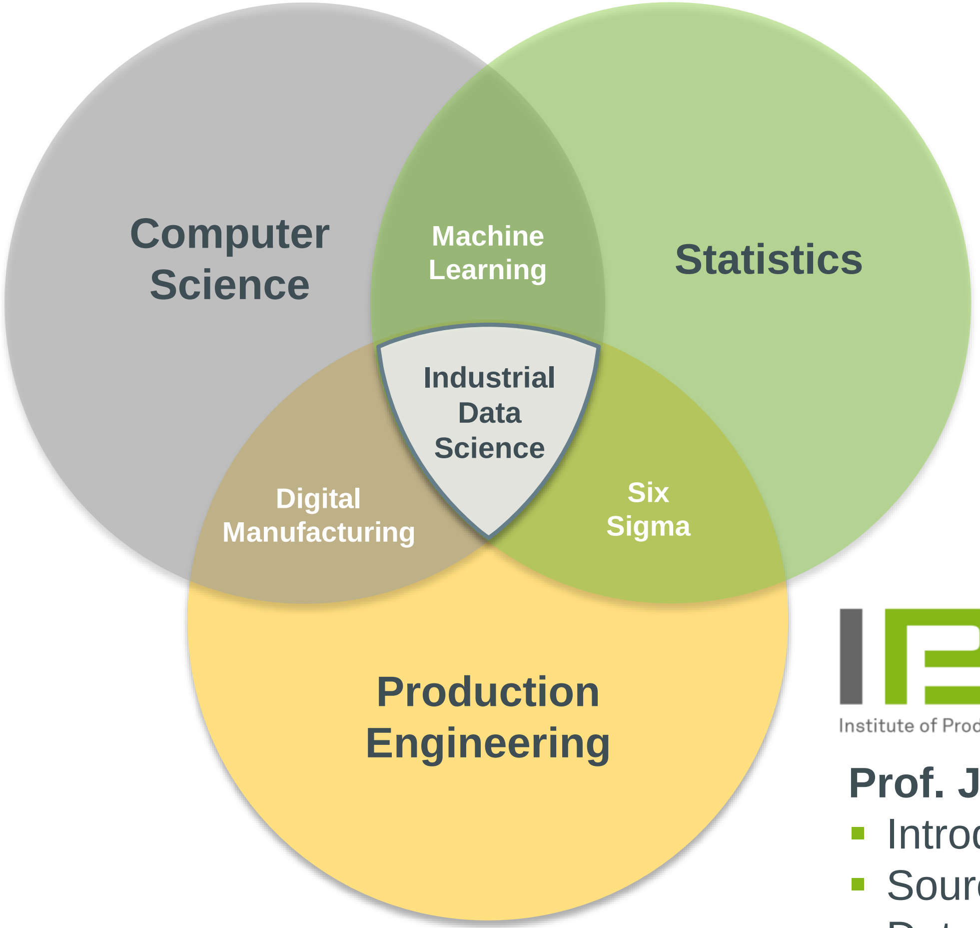
Institute of Production Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

Prof. Kristian Kersting:
- Deep Learning
- Tree-based procedures
- Ensemble strategies

fakultät statistik

Prof. Claus Weihs:
- Association Analysis
- Data Transformations
- Concepts for Model Selection



fakultät für informatik

Prof. Jens Teubner:
- Basics of Data Management
- Database Systems
- Data Warehouses

Institute of Production Systems

Prof. Jochen Deuse:
- Introduction to Industrial Data Science
- Sources of industrial data
- Data analysis in industrial environments

Students from all three Disciplines collaborating on Industrial Use Cases

Institute of Production Systems

Project coaching

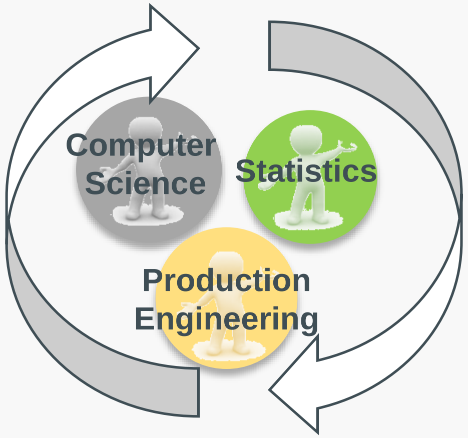
TECHNISCHE UNIVERSITÄT DARMSTADT

fakultät für informatik

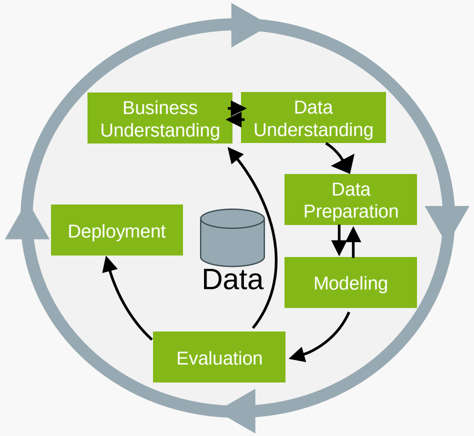
fakultät statistik

Institute of Production Systems

Use case driven competence development



Interdisciplinary project teams



Common project structure

Industrial use cases

Weidmüller

Quality prediction for injection molding

EVONIK KRAFT FÜR NEUES

Yield prediction for chemical processes

ebmpapst

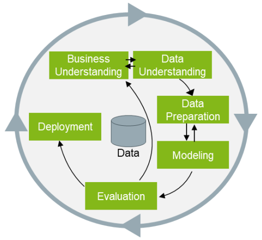
Quality prediction for fan assembly

BMW GROUP

Quality prediction for engine assembly

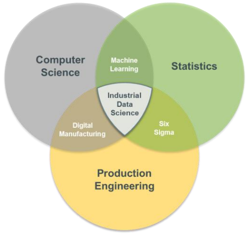
Conclusion

- Defining the Process



- Dealing with Data Immaturity

- Combining Domain Knowledge with Data Science



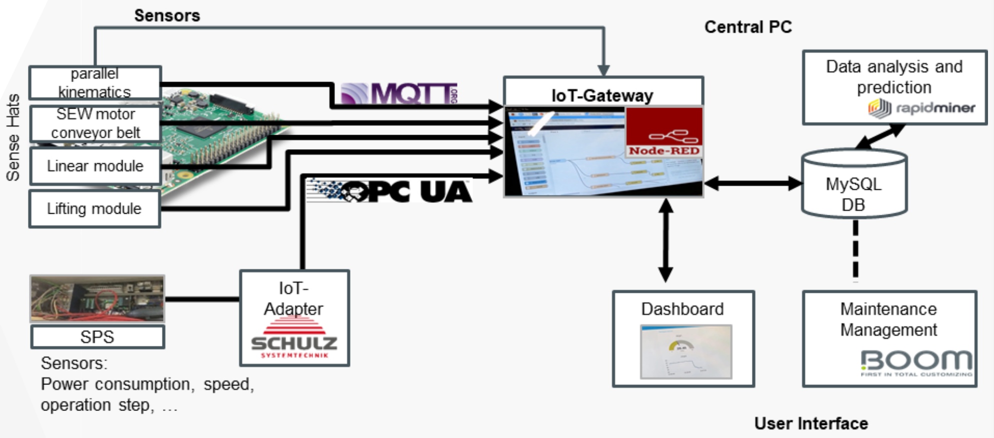
- **Conclusion**

Different Approaches can overcome Data Maturity Challenges

Institute of Production Systems

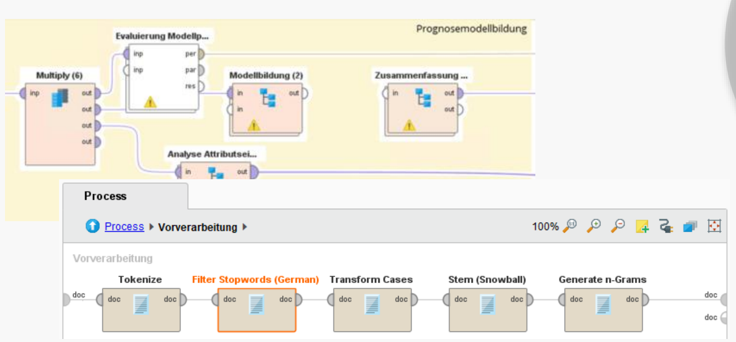
Criteria	Maturity Level			
	1	2	3	4
Data Collection	manual entry	electronical, must be triggered manually	data acquisition is carried out automatically in most cases	fully automated data collection
Completeness of data collection	unilateral and incomplete recording of relevant characteristics	recording of the essential characteristics	recording of a large part of the relevant characteristics	recording of all relevant, (un)influenceable characteristics
Sample Size	no historic data	small sample per object group	large sample per object group, but unbalanced data	large sample with large number per object group and class
Data Sources	paperbased records	decentralised data storage with simple software (e.g. Excel)	different data management systems with central data storage	comprehensive Data Warehouse
Data Format	formats that are difficult to process (e.g. scans, photos)	formats with limited processability (e.g. PDF)	different, directly processable formats (e.g. CSV, XML)	comprehensive standard format
Data Structure	unstructured text or images	Semi-structured data (e.g. XML, JSON)	structured, mixed-scaled data	structured, metrically scaled data and standardized codes
Feature Type	only set points	highly aggregated actual values	aggregated actual values or raw data with low sampling rate	raw data in real
Reference Level	attributes of the values at the highest reference level	attributes of the values at the upper reference level	attributes of the values at the next higher level	attributes of the values at individual element level
Consistency of data	No consistency/integrity	Massive amount of logical differences	Few logical differences	Full integrity/consistency
Traceability	No ID/ time stamp	Different ID/ timestamp	Comprehensive ID/ time stamp	Comprehensive ID/ timestamp on same reference level

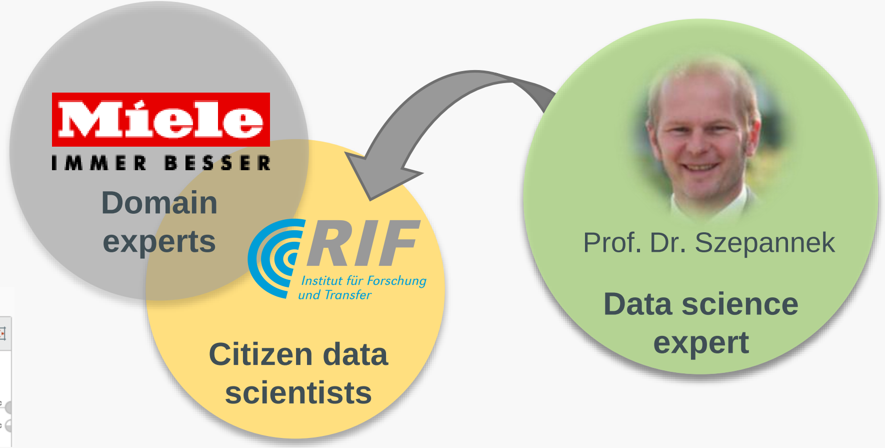
Improving data maturity (e.g. retrofitting)



Criteria	Maturity Level			
	1	2	3	4
Data Collection	manual entry	electronical, must be triggered manually	data acquisition is carried out automatically in most cases	fully automated data collection
Completeness of data collection	unilateral and incomplete recording of relevant characteristics	recording of the essential characteristics	recording of a large part of the relevant characteristics	recording of all relevant, (un)influenceable characteristics
Sample Size	no historic data	small sample per object group	large sample per object group, but unbalanced data	large sample with large number per object group and class
Data Sources	paperbased records	decentralised data storage with simple software (e.g. Excel)	different data management systems with central data storage	comprehensive Data Warehouse
Data Format	formats that are difficult to process (e.g. scans, photos)	formats with limited processability (e.g. PDF)	different, directly processable formats (e.g. CSV, XML)	comprehensive standard format
Data Structure	unstructured text or images	Semi-structured data (e.g. XML, JSON)	structured, mixed-scaled data	structured, metrically scaled data and standardized codes
Feature Type	only set points	highly aggregated actual values	aggregated actual values or raw data with low sampling rate	raw data in real
Reference Level	attributes of the values at the highest reference level	attributes of the values at the upper reference level	attributes of the values at the next higher level	attributes of the values at individual element level
Consistency of data	No consistency/integrity	Massive amount of logical differences	Few logical differences	Full integrity/consistency
Traceability	No ID/ time stamp	Different ID/ timestamp	Comprehensive ID/ time stamp	Comprehensive ID/ timestamp on same reference level

Involving senior data scientists





Thank you for your kind Attention!

Institute of Production Systems