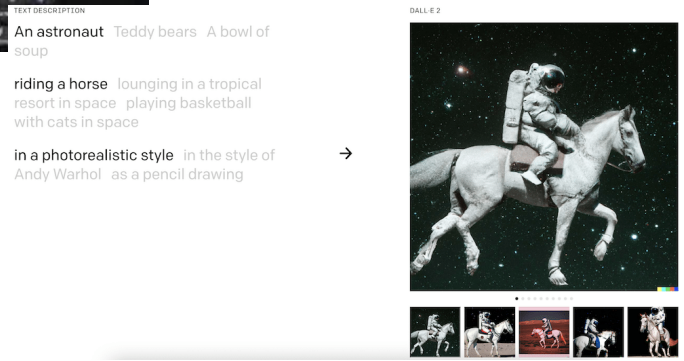
BAVARIAN CENTER FOR
SOFTWARE
INNOVATION

Zur Einführung

Prof. Dr. Alexander Pretschner
fortiss, TUM, bidt, CDTM

Ausgangspunkt

- Beeindruckende Beispiele für KI (Maschinenlernen) häufig ohne Bewertung der Güte



Was meine Frau und die Debatte um Maschinenlernen unterscheidet

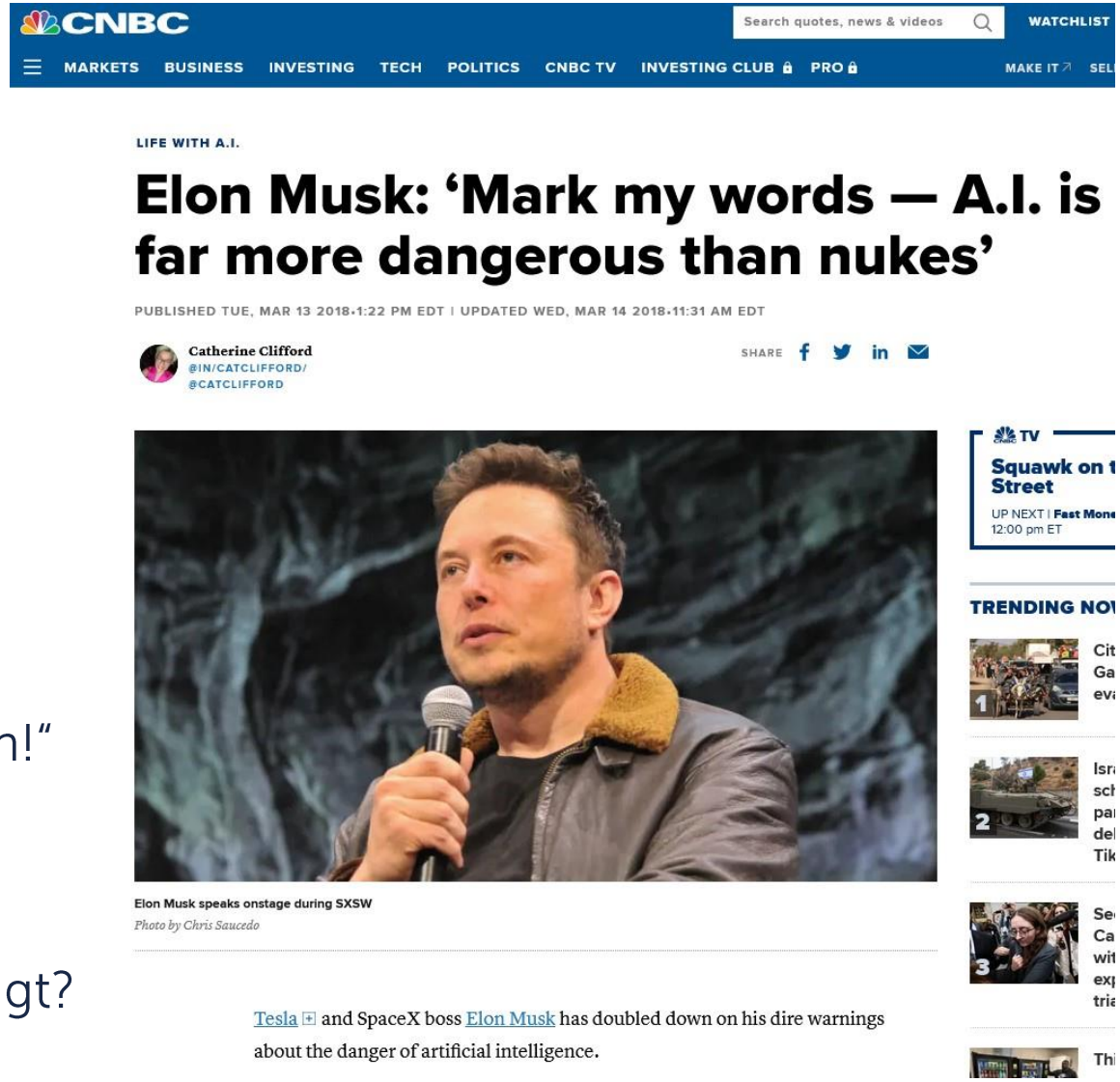
Diskursverschiebung

Von „kann KI das“ bzw. „wie gut kann KI das“

Zu

„KI kann alles! Deswegen müssen wir sie einhegen!“

… wo ist denn das Tool, das Börsenkurse vorhersagt?



Zur Einführung

Wie gut ist KI?

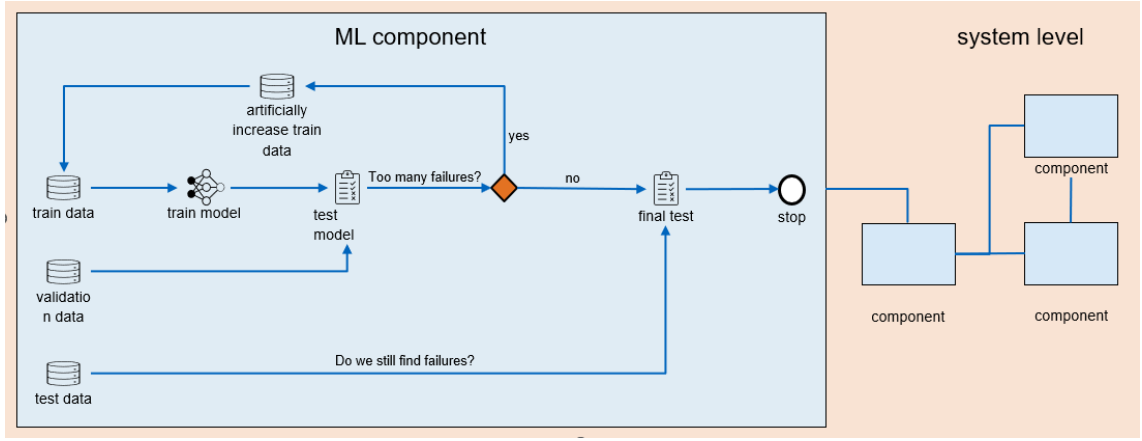
- Deep Fake Challenge 2020 mit $1.000.000 Preisgeld
- Erkennung von Deep Fakes mit Accuracy 65%:
  Zwei Drittel der Videos korrekt als „Deep Fake“ oder „kein Deep Fake“ erkannt
  Ein Drittel der Videos inkorrekt erkannt: Falschpositiv und Falschnegativ
- Ist KI „gut“?

Zwei Beispiele

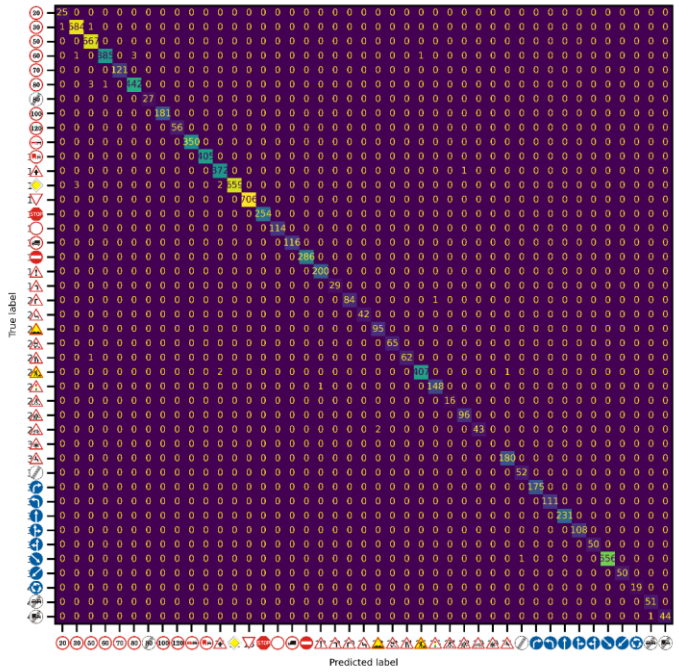
Maschinenlernen

- Maschinenlernen berechnet Vorhersagen, Klassen, beliebige Funktionen
- Immer dann attraktiv, wenn Problembereich nicht präzise fassbar
- Statt präzisen Schritt-für-Schritt-Berechnungen Ermittlung des Ergebnisses über Beispiele (Training) und Ähnlichkeiten zu Beispielen (Verwendung)

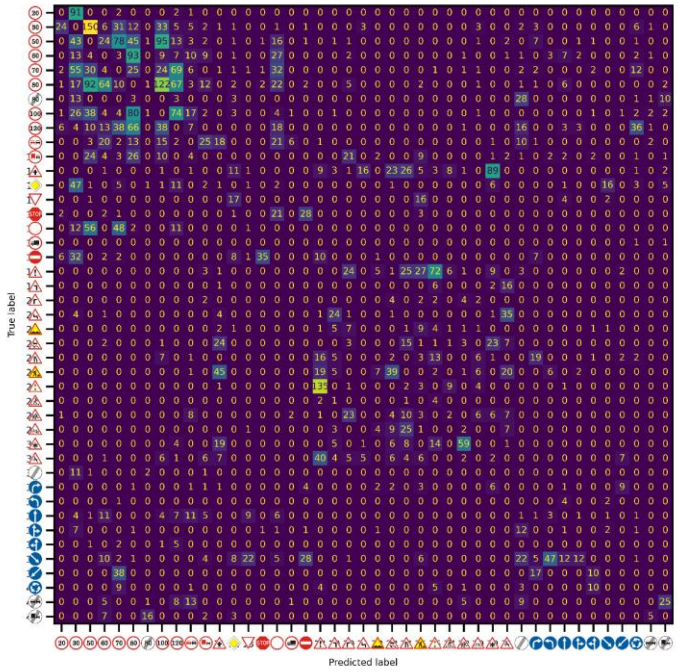
- Wenn Problembereich nicht präzise beschreibbar, wann ist das Ergebnis gut?
  - Variante der Konfusionsmatrix
  - Robustheit bzgl. out-of-distribution-Daten
  - Daten repräsentativ
  - Daten ohne Bias



Testdaten?



Original Test Dataset



Corner Case Images

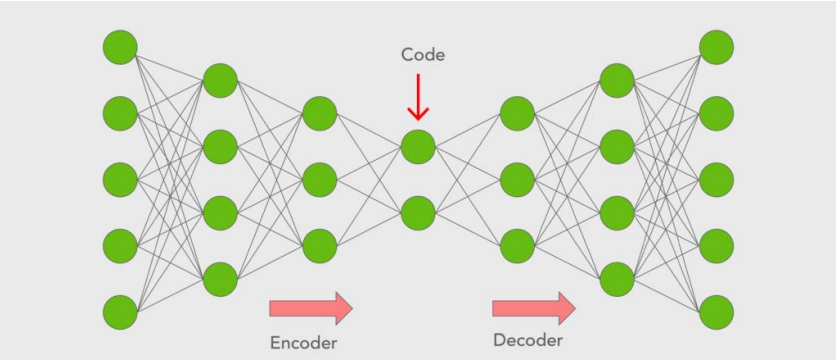
Gut genug?

- ChatGPT als Tutor für Schüler
- 20% Halluzination: gut genug?

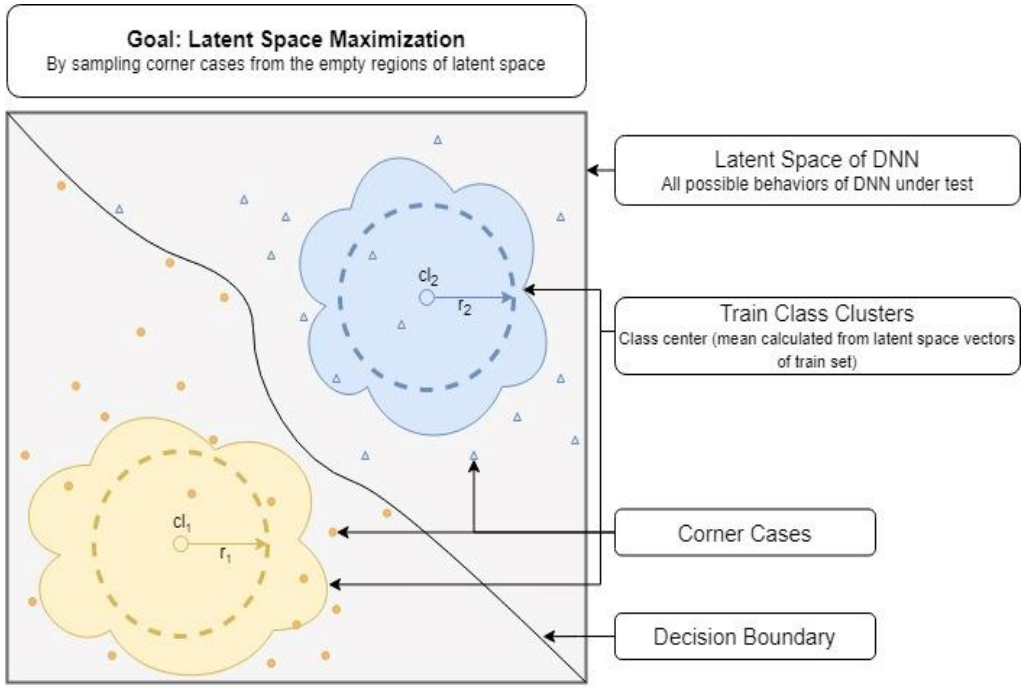
- KI für Diagnostik in der Medizin
- Gut genug, wenn so gut wie Spitze/Durchschnitt/unteres Ende?

- Weniger Tote durch Teslas als durch menschliche Fahrer?

Warum schwierig?



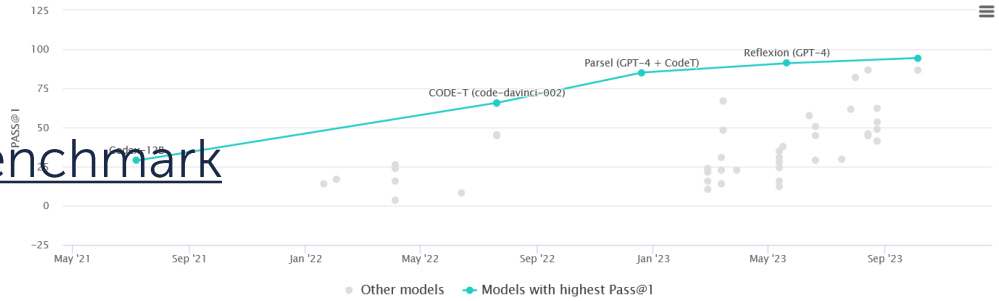
- Beispielbasiert
- Daten vorhanden, gelabeled, korrekt, vollständig, nicht kompromittiert
- Daten repräsentativ und ohne Bias
- Concept Drift
- Erklärbarkeit
- Gütekriterium/-maß

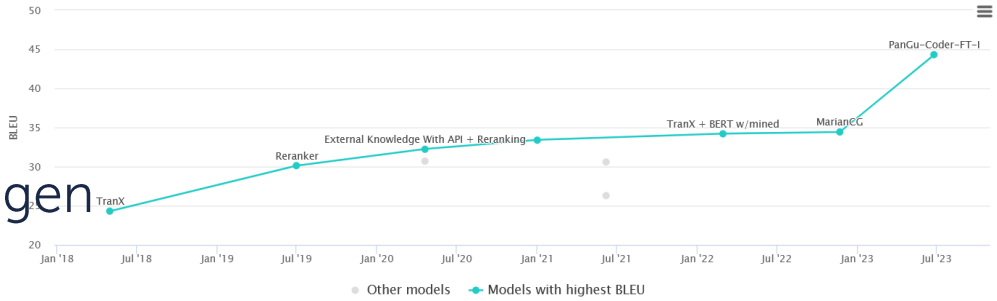


Wie und was messen?

Beispiel: Benchmarks für LLMs und Codeerzeugung

https://paperswithcode.com/task/code-generation
https://tech.ebu.ch/publications/overview-of-some-llm-benchmark
https://github.com/LudwigStumpp/llm-leaderboard
https://arxiv.org/pdf/2202.13169.pdf
https://arxiv.org/abs/2107.03374





Korrektheit? Mit Tests
Robustheit: ähnliche Prompts sollten ähnlichen Code erzeugen
Erklärungen
Determinismus
Sicherheit

So What?

Enttäuschung als Funktion der Erwartung

- Hype Cycle AI reloaded: generative KI
- Kurz vorher Abflachen beobachtbar: Herausforderungen in der Praxis!
  - Daten: Verfügbarkeit von Labels, Korrektheit, Vollständigkeit, Concept Drift
  - Güte der Berechnung

- LLM-Hype löst diese Probleme nicht!
  - Manchmal Qualität egal bzw. „you know it when you see it“
  - Oft nicht: **Dann vorher darüber nachdenken, um Enttäuschungen zu vermeiden!**
  - Und zwar bitte optimistisch!
  - Technische und methodische Ansätze u.a. bei fortiss

Vielen Dank!