

Programm

- **Tag 1:**
- Generelle Einführung in ML / AI
- Grundlagen NLP
- Praxisblock NLP
- Embeddings
- Praxisblock Embeddings
- Abschluss mit Fragen und Diskussion
- **Tag 2:**
- Kurzes Recap
- RAG
- Praxisblock naive RAG
- Advanced RAG und Agents
- Praxisblock Agent
- MLOps / Abschluss mit Fragen und Diskussion

Eure Erwartungen

Menti.com
7131 8194





ML/AI Schulung

AGENDA

Einführung

NLP

Embeddings

Abschluss



Begrüßung

Daniel Dutli

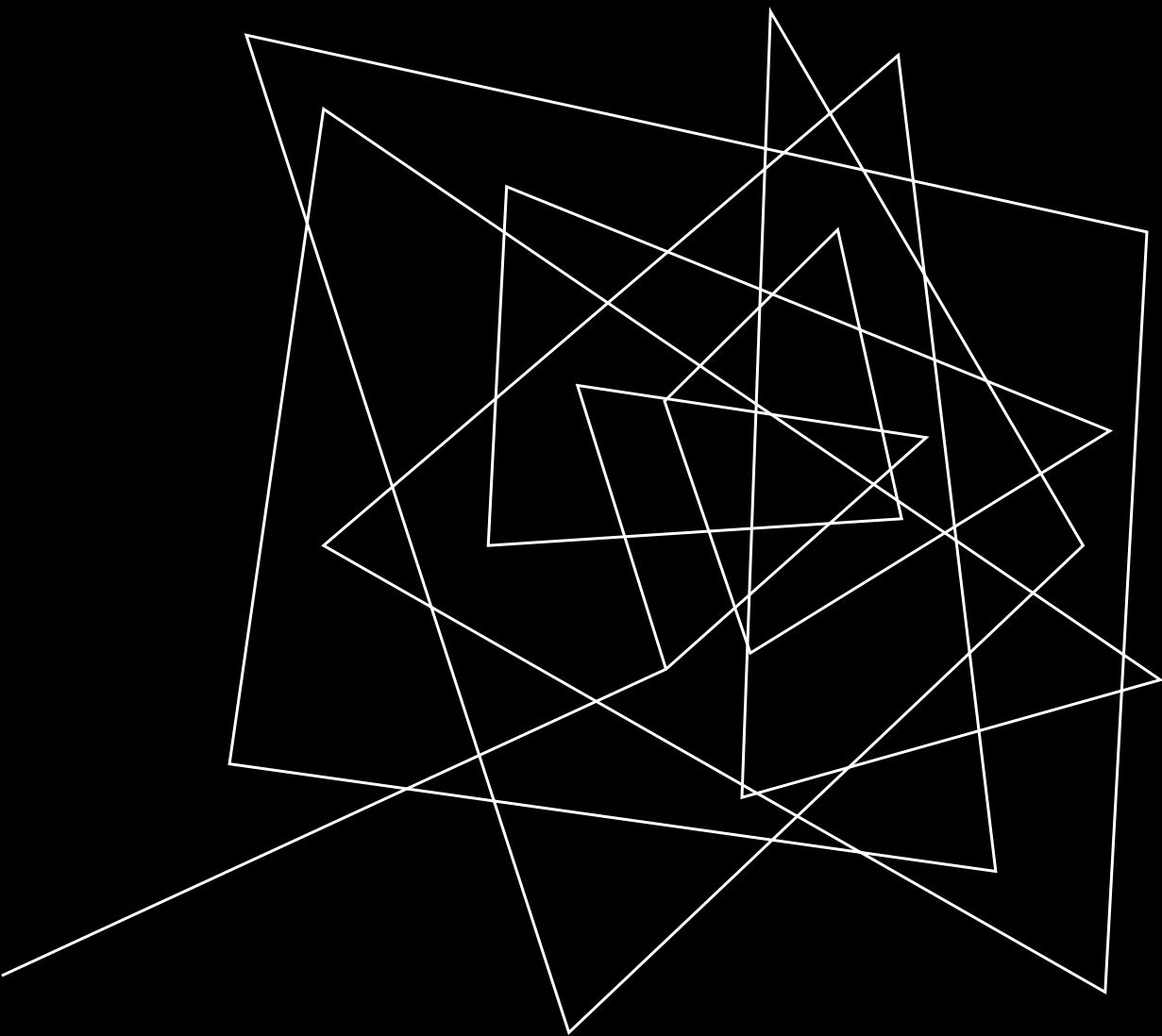
Typ der das Zeug erledigt

- Data Scientist aus Leidenschaft
- Bekennender ~~Vibe~~-Coder Weintrinker
- Feels the AGI
- Mitgründer der enki GmbH

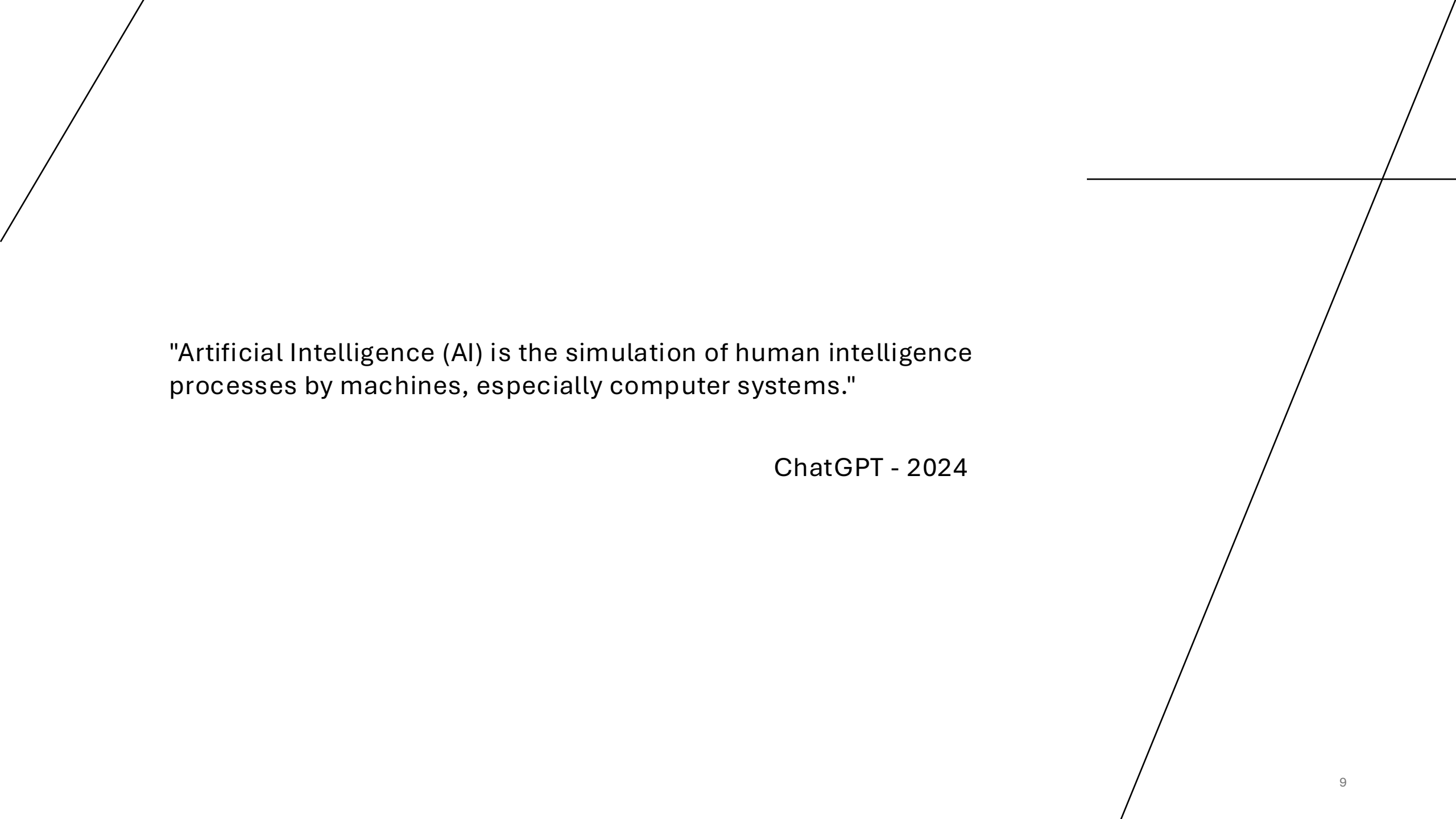
Marco Crisafulli

Vater, ML Enthusiast und manchmal lustig

- Studium in Informatik an der HSR
- Software-Entwickler bei HxGN Schweiz
- Data-Engineer bei Avobis Data
- Hat Cursor immer noch nicht installiert
- Mitgründer der enki GmbH

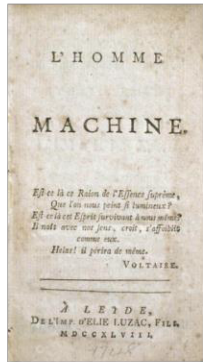


Geschichte

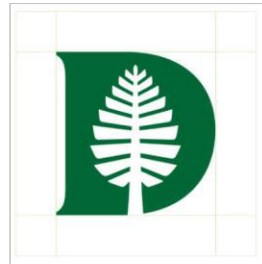


"Artificial Intelligence (AI) is the simulation of human intelligence processes by machines, especially computer systems."

ChatGPT - 2024



La Mettrie - 1748



Dartmouth Workshop - 1956

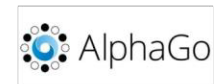
Der erste Winter - 1974

Expert System Boom - 1980

Der zweite Winter - 1987



Deep Blue - 1996

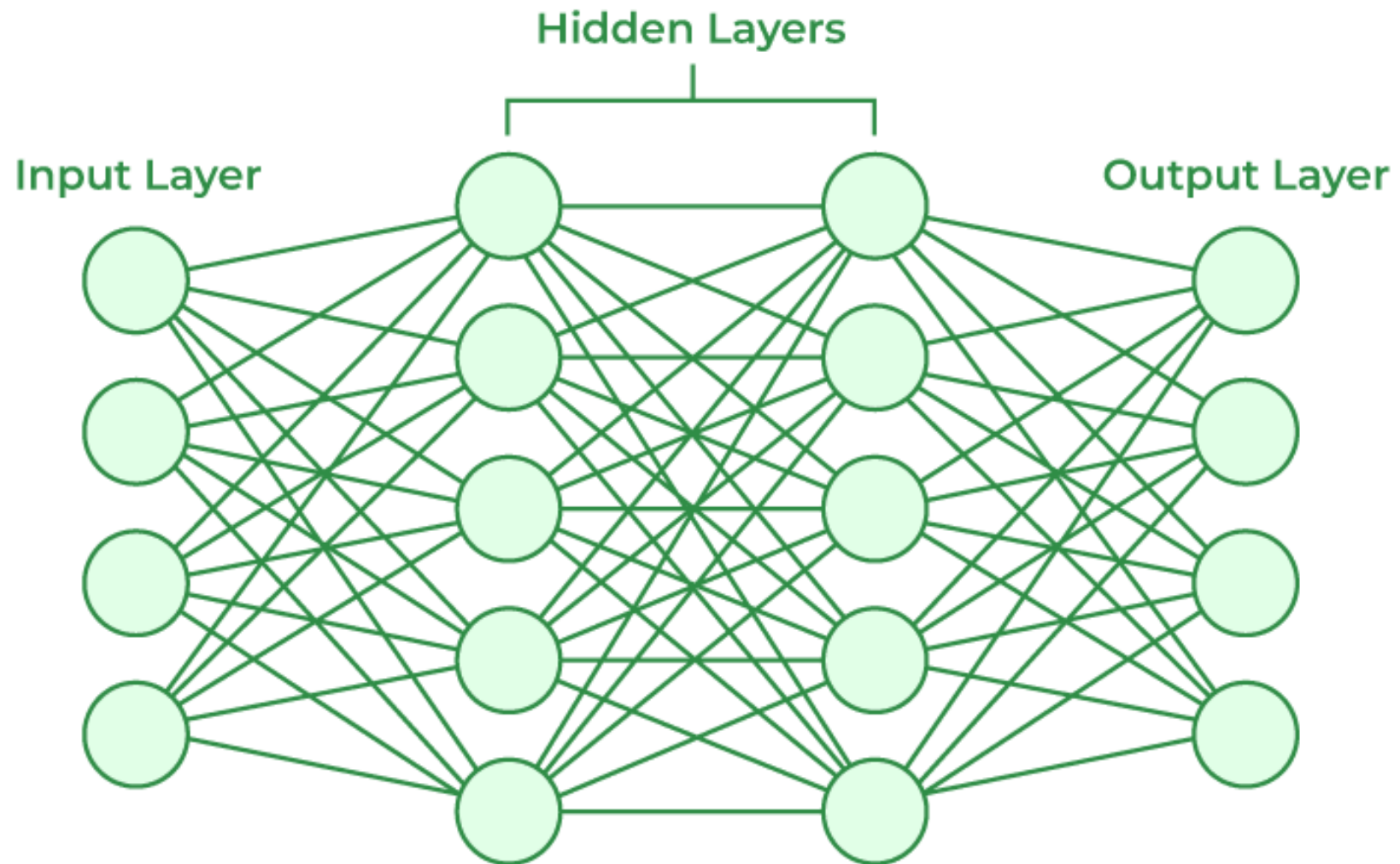


AlphaGo - 2015



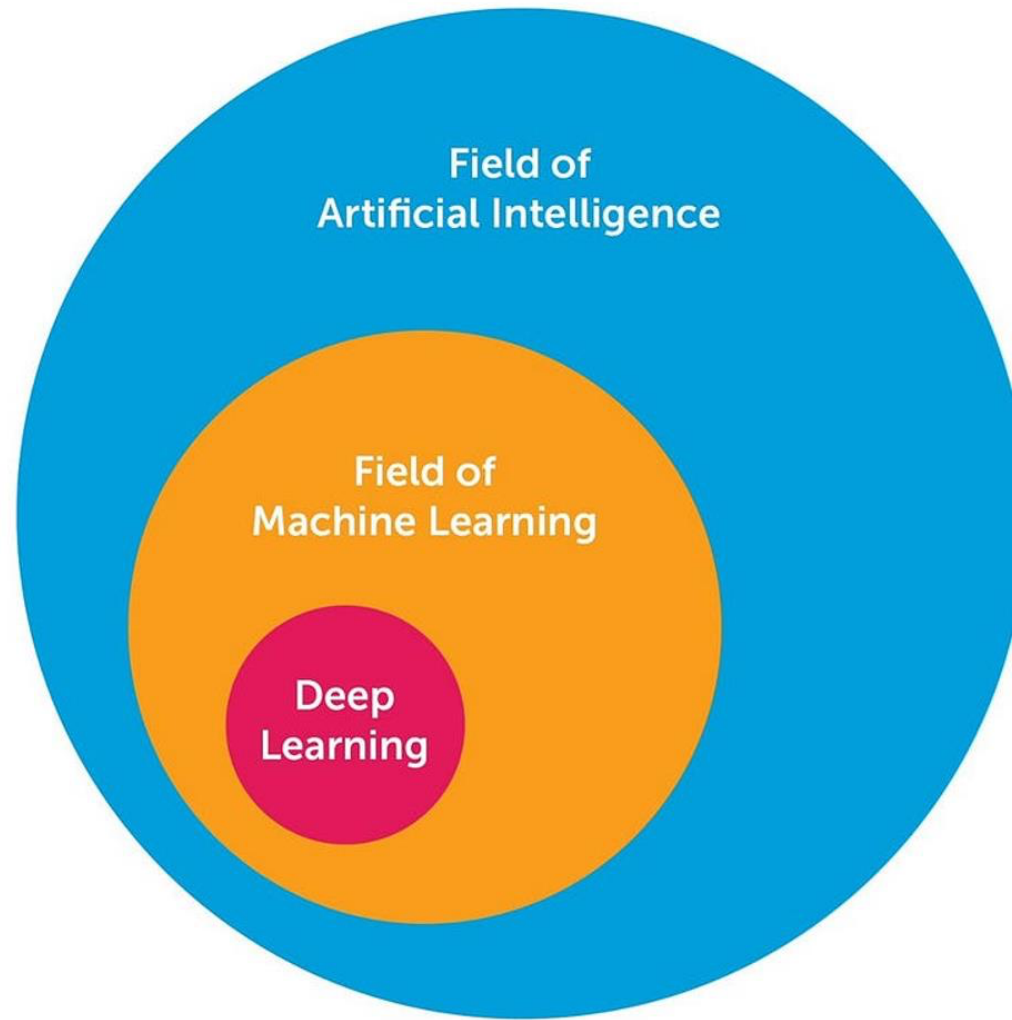
ChatGPT - 2022







Einordnung



Unterschiede

Aspekt	Klassische Programmierung	Maschinelles Lernen
Problemlösung	Explizite Regeln und Logik	Lernen aus Daten
Umgang mit Unsicherheit	Deterministisch	Probabilistisch
Datenabhängigkeit	Gering	Hoch
Anpassungsfähigkeit	Statisch	Dynamisch
Entwicklungsprozess	Linear	Iterativ

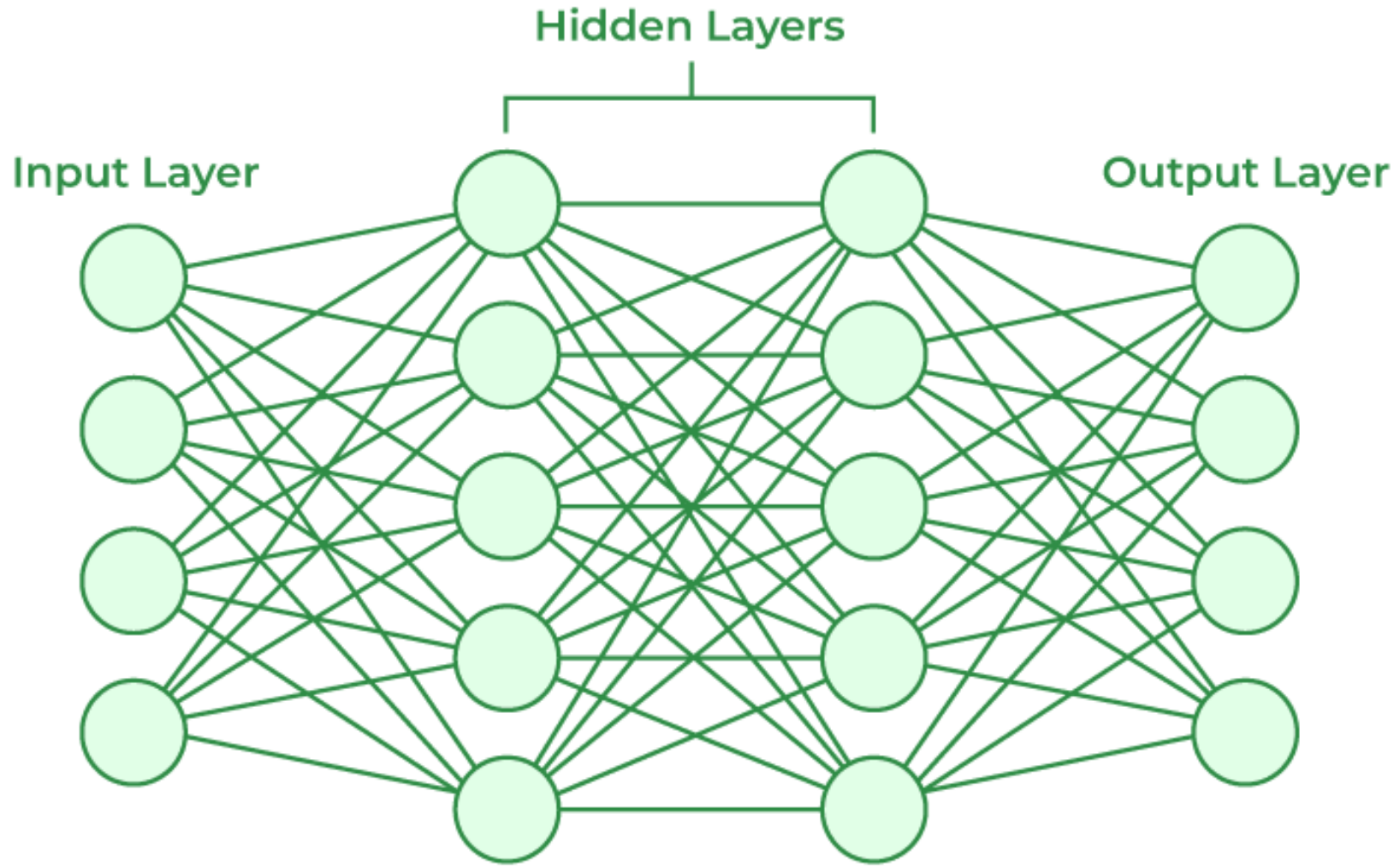
Funktionsweise

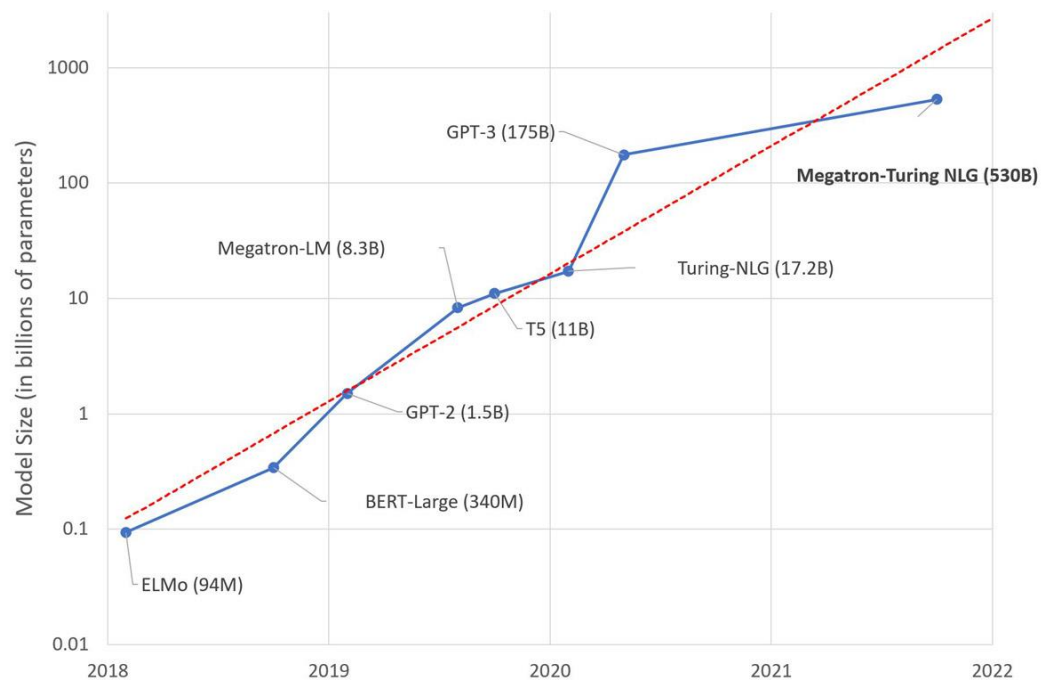
Training

- Das Modell wird angepasst
- Viele Daten werden dem Modell "gezeigt"
- Label (das was gelernt werden soll) muss bekannt sein
- Das Modell lernt aus Fehler

Inference

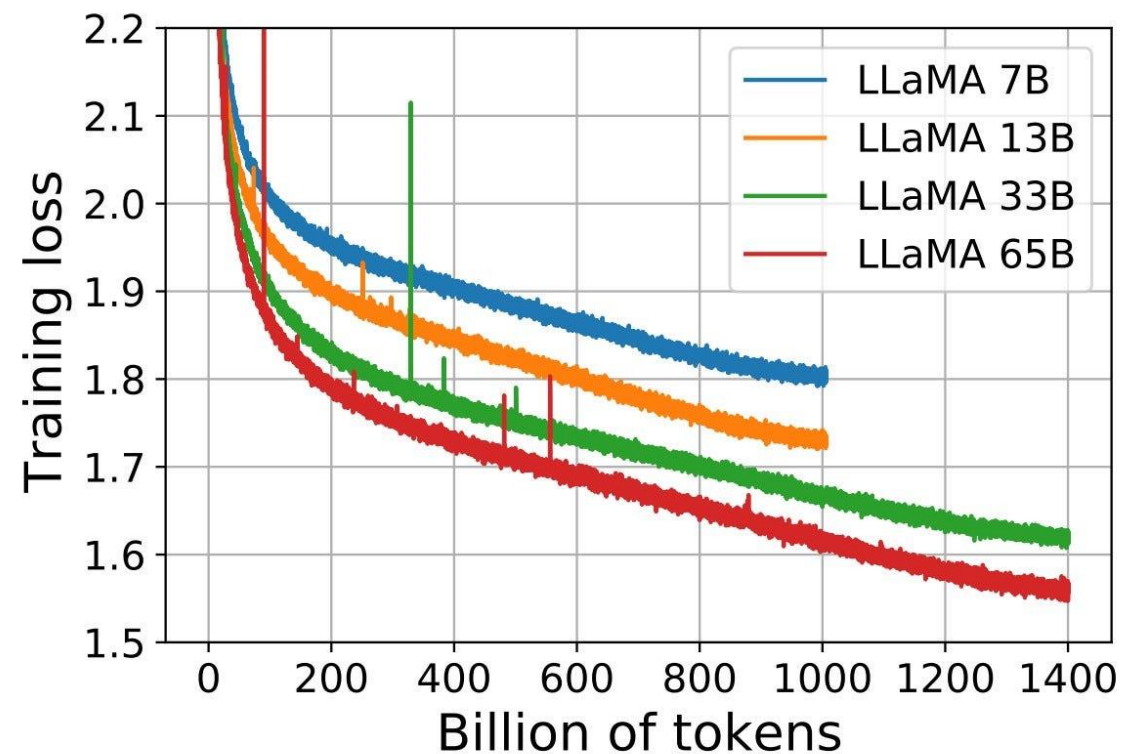
- Das Modell wird nicht angepasst
- Einzelne Daten werden dem Modell "gezeigt"
- Label ist nicht bekannt
- Das Modell lernt nicht

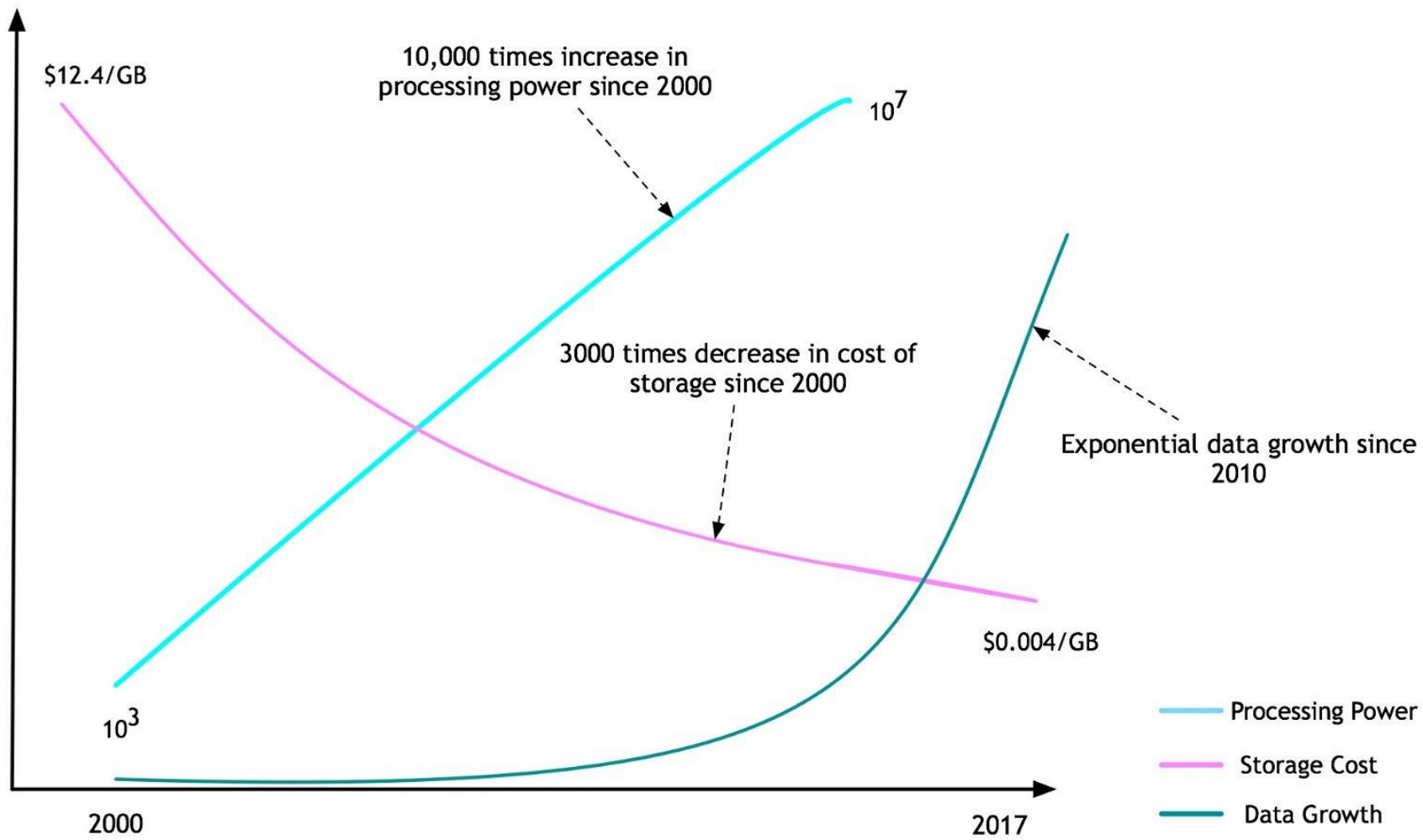


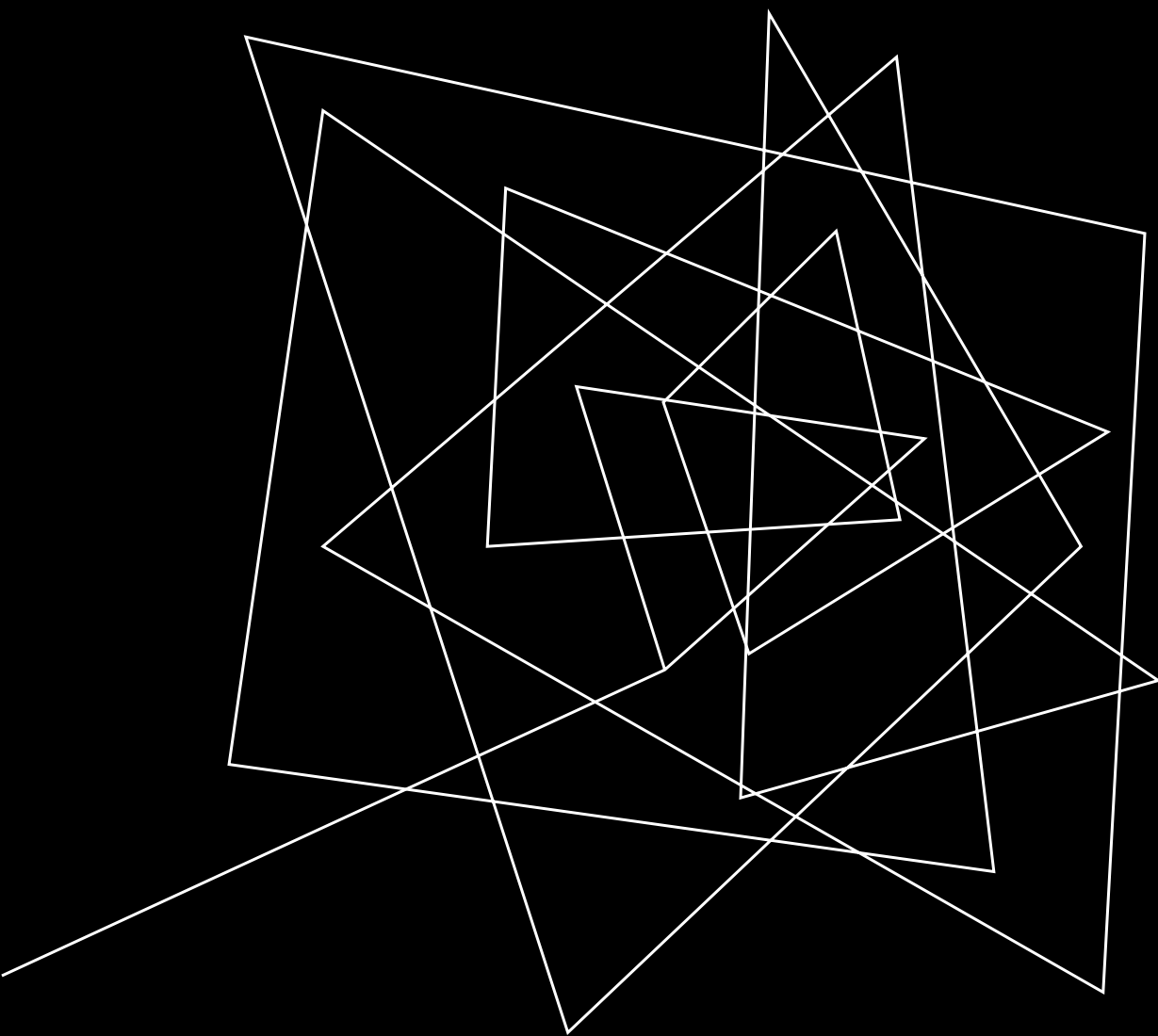


Modellgröße zu Zeit

Modell Qualität zu Anzahl Datenpunkte





















































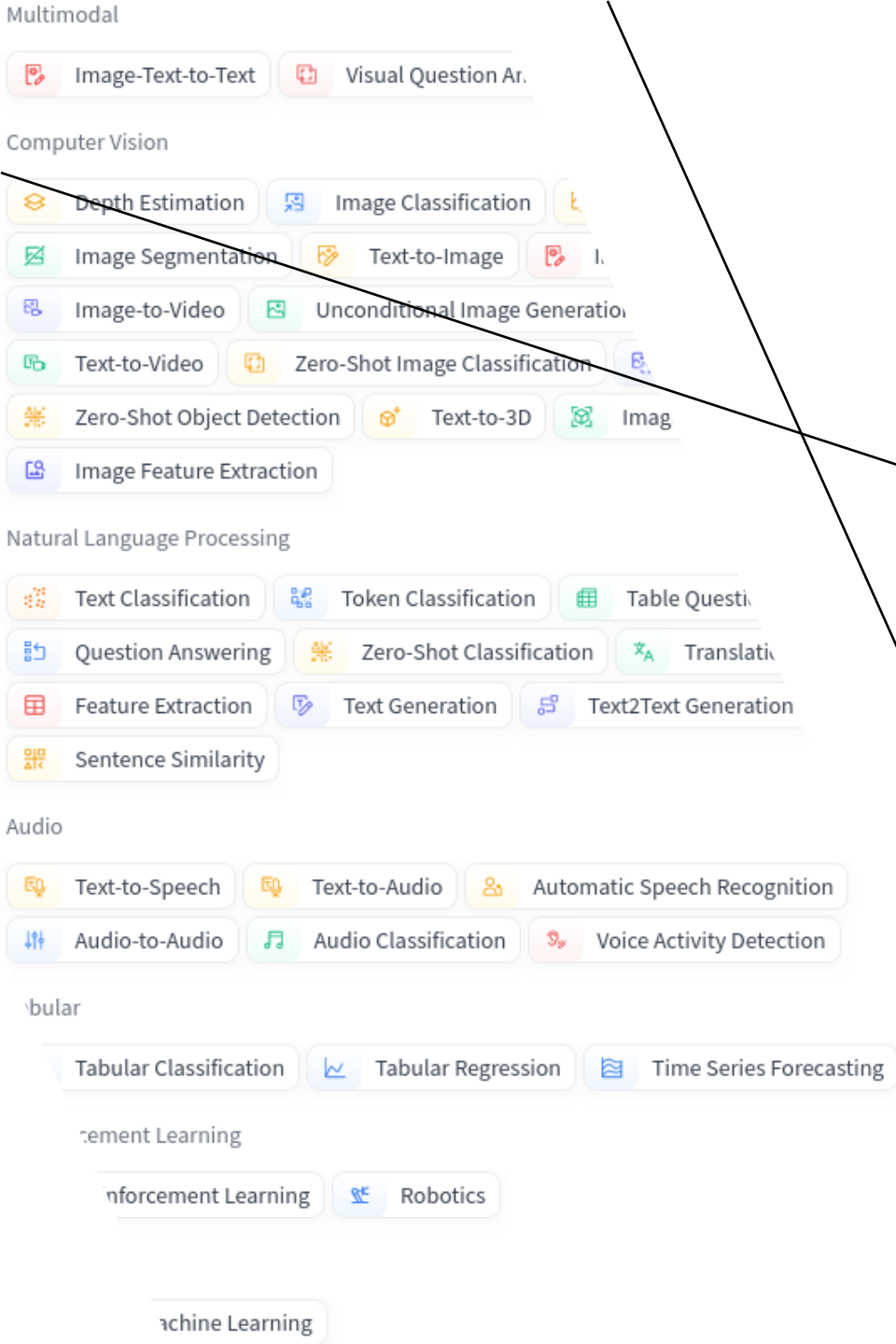




Möglichkeiten

The Top 50 Gen AI Web Products, by Unique Monthly Visits

1.  ChatGPT	11.  IIElevenLabs	21.  PhotoRoom	31.  PIXAI	41.  MaxAI.me
2.  Gemini*	12.  Hugging Face	22.  YODAYO	32.  ideogram	42.  Craiyon
3.  character.ai	13.  Leonardo.Ai	23.  Clipchamp	33.  invideo AI	43.  OpusClip
4.  liner	14.  Midjourney	24.  runway	34.  Replicate	44.  BLACKBOX AI
5.  QuillBot	15.  SpicyChat	25.  YOU	35.  Playground	45.  CHATPDF
6.  Poe	16.  Gamma	26.  DeepAI	36.  Suno	46.  PIXELCUT
7.  perplexity	17.  Crushon AI	27.  Eightify	37.  Chub.ai	47.  Vectorizer.AI
8.  JanitorAI	18.  cutout.pro	28.  candy.ai	38.  Speechify	48.  DREAMGF
9.  CIVITAI	19.  PIXLR	29.  NightCafe	39.  phind	49.  Photomyne
10.  Claude	20.  VEED.IO	30.  VocalRemover	40.  NovelAI	50.  Otter.ai



Tasks

Bilder / Videos

Text

Audio

Tabellen

Spiele

Graphen

Large Language Model (LLM)

Einordnung

- Gehören zum Task Text -> Text Generation
- Ziel ist den nächsten Token vorherzusagen
- Sehr teuer

Anwendung

- Chatbots
- RAG
- Spezialisierte Text Aufgaben

Anbieter

OpenAI (Microsoft) - ChatGPT

Google - Gemini

Anthropic (Amazon) - Claude

Meta – Llama

Image Generator

Einordnung

- Gehören zum Task Bild-> Text-To-Image (u.a.)
- Ziel ist Bildstörungen zu entfernen
- Teuer

Anwendung

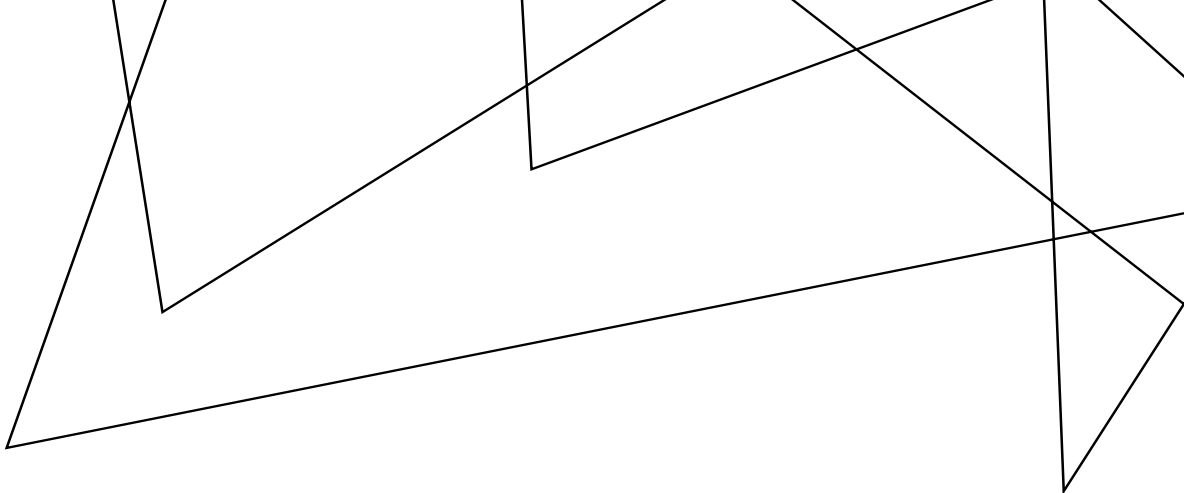
- Kunst
- Bildbearbeitung
- Unterhaltung

Anbieter

OpenAI (Microsoft) - Dall-E

Midjourney Inc. - Midjourney

StabilityAI – Stable Diffusion



Speech synthesis

Einordnung

- Gehören zum Task Audio-> Text-To-Speech
- Ziel ist das nächste Stück Ton vorherzusagen
- Teuer wenn sehr gut

Anwendung

- Sprachassistent
- Service-Line
- Hilfe für Personen mit visuellen Einschränkungen

Anbieter

OpenAI (Microsoft) - GPT-4o

ElevenLabs - ElevenLabs

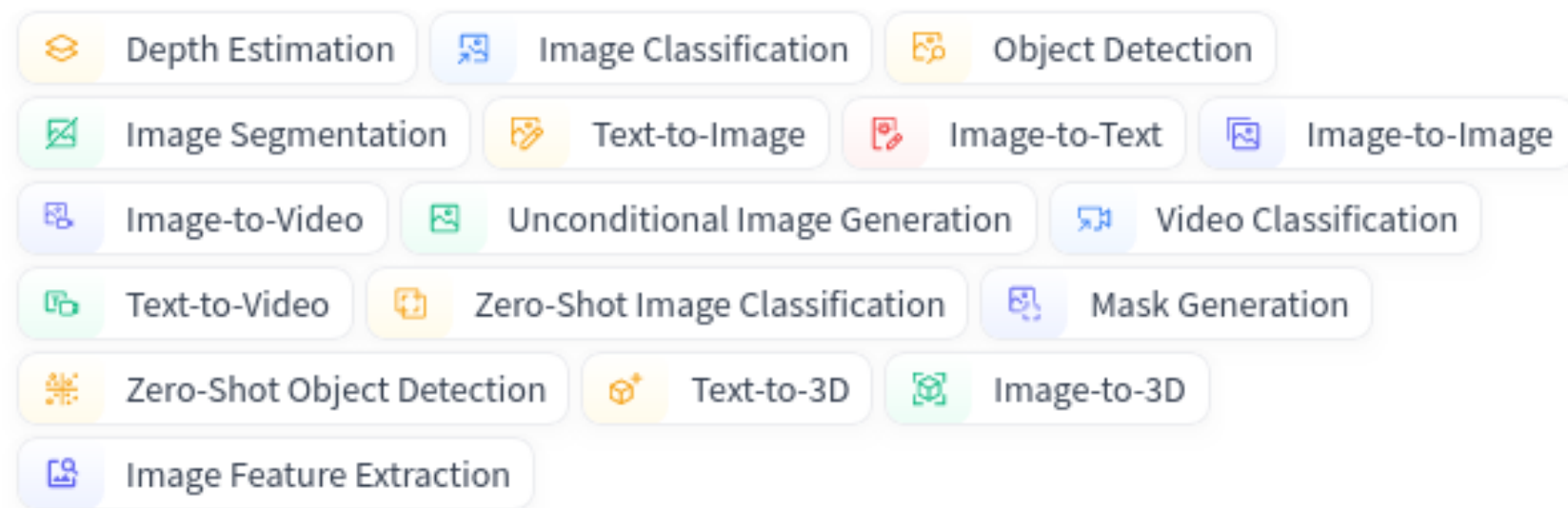
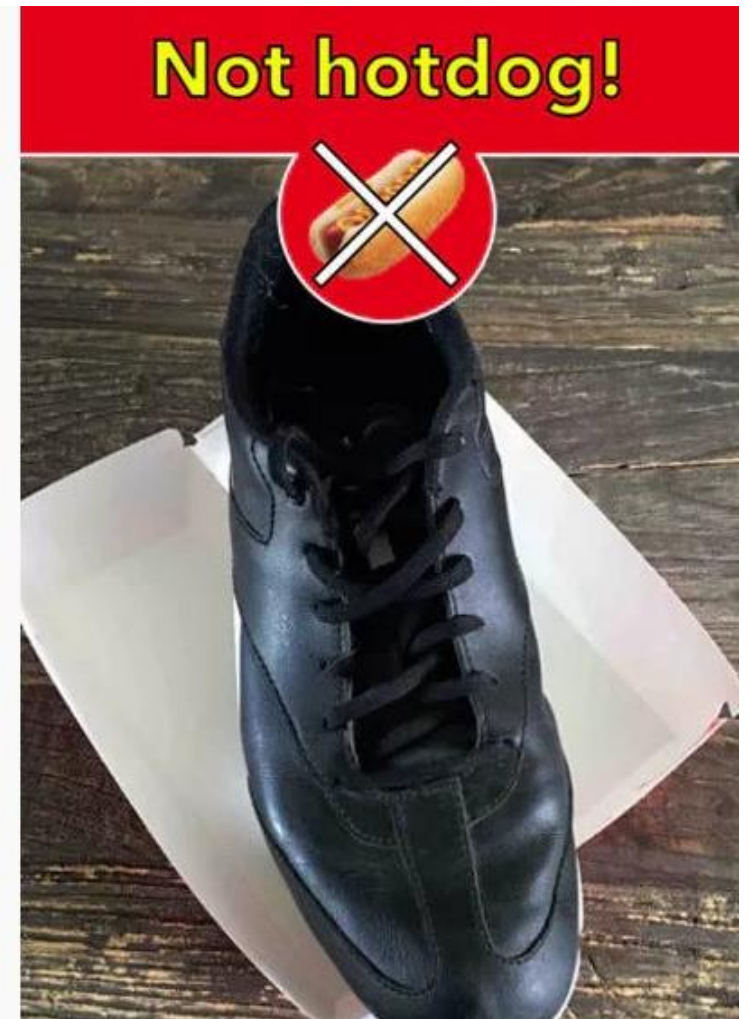


Image Classification

- Zu welcher Klasse gehört dieses Bild?



Object Detection

- Welche Klassen sind im Bild und wo?

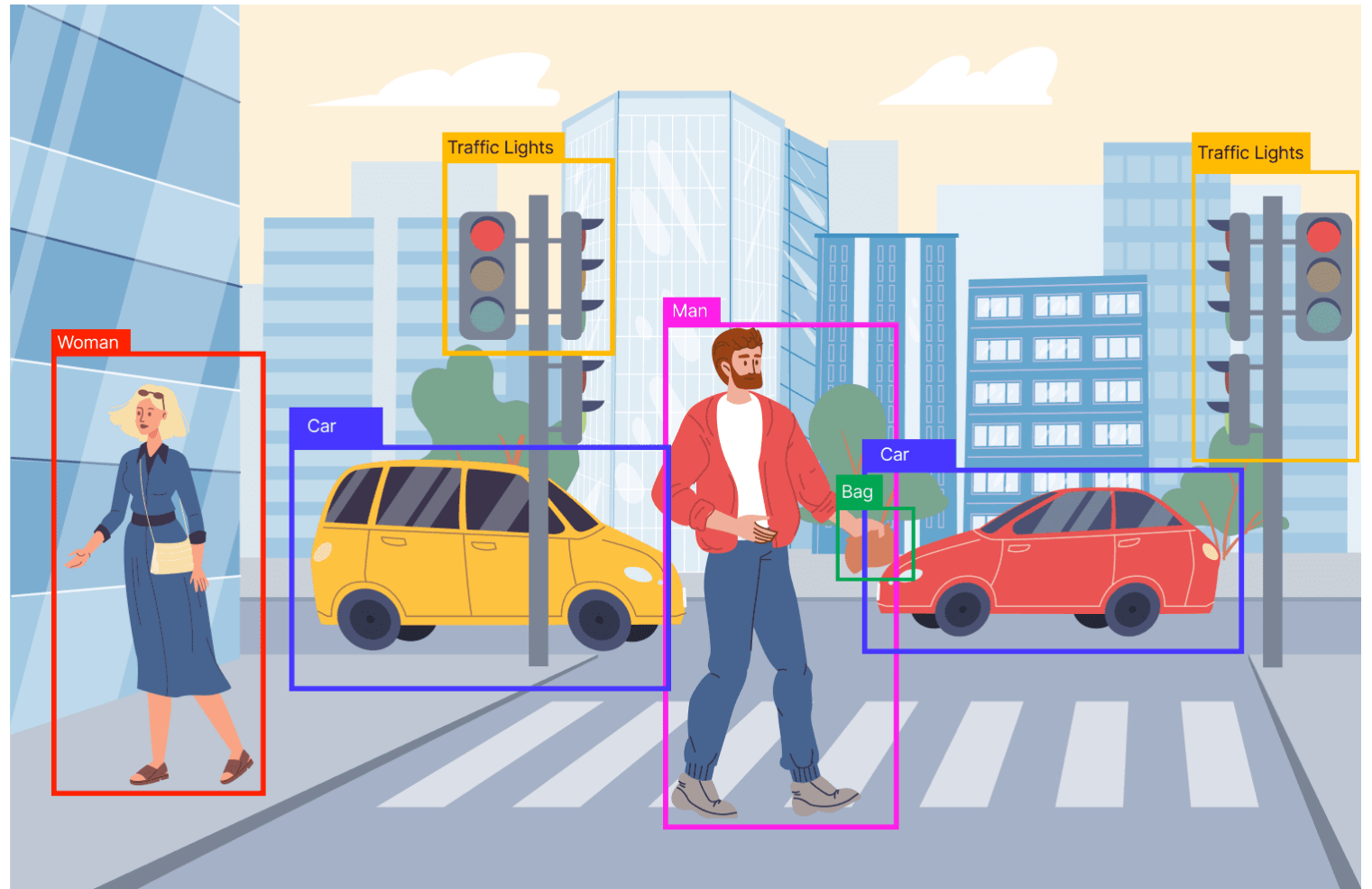
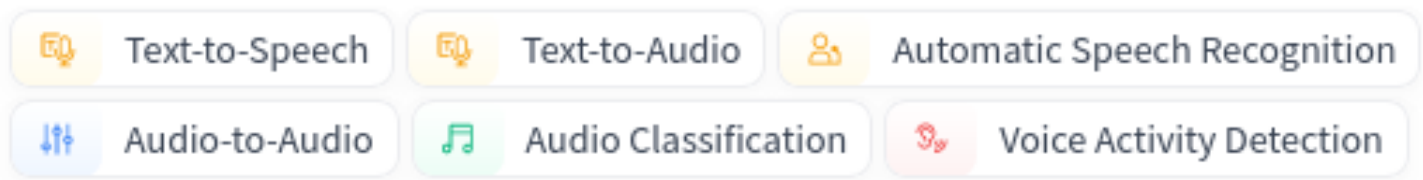


Image segmentation

- Welche Klassen sind im Bild und wo GENAU?

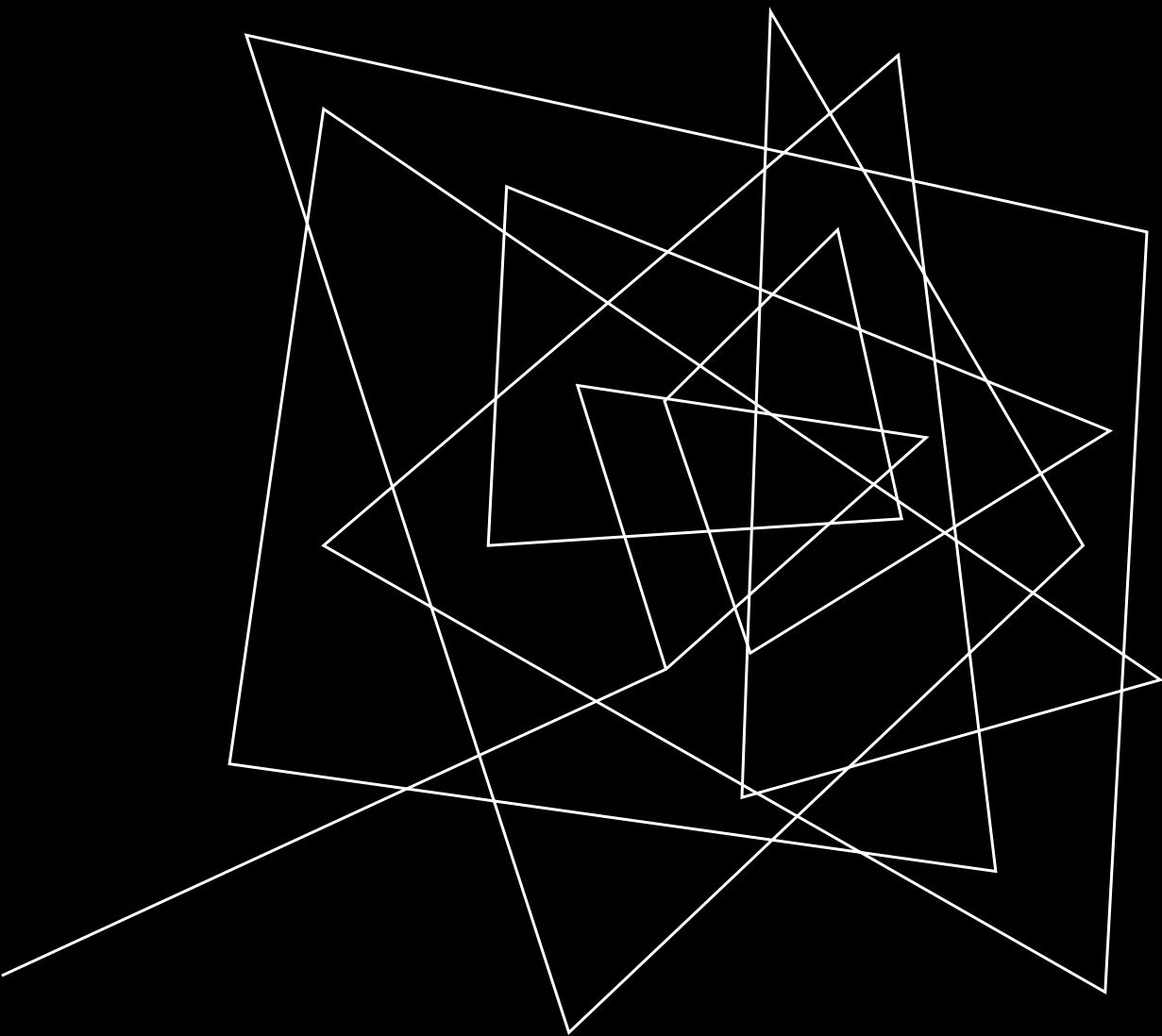




Automatic speech recognition (ASR)

- Wer sagt was?





NLP

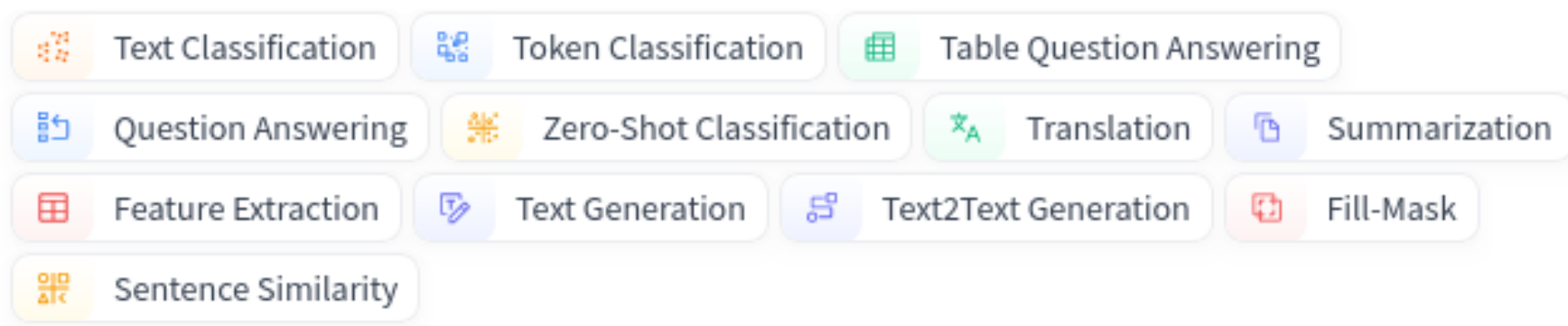
Natural Language Processing (NLP)

Einordnung

- Computer sollen natürliche Sprache verstehen
- Bereits Alan Turing hat sich 1950 damit befasst
- Symbolic, Statistical und Neural Networks

Abgrenzung

- Alles was irgendwie mit Sprache zu tun hat (auch OCR, Speech Recognition, Text-to-Speech)
- Heute eher alles was mit (geschriebenem) Text zu tun hat



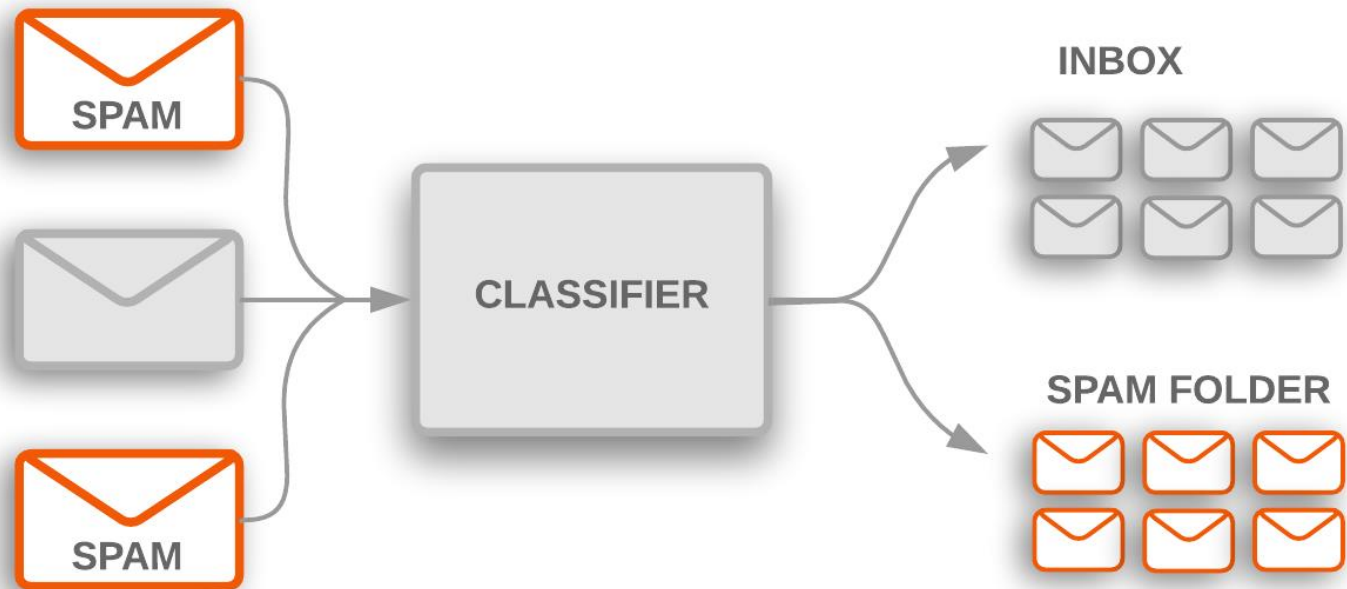
Translation

- Eines der ältesten ML Probleme
- "Urmutter" heutiger Modelle



Text Classification

- Zu welcher Klasse gehört dieser Text?



Named Entity Recognition (NER)

- Welche "Objekte" kommen in diesem Text vor?

The screenshot displays a Named Entity Recognition (NER) interface. At the top, a blue header bar contains a legend with six categories: Person (p, blue), Loc (l, yellow), Org (o, black), Event (e, green), Date (d, red), and Other (z, purple). Below the legend, a text snippet is shown with various entities highlighted in colored boxes, each followed by a small 'x' icon. The text is: "Barack Hussein Obama II (born August 4, 1961) is an American attorney and politician who served as the 44th President of the United States from January 20, 2009, to January 20, 2017. A member of the Democratic Party, he was the first African American to serve as president. He was previously a United States Senator from Illinois and a member of the Illinois State Senate." The entities are: "Barack Hussein Obama II" (Person, blue), "August 4, 1961" (Date, red), "American" (Other, purple), "the United States" (Loc, yellow), "January 20, 2009" (Date, red), "January 20, 2017" (Date, red), "Democratic Party" (Org, black), "African American" (Other, purple), "United States Senator" (Other, purple), "Illinois" (Loc, yellow), and "Illinois State Senate" (Org, black).

Definitionen

- **Token**
 - Die kleinste Einheit Text welche verarbeitet wird
 - Meistens ein Wort. Aber auch Satzzeichen, Zahlen oder Silben
- **Document**
 - Ein zusammenhängendes Textstück welches als Einheit betrachtet wird
 - z.b ein Satz, Absatz, Artikel oder ein ganzes Buch
- **Corpus**
 - Eine Sammlung von Dokumenten als Datengrundlage
 - z.b 100 Artikel oder 1 Million Tweets



Computer verstehen Token nicht. Was also tun?



Bag of ~~Cats~~ Words

- Einfaches aber effektives Verfahren
- Wie oft kommt ein Wort in einem Text vor
- Ergibt ein Vektor
 - (Der länge des Vokabulars)

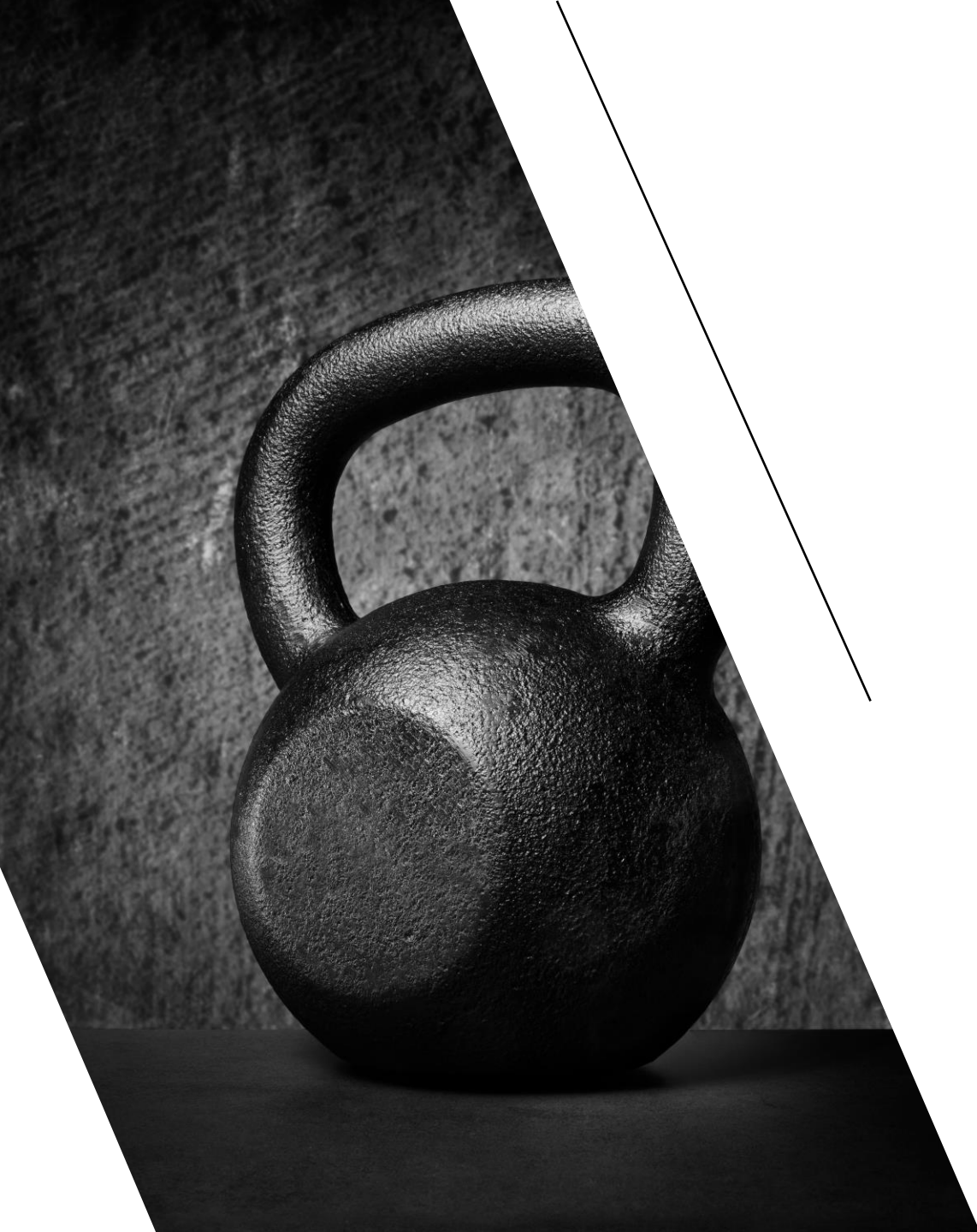
Dokument 1

Das Wetter ist heute
schön

Dokument 2

Heute ist das
Wetter schlecht

Wort	Dokument 1	Dokument 2
das	1	1
wetter	1	1
ist	1	1
heute	1	1
schön	1	0
schlecht	0	1



TF-IDF

- Term frequency – inverse document frequency
- Methode zur Gewichtung von Wörter
- Ziel ist es wichtige Wörter hoch zu gewichten und unwichtige niedrig

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

Beispiel

Dokument 1

Das Wetter ist heute
schön

Dokument 2

Heute ist das
Wetter schlecht

Wort	Dokument 1 (TF-IDF)	Dokument 2 (TF-IDF)
das	0.3	0.3
wetter	0.5	0.5
ist	0.3	0.3
heute	0.3	0.3
schön	0.9	0
schlecht	0	0.9

BoW Cosine Similarity: 0.7999999999999999

TF-IDF Cosine Similarity: 0.6694188517266485

Bag of Words & TF-IDF

Vorteile

- Gut verstanden und einfache implementierung
- Effizient bei kleinen bis mittleren Texten
- Gute Baseline zum Vergleich
- Sehr nützlich für Suche (Google's early days) und Recommendation Engines (Heute noch)
- Lightweight ML pipelines (kein Deep Learning)

Nachteile

- Verlust von Kontext und Semantik
 - Wörter mit verschiedenen Bedeutungen
- Hohe Dimensionalität (skaliert schlecht)
- Verschwenderisch
 - Sparse Matrix Problem

Limitationen BoW & TF-IDF

- **Reihenfolge**
 - "Dog bites Man" vs. "Man bites dog"
- **Negierungen**
 - "I like this movie" vs. "I don't like this movie"
- **Synonyme & Polysemie (Mehrdeutigkeit)**
 - "I love my automobile" vs. "I love my car"
 - "He opened a bank account" vs. "He sat by the river bank"
- **Kontext**
 - "Apple releases new iPhone" vs. "Apple pie recipe for all"

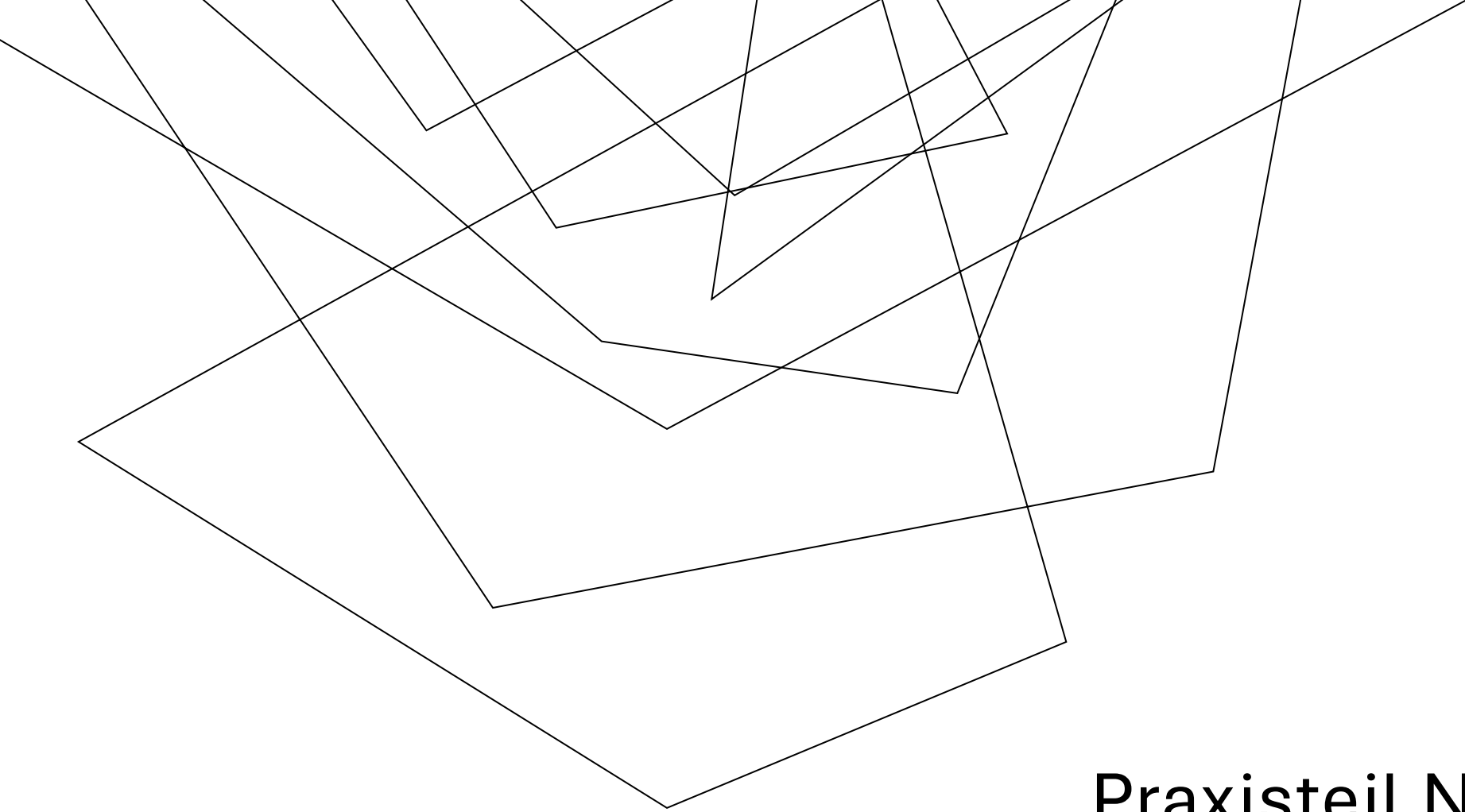


Preprocessing

- Wörter sind nicht nur Wörter
- Vocabulary verkleinern

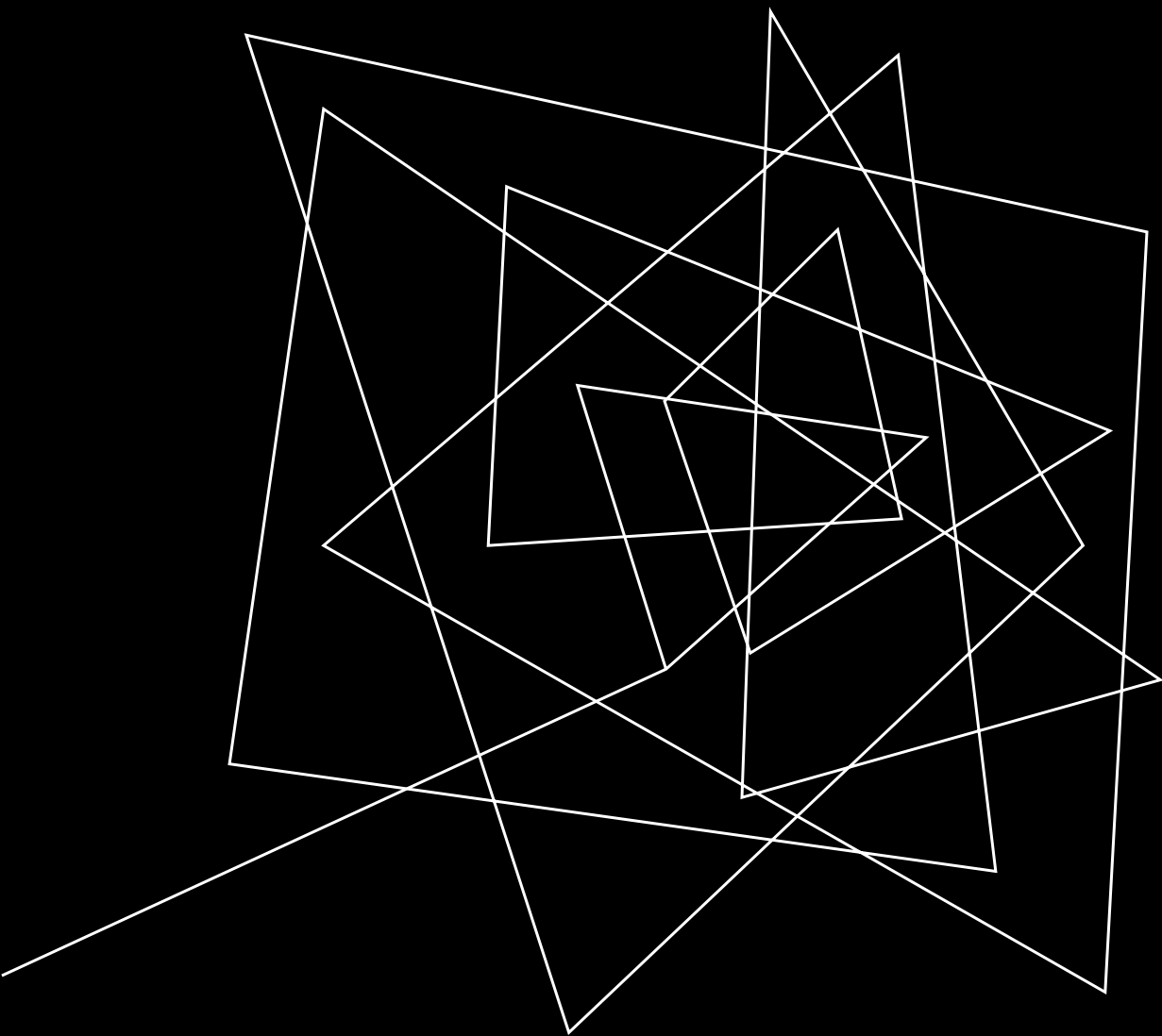
Preprocessing Methoden

- **Text Cleaning**
- **Stopword Removal**
 - [This, is, an, example, for, stop, word, removal]
 - [This, example, stop, word, removal]
- **Stemming**
 - Automate, automatic, automation -> automat
- **Lemmatization**
 - Car, cars, cars', car's -> car
 - Am, is, are -> be




Praxisteil NLP

<https://github.com/enki-farm/MLCourse-I>

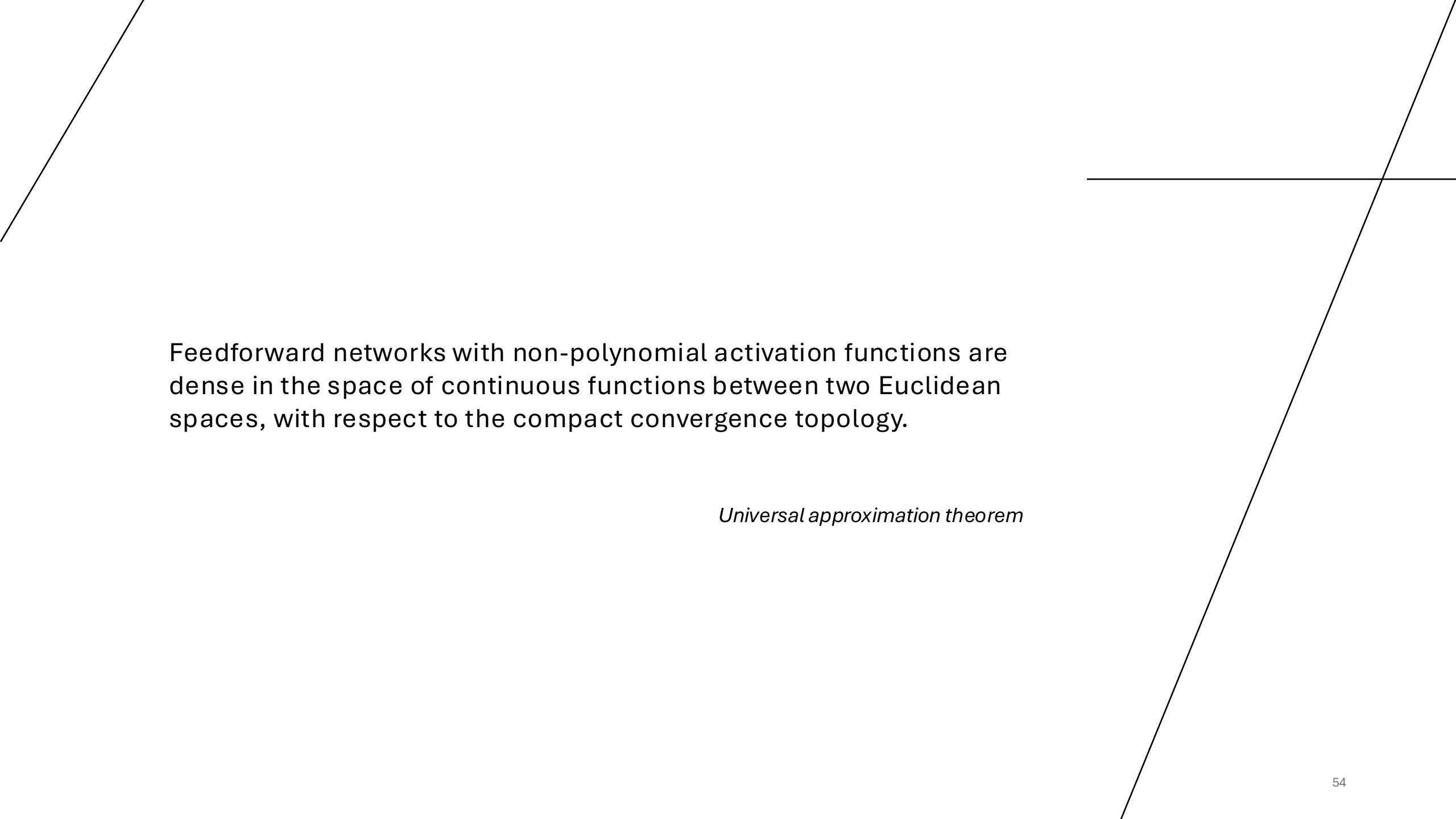


Embeddings



TF-IDF bewertet Wörter basierend auf ihrer Häufigkeit – aber es versteht nicht, was Wörter bedeuten.

Um diese Bedeutung zu verstehen muss sie gelernt werden.

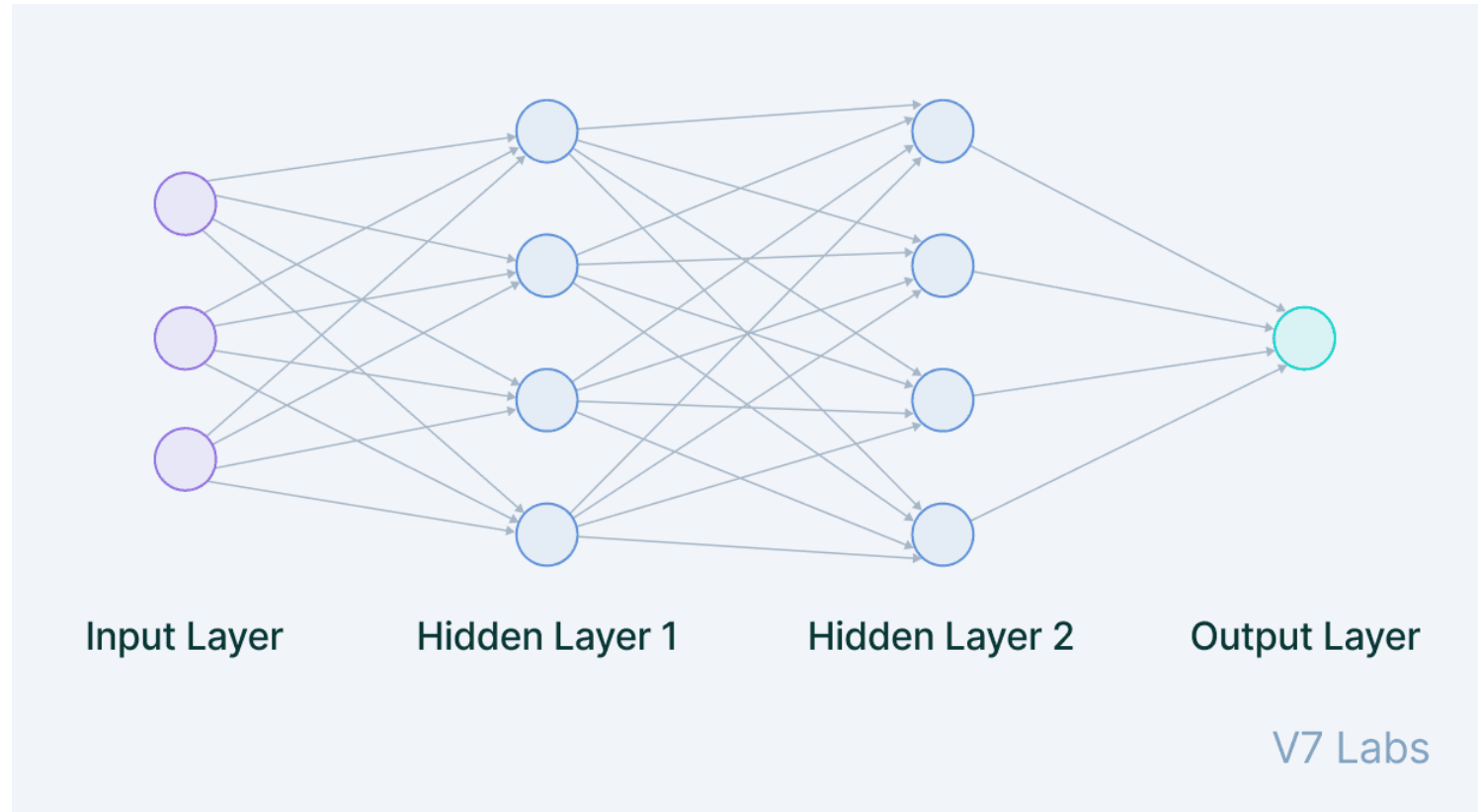


Feedforward networks with non-polynomial activation functions are dense in the space of continuous functions between two Euclidean spaces, with respect to the compact convergence topology.

Universal approximation theorem

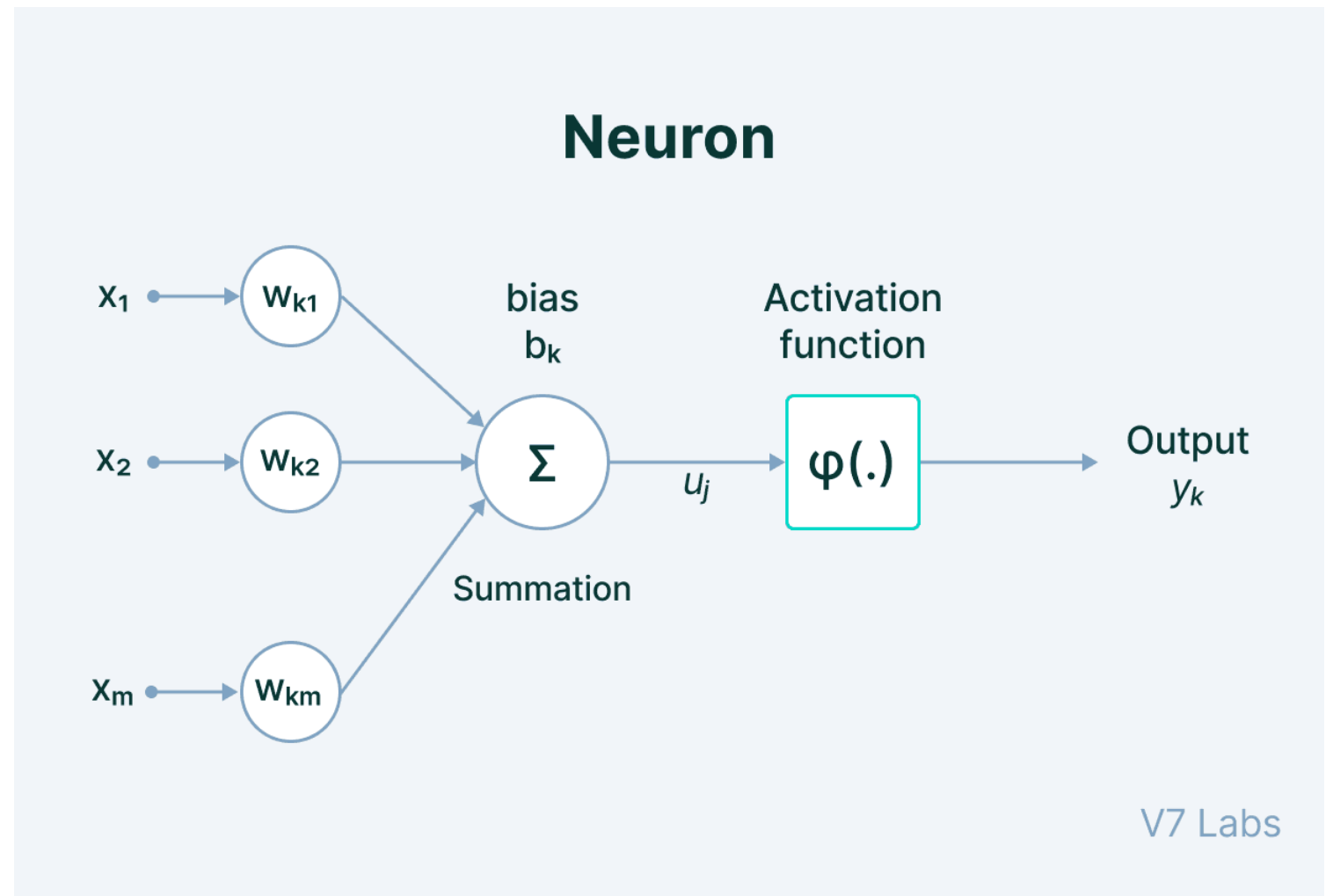
Neuronals Netzwerk

- Viele Neuronen in Schichten
- Breite: Anzahl Neuronen in Schicht
- Tiefe: Anzahl Schichten
- Anzahl Neuronen bestimmt komplexität der Funktion (grob vereinfacht)



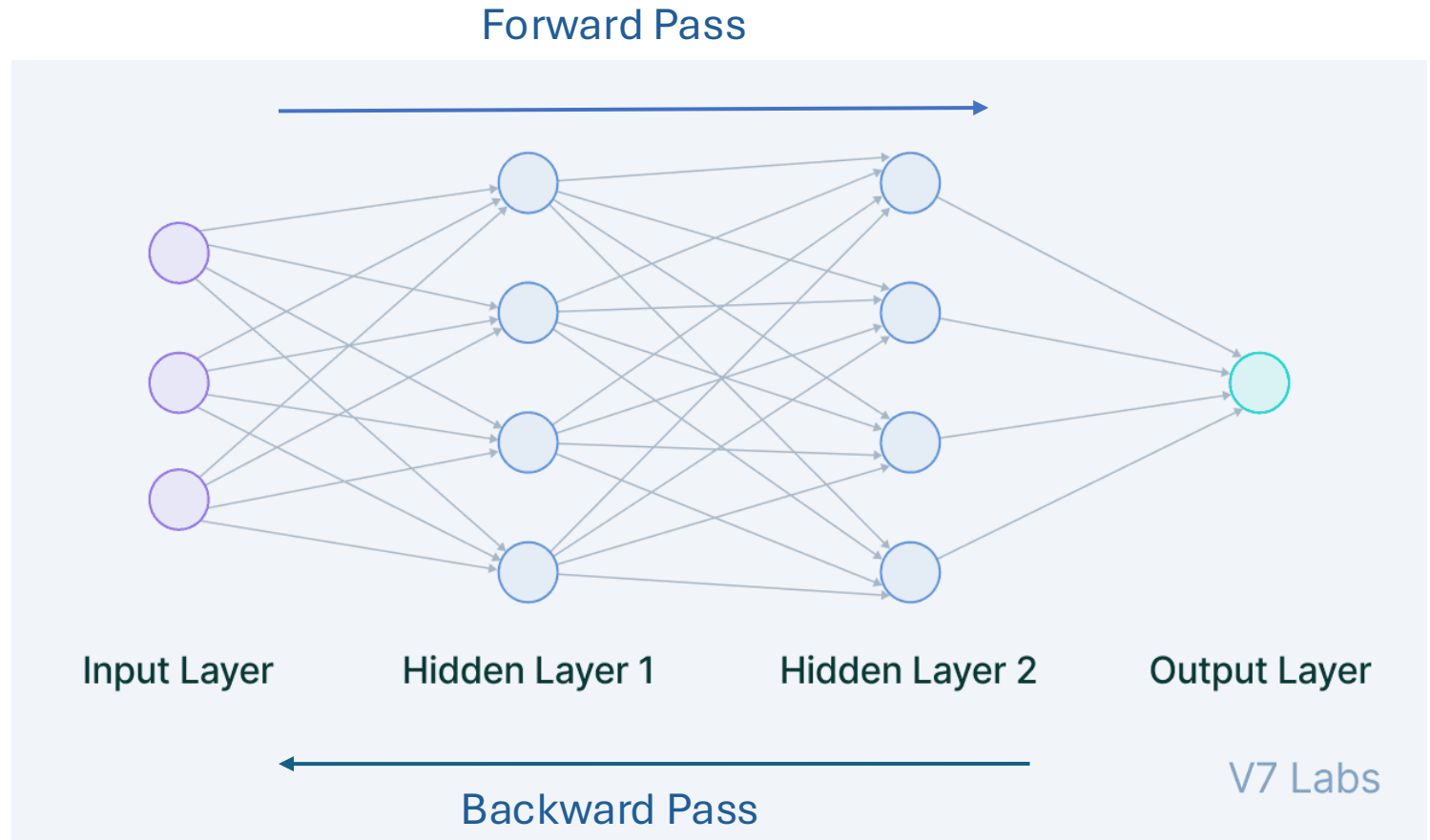
Neuron

- Inspiriert vom menschlichen Neuron
- Grundstein von Neuronalen Netzwerken

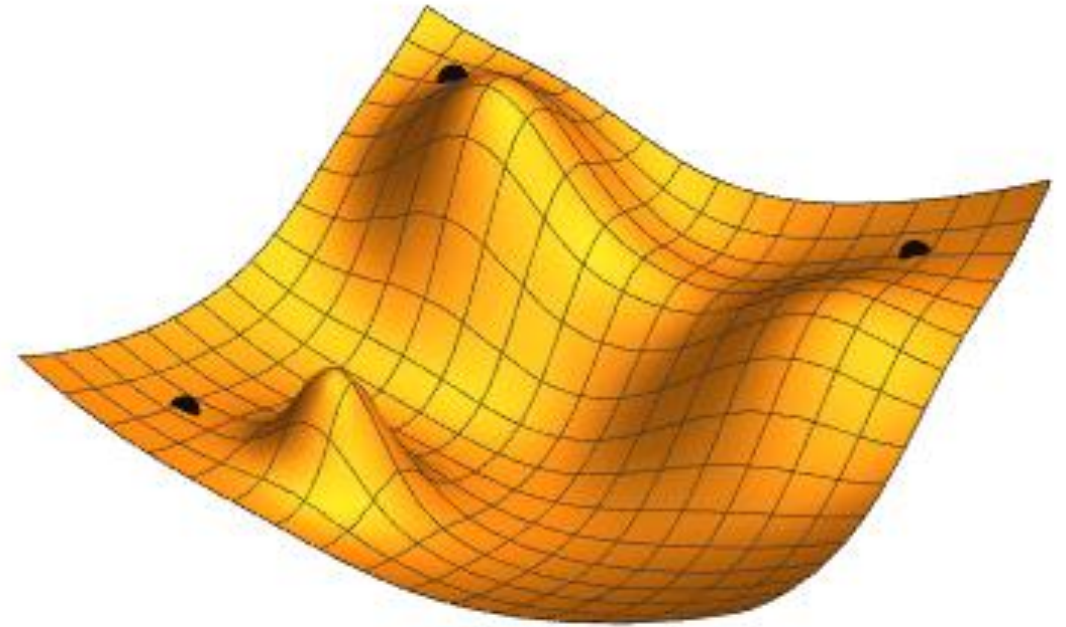
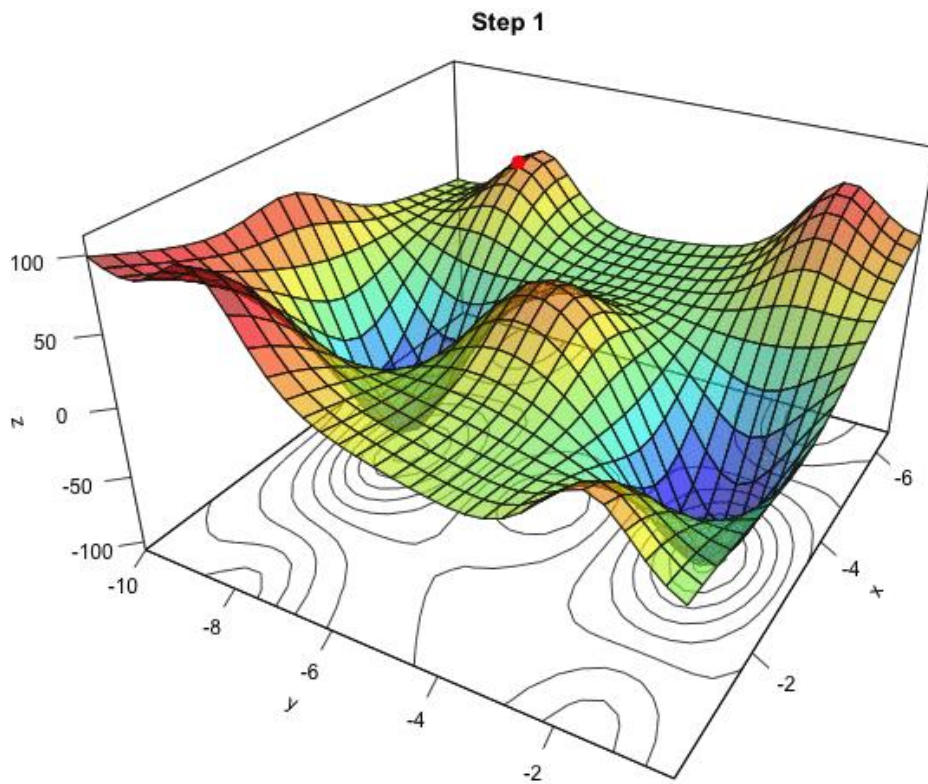


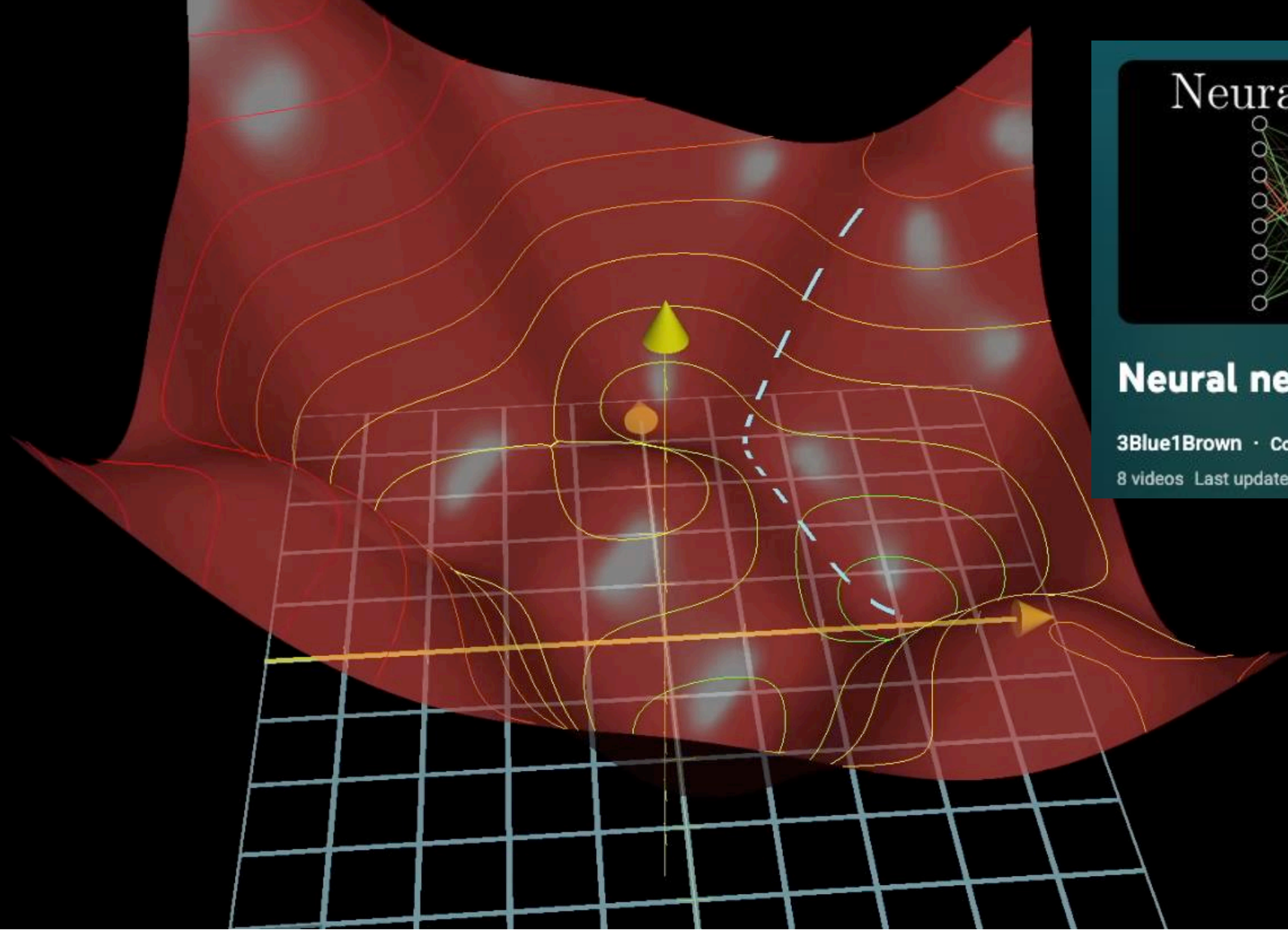
Backpropagation

- Forward Pass wird mit Ground Truth verglichen
- Fehler wird rückwärts durch Netzwerk propagiert
- Einfluss jedes Neurons wird berechnet
- Gewicht wird verändert um Fehler zu minimieren (Gradient Descent)

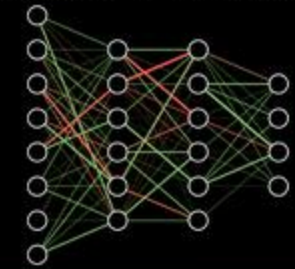


Gradient Descent





Neural Networks



Neural networks

3Blue1Brown · Course

8 videos Last updated on Feb 18, 2025

https://youtube.com/playlist?list=PLZHQObOWTQDNU6R1_67000Dx_ZCJB-3pi&si=d6wubzIPp5E7ZE

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



Epoch
000,519

Learning rate

0.03

Activation

Tanh

Regularization

None

Regularization rate

0

Problem type

Classification

DATA

Which dataset do you want to use?



Ratio of training to test data: 50%



Noise: 0



Batch size: 10



REGENERATE

FEATURES

Which properties do you want to feed in?



+ - 3 HIDDEN LAYERS



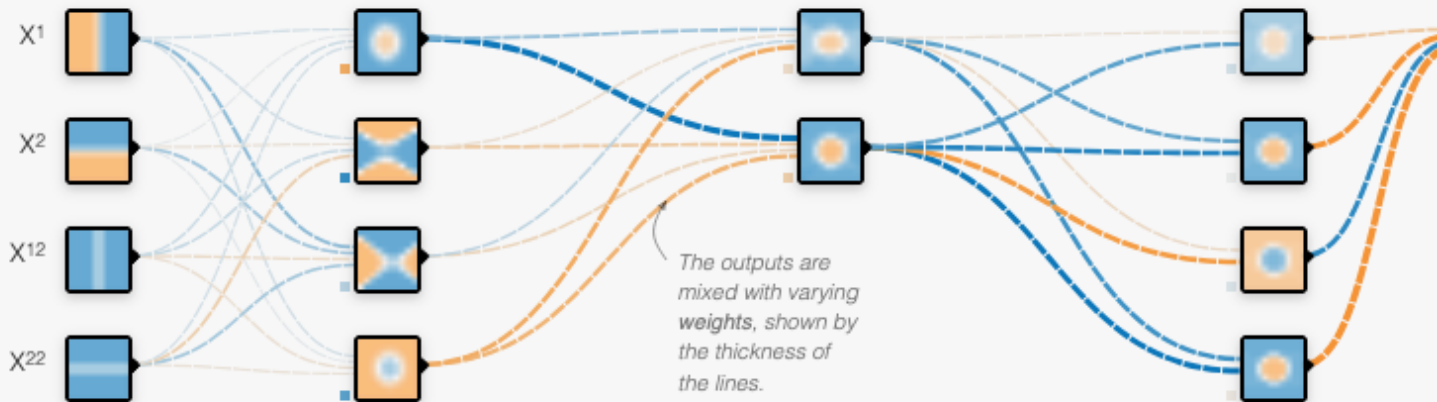
4 neurons



2 neurons



4 neurons



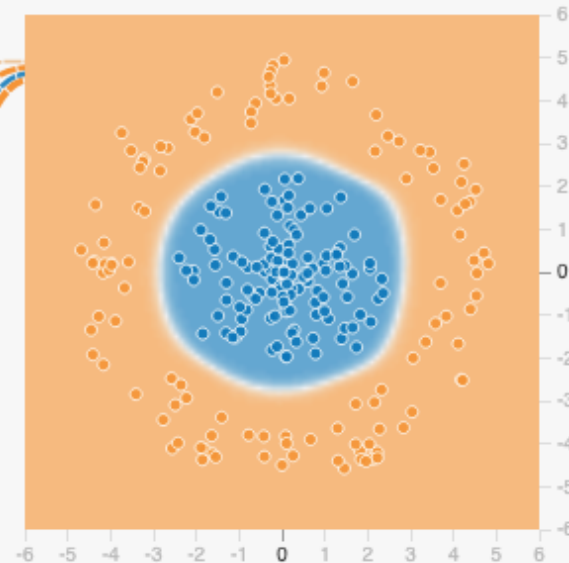
The outputs are mixed with varying weights, shown by the thickness of the lines.

This is the output from one neuron. Hover to see it larger.

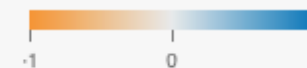
OUTPUT

Test loss 0.000

Training loss 0.000



Colors shows data, neuron and weight values.

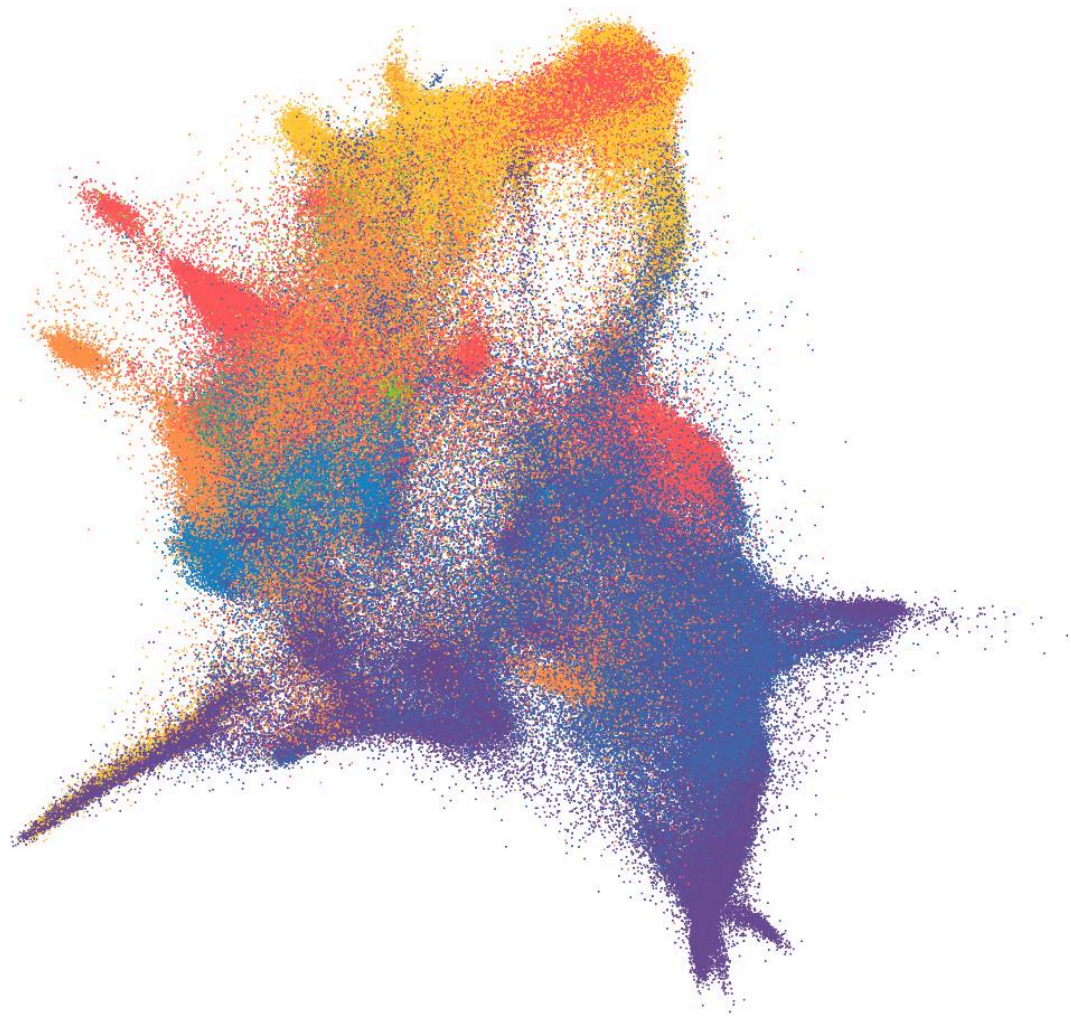


Embeddings

- Wörter können als Vektoren dargestellt werden
- Ähnliche Bedeutung -> Ähnliche Vektoren
- Embeddings erfassen semantische Beziehung zwischen Wörtern
- Die Vektoren werden gelernt

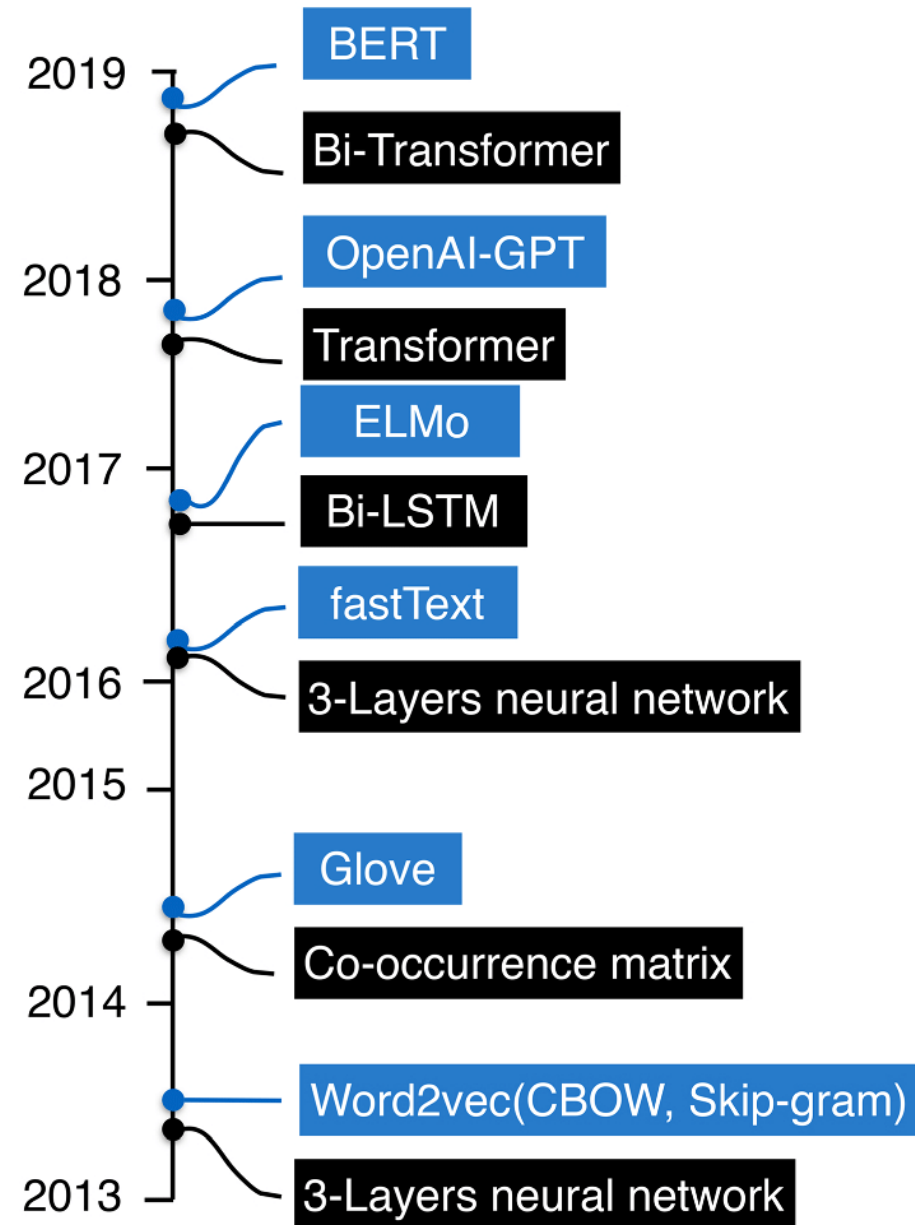
Wozu ist das gut?

- Semantic Search
- Grundlage für LLM
- Hilft Zusammenhänge von Daten zu verstehen
- Herrlich zu visualisieren



Geschichte der Embedding Modelle

- Seit 2005 werden Embedding Modelle gelernt (Dank Yoshua Bengio)
- Word2Vec von Google hat das Interesse massiv gesteigert
- ELMo legt den Grundstein für LLM



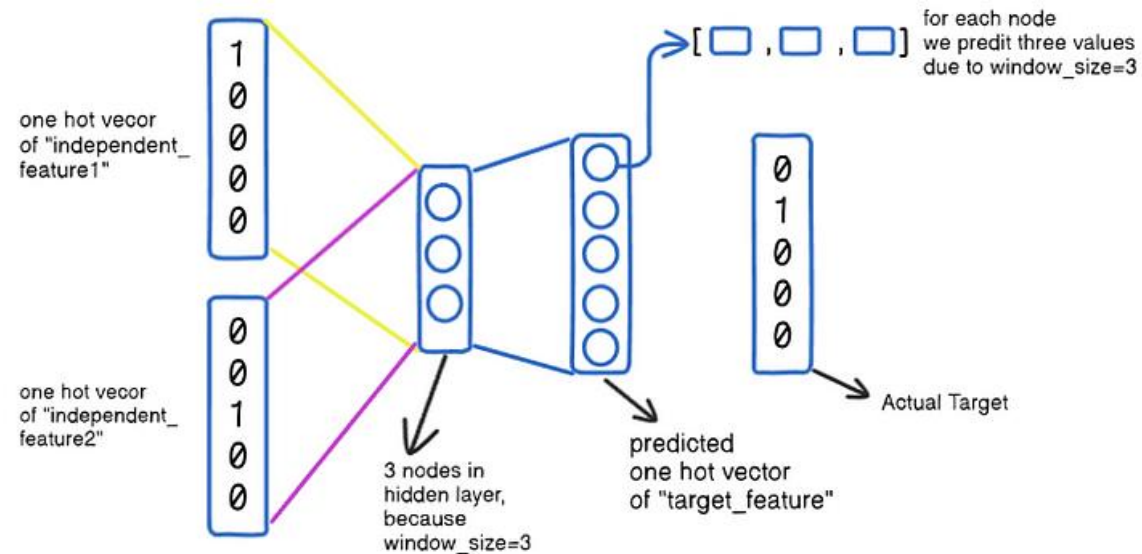


Word2Vec

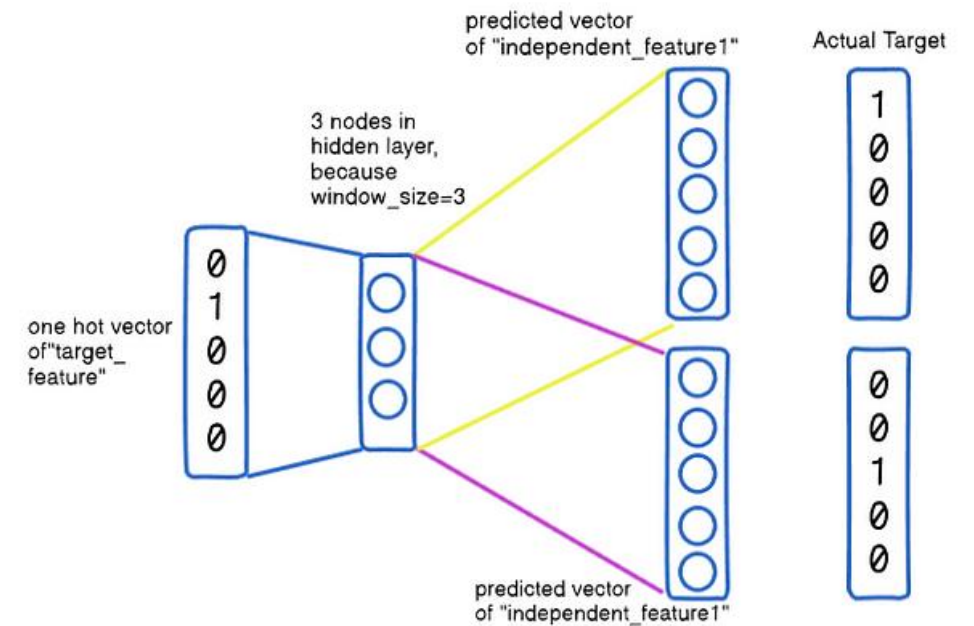
- Wann: 2013 - Veröffentlichung der Original-Paper
 - Efficient Estimation of Word Representations in Vector Space
 - Distributed Representations of Words and Phrases
- Entwickler: Tomas Mikolov bei Google Research
- Ziel: Wörter als dichte, kontinuierliche Vektoren repräsentieren, statt als One-Hot-Vektoren

Word2vec

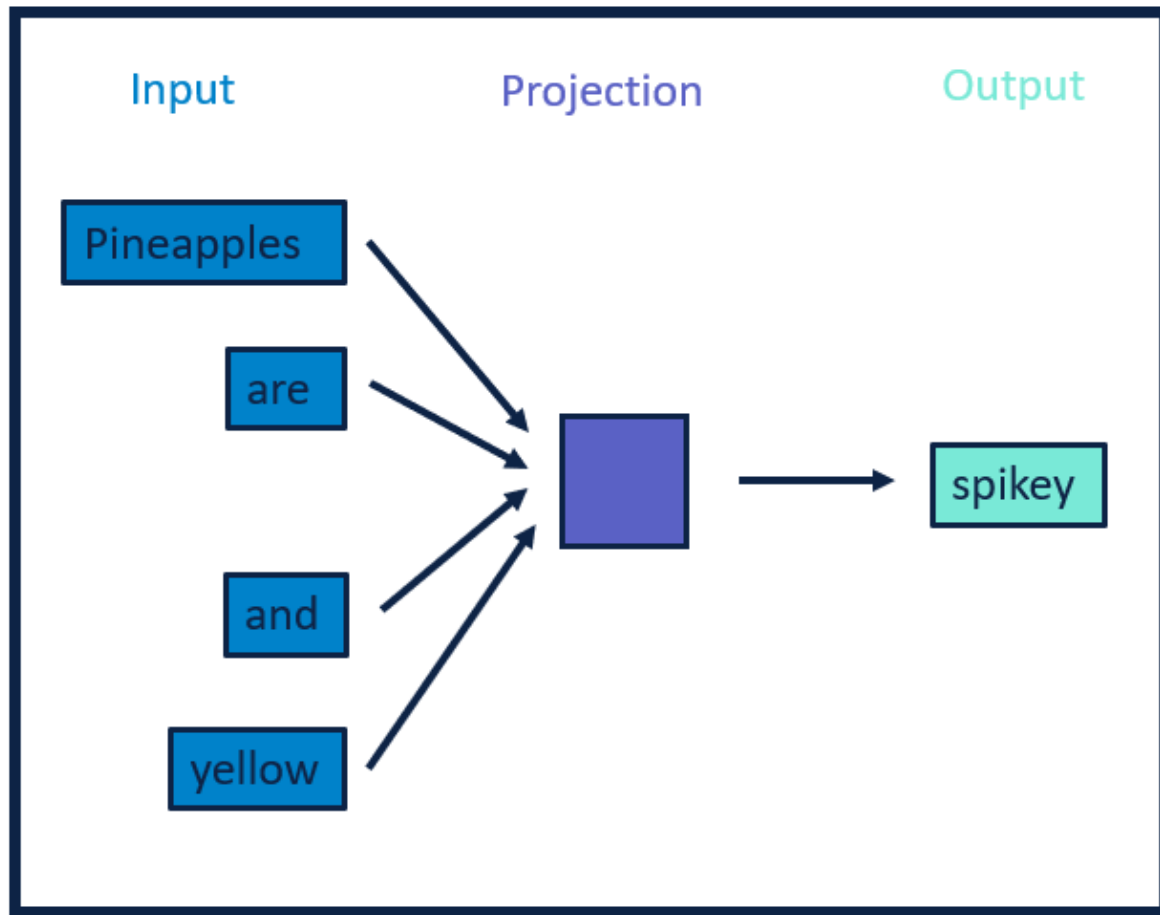
CBOW



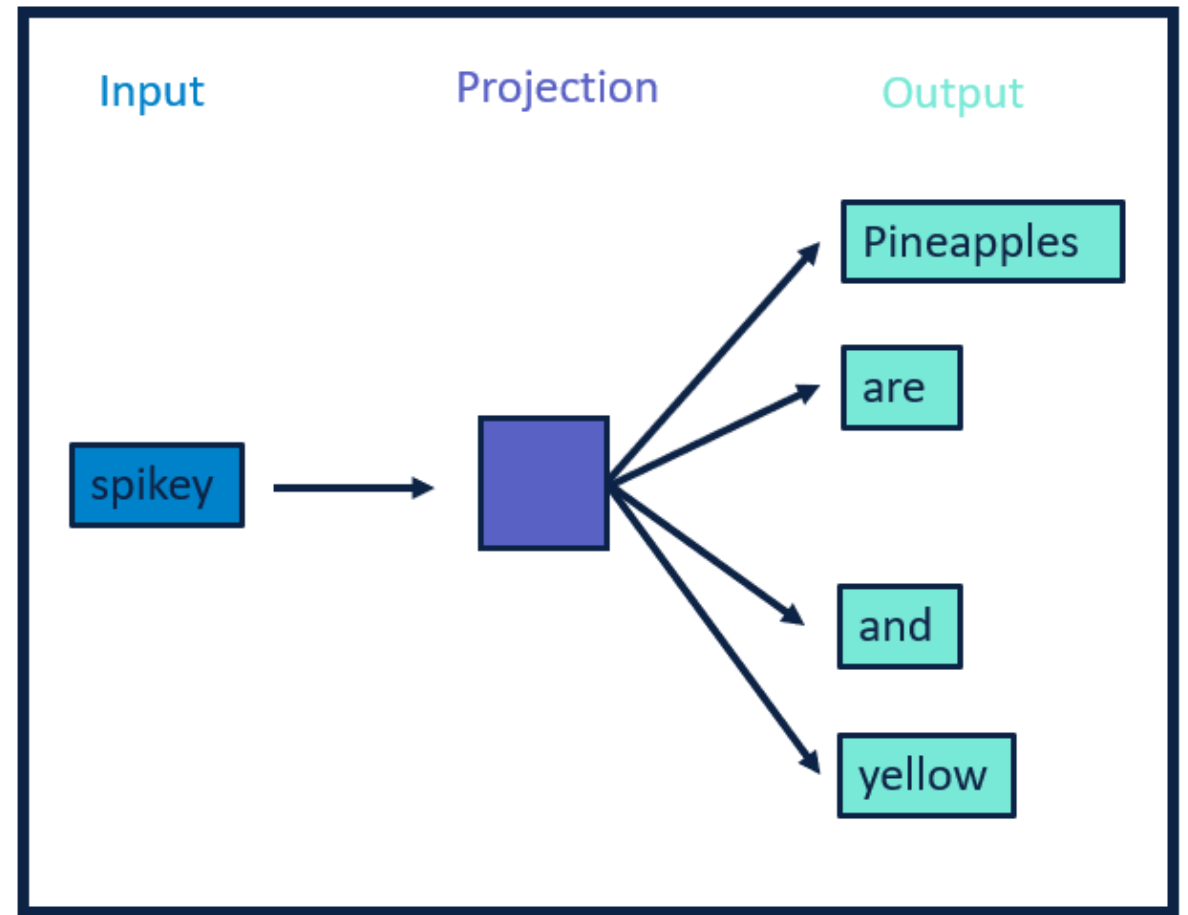
Skip-gram



- CBOW (Continuous Bag of Words) - Ein Wort aus dem Kontext vorhersagen
- Skip-Gram - Kontextwörter aus einem Wort vorhersagen

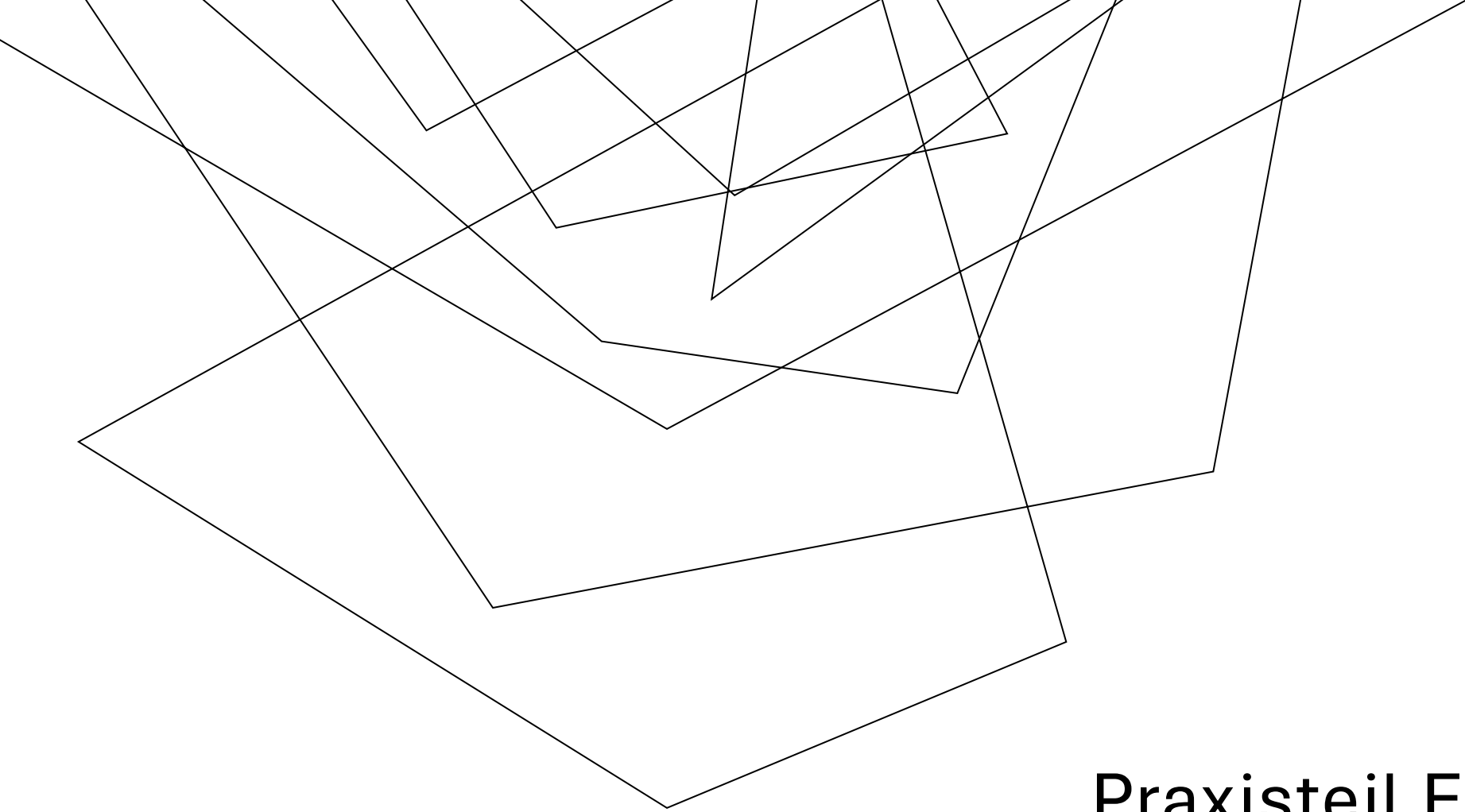


CBOW



Skip-gram

- CBOW (Continuous Bag of Words) - Ein Wort aus dem Kontext vorhersagen
- Skip-Gram - Kontextwörter aus einem Wort vorhersagen



Praxisteil Embeddings

<https://github.com/enki-farm/MLCourse-I>



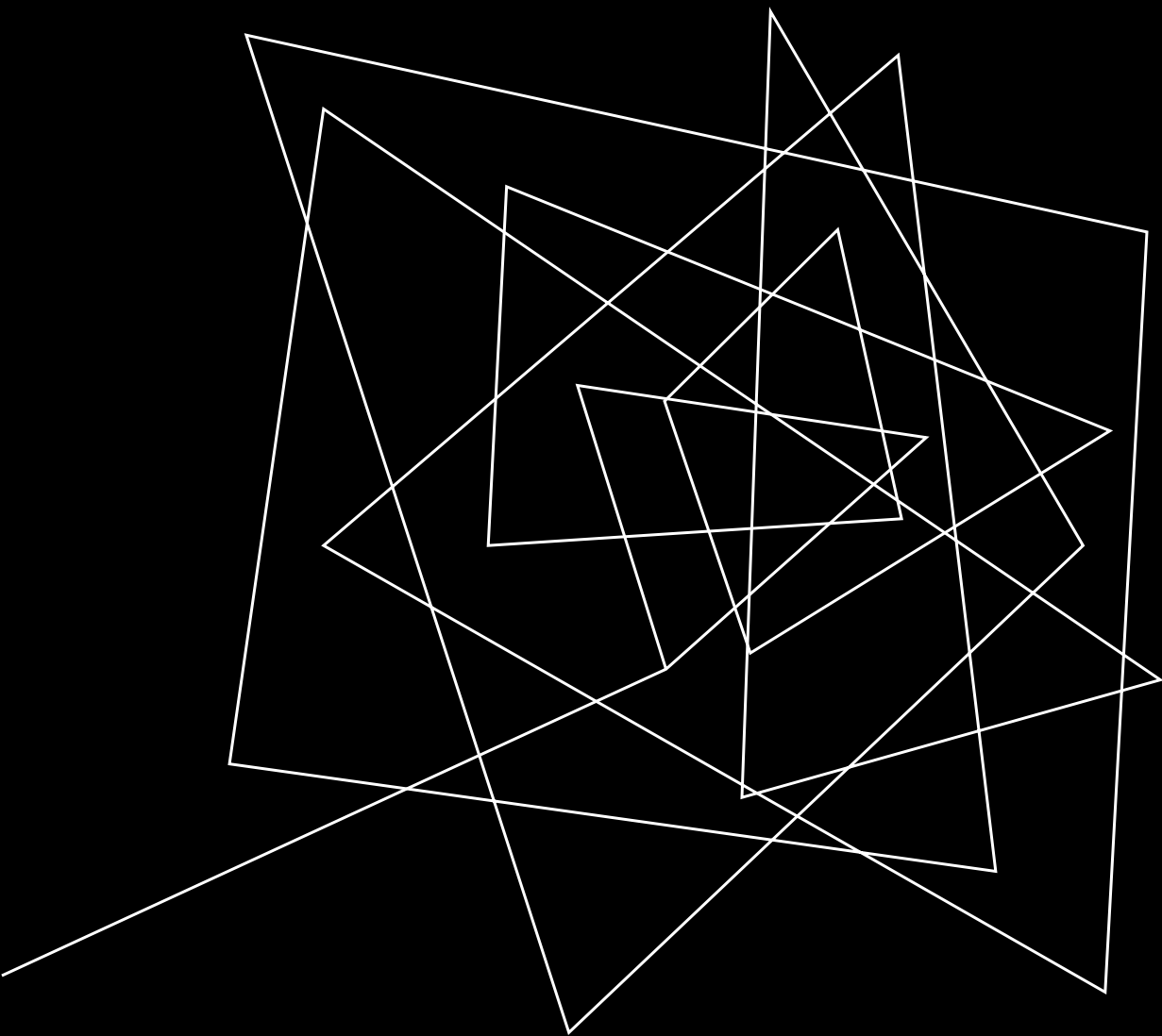
Limitationen Embeddings

- Out-of-Vocabulary
- Kontext fehlt (auch wenn es manchmal nicht so scheint)
- Interpretierbarkeit ist schwierig ($D > 4$)



Abschluss

- Embeddings sind eine Kern-Idee von RAG
- Nächstes Mal: RAG und Agents
- Fragen?



Ende Tag 1