

Künstliche Intelligenz (KI/AI) - Basiswissen und Anwendungen



Dr. Gerrit Korff

exeta





MATTI GERRIT KORFF

Senior Data Scientist

Dr. Matti Gerrit Korff verfügt über mehr als 6 Jahre Berufserfahrung als Experte für Data Science und maschinelles Lernen. Erfahrungen sammelte er durch seine Tätigkeit als Machine Learning Consultant in einer Vielzahl von Projekten. Hierbei hat er sich in einen weiten Bereich an Branchen und Fragestellungen eingearbeitet. Seine Erfahrung macht ihn zu einem hochqualifizierten Problemlöser und Projektleiter.

Biografie

- Tätigkeiten im Bereich Versicherungen, Telekommunikation, Webportale
- Dr. rer. nat. Chemistry, Freie Universität Berlin Biochemistry, Universität Bielefeld

Beratungskompetenz

- Data Science, Machine Learning, Artificial Intelligence, DevOps, Software Engineering, Entrepreneurship, Agile nach Scrum
- IT Expertise u.a. in Python (Kedro, Sklearn, numpy, scipy); DevOps (Docker, AWS, Apache Airflow, grafana, Elasticsearch, logstash, Kibana)

Sprachen

- Deutsch, Englisch

Auszug relevante Projekterfahrung

Machine Learning Engineer, Entwicklung einer KI-Plattform, Versicherung

- Containerisierung der KI-Modelle
- Aufstellen der Daten und CI/CD-Pipelines

Machine Learning Engineer, Umsetzung von Legal-Tech Usecases, Legal-Tech Start-up

- Klassifikation von Verträgen auf ihre Gültigkeit mittels NLP (Spacy, Transformers)

Data Scientist, KI-Assistenzsystem im Kundenservice, großer deutscher Transportdienstleister

- Entwicklung einer KI-Lösung im Bereich Question Answering
- Entwicklung eines Anonymisierungsservices für Sprachtranskripte

Data Scientist, Smart Speaker Entwicklung, dt. Telekommunikationskonzern

- Verbesserung des natürlichen Sprachverständnisses (NLU)
- Entwicklung einer Analyse- und Entwicklungsumgebung für die NLU

Agenda Tag 1

Vorstellungsrunde

Daten und Information

Wissen aus Daten generieren

Machine Learning Workflow

Agenda Tag 2

Machine Learning Workflow (übrige Schritte)

KI-Projekte in der Umsetzung

Projektteams

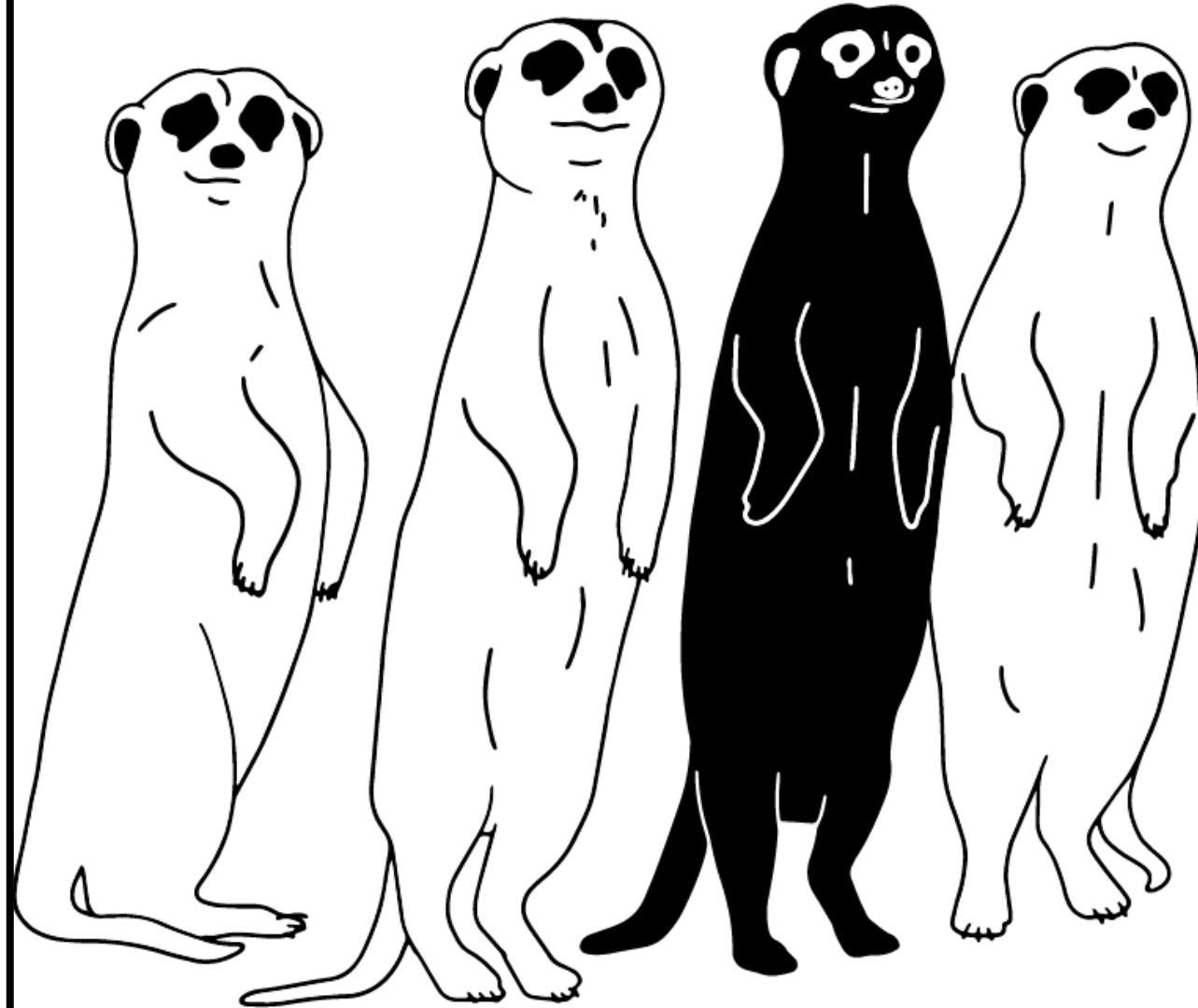
Projektphasen

KI-Ethik und Datenschutz

Anonymisierung / Pseudonymisierung

Vorstellungsrunde

- Hintergrund - Was mache ich meinem Beruf / Tätigkeit?
- Habe ich bereits Erfahrung mit KI / AI gemacht?
- Warum besuche ich diese Schulung?
- Welche Vorkenntnisse habe ich?
- Was erwarte ich mir von dieser Schulung?

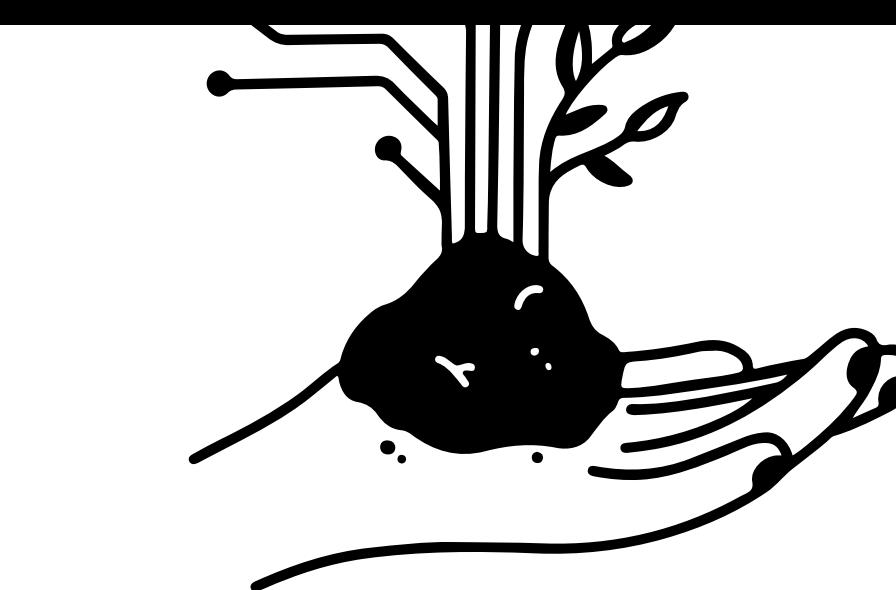


Lernziele

- Grundlegendes Verständnis über AI/KI sowie verwandte Themengebiete wie Data Science & Machine Learning (ML)
- Herangehensweise an Data Science / ML – Problemstellungen anhand eines Machine Learning Workflows
- Überblick über gängige ML Methoden & Konzepte, sowie ausgewählte Algorithmen des Supervised & Unsupervised Learnings
- Praktische Umsetzung von einer typischen ML-Pipeline für ausgewählte Algorithmen
- Wo liegen die Grenzen von KI-Modellen und welche Risiken gibt es

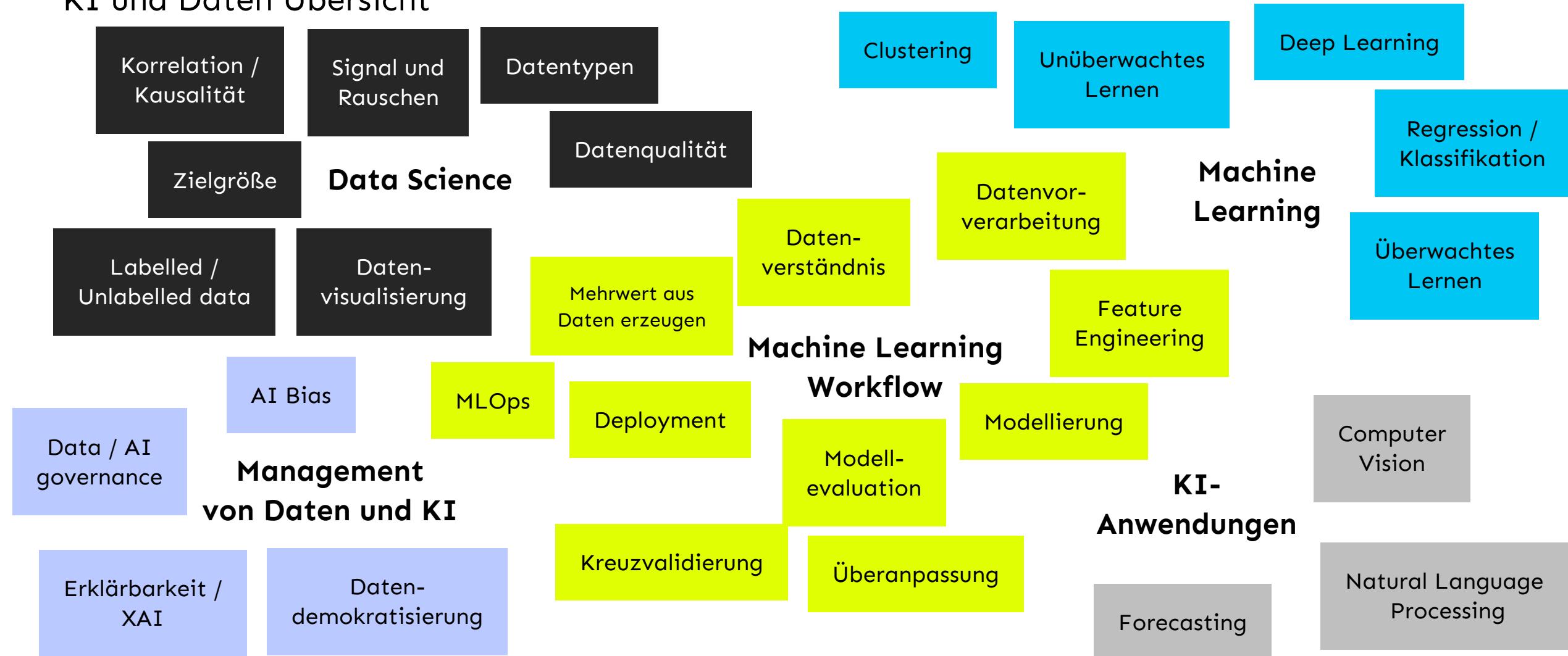
01

Daten und Information



Wissen aus Daten generieren

KI und Daten Übersicht



Daten und Information

Daten sind die Grundlage

Viele Anwendungen von künstlicher Intelligenz (KI) basieren auf Methoden aus dem maschinellen Lernen (ML). Hierbei soll die KI selbstständig Muster aus historischen Daten erlernen.

Je mehr Daten verfügbar sind, desto **komplexere Muster** können erkannt werden. Die **digitale Verfügbarkeit** ist die Grundlage für die Anwendungen von KI.

Das **Verständnis** von Daten und ihre Haltung bilden den **Ausgangspunkt** für KI-Anwendungen

Daten und Information

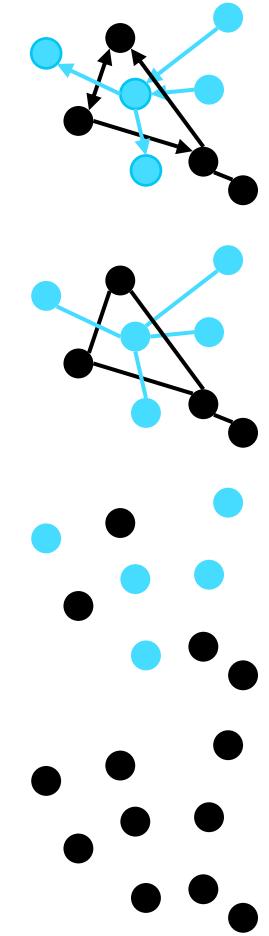
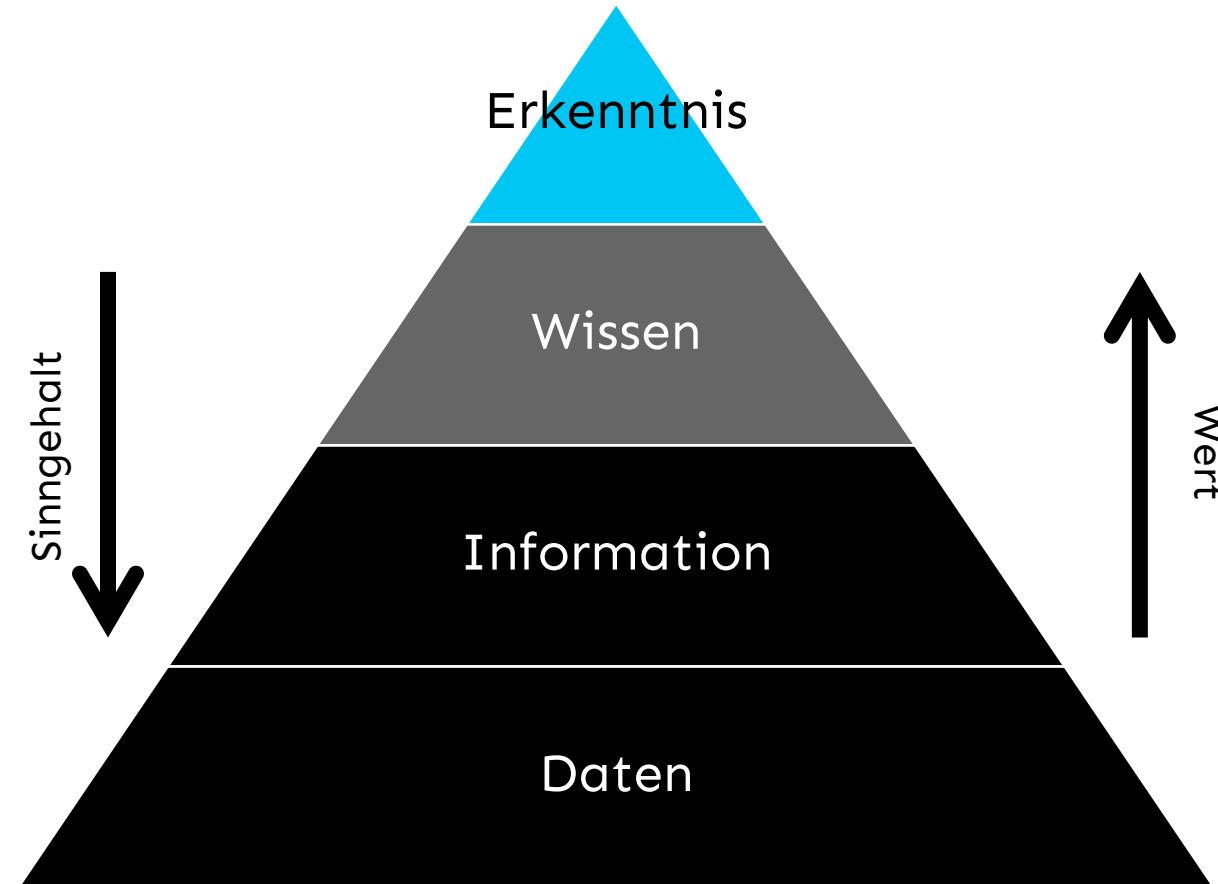
Verarbeitbarkeit und Sinngehalt

Information beschreibt den Verarbeitbarkeitsgrade und den Sinngehalt.

Daten bezieht sich auf die Art und Weise wie Information gespeichert werden.

Was sind Daten?

Repräsentationen von Fakten, welche in unterschiedlichen digitalen Formen gespeichert werden kann, was eine Verarbeitung mittels Computer und Algorithmen erlaubt.



Daten und Information

Datenqualität

Integrität			Konformität			Encoding Fehler
CUSTOMER_ID	ORDER_ID	ORDER_DATE	DELIVERY_DATE	RATING_Q1	RATING_Q2	COMMENT
US733847		Fehlwerte	18.12.2018	5	5	
US648202	75938375	20.12.2018	23.12.2018	0	0	Nicht plausible Werte Please stop sending me emails
FR007492	77774948	23.12.2018	N/A	4	3	Je suis très content. Merci
UK849372	78883745	02.01.2019	05/01/2019	4	5	
DE334839	11396723	03.01.2019	08-01-2019	5	5	
DE334839	11396723	03.01.2019	08-01-2019	1	1	Völlig Unzufrieden!
CN223475	44637948	03.01.2019	2019-01-15	4	0	oooooooooooo
CN223475	44637948	03.01.2019	2019-01-15	4	0	oooooooooooo

Mögliche Antworten für RATING_Q1 und RATING_Q2:

1 = Gar nicht, 2 = Nicht wirklich, 3 = Etwas, 4 = Ja, bedingt, 5 = Ja

Duplikate

Daten und Information

Datentypen

CUSTOMER_ID	LAST_NAME	FIRST_NAME	ORDERS	CITY	ZIP_CODE	COUNTRY
10302	Boucher	Peter	1	Nantes	44000	France
11244	Smith	Maryam	53	Berlin	83030	Germany
11405	Han	Sun-He	2	Sydney	3004	Australia
11993	Mueller	Gisela	13	Tamm	71732	Germany



```
{
  "EMPLOYEES": {
    "SALES": {
      "648229": {
        "NAME": "Olivia Johnson",
        "DOB": "1989-08-06"
      },
      "648666": {
        "NAME": "Frank Mueller",
        "DOB": "1985-05-11",
        "MISC": "On paternal leave"
      }
    }
  }
}
```

Daten und Information

Datentypen

	Strukturierte Daten	Semi-strukturierte Daten	Unstrukturierte Daten
Was ist es?	<ul style="list-style-type: none">Daten mit hohem Organisationsgrad, die normalerweise tabellarischer Form gespeichert werden	<ul style="list-style-type: none">Daten mit einem gewissen Organisationsgrad	<ul style="list-style-type: none">Daten ohne vordefinierte Organisationsform und kein bestimmtes Format
Bespielformate	<ul style="list-style-type: none">Excel TabellenComma-separated value (.csv) DateinRelationale Datenbank Tabellen	<ul style="list-style-type: none">Hypertext Markup Language (HTML) DateinJavaScript Object Notation (JSON) DateinExtensible Markup Language (XML) Datein	<ul style="list-style-type: none">Bilddatein(.jpeg, .png)Videodatein(.mp4, m4a)Sounddatein(.mp3, .wav)TextdateinWord DateinPDF Datein
Merkmale	<ul style="list-style-type: none">Daten sind tabellarisch strukturiertEinträge haben ein einheitlichen FormatEinfach maschinenlesbar	<ul style="list-style-type: none">Dateien haben einen gewissen Grad an Organisation und StrukturTags/Marker trennen Elemente und erzwingen Hierarchien, aber die Einträge können im Format variierenBenötigt einige Vorverarbeitungen	<ul style="list-style-type: none">Daten können eine beliebige Form annehmenInnerhalb der Datei gibt es keine InhaltsstrukturBenötigt in der Regel umfangreiche Vorverarbeitung, kann aber oft von Menschen leicht verstanden werden

Daten und Information

Stamm- und Transaktionsdaten

Stammdaten (Master Data)	Transaktionsdaten
<ul style="list-style-type: none">• Daten zu Geschäftsobjekten, die unternehmensweit gemeinsam genutzt werden• Normalerweise statische Daten, die sich selten ändern• Beispiele:<ul style="list-style-type: none">• Kundendaten• Produktdaten• Mitarbeiterdaten	<ul style="list-style-type: none">• Daten, die Ereignisse und Transaktionen beschreiben• Nicht statisch und haben typischerweise eine zeitliche Dimension• Beispiele:<ul style="list-style-type: none">• Buchungen eines Online Shops• Website Logdaten• Sensordaten

Daten und Information

Datenmenge

ABSCHÄTZUNG

Allgemein: Je mehr Daten desto besser

Benötigte Datenmenge abhängig von:

- Anwendungsfall
- Datenqualität

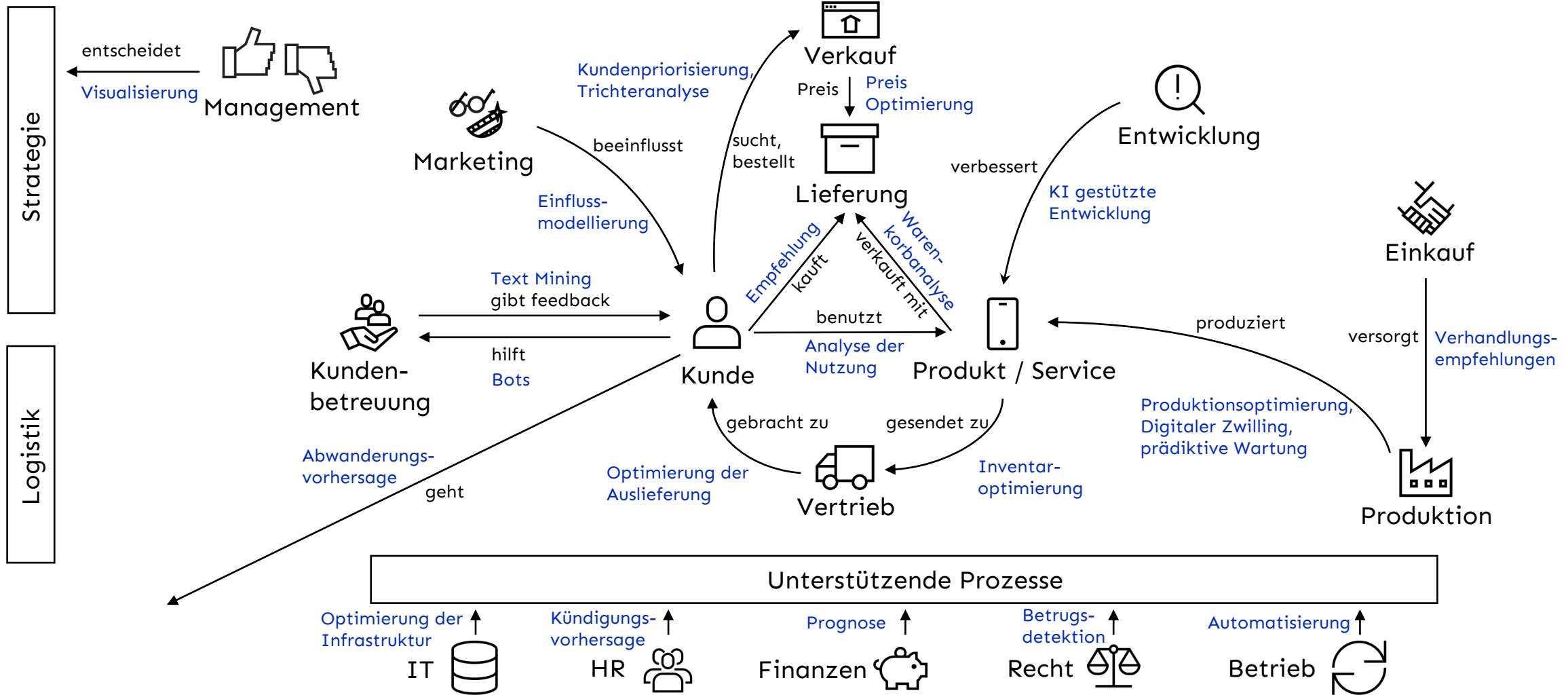
EIGENSCHAFTEN

- Intuitive Hypothese -> weniger Daten
- Seltene Ereignisse -> mehr Daten
- Viele Eigenschaften -> mehr Daten
- Mehr Modellparameter -> mehr Daten
- Nicht-Lineare Zusammenhänge -> mehr Daten



Wissen aus Daten generieren

Data Science Use Cases



Wissen aus Daten generieren

Joghurt Bestand

Wie ließe sich das verhindern?

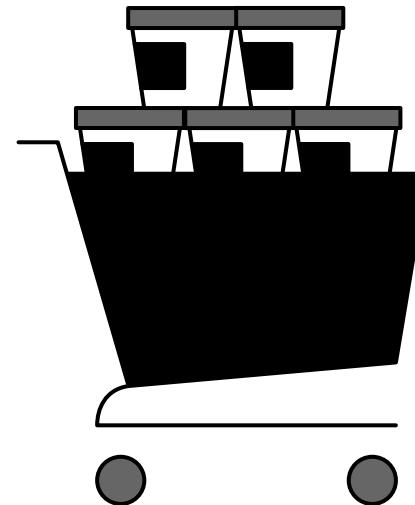


Wissen aus Daten generieren

Machine Learning in der Bestandsoptimierung

Einflussfaktoren für die Nachfrage

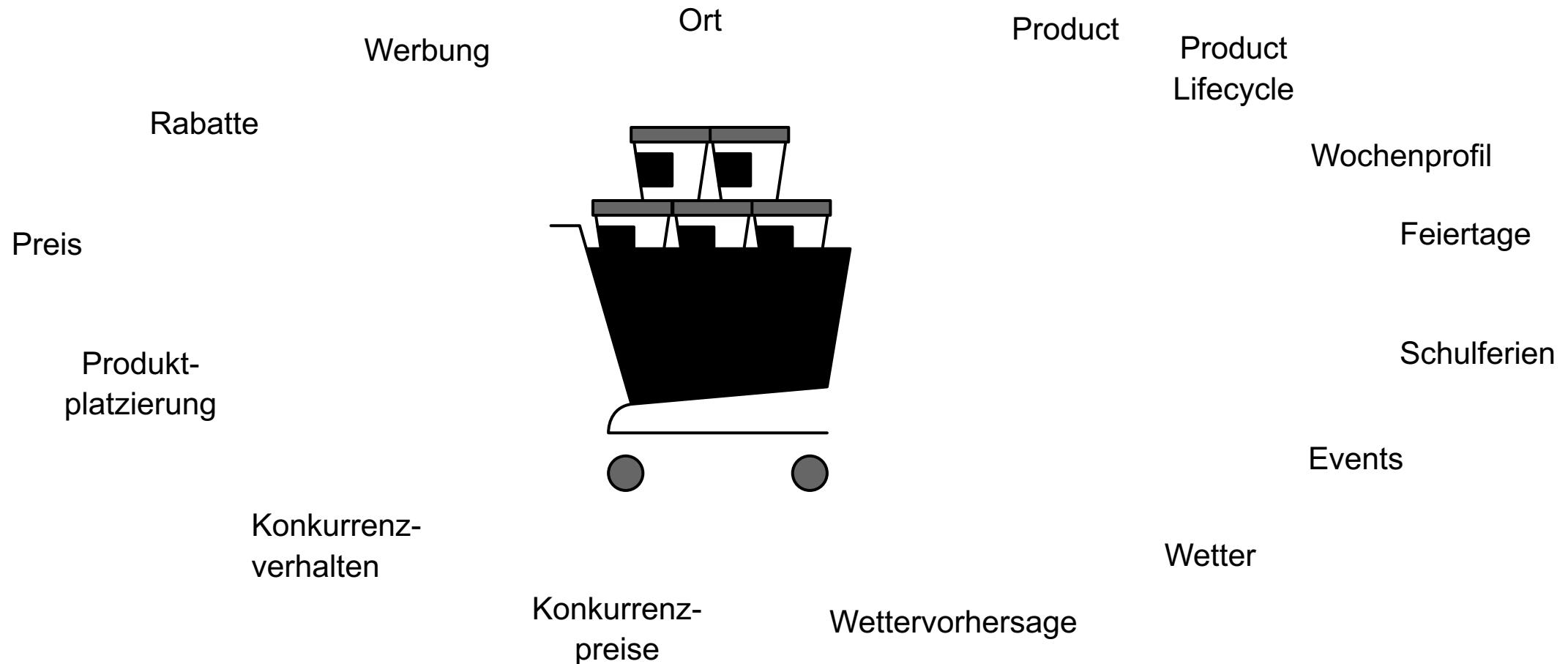
Problemstellung: Vorhersage der Nachfrage so genau wie möglich



Wissen aus Daten generieren

Machine Learning in der Bestandsoptimierung

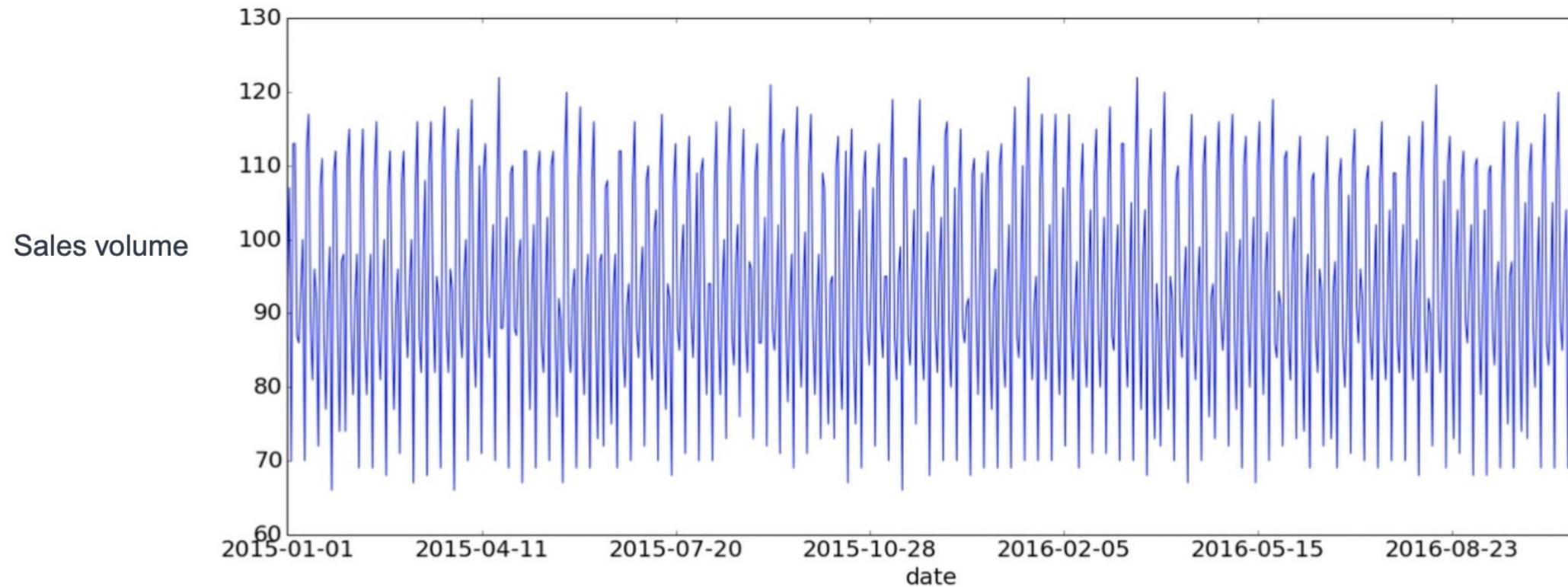
Einflussfaktoren für die Nachfrage



Wissen aus Daten generieren

Machine Learning in der Bestandsoptimierung

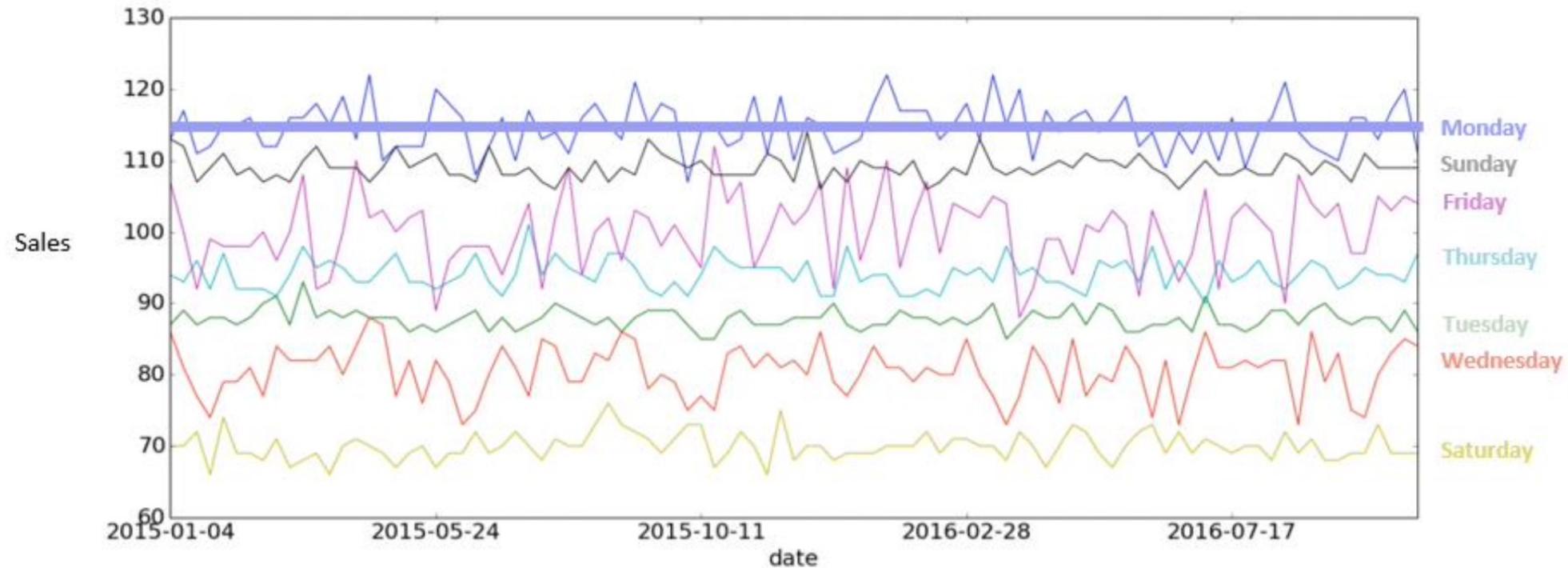
Herangehensweise: Lernen aus historischen Daten



Wissen aus Daten generieren

Machine Learning in der Bestandsoptimierung

- **Wochenprofil (Wochensaisonalität):** Die meisten Joghurts werden Montags gekauft
- **Entscheidung:** Mehr Joghurt für Montags



Wissen aus Daten generieren

Analyse der Joghurtverkäufe

Live Demo

Wissen aus Daten generieren

KI und ML

Künstliche Intelligenz

Computer / Maschinen die intelligente, menschenähnliche Verhaltensweisen und Fähigkeiten aufweisen, welche es ihnen erlauben Aufgaben auszuführen die menschliche Intelligenz benötigen

Machine Learning

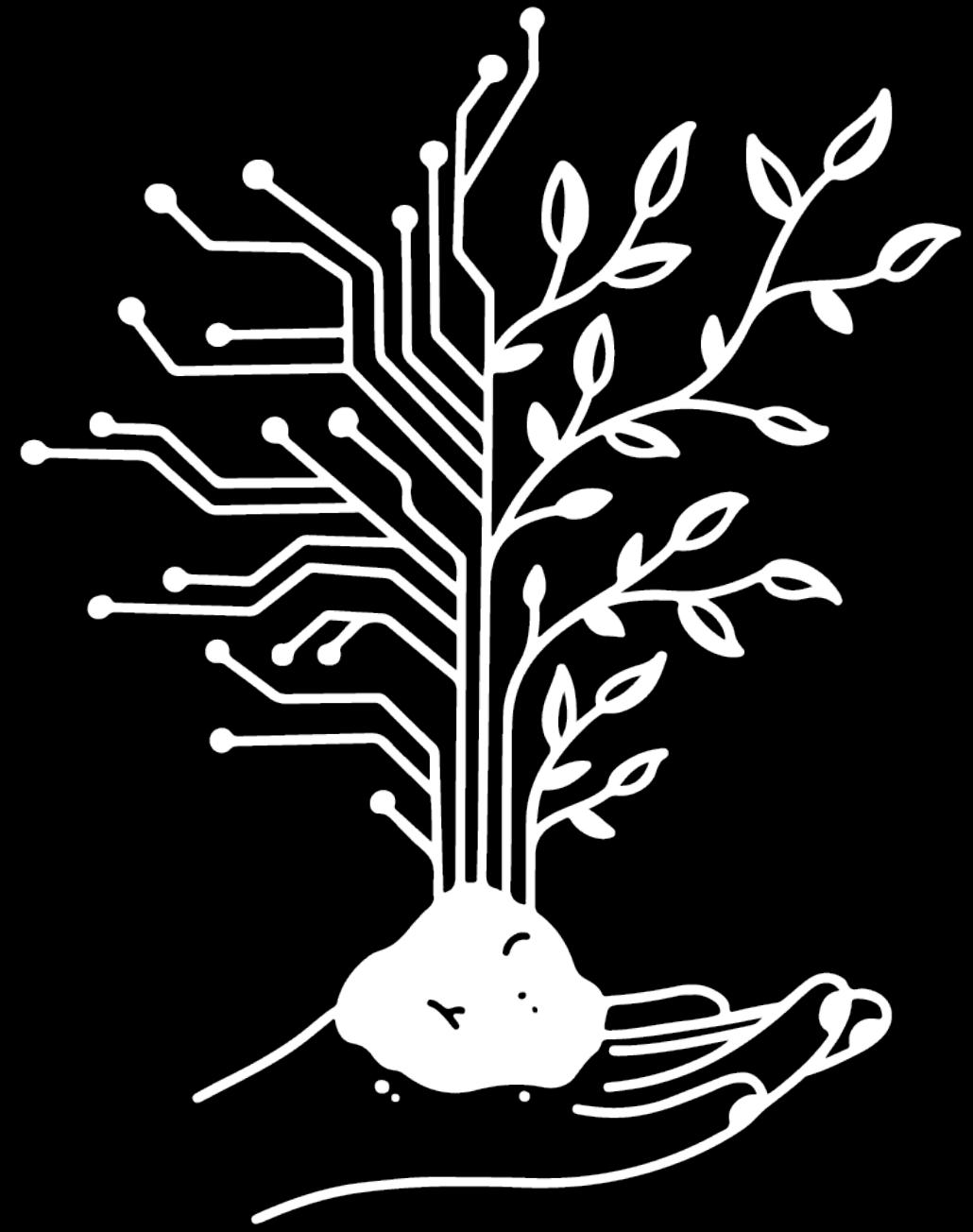
Ein Forschungsbereich zu Algorithmen, statistischen Modellen und Computersystemen, mit dem Ziel, das ein Computer lernt eine Aufgabe auszuführen ohne das er explizit programmiert wurde.

Neuronale Netzwerke (Deep learning)

Wissen aus Daten generieren

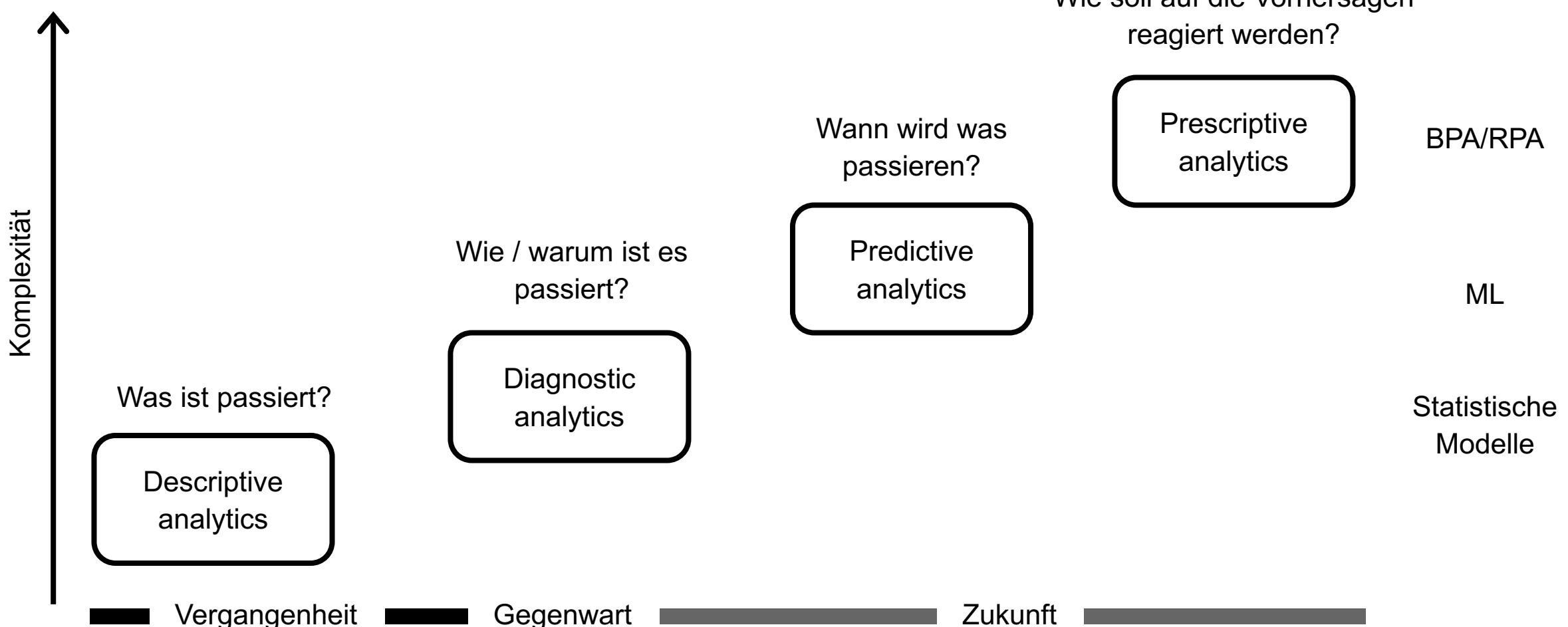
Was ist Data Science?

- Data Science ist ein Wissenschaftsfeld welches wissenschaftlich fundierte Methoden, Prozesse, Algorithmen und Systeme zur **Extraktion von Erkenntnissen, Mustern und Schlüssen** sowohl aus Daten ermöglicht.
- Die **Machine Learning Algorithmen sind ein Schlüsselement** von Data Science, da sie es erlauben auch komplexe Muster aus Daten zu erlernen.



Aus Daten Mehrwert erzeugen

Die verschiedenen Formen von Data Analytics



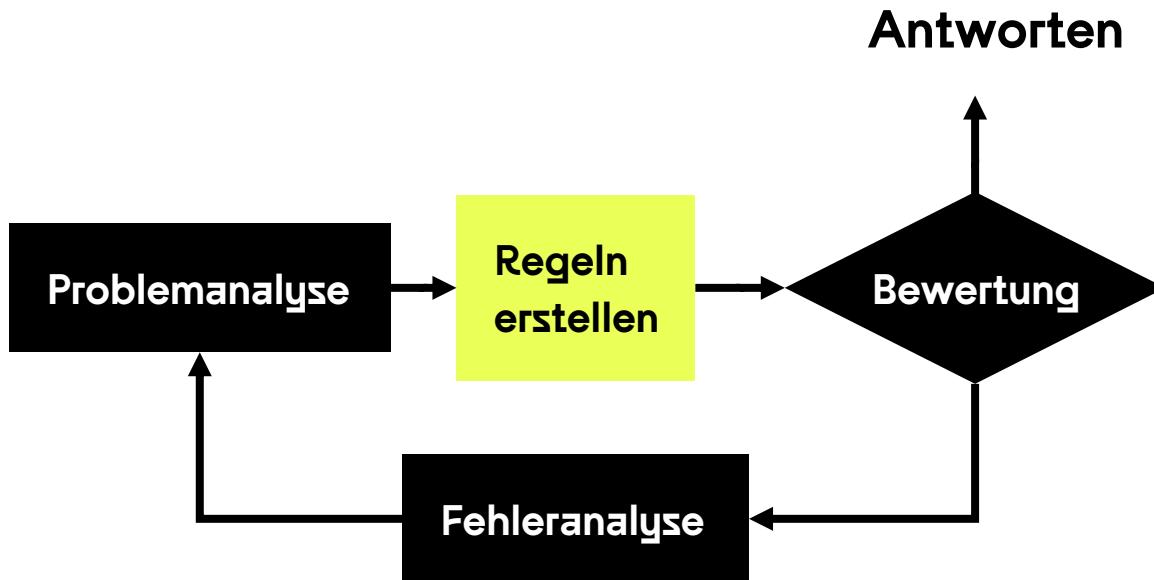
Wissen aus Daten generieren

Was ist Machine Learning (ML)?

- Die Erforschung von Algorithmen, statistischen Modellen und Computersystemen, um **Aufgaben ohne explizite Anweisungen, zu lösen**. Ein ML-Modell **lernt selbstständig** aus historischen Daten.
- ML-Algorithmen **brauchen viele Daten**, um zu funktionieren, je mehr desto besser.
- Immer mehr Daten sind verfügbar, wodurch die Performance der Algorithmen sich verbessert.
- Mit der Menge der Daten ist auch die **Verfügbarkeit von Rechenleistung** gestiegen. Viele ML-Ideen sind schon länger bekannt, aber es fehlte früher die Rechenleistung.
- ML hat sich weiterentwickelt, speziell Modelle aus dem Bereich der neuronalen Netze haben für Durchbrüche gesorgt.

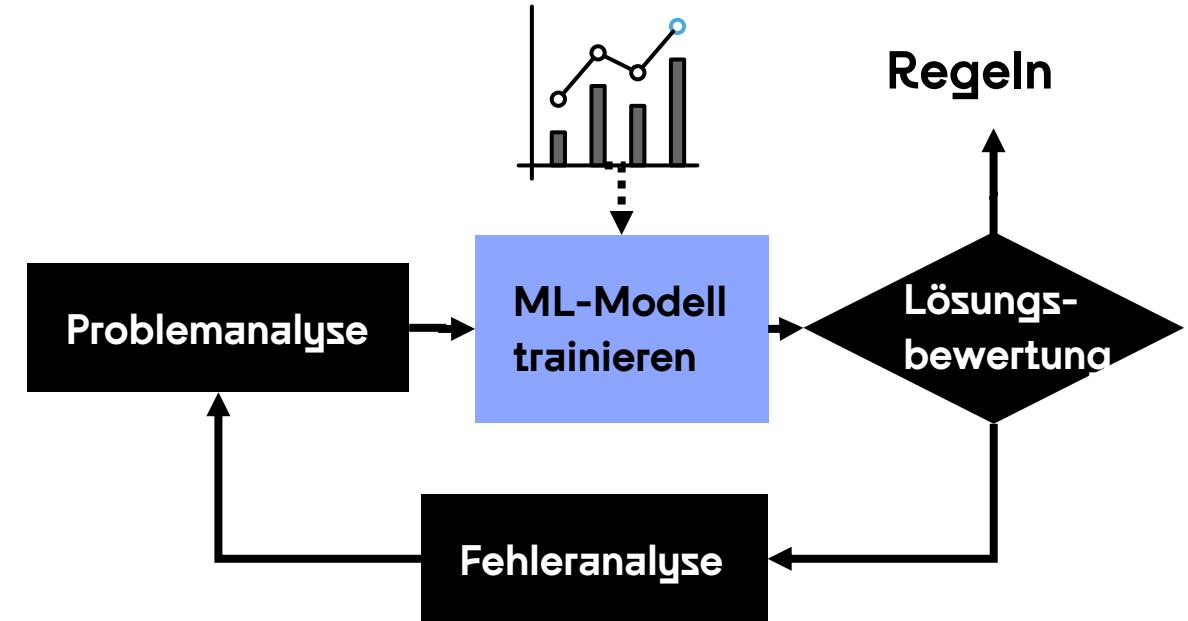
Wissen aus Daten generieren

Was bedeutet lernen aus Daten?



Traditioneller Ansatz

- Komplex, *hard coded*
- Schwer zu warten



Machine Learning Ansatz

- Automatisches lernen aus Daten
- Automatisches neu trainieren

Wissen aus Daten generieren

Mechaniker bei der Rally Dakar - Übung

Ein Fahrzeug um die Ralley Dakar zu gewinnen.

Daten zu den **Schäden** bei den **letzten 50 Siegerfahrzeugen** liegen vor.

=> Wo muss ein Fahrzeug verstärkt werden, um nicht während der Rallye auszufallen?

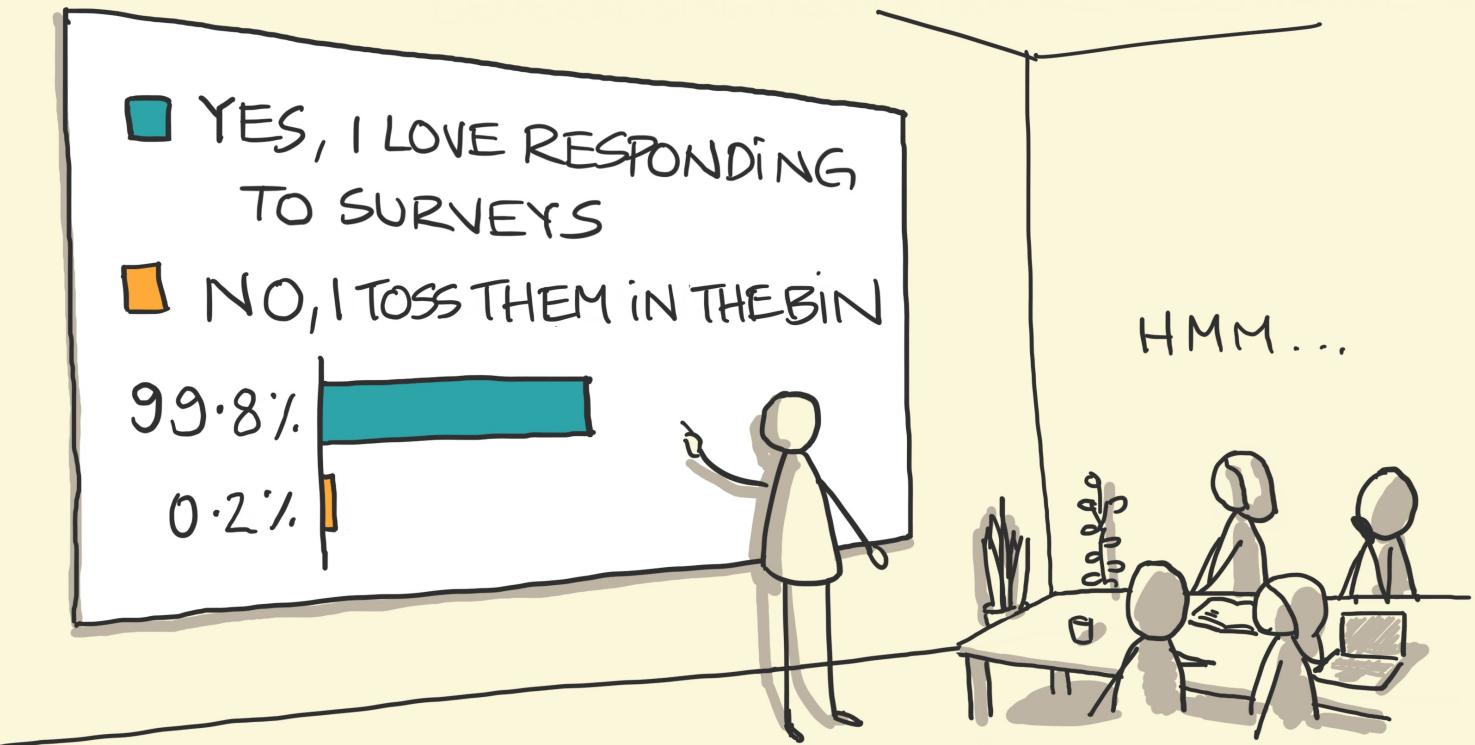


Wissen aus Daten generieren

Sampling Bias

- Die Zusammensetzung eines Datensatzes ist entscheidend.
- Gilt nicht nur für die Zielgröße.

SAMPLING BiAS



" WE RECEIVED 500 RESPONSES AND FOUND THAT PEOPLE LOVE RESPONDING TO SURVEYS "

sketchplanations

Wissen aus Daten generieren

Herausforderungen bei Machine Learning

- Machine Learning bringt im Vergleich zu traditionellen Software System **neue Herausforderungen** mit sich
- Unzureichende Trainingsdaten
 - Nicht repräsentative Trainingsdaten (*sampling bias*)
 - Schlechte Datenqualität (Fehlwerte oder welche die nicht als solche erkennbar sind)
 - Irrelevante oder unzureichende Merkmale
 - Überanpassung

Es gilt: „Garbage in, garbage out“

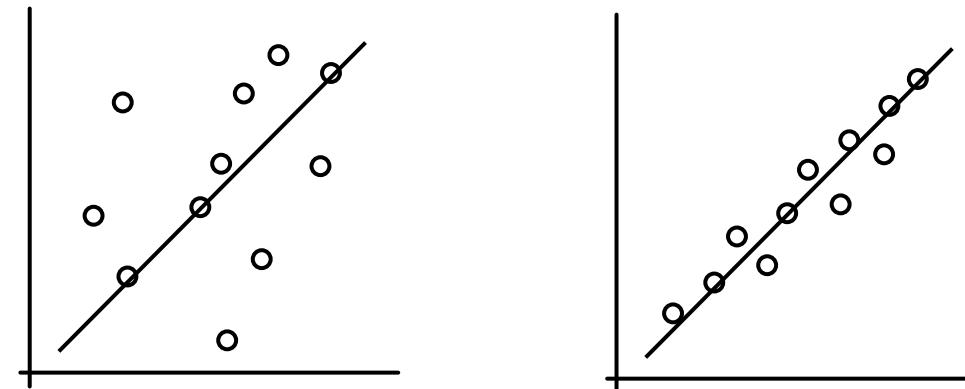


Wissen aus Daten generieren

Computer lernen Muster - Signal und Rauschen

Signal: Das **zugrunde liegende Muster** eines Prozesses. Muster bedeutet hier etwas sich wiederholendes oder vorhersagbares. Mit einem Modell soll das Muster vorhergesagt werden.

Rauschen: Zufällige und unbekannte Einflussfaktoren, welche das zugrunde liegende **Muster verfälschen**.



Je stärker das Signal und je schwächer das Rauschen eines Prozesses, desto einfacher und besser lässt er sich über ein ML-Modell vorhersagen.

Wissen aus Daten generieren

Computer lernen Muster - Voraussetzungen

Prozess / Ergebnis muss überhaupt ein **Signal haben**

- Ein zu komplexes Signal lässt sich nicht von Rauschen unterscheiden

Das Signal muss **möglichst rauschfrei** sein

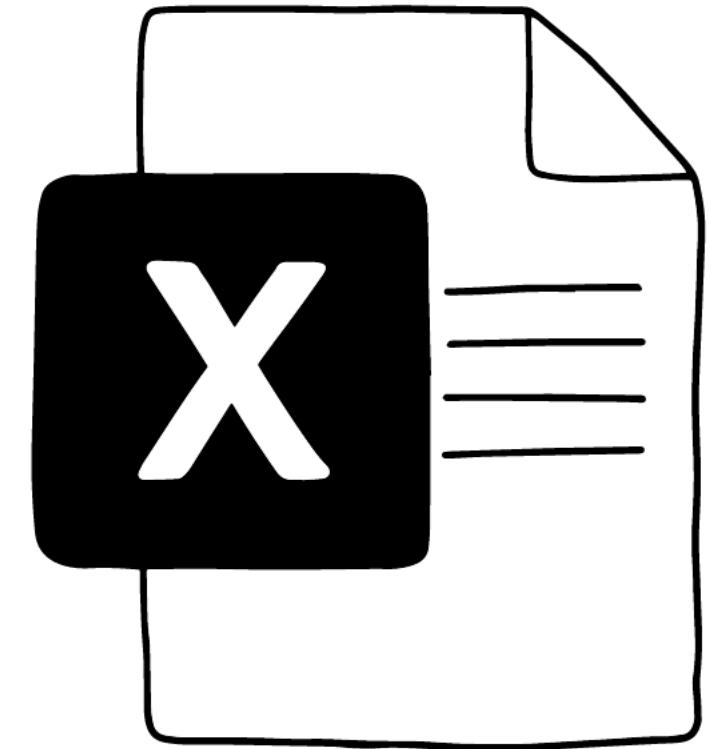
- Je mehr (unbekannte) Einflussfaktoren, desto schwerer zu erkennen

Es muss **ausreichend Daten** geben, um das Signal zu erkennen

- Es kann nur vorhergesagt werden, was auch in den Daten ist

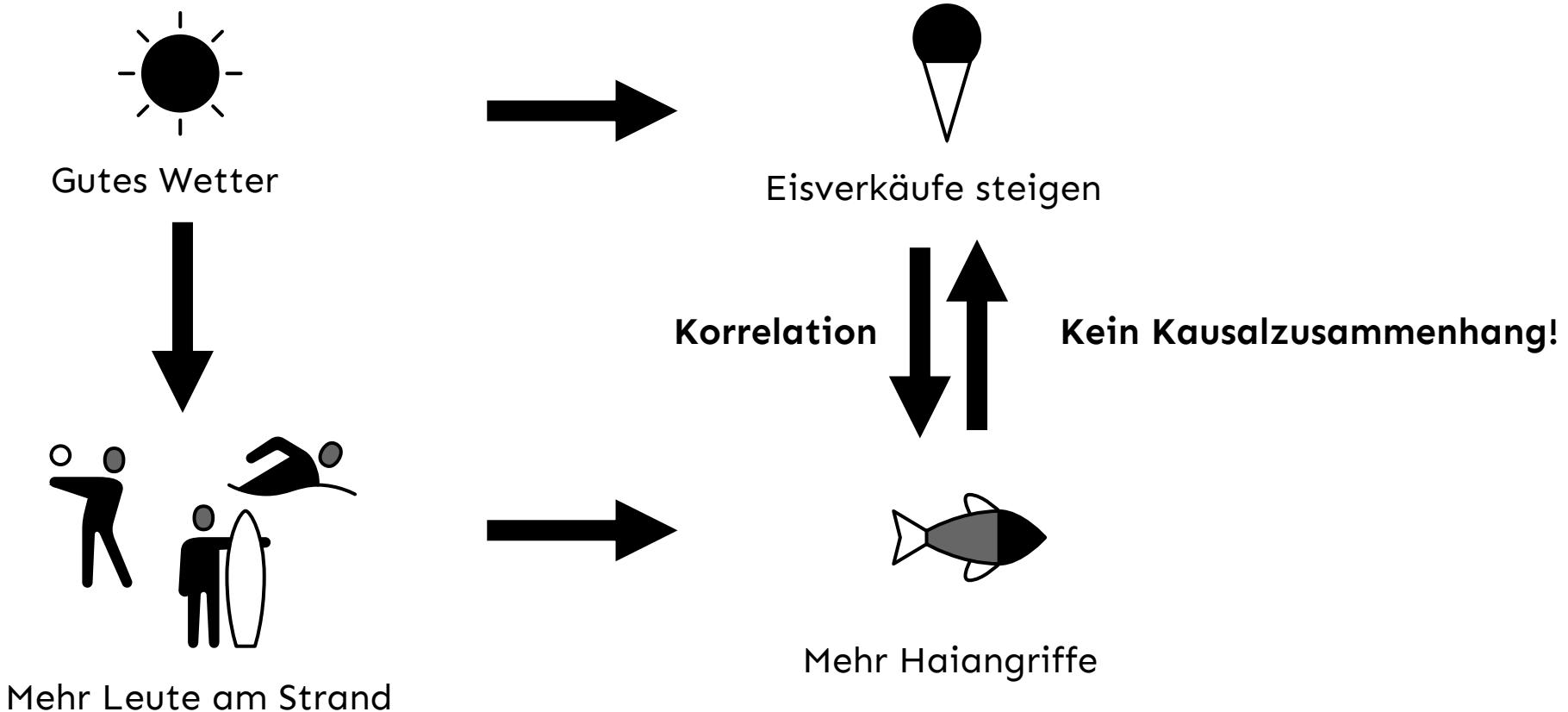
Es werden nur **Wahrscheinlichkeiten** ausgegeben

- Keine Dichotomie in vorhersagbar und unvorhersehbar sondern ein Spektrum



Wissen aus Daten generieren

Korrelation ≠ Kausalität



Wissen aus Daten generieren

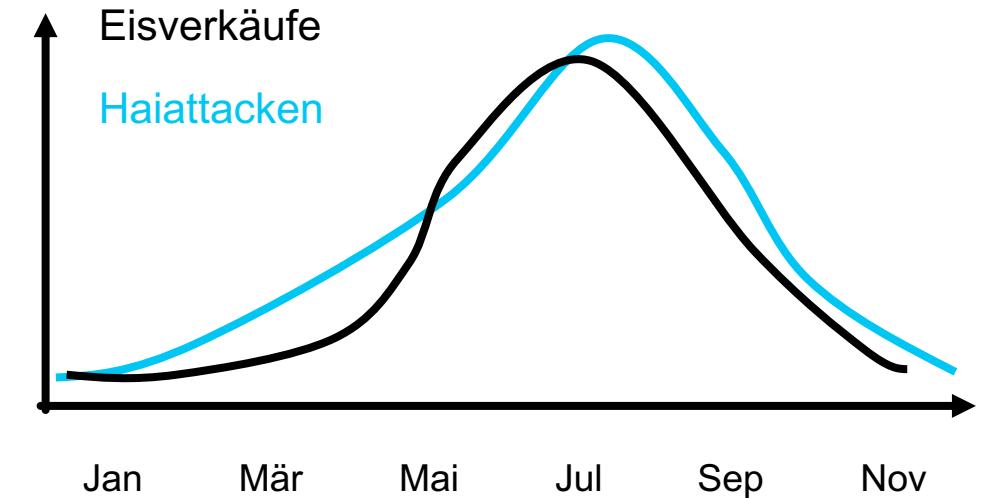
Korrelation ≠ Kausalität

Wenn ein Merkmal A und B korrelieren, kann man nicht zwangsläufig davon ausgehen, dass A der Auslöser / Treiber für B ist.

Beispiel:

Bei der Untersuchung einer demografischen Datenbank, wird eventuell ein Zusammenhang zwischen „Anzahl von Krankenhäusern“ und der „Anzahl von Autodiebstählen in der Region“ gefunden, welche korrelieren.

- Dies bedeutet nicht, dass das eine die Ursache für das andere ist
- Beide sind offensichtlich verbunden, jedoch durch ein drittes Attribut, nämlich „Bevölkerung“



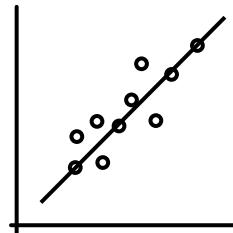
Wissen aus Daten generieren

Typen von Machine Learning

Meist wird unterteilt in 2 Typen, die sich hinsichtlich der **Art wie gelernt** wird und der Form und Vorhandensein der sogenannten **Zielgröße** unterscheiden

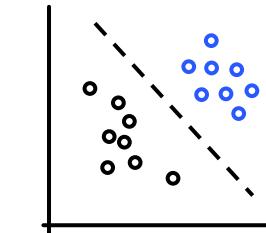
SUPERVISED LEARNING (ÜBERWACHTES LERNEN)

Regression



Vorhersage einer
kontinuierlichen
Zielgröße, z.B.
Immobilienpreis

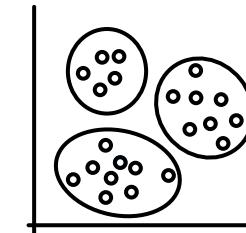
Klassifikation



Vorhersage einer
kategorischen
Zielgröße, z.B.
Schadensklasse

UNSUPERVISED LEARNING (UNÜBERWACHTES LERNEN)

Clustering



Einteilen von
Beobachtungen in
Gruppen, z.B.
Kundensegmente

Wissen aus Daten generieren

Zielgröße und Merkmale

- Die **Zielgröße** (abhängige Variable, *target variable*) beschreibt den Gegenstand des Interesses.
- Eine **Zielgröße** gibt es nur beim *Supervised learning*, nicht beim *unsupervised learning*
- Die **Merkmale** (Einflussgrößen, unabhängige Variablen, *input variables*) werden von dem Modell verwendet, um die Zielgröße vorherzusagen

Beispiel:

- Vorhersage des Immobilienwertes (Zielgröße) aus der Ausstattung und Lage (Merkmale)

Wissen aus Daten generieren

Supervised learning

- *Supervised learning* ist die **verbreitetste** und **wichtigste** Form von Machine Learning.
- Die Voraussetzung ist, dass es **Faktoren gibt, welche die Zielgröße beeinflussen** oder in Beziehung zu ihr stehen, die **Merkmale**.
- Ziel ist die Vorhersage der Zielgröße über Bestimmung des Musters wie Merkmale und Zielgröße zusammenhängen.
- Es wird *supervised learning* genannt, da es **während des Trainings Feedback** gibt.

Beispiel:

- Klassifikation von E-Mails
- Zukünftige Absatzzahlen

Wissen aus Daten generieren

Supervised learning: Regression und Klassifikation

Regression

Es soll eine numerische Zielgröße vorhergesagt werden. Die möglichen Werte sind **kontinuierlich**.

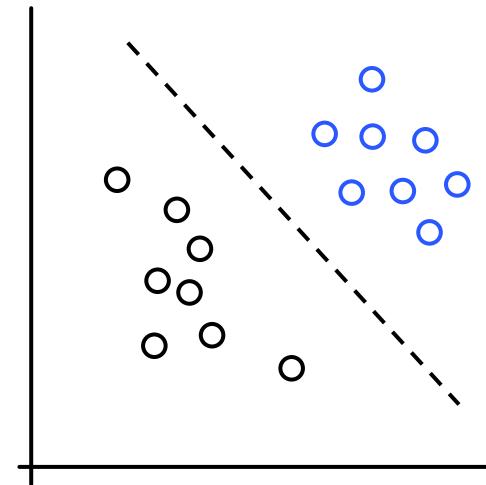
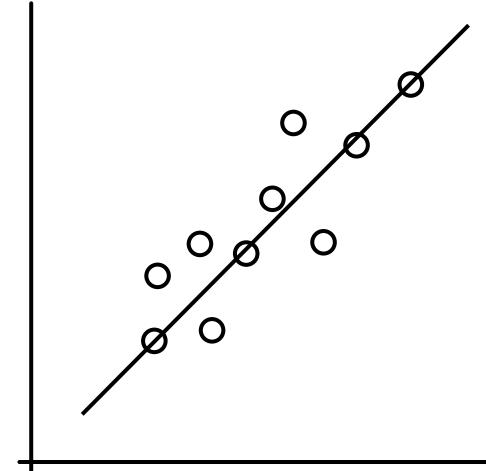
Beispiele: Wahrscheinlichkeiten, Verkäufe

Klassifikation

Es soll eine kategoriale Zielgröße vorhergesagt werden. Die Zielgröße kann nur **definierte Werte** annehmen. Das Modell erlernt eine **Entscheidungsgrenze**

Beispiele: Wahr / Falsch, Ja / Nein, Steuerklassen, Noten

Manche Anwendungsfälle können sowohl als Klassifikations- als auch als Regressionsproblem angegangen werden



Wissen aus Daten generieren

Unsupervised learning

- Die **zweitwichtigste** Form von Machine Learning.
- Zu der Zielgröße, die uns interessiert, gibt es keine Daten. Es kann somit **kein Modell erzeugt werden, dass eine Zielgröße vorhersagt**.
- *Unsupervised learning* sucht nach Mustern in den Daten.
- Ohne Zielgröße gibt es beim Trainieren des Modells **kein Feedback**

Beispiel

- Gruppieren von Fahrzeugen mit ähnlichen Symptomen, um die Erstdiagnose zu beschleunigen.

Wissen aus Daten generieren

Unsupervised learning: Clustering, Anomalie Erkennung und Autoencoder

Clustering

Gruppierung von Beobachtungen so dass es innerhalb der Gruppe große und zwischen den Gruppen geringe Ähnlichkeiten gibt

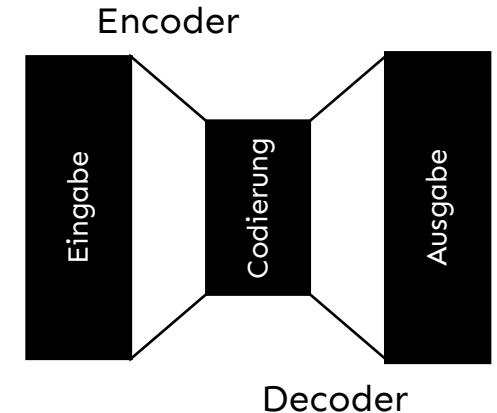
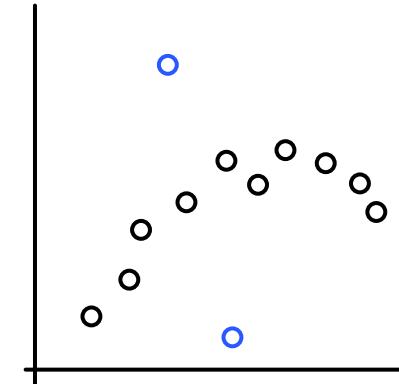
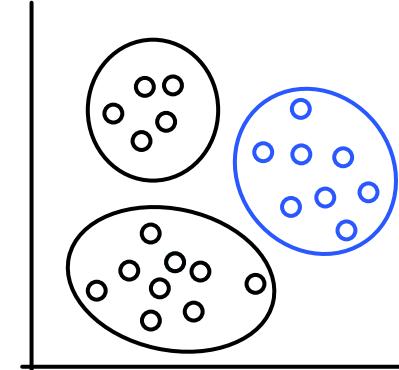
Anomalieerkennung

Erkennen von Ausreißern: Seltene und / oder signifikant unterschiedliche Beobachtungen

Autoencoder / Dimensionsreduktion

Erlernen von einer effizienten Repräsentation eines Datensatzes

Was für Muster lassen sich in einem Datensatz finden?



Wissen aus Daten generieren

Labelled und *unlabelled* data

In einem ***labeled dataset*** gibt es die Zielgröße

In einem ***unlabelled dataset*** ist die Zielgröße nur theoretisch vorhanden, wurde aber nicht zugewiesen

Über unsupervised learning können *labels* für *unlabelled data* erzeugen werden

Zielgröße: Betrugsfälle

Überweisungsdaten

Spalte vorhanden ob Betrug vorliegt
oder nicht
⇒ **Labelled** data

Direkte Einteilung möglich

Keine Spalte ob Betrug vorliegt oder nicht
⇒ **Unlabelled** data

Betrugsfälle müssen erst markiert werden bzw.
Einteilung der Daten in ähnliche Gruppen

Wissen aus Daten generieren

Regression, Klassifikation und Clustering - Übung

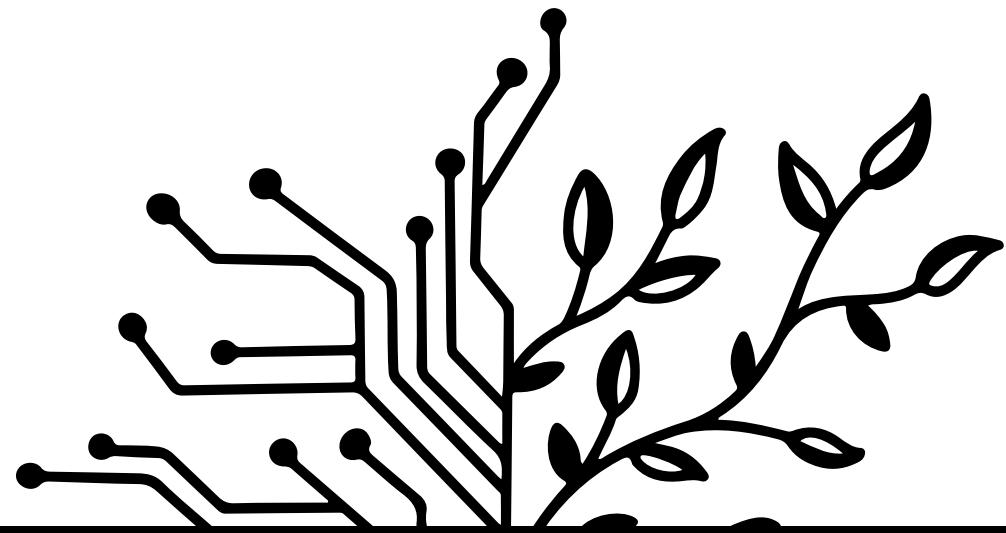
Welche Anwendungsfälle fallen Ihnen zu Regression, Klassifikation und Clustering ein?

Wissen aus Daten generieren

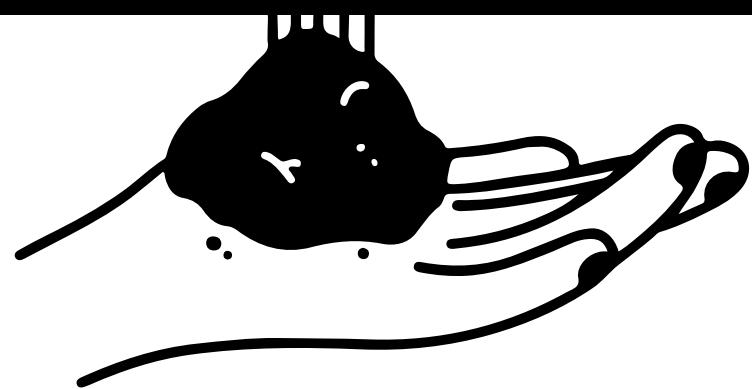
Gänge Data Science Tools

Je nach Anwendungsfall ist eine unterschiedliche Auswahl relevant.



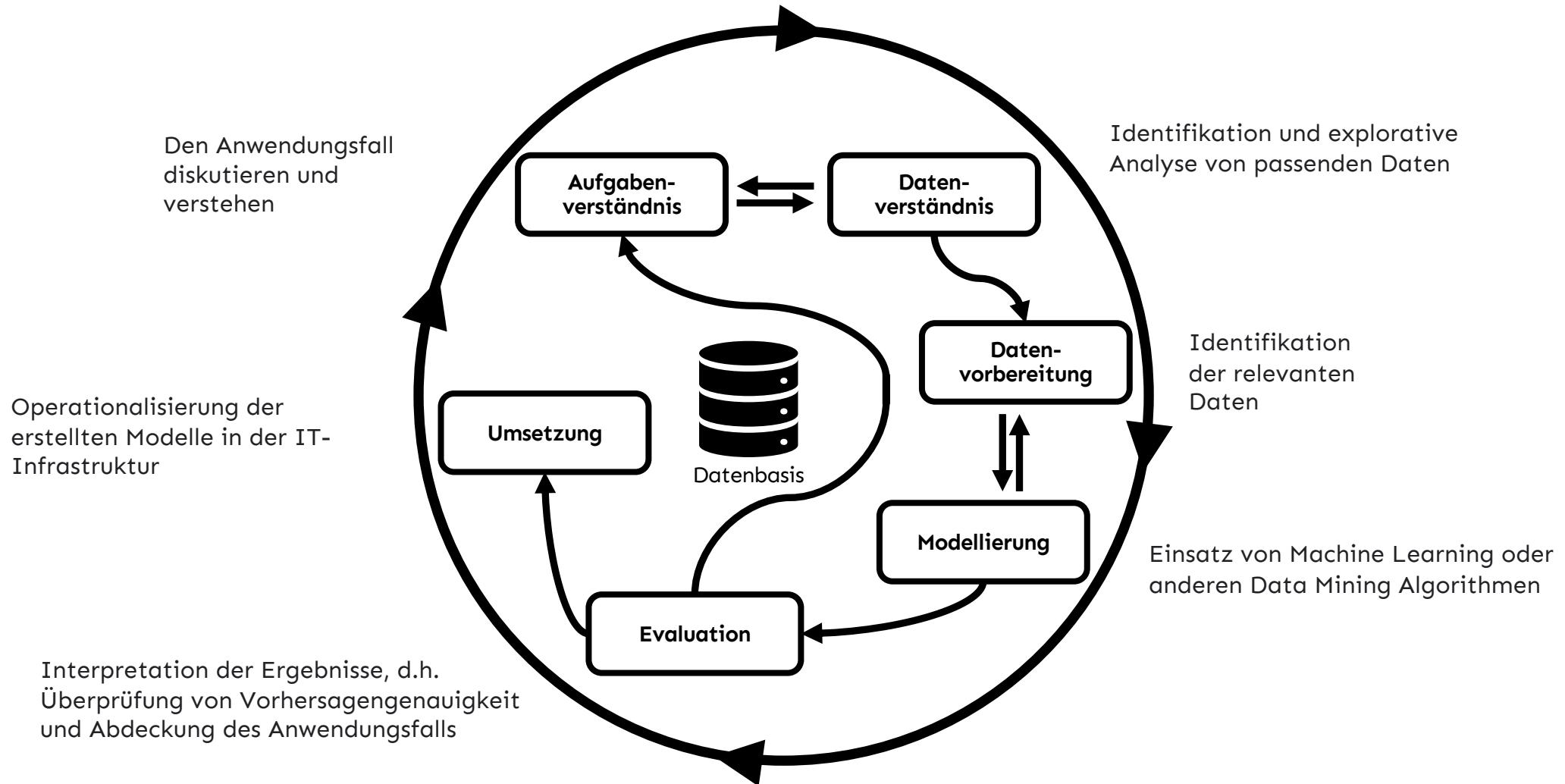


Machine Learning Workflow

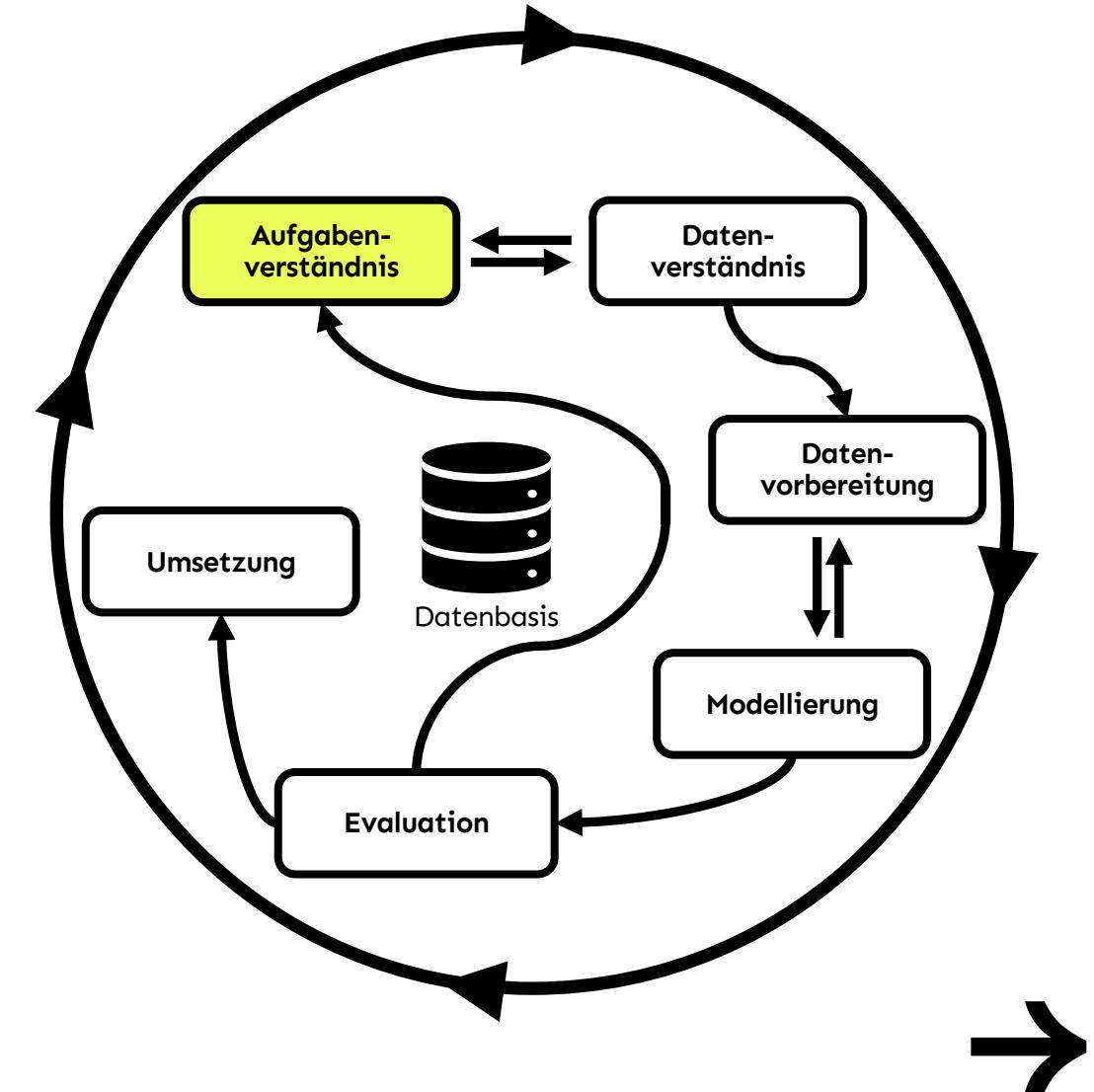


Der Machine Learning Workflow

Die verschiedenen Schritte innerhalb eines KI-Projekts



Aufgaben-verständnis



Der Machine Learning Workflow

Aufgabenverständnis - Verständnis des Anwendungsfall

Daten zu Immobilienangeboten in Melbourne.

Daten bezogen über *web scraping* von Domain.com.au im Zeitraum von 2016 - 2018.

Es soll eine Immobilienblase in diesem Zeitraum in Melbourne gegeben haben.

Lässt sich der Preis vorhersagen?

Was hat den größten Einfluss auf den Preis gehabt?

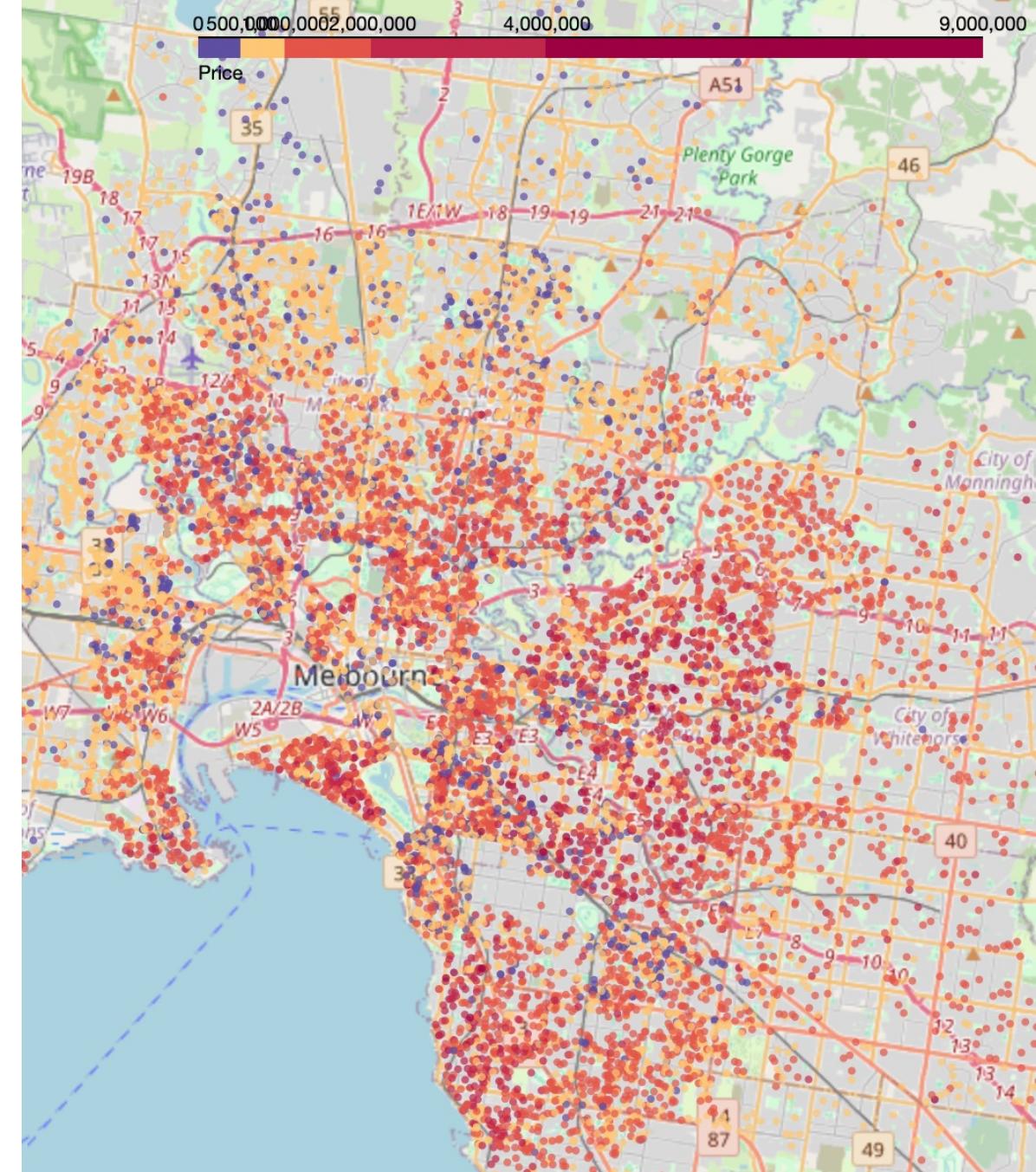
Hat sich dieser Einfluss mit der Zeit geändert?

<https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market>

Der Machine Learning Workflow

Aufgabenverständnis - Merkmale

Merkmal	Beschreibung
Longitude	Position Ost - West
Latitude	Position Nord - Süd
YearBuilt	Baujahr
Price	Preis
Method	Verkaufsart
Type	Immobilientyp
SellerG	Makler
Date	Einstellungsdatum
Distanz	Distanz zum CBD
Regionname	Himmelsrichtung
Propertycount	Immobilien im Bezirk
Bedroom2	# Schlafzimmer
Bathroom	# Badezimmer
Landsize	Grundstück
BuildingArea	Wohnfläche
CouncilArea	Bezirk



Der Machine Learning Workflow

Aufgabenverständnis - Entscheidende Fragen

Was ist der Anwendungsfall?

Welche Datenquellen gibt es?

Wie können die Datenquellen erreicht werden?

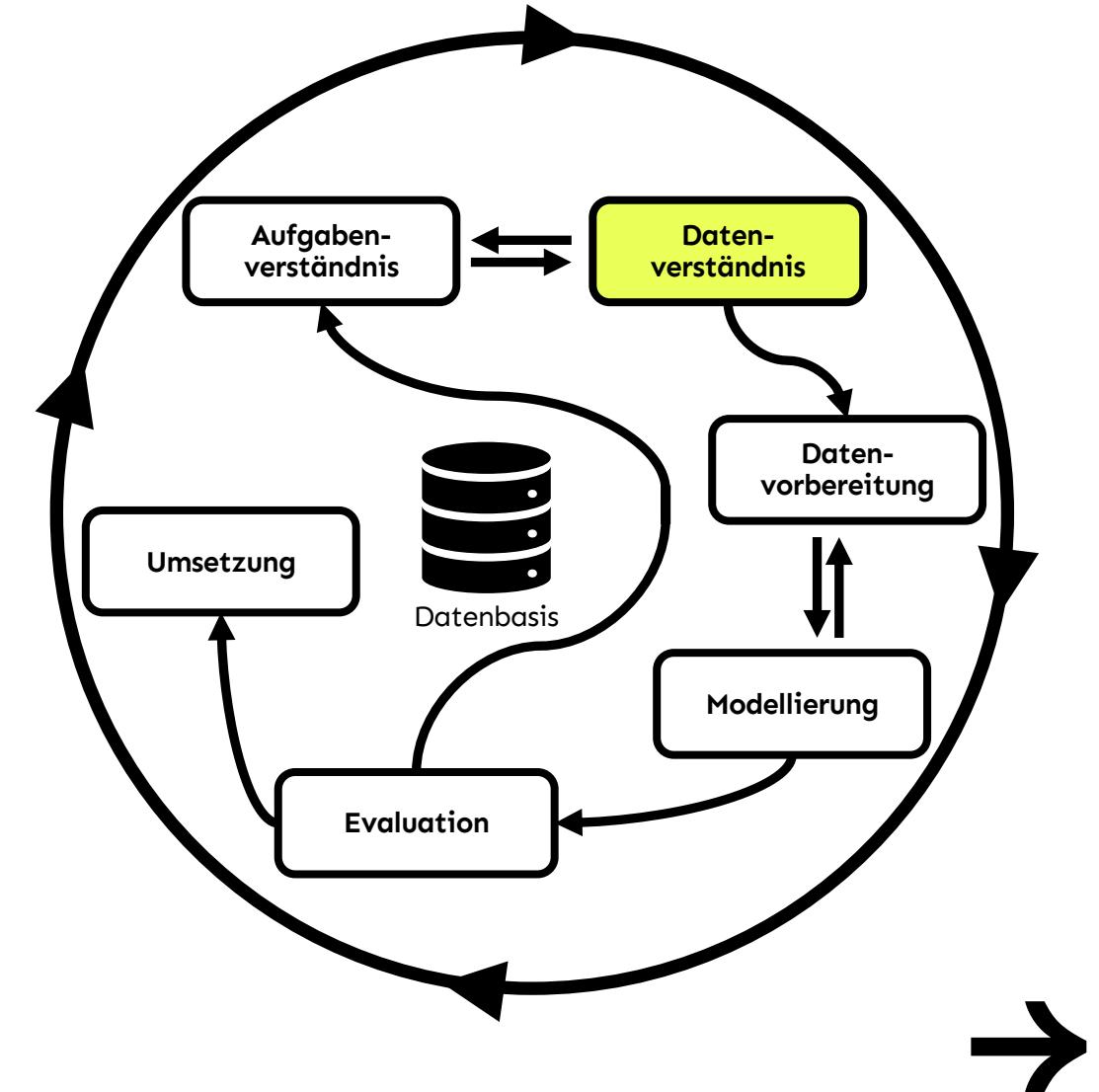
Gibt es bereits Anforderungen an das Modell (Genauigkeit, Interpretierbarkeit, etc.)?

Betrifft es vor- oder nachgelagerte Systeme?

Lässt sich die Lösung in einen bestehenden Workflow einbetten?

Was ist der Kundennutzen?

Daten-verständnis



Der Machine Learning Workflow

Datenverständnis - Was ist explorative
Datenanalyse?

Bevor man mit Machine Learning beginnen kann, muss man zuerst seine Daten besser verstehen!

„Explanatory data analysis is detective work“¹

„Die Arbeit eines guten Ermittlers zeichnet sich dadurch aus, dass er weiß, wonach es sich an einem Tatort zu suchen lohnt und welche Hilfsmittel er dazu benötigt.“

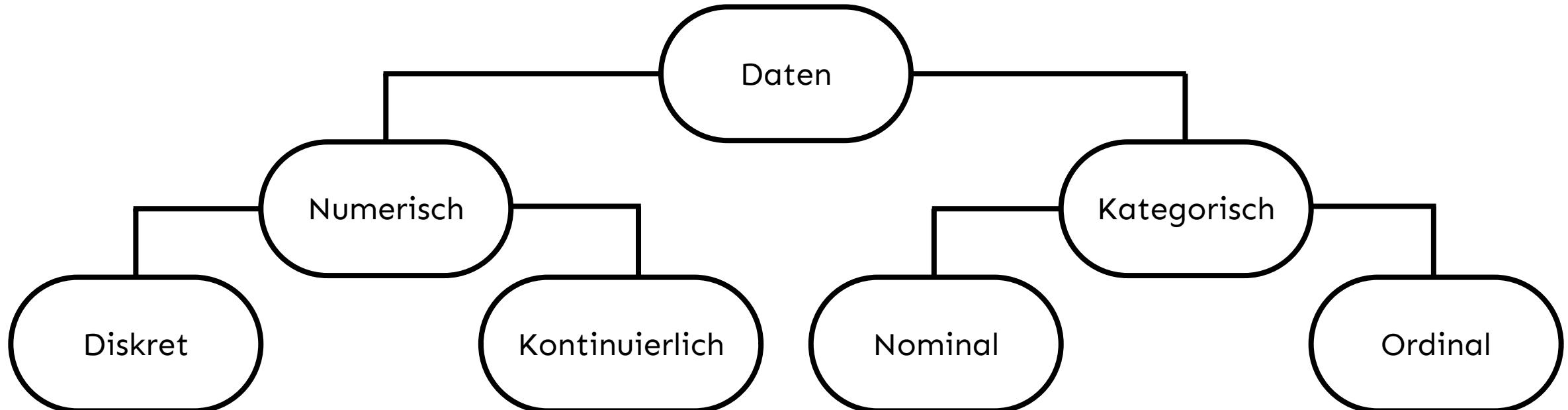
¹Tukey, J., W., Explanatory data analyses, 1977, S 1.

²Burkhardt, M., Sedlmeier, P., Explorative und deskriptive Datenanalyse mit R, 2015, S.9



Der Machine Learning Workflow

Datenverständnis - Arten von Merkmalen



Werte sind ganzzahlig:

- Anzahl Studierende
- Anzahl Bestellungen

Werte sind reelle Zahlen, üblicherweise innerhalb eines Bereichs:

- Temperatur
- Alter

Keine natürliche Reihenfolge zwischen den Kategorien:

- Geschlecht
- Länder
- Farbnamen

Eine Reihenfolge zwischen den Kategorien:

- T-Shirt Größen (S, M, L)
- Tageszeit (morgens, mittags, abends)

Der Machine Learning Workflow

Datenverständnis - Merkmale des Melbourne House Price Datensatz

Merkmal	Daten	Datentyp
Longitude	144.9, 145.1, 144.8	Numerisch
Latitude	-37,9, -37,7, -37,8	Numerisch
YearBuilt	1970, 2000, 1981	Numerisch
Price	880000, 541000, 535000	Numerisch
Method	S, S, NB	Kategorisch
Type	u, t, h	Kategorisch
SellerG	Fletchers, Nelson	Kategorisch
Date	2017-02-11, 2017-07-29	Datetime
Distanz	11.2, 8.5, 4.6	Numerisch
Regionname	Southern Metropolitan	Kategorisch
Propertycount	5457, 7485	Numerisch
Bedroom2	3, 2, 4	Numerisch
Bathroom	1, 1, 2	Numerisch
Landsize	217, 133, 771	Numerisch
BuildingArea	110, NaN, 312	Numerisch
CouncilArea	Moreland, Whitehorse	Kategorisch

Der Machine Learning Workflow

Datenverständnis - Ziele der explorativen Datenanalyse

Explorative Datenanalyse ist ein wichtiger erster Schritt bei der Analyse von Daten und der Erstellung von Predictive Applications

- Daten näher kennenlernen (Muster erkennen)
- Datenverteilung, Datenqualitätsprobleme, *Outlier*, Korrelationen / Beziehungen
- Aufstellung und prüfen von Thesen / Annahmen

Entscheidend um **relevante Merkmale** für eine Vorhersage zu finden

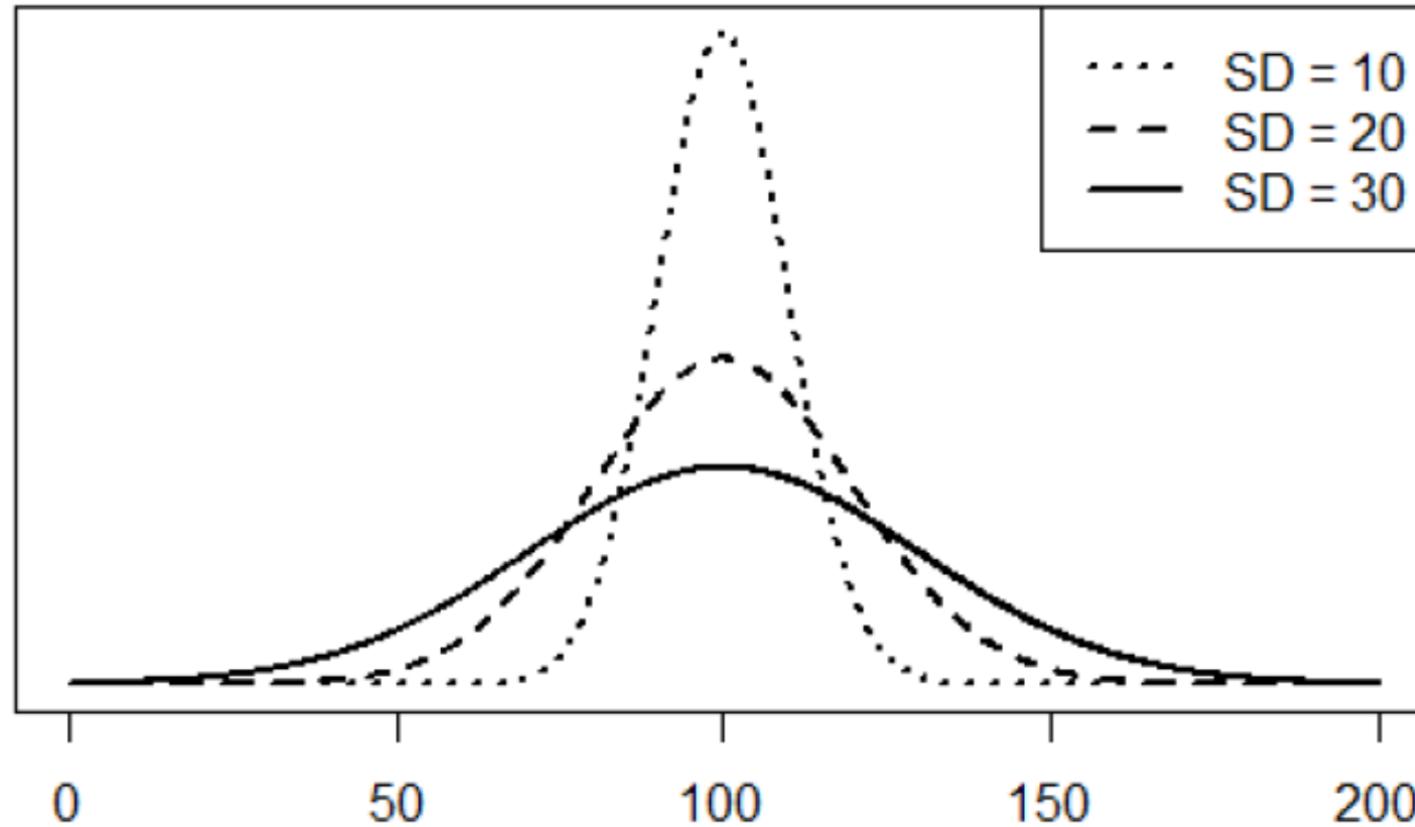
Ziel ist es folgende Aspekte so früh wie möglich zu adressieren

- Feststellen von **Fehlern** (evtl. in der Datensammlung / Verteilung)
- Zutreffen von **Annahmen**
- Grobe Untersuchung der **Beziehung** zwischen unabhängigen Variablen (möglichen Merkmalen) und abhängiger Variable (Zielgröße)

Der Machine Learning Workflow

Datenverständnis - Beispiel Varianz

3 Verteilungen mit gleichem Mittelwert und unterschiedlicher Varianz



Der Machine Learning Workflow

Datenverständnis - Kovarianz

Ist eng verwandt mit der **Korrelation**.

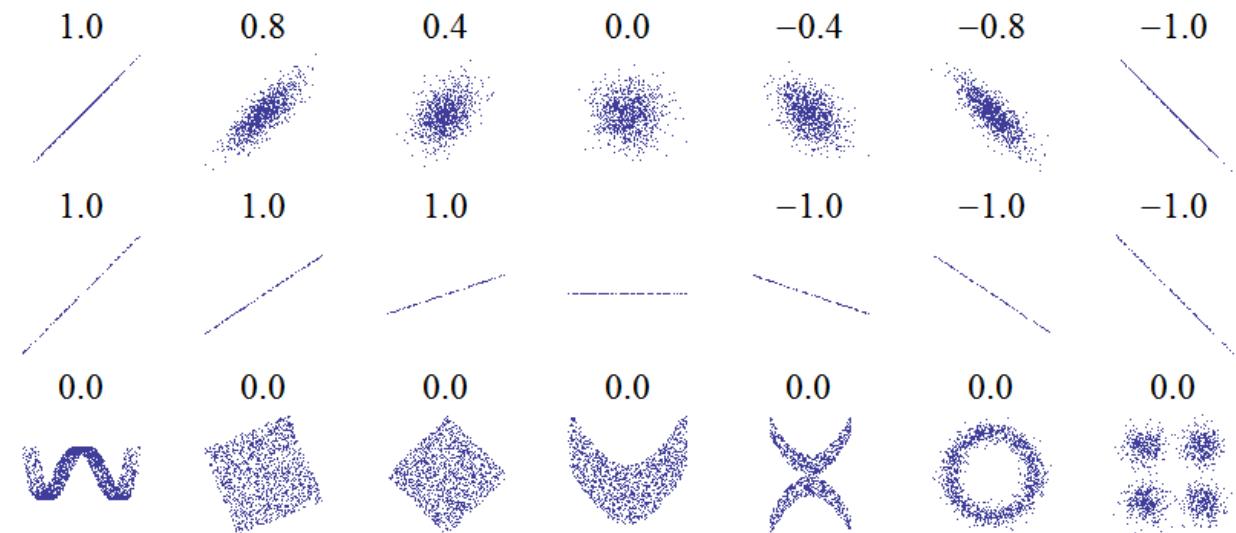
Maß für den **linearen Zusammenhang** zweier Variablen.

Nicht standardisiert, was es erschwert Rückschlüsse aus den Werten zu schließen.

Positives Vorzeichen: Beide Variablen bewegen sich in die **gleiche** Richtung.

Negatives Vorzeichen: Beide Variablen bewegen sich in **entgegengesetzte** Richtung.

Standardisierte Kovarianz ergibt Korrelation.

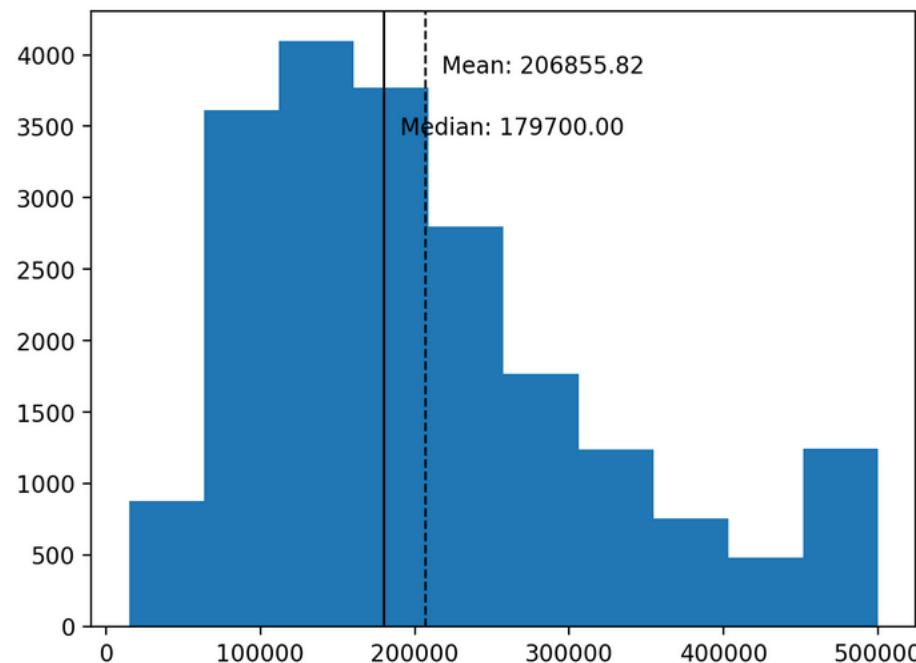


Der Machine Learning Workflow

Datenverständnis - Univariate Visualisierungsmethoden - Histogram

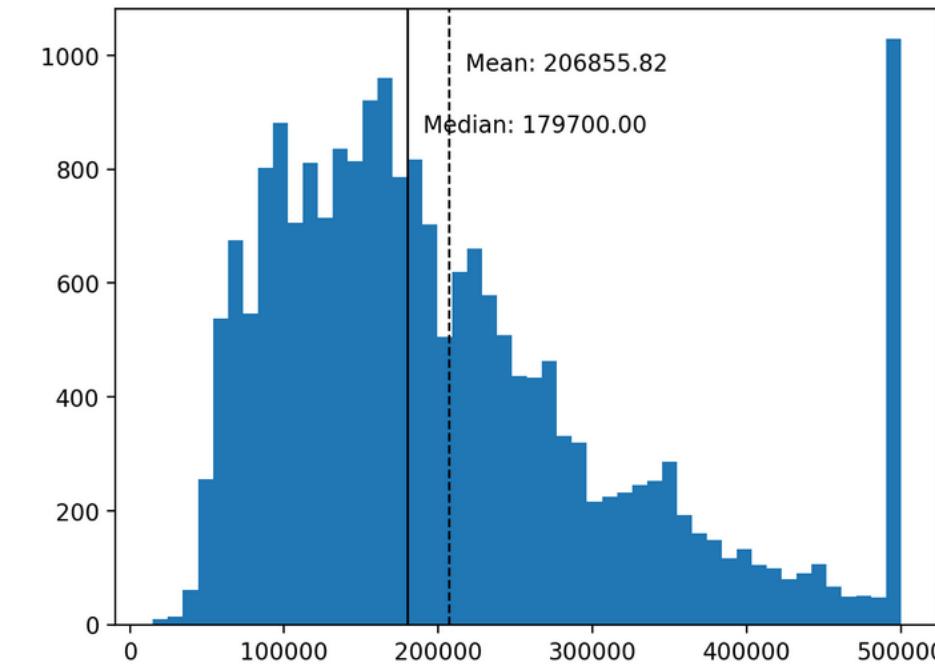
Stellt die **Häufigkeit** eines Merkmals dar

Gibt einen ersten Eindruck über die **Verteilung** der Daten



Es lassen sich die häufigsten Werte finden

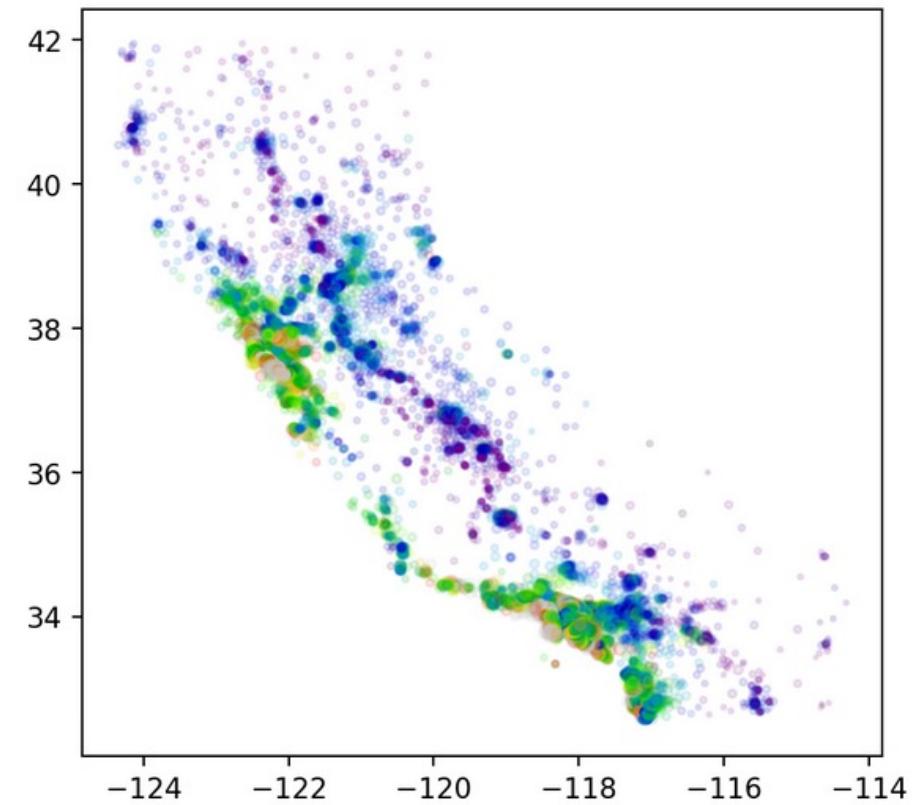
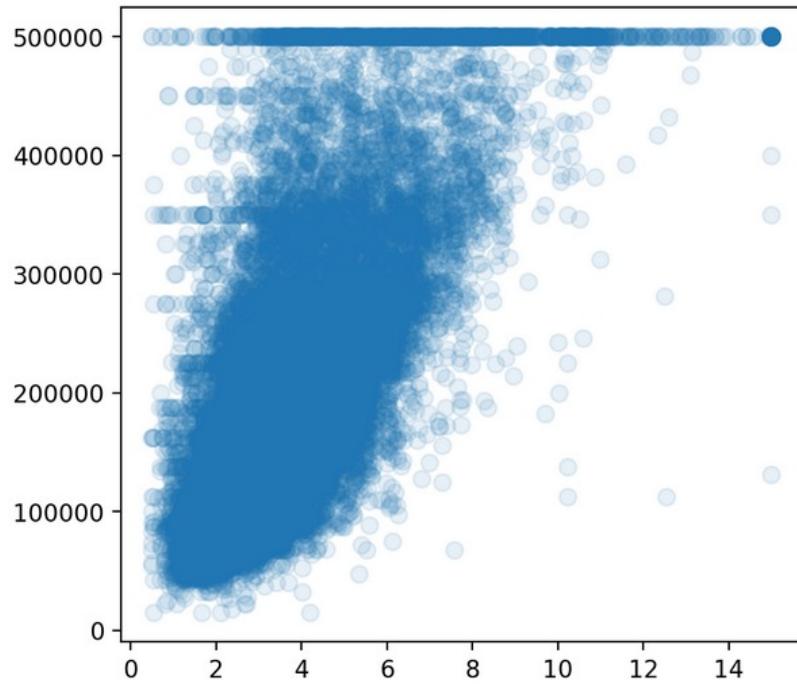
Durch *Binning* lässt sich die **Genauigkeit** und das **Rauschen** regeln



Der Machine Learning Workflow

Datenverständnis - Multivariate Visualisierungsmethoden - Scatterplot

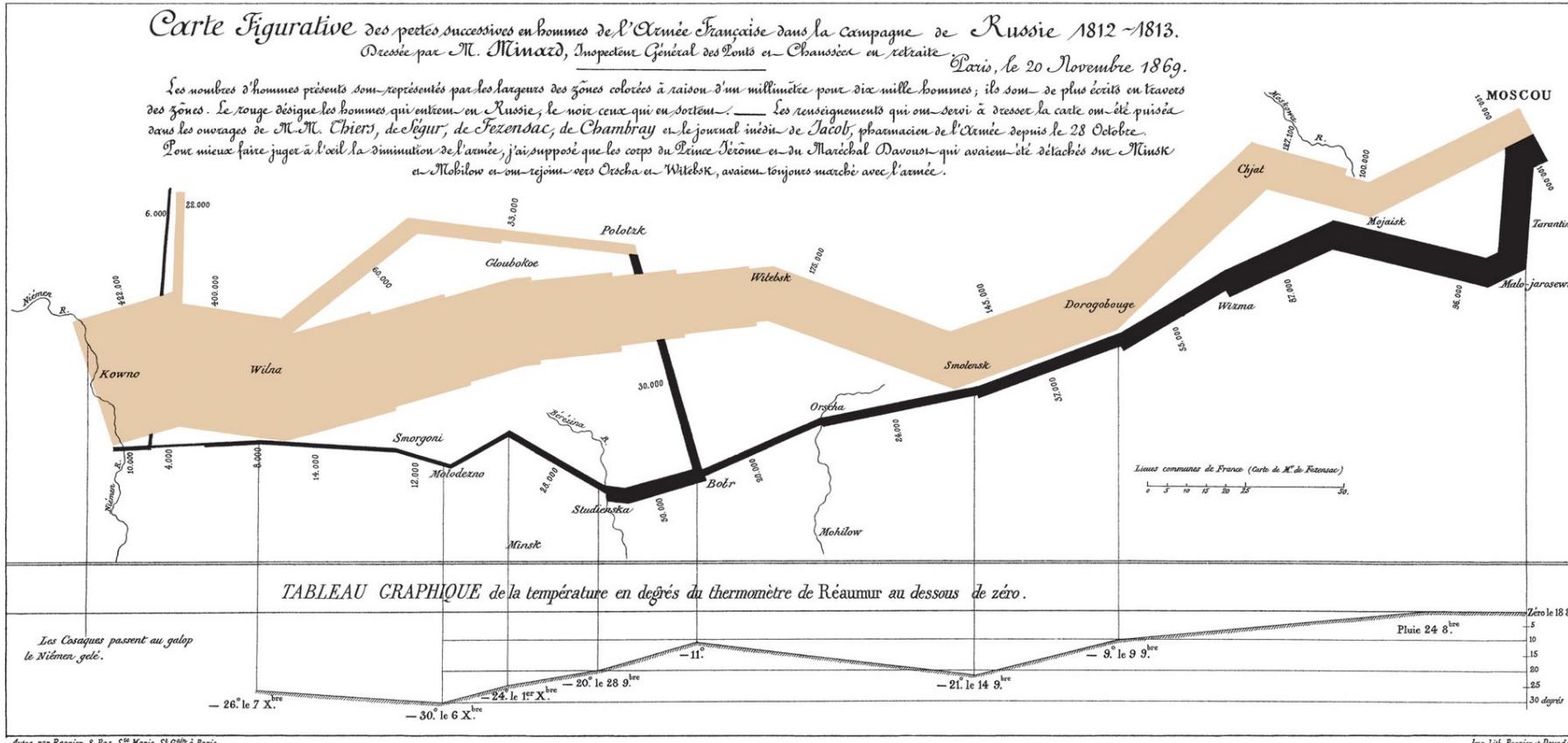
- Jeder Datenpunkt wird einzeln auf einem Graph abgetragen
- Durch Kombination von x- und y-Achse sowie Farbe und Größe können mehrere Merkmale gleichzeitig betrachtet werden
- Bietet sich zur Untersuchung von Beziehungen an



Der Machine Learning Workflow

Datenverständnis - Multivariate Visualisierungsmethoden - Data Storytelling

Karte von Charles Minards aus dem Jahre 1869 über den Russlandfeldzug Napoleons



https://de.wikipedia.org/wiki/Multivariate_Verfahren#/media/Datei:Minard.png

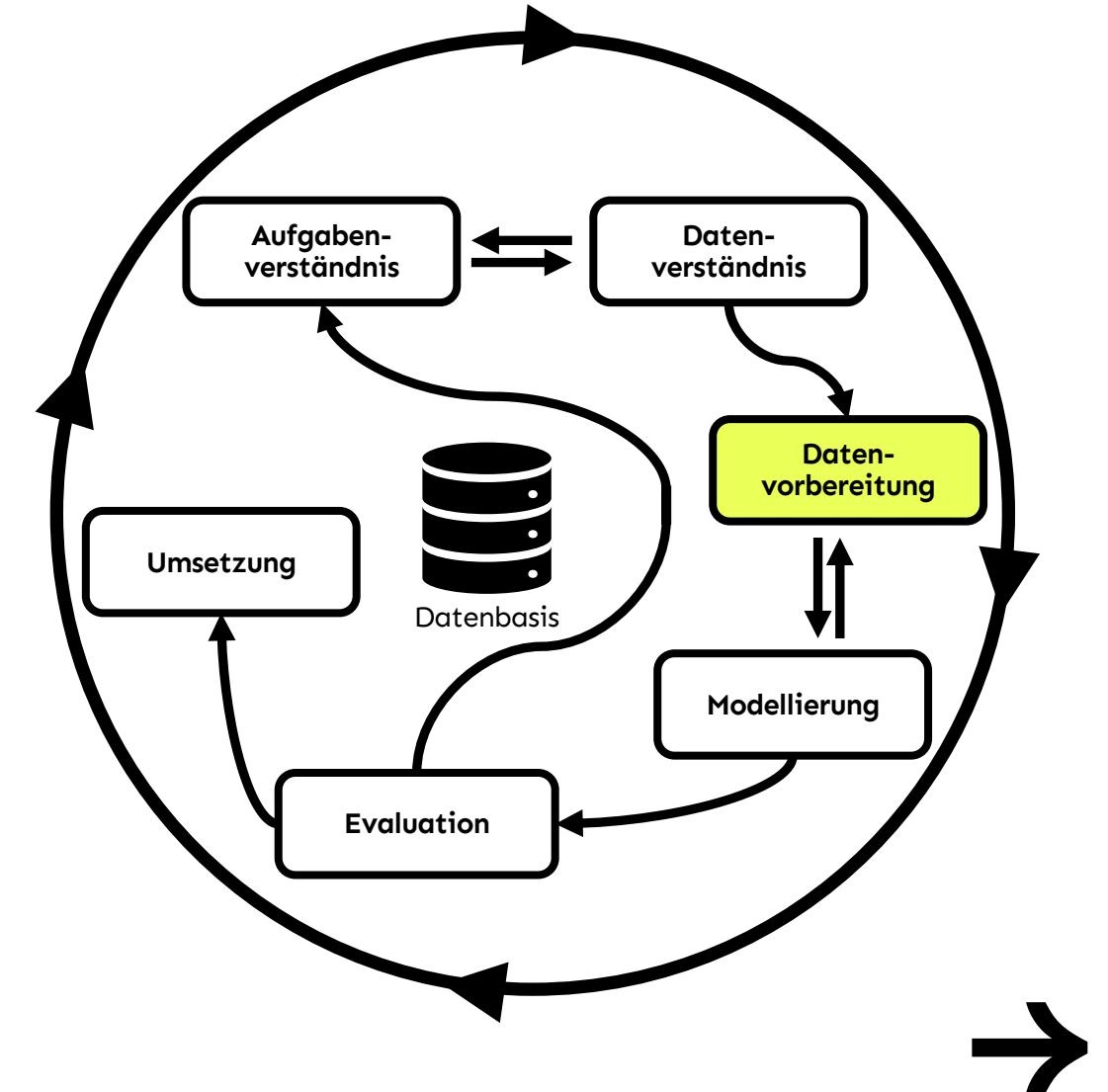
Interaktive Version: <https://www.masswerk.at/minard/>

Der Machine Learning Workflow

Datenverständnis - Entscheidende Fragen

- Wie können die Daten interpretiert werden?
- Was steckt hinter den Verteilungen?
- Was stellen Ausreißer da?
- Was sind sinnvolle Wertebereiche?
- Gibt es Verbindungen zwischen verschiedenen Merkmalen?

Daten- vorbereitung



Der Machine Learning Workflow

Datenvorbereitung - Überblick

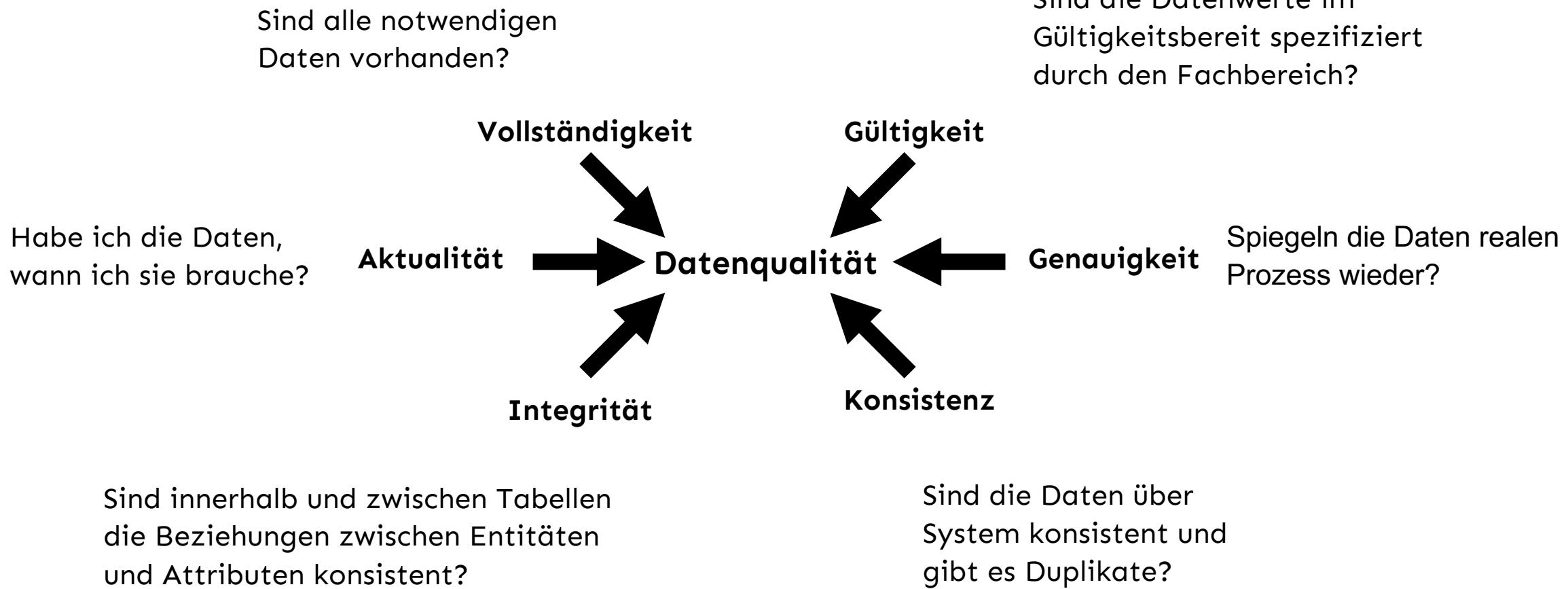
Reale Daten sind „unrein“: Möglicherweise enthalten sie inkorrekte Daten (Messfehler, menschliches Versagen oder Computerfehler, Übertragungsfehler, etc.):

- **Unvollständig:** Attribute fehlen
- **Verrauscht:** Verzerrte Daten, Fehler (Größe = -100 cm), Ausreißer
- **Inkonsistenzen:** Diskrepanzen zwischen verschiedenen Einträgen (Alter und Geburtstag).
- **Intentional Errors:** Versteckte Fehlwerte (Jeder fehlende Eintrag bei Geburtstag ist 01.01).

Der Ursprung von Fehlwerten kann vielfältig sein!

Der Machine Learning Workflow

Datenvorbereitung - Datenqualitätsmerkmale



Der Machine Learning Workflow

Warum Datenvorbereitung?

- Schlechte Datenqualität führt zu **geringer Qualität** der darauf aufbauenden Machine Learning **Ergebnisse**
- **Verbessert die Performance** von Vorhersagen
- **Modellierung setzt gute Datenqualität voraus.** Klassifikationsalgorithmen können grundsätzlich nicht mit Fehlwerten umgehen
- Datenaufbereitung, -säuberung und -transformation beanspruchen den **Hauptteil der Arbeit** bei Machine Learning Projekten

Der Machine Learning Workflow

Datenvorbereitung - Hauptaufgaben

Datenbereinigung

- Füllen von Fehlwerten, glätten von verrauschten Daten, Identifizieren und entfernen von Ausreißern und verrauschten Daten, auflösen von Inkonsistenten

Datenintegration

- Integration von mehreren Datenbanken und Dateien

Datentransformation

- Normalisierung und Aggregation
- Datendiskretisierung

Datenreduktion

- Reduzieren des Datenvolumens unter beibehalten derselben analytischen Ergebnisse

Der Machine Learning Workflow

Data Preprocessing - Fehlwerte

Aus der Statistik werden grundsätzlich 3 verschiedene Arten unterschieden:

Missing Completely at Random

- Das ein Wert fehlt ist **unabhängig** von den vorliegenden und fehlenden Datenwerten
- Beispiel: Ein Temperatursensors fällt wegen eines technischen Problems aus

Missing at Random

- Das ein Wert fehlt ist **teilweise** von den vorliegenden aber nicht den fehlenden Datenwerten **abhängig**
- Beispiel: Ein Temperatursensors fällt zufällig aus, aber meist Nachts

Missing Not at Random

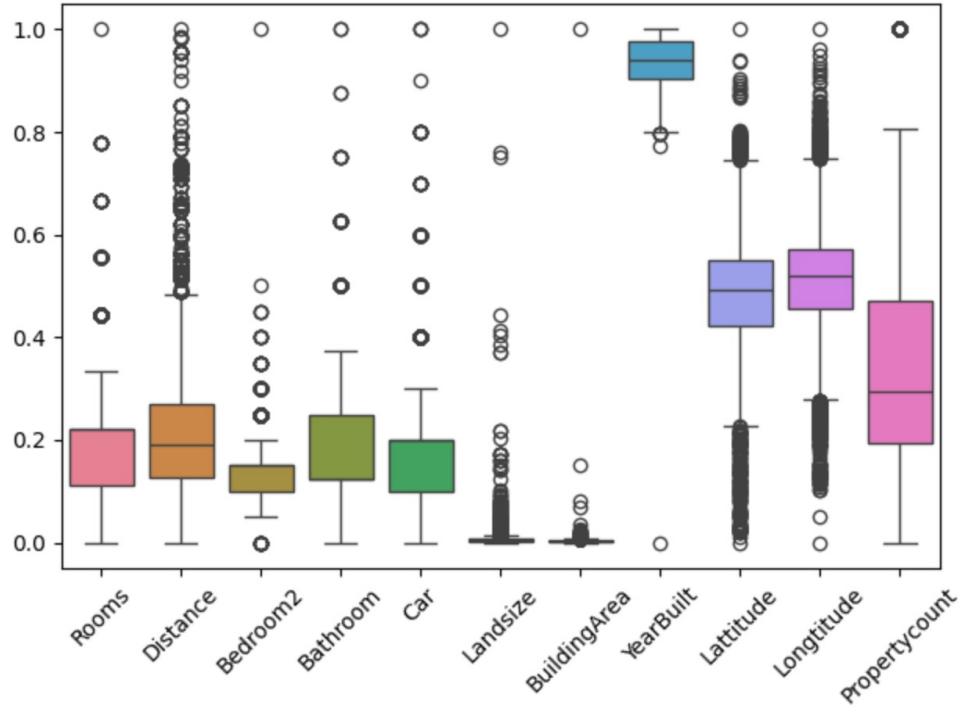
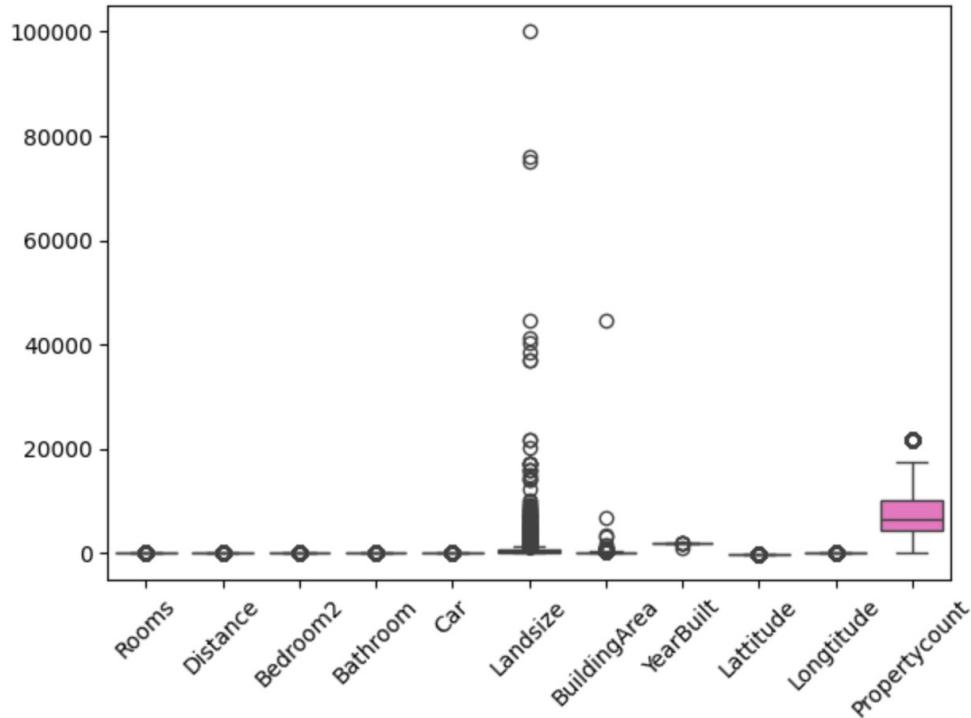
- Das ein Wert fehlt **hängt von den Fehlwerten ab**
- Beispiel: Ausfall des Temperatursensors bei extremen Temperaturen

Der Machine Learning Workflow

Datenvorbereitung - Normalisierung

Normalisierung überführt Werte in einen definierten Wertebereich, z.B. 0 - 1, Varianz 1 und Mittelwert 0.

- Relevant wenn Merkmale in sehr unterschiedlichen Wertebereichen liegen.
- Erhöht die Robustheit eines Modells, sprich den Einfluss, den einzelne Beobachtungen haben.
- Erhöht die Trainingsgeschwindigkeit.



Der Machine Learning Workflow

Datenvorbereitung - Encoding

Viele Machine Learning Algorithmen können nur mit **numerischen** Merkmalen umgehen

- Encoding zielt darauf ab, kategorische in numerische Merkmale zu konvertieren
- Es gibt verschiedene Möglichkeiten:
 - **Label Encoding**
 - **One-Hot Encoding**

Der Machine Learning Workflow

Datenvorbereitung - Encoding

Viele Machine Learning Algorithmen können nur mit **numerischen** Merkmalen umgehen

Stichprobe	Kategorie
1	Whitehorse
2	Whitehorse
3	Moreland
4	Yarra
5	Coburg

Label
Encoding

Stichprobe	Kategorie	Numerisch
1	Whitehorse	1
2	Whitehorse	1
3	Moreland	2
4	Yarra	3
5	Coburg	4

Algorithmen könnten das Merkmal als ordinal skaliert interpretieren

One Hot
Encoding

Stichprobe	Whitehorse	Moreland	Yarra	Coburg
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1

One-Hot Encoding ist ähnliche zu Pivottabellen

Der Machine Learning Workflow

Datenvorbereitung - Probleme mit zu vielen Merkmalen

Eine (zu) große Zahl an Variablen kann zu nachteiligen Effekten führen:

Überanpassung (Overfitting): Das trainierte Model generalisiert nicht auf unbekannte Daten, da neben dem Signal auch das Rauschen gelernt wurde.

Kolinearität: Merkmale, mit einer hohen linearen Korrelation zueinander, enthalten praktisch die gleiche Information und sind somit redundant. Dies kann zu instabilen Modellen führen, sprich man erhält sehr unterschiedliche Modelle auch wenn sie auf dem gleichen Datensatz trainiert wurden.

Fluch der Dimensionen: Je mehr Merkmale vorhanden sind, desto mehr Kombinationen sind möglich.

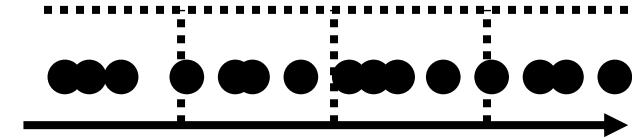
Schlechte Interpretierbarkeit: Es ist schwerer zu erkennen, welche Merkmale für eine Entscheidung des Modells relevant waren.

Der Machine Learning Workflow

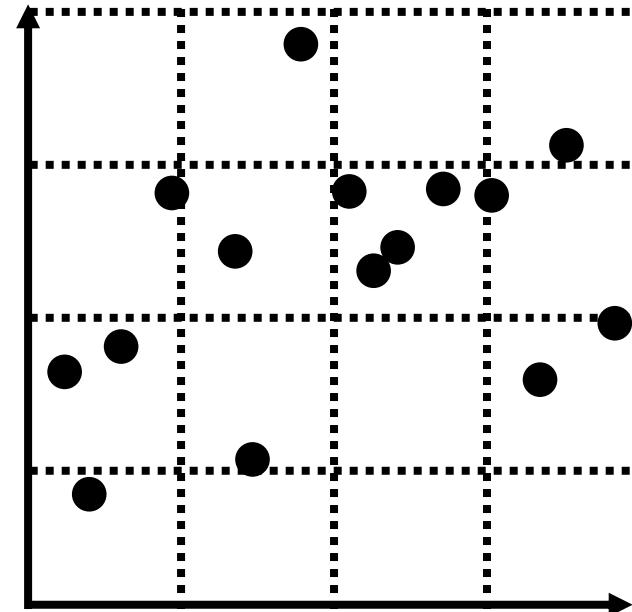
Datenvorbereitung - Fluch der Dimensionalität

- Jede Einflussgröße hat verschiedene Ausprägungen
- Mit **steigender Dimensionalität schrumpft** der Raum, den wir mit gleichbleibender Datenmenge abdecken können, **exponentiell**.

1 Dimension



2 Dimensionen



Der Machine Learning Workflow

Datenvorbereitung - Auswählen von Merkmale

Auswahl der Merkmale nach **Domänenwissen**

Verwendung von Algorithmen zur **Feature Selection**. Manche ML-Modelle können relevante Merkmale auswählen. Sehr verbreitet ist z.B. Regularisierung, wobei jede Einflussgröße mit Kosten verbunden ist.
Beispiele: Lasso (L1), Ridge-Regression (L2)

Einsatz von Algorithmen zur **Dimensionsreduktion**. Durch Kombination von Merkmalen miteinander wird die Anzahl an Merkmalen verringert.
Beispiele: Hauptkomponenten Analyse (PCA)

Der Machine Learning Workflow

Datenvorbereitung - Feature Engineering

Feature Engineering ist ein Prozess indem neue sehr **relevante Merkmale erzeugt** werden, um das Model zu trainieren.

Es geht darum die Faktoren zu identifizieren, welche die Zielgröße beeinflussen.

Die Qualität der Merkmale hat einen bedeutenden Einfluss auf die Performance und Qualität des ML-Modells.

- Erzeugen gänzlich neuer Merkmale
- Veränderung an existierenden Merkmalen
- Extraktion von Informationen aus existierenden Merkmalen
- Aggregation von existierenden Merkmalen

Im Gegensatz zu vielen anderen Schritten lässt sich Feature Engineering nicht automatisieren, da es sich maßgeblich um einen kreativen Prozess handelt, der Domänenwissen voraussetzt.

Der Machine Learning Workflow

Datenvorbereitung - Arten des Feature Engineerings

Methode	Beispiel
Erweiterung um externe Merkmale	Hinzuziehen neuer Datenquellen (bspw. Wetterdaten zu einer bestimmten Geo-Position)
Erstellung neuer Variablen	Erweitern einer vorhanden Datenquelle (z.B. Urlaubstage, Wochenenden)
Information aus vorhandenen Variablen ziehen	Informationen aus E-Mail-Adressen (Name, Land, kostenpflichtig, etc.) Extraktion von Jahr, Monat, Tag, Uhrzeit, Wochentag, Wochenende, Ferien aus einer Datumsangabe
Modifizieren von vorhandenen Variablen	Alter statt Geburtsdatum, Skalierung von Variablen, Nicht-lineare Transformationen wie Potenz, Logarithmus, Wurzel, Trigonometrische Funktionen
Aggregation von Variablen	Zusammenfassen von granularer Information (Sekunden zu Tagen)

Der Machine Learning Workflow

Datenvorbereitung - Entscheidende Fragen

- Welche Daten liegen vor?
- Gibt es ergänzende Datenquellen?
- Wie soll mit Fehlwerten umgegangen werden?
- Was sind wichtige und unwichtige Merkmale?

Der Machine Learning Workflow

Datenvorbereitung - Übung

Mit welchen Daten wird gearbeitet?
Wie sollten diese Verarbeitet werden?

Der Machine Learning Workflow

Datenvorbereitung - Übung zu komplexen Daten

Message-ID: <23440430.1075840343530.JavaMail.evans@thyme>

Date: Tue, 5 Feb 2002 16:40:23 -0800 (PST)

From: infrastructure.ubsw@enron.com

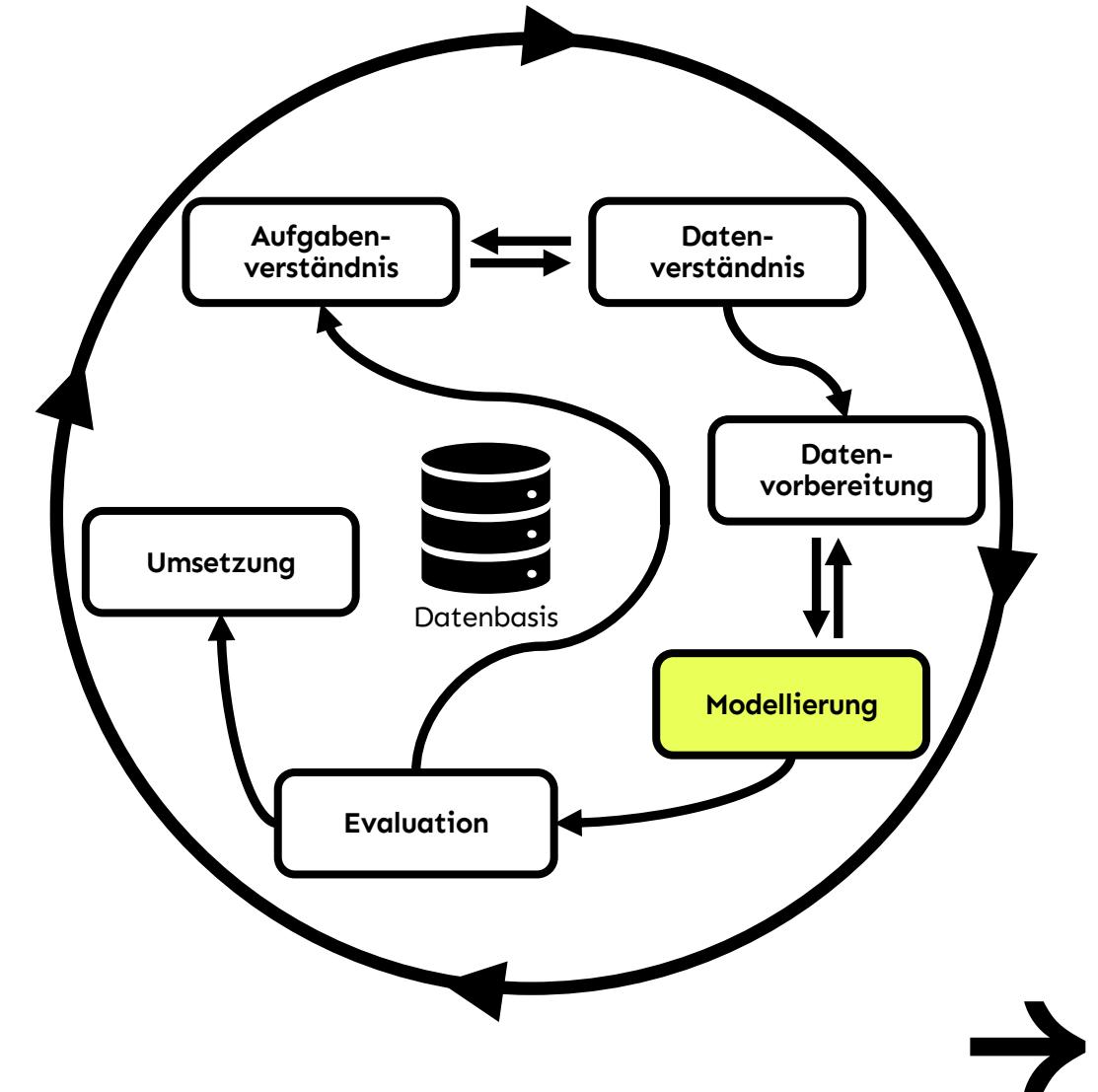
To: canada.dl-ubsw@enron.com, houston.dl-ubsw@enron.com, portland.dl-ubsw@enron.com

Subject: Quick Tips for the UBSWE migration

As of start of business, Wednesday, February 6th, you will have been migrated to the UBSW Energy environment. Here are a couple of quick tips and reminders to get you going:

- ? You will log in with your Enron NT ID, this will not change
- ? You will be asked to change your password, follow the standard Enron rules
- ? Your desktop will look the same
- ? Email will not be affected until Day 1, on which you will have your new UBSWE email address
- ? All compliant data and email should be copied to the UBSWE environment no later than Midnight (CST), Thursday, February 7, 2002 (see Data Migration Protocol email for compliancy direction)
- ? No data or emails are to be deleted from the system

Modellierung



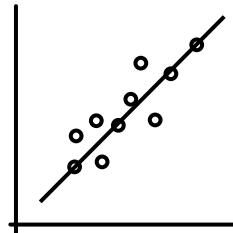
Der Machine Learning Workflow

Modellierung - Die 2 verbreitetsten Typen von Machine Learning

Supervised und unsupervised learning sind die verbreitetsten Typen von Machine Learning

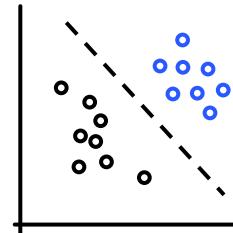
Supervised Learning
(Überwachtes Lernen)

Regression



Vorhersage einer
kontinuierlichen
Variable

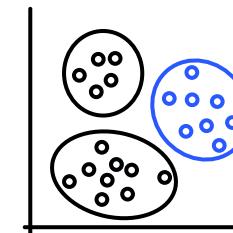
Klassifikation



Vorhersage einer
kategorischen
Variable

Unsupervised Learning
(Unüberwachtes Lernen)

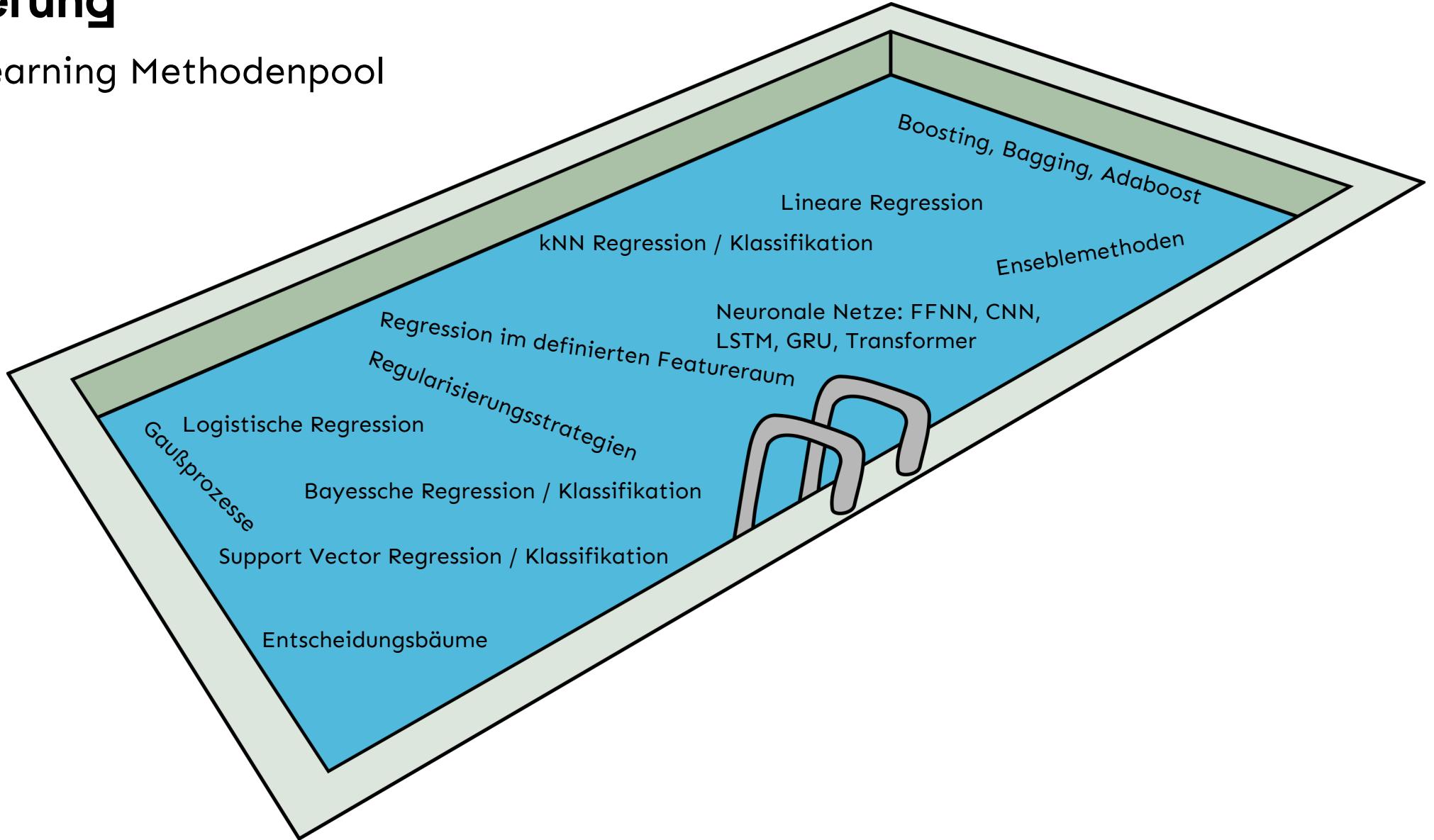
Clustering



Einteilen von
Beobachtungen in
Gruppen

Modellierung

Machine Learning Methodenpool



Der Machine Learning Workflow

Modellierung - Machine Learning Methoden

Supervised Learning:

- Lineare Regression
- Entscheidungsbäume

Clustering:

- K-Means
- DBSCAN

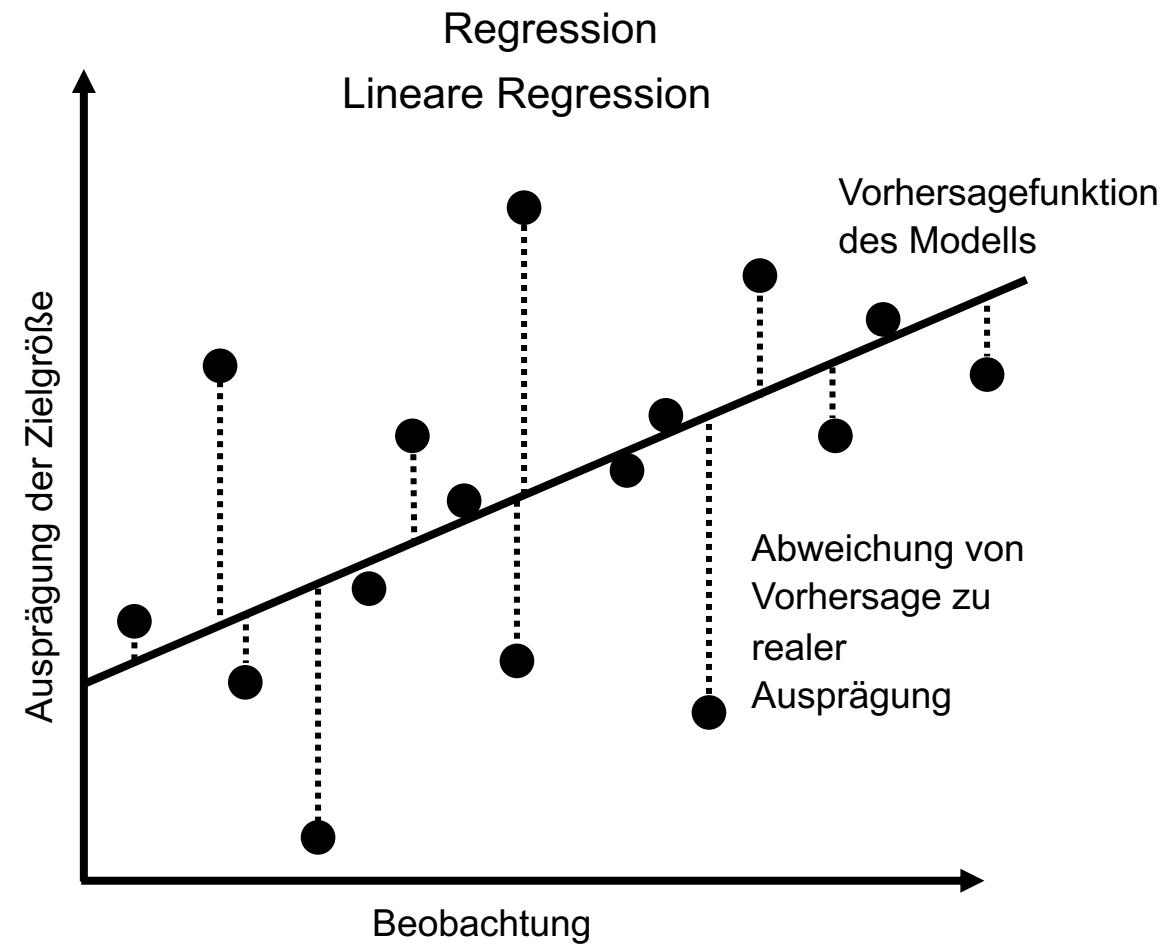
Artifizielle neuronale Netzwerke

- Perzepron
- Deep Learning

Der Machine Learning Workflow

Modellierung - Lineare Regression

- Lineare Regressionsalgorithmen zeigen oder prognostizieren die Beziehung zwischen zwei Variablen oder Faktoren.
- Passen eine kontinuierliche gerade Linie an die Daten an.
- Die Linie wird häufig mit der Kostenfunktion "quadratischer Fehler" berechnet.



Lineare Regression

Regressionsmodell

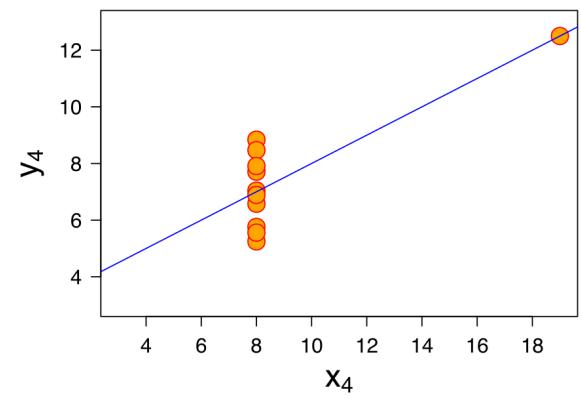
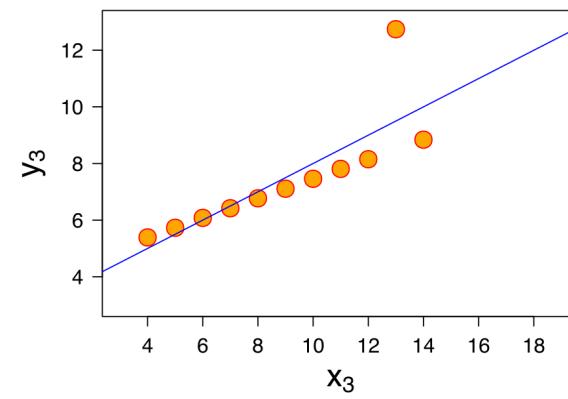
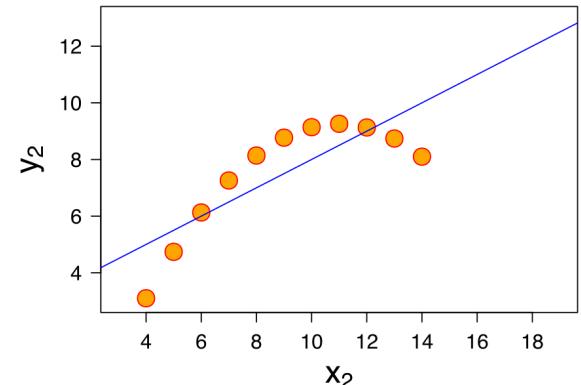
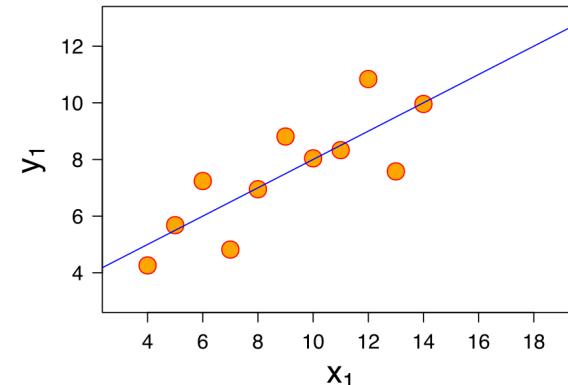
- Wir haben eine Sammlung von gelabelten Daten $\{(x_i, y_i)\}_{i=1}^N$, wobei N die Größe der Sammlung ist, x_i der D -dimensionale Merkmalsvektor der Daten $i = 1, \dots, N$, y_i ist ein reell-wertiges Ziel und jedes Merkmal $x_i^{(j)}, j = 1, \dots, D$ ist ebenfalls eine reelle Zahl.
- Wir wollen ein Modell $f_{w,b}(x)$ als lineare Kombination von Merkmalen des Datensatzes x erstellen:

$$f_{w,b}(x) = wx + b$$

wobei w ein D -dimensionaler Vektor von Parametern ist und b eine reelle Zahl ist.

- Wir werden das Modell verwenden, um die Unbekannte y für ein gegebenes x wie folgt vorherzusagen: $y \leftarrow f_{w,b}(x)$. Wir wollen die optimalen Werte (w^*, b^*) finden.
- Die Zielfunktion, die wir minimieren wollen, ist

$$\frac{1}{N} \sum_{i=1}^N (f_{w,b}(x_i) - y_i)^2 \text{ (squared error loss).}$$



Modellierung

Lineare Modelle

- Regressionsmethode
- Modellierung durch gewichtete Summe, alle Merkmale sind unabhängig voneinander
- Grundlegend **lineares Modell**: ungeeignet für nicht lineare (komplexe) Zusammenhänge
- Bieten eine einfache und schnelle **Baseline**, um die Performance komplexerer Modelle zu bewerten
- Können **underfitten** bei sehr vielen Beobachtung aber wenigen Merkmalen
- Sind schwer zu schlagen, wenn sehr viele Merkmale vorliegen
- Merkmale müssen vorher **normalisiert** werden
- Sind sehr gut zu interpretieren und geben darüber hinaus eine Aussage zur Feature Importance

Logistische Regression

Klassifikationsmodell

- Bei der logistischen Regression wollen wir y_i immer noch als lineare Funktion von x_i modellieren, aber jetzt ist y_i binär
- Wenn wir das negative Label als 0 und das positive Label als 1 definieren, müssten wir nur eine einfache kontinuierliche Funktion finden, deren Kodomäne (0,1) ist
- Eine Funktion mit dieser Eigenschaft ist die logistische Standardfunktion (auch bekannt als Sigmoid):

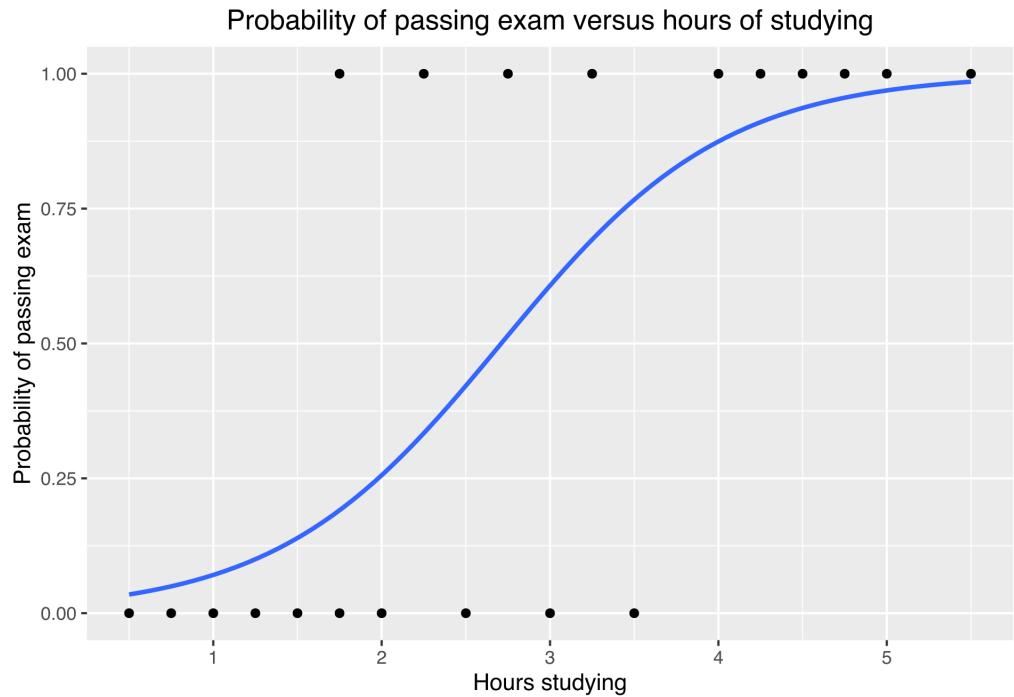
$$f(x) = \frac{1}{1 + e^{-x}}$$

- Das logistische Regressionsmodell sieht wie folgt aus

$$f_{w,b}(x) = \frac{1}{1 + e^{-(wx+b)}}.$$

- Wenn wir die Werte von w und b entsprechend optimieren, können wir die Ausgabe von $f(x)$ als die Wahrscheinlichkeit interpretieren, dass y_i positiv ist.
- Die Zielfunktion der logistischen Regression ist die Wahrscheinlichkeit unserer Trainingsmenge gemäß dem Modell:

$$L_{w,b} = \prod_{i=1 \dots N} f_{w,b}(x_i)^{y_i} \left(1 - f_{w,b}(x_i)\right)^{1-y_i} \text{ (Annahme iid)}$$



Logistische Regressionsalgorithmen zeigen oder prognostizieren die Beziehung zwischen zwei Variablen oder Faktoren, indem sie eine kontinuierliche S-Kurve an die Daten anpassen.

Modellierung

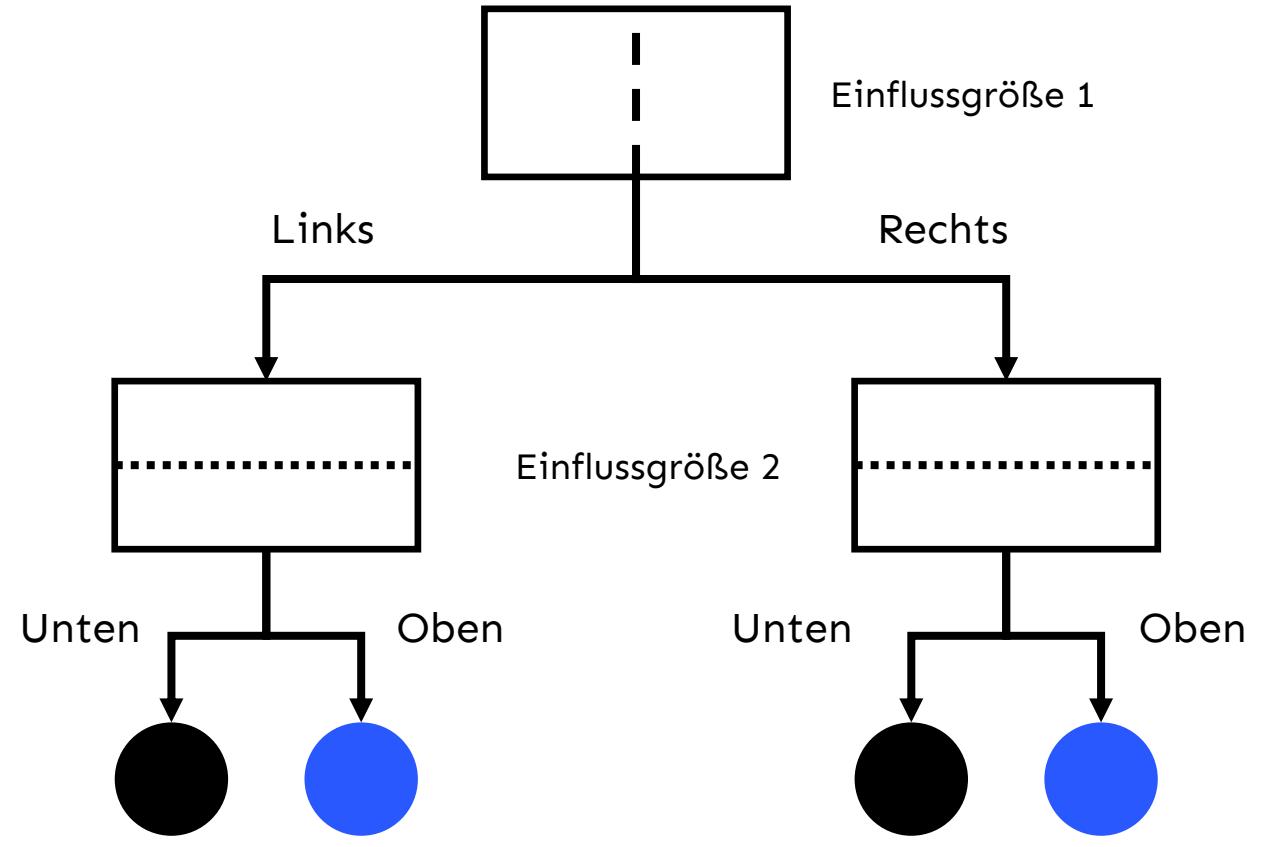
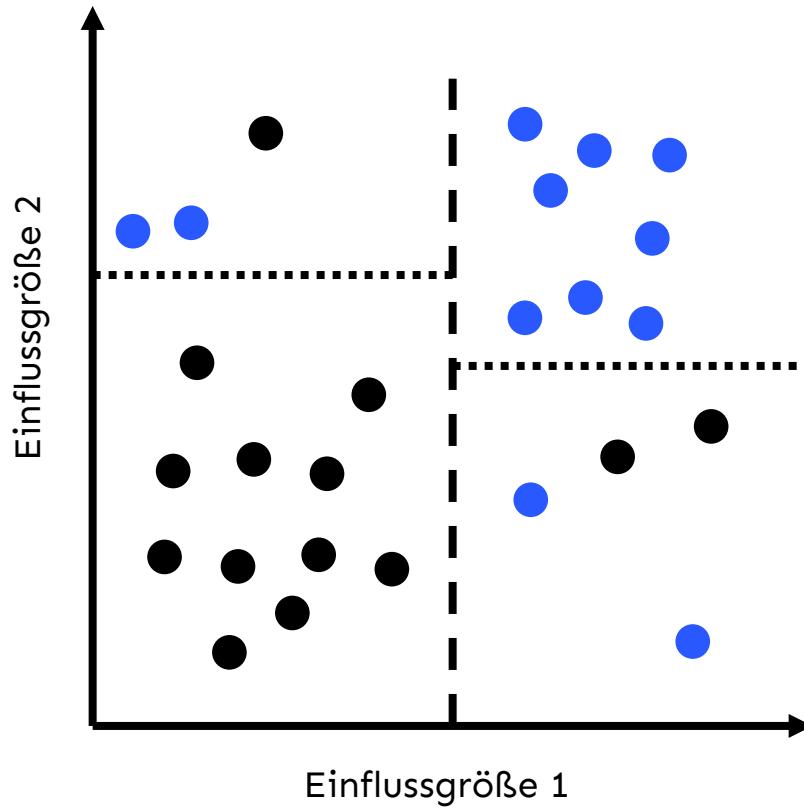
Logistische Regression

- Klassifikationsmodell, dass ein lineares Modell in Wahrscheinlichkeiten umwandelt
- Modellierung durch gewichtete Summe, alle Merkmale sind unabhängig voneinander
- **Nicht-lineares Modell:** kann komplexe Zusammenhänge abbilden
- Bieten eine einfache und schnelle **Baseline** um die Performance komplexerer Modelle zu bewerten
- Es kann sich leicht auf mehrere Klassen (multinomiale Regression) und eine natürliche probabilistische Ansicht von Klassenvorhersagen erstrecken
- Sind schwer zu schlagen, wenn sehr viele Merkmale vorliegen
- Gute Genauigkeit für viele einfache Datensätze und gute Leistung, wenn der Datensatz linear trennbar ist
- Merkmale müssen vorher **normalisiert** werden
- Sind sehr gut zu interpretieren und geben darüber hinaus eine Aussage zur Feature Importance
- Wenn die Anzahl der Beobachtungen geringer ist als die Anzahl der Merkmale, kann es zu einer Überanpassung kommen
- Es ist schwierig, komplexe Beziehungen mithilfe der logistischen Regression zu erhalten. Leistungsfähigere Algorithmen wie neuronale Netze können diesen Algorithmus leicht übertreffen.

Der Machine Learning Workflow

Modellierung - Entscheidungsbäume

Einen Entscheidungsbaum wachsen lassen.



Der Machine Learning Workflow

Modellierung - Entscheidungsbäume

Folge von einfaches Entscheidungsregeln: Eine Einflussgröße und ein Schwellenwert nach dem anderen.

Keine Skalierung und Normalisierung der Merkmale notwendig.

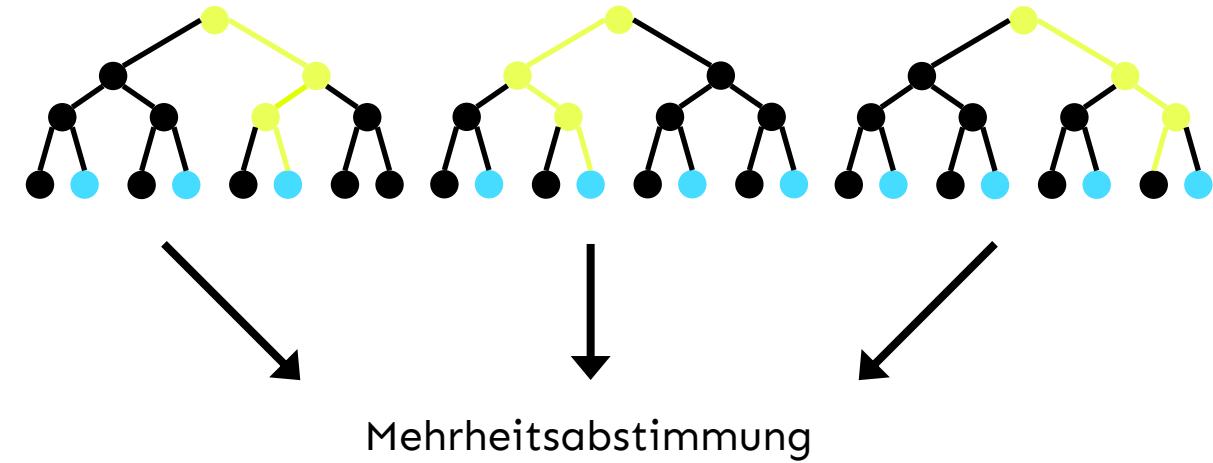
Etwaiges *under-* oder *overfitting* kann durch Anpassung der Modellparameter verhindert werden.

Klassifikationen sind **einfach nachzuvollziehen** und zu erklären.

Der Machine Learning Workflow

Modellierung - Ensemble Methoden

- Ensemble Methoden kombinieren mehrere einfache Modelle zusammen. Die letztliche Vorhersage ist die Klasse, die von der **Mehrheit der Modelle vorhergesagt** wird (oder der Durchschnitt für die Regression).
- Jedes einfache Modell wird auf einer unterschiedlichen **Teilmenge** aller **Beobachtungen und Merkmale** trainiert.



Der Machine Learning Workflow

K-Means (<https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>):

- Welche Rolle spielt die Auswahl des initiales Centroids? Probiert „I'll Choose“ und „Randomly“ aus.
- Was ist eine gute Anzahl an Centroiden?
- Ohne visuelle Überprüfung, was könnte eine Metrik sein diese Anzahl zu bestimmen?

DBSCAN (<https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>):

- Welche Auswirkungen haben epsilon und minPoints?
- Ohne visuelle Überprüfung, was könnte eine Metrik sein, um gute Werte für epsilon und minPoints zu finden?

Betrachtet die Datensätze:

- Uniform Points
- Gaussian Mixture
- Packed Circles
- Pimpled Smiley

Der Machine Learning Workflow

Modellierung - Artifizielle neuronale Netze und Deep Learning

Warum jetzt?

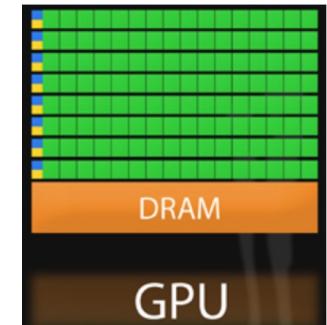
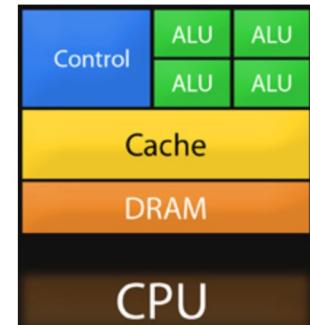
- Bessere Hardware (**GPUs**)
- **Mehr Daten** vorhanden

Neue Erkenntnisse zum Training neuronaler Netzwerke

- Parameter Initialisierung
- Aktivierungsfunktionen
- Optimierer



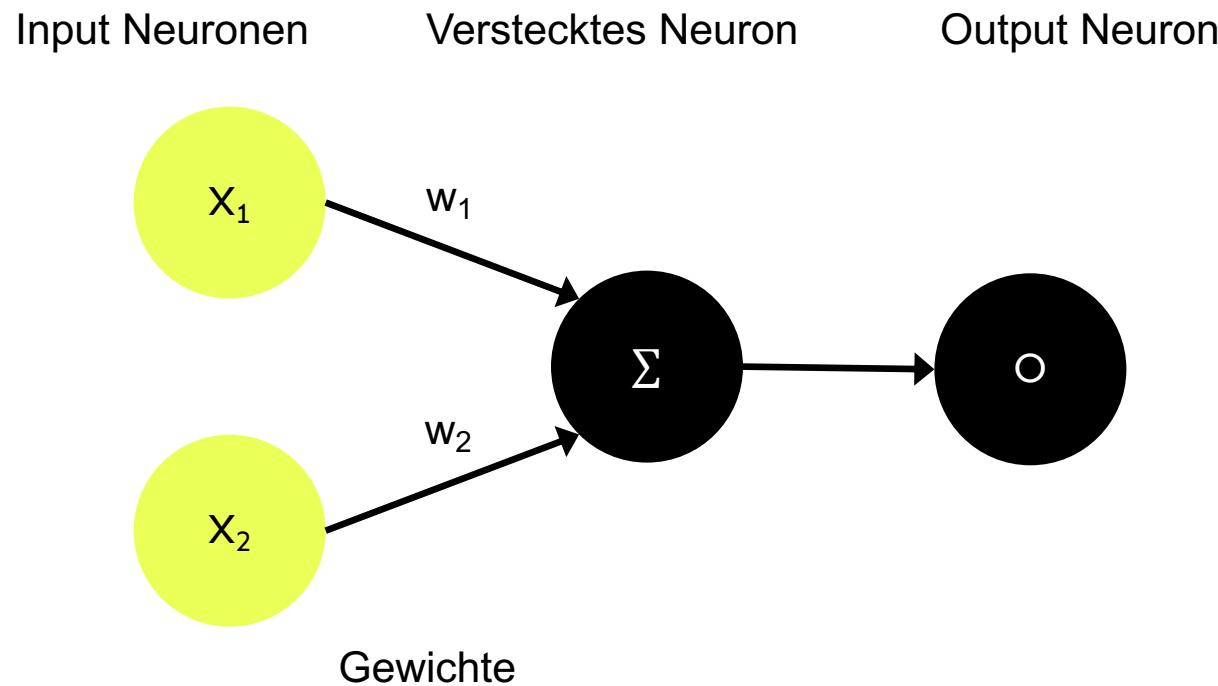
Architektur von **CPU** (Central Processing Unit) und **GPU** (Graphic Processing Unit)



Der Machine Learning Workflow

Modellierung - Perzeptron

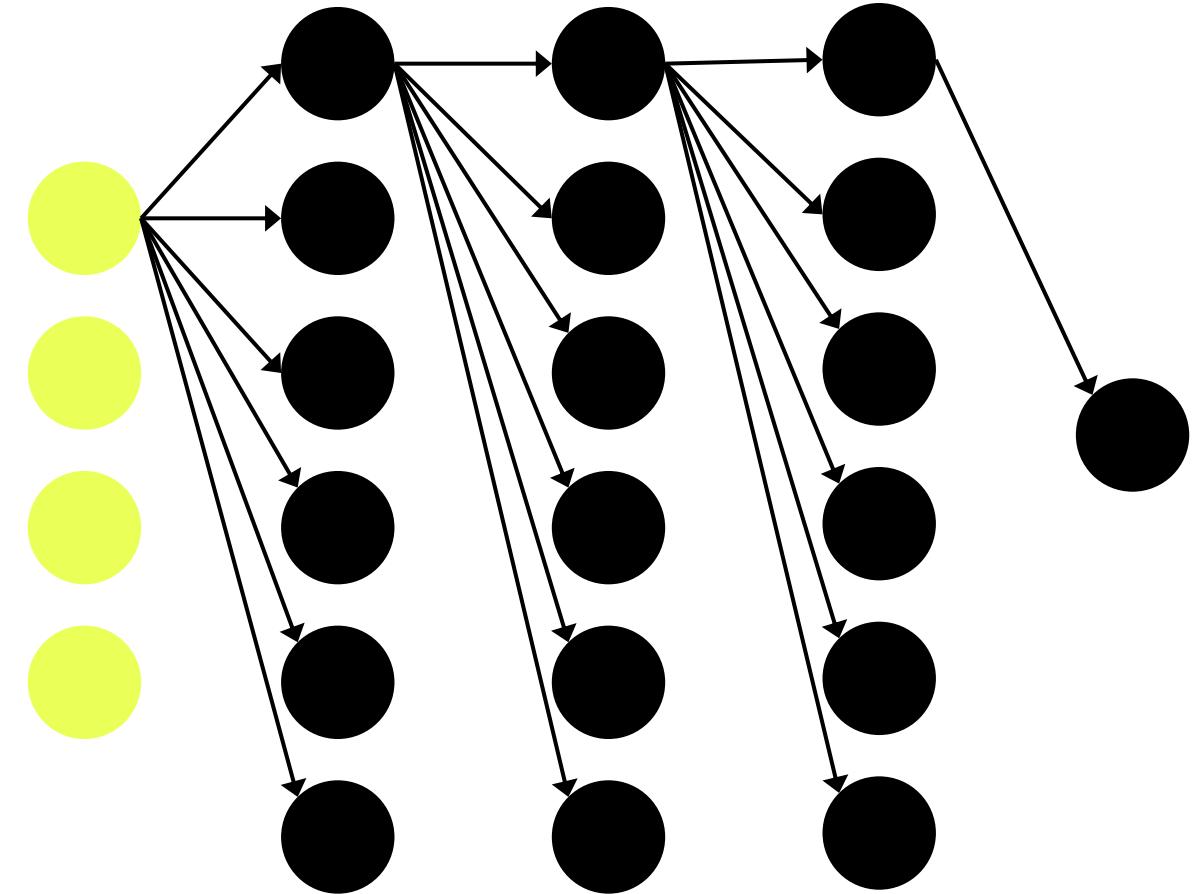
Eine der einfachsten neuronalen Netzwerkarchitekturen die auf einem künstlichen Neuron basiert.
Funktioniert als einfacher linearer Klassifizierer, da die Funktion nur 0 oder 1 zurück geben kann



Der Machine Learning Workflow

Modellierung - Deep Learning

- Tiefe Netzwerke
- Komplexe Architekturen
- Komplexe Summenfunktionen
- Sehr viele Parameter

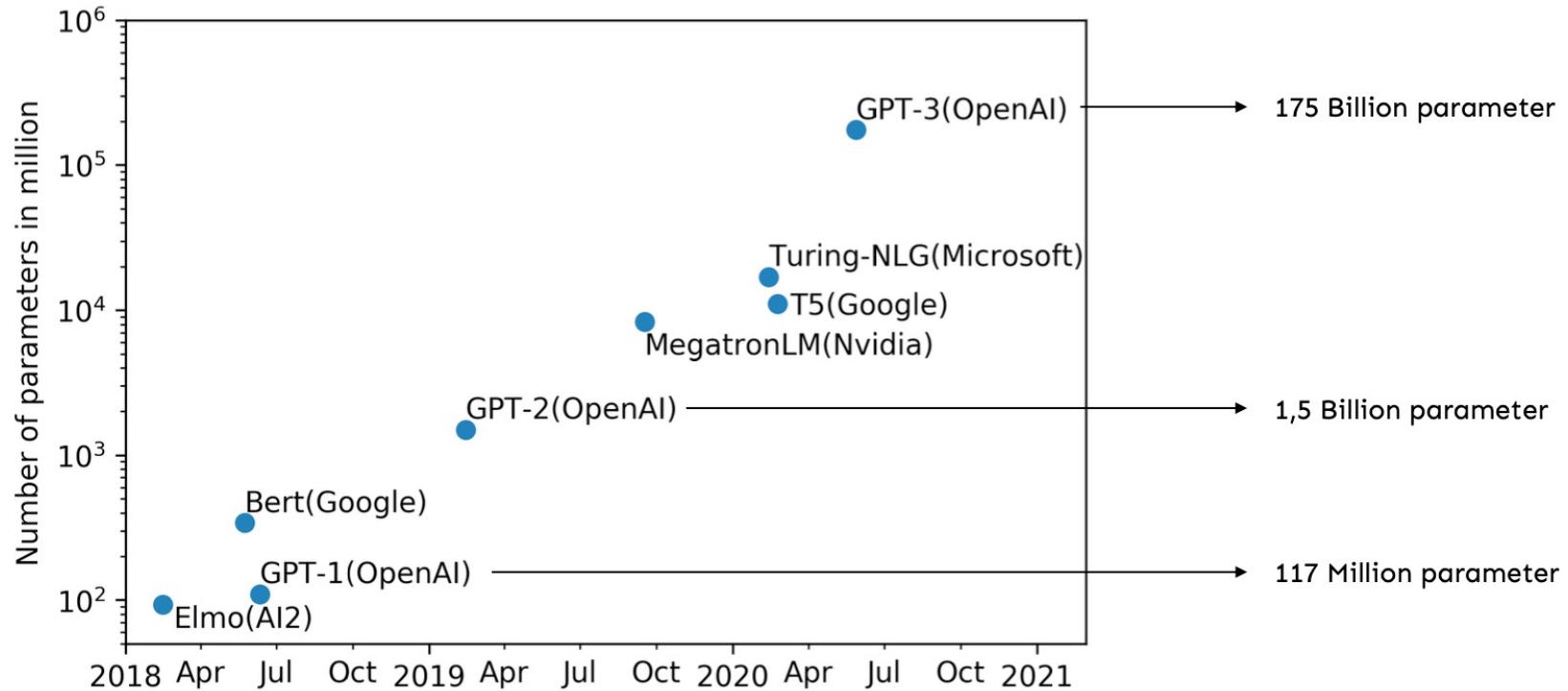


102 Gewichte müssen angepasst werden

<https://playground.tensorflow.org/>

Der Machine Learning Workflow

Modellierung - Deep Learning

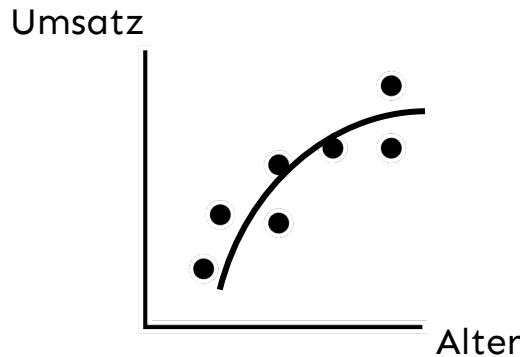


Der Machine Learning Workflow

Modellierung - Zusammenfassung Supervised Learning

Linearen Modellen

Einfach, interpretierbar,
kurze Trainingszeit

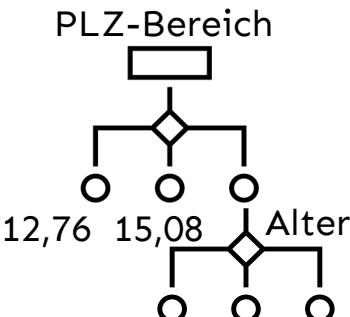


Beispiele

- lineare Regression
- logistische Regression

Entscheidungsbäumen

Einzelner Baum: Einfach,
leicht interpretierbar, kurze
Trainingszeit

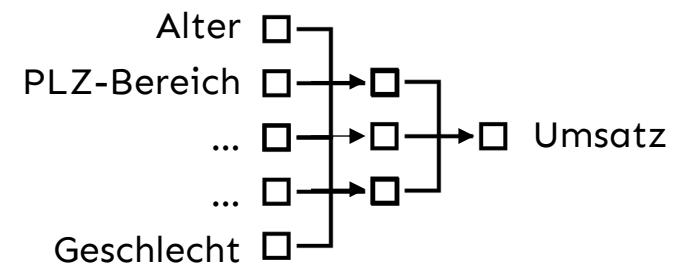


Umsatz (€): 6,67 13,45 17,92

- Decision Tree
- Random Forest
- Gradient Boosted Decision Trees

Neuronalen Netzen (NN)

Für komplexe Fragestellungen
geeignet, Blackbox, lange
Trainingszeit



- Autoencoder für Anomalie-Erkennung
- Convolutional NN für Bild-Erkennung
- Recurrent Neural Network für Zeitreihen

Der Machine Learning Workflow

Modellierung - K-Means und DBSCAN

Clustering

K-Means

- Jeder Punkt eines Clusters sollte nah zum Zentrum des Clusters sein.
- Die Anzahl an Clustern, die gefunden werden soll (K), wird vorgegeben.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- Cluster sind dichte Gruppen an Punkten.

Der Machine Learning Workflow

Modellierung - Entscheidende Fragen

Supervised Learning (Klassifikation und Regression):

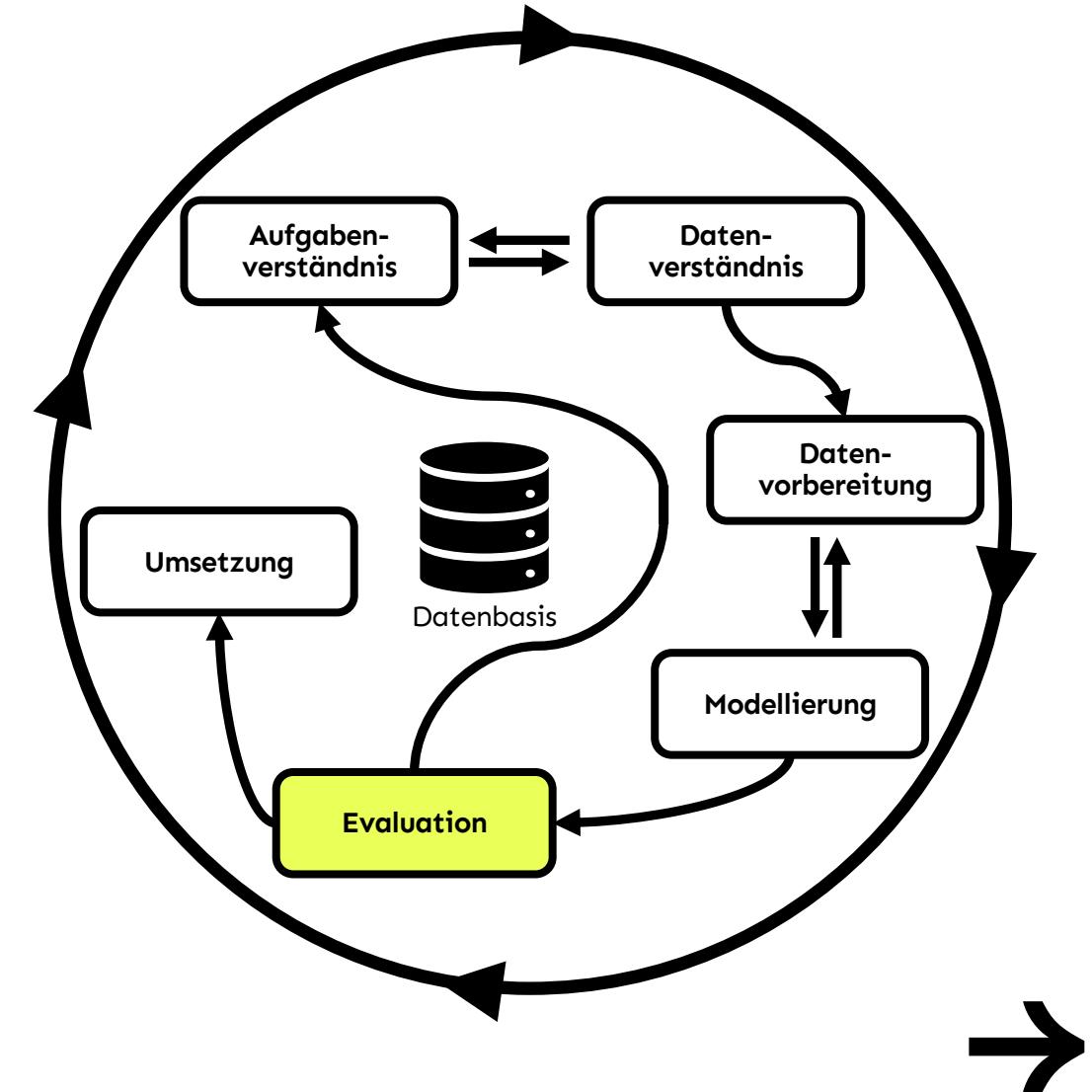
- Wieviel Beobachtungen sind vorhanden?
- Ist die Erklärbarkeit des Modells wichtig?
- Ist Geschwindigkeit wichtig?
- Ist Genauigkeit wichtig?

Clustering:

- Ist die Anzahl an Clustern bekannt?
- Liegen Kategorische Daten vor?

Modell	Anwendungsbereich
Lineare Modelle	<ul style="list-style-type: none">• Erstellung einer Baseline für komplexere Modelle• Nur für einfache Probleme geeignet
Entscheidungsbäume und Random Forest	<ul style="list-style-type: none">• Gute Nachvollziehbarkeit• keine Skalierung nötig
K-Means und DBSCAN	<ul style="list-style-type: none">• Clustering,• Feature Engineering (Erzeugung von zusätzlichen Merkmalen)
Deep Learning	<ul style="list-style-type: none">• Geeignet für komplexe Probleme (Sprachverarbeitung, Bildverarbeitung)• Braucht große Datenmengen, Komplexer Aufbau,• Hohe Anforderungen an Hardware

Evaluation



Der Machine Learning Workflow

Wichtiger Schritt nach / während dem Model Training bei dem die Güte des Models anhand verschiedener Metriken bewertet wird.

Ist teilweise in Model Training eingebunden.

Kann sowohl rein numerisch als auch graphisch erfolgen.

Kann automatisiert werden, was nicht immer zu empfehlen ist.

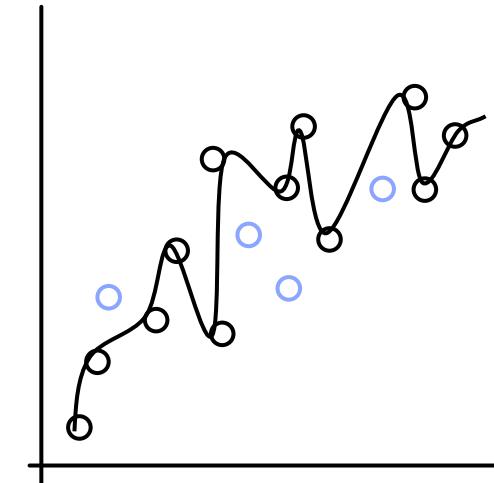
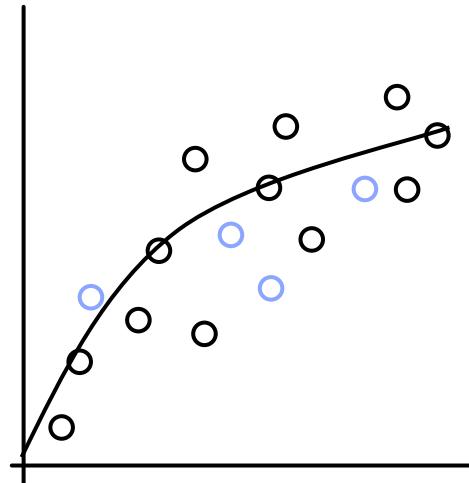
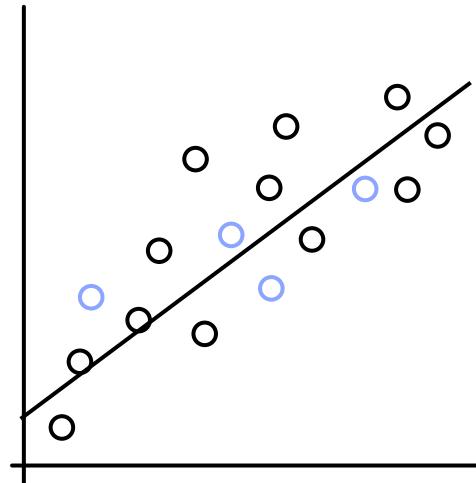
Erfüllt das Model die Problemstellung?

Der Machine Learning Workflow

Evaluation - Overfitting und Underfitting

Overfitting: Modell zu stark an Trainingsdaten angepasst und generalisiert daher zu wenig. (Vorhersagen auf Trainingsdaten besser als auf unabhängigem Datensatz).

Underfitting: Das Modell ist zu einfach und kann die Abhängigkeiten der Daten nicht wiedergeben.



Ohne unabhängige Testmenge wählt ein Modell Merkmale, welche die Daten tendenziell *overfitten*. Aus diesem Grund sollten Modelle mit einem **unabhängigen Datensatz bewertet** werden.

Der Machine Learning Workflow

Evaluation - Train- / Test-Split

Zurückhalten von Daten um „neue Daten“ zu simulieren, welche nicht in die Modellbildung eingeflossen sind sind.



Ziel:

- **Verallgemeinerungsfähigkeit** des Models beurteilen
- *Overfitting* erkennen

Nachteil:

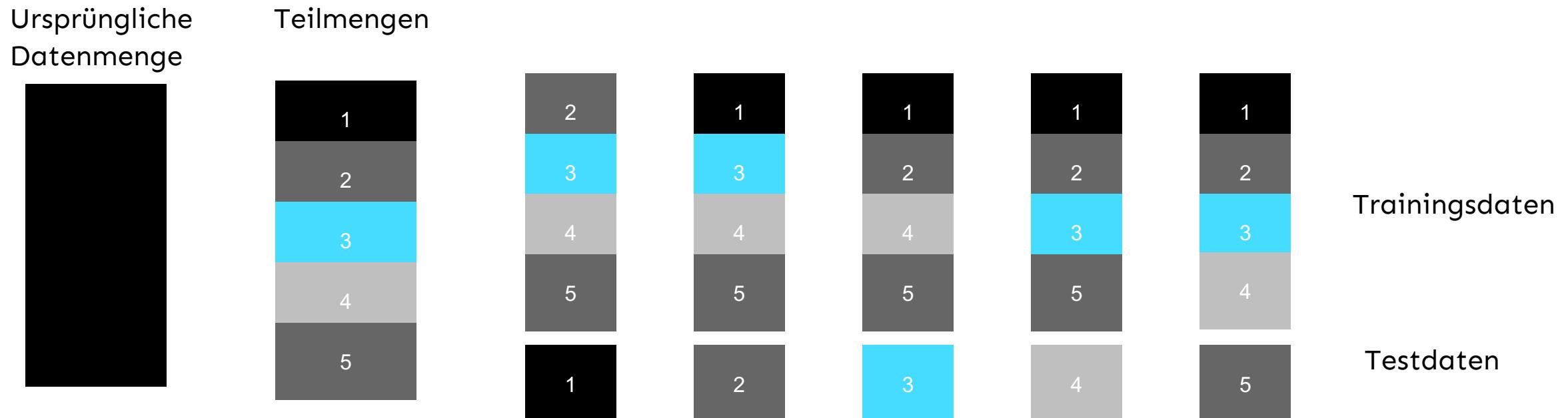
- Testdaten müssen Trainingsdaten abgezogen werden, was bei einer geringen Datenmenge problematisch ist.

Der Machine Learning Workflow

Evaluation - k-fache Kreuzvalidierung

Mehrere Kombinationen von Trainings- und Testdaten werden geprüft

- Mehrere Werte fließen in finale Beurteilung ein
- Durchschnittswert und Standardabweichung bestimmen finales Ergebnis



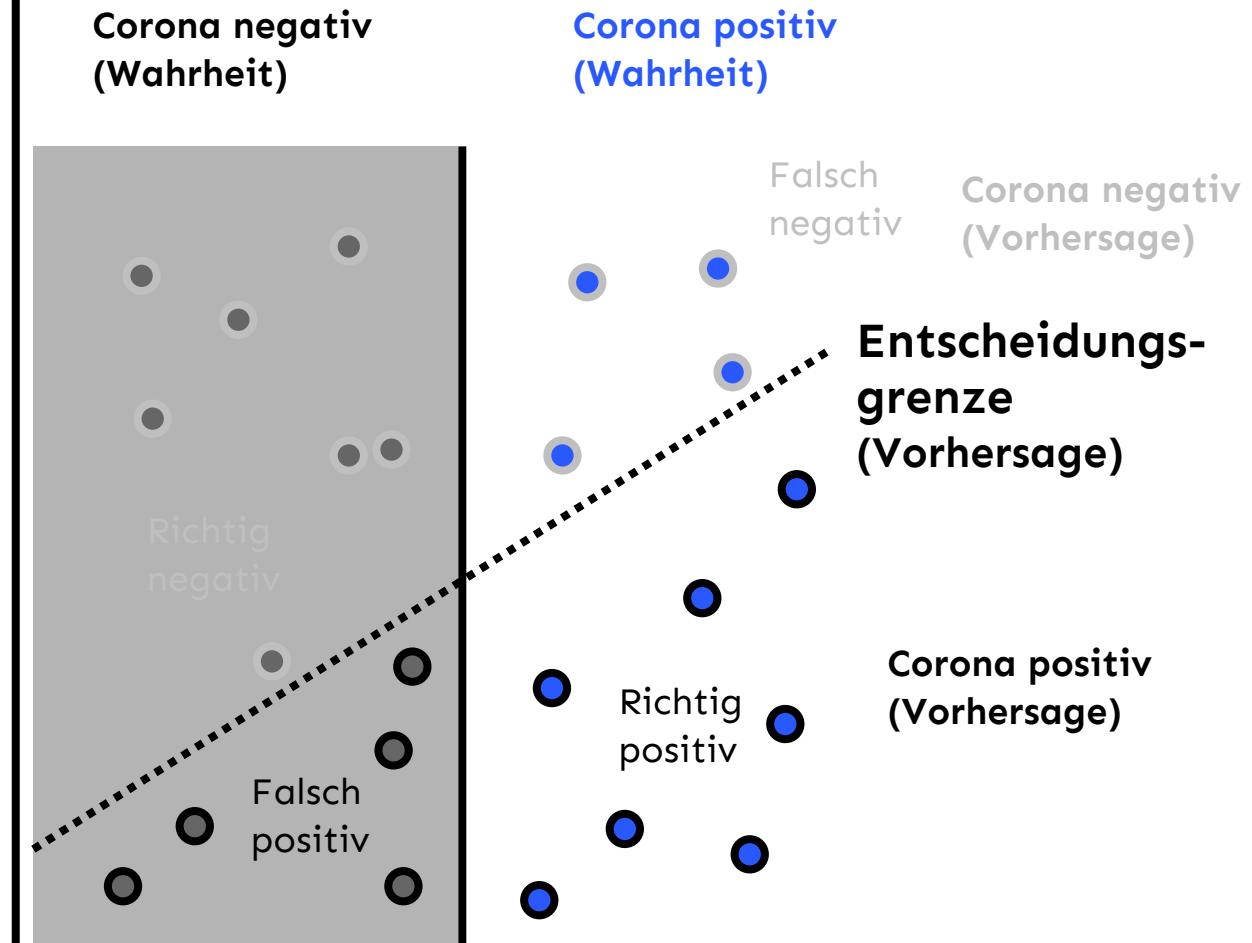
Der Machine Learning Workflow

Evaluation Klassifikation - Konfusionsmatrix

Beispiel Corona Tests

	Corona negativ (Wahrheit)	Corona positiv (Wahrheit)
Corona positiv (Vorhersage)	Falsch positiv	Richtig positiv
Corona negativ (Vorhersage)	Richtig negativ	Falsch negativ

	Corona negativ (Wahrheit)	Corona positiv (Wahrheit)
Corona positiv (Vorhersage)	5	7
Corona negativ (Vorhersage)	4	6



Der Machine Learning Workflow

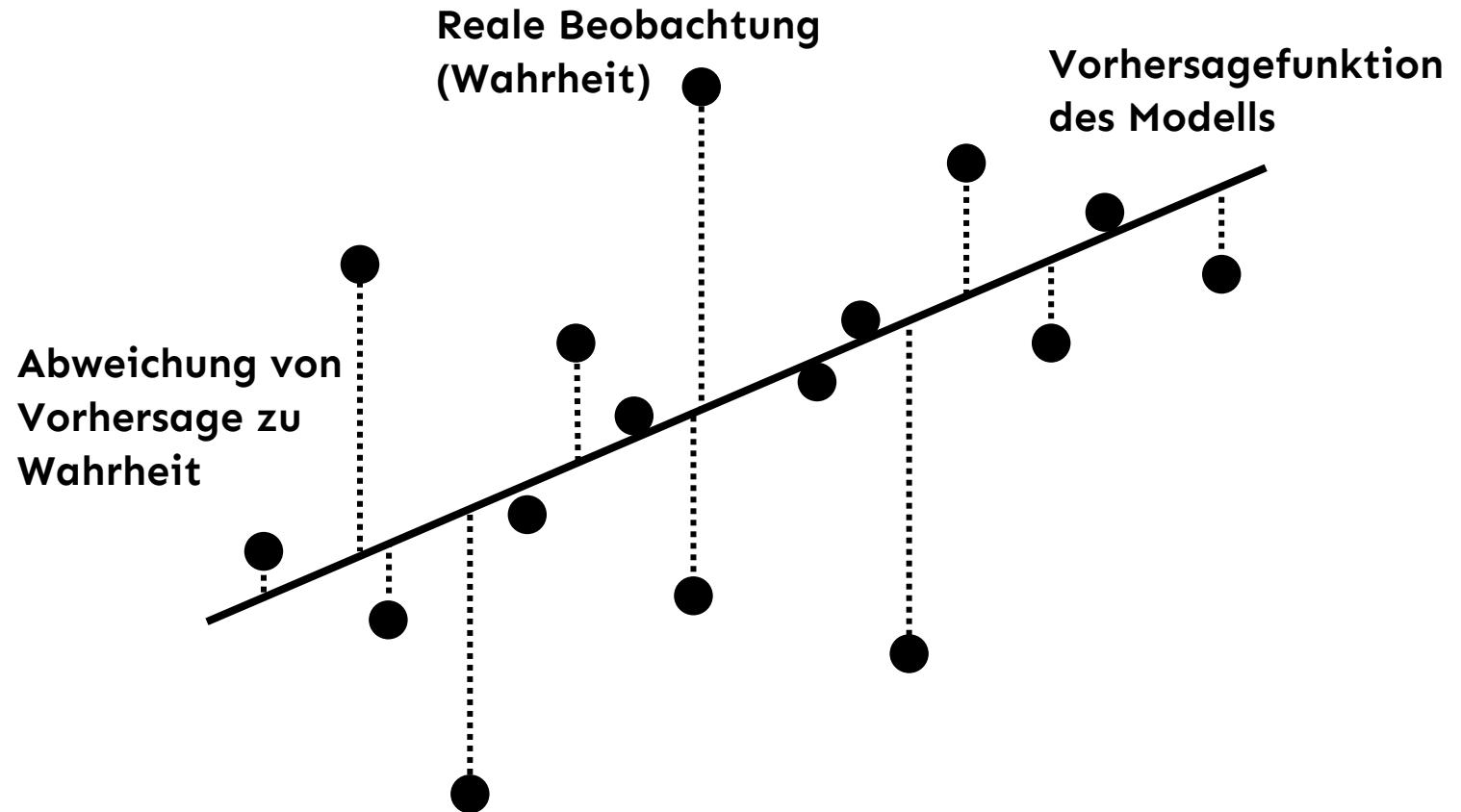
Evaluation Regression - Distanzen

Berechnung von verschiedenen Distanzen,
je nach Fragestellung:

Wie sollen Ausreißer bewertet werden?

Sind negative und positive Abweichung
unterschiedlich zu behandeln?

Sind große und kleine Abweichungen
unterschiedlich zu behandeln?

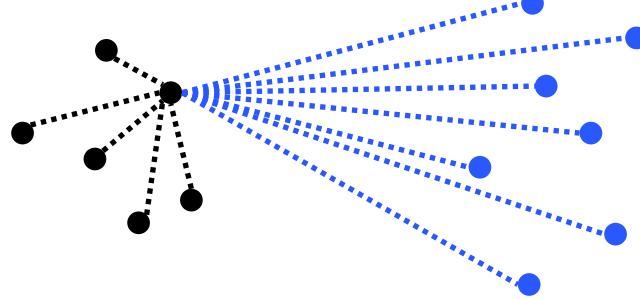


Der Machine Learning Workflow

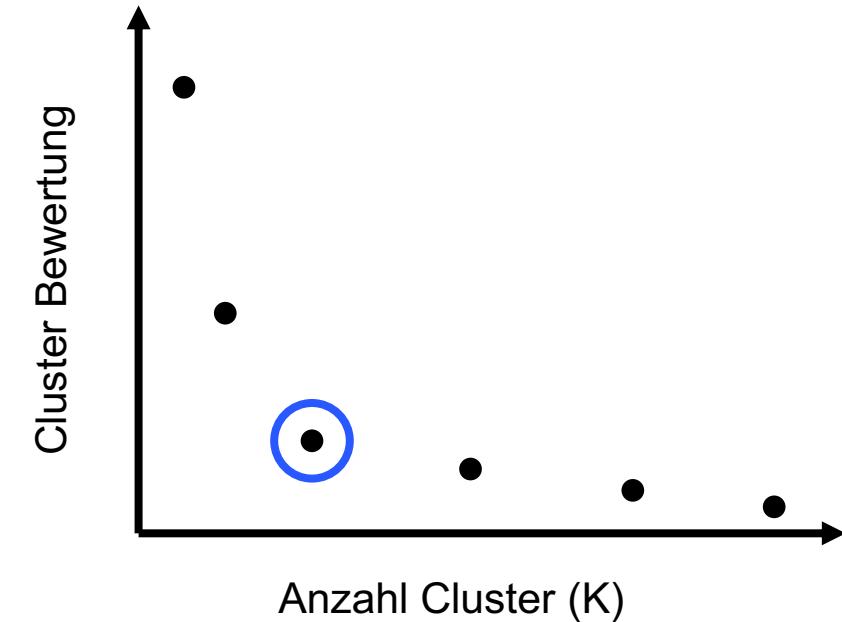
Evaluation Clustering - Ähnlichkeiten

Bewertung wie ähnlich sich Beobachtungen des gleichen Clusters durchschnittlich sind.

Ähnlichkeit kann das Verhältnis der Abstände innerhalb des Clusters zu den Abständen zwischen den Clustern sein.



Über die **Elbow-Method** lässt sich das optimale Clustering identifizieren.
Beispiel: Optimales K bei K-Means



Der Machine Learning Workflow

Evaluation - Erklärbarkeit

Explainable Artificial Intelligence (XAI) soll eindeutig nachvollziehbar machen, auf welche Weise dynamische und nicht linear programmierte Systeme zu Ergebnissen kommen.

Lime (Local Interpretable Model-Agnostic Explanations)

- *Local*: Erklärung spiegelt das Verhalten des Klassifikators "um" die vorhergesagte Instanz wirklich wider
- *Interpretable* : Die Erklärung muss für Menschen verständlich sein
- *Model-Agnostic*: Es muss unabhängig vom Modell funktionieren

Trainieren eines einfachen linearen Modells auf Basis von Störungen von Beobachtungen

Der Machine Learning Workflow

Evaluation - LIME Beispiel

Warum erkennt ein KI-Modell einen Frosch?

Welche Bildausschnitte sind relevant?



Originalbild

Wird zu 52% als Frosch erkannt



Einführen von
Bildstörungen



Erkennungs-
Wahrscheinlichkeit
eines Froschs

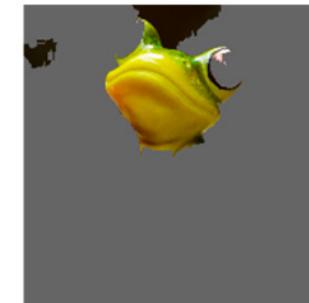
85%

0,01%

52%



Trainieren eines linearen
Modells zur Vorhersage der
Vorhersage des KI-Modells



Bestimmung der
relevanten
Bildausschnitte

Der Machine Learning Workflow

Evaluation - Entscheidende Fragen

Ist die Qualität ausreichend für den Anwendungsfall?

Welche Art Abweichungen sind relevant?

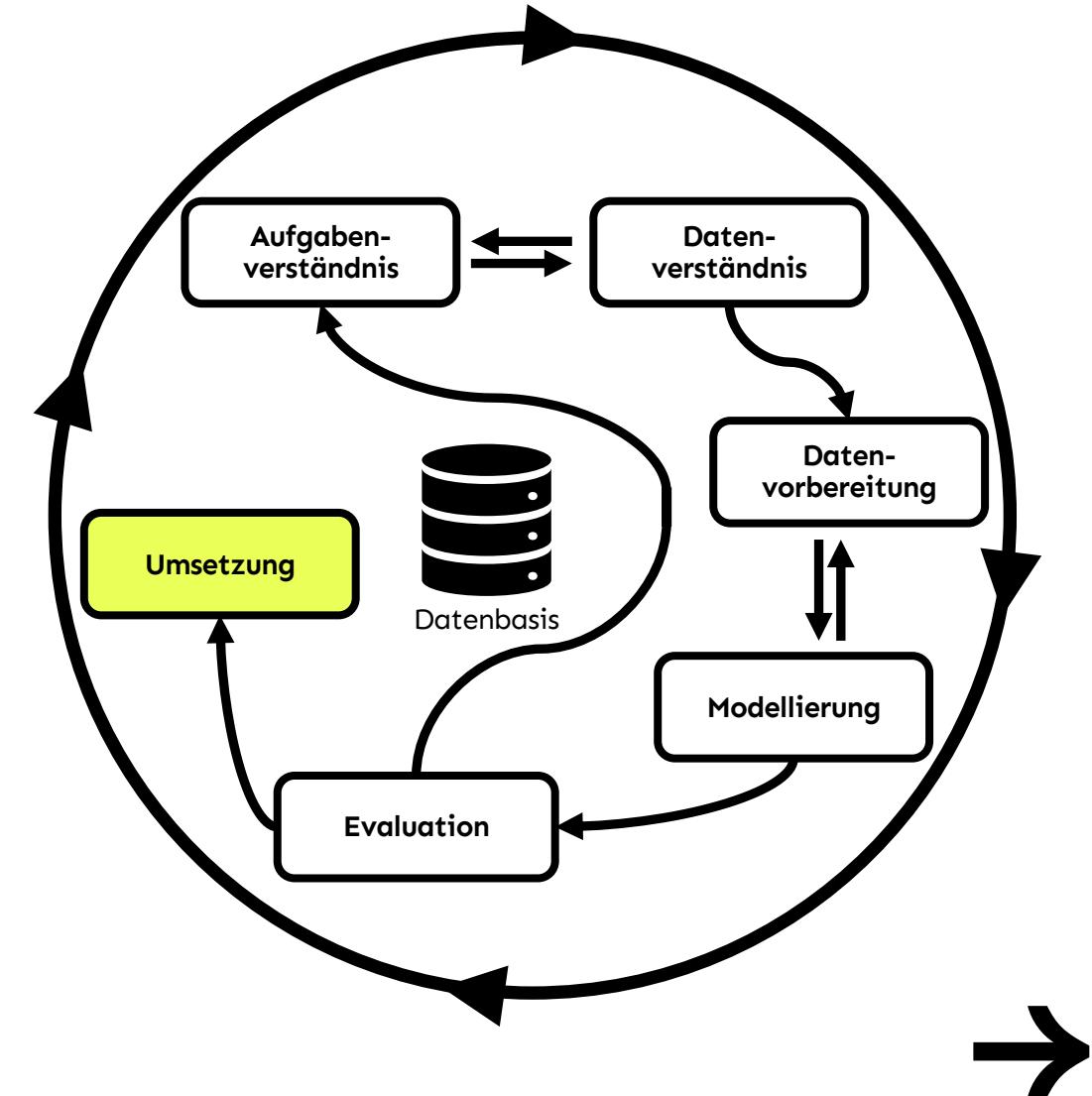
Wie schwerwiegend sind die Auswirkungen von Fehlern des Modells?

Kommt das Modell mit Grenzfällen zurecht?

Erfüllt das Modell Anforderungen bezüglich Nachvollziehbarkeit?

Auch bei einem trainierten Modell muss die Qualität **immer überwacht** werden, um **Data Drift** zu erkennen.
Daten verändern sich mit der Zeit, was zu einer **Verschlechterung der Modellqualität** führt.

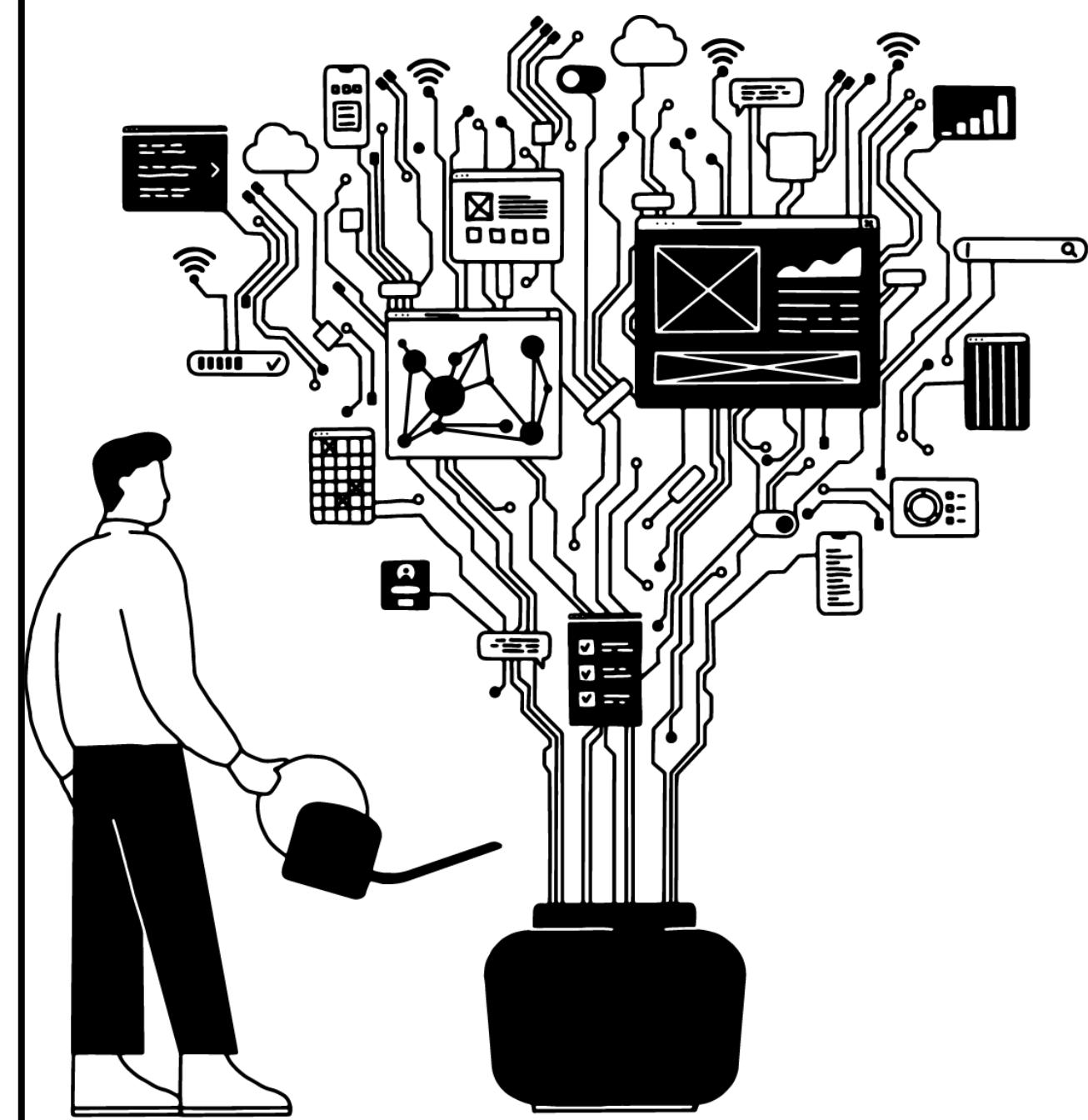
Umsetzung



Der Machine Learning Workflow

Umsetzung - Deployment

- Deployment eines Machine Learning Modells bedeutet es für die **Nutzer verfügbar zu machen.**
- Das Modell muss in die entsprechenden **Geschäftsprozesse und IT-Infrastruktur eingebunden** werden.
- Aus einem Data Science Projekt wird ein **Software Entwicklungsprojekt**. Gänzlich andere Fähigkeiten und Prozesse sind notwendig.



Der Machine Learning Workflow

Umsetzung - Übung

Was muss beim Wechsel in die Produktion beachtet werden?

Was muss beim weiteren Betrieb beachtet werden?

Der Machine Learning Workflow

Umsetzung - Deployment in Produktion

Integration des Modells in eine CI/CD Pipeline.

Ablegen des Modells in eine Model Registry zur Versionierung.

Integration des Modells im Serving und den jeweiligen Anwendungen.

Anlegen der Datenpipeline für den Produktivbetrieb.

Klären der Verantwortlichkeiten für Wartung und Betrieb, fachlichem Logging, technischem Logging.



Der Machine Learning Workflow

Umsetzung - Entscheidende Fragen

Wie werden Fehler gehandhabt?

Verschlechtert sich die Modellqualität?

Muss das Modell neu trainiert werden?

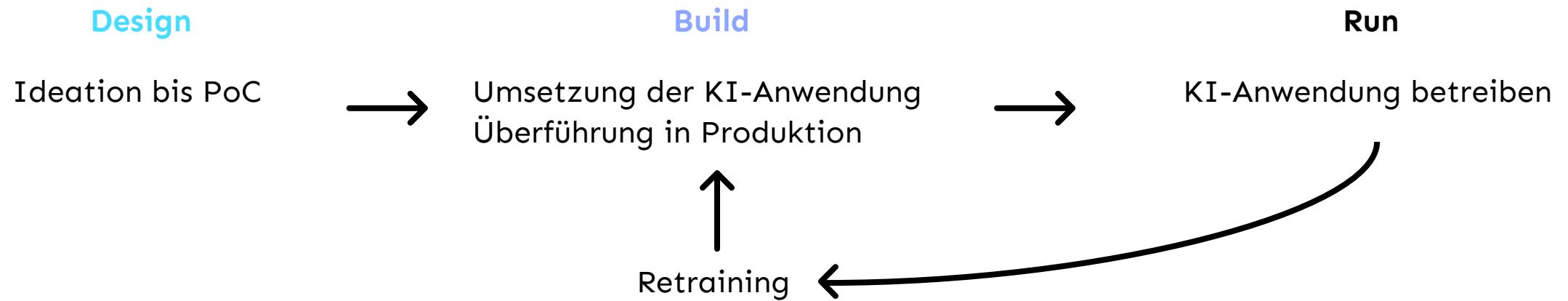
Ändert sich das Datenformat?

Erfüllt das Modell die Anforderungen aus dem Fachbereich?



KI-Projekte

KI-Wertschöpfungsprozess



KI-Projekte

Innovate, Manage, Plan



- Anbahnung und Ideation
- Datenbeschaffung für den PoC
- PoC Durchführung und Evaluation

- Vorbereitung der Build-Phase
- Aufnehmen der Anforderungen an die KI-Plattform
 - Benötigte Infrastruktur, Technologien, Prozesse
 - Neuerungen planen, verproben und einbauen
- Bereitstellung der Arbeitsumgebung

KI-Projekte

Build



- Datenintegration
- Bestandteile (Modelle, Schnittstellen) sind deploybar
- Modell Training und Serving in Produktion ermöglichen

KI-Projekte

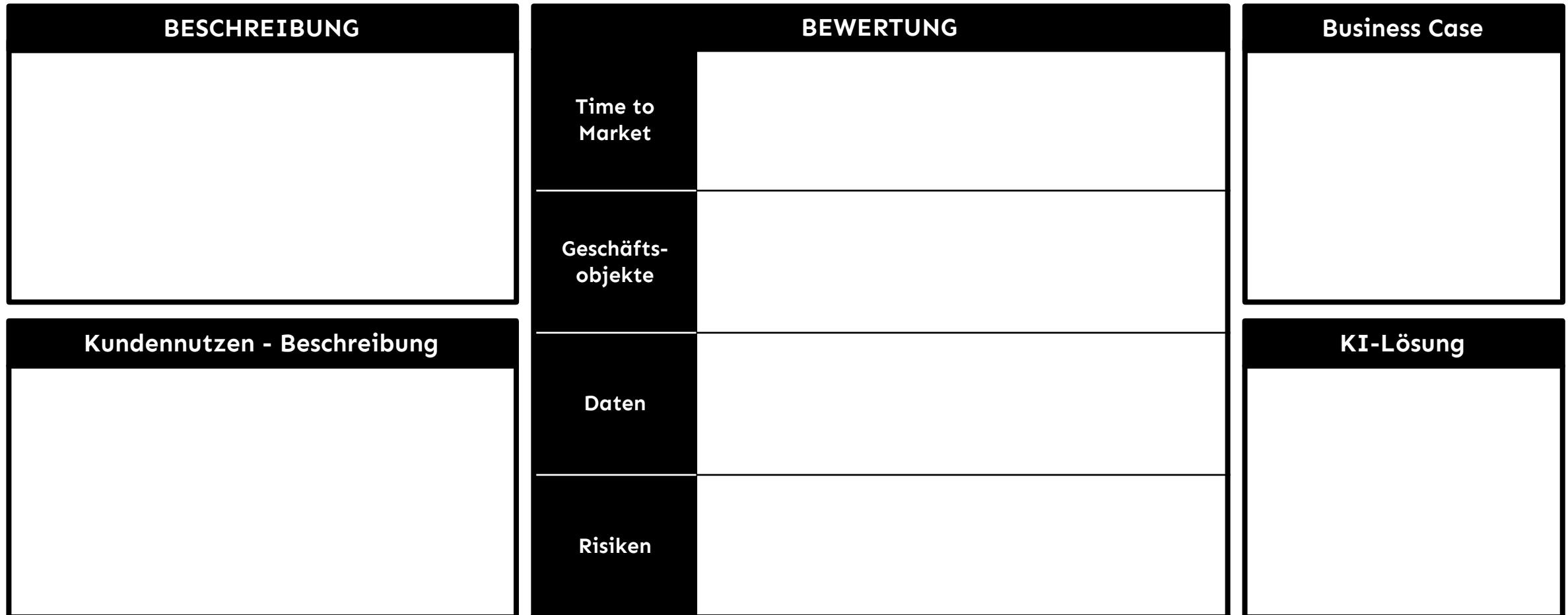
Run



- Monitoring und Support (Modelle, Daten, Infrastruktur, etc.).
- Aufsetzen von Tools und Prozessen für Monitoring.
- Aufsetzen von Prozessen für Daten- und Modellaktualisierung.
- Aufsetzen von Error Handling und Neustartprozessen.

KI-Projekte

KI-Canvas zur Planung von KI-Projekten



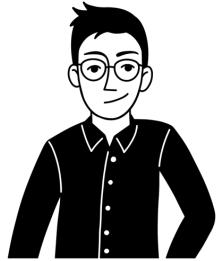
KI-Teams

Rollen und Aufgaben



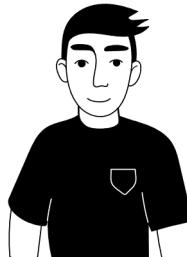
Architekt

- Entwicklung der übergeordneten Software-, Daten- und Prozessarchitektur



Product Owner

- Kümmerer und rechtliche Klärung



Data Engineer

- Data Ingestion
- Verbindung von Anwendung und Data Science



Software Engineer

- Anwendungsentwicklung (Front- und Backend)

- Deployment und Produktivsetzung
- Betrieb und Monitoring



Data Scientist

- Datenexploration
- Datenaufbereitung
- Entwicklung der der Data Science Lösung
- Modellerstellung

01



Data / AI Governance



Data / KI Governance

Ethik in KI

Erklärbarkeit: Es ist nachvollziehbar, wie ein Modell entscheidet und auf Basis welcher Daten.

Gerechtigkeit: Ein Modell behandelt alle Gruppen und Individuen ohne Bias.

Leistungsfähigkeit: Ein Modell erzeugt keinen Schaden bei Angriffen oder Anomalien in den Eingangsdaten.

Transparenz: Es muss für Nutzer ersichtlich sein, wie ein Modell arbeitet, was Stärken und Limitierungen sind.

Datenschutz: Persönliche Daten werden geschützt.

Data / KI Governance

KI-Bias



 **diri noir avec banan** @jackyalcine · Jun 29
Google Photos, y'all [REDACTED] My friend's not a gorilla.

813 394 ...

TWITTER

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ



 BUSINESS INSIDER

Microsoft Took Its New A.I. Chatbot Offline After It Started Spewing Racist Tweets

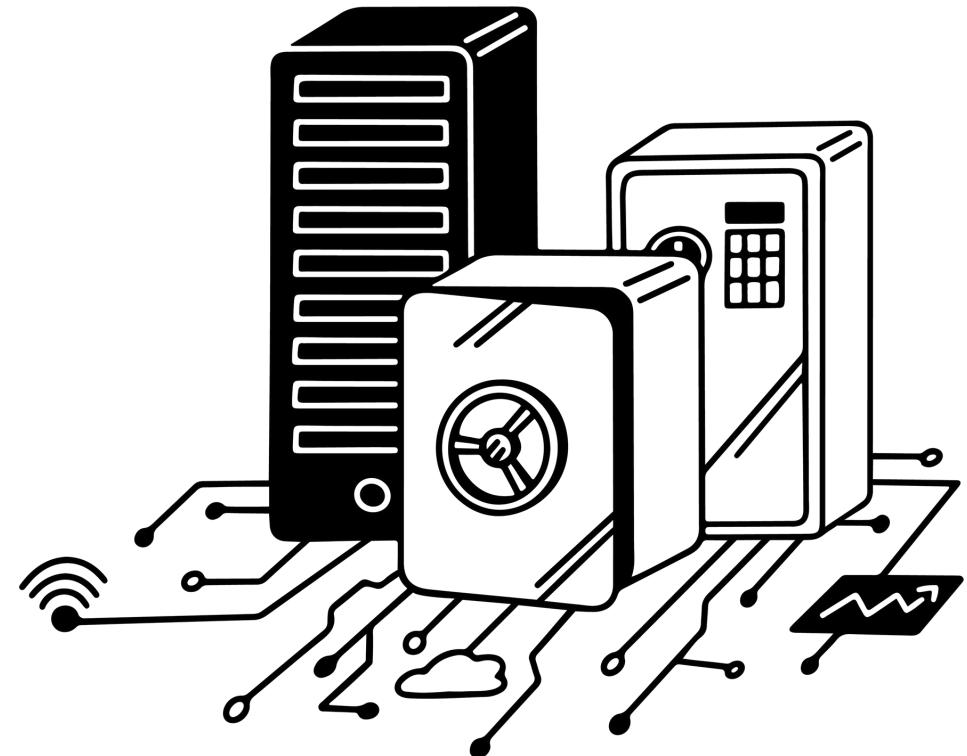
BY ROB PRICE

MARCH 24, 2016 • 12:21 PM

Data / KI Governance

Data Minimization

- Alle Produkte und Services sollten so gestaltet sein, dass so **wenig personenbezogene Daten** verarbeitet werden wie möglich.
- Es werden nur Daten verarbeitet, die eine **indirekte Identifikation** zulassen..
- Datensammlung auf **wenig sensitive Daten** begrenzen.
- Namen durch **Pseudonymen** ersetzen.
- Personenbezogene Identitätsnummern sollten **kein Routinefeld** in Datenbanken sein.



Data / KI Governance

Personenbezogenen Daten (PII)

- Name
- Adresse, inklusive Postleitzahl
- Telefonnummer
- E-Mail-Adresse
- Account Nummern
- Personalausweisnummern
- Kreditkartennummern
- Bankinformationen
- Geburtsdatum
- Alter
- Nationalität
- Lebenslauf
- Partielle Daten die eine Identifikation zulassen

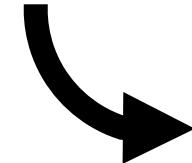


Data / KI Governance

Schwärzung

- Einfachste Technik zur Datenanonymisierung
- Entfernen oder Ersetzen der relevanten Inhalte

Users	
user_id	4759
first_name	Engrimm
last_name	von Horstman
job_title	Großmeister
credit_card	2450 2365 6006 4558



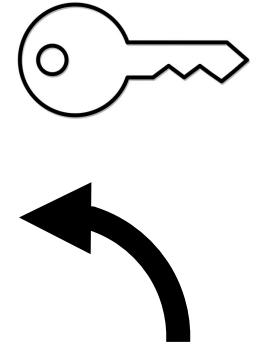
Users	
user_id	4759
first_name	Engrimm
last_name	v*****
job_title	Xxxxxxx
credit_card	xxxx xxxx xxxx xxxx

Data / KI Governance

Format-erhaltende Verschlüsselung

- Informationen mithilfe eines Geheimnisses, dem Verschlüsselungsschlüssel, codieren.
- Das Format der ursprünglichen Eingabe bleibt im Ergebnis erhalten.
- Verwendet in der Regel den *Advanced Encryption Standard* (AES).
- Lässt sich mit Hilfe des Schlüssels umkehren.
- Der Schlüssel wird getrennt von den Daten aufbewahrt.

Users	
user_id	4759
first_name	Engrimm
last_name	von Horstman
job_title	Großmeister
credit_card	2450 2365 6006 4558



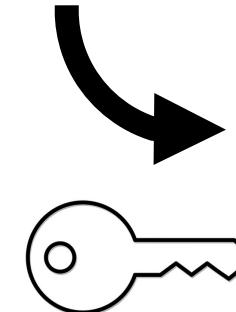
Users	
user_id	4759
first_name	3n9r1mm
last_name	v4n H0rstm4n
job_title	Gr05ßm31st3r
credit_card	5632 6547 3398 4728

Data / KI Governance

Scrambling

- Ist im Gegensatz zur Verschlüsselung **dauerhaft und nicht umkehrbar**.
- Für die Scrambling-Methoden gibt es **keine vordefinierten Regeln oder Vorgaben**.
- Beispiele für Scrambling-Methoden sind die teilweise Ersetzung von Namen (z.B. "John Smith" zu "Jxxx Sxxxx") oder die Randomisierung von Innenzeichen.
- Häufig verwendet beim Klonen von Datenbanken von einer Umgebung in eine andere, damit Daten während des Klonprozesses geschützt sind.
- Datenbanken mit scrambled Data können dann für Stress- und Integrationstests verwendet werden.

Users	
user_id	4759
first_name	Engrimm
last_name	von Horstman
job_title	Großmeister
credit_card	2450 2365 6006 4558



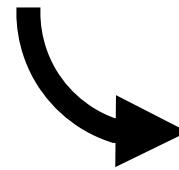
Users	
user_id	4759
first_name	3n9r1mm
last_name	v4n H0rstm4n
job_title	Gr05ßm31st3r
credit_card	5632 6547 3398 4728

Data / KI Governance

Pseudonymisierung

- Weniger umfassende Anonymisierungstechnik, fokussiert auf personenbezogene Informationen (PII).
- Ersetzt die Informationen einer Person durch ein **Alias**.
- Gemäß der DSGVO ist Pseudonymisierung die "*Verarbeitung personenbezogener Daten in einer Weise, dass die Daten nicht mehr ohne zusätzliche Informationen einem bestimmten Betroffenen zugeordnet werden können.*"
- Pseudonymisierte Daten können jedoch weniger identifizierbare Informationen behalten, um die **statistische Nützlichkeit der Daten zu gewährleisten**.
- Kann, wie die Verschlüsselung, **bei Bedarf rückgängig** gemacht werden.
- Erfordert sorgfältige Handhabung, da scheinbar nicht identifizierbare **Daten in Kombination identifizierbar** werden können.

Users	
user_id	1456
first_name	Heinz
last_name	Meissner
job_title	COO
salary	120.000
zip	13400
data_of_birth	10-05-1970
gender	m
robe_size	XL
hire_data	01-10-2019
manager_id	23



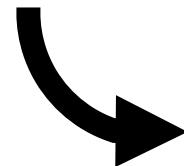
Users	
user_id	1456
first_name	Max
last_name	Muster
job_title	COO
salary	120.000
zip	77100
data_of_birth	null
gender	m
robe_size	XL
hire_data	01-10-2019
manager_id	23

Data / KI Governance

Statistical Data Replacement

- Basiert auf Statistiken über die zugrunde liegenden Daten und deren **realistischen Ersatz basierend auf diesen Statistiken**.
- Der neue Inhalt bewahrt oft die Nützlichkeit der Originaldaten (z.B. echte Namen werden durch fiktive Namen ersetzt), kann aber auch das Original einfach mit zufälligen Zeichen durchmischen oder nullen, wenn die Nützlichkeit nicht erhalten bleiben muss.
- Kompromiss aus Datenschutz und Nützlichkeit.

Users	
user_id	1456
first_name	Heinz
last_name	Meissner
job_title	COO
salary	120.000
zip	13400
data_of_birth	10-05-1970
gender	m



Users	
user_id	1456
first_name	Max
last_name	Muster
job_title	C-Level
salary	116.182
zip	13000
data_of_birth	null
gender	d

Data / KI Governance

Text-Anonymisierung

- Mit strukturierten Daten ist alles schön und gut, bei unstrukturierten Daten wird es schwierig.
- Die relevanten Stellen zu finden ist die entscheidende Herausforderung.
- Einsatz von Sprachverarbeitung auf Basis von Machine Learning Modellen (Transformer) um die relevanten Stellen zu identifizieren.

Überprüfung von GitHub Repositories

Dukovic, Duba (567)
Veröffentlicht 2.2.2024

Other Authors

HA Hollmann, Anna-Lena (567)
DK Derres, Kevin (567)

Organizational Unit

Data Governance & Data Protection & Information Security (DGI)

*Tl;dr**

Aus gegebenem Anlass: Prüft eure GitHub Repositories auf personebezogene Daten, sowie Secrets. Führt zudem einen Secret Scan durch.

Bei Problemen meldet euch unter irs@mercedes-benz.com

Anonymization

Strategien für die Textanonymisierung

- **Mustererkennung/Checksum:** Einsatz von Regex oder Kontext. (z.B. Kreditkarten-, Telefonnummern)
- **Blacklist:** Gibt es eine Liste von Optionen nach denen gesucht wird. (z.B. alle Anreden: Herr, Frau, Dr., Prof., ...)
- **Regelbasiert:** Entitäten lassen sich durch Regeln identifizieren.
- **Named Entity Recognition:** Einsatz von Sprachverarbeitung mittels Machine Learning. (z.B. Städte, Straßen, Namen)

Hallo Marcus,

hier ist Meyer, du kannst mich unter der Nummer 0168 77311345 erreichen.

Anonymization

Named Entity Recognition

Part-of-Speech-Tagging (POS):

Erkennung von Wortarten mittels Machine Learning Modellen.
(z.B. Nomen, Adjektive, Adverb)

Named-Entity-Recognition (NER):

Lokalisieren und klassifizieren von Entitäten, in vordefinierte Kategorien.
(z.B. Personennamen, Organisationen, Orte)

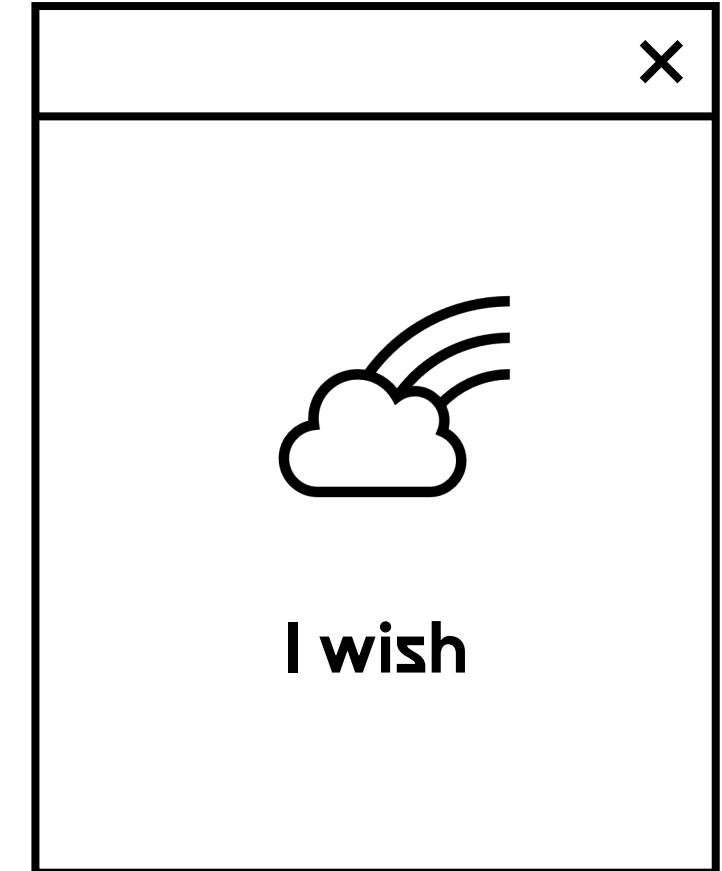
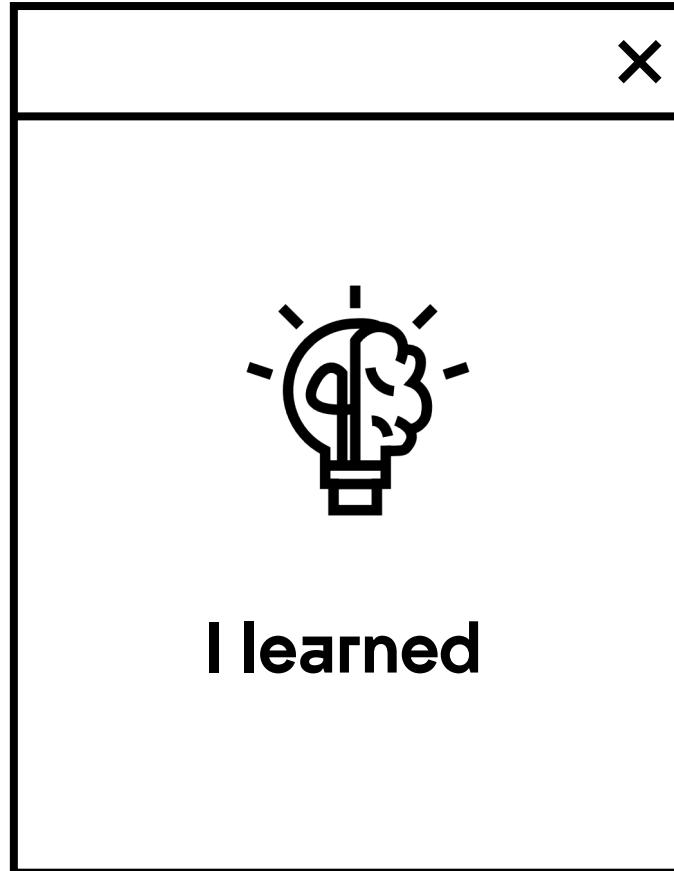
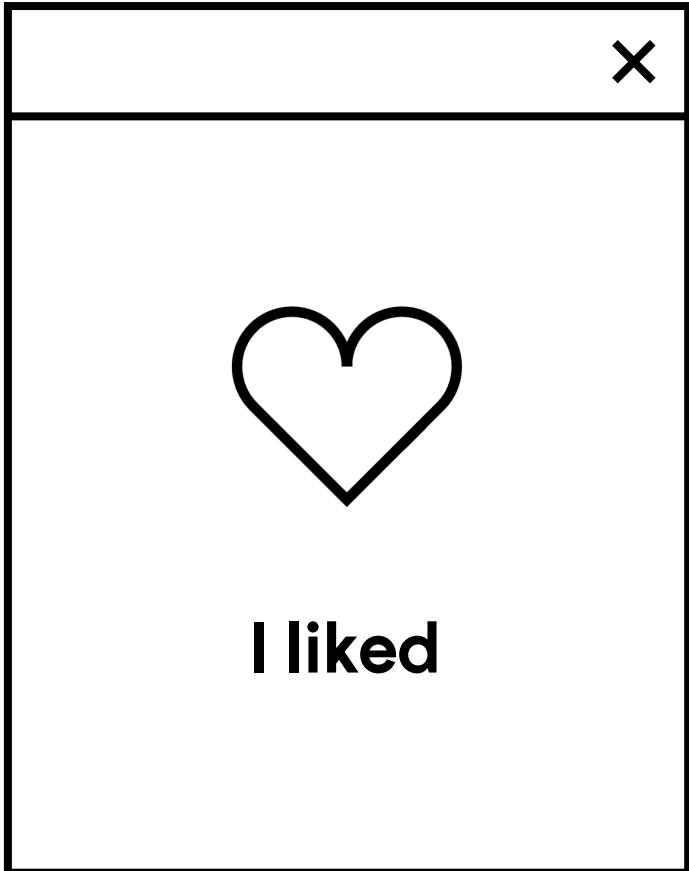
PER	PROPN	PROPN										
Hallo Marcus,												
MISC	hier ist	Meyer,	du	kannst	mich	unter	der	Nummer	0168	77311345	erreichen.	
	ADV	AUX	PROPN	PRON	AUX	PRON	ADP	DET	NOUN	NUM	NUM	VERB

Data / KI Governance

KI-Bias - Übung

Wie kommt es zu KI-Bias und wie ließe es sich verhindern?
Wo sieht ihr Grenzen der Machbarkeit?

Feedback



Feedback gerne auch an: matti.gerrit.korff@exxeta.com