

Saubere Daten Starke Modelle

Datenvorverarbeitung für KI-Systeme in industriellen
Entwicklungsprozessen

Patrick Kraus-Füreder

30.10.2025



Contact

DI Dr. techn. eur. Patrick Kraus-Füreder

AI Product Manager

Mobile: +43 664 1282849

E-Mail: patrick.kraus-fuereder@risc-software.at

RISC Software GmbH

Softwarepark 32a • 4232 Hagenberg • Austria

office@risc-software.at • <https://www.risc-software.at>





SITUATED IN...

Softwarepark Hagenberg

FOUNDED BY UNIV. PROF. DR. BRUNO BUCHBERGER

>3100

people at the site –
working, researching,
teaching, learning

11

research
institutions

25

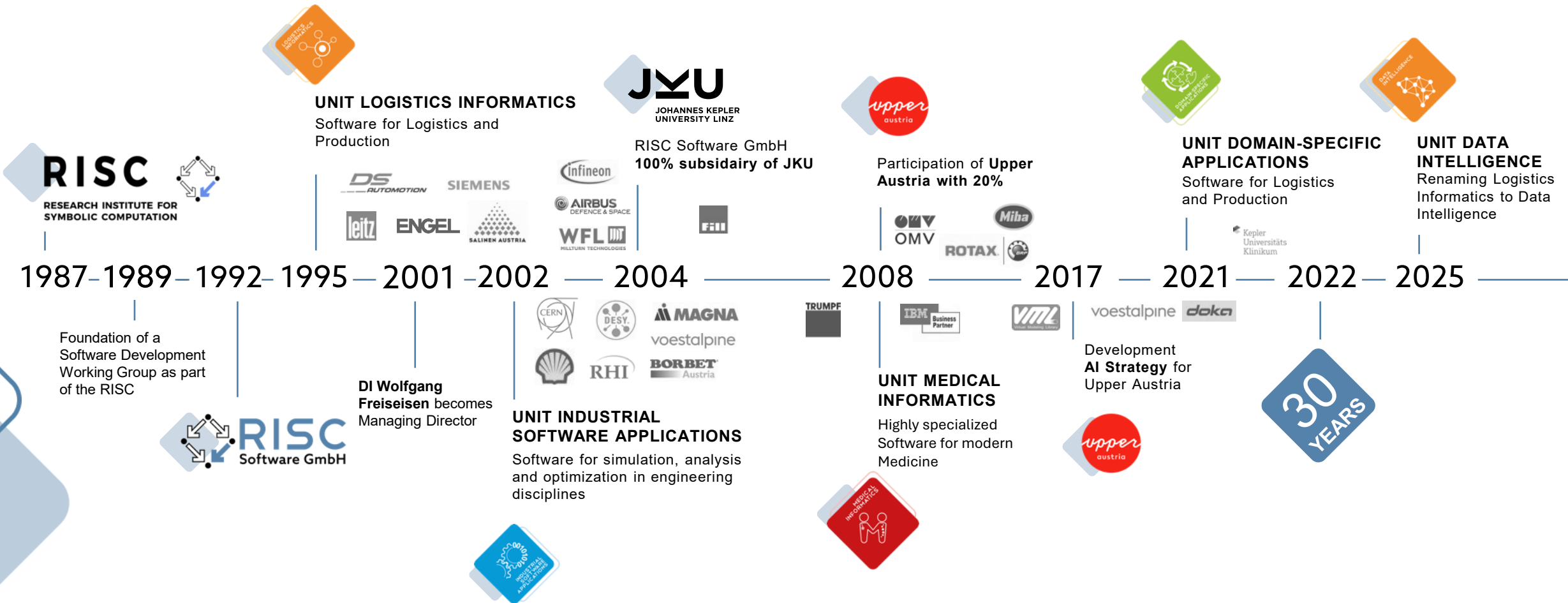
study
programs

75

companies



Our History



FACTS

RISC Hagenberg

130 
Employees
from research and
development

200 
Cooperation
Partners


> 12 million EUR*
Research
Budget
*est. 2025


~100
Projects


130
Publications


52
Participations

2024

COMPANY

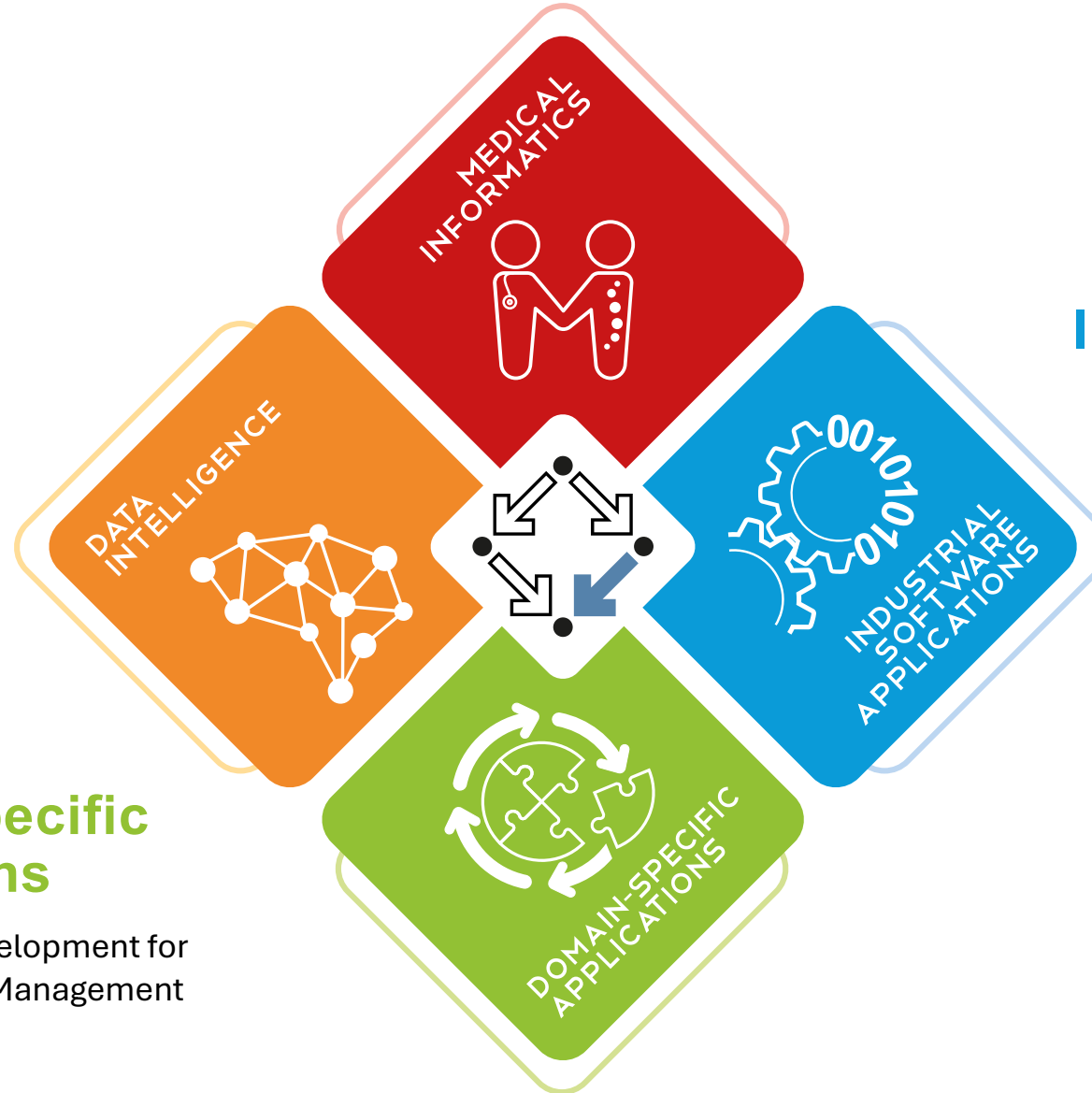
Units

Data Intelligence

Intelligent data solutions for production, energy and mobility

Domain-specific Applications

Agile Software Development for Data and Process Management



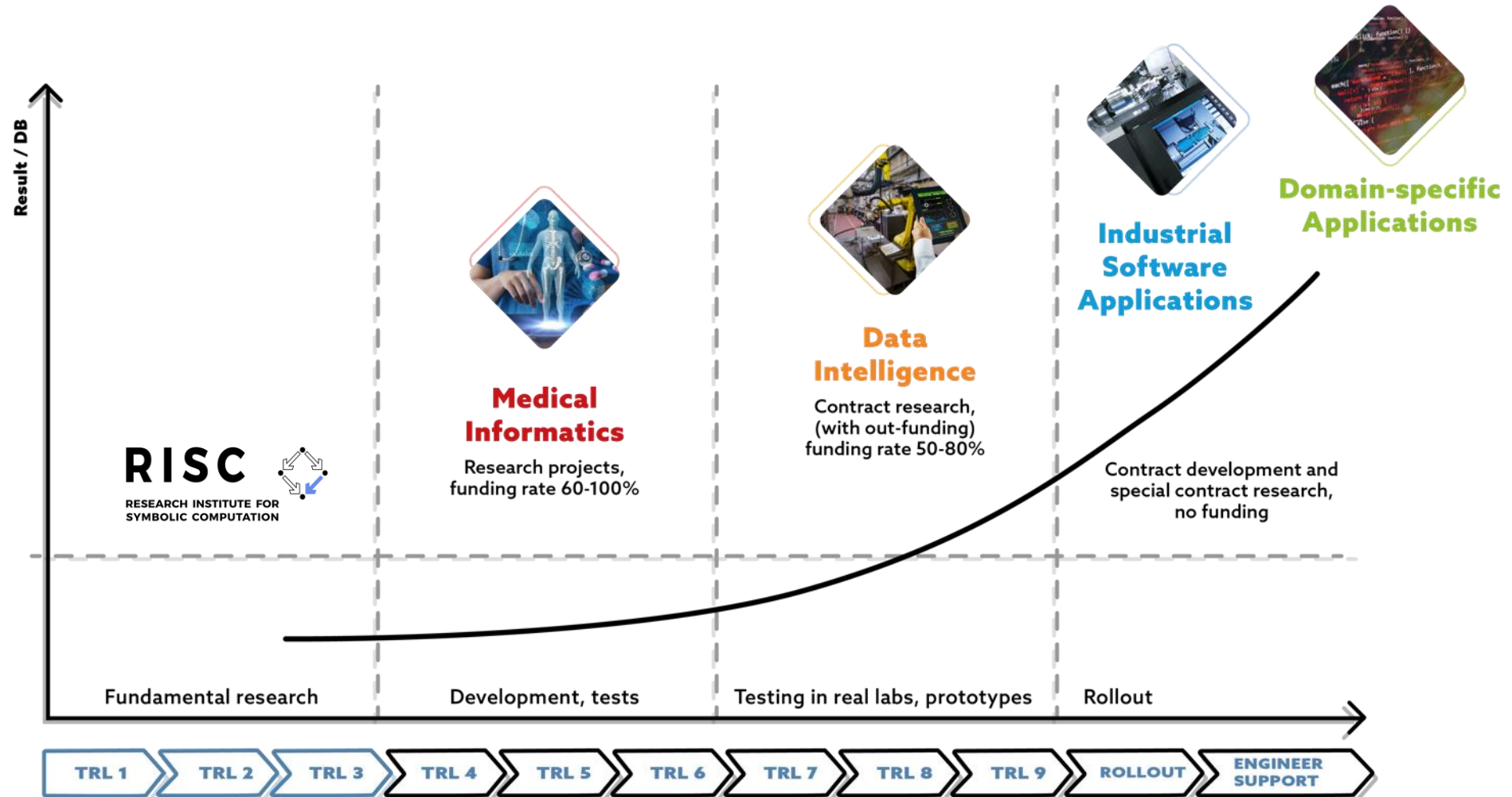
Medical Informatics

Highly specialized Software for modern Medicine

Industrial Software Applications

Software for Simulations, Analyses and Optimizations in technical Disciplines

Technology Readiness Level



OUR CORE COMPETENCE

Software



MODELING AND MULTIDISCIPLINARY STRUCTURAL OPTIMIZATION SYSTEMS

Virtual product development, CAD/CAE,
Customers/Partners: Airbus, LCM

For over 20 years, several hundred installations

DIGITAL-TWIN SIMULATION SYSTEMS

CAM, virtual modeling library, holistic simulations
Customers/Partners: WFL Millturn Technologies

For over 20 years, > 300 installations



OPTIMIZATION AND ANALYTICS

Detailed planning, forecasts, production optimization math. Optimization, AI
forecasts, real time, customers: VA Stahl, OMV, Borbet, 3eAG, Sprecher Autom.

For over 25 years,
Sprecher real-time optimization approx. 35000 blocks/week
3eAG approx. 15,000 forecasts/week

CONTROL SOFTWARE FOR DRIVERLESS TRANSPORT SYSTEMS

Automation, Autonomy, AGV, AMR, Customers/Partners: DS Automation

For over 20 years, > 150 installations



Reliability

- ◆ Reliable partner RISC
- ◆ Realize sustainable software solutions that are easy to maintain in the long term

Continuity

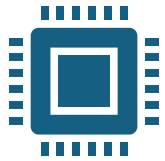
- ◆ The best proof:
our highly complex customer projects,
which have been successfully developed
and maintained for many years

Domain Know-how

- ◆ Successfully solved challenging
problems for our customers in the past
- ◆ Built up extensive domain expertise in
the industrial and medical environment

1959 – IBM 704

Die Geburt des Machine Learning



12k FLOP
18 kB RAM



Einige Tausend
annotierte
Positionen



Tausende Spiele
gegen sich
selbst

Anpassung von **32 internen Parametern**
zur Optimierung der Spielleistung



1959 – IBM 704

Die Geburt des Machine Learning

1962

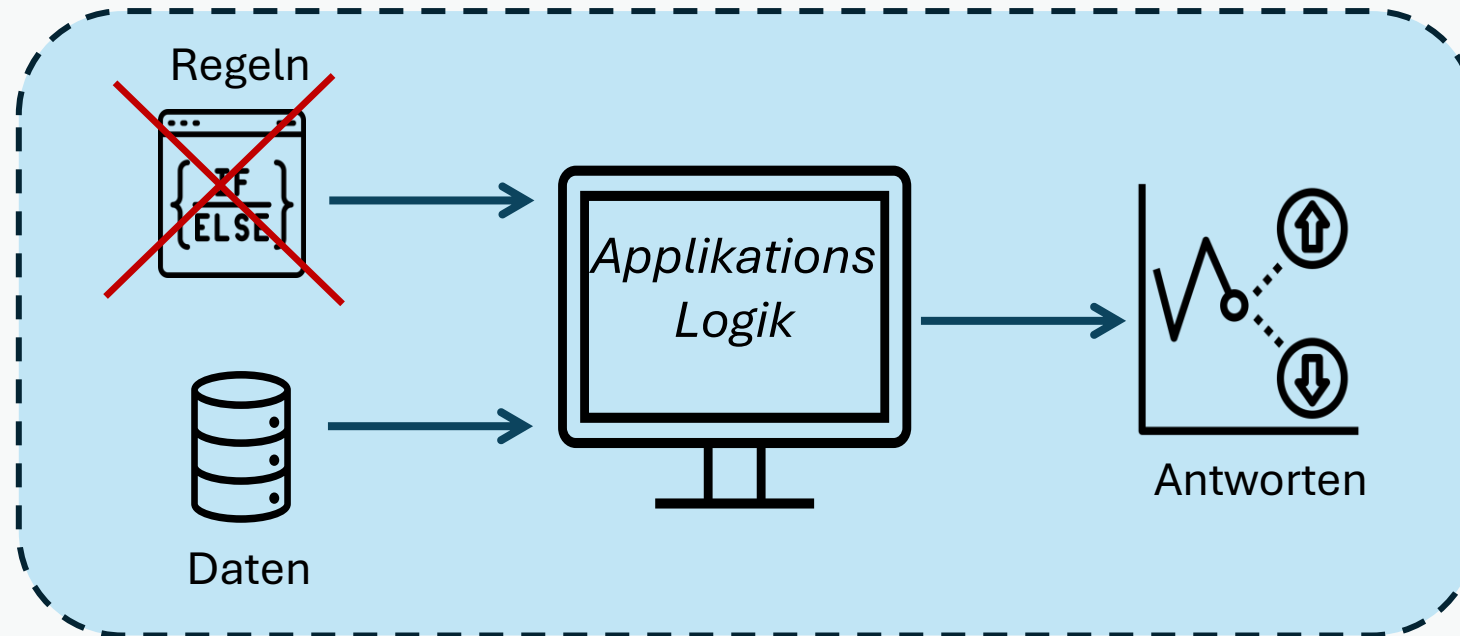
„Samuel Checkers“ schlägt (den selbsternannten) Dame-Meister Robert Nealey



Lernen – ohne zu wissen WAS zu lernen ist....

Was ist Machine Learning?

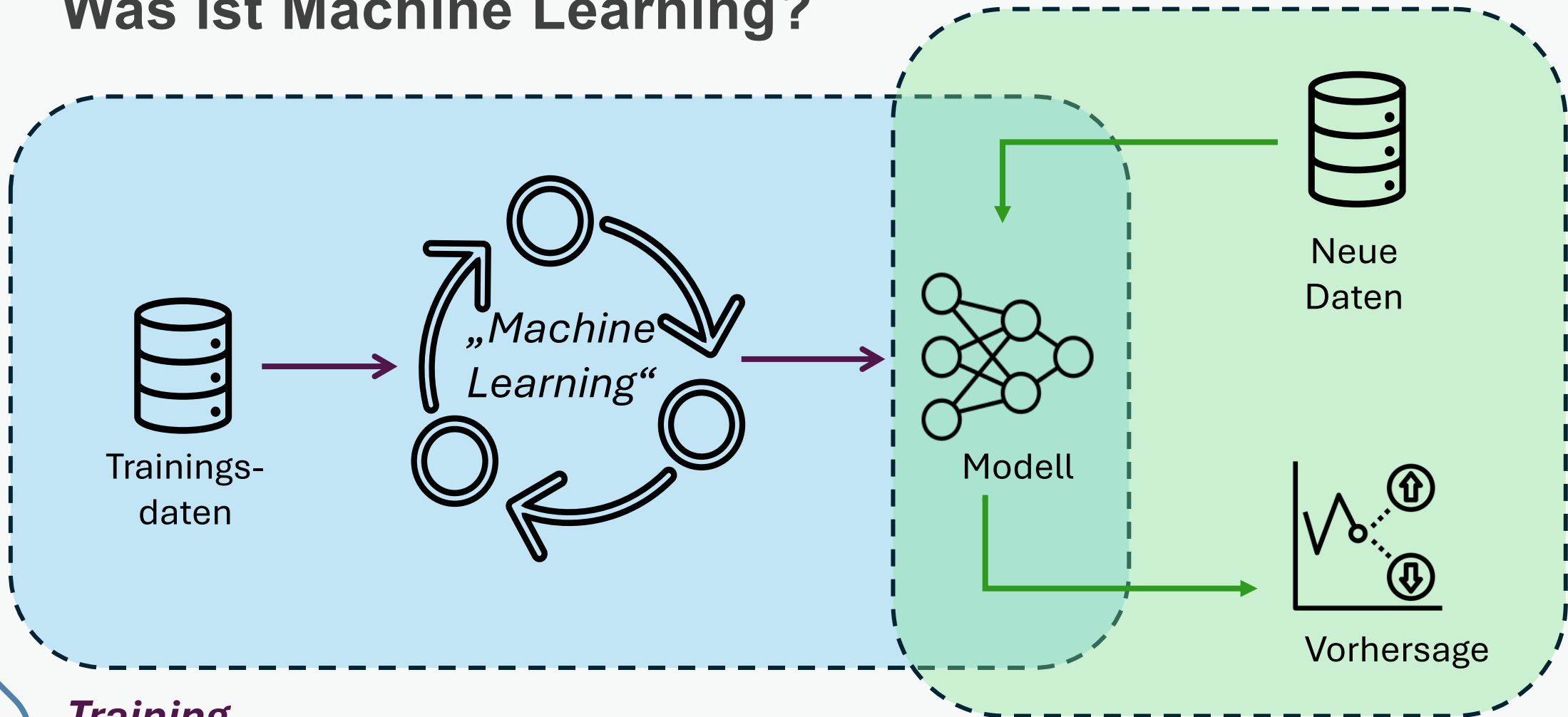
*Klassischer
Ansatz*



Lernen – ohne zu wissen WAS zu lernen ist....

Was ist Machine Learning?

Vorhersage „Inferenz“



Training

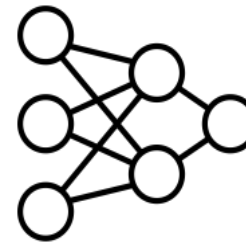
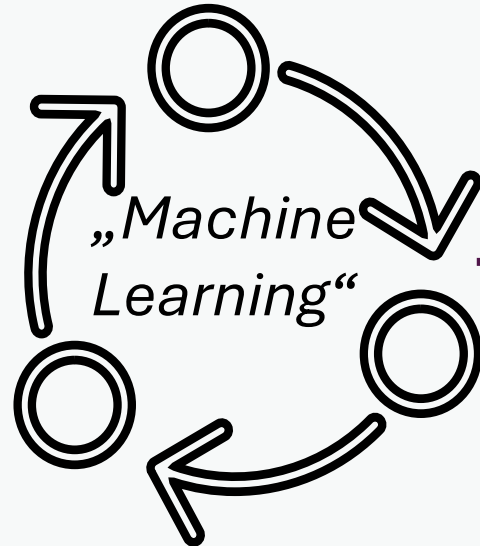
Lernen – ohne zu wissen WAS zu lernen ist....

Was ist Machine Learning?

Inferenz

Daten	Label
	Auto
	Boot
	Flugzeug
	...

~3k Beispiele



Modell



90% Auto

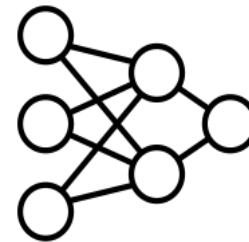
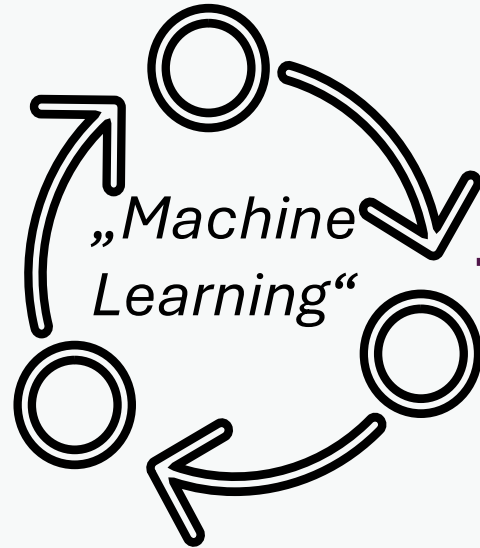
Training

Lernen – ohne zu wissen WAS zu lernen ist....

Was ist Machine Learning?

Daten	Label
Pos. 1 e4-f5...	Weiß gewinnt
Pos. 2 e4-f7...	Weiß gewinnt
Pos. 3 e3-f5...	Schwarz gewinnt
Pos. 4 b4-f5...	Weiß gewinnt

~3k Beispiele



Model

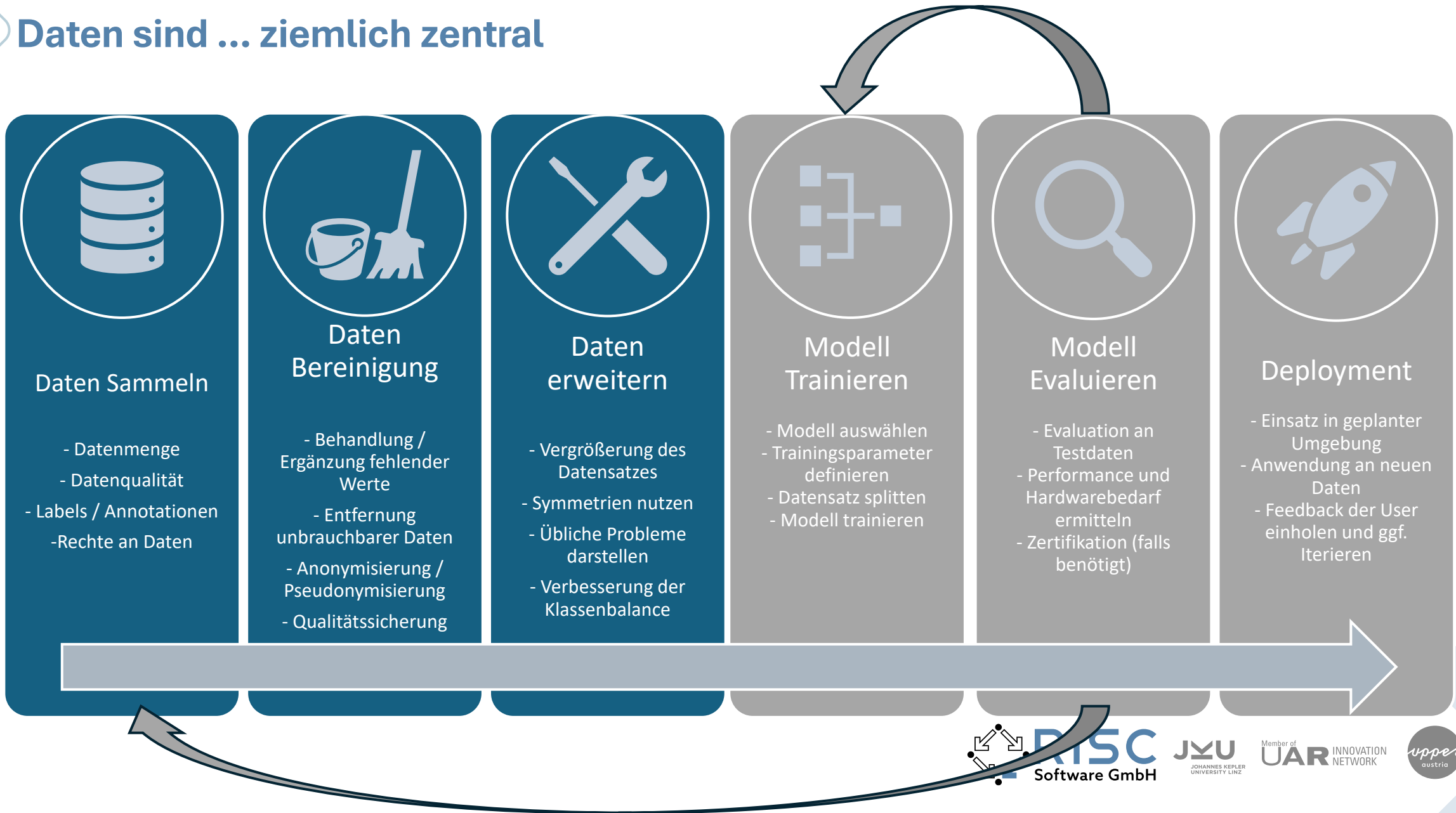


Inferenz

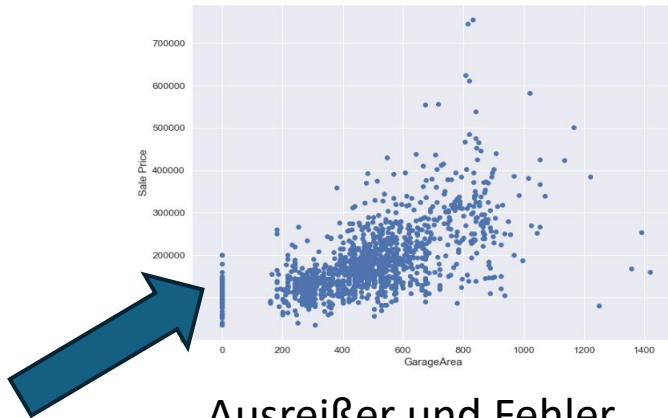
*90% Black
wins*

Training

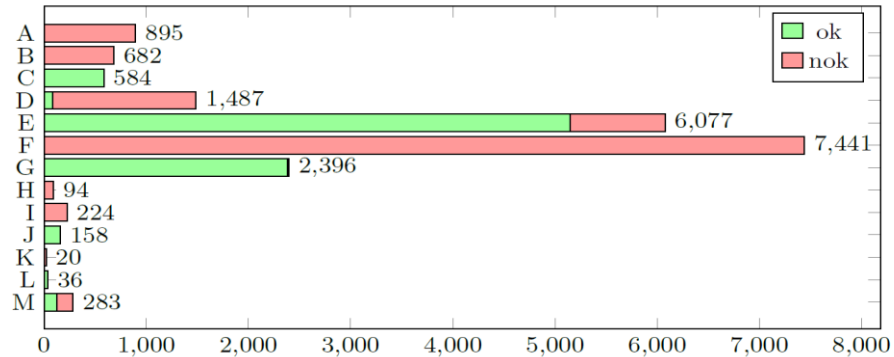
Daten sind ... ziemlich zentral



... und sind der Quell vieler Probleme



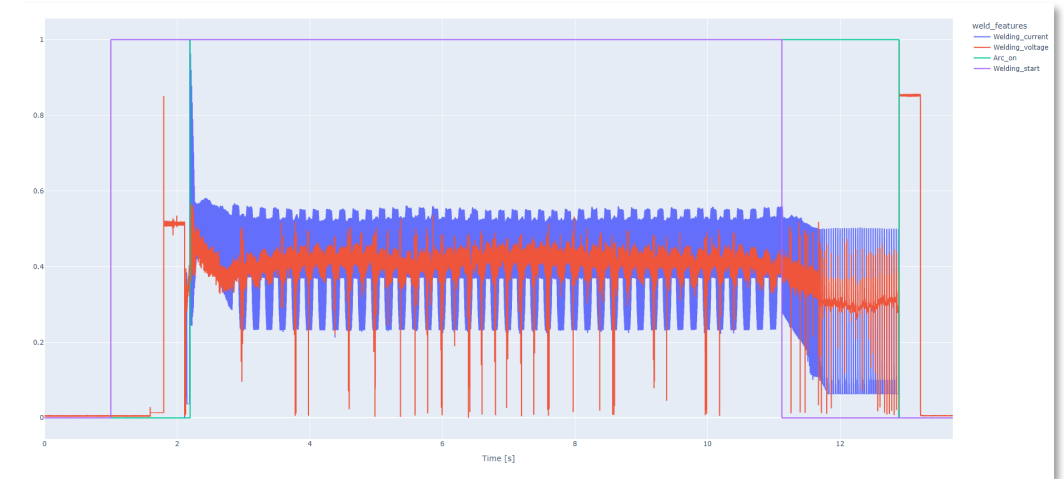
Ausreißer und Fehler



Overview: Distribution of defect labels

Unbalancierte Labels

205	Johann-Wilhelm-Klein Str.	25	Fehlende oder unstrukturierte Information
3758	Johann-Wilhelm-Klein-Straße	30	
1880	J.-W.-K. Straße	20	
289	J.-W.-K. <u>Strasse</u>	11	
735	J.-W.-Klein Straße	22	
671	Johann-W.-Klein Straße	12	

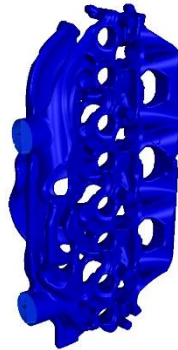


Überlagerung verschiedener Effekte

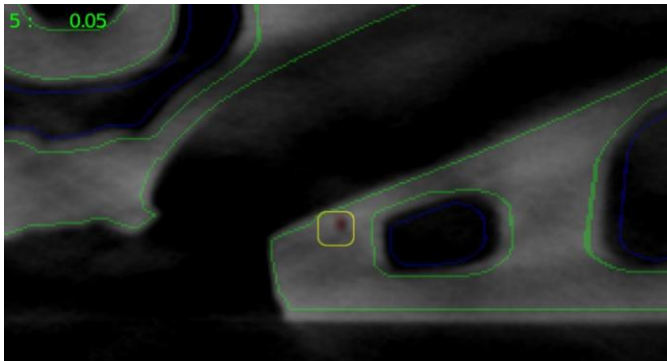
Serielle inline-XCT Prüfung bei Nemak



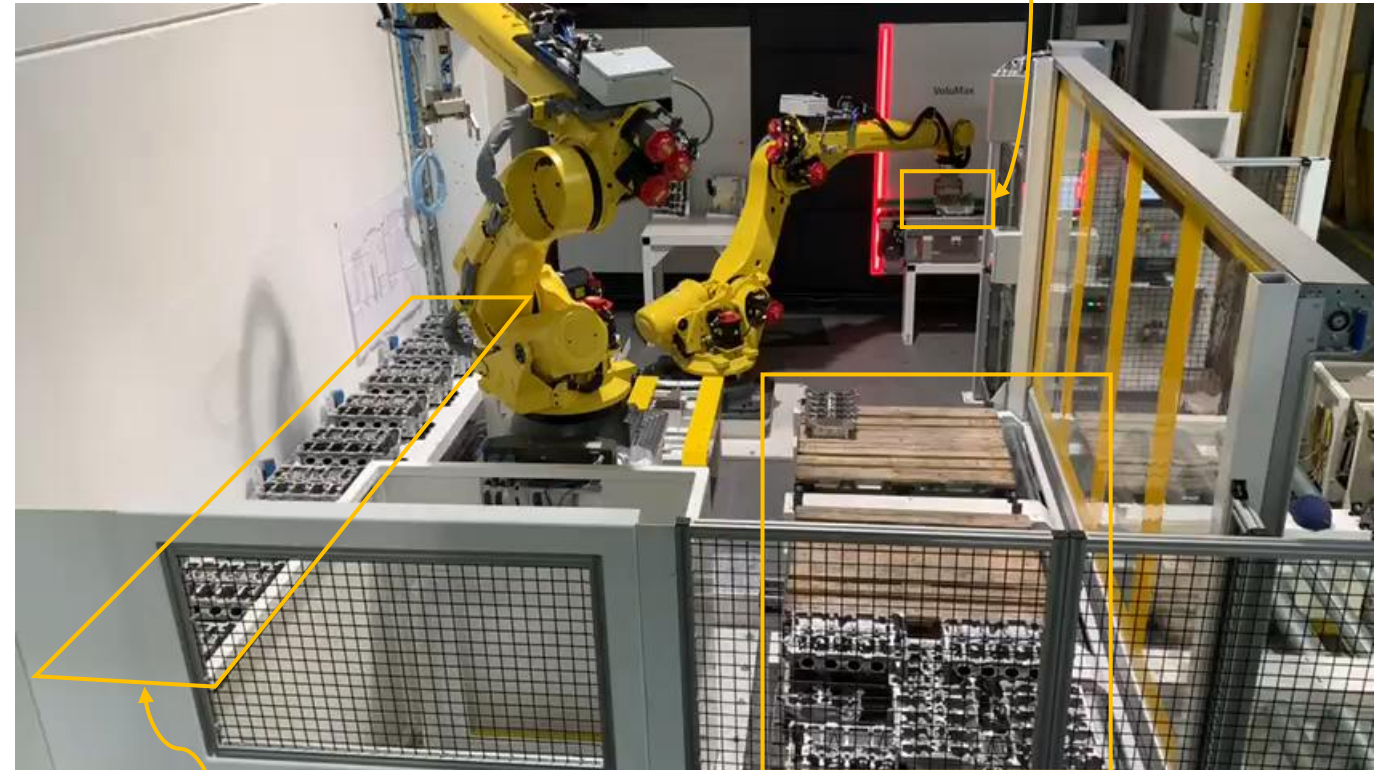
Raw material



Water jacket



Defect class: B

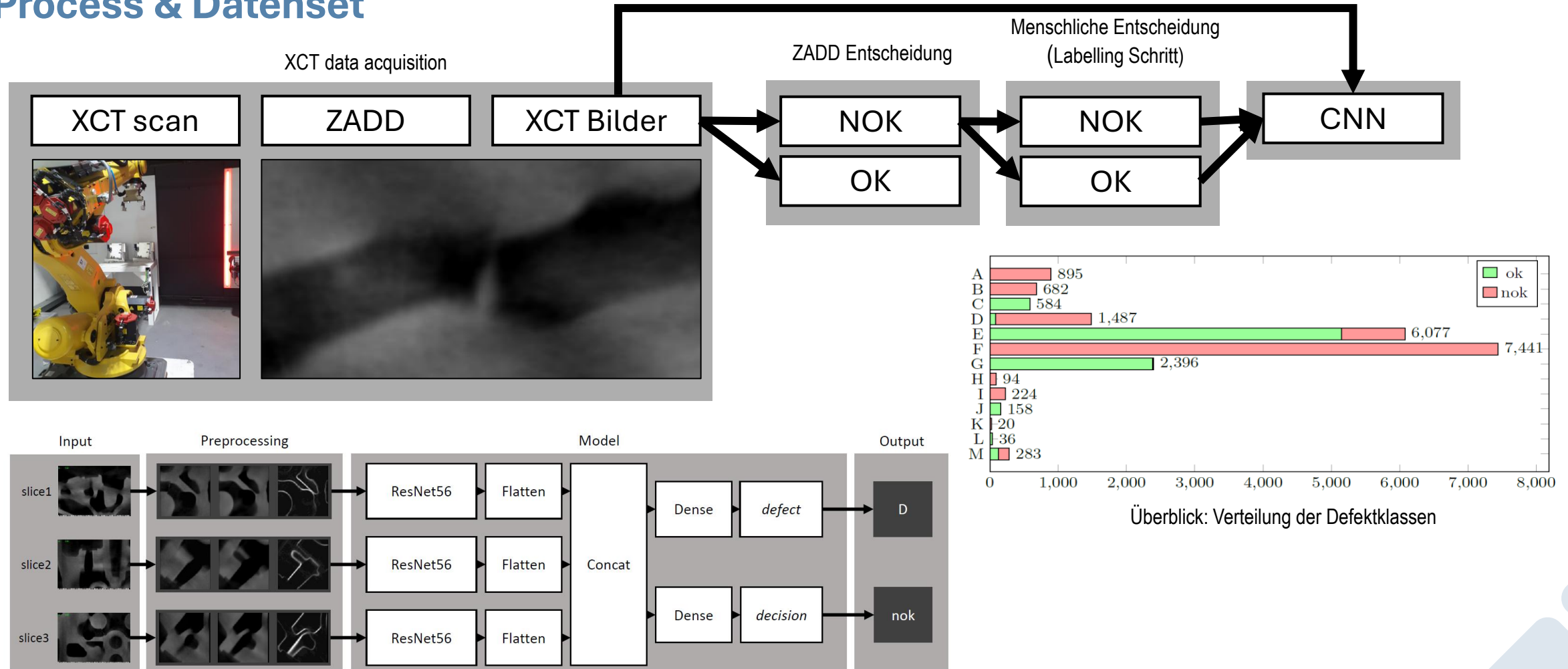


Neue Gussteile

Gescannte Teile (warten auf ZADD Analyse)

Gussteile, bereits durch ZADD analysiert

Process & Dataset



Training & post-processing

Datenvorverarbeitung & Training

- Es wurden die 7 häufigsten Klassen ausgewählt (A-G)
- Total 19,459 defekte (95.5%)

Accuracy	Training	Validation	Testing
defect	98%	92%	93%
decision	97%	93%	94%

Predicted-Defect	A	B	C	D	E	F	G	Total
Accuracy	98.6%	97.3%	100.0%	93.7%	72.5%	99.0%	85.6%	92.0%
FPS	0%	0%	0%	0%	3.2%	0%	14.4%	2.2%
CFPS	0%	0%	0%	0%	1.6%	0%	13.7%	1.8%
Overall Rate	6%	6%	11%	10%	17%	41%	9%	100%
Rate of TPs	0%	0%	11%	0%	11%	0%	9%	31%
Rate of TNs	5%	5%	0%	8%	3%	39%	0%	61%

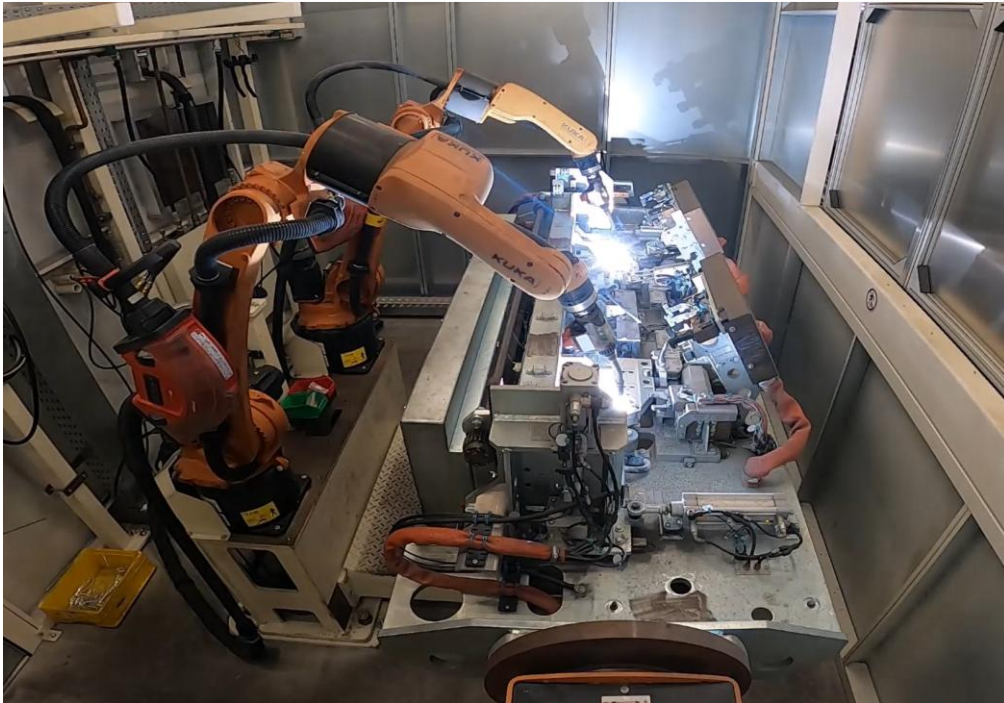
Gute Genauigkeit, aber...

- Weiterhin 2.9% *False-Positive* (Entscheidung ist "ok" sollte aber "nok" sein) Ergebnisse

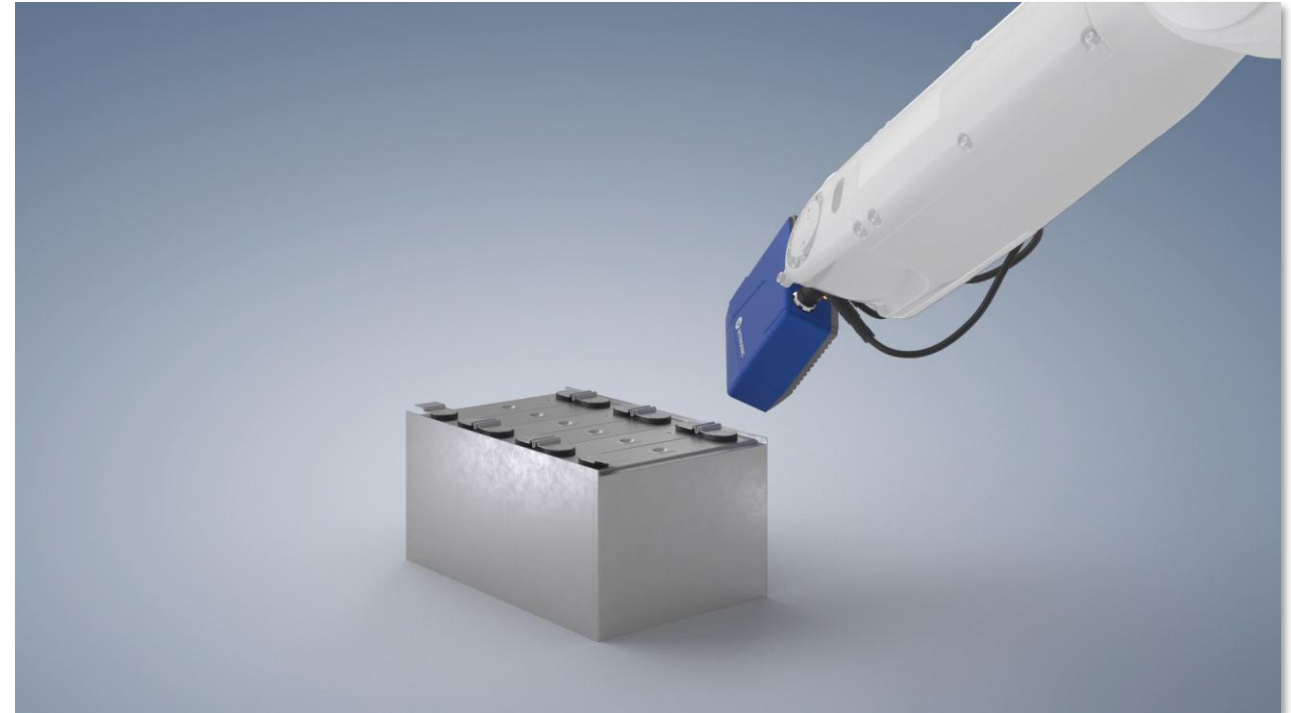
! Domänenwissen

- Manche Kombinationen von Defektklasse und OK/nOK Entscheidung sind nicht plausibel
- Vordefinierte Regeln** um False-Positives zu vermeiden
 - Klassen A, B und F wenn erkannt, werden immer als "nok" gelabelt.
 - False-Negatives* ("nok" gelabelt, sollten aber "ok" sein) sind weniger kritisch bei diesen Klassen

Projekt Metafacturing – Use Case Structural Welding



© BENTELER Automobiltechnik GmbH



© VITRONIC Dr.-Ing. Stein Bildverarbeitungssysteme GmbH

Probleme der Datenanalyse

Gute Datenanalyse erfordern hochqualitative Daten – das kostet Zeit

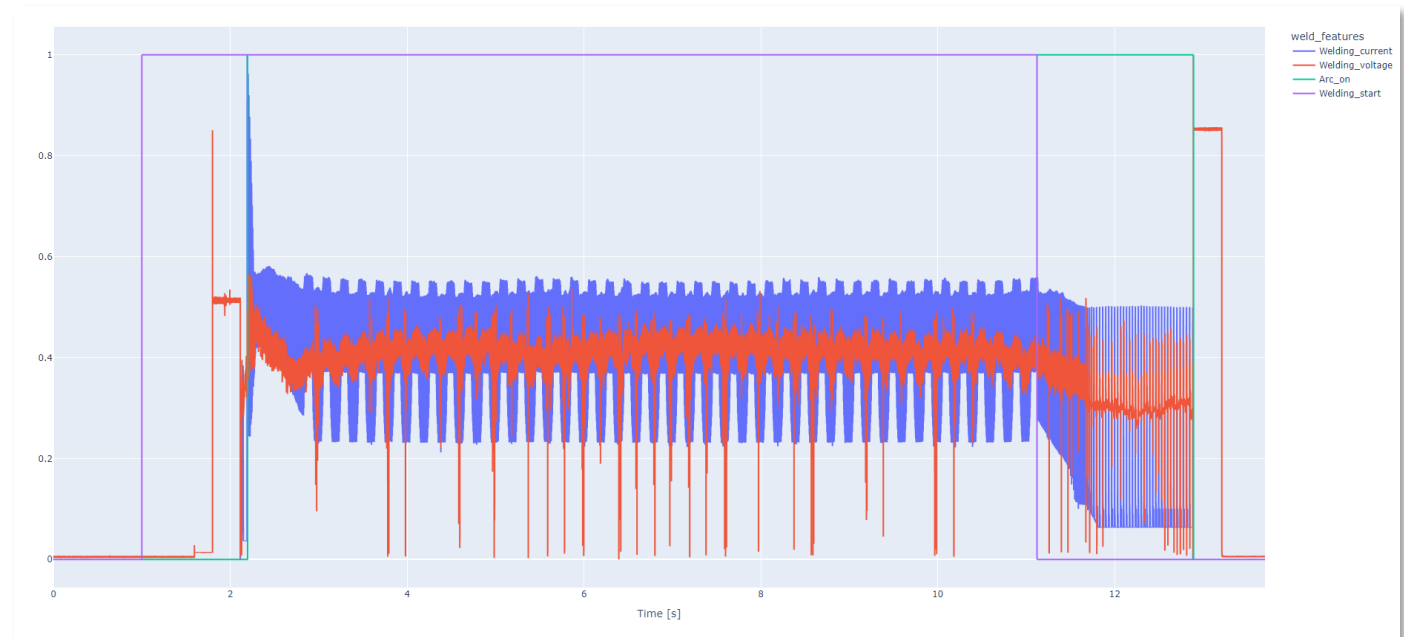
I. Analyse der Schweißdaten

- Keine Labels vorhanden
- Hochfrequente Zeitserie (sampling Rate ~ 0.1 ms) $\rightarrow$ VIELE Datenpunkte

*„Während einem Messtag können ‚ein paar‘ Messpunkte
millionen von Datensätzen werden“ (Benteler)*

II. Analyse der kombinierten Schweißdaten / Analysedaten

- Labels vorhanden



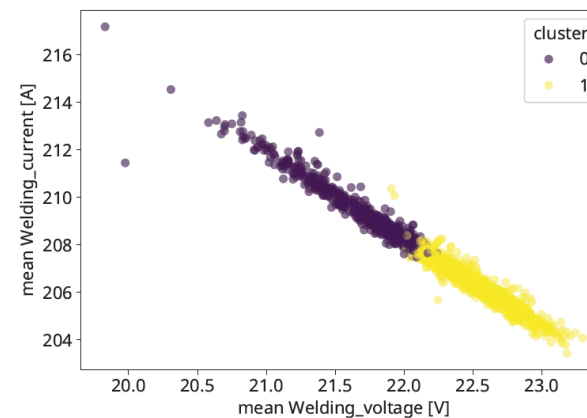
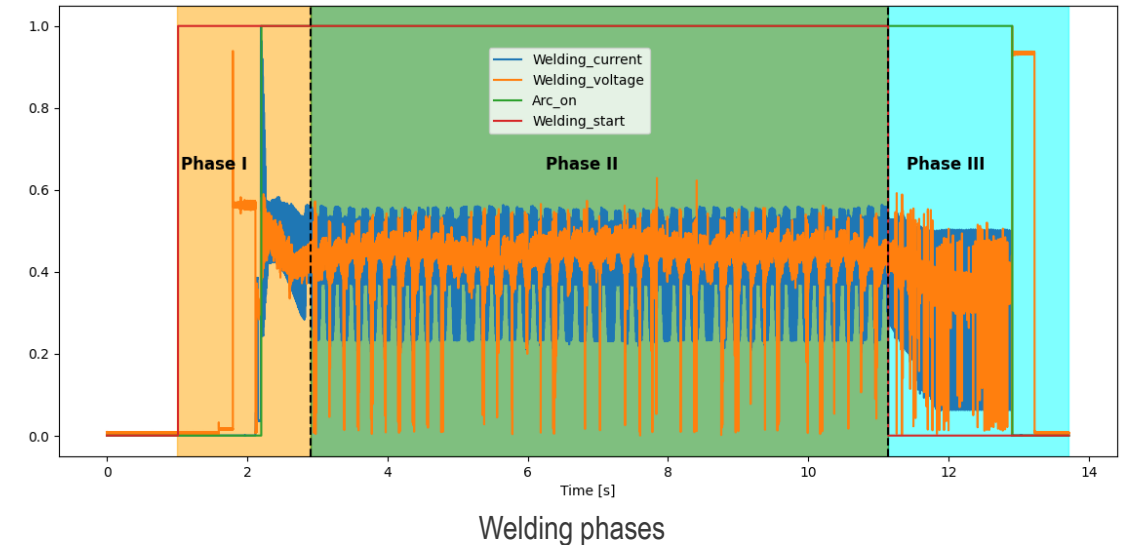
KI basierte Anomalieerkennung der Schweißdaten

Ergebnis

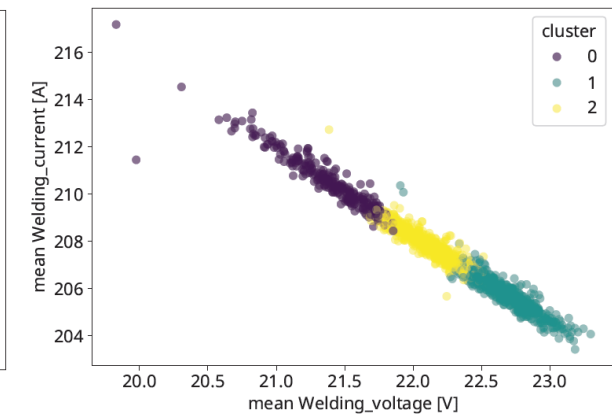
- Cluster analyse einer begrenzten Anzahl von Schweißdaten um zu ermitteln ob Ausreißer existieren
 - Identifikation bedeutender Frequenzspektren, welche die Qualität der Schweißnaht determinieren

Analyseworkflow

- Zu viele Datenpunkte für einfache Analyse (~1500)
 - Vorverarbeitung der Daten: Berechnung der Leistungsdichtespektren (Power Spectral Density, PSD) für jede Schweißnaht
- Anwendung von k-means-clustering auf mehrere Vorgehensweisen
 - Pipeline A → Datenset enthält vollständige Zeitserie
 - Pipeline B (Teilung der Zeitserie in Phasen) → **Phase II** des Schweißprozesses wird analysiert



(a)

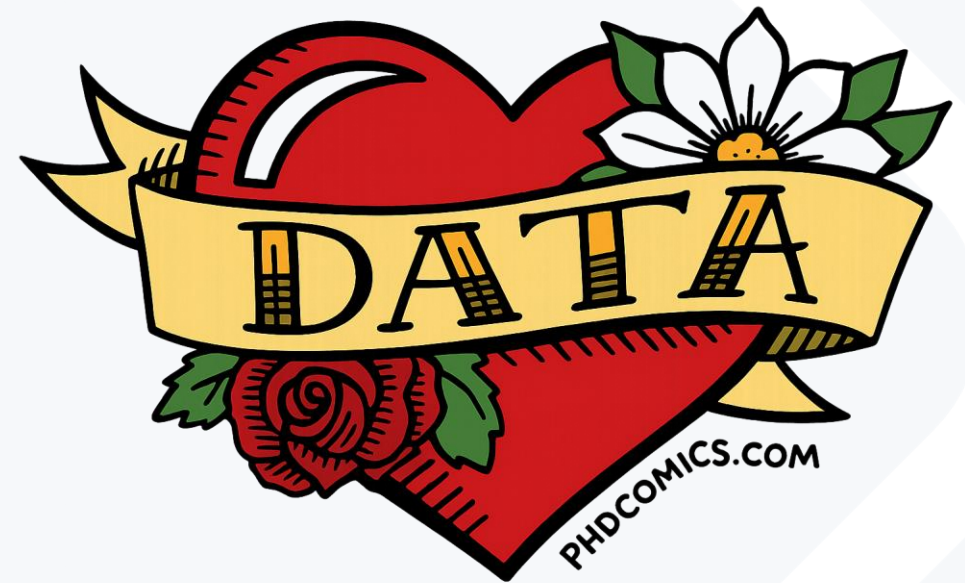


(b)

Scatter plots of mean *welding_current* and mean *welding_voltage* for Pipeline B: Results for (a) two clusters and (b) three clusters

Saubere Daten = Starke Modelle

- ◆ **Datensätze sind das Herz jedes Machine-Learning Prozesses**
→ Ohne Qualität keine Generalisierung
- ◆ **Vorverarbeitung entscheidet über Erfolg oder Scheitern eines Modells**
→ Rauschen, Ausreißer, Bias müssen vor jedem Training erkannt werden
- ◆ **Selbst bei perfekten Datensätzen bleibt Domänenwissen entscheidend**
→ Wer das Experiment versteht, erkennt Muster, die kein Algorithmus sieht
- ◆ **Die Zukunft industrieller KI liegt in der Verknüpfung von Datenqualität, Prozessverständnis und Modellkompetenz**



Danke!

Zeit für Fragen ... ?

