TKFDM Coffee Lecture:

Was kann KI und warum sind strukturierte Daten dafür wichtig?

Thüringer Zentrum für Lernende Systeme und Robotik

Oliver Mothes

23.02.2022

oliver.mothes@tzlr.de

Oliver Mothes

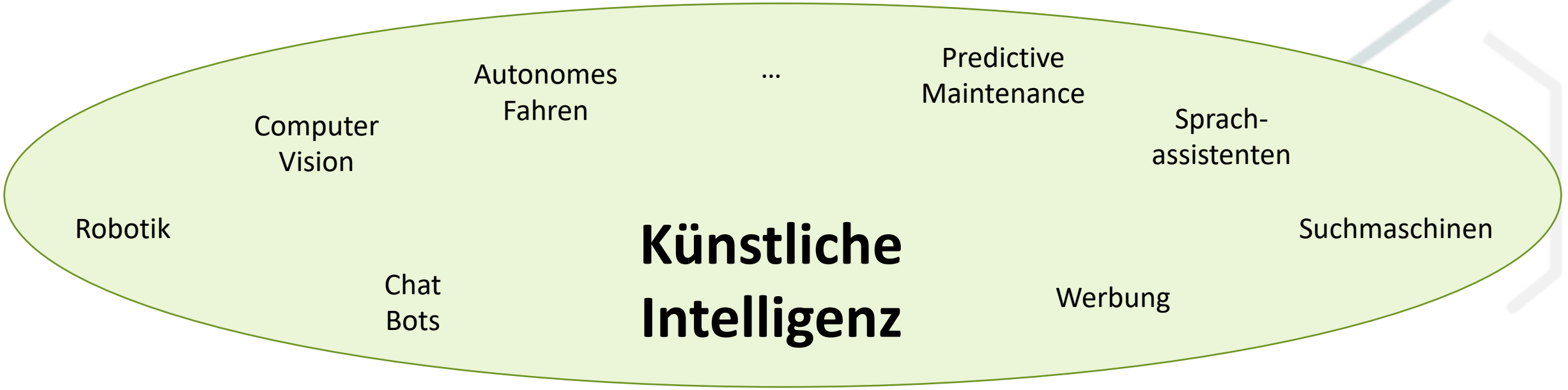
- Wissenschaftlicher Mitarbeiter und Doktorand in der **Computer Vision Group** der FSU Jena (Prof. Dr. Joachim Denzler)
  - Machine Learning / Deep Learning
  - Multiple Object Tracking
  - Anwendungen in verschiedenen biomedizinischen und industriellen Projekten
- Transferkoordinator Wissenschaft des **TZLR**
- Mitglied der AG **WISSENsAllmende** des Smart City Modellprojektes Jena
- Organisator des Jenaer KI-Stammtiches **JENA.AI**

Agenda

- **Was ist KI ?**
  - Definition, Einordnung, Teilgebiete
- **Was kann KI ?**
  - Die Rolle von Daten, Datenqualität, Aus Daten lernen
- **Wie kann (Forschungs-)Datenmanagement die KI unterstützen ?**
  - Wichtige Aspekte aus Sicht von Datenwissenschaftlern

Was ist KI?

→ “Teilgebiet der Informatik, das sich mit der Automatisierung *intelligenten* Verhaltens und dem maschinellen Lernen befasst“ (Quelle: Wikipedia)

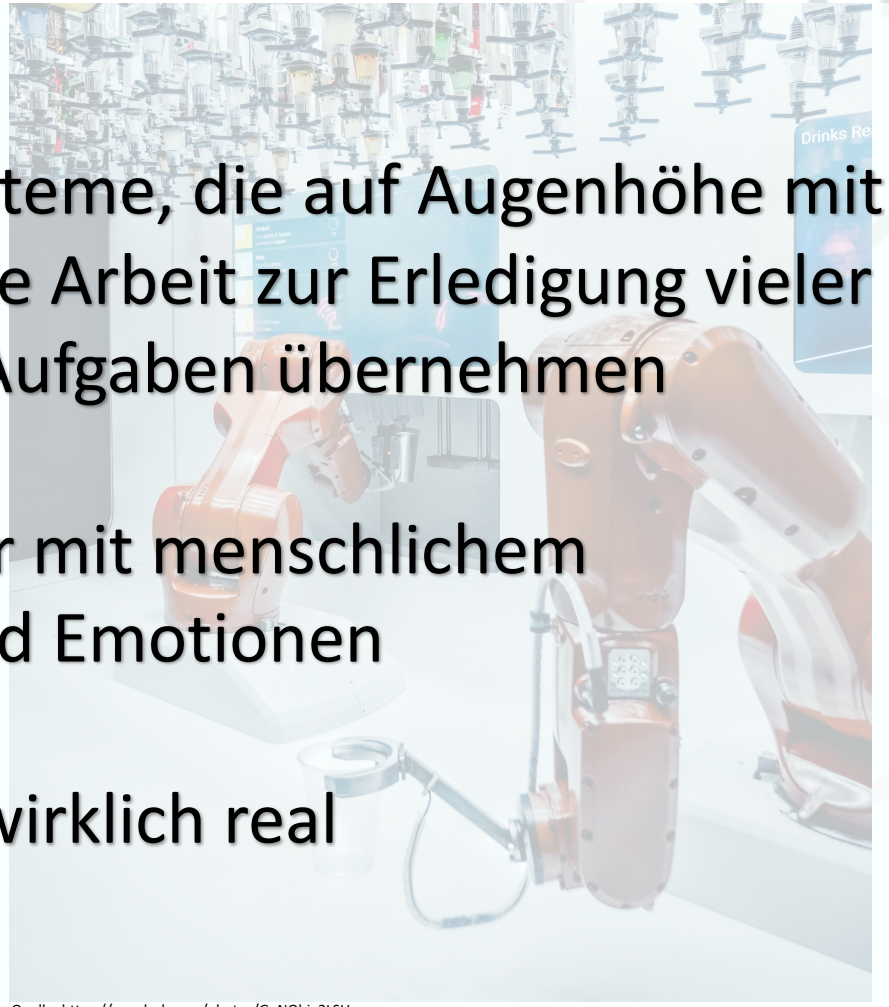


Starke KI vs. Schwache KI

Quelle: https://unsplash.com/photos/fv1EFjgIb94

- Computersysteme, die auf Augenhöhe mit Menschen die Arbeit zur Erledigung vieler schwieriger Aufgaben übernehmen
- Bsp.: Roboter mit menschlichem Verhalten und Emotionen

→ noch nicht wirklich real



Quelle: https://unsplash.com/photos/GpNOhig3LSU

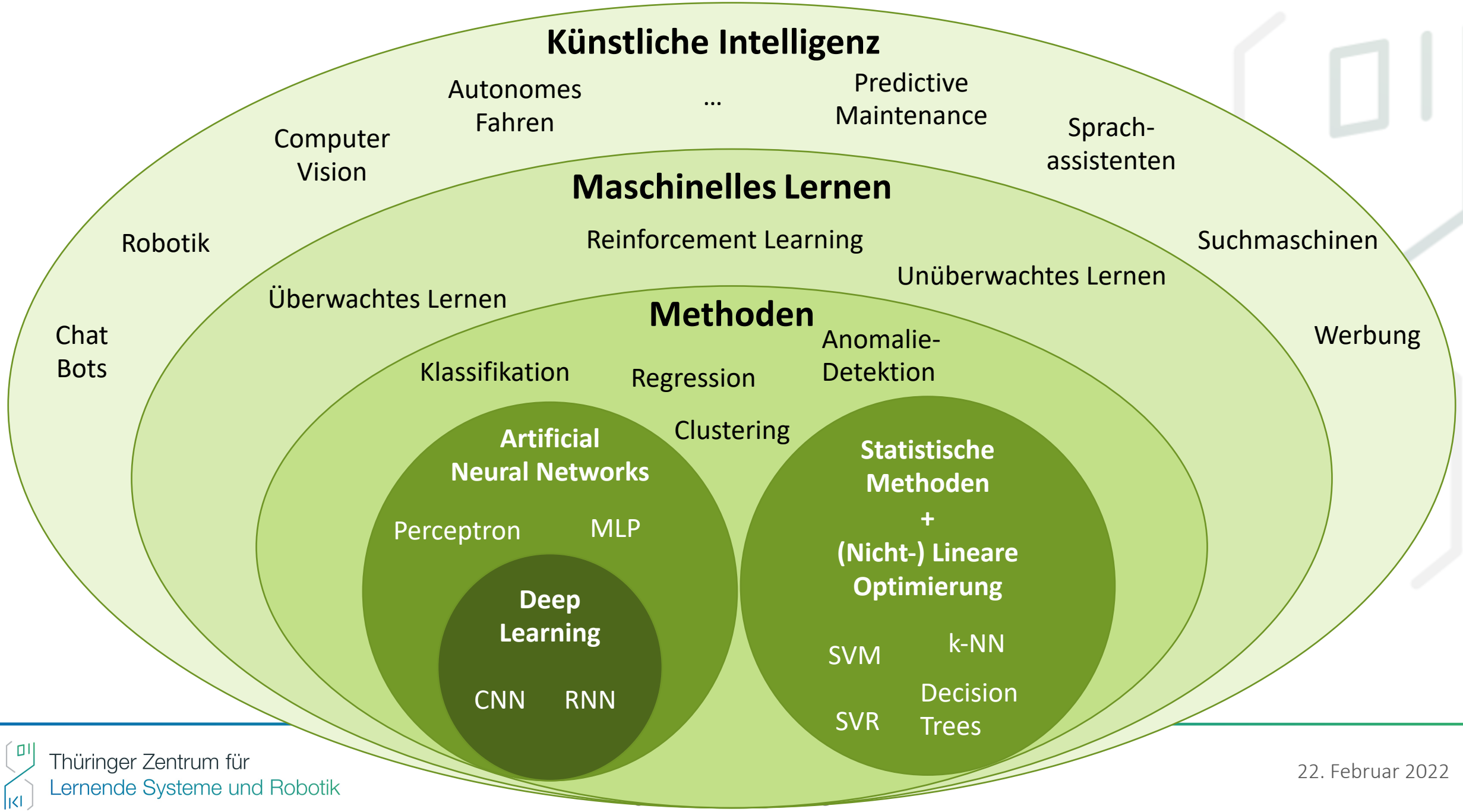
Starke KI vs. Schwache KI

- Begrenzt auf die gleichzeitige Lösung eines oder weniger Anwendungsprobleme
  → leistungsstark
- Bsp.: Siri, Gesichtserkennung, Serviceroboter

→ **Maschinelles Lernen** als Lösungsansatz

Quelle: https://unsplash.com/photos/fv1EFjgIb94

Quelle: https://unsplash.com/photos/GpNOhig3LSU

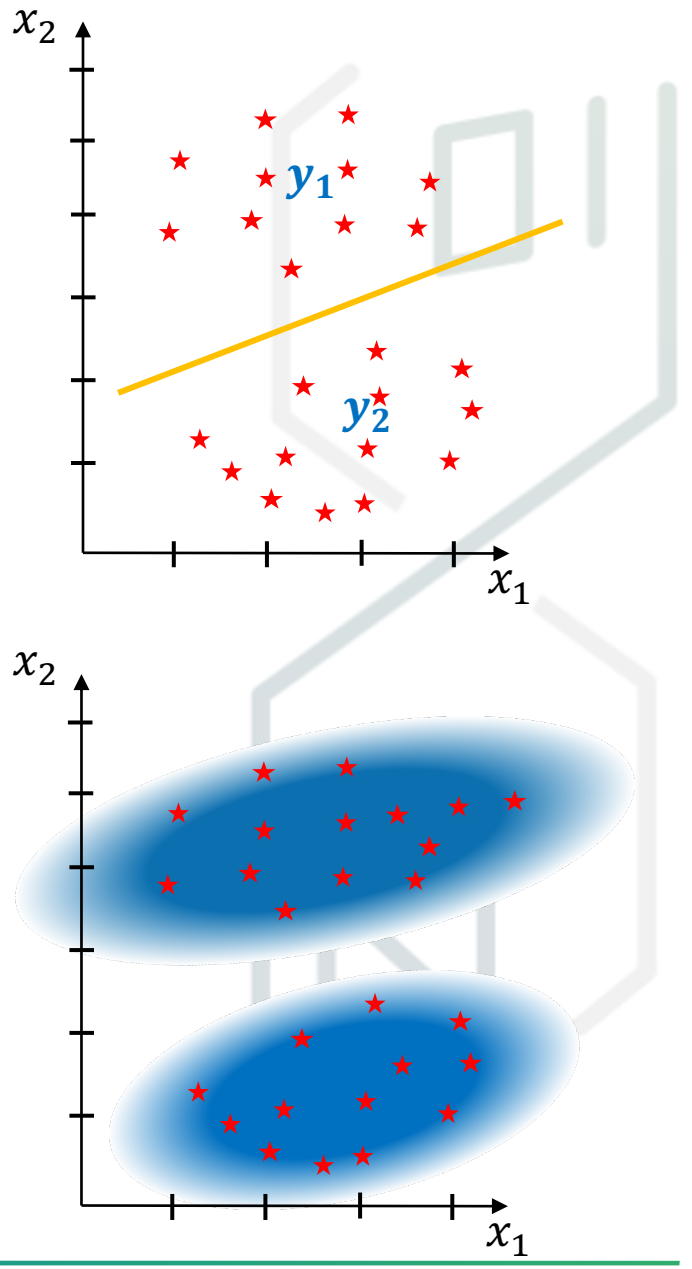

Künstliche Intelligenz
Autonomes Fahren
...
Predictive Maintenance
Computer Vision
Sprach-assistenten
Maschinelles Lernen
Robotik
Reinforcement Learning
Suchmaschinen
Unüberwachtes Lernen
Überwachtes Lernen
Methoden
Chat Bots
Anomalie-Detektion
Werbung
Klassifikation
Regression
Clustering
Artificial Neural Networks
Statistische Methoden + (Nicht-) Lineare Optimierung
Perceptron
MLP
Deep Learning
k-NN
SVM
Decision Trees
CNN
RNN
SVR

Was kann KI?

- **Überwachtes** Lernen aus Daten $\boldsymbol{x}$ zusammen mit Expertenwissen $\boldsymbol{y}$

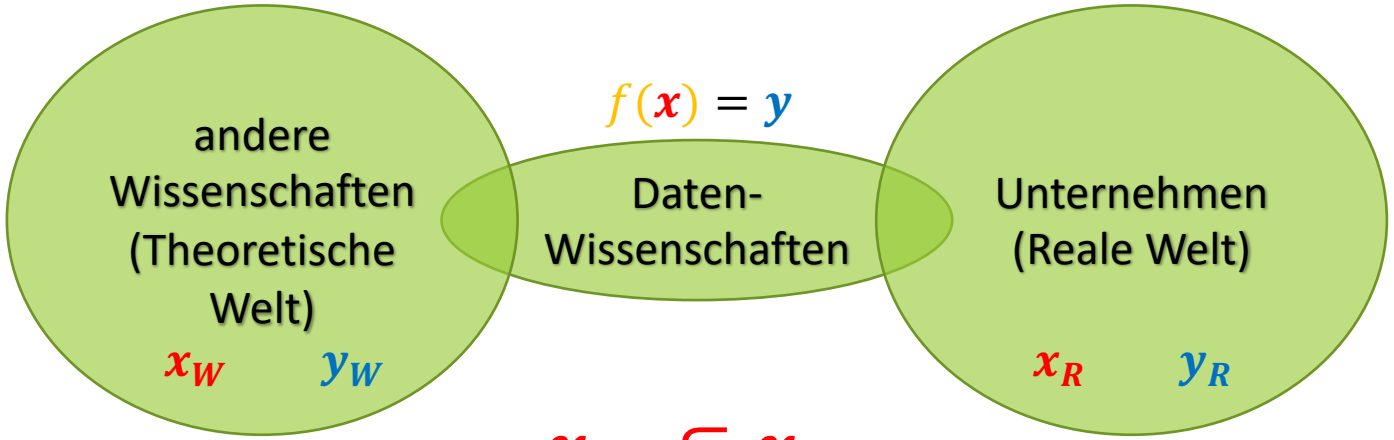
$$f(\boldsymbol{x}) = \boldsymbol{y}$$

- **Unüberwachtes** Lernen aus Daten $\boldsymbol{x}$ ohne Expertenwissen $\boldsymbol{y_1}$ und $\boldsymbol{y_2}$



Daten vs. Daten

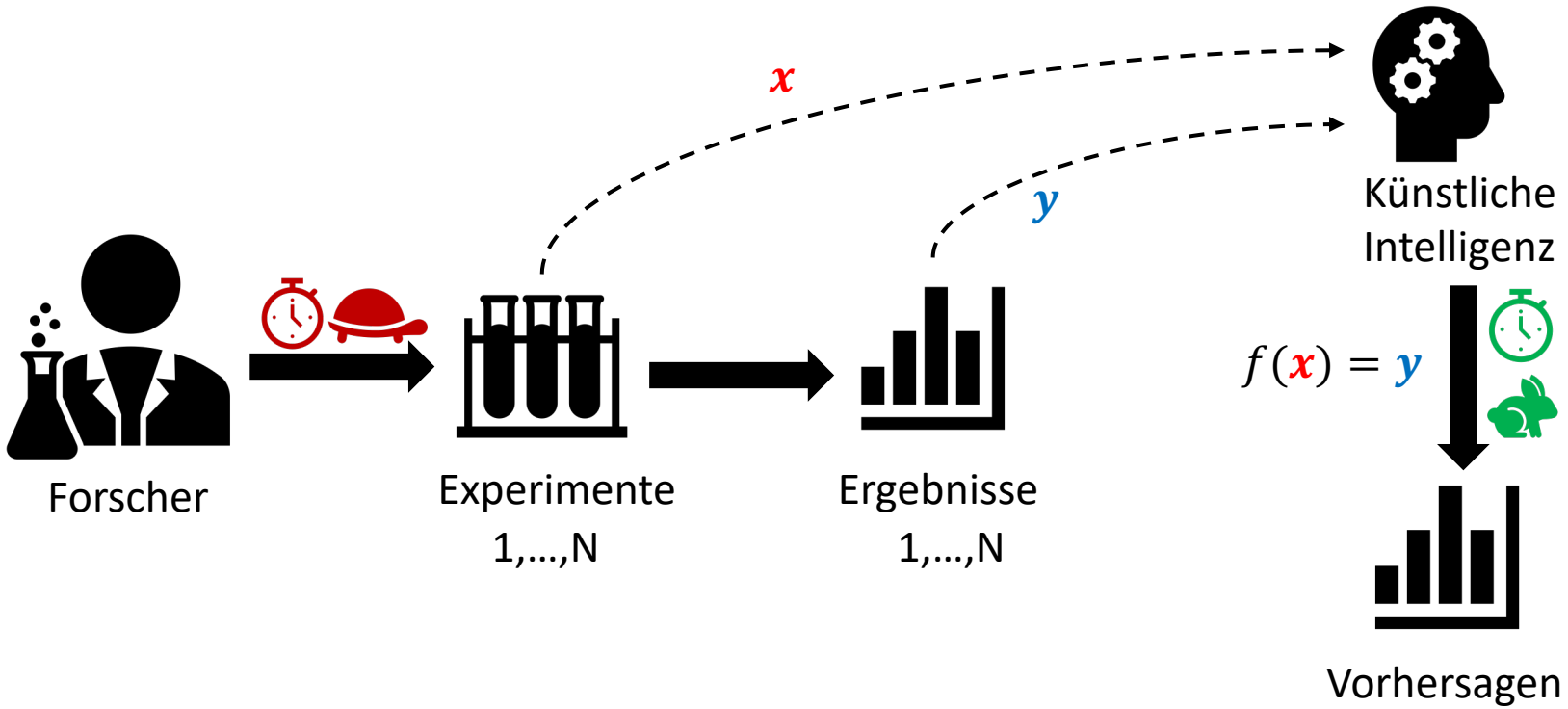
- Sind die **„wissenschaftliche“** Daten vergleichbar zu **„realen“** Daten ?



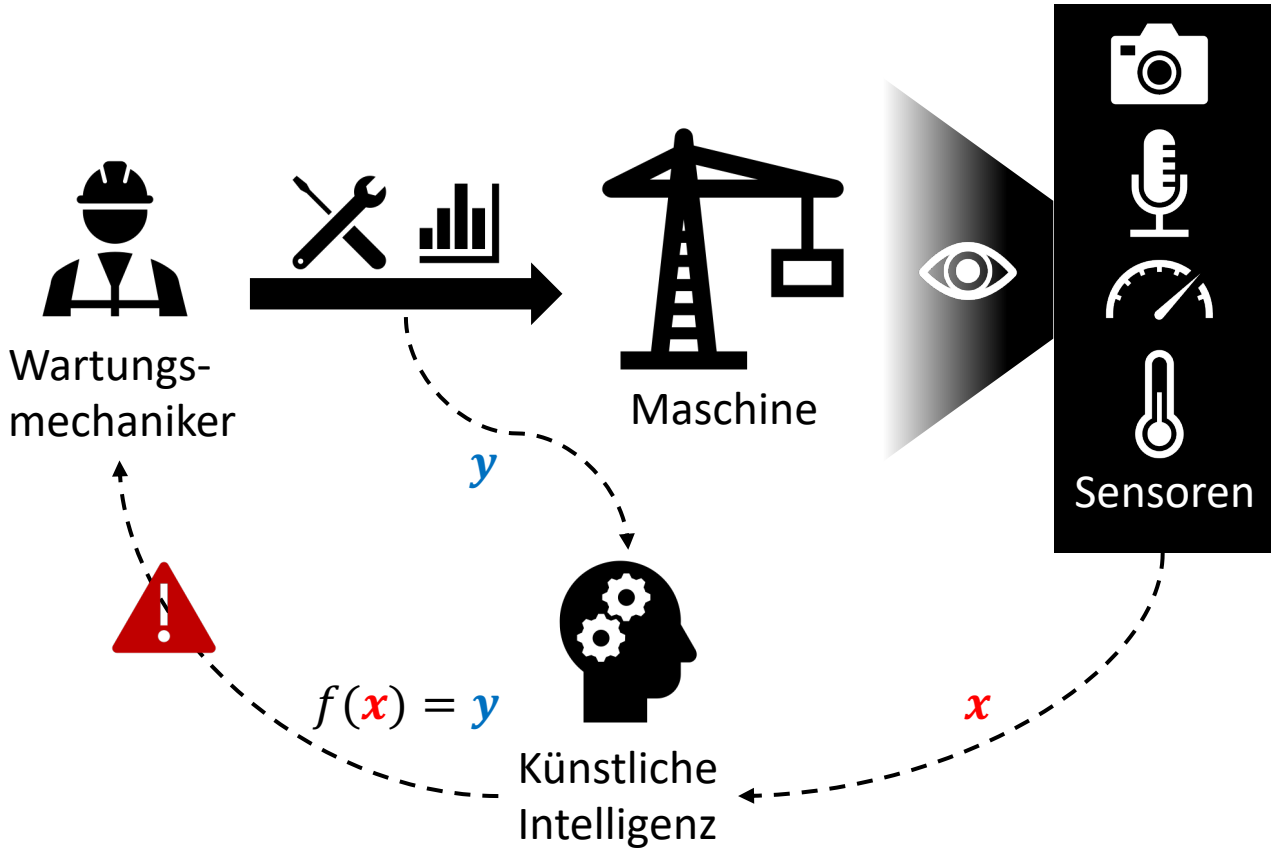
$$\boldsymbol{x_W} \subseteq \boldsymbol{x_R}$$
$$\boldsymbol{y_W} \subseteq \boldsymbol{y_R}$$

- Daten $\boldsymbol{x}$ entstehen „von allein“, die Annotationen $\boldsymbol{y}$ können teuer sein

Ein Beispiel aus der Wissenschaft

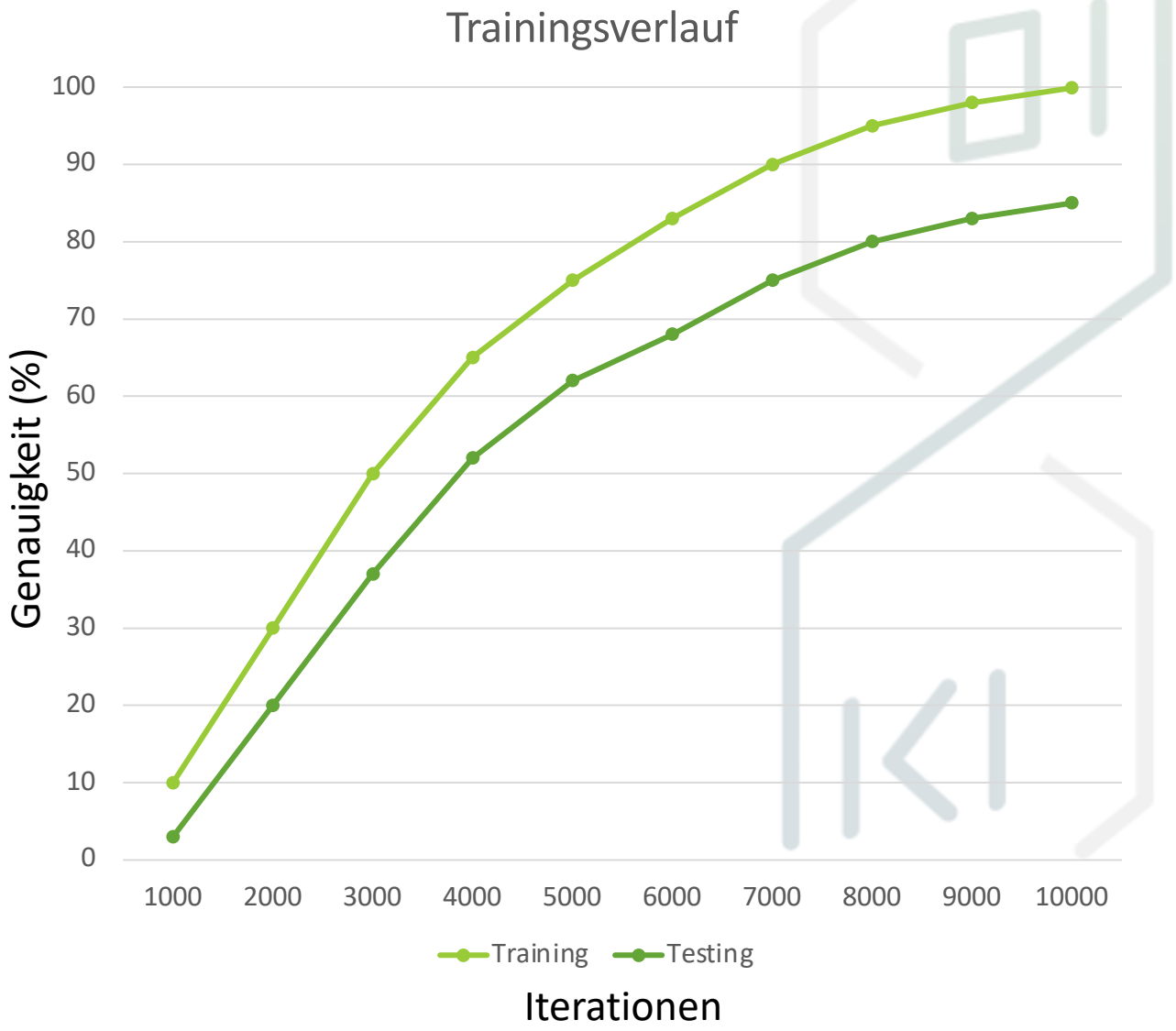


Ein Beispiel aus der Industrie



Aus Daten lernen

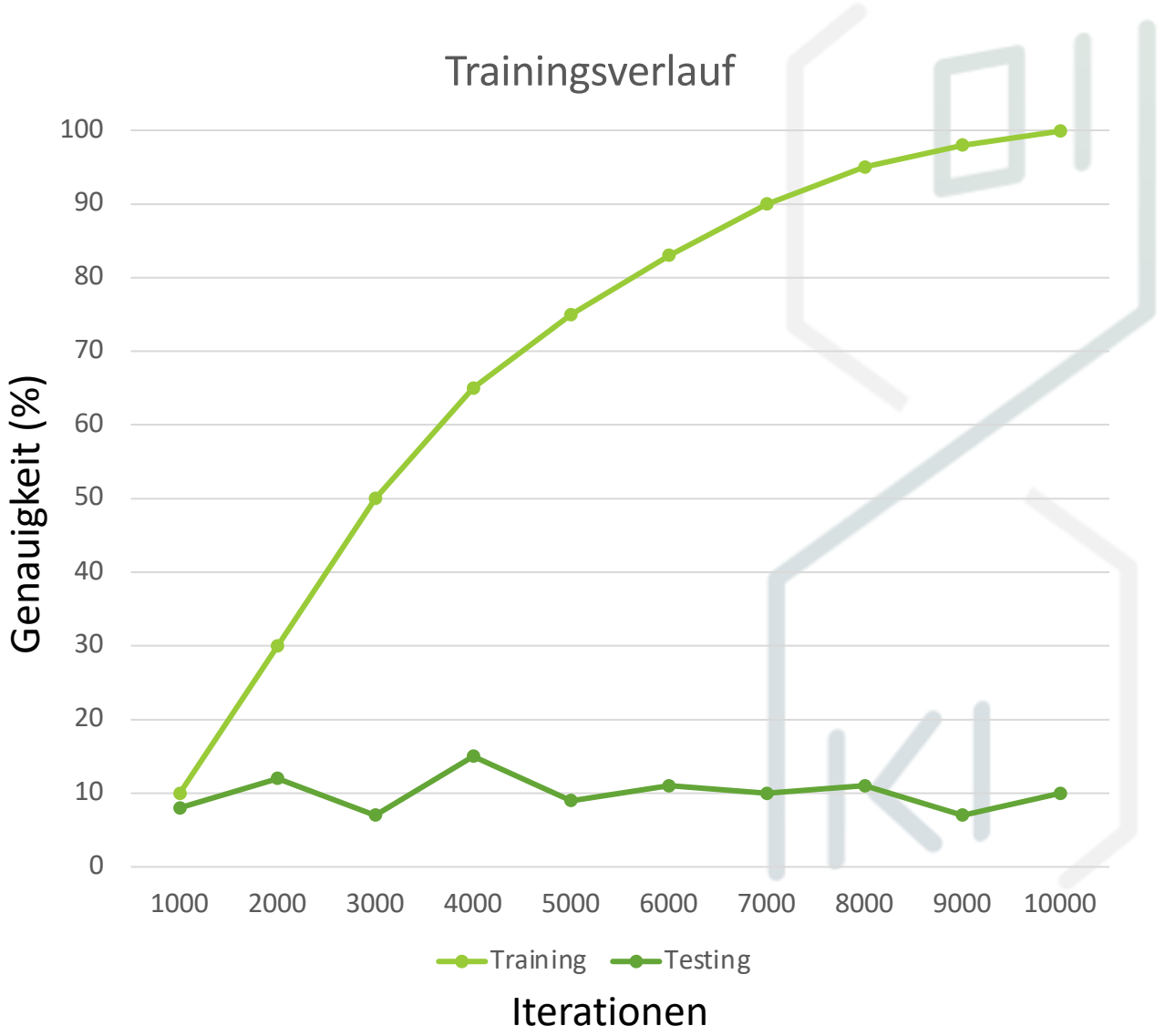
- Generalisierung
  - Modell „erfüllt“ Aufgabe an ungesehen Daten (Testdaten)
- **ACHTUNG**:
  - Testdaten zu ähnlich zu Trainingsdaten?
  - Entsprechen die Testdaten den „realen“ Daten?



Aus Daten lernen

- Data Splitting
  - 80 % Training
  - 10 % Testing
  - 10 % Validation

- Überanpassung (*Overfitting*)
  - Modell lernt Trainingsdaten „auswendig“
  - generalisiert nicht
  - Am Beispiel: 10 Klassen



Wie kann (Forschungs-) Datenmanagement die KI unterstützen?

- Einhaltung der **FAIR**-Prinzipien
  - **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
  - Konsistenz in der Datenstruktur
- Jede Information zu Daten sind hilfreich
- Statistiken zu erhobenen Datensätzen

  - WARUM SIND EINFACHE STATISTIKEN ZU DEN DATEN WICHTIG ?

  → 2 Beispiele

Daten und ihre Verteilung

- Balancierte Daten
  - Beispiel 1: Häufigkeit der Buchstaben in einem deutschen Text (8.953.540 Zeichen)

Zeichen	A	B	C	D	E	…	Z	…	Ä	Ö	Ü
Häufigkeit (%)	5,58	1,96	3,16	4,98	16,93		1,211		0,51	0,30	0,65

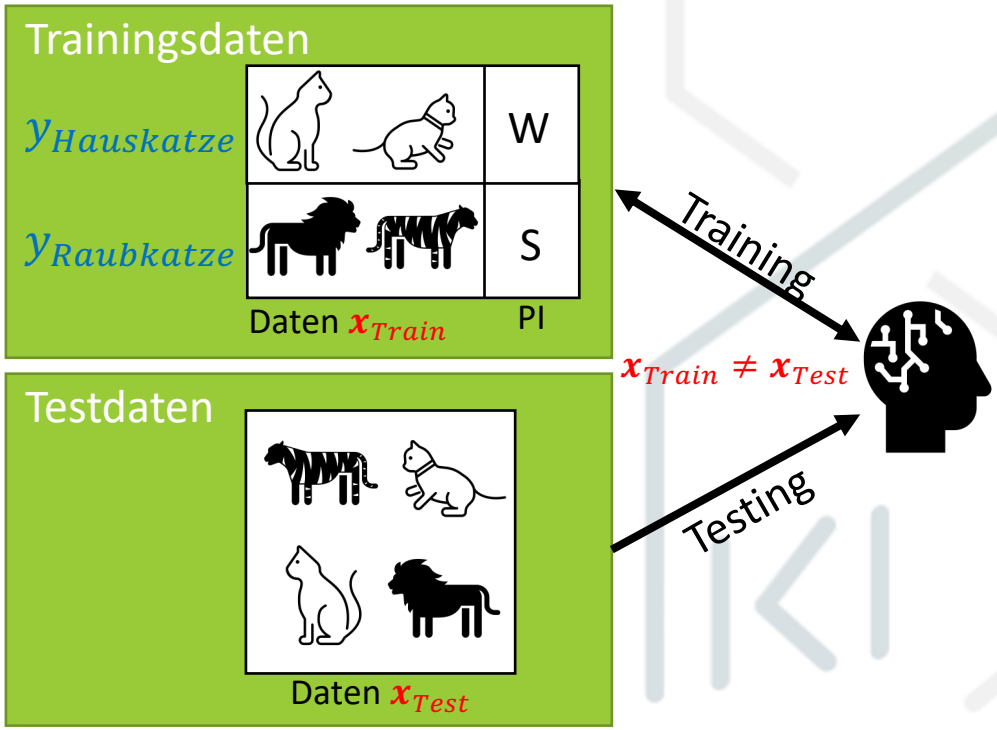
  - Beispiel 2: „gute“ und „schlechte“ Produktionserzeugnisse

Zustand	gut	schlecht
Häufigkeit (%)	99,99	0,01

→ Wichtig für die Modellwahl

Wie kann (Forschungs-) Datenmanagement die KI unterstützen?

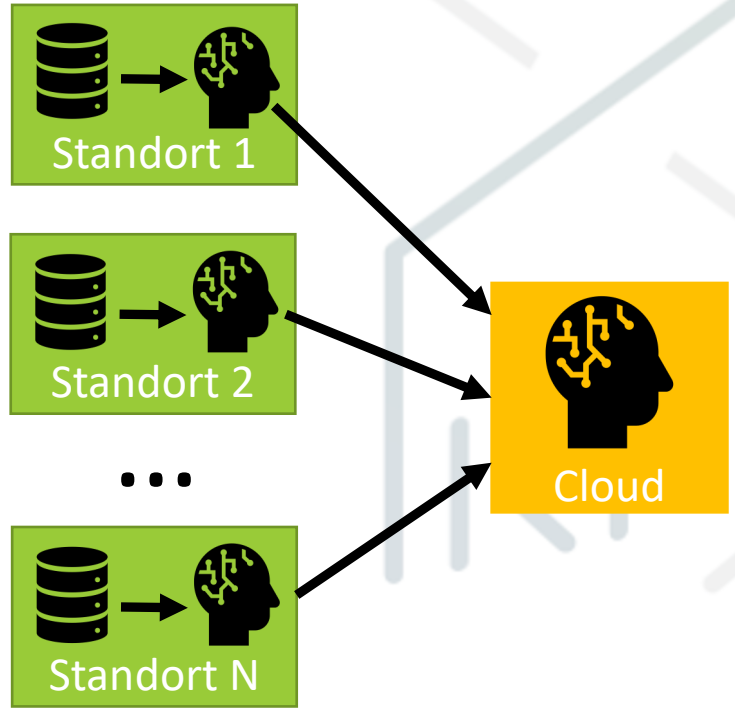
- **Learning using Privileged Information (LUPI)**
  - Metadaten zusätzlich als Privileged Information (PI) nutzen
  - PI unterstützt beim Trainieren des Modells



Vladimir Vapnik, Akshay Vashist, A new learning paradigm: Learning using privileged information, Neural Networks, Volume 22, Issues 5–6, 2009, Pages 544-557,

Wie kann (Forschungs-) Datenmanagement die KI unterstützen?

- **Federated Learning**
  - (Sensible) Daten dezentral kollaborativ nutzen
  - Bsp.: nicht-anonymisierbare Patientendaten (Gesichtsbilder)



Konečný, Jakub; McMahan, Brendan; Ramage, Daniel (2015). "Federated Optimization: Distributed Optimization Beyond the Datacenter". arXiv:1511.03575

Wie kann (Forschungs-) Datenmanagement die KI unterstützen?

- **Weitere wichtige Punkte:**
  - Reproduzierbarkeit von KI-Methoden
  - Verstehen von KI-Methoden
  - uvm.

Vielen Dank für Ihre Aufmerksamkeit!

oliver.mothes@tzlr.de