



Technische Hochschule Georg Agricola

INDUSTRIEAUTOMATION

DATA SCIENCE IN DER PRODUKTION

Bochum

Prof. Dr.-Ing. Lars N. Josler

AGENDA



1. Einordnung von Data Science für Ingenieure
2. Datenspeicherung in Datenbanken
3. Datenverarbeitung in Python
- 4. Machine Learning Basics**
- 5. Ablauf einer Machine Learning Pipeline**
- 6. Anwendungsbeispiel im Ingenieurskontext**

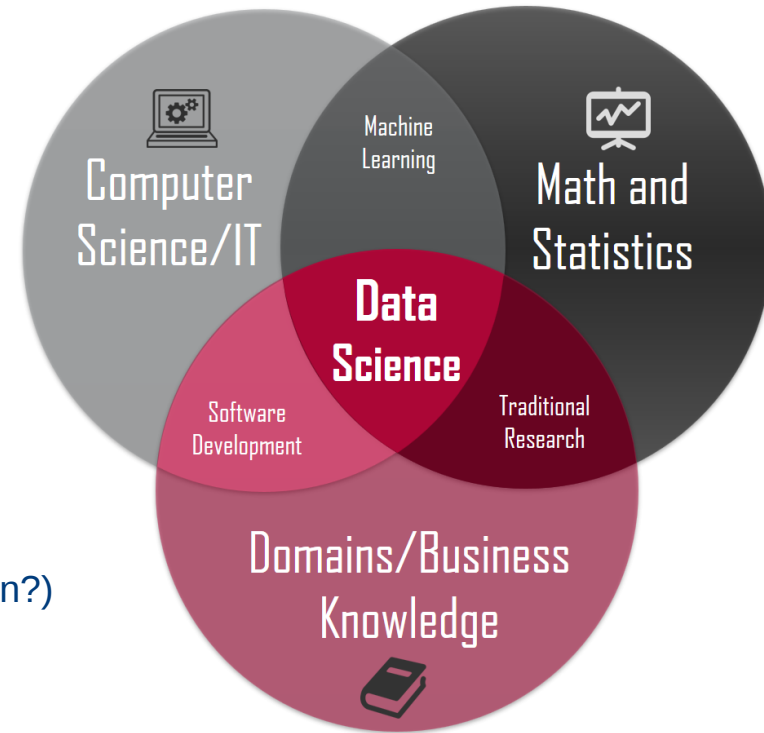
AGENDA



1. Einordnung von Data Science für Ingenieure
2. Datenspeicherung in Datenbanken
3. Datenverarbeitung in Python
- 4. Machine Learning Basics**
5. Ablauf einer Machine Learning Pipeline
6. Anwendungsbeispiel im Ingenieurskontext



- Generierung von Wissen aus Daten
- *"use the most powerful hardware, the most powerful programming systems, and the most efficient algorithms to solve problems"*
(Lescovec et al., Mining of Massive Datasets)
- **Beispiele:**
 - Text- und Bilderkennung
 - Churn Prediction (Wird der Kunde den Vertrag verlängern?)
 - Empfehlungssysteme (Spotify, Netflix, Amazon)
 - Predictive Maintenance



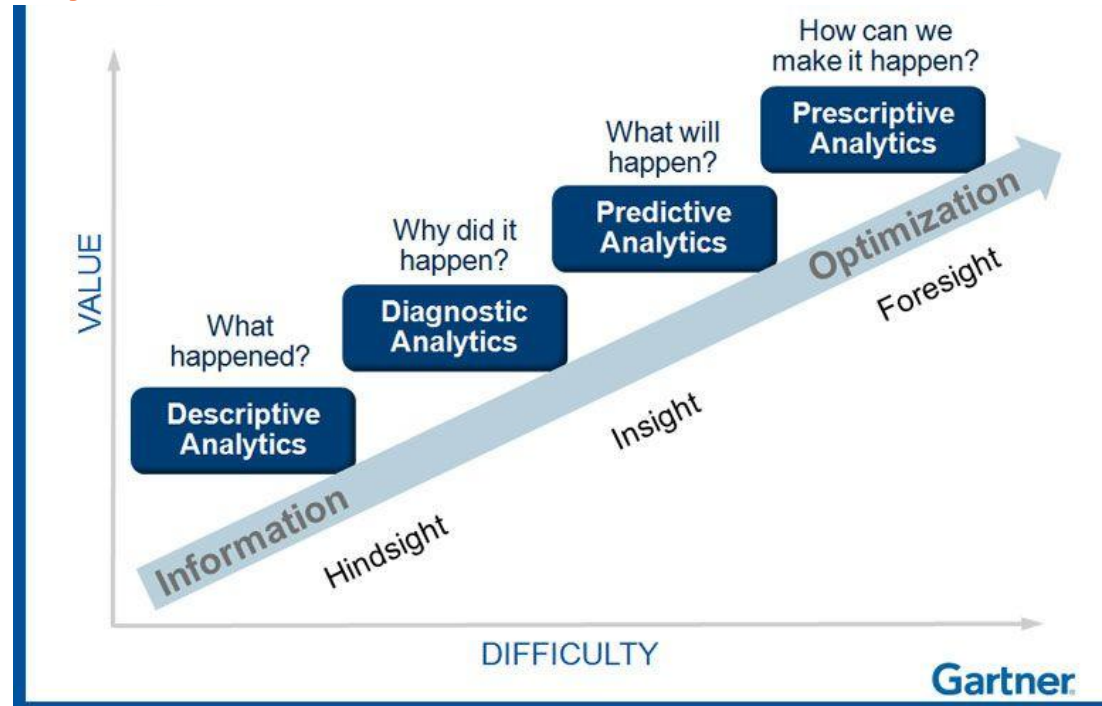
DATA SCIENCE VS. DATA ANALYTICS



Advanced Analytics

Data Analytics

Data Science



<https://i.pinimg.com/originals/04/df/66/04df66a10529b687d7799429afe9df4a.jpg>



- **Analyse von Unternehmensdaten**
 - Datensammlung
 - Datenaggregation
 - KPI-Darstellung für Dashboards und Reporting
- **Software für Dashboarding**



Power BI



Sense



+a|b|eau



Beispiel Dashboard

https://upload.wikimedia.org/wikipedia/commons/b/b7/Infroid%27s_Self-Service_BI_Tool_Dashboard.jpg

EINLEITUNG MASCHINELLES LERNEN

GRUNDLAGEN

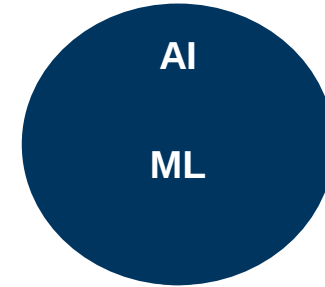


- **Artificial Intelligence (Künstliche Intelligenz)**

- „smarte“ Problemlösung durch Maschinen
 - Wissensdarstellung
 - Sprachanalyse
 - (Ablauf-)Planung

- **Maschinelles Lernen**

- Teilgebiet der Artificial Intelligence
- Mathematische Modellbildung für die Extraktion von Wissen aus Daten und Erfahrungen
- Verwendung von statistischen Methoden
- Erzeugung künstlicher Intelligenz ohne explizite Programmierung (engl. *hardcoding*)
- Unterteilung in Supervised Learning, Unsupervised Learning und Reinforcement Learning



WAS IST MASCHINELLES LERNEN

DEFINITION



“Machine Learning is the study of computer algorithms that improve automatically through experience.”

[Machine Learning; Mitchel, Tom; 1997]

EINLEITUNG MASCHINELLES LERNEN

GRUNDLAGEN



- **Artificial Intelligence (Künstliche Intelligenz)**

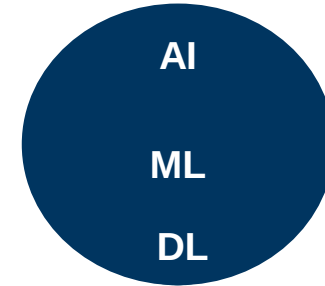
- „smarte“ Problemlösung durch Maschinen
 - Wissensdarstellung
 - Sprachanalyse
 - (Ablauf-)Planung

- **Maschinelles Lernen**

- Teilgebiet der Artificial Intelligence
- Mathematische Modellbildung für die Extraktion von Wissen aus Daten und Erfahrungen
- Verwendung von statistischen Methoden
- Erzeugung künstlicher Intelligenz ohne explizite Programmierung (engl. *hardcoding*)
- Unterteilung in Supervised Learning, Unsupervised Learning und Reinforcement Learning

- **Deep-Learning**

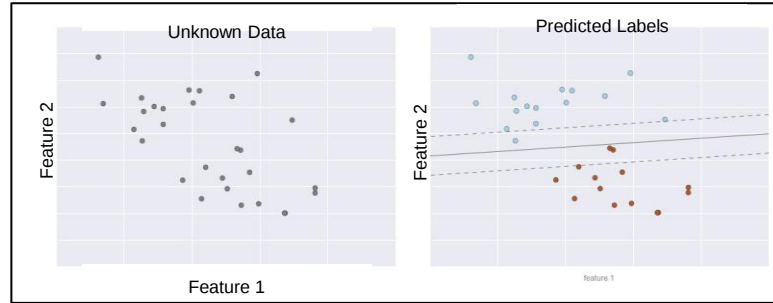
- Verwendung von Neuronalen Netzen mit mehreren „versteckten“ Schichten



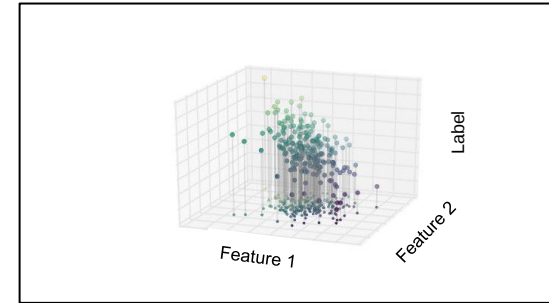


- **Motivation:**

Klassifikation



Regression



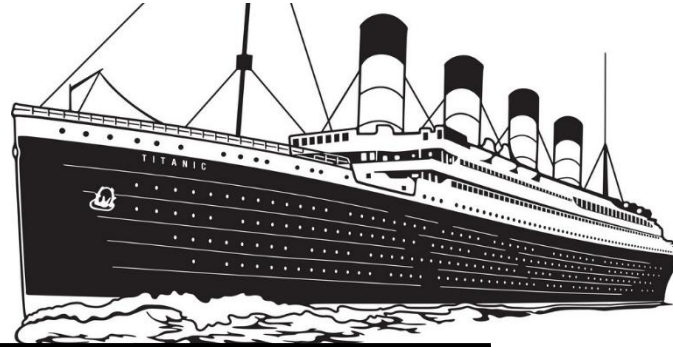
- **Beispiele:**

- Produktqualität
- Konsumentenverhalten vorhersagen
- Prädiktive Wartung
- Medizintechnik

- **Methode: Überwachtes Lernen**

- Klassierte Daten (engl. *Labeled Data*)

TITANIC DATENSATZ



■ Passagierliste der Titanic

Name	Sex	Fare	Age	Survived
Braund, Mr. Owen Harris	male	725	22	no
Cumings, Mrs. John Bradley	female	712833	38	yes
Heikkinen, Miss. Laina	female	7925	26	yes
Futrelle, Mrs. Jacques Heath	female	531	35	yes
Allen, Mr. William Henry	male	805	35	no

Labels, beschreiben die Klasse(z.B. cat, dog, dead, alive)

or

Output variable
(Nummer)

Features, beschreiben und repräsentieren die Eingaben

TITANIC DATENSATZ 2



Name	Sex	Fare	Age	Survived
Braund, Mr. Owen Harris	male	725	22	no
Cumings, Mrs. John Bradley	female	712833	38	yes
Heikkinen, Miss. Laina	female	7925	26	yes
Futrelle, Mrs. Jacques Heath	female	531	35	yes
Allen, Mr. William Henry	male	805	35	no

Name	Sex	Fare	Age	Survived
McCarthy, Mr. Timothy J	male	518625	54	?
Palsson, Master. Gosta	male	21075	2	?
Johnson, Mrs. Oscar	female	111333	27	?
Nasser, Mrs. Nicholas	female	300708	14	?
Sandstrom, Miss. Marguerite	female	167	4	?

Training

Finden von
Modellparametern, welche
das bekannte Label durch
die Features erzeugen

→ langsam



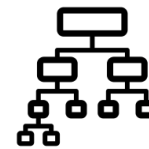
Predicting

Vorhersage des Labels
neuer, ungesehener Daten
durch die Verwendung des
trainierten Modells

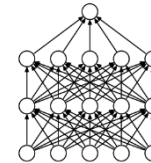
→ schnell

Model

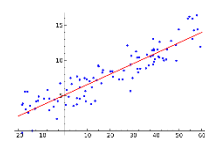
aka Der Algorithmus



decision
trees



neural
networks



regression



- **Features:**

- Features stellen die Input-Größen des Datensatzes dar

- Beispiele:

- Gestaltsausprägungen (Iris-Datensatz)
- Produktqualität (Produktionsprozess)
- Produktpreise (BostonHousingPrizes)

Ausprägung: Breite und Länge des Kelch- und Kronblatts

Ausprägung: physikalische Größen (Kraft, Drehmoment, Leistung)

Ausprägung: Kriminalitätsrate, Raumanzahl, Lehrer-Schüler-Rate

- **Labels:**

- Labels stellen die Zielgröße des zugrundeliegenden Problems dar

- Beispiele:

- Biologische Klasse (Iris-Datensatz)
- Produktqualität (Produktionsprozess)
- Produktpreise (BostonHousingPrizes)

Ausprägung: Iris setosa, Iris virginica und Iris versicolor

Ausprägung: IO und nIO

Ausprägung: kontinuierliche Preisspanne [€ / \$ / ...]



- **Schwertlinien-Datensatz:**

- **Sepalum: Kelchblatt**

- Breite
- Länge

- **Petalum: Kronblatt**

- Breite
- Länge



Iris setosa



Iris versicolor



Iris virginica

FEATURES UND LABELS



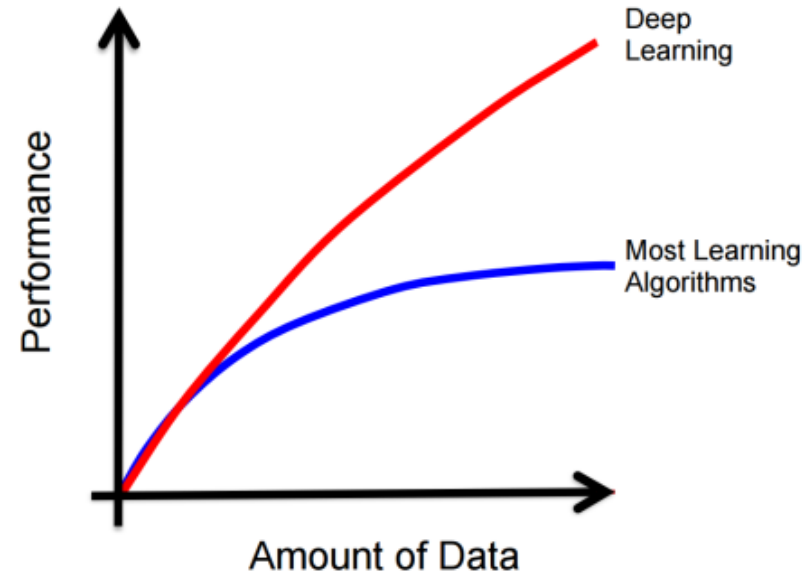
		Feature				Label
Samples		Sepal_ length	Sepal_ width	Petal_ length	Petal_ width	species
	43	5.0	3.5	1.6	0.6	setosa
	64	5.6	2.9	3.6	1.3	versicolor
	23	5.1	3.3	1.7	0.5	setosa
	115	6.4	3.2	5.3	2.3	virginica
	27	5.2	3.5	1.5	0.2	setosa



- **Möglichkeiten heutiger Bilderkennung mittels neuronaler Netze:**
 - https://www.youtube.com/watch?v=z5X_s2Erusc
- **Interaktives Klassifikationsbeispiel:**
 - <https://cs.stanford.edu/people/karpathy/convnetjs/demo/classify2d.html>
- **Regressionsbeispiel:**
 - <https://cs.stanford.edu/people/karpathy/convnetjs/demo/regression.html>

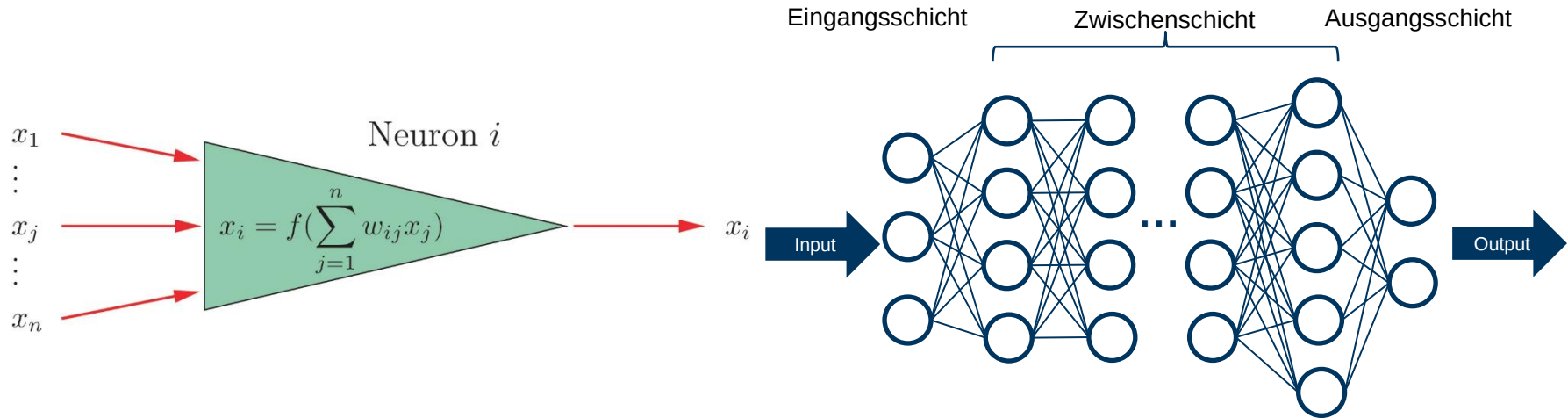


- **Deep Learning**
 - Teilbereich des Machine Learnings
 - Verwendung von Künstlichen Neuronalen Netzen als Modelle
 - Hauptsächlich für komplexe Probleme wie Bild- und Spracherkennung geeignet
 - Modelle durch das menschliche Gehirn inspiriert (sinnvolle Analogie?)
 - Beispiele für Produkte mit Deep Learning Ansatz
 - Siri, Alexa, Google Home
 - Recommender Systems von Amazon, Facebook, Instagram, YouTube, Netflix
 - Tesla Auto Pilot





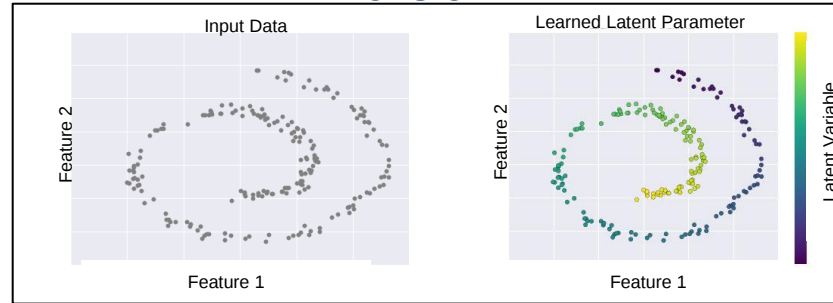
- **Basismodell ist ein Feed-Forward-Netz**
 - Aufbau besteht aus Knoten (Neuronen) und Verbindungen
 - Berechnung erfolgt in Knoten
 - Ergebnis wird über Verbindung an Knoten geleitet



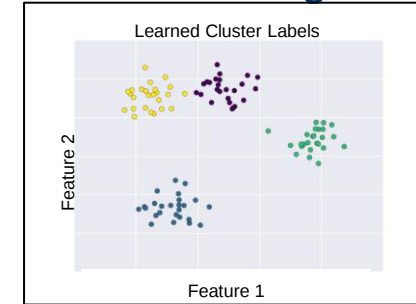


- **Motivation:**

Dimensionsreduktion



Clustering



- **Beispiele:**

- Informationsgewinnung
- Marktanalyse
- Visualisierungen

- **Methode: Unüberwachtes Lernen**

- Unklassierte Daten (engl. Unlabeled Data)

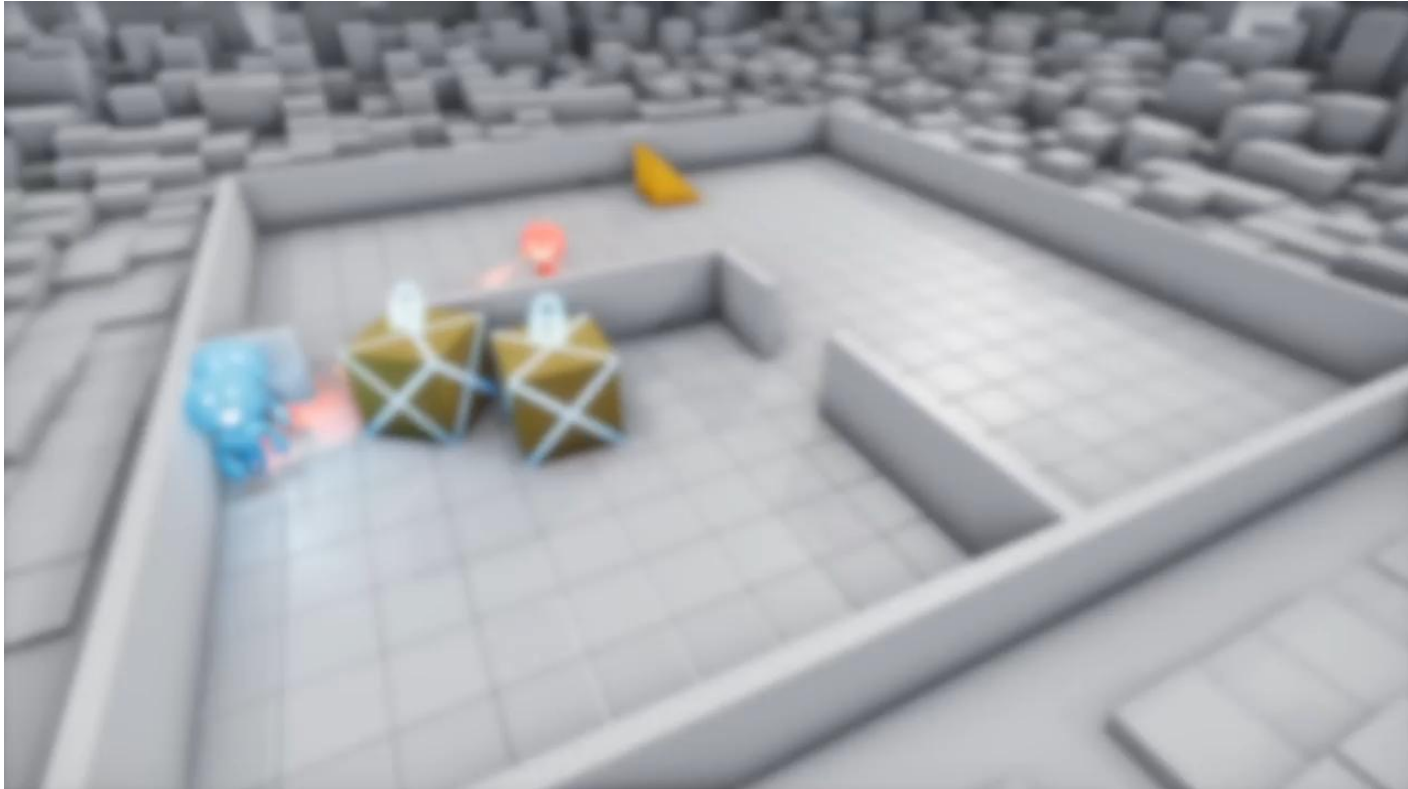
EINLEITUNG MASCHINELLES LERNEN

BESTÄRKENDES LERNEN



- **Motivation:**
 - Bei Fragestellungen mit positiven oder negativen Aussagen eine optimale „Strategie“ entwickeln
- **Beispiele:**
 - Derzeit noch in der Forschung und lediglich vereinzelte praktische Umsetzungen
- **Methode: Bestärkendes Lernen**
 - Positive und negative Belohnungen
 - Agent erlernt Strategie (Schach, Arcade-Spiel)
- Demos:
 - Open-AI-Hide-And-Seek

REINFORCEMENT LEARNING





Anteile eines Data Science Workflow



- Aufbau eines Trainingsdatensatz
- Datensatzsammlung
- Säuberung und Organisation von Daten
- Data-Mining / Informationsextraktion
- Verfeinerung des Algorithmus auf das vorliegende Problem
- Andere

<https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#24d1efae6f63>

ALGORITHMEN ÜBERSICHT (AUSWAHL)

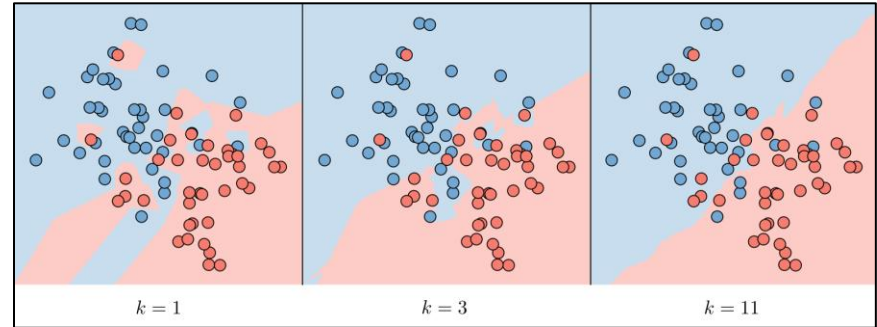


- **Grundlegende maschinelle Lernverfahren:**
 - **KNN** – K-nearest-neighbor – k-nächste-Nachbarn
 - **RandomForest** – Ensemble aus verschiedenen Baumalgorithmen
 - **SVM** – Support-Vector-Machine – Stützvektormaschine

ALGORITHMEN: KNN

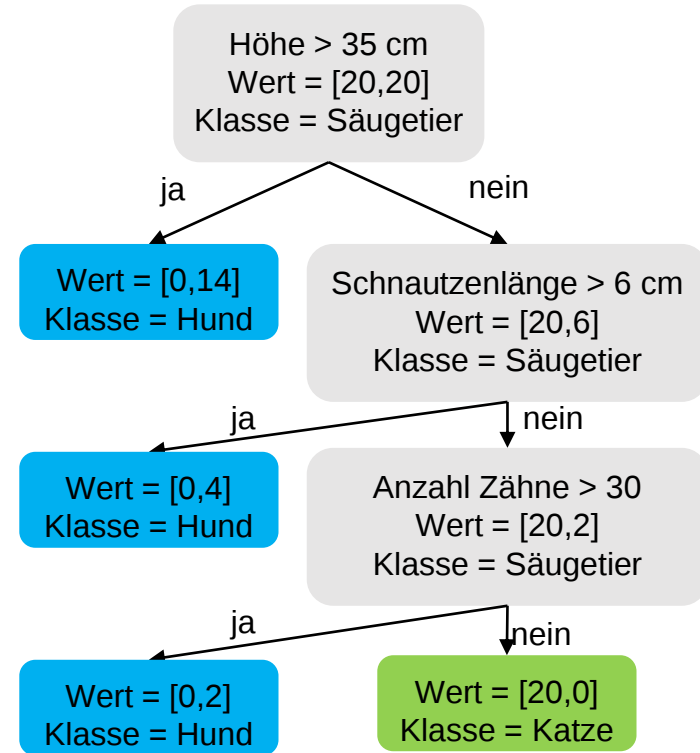


- träges Lernverfahren („lazy“)
- Simpler-Algorithmus: Vergleich eines neuen Datenpunktes mit seinen k -Nachbarn
- Bei steigender Dimension und Umfang der Daten Problem bei Speicher- und Rechenaufwand



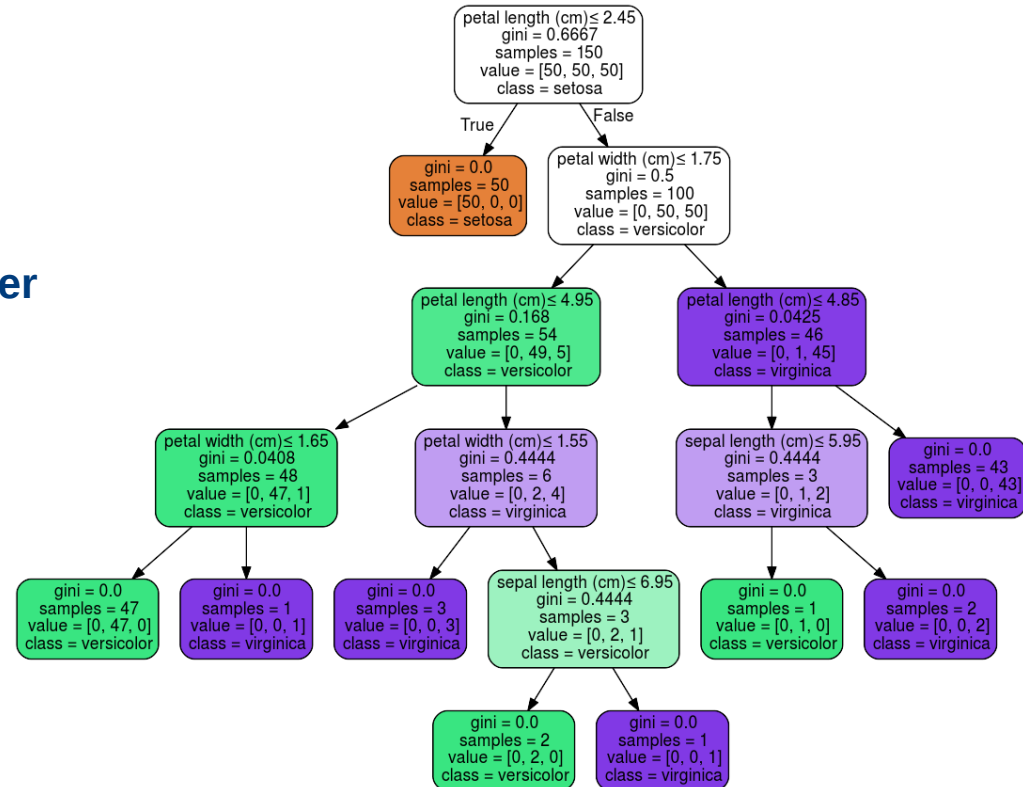


- Ensemble aus unkorrelierten Entscheidungsbäumen
- Sehr robust aufgrund Summe vieler verschiedener Klassifizierer
- Schnelles Training; gut für große Datensätze



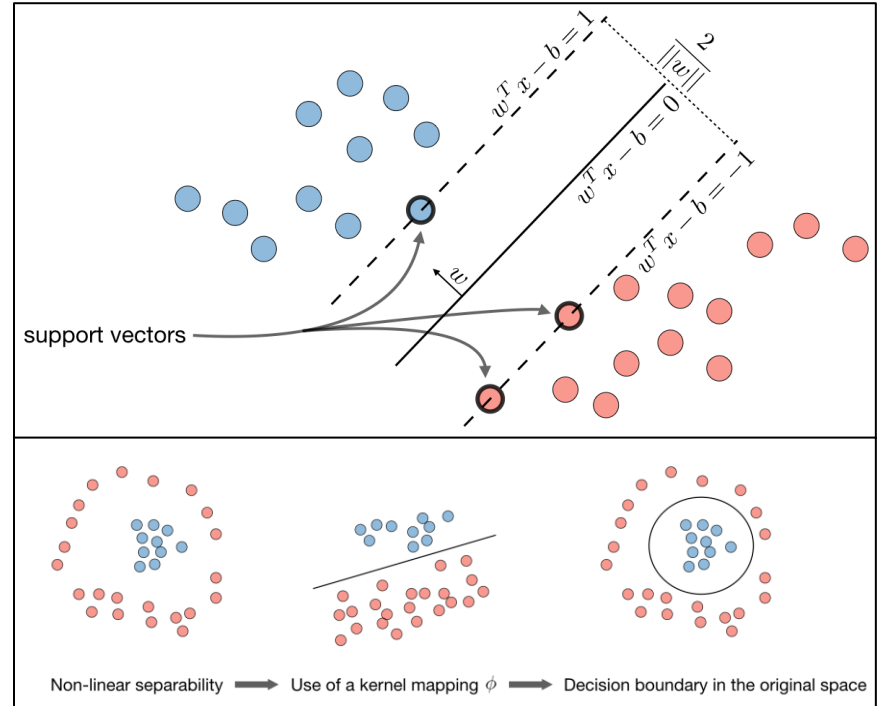


- Ensemble aus unkorrelierten Entscheidungsbäumen
- Sehr robust aufgrund Summe vieler verschiedener Klassifizierer
- Schnelles Training; gut für große Datensätze





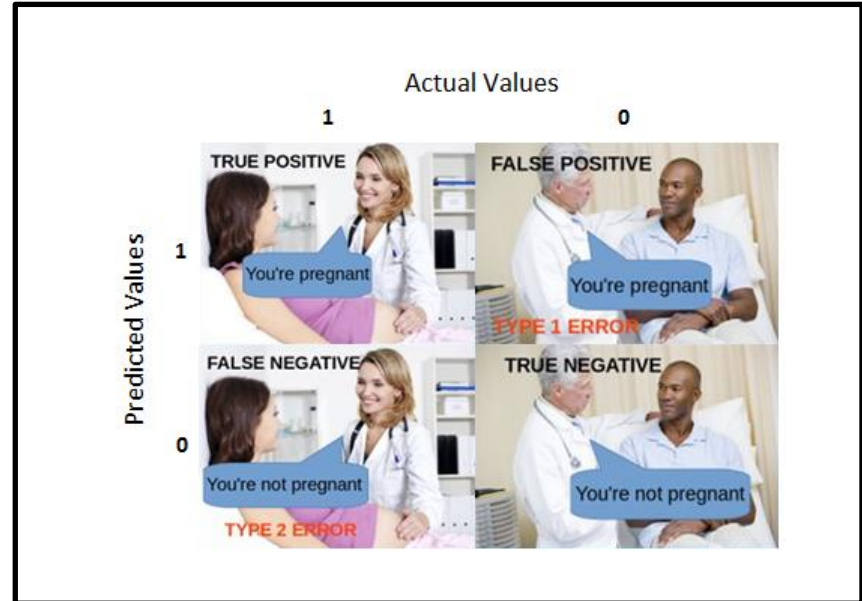
- SVMs (SVR) versuchen einen Datensatz verschiedener Klassen durch Stützvektoren möglichst optimal zu separieren (regressieren)
- 2-Dimensional gut vorstellbar; Probleme bei höheren Dimensionen
- „Kernel-Trick“ für Separierung nicht linear-separierbarer Datensätze





■ Metriken

- „Eine Softwarequalitätsmetrik ist eine Funktion, die eine Software-Einheit in einen Zahlenwert abbildet, welcher als Erfüllungsgrad einer Qualitätseigenschaft der Software-Einheit interpretierbar ist.“ - IEEE Standard 1061, 1998
- Klassifikation:
 - Accuracy
 - Recall
 - Precision
 - F1-Score
- Regression
 - MAE – Mean Absolute Error
 - MAPE – Mean Absolute Percentage Error
 - MSE – Mean Squared Error
 - RMSE – Root Mean Squared Error



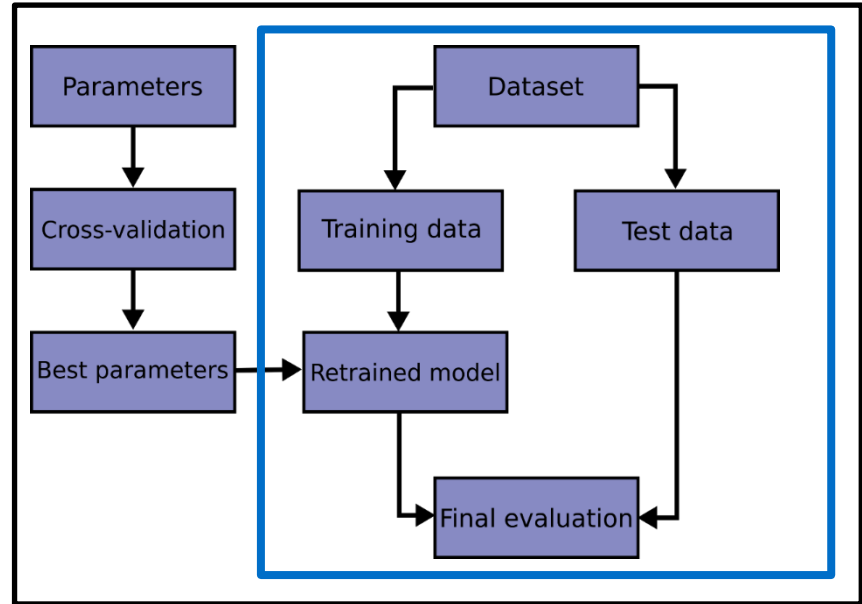
TRAIN-TEST-SPLIT



- „The central challenge in machine learning is that our algorithm must perform well on new, previously unseen inputs—not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called *generalization*.“

Quelle: Deep Learning S. 108; GOODFELLOW ET AL.

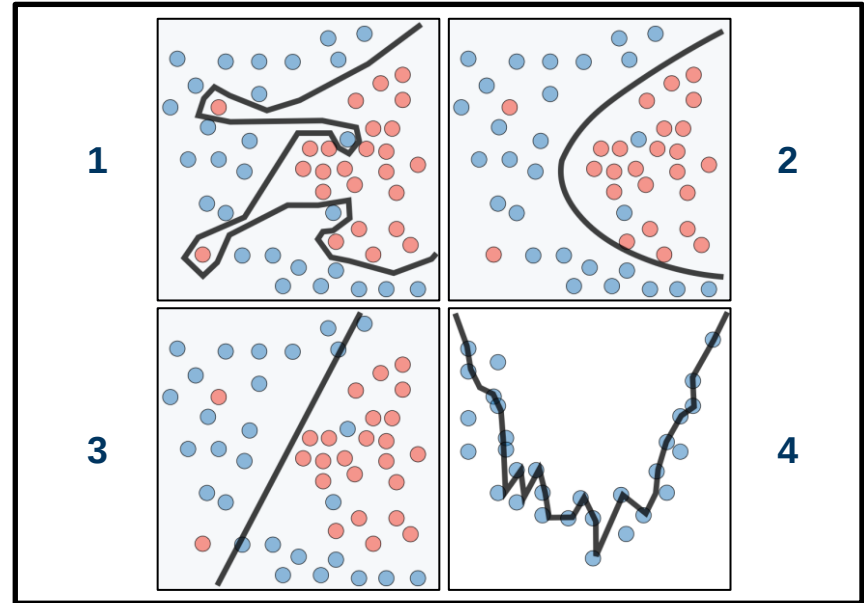
- Ziel: Generalisierung**
 - Aufteilen des Datensatzes
 - Faustregel: Train-Test-Split-Ratio ~ 80/20
 - Variation je nach Umfang des Datensatzes



OVER- UND UNDERFITTING



- Ziel: Classifier soll durch Training auf unbekannte, neue Daten generalisieren
- Einordnung der rechts stehenden Classifier:
 1. Overfit
 2. Good-fit
 3. Underfit
 4. Overfit



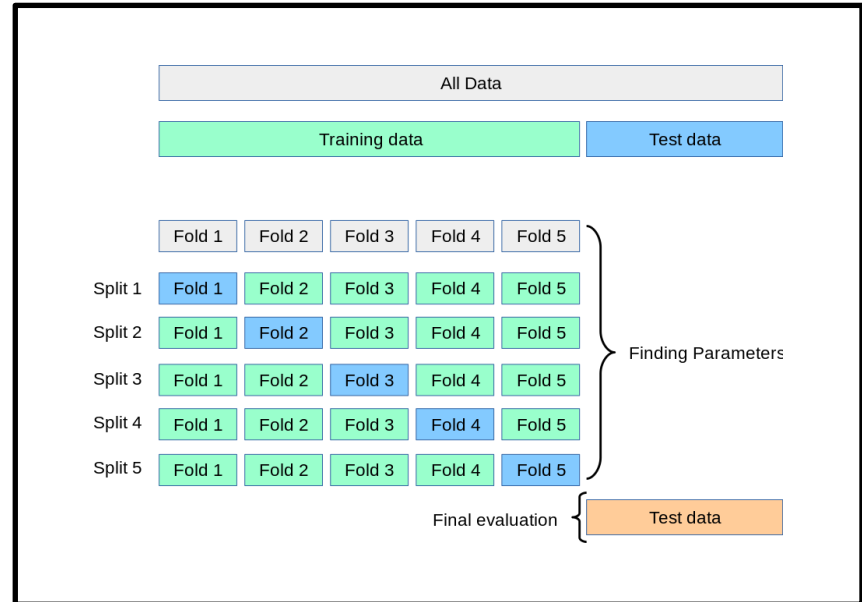


- Werden VOR Beginn des Trainings durch den Anwender gesetzt
- Werden während des Trainings nicht verändert bzw. erlernt
- **Beispiele:**
 - Anzahl K-nächster Nachbarn [KNN]
 - Kernel [SVM]
 - Gamma [SVM]
 - Lernrate [NN]
 - ...

K-FOLD-CROSSVALIDATION



- Für die Optimierung von Hyperparametern können „Validations-Splits“ erzeugt werden
- **Nachteil:** Noch weniger Trainingsdaten
- **Alternative: K-Fold-Crossvalidation**
- **Vorteil:**
 - Validations-Split entfällt
 - Durchschnittliche Performance anhand k-facher Ausführung
 - Train-Test-Split bleibt unangetastet



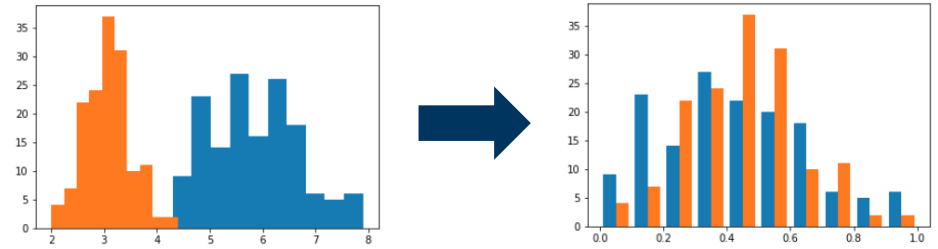
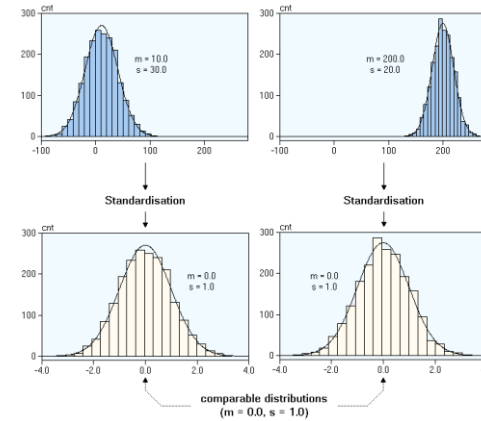


- **Standardisierung:**

- Erwartungswert Null
- Varianz 1
- Meist bei gauß-verteilten Daten gut anwendbar

- **Normalisierung:**

- Alle Werte in Bereich [0,1] bzw. [-1,1] skalieren
- Vorteilhaft wenn Verteilung der Daten unbekannt oder nicht gauß-verteilt ist
- Möglichkeiten:
 - Scaling
$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$
 - Clipping
 - Z-Score



AGENDA



1. Einordnung von Data Science für Ingenieure
2. Datenspeicherung in Datenbanken
3. Datenverarbeitung in Python
4. Machine Learning Basics
- 5. Ablauf einer Machine Learning Pipeline**
6. Anwendungsbeispiel im Ingenieurskontext

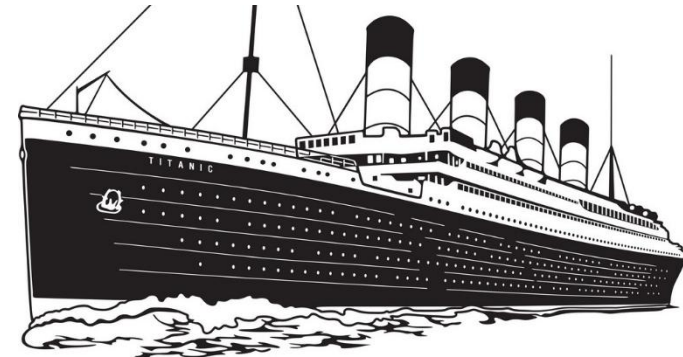
PROBLEM DEFINITION



■ Titanic Dataset

- Vorhersage, welche Passagiere das Titanic Unglück überlebt haben
- Wird häufig als „Hello-World“ für Machine Learning Probleme bezeichnet
- Supervised Learning, Binäres Klassifikationsproblem

	sex	age	class	fare	embark_town	survived
0	male	22.0	Third	7.2500	Southampton	0
1	female	38.0	First	71.2833	Cherbourg	1
2	female	26.0	Third	7.9250	Southampton	1
3	female	35.0	First	53.1000	Southampton	1
4	male	35.0	Third	8.0500	Southampton	0





DEMO JUPYTER NOTEBOOKS

KAGGLE-COMPETITIONEN



- Bosch-Competition
- Mercedes-Competition
- Power_Line-Competition
- Severstal-Steel-Competition

WEITERES MATERIAL



- [Google-ML-Glossar](#)
- [Stackoverflow](#)
- [Library-Dokumentationen \(z.B. Scikit-Learn\)](#)
- [DataScienceHandbook](#)
- [The-Elements-Of-Statistical-Learning](#)
- [Ian Goodfellow: Deep Learning](#)