Come2Data

TRAINING: Python für die Datenanalyse in den Sozialwissenschaften

Teil 3: Grundlagen Maschinelles Lernen mit Python

Grundlagen und Konzepte des Maschinellen Lernens

GEFÖRDERT VOM
Bundesministerium für Forschung, Technologie und Raumfahrt

SACHSEN
Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

SPEAKER: Matthias Täschner

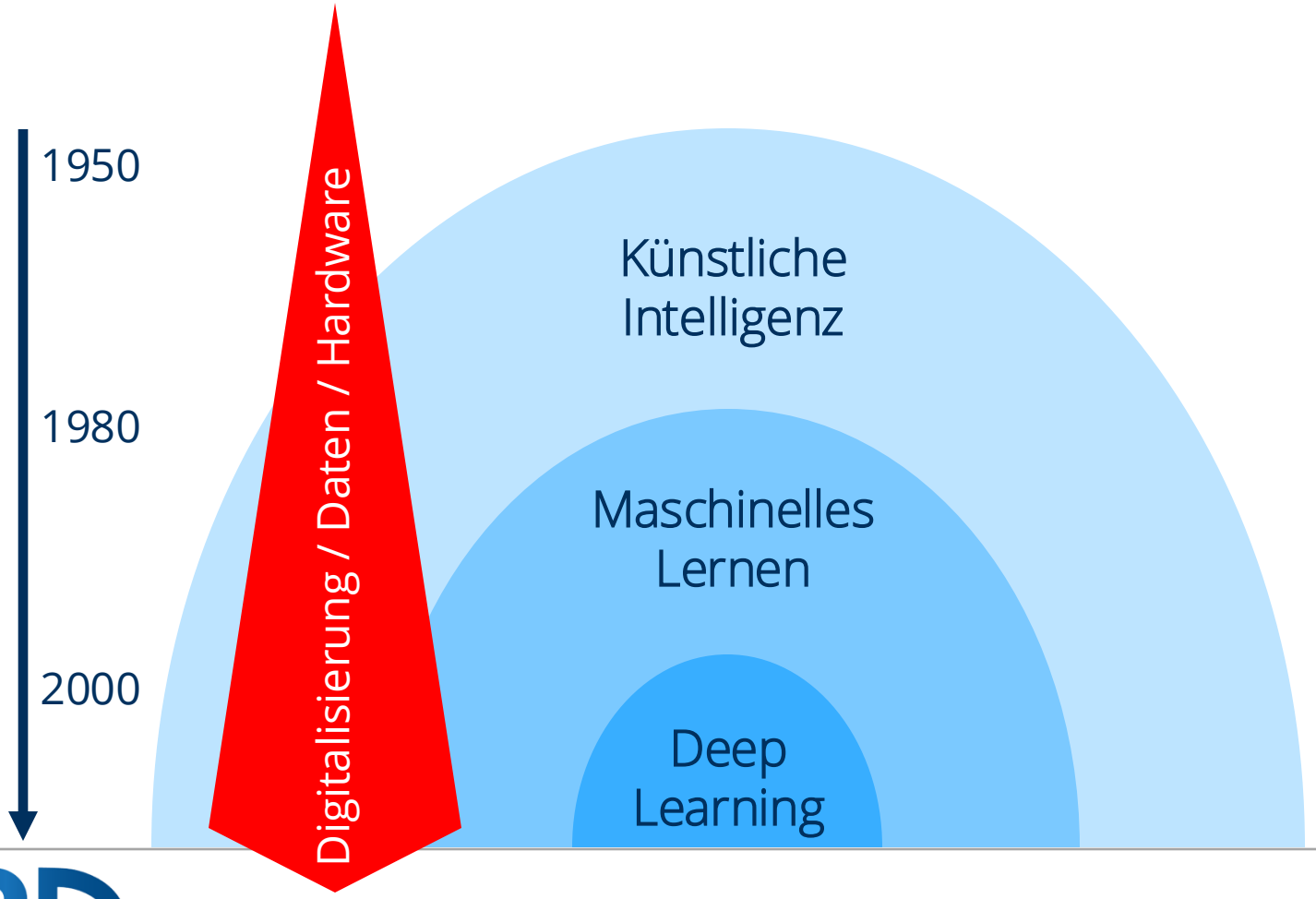
Mit Verwendung von Materialien von Robert Haase (ScaDS.AI / Universität Leipzig)

AGENDA

- Überblick zu den Gebieten der Künstlichen Intelligenz (KI)
- Arten des Machinellen Lernens (ML)
- Modeltypen im ML
- Praktische Übung: Handhabung von ML-Bibliotheken in Python

Gebiete der Künstlichen Intelligenz (KI)

Historische Phasen von KI



Programme ahmen intelligentes menschliches Verhalten nach

Programme erschließen sich selbstständig Zusammenhänge und Muster aus (strukturierten) Daten

Nutzung neuronaler Netze mit (sehr) vielen Schichten

Gebiete der Künstlichen Intelligenz (KI)

Spezialisierte (schwache / „narrow“) KI

- Anwendungsspezifisch
- Trainiert auf gelabelten Daten
- Adaption für andere Anwendung nicht / schwer möglich
- Kann nicht extrapolieren

Hervorragend geeignet für Datenanalyse

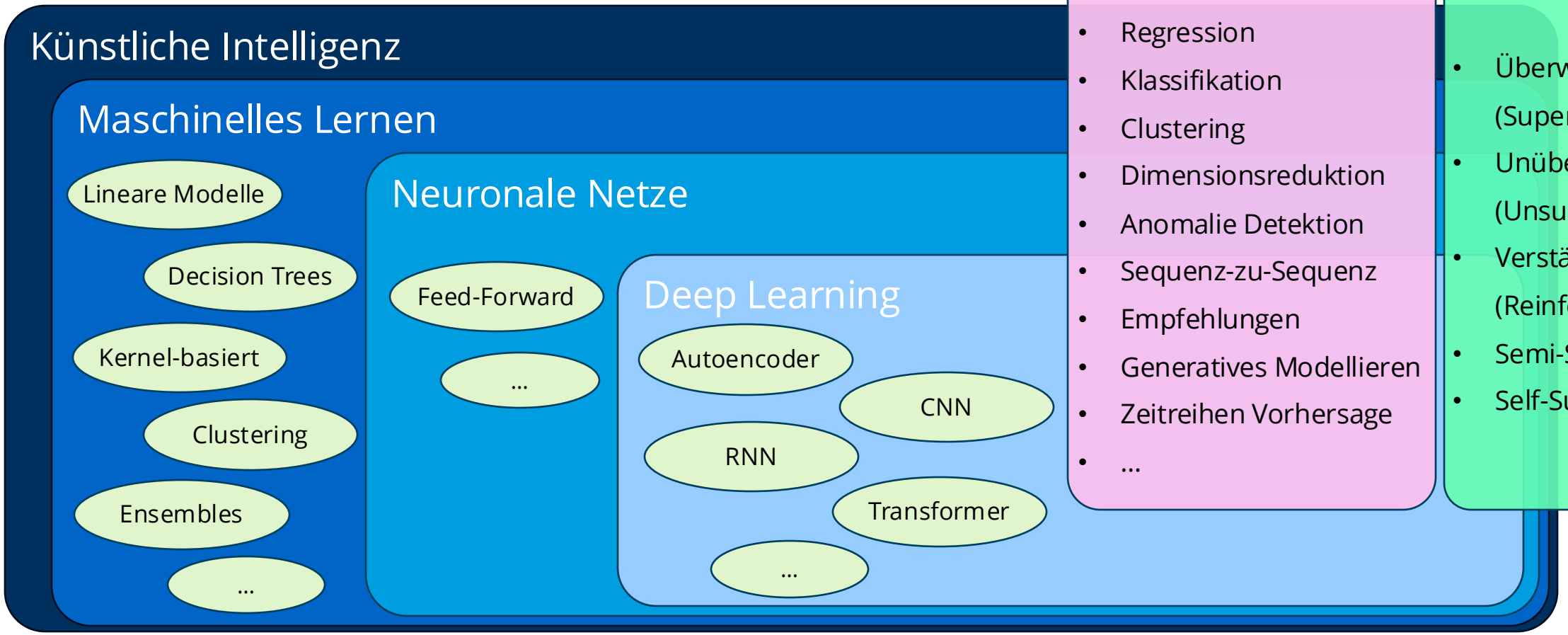
Allgemeine (starke / „general“) KI

- Menschliche Fähigkeiten
- Zugang zum Wissen der Menschheit, über den Einzelnen hinaus
- Kann kreativ arbeiten und neue Lösungen für universelle Aufgaben schaffen

Gebiete der Künstlichen Intelligenz (KI)

Modellfamilien

Bereiche, Paradigmen, Modelfamilien (nicht vollständig)



Problemstellungen

- Regression
- Klassifikation
- Clustering
- Dimensionsreduktion
- Anomalie Detektion
- Sequenz-zu-Sequenz
- Empfehlungen
- Generatives Modellieren
- Zeitreihen Vorhersage
- ...

Paradigmen

- Überwacht (Supervised)
- Unüberwacht (Unsupervised)
- Verstärkend (Reinforcement)
- Semi-Supervised
- Self-Supervised

Gebiete der Künstlichen Intelligenz (KI)
Anwendung auf Problemstellungen

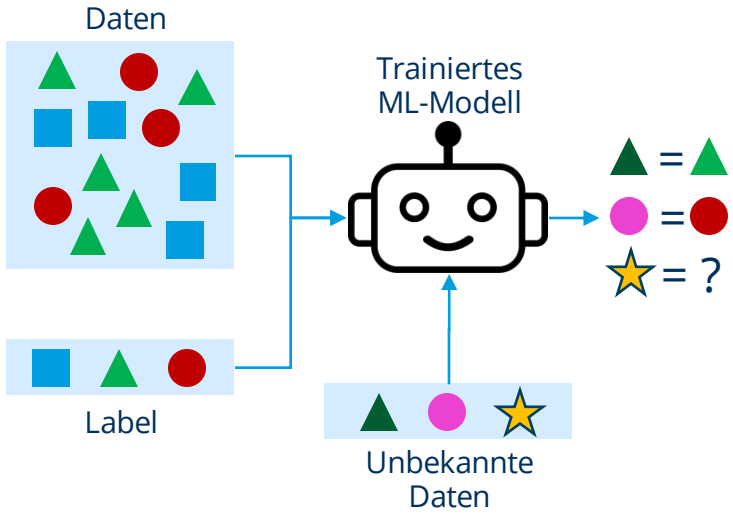
- Regression
  - Vorhersage eines kontinuierlichen (numerischen) Wertes
- Klassifikation
  - Vorhersage eines diskreten Labels / Klasse / Kategorie
- Clustering
  - Gruppierung von Datenpunkten auf Basis ihrer Eigenschaften
- Dimensionsreduktion
  - Komprimieren von hochdimensionalen Daten auf wenige informative Dimensionen
- Anomalie Detektion
  - Erkennen von Datenpunkten, die vom „normalen“ Muster abweichen
- Sequenz-zu-Sequenz
  - Eine geordnete Folge von Datenpunkten in eine andere umwandeln
- Empfehlungen
  - Daten bzgl. Relevanz ordnen oder Bewertung durch Nutzer vorhersagen
- Generatives Modellieren
  - Datenverteilung lernen und daraus neue, synthetische Daten erzeugen

Arten des Maschinellen Lernens (ML)
Überwachtes (Supervised) Lernen

Vorgehen
- ML-Modelle werden mit vorab gelabelten („etikettierten“) Daten trainiert
- Für das Training werden Eingabedaten und die gewünschten Zielwerte bereitgestellt
- Vorhersage der Zielwerte auf neuen, bislang unbekannten Daten des gleichen Formats

Anwendungsbeispiele
- Klassifikation, Regression
- Anomalie Detektion
- Generatives Modellieren

Algorithmen / Modelarten - Beispiele
- Lineare Regression
- Decision Trees & Random Forest (DT & RF)
- Support-Vektor-Machines (SVM)
- Künstliche Neuronale Netze (KNN)

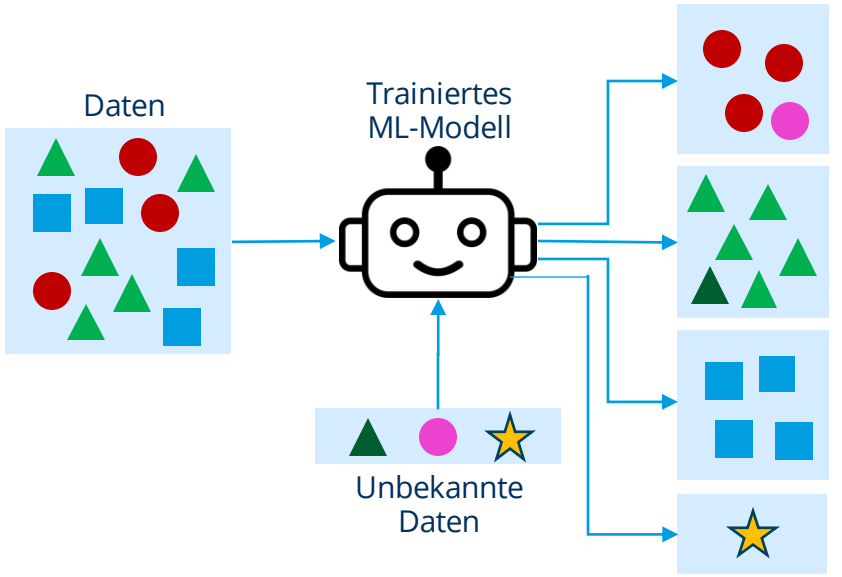


Arten des Maschinellen Lernens (ML)
Unüberwachtes (Un-supervised) Lernen

Vorgehen
- ML-Modelle werden mit nicht gelabelten Daten trainiert
- Modelle erkennen Muster, Strukturen, Zusammenhänge selbstständig

Anwendungsbeispiele
- Clustering
- Dimensionsreduktion
- Anomalie Detektion

Algorithmen / Modelarten - Beispiele
- K-Means Clustering
- Principal Component Analysis (PCA)
- Autoencoder



Arten des Maschinellen Lernens (ML)
Verstärkendes (Reinforcement) Lernen

Vorgehen

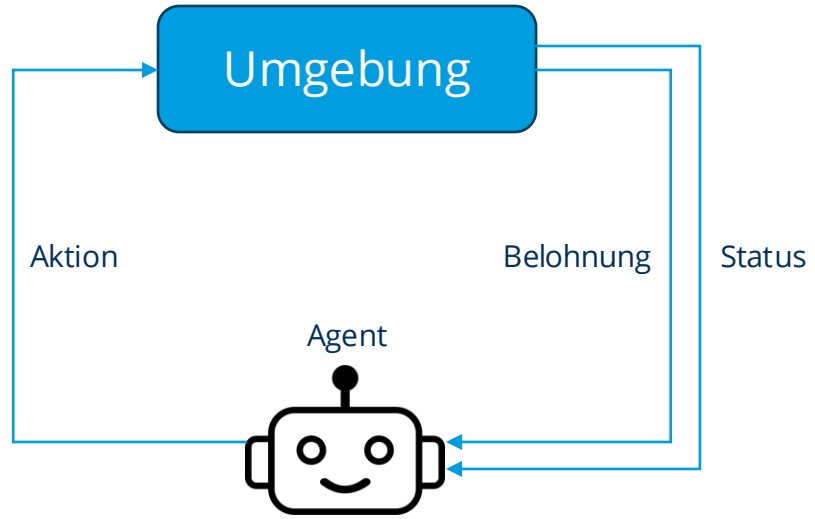
- Ein Agent wird trainiert, in einer Umgebung mit seinen Entscheidungen eine bestimmte Belohnung zu maximieren – „Versuch und Irrtum“
- Regeln legen mögliche Schritte des Agenten fest
- Belohnungen und Bestrafungen beeinflussen das Verhalten des Agenten

Anwendungsbeispiele

- Spiele-Agenten, z.B. im Schach oder Go
- Automatisierungssysteme, Robotik
- Simulationen

Algorithmen / Modelarten - Beispiele

- Q-Learning
- Markov Decision Processes (MDP)
- Monte Carlo Methods



Arten des Maschinellen Lernens (ML)
Mischformen

Semi-Supervised

- Ein kleiner Datensatz mit gelabelten Stichproben
- Ein größerer Datensatz mit nicht gelabelten Stichproben aus derselben Verteilung
- Trainieren auf gelabelten Daten → Modell weist nicht gelabelten Daten „Pseudo-Label“ zu
- Daten mit sichersten „Pseudo-Label“ in weiteres Training einbeziehen (Human-in-the-Loop)

Self-Supervised

- Daten sind nicht gelabelt, benötigte Labels werden vom Modell selbst erzeugt
- Das Modell erzeugt seine eigenen „Übungsfragen“ und Antworten aus den Rohdaten
- Lernt dabei nützliche Muster aus den Daten
- Bsp: Wörter in einem Satz abdecken und dann erraten - kein Label nötig

Modeltypen des Maschinellen Lernens (ML)
Lineare Modelle vs. Nicht-lineare Modelle

- Lineares Modell ≠ geradlinig gezeichnete Kurve
- Linearität bezieht sich auf Modell-Parameter

Ein Modell ist linear, wenn seine Vorhersage eine lineare Kombination der Features ist.
(z.b. gewichtete Summe der Eingangsdaten)

- Lineare Modelle
  - Leicht interpretierbar, basieren auf gut verstandenen Prinzipien, schnelles effizientes Training
  - Können nur wenig komplexe, lineare Beziehungen in den Daten modellieren
  - Bsp: Lineare bzw Logistische Regression, ARIMA, Lineare SVM, …
- Nicht-lineare Modelle
  - Modellieren auch komplexe, nicht-lineare Beziehungen in den Daten
  - Schwerer interpretierbar, aufwendiges Training mit mehr Parametern, oft mehr Daten nötig
  - Bsp: Decision Tree, Kernel-SVM, k-NN, Neuronale Netze, …

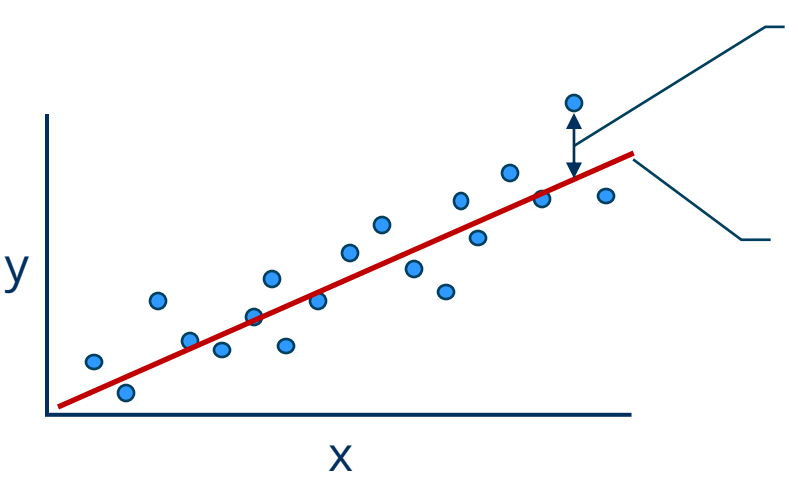
Modeltypen des Maschinellen Lernens (ML)
Probabilistische Modelle

- Lernen die (statistische) Verteilung in den Daten, anstatt Einzelpunkte vorherzusagen
- Vorwissen (Prior), kombiniert mit Eingangsdaten, für die Vorhersage einer nachgelagerten Verteilung (Posterior)
- Modelliert Unsicherheit / Konfidenz – gibt „Was“ und „Wie sicher“ zurück
- Bsp: Markov Ketten, Naive Bayes, Bayesian Neural Networks, …

Modeltypen des Maschinellen Lernens (ML)

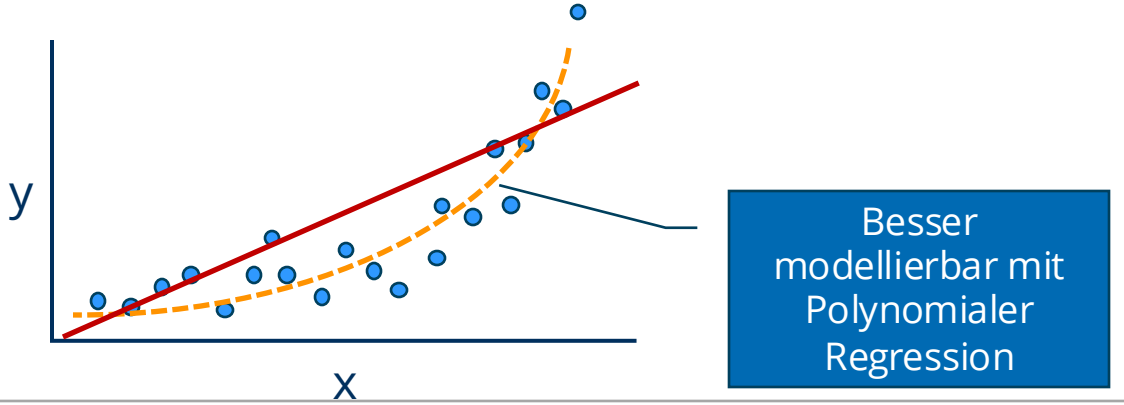
Lineare Regression

- Modelliert eine gerade Linie → $y = m*x + c$ mit:
  - y = Zielvariable
  - x = Eingangsdaten / Features
  - m = Regressions-Koeffizient
  - c = Konstante
  - *[x und m können auch Vektoren sein]*



- Versucht den mittleren Abstand (Fehler / Residual) zu den Punkten zu minimieren
- Vorhersage eines numerischen Wertes, welcher sich linear zu den Eingabedaten verhält
- Schnell, leicht interpretierbar
- Probleme bei komplexen, nicht-linearen Zusammenhängen in den Daten (z.b. Kurven)
- Erweiterung: Polynomiale Regression →

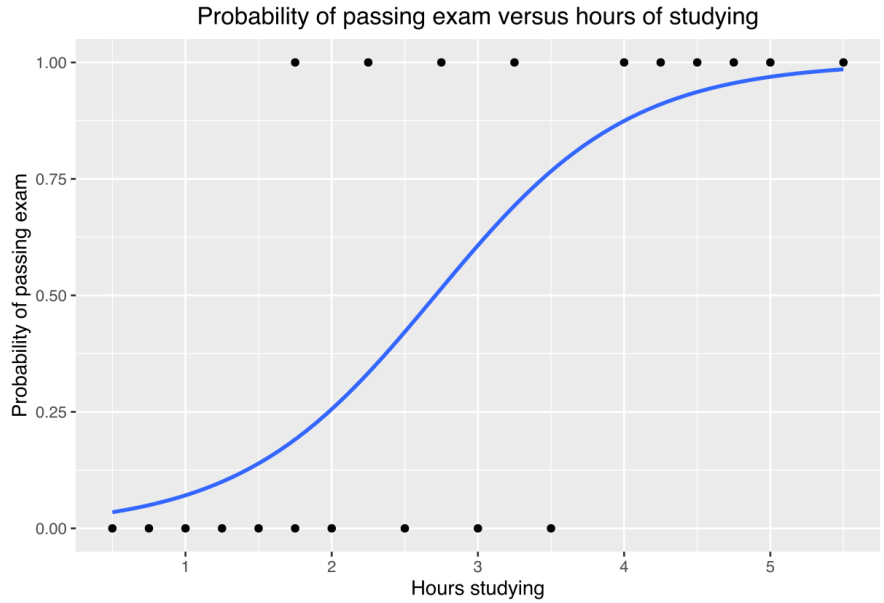
$$y = m_1*x + m_2*x^2 + \ldots + m_d*x^d + c$$



Modeltypen des Maschinellen Lernens (ML)

Logistische (“*Logit*“) Regression

- Lineares Modell für (binäre) Klassifikation
- Modelliert eine Sigmoid-Funktion (S-Shape) für Ausgabe einer Wahrscheinlichkeit zw. 0.0 - 1.0
- Leicht zu interpretieren, effizient auch auf hochdimensionalen Daten
- Passend bei linear trennbaren Daten
- Probleme bei komplexen, nicht-linearen Zusammenhängen oder unausgewogenen Daten (Verhältnis wahr/falsch)

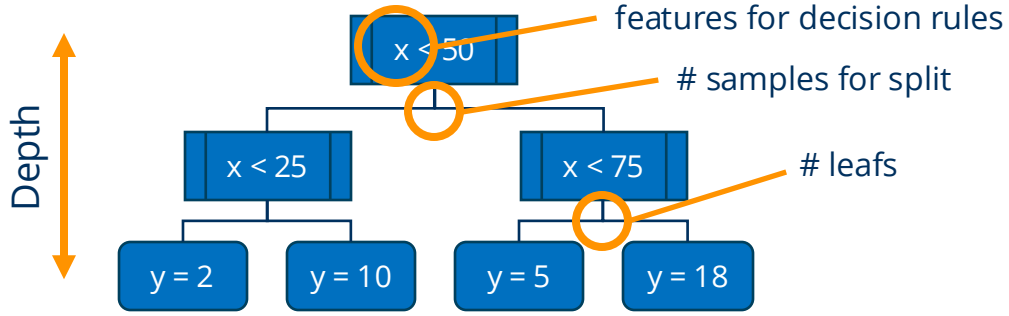


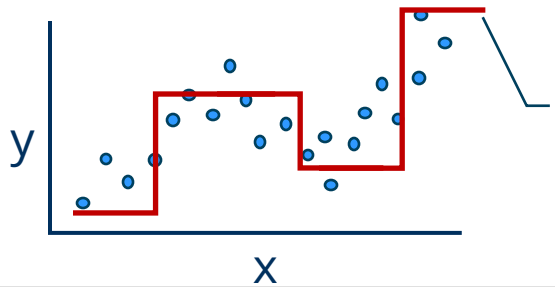
Quelle: Canley, https://en.wikipedia.org/wiki/Logistic_regression, CC BY-SA 4.0

Modeltypen des Maschinellen Lernens (ML)

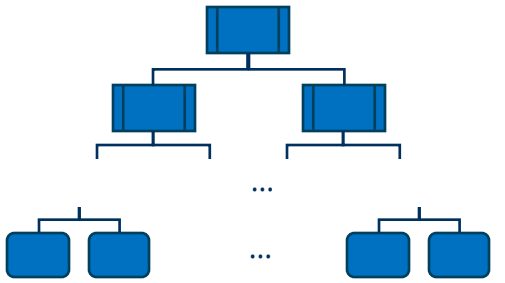
Decision Tree

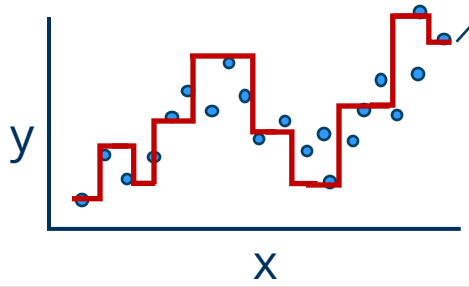
- Unterteilt Eingangsdaten (Feature-Space) in Regionen, basierend auf erlernten Entscheidungs-Regeln → kann für z.B. Regression, Klassifikation, Anomalie Detektion genutzt werden
- Jeder Region wird eine Zielvariable zugewiesen, dabei Approximation der Features
- Kann komplexe nicht-lineare Zusammenhänge erfassen, bleibt dabei gut interpretierbar





Modellierter Decision Tree



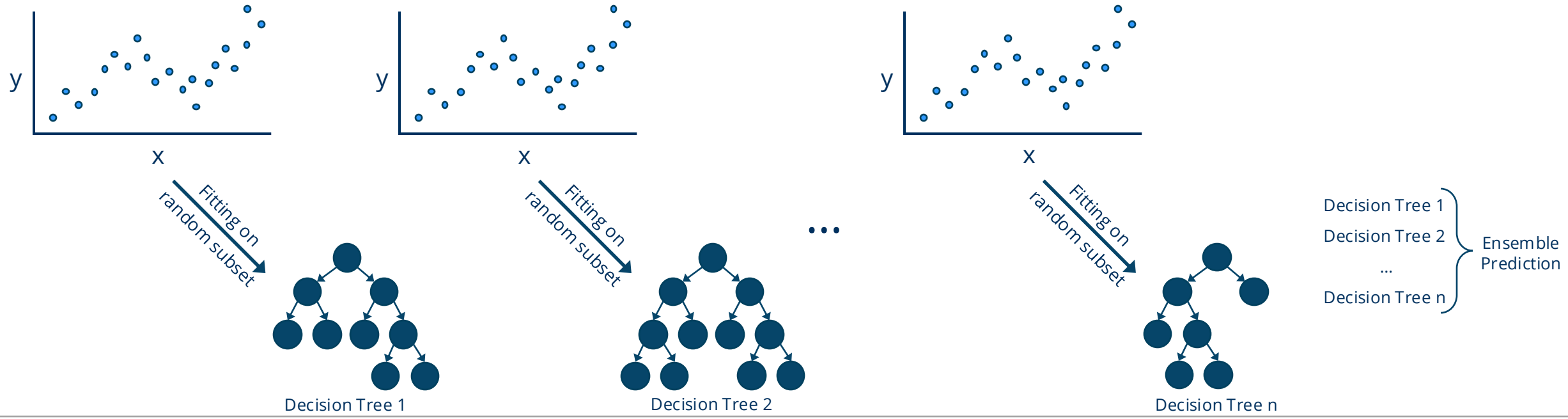


Decision Tree mit größerer Tiefe und mehr Splits für bessere Approximation... tendiert aber zu Überanpassung und schlechter Generalisierung auf neuen Daten

Modeltypen des Maschinellen Lernens (ML)

Random Forest

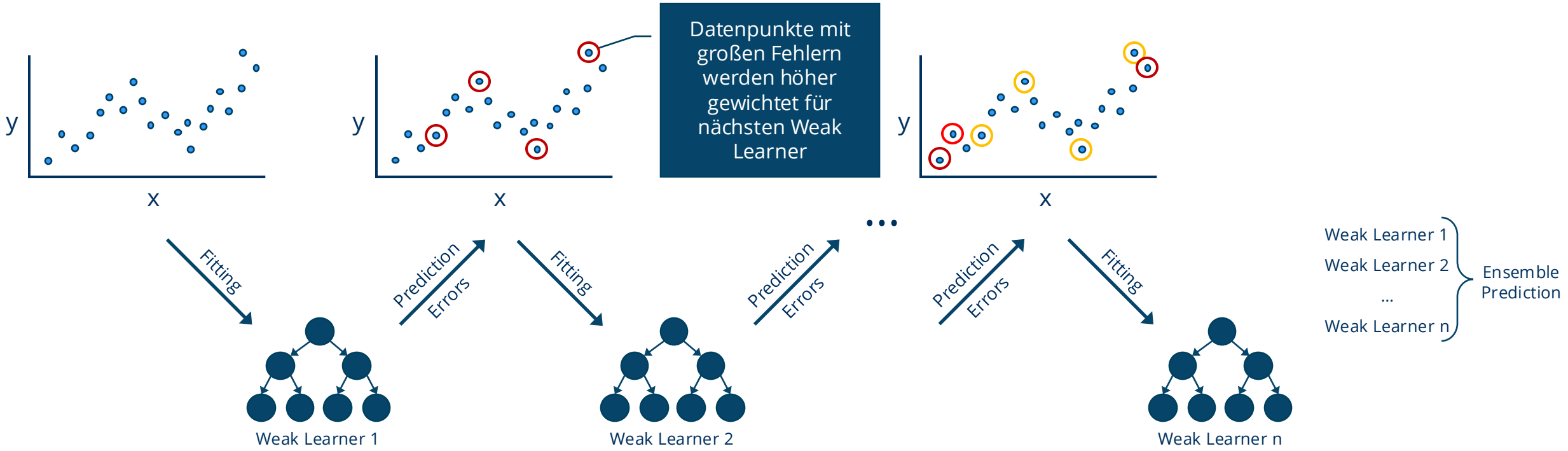
- Ensemble Model – mit *n* Decision Trees (DT)
- Jeder DT wird entweder auf zufällig gewähltem Teilbereich oder allen Daten trainiert
- Vorhersage dann z.b. via Mittelwert über alle DT (Regression) oder Majority Vote (Klassifikation)



Modeltypen des Maschinellen Lernens (ML)

Gradient Boosting

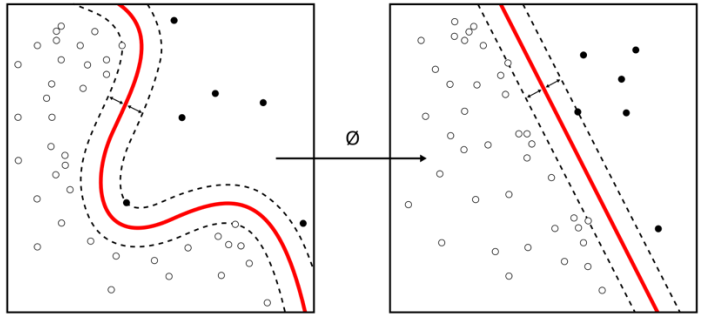
- Ensemble Model – mit *n* (relativ flachen) Decision Trees (DT), sogenannte „Weak Learner“ (WL)
- WL werden sequenziell aufgebaut – Fehler der Vorgänger bereinigen und Gesamtfehler minimieren (oft mittels Gradienten einer Fehlerfunktion „Gradient-boosting“)



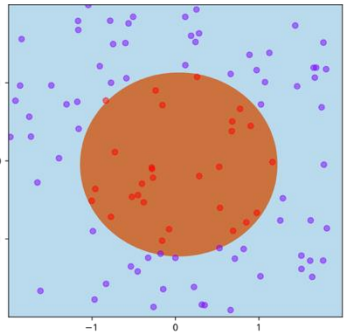
Modeltypen des Maschinellen Lernens (ML)

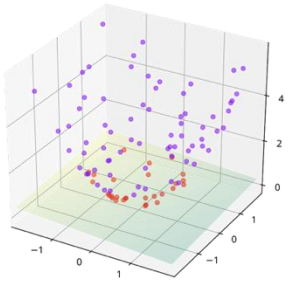
Support Vector Machine (SVM)

- Mathematisch-statistischer Ansatz mit linearen und nicht-linearen (Kernel) Varianten
- Primär für Klassifikation, über Anpassungen auch für Regression (Support Vector Regression)
- Modelliert eine Trennlinie bzw Hyperebene *H*, welche den Abstand (Margin) zw. Datenpunkten maximiert → siehe Abb. rechts, in welcher
  - H_1 die Klassen nicht separiert
  - H_2 die Klassen separiert, aber nicht mit max. Margin
  - H_3 die Klassen mit max. Margin separiert
- Kernel-SVM (Abb. unten) für Modellierung nicht-linearer Probleme

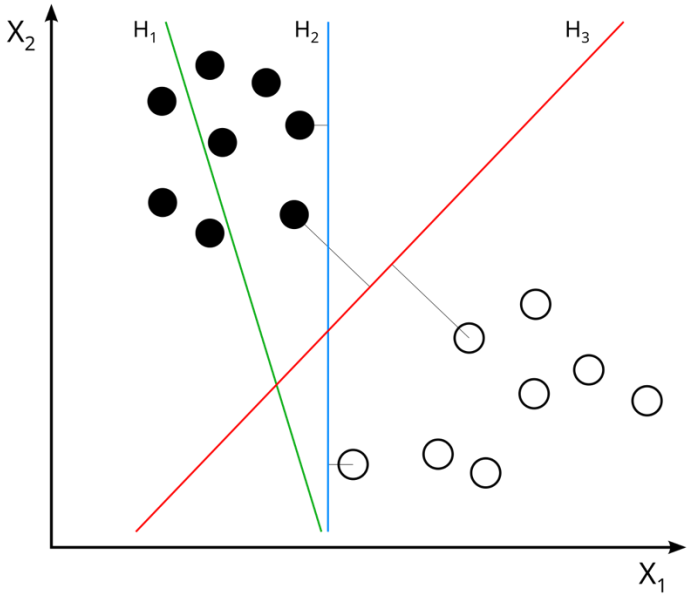


Quelle: Zirguezi, https://en.wikipedia.org/wiki/Support_vector_machine, CC BY-SA 4.0



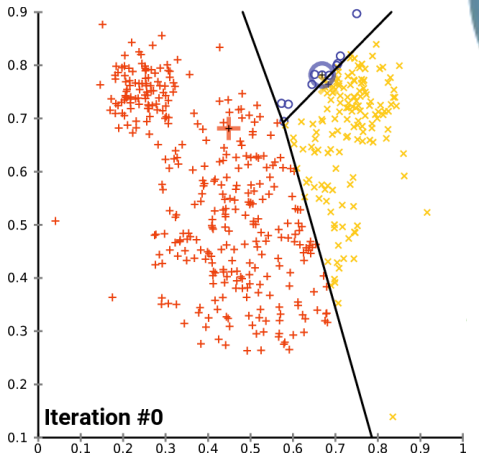


Quelle: Shiyu Ji, https://en.wikipedia.org/wiki/Support_vector_machine, CC BY-SA 4.0



Quelle: ZackWeinberg, https://en.wikipedia.org/wiki/Support_vector_machine, CC BY-SA 3.0

Modeltypen des Maschinellen Lernens (ML)

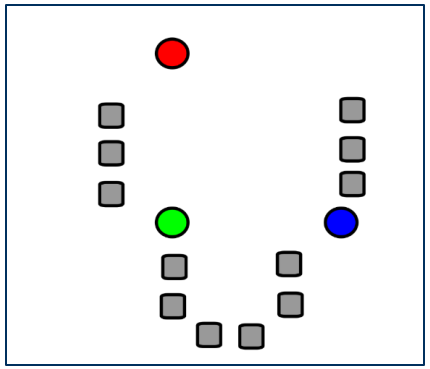


Quelle: Chire, https://de.wikipedia.org/wiki/K-Means-Algorithmus, CC BY-SA 4.0

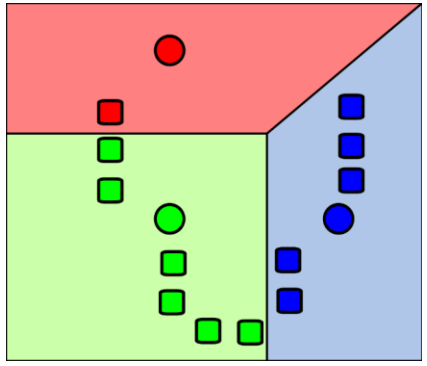
k-Means Clustering

- Einfacher und schneller Clustering-Algorithmus
- Anzahl der Cluster *k* muss angegeben werden
- Erwartet ähnlich große, konvexe Cluster
- Anfällig für Ausreißer und unterschiedliche Skalierung der Feature
- Vergleich von Clustering-Algorithmen: https://scikit-learn.org/stable/modules/clustering.html

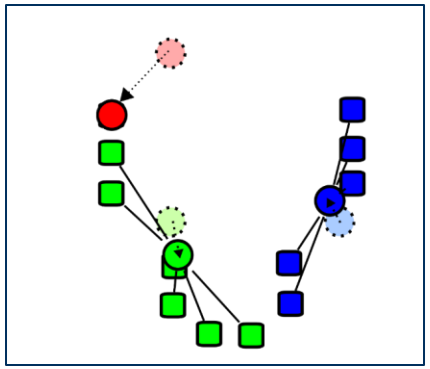
1: Zufällige Auswahl der Cluster-Zentren (Centroids)



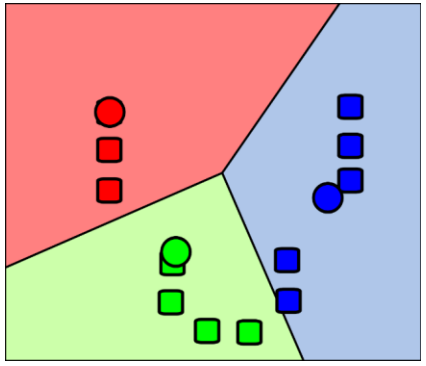
2: Zuordnung von Objekten zum nächsten Zentrum



3: Neuberechnung der Cluster-Zentren



4: Wiederholung von 2 & 3, bis Konvergenz eintritt oder max. Durchläufe erreicht



Quelle: I, Weston.pace, https://de.wikipedia.org/wiki/K-Means-Algorithmus, CC BY-SA 3.0

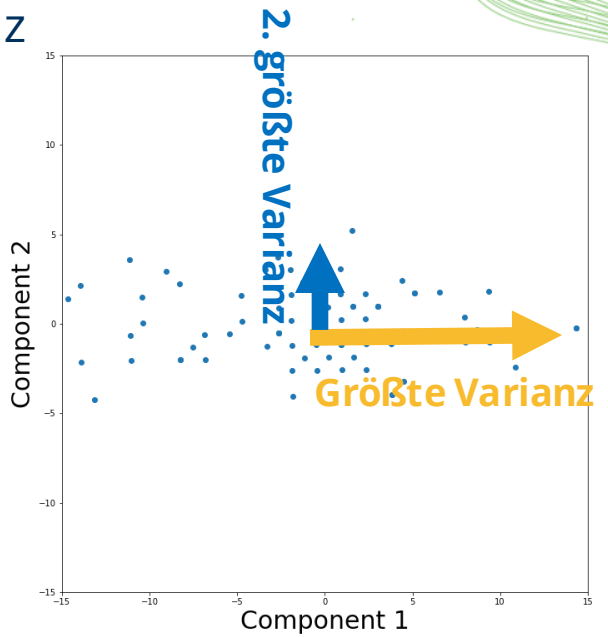
Modeltypen des Maschinellen Lernens (ML)

Principal Component Analysis (PCA)

- Dimensionsreduktion durch lineare Transformation hochdimensionaler Daten
- Erzeugt niedrig-dimensionale Einbettung der *n* Features mit höchster Varianz

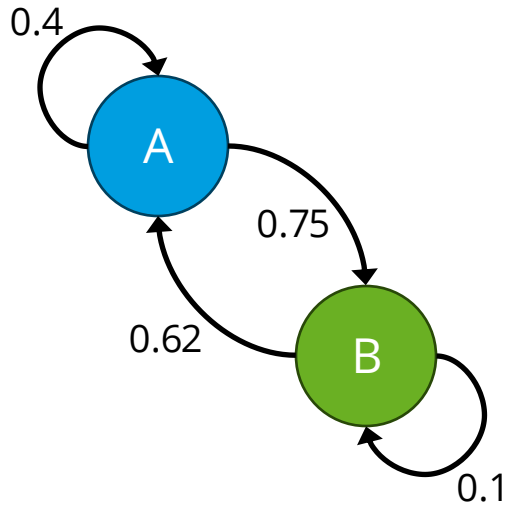
height	width	depth
0.649060	0.213074	0.032167
0.983763	0.533933	0.026125
0.826448	0.223712	0.048805
0.610540	0.574425	0.116101
0.383580	0.042504	0.973645
0.222935	0.842952	0.152771
0.946367	0.780378	0.565486
0.580490	0.001958	0.945884
0.005322	0.019889	0.455281
0.359661	0.426161	0.369291

PCA(2)



- Alternative Ansätze
  insbesondere für komplexere nicht-lineare Zusammenhänge
  - t-SNE (t-Distributed Stochastic Neighbor Embedding)
  - UMAP (Uniform Manifold Approximation and Projection)
  - Autoencoder (Neuronales Netz)

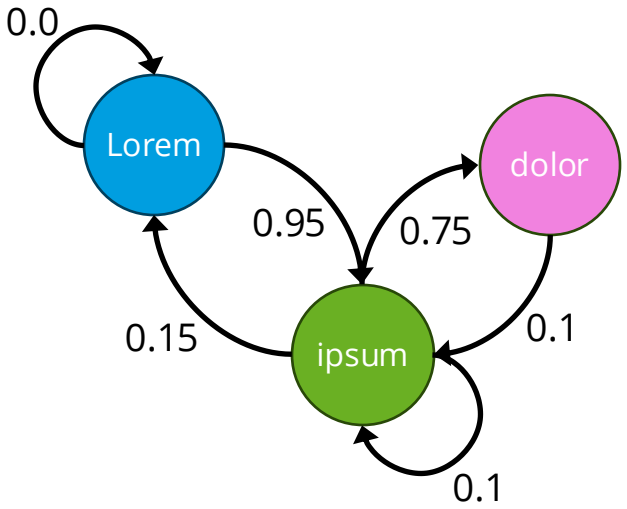
Modeltypen des Maschinellen Lernens (ML)



Markov Chain / Hidden Markow Model

- Stochastischer Prozess zur Modellierung von Zuständen und den Wahrscheinlichkeiten von Zustandsübergängen
- Anwendbar als Generatives Modell bzw. für Sequenz-Bildung
- Bsp: Text-Vervollständigung mit 1-stufiger Markov-Kette

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus facilisis, quam non tempor vulputate, ipsum arcu eleifend dui, vitae sagittis sem lorem a metus. In et efficitur eros. In placerat neque ac odio congue pretium. Sed quis nulla sit amet metus feugiat hendrerit. Pellentesque quis tristique est. Maecenas sodales purus ut tortor...



Lorem ipsum dolor ...

sit = 0.9
non = 0.1
Lorem = 0.0
...

Text-Input

Modellierung der Zustände (Wörter) mit Übergangswahrscheinlichkeiten

Wahrscheinlichkeit des nächsten Wortes

Modeltypen des Maschinellen Lernens (ML)

Bayessche Modelle

- Probabilistische Modelle, basierend auf Bayes Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- Modelparameter als bedingte Verteilungen, basierend auf den Verteilungen der Trainingsdaten
- Berechnung der bedingten Wahrscheinlichkeiten von Ereignissen
- Am Beispiel:

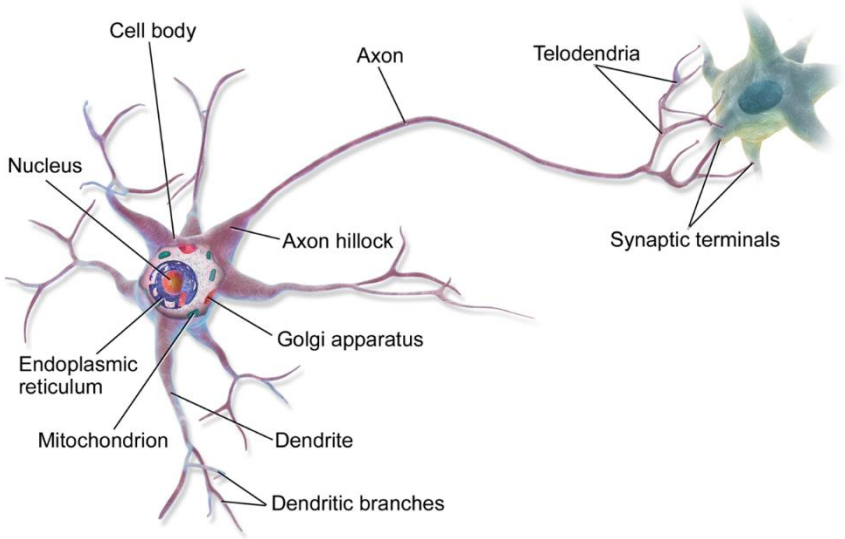
„Wie wahrscheinlich ist eine Klasse (z.B. Spam) für einen Text, nachdem die Wörter des Textes bekannt sind?“

- P(A): sog. „Prior“ (Vorwissen)…wie oft kommt die Klasse generell vor (z.B. 1 von 10 Texten → 0.1)
- P(B|A): sog. „Likelihood“…wie gut passt der Text zur Klasse - bspw. wenn es Spam ist, wie typisch wären diese Wörter dafür? (Naive Bayes: jedes Wort ist unabhängig, Einzelwahrscheinlichkeiten multiplizieren)
- P(B): sog. „Marginal“…„durchschnittliche Seltenheit“ des Textes unabhängig von Klassen
- P(A|B): sog. „Posterior“…aktualisierte Klassenwahrscheinlichkeit, nachdem Text gelesen wurde und Wörter bekannt sind
- Höhere „Posterior“-Wahrscheinlichkeit einer Klasse gewinnt (z.B. Spam)

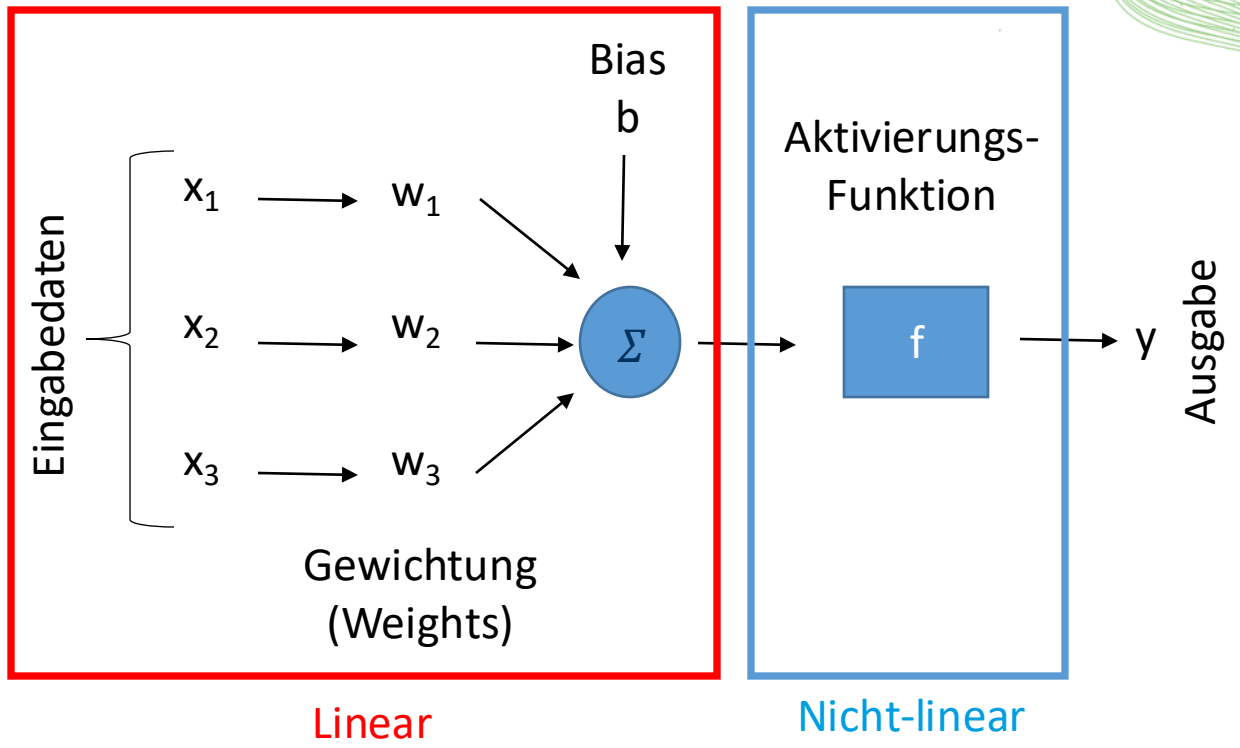
Modeltypen des Maschinellen Lernens (ML)

Neuronale Netze

- Neuron aus Sicht der Biologie



- Neuron in der Data Science



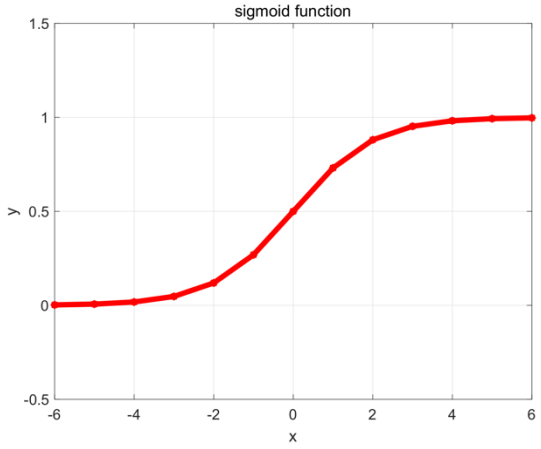
$$y = f(w_1x_1 + w_2x_2 + w_3x_3 + \mathrm{b})$$

Quelle: https://commons.wikimedia.org/wiki/File:Blausen_0657_MultipolarNeuron.png
Lizenz CC-BY 3.0, BruceBlaus

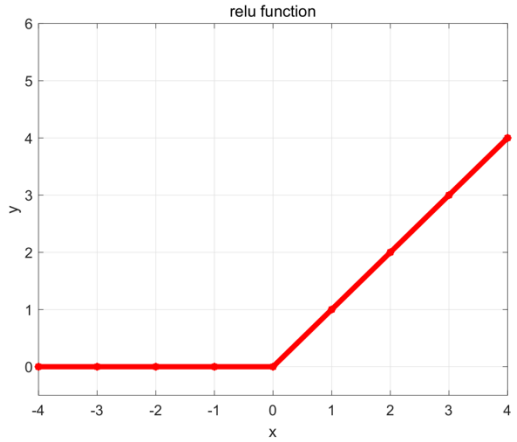
Modeltypen des Maschinellen Lernens (ML)

Neuronale Netze – Aktivierungsfunktion

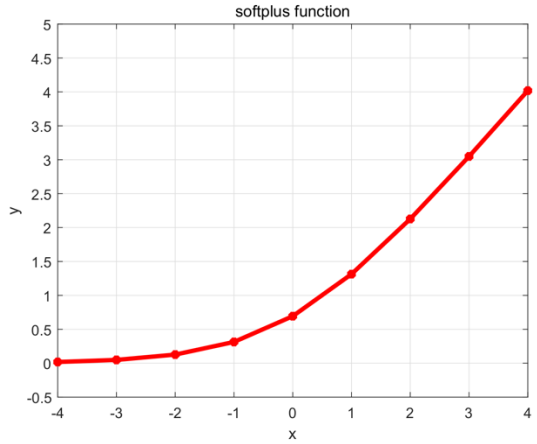
- Nicht-lineare Transformation der linearen Kombination der Features



(a) The curve of sigmoid function



(c) The curve of ReLu function



(d) The curve of softplus function

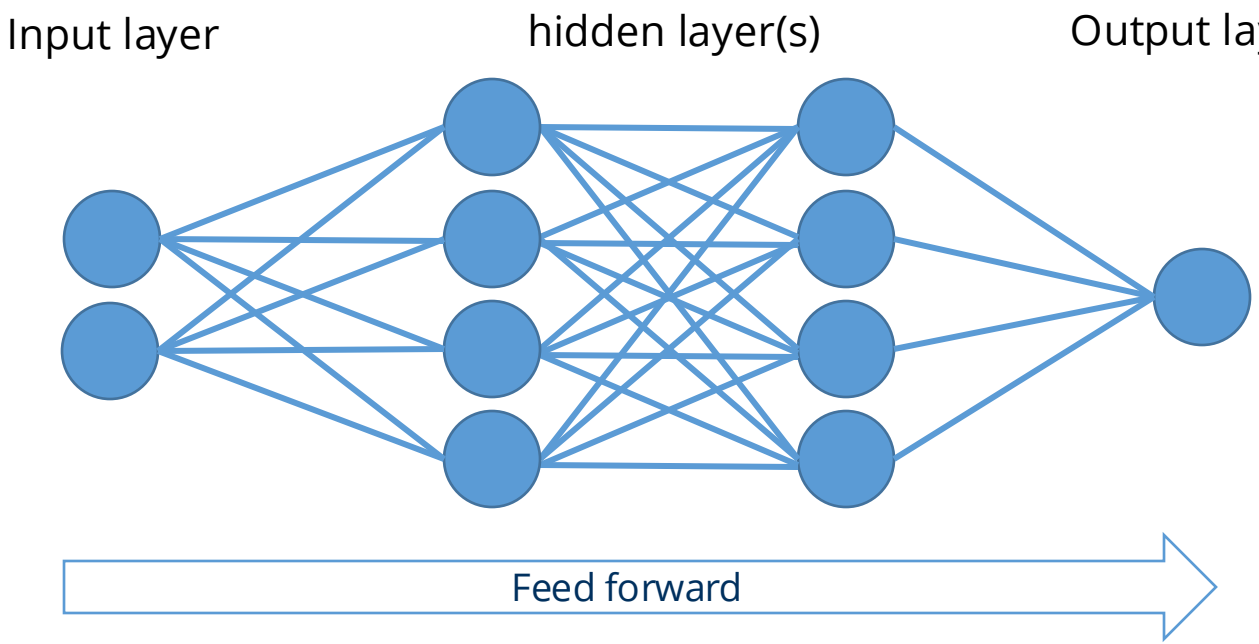
Quelle: Wang, Y., Li, Y., Song, Y., & Rong, X. (2020). The Influence of the Activation Function in a Convolution Neural Network Model of Facial Expression Recognition. Applied Sciences, 10(5), 1897. https://doi.org/10.3390/app10051897, OpenAccess

- Weitere Bsp und Erläuterungen bei Wikipedia: https://en.wikipedia.org/wiki/Activation_function

Modeltypen des Maschinellen Lernens (ML)

Neuronale Netze – Multi-Layer Perceptron

- Mehrere Schichten (Layer) mit Neuronen: Eingabe → Hidden → Ausgabe
- Neuronen einer Schicht sind vollständig mit denen der nächsten verbunden (Fully-Connected)
- Information durchfließen das Netz von Eingabe zu Ausgabe (Feed-Forward)

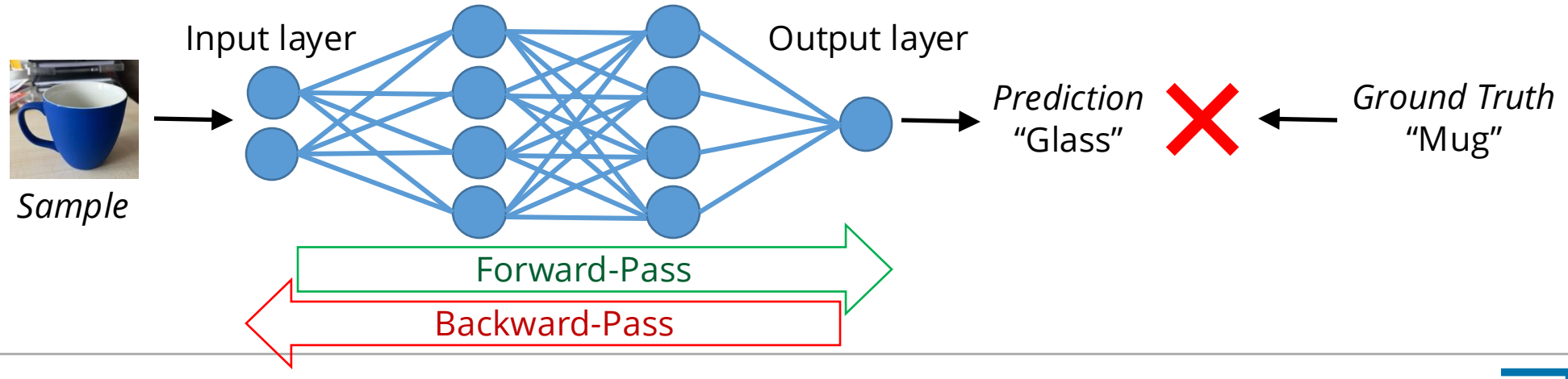


„Deep Learning“
Neuronale Netze mit sehr vielen Hidden Layers

Modeltypen des Maschinellen Lernens (ML)

Neuronale Netze – Lernen durch „Back Propagation“

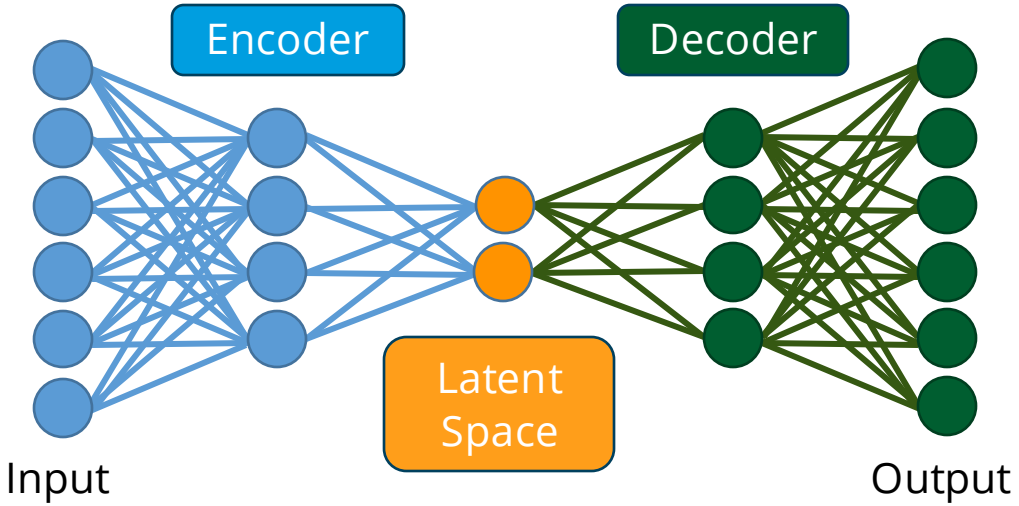
- Schritt 1: Initialisiere *Weights* aller Neuronen mit zufälligen Werten (zw. 0.0 – 1.0)
- Schritt 2: Schicke Datenpunkt (*Sample*) durch das Netz für Vorhersage / Prediction (*Forward-Pass*)
- Schritt 3: Vergleiche Prediction mit bekanntem Label (*Supervised Learning* mit *Ground Truth*), berechne Fehler zwischen Prediction und Ground Truth (via *Loss Function*)
- Schritt 4: Anpassung der Weights in den Neuronen, beginnend bei Ausgabe-Layer (*Backward-Pass*), um berechneten Fehler zu minimieren (z.B. via *Gradienten-Abstiegsverfahren*)
- Wiederhole 2-4 für jedes Sample in den Trainingsdaten, durchlaufe dies für mehrere *Epochen*



Modeltypen des Maschinellen Lernens (ML)

Autoencoder

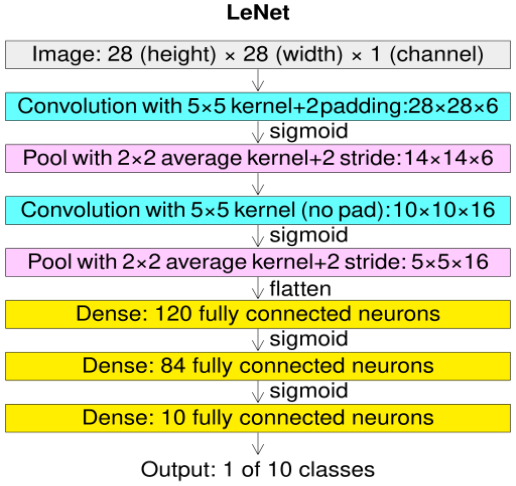
- Neuronales Netz zur Kompression (*Encoder*) und Rekonstruktion (*Decoder*)
- Wichtigste Eigenschaften der Features eingebettet im niedrigdimensionalen *Latent Space*
- Training zielt darauf ab, den Fehler bei der Rekonstruktion zu minimieren
- Anwendung z.B. Dimensionsreduktion oder Anomalie Detektion (großer Fehler = Anomalie)



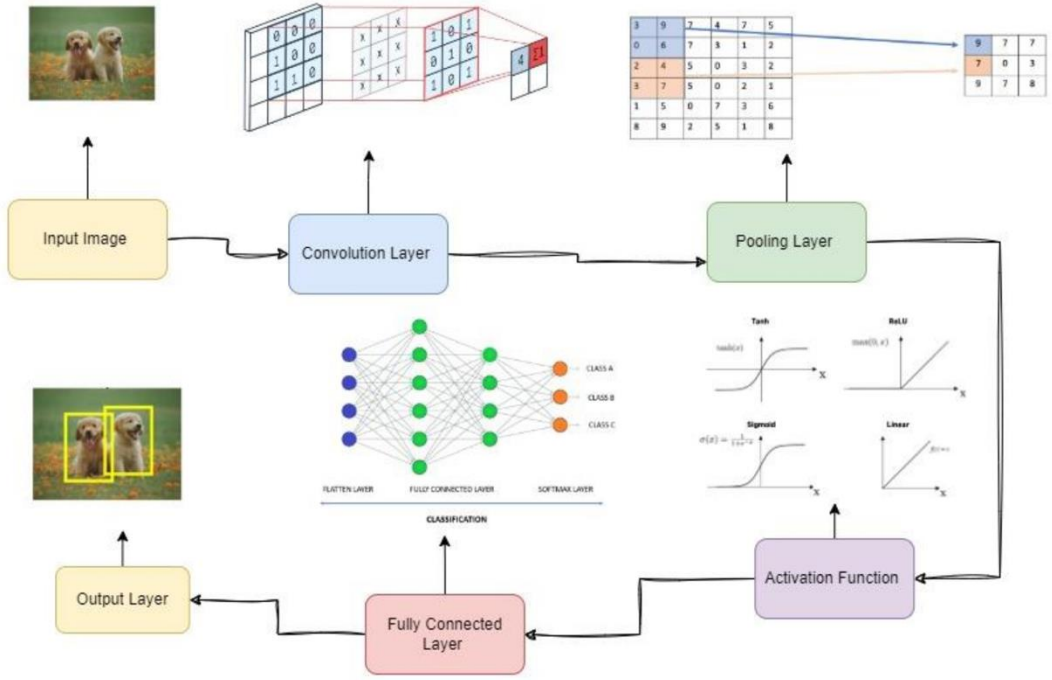
Modeltypen des Maschinellen Lernens (ML)

Convolutional Neural Network (CNN)

- Neuronales Netz mit verschiedenen Layer-Arten zur Extraktion lokaler bis globaler Merkmale
  - Convolutional Layer (Convolution = Faltung)
  - Pooling Layer (Aggregation)
  - Fully Connected Layer (Klassifikation)
- Anwendung bei Bildverarbeitung (Objekterkennung, Klassifikation), aber auch Zeitreihenanalyse



Quelle: Cmglee, https://en.wikipedia.org/wiki/Convolutional_neural_network, CC BY-SA 4.0



Quelle: Taye, M. M. (2023). Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions. *Computation*, *11*(3), 52. https://doi.org/10.3390/computation11030052, OpenAccess

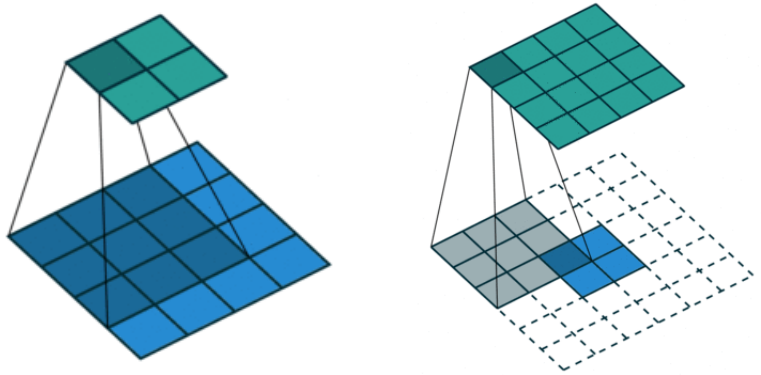
Modeltypen des Maschinellen Lernens (ML)

Convolutional Neural Network (CNN)

- Convolutional Layer – nicht fully-connected, sog. Kernel wandert über Input und "faltet" diesen, extrahiert dabei Feature und hat anpassbare Weights

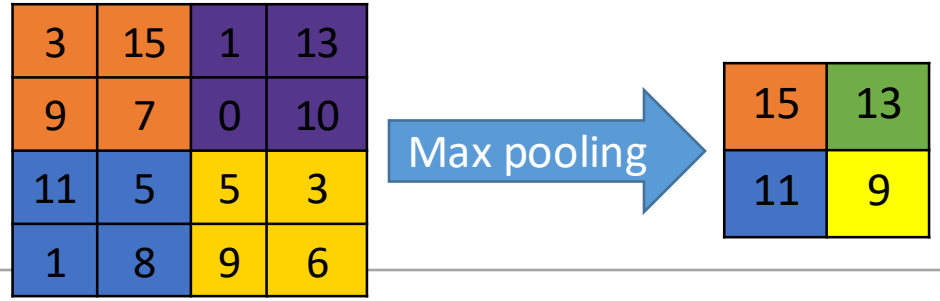
0	0	0
0	1	0
0	0	0

3x3 Identity Kernel

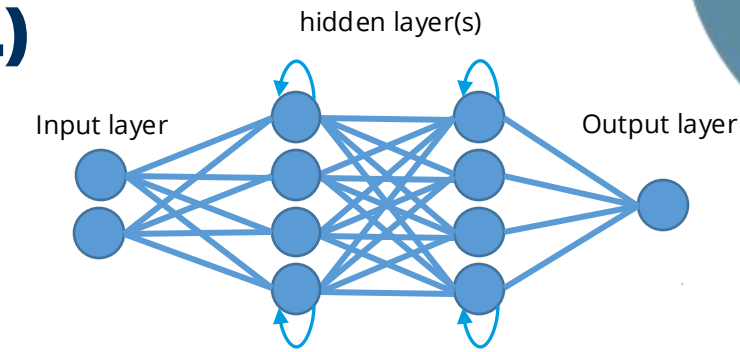


Quellen:
Vincent Dumoulin and Francesco Visin (2018). A guide to convolution arithmetic for deep learning. *arXiv*. https://doi.org/10.48550/arXiv.1603.07285
Vincent Dumoulin, Francesco Visin, https://github.com/vdumoulin/conv_arithmetic, MIT

- Pooling Layer – aggregiert Input (z.b. max, avg, ...)



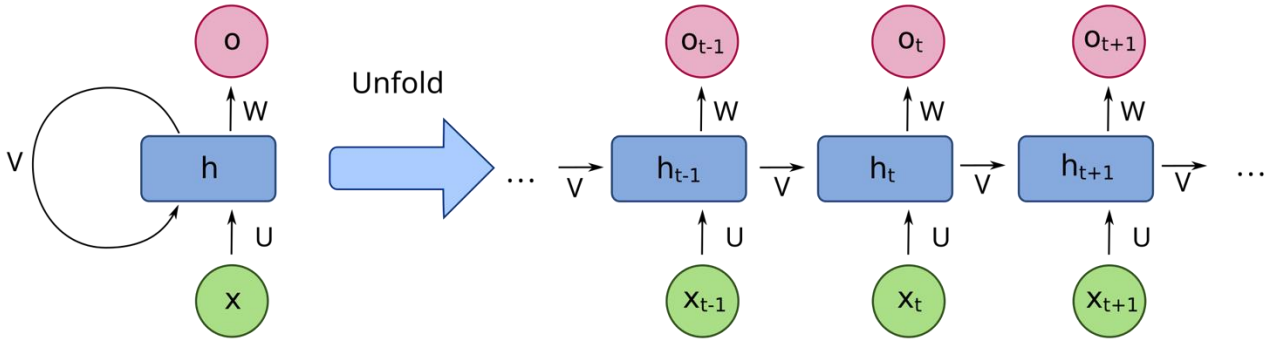
Modeltypen des Maschinellen Lernens (ML)



Stark vereinfachte Darstellung

Recurrent Neural Network (RNN / LSTM / GU)

- Neuronales Netz zur Modellierung von Sequenzen
- Verfügt über „Gedächtnis“ mithilfe von „Feedback Loops“
- Informationen fließen nicht mehr nur von Ein- zu Ausgabe (Feedforward), sondern werden auch wieder in die Neuronen der Hidden Layer zurückgeführt
- Varianten davon ermöglichen u.a. „Langzeit-Gedächtnis“ (Long Short Term Memory, LSTM)



x – Eingabe
h – Hidden (Memory)
o – Ausgabe
U, V, W – entsprechende Weights

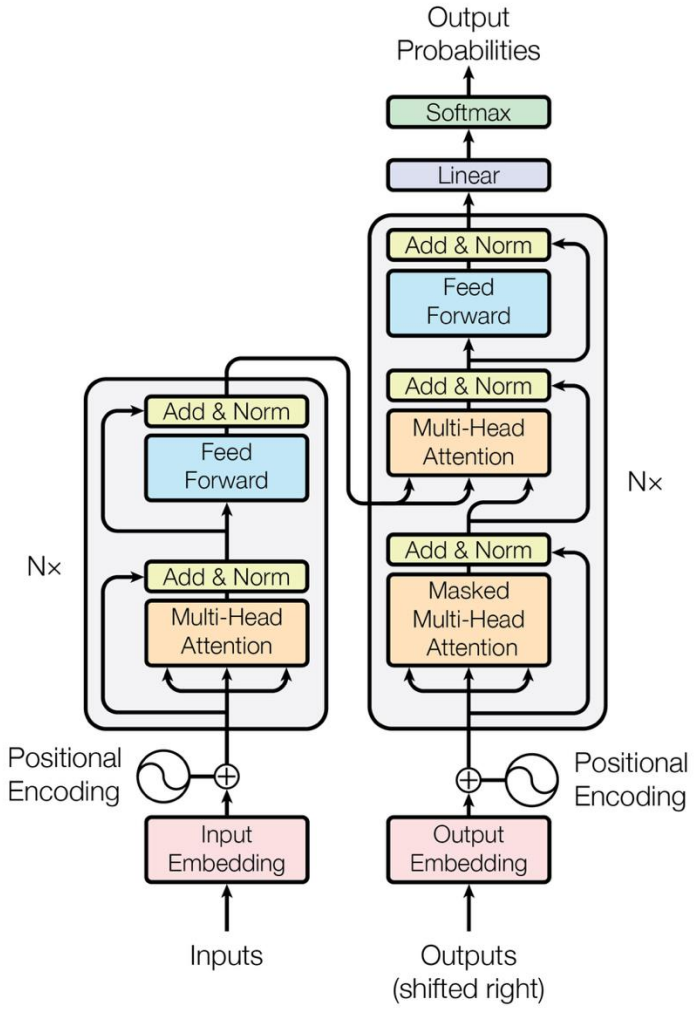
Unfold…”Aufrollen” über die Zeit:
Bei Zeitpunkt t fließt vorheriger Zustand t-1 über V ein
Bei Zeitpunkt t+1 fließt vorheriger Zustand t über V ein
…

Quelle: fdeloche, https://en.wikipedia.org/wiki/Recurrent_neural_network, CC BY-SA 4.0

Modeltypen des Maschinellen Lernens (ML)

Transformer

- Komplexes, auf Deep-Learning basierendes Modell für Generative KI, Sequenz-Modellierung…
- Besteht aus verschiedenen Komponenten, gruppiert in Encoder und Decoder

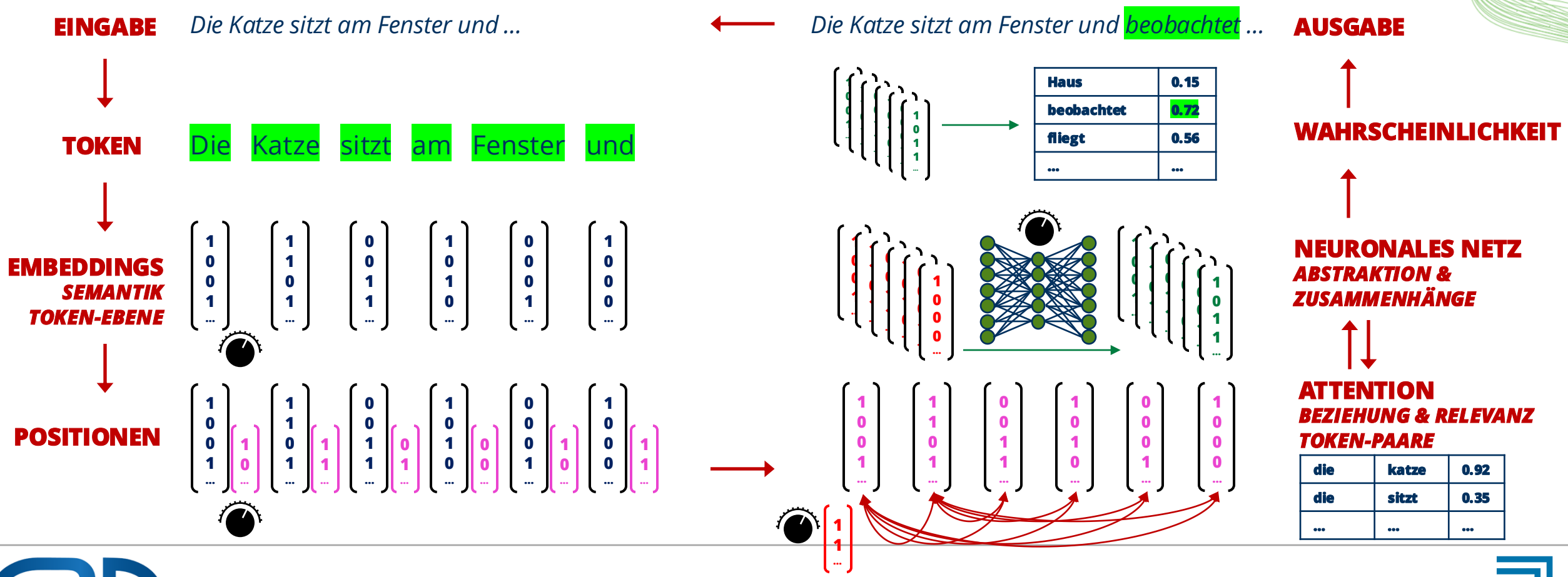


Quelle: Vaswani et al (2017), https://arxiv.org/abs/1706.03762

Modeltypen des Maschinellen Lernens (ML)

Gewichte (weights) bei fast allen Operationen
Anpassbare Multiplikatoren
„Stellschrauben“

Transformer - Ablauf



Modeltypen des Maschinellen Lernens (ML)

Transformer – Architektur erklärt

Encoder Block
Eingangsdaten verarbeiten
Bedeutung extrahieren

Add & Norm
Trainingsprozess stabilisieren

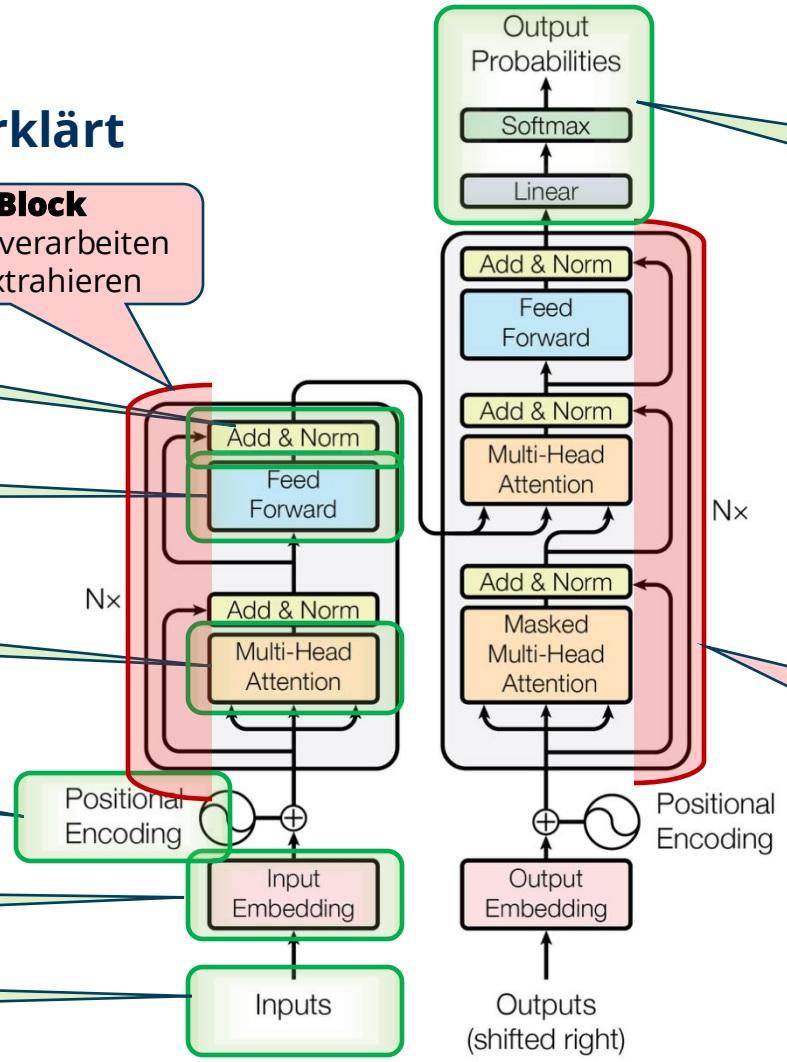
Feed Forward
Neuronales Netz, Nicht-lineare Transformation

Multi-Head Attention
Beziehungen zwischen Embeddings berechnen

Positional Encoding
Embeddings mit Positions-Info anreichern

Embeddings
Embeddings aus Tokens erzeugen

Tokenizer
Tokens aus Eingangsdaten erzeugen



Output
Wahrscheinlichkeit für nächsten Token

Decoder Block
Vorangegangenes Token bzw Vorhersage und aktuellen Encoder Output zusammenführen, um nächsten Token vorherzusagen

Source: Vaswani et al (2017), https://arxiv.org/abs/1706.03762

Praktische Übung

Handhabung von Bibliotheken für Machine Learning

- scikit-learn 

  - Werkzeuge für prädikative Datenanalyse
  - Methoden des überwachten und unüberwachten ML (CPU-basiert, nur kleiner Teilbereich von Neuronalen Netzen, kein Deep Learning)
  - Werkzeuge für Datenvorverarbeitung, Feature Selection, Parameter-Tuning und Evaluation
  - https://scikit-learn.org

- NLTK und Gensim (bereits in Teil 2 der Schulung kennengelernt)
  - Natural Language Processing und unüberwachten ML
  - https://www.nltk.org/index.html
  - https://radimrehurek.com/gensim/index.html