DOXNET – Intelligente Textverarbeitung mit Künstlicher Intelligenz

Eine Reise durch die neuesten Ergebnisse und Durchbrüche aus der Forschung

DOXNET® THE DOCUMENT X-PERTS NETWORK

ZUR PERSON …

Arthur Brack

Master of Engineering, Diplom-Informatiker (FH)

Geschäftsführer bei SET GmbH

Persönliche Angaben

- 38 Jahre
- Über 20 Jahre Erfahrung in der Software-Entwicklung, davon 14 Jahre in der SET GmbH
- Honorardozent an der FHDW Hannover zu „Data Analytics“
- Seit 2018: Forschung im Bereich Natural Language Processing und Deep Learning an der Leibniz Universität Hannover / TIB

Ziele des Vortrags

- Grundverständnis zum Thema Machine Learning vermitteln
- Überblick über ausgewählte relevante Forschungsergebnisse und Trends in der KI-Forschung zu Textverarbeitung geben
- Impulse und Inspiration zur Lösung von potenziellen Problemen in der Praxis mitgeben
- Einblick in einige Herausforderungen beim Einsatz von KI in der Praxis und mögliche Lösungsansätze vermitteln

Agenda

1. Motivation
2. Einführung in Machine Learning
3. Aktuelle Ergebnisse und Trends aus der Forschung
4. Herausforderungen für die Praxis und Lösungsansätze
5. Fazit

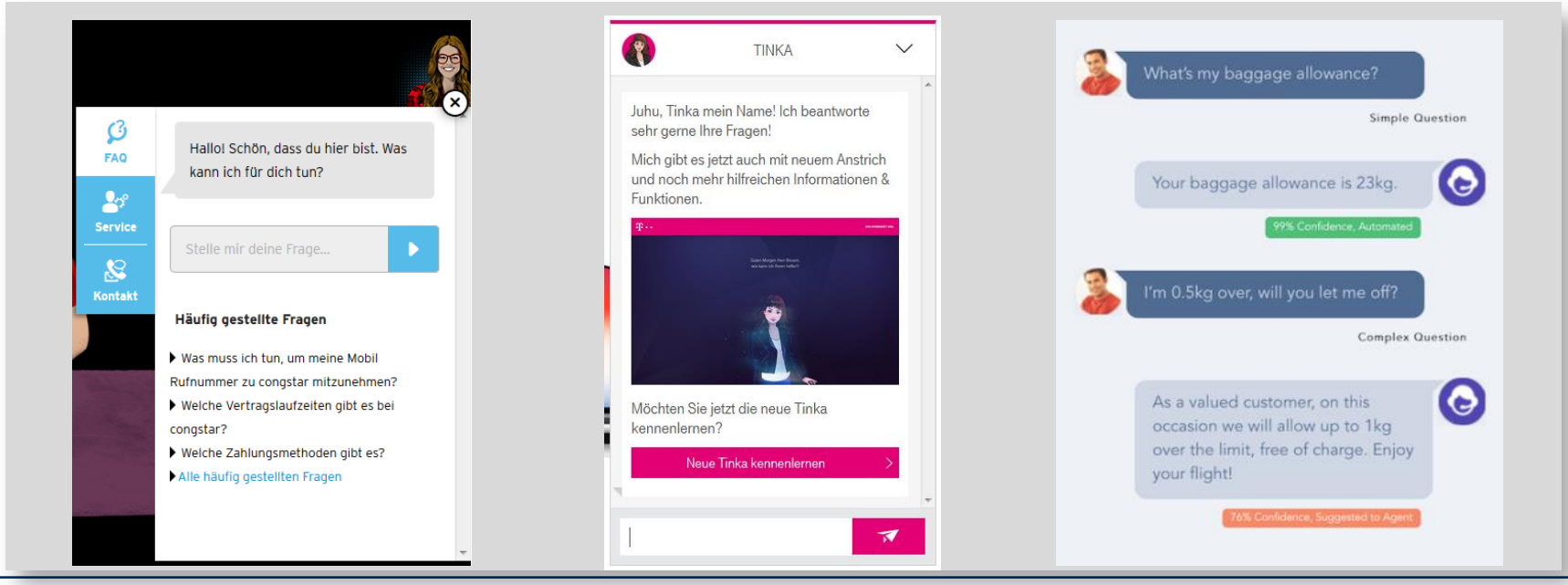
Motivation

Zunehmend elektronische Kommunikation

- Ca. 90% der Zeit auf dem Smartphone wird mit Messaging Apps verbracht
- Über 41 Millionen Nachrichten pro Minute, über. 2.9 Mrd. Anwender (Statista, 2020)

Elektronische Kommunikation zwischen Kunden und Unternehmen wird immer beliebter

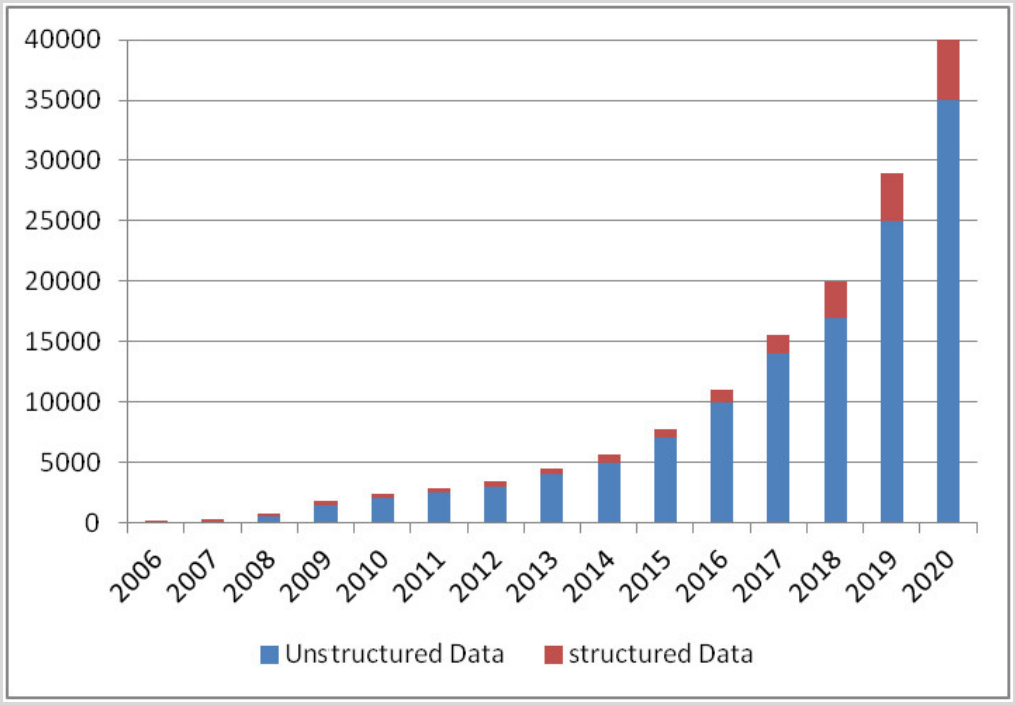
- Email, Chatbots, Webseiten, Apps, Social Media, …



Motivation

Elektronische Kommunikation ist ein Treiber für immer mehr unstrukturierte Daten in Unternehmen

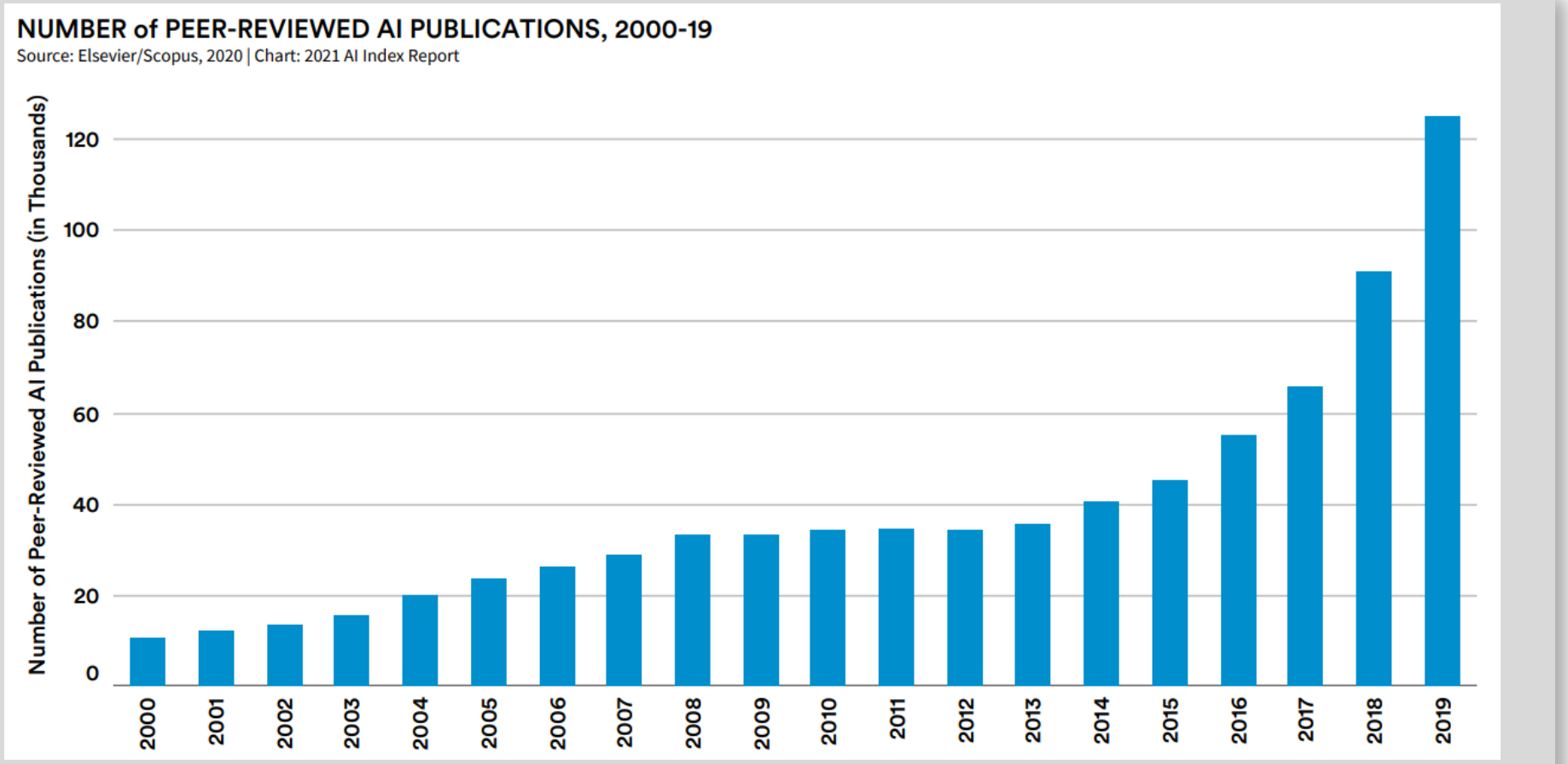
- Unstrukturierte Daten (Texte, Nachrichten, Bilder) erschweren Datenanalyse und Automatisierung von Prozessen
- Kunden erwarten schnelle Bearbeitung ihrer Anfragen
- KI-Methoden haben ein hohes Potenzial unstrukturierte Daten zu strukturieren



https://www.linkedin.com/pulse/integrate-digital-analytics-your-business-romit-atta/

Motivation

Die Forschungsausgaben im KI-Bereich steigen rasant



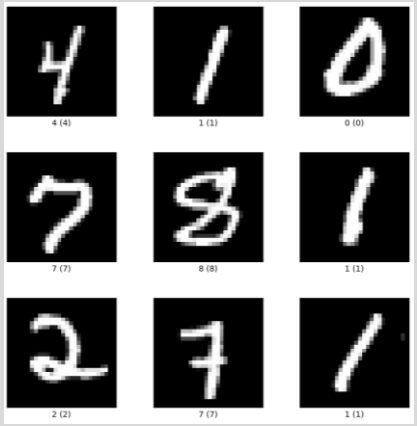
https://aiindex.stanford.edu/wp-content/uploads/2021/03/2021-AI-Index-Report_Master.pdf

Motivation

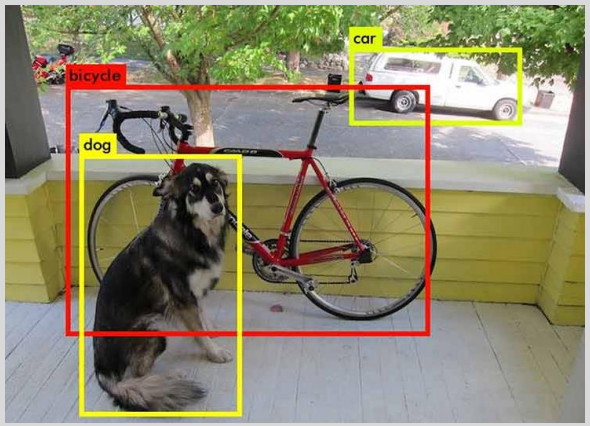
KI-Benchmarks sind ein wesentlicher Baustein, um den Fortschritt in der KI-Forschung zu messen

- Messen die Leistung von KI-Verfahren für bestimmte Tasks (z. B. OCR-Erkennung)
- Bestehen aus einem Testdatensatz und optional einem Trainingsdatensatz
- Testdatensatz repräsentiert die "ground truth": zu erwartende Ausgaben zu konkreten Eingaben
- Testdatensatz wird durch Menschen annotiert (z. B. Experten oder Mehrheitsentscheid)
- "Human Performance" repräsentiert die Leistung von Menschen auf dem entsprechenden Task

MNIST: OCR-Erkennung



ImageNet: Objekterkennung



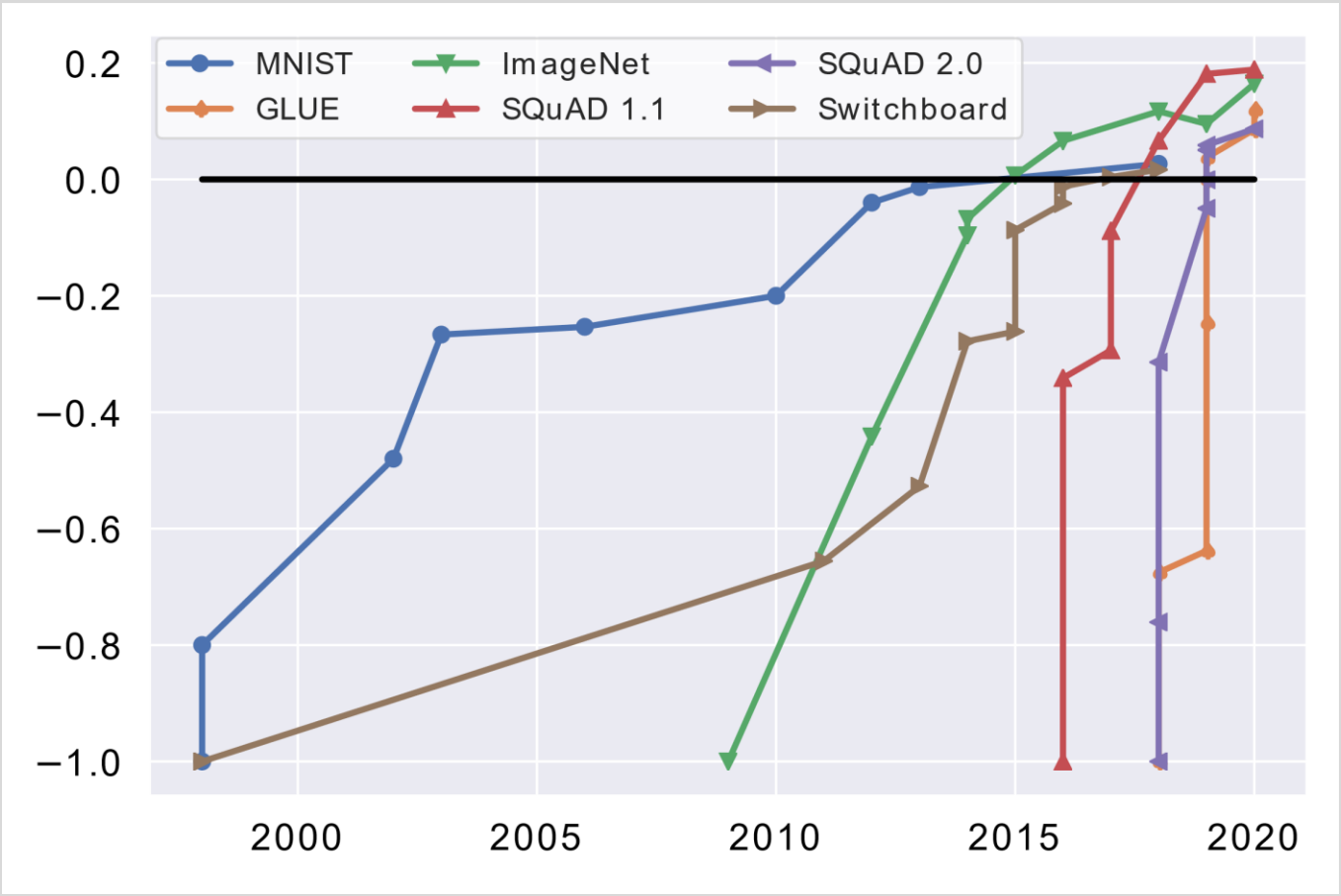
SQuAD: Question Answering

Question: What is Nigeria's official language?

Answer in Context: [...] Nigeria has one of the largest populations of youth in the world. The country is viewed as a multinational state, as it is inhabited by over 500 ethnic groups, of which the three largest are the Hausa, Igbo and Yoruba; these ethnic groups speak over 500 different languages, and are identified with wide variety of cultures. **The official language is English.** [...]

Motivation

KI-Benchmarks aus der Forschung werden immer schneller gelöst



Time to saturation

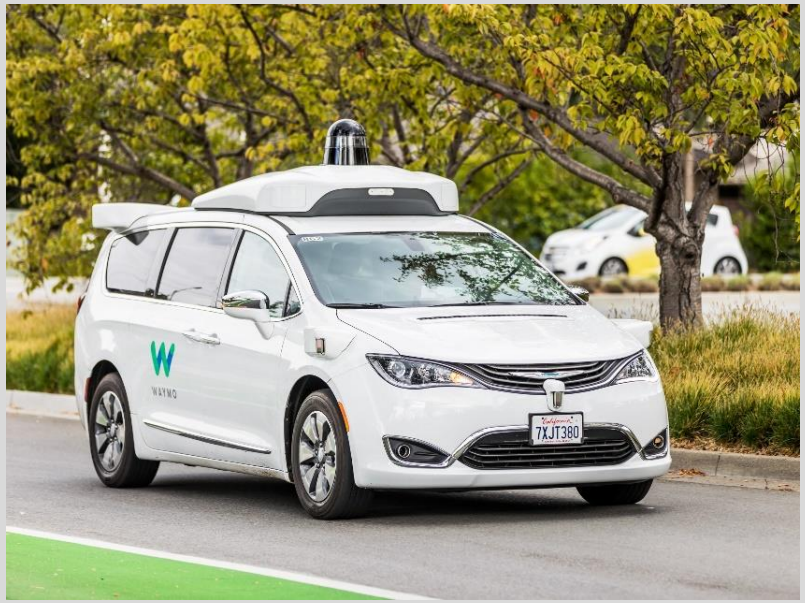
- MNIST 15 years
- ImageNet 6 years
- SQuAD 1.1 2 years
- SQuAD 2.0 1 year
- GLUE 1 year

Kiela et al.: Dynabench: Rethinking Benchmarking in NLP, NAACL-HLT 2021

Motivation

Aus der KI-Forschung resultieren immer mehr erfolgreiche KI-Anwendungen in der Wirtschaft

Sprachassistenten



Selbstfahrende Autos

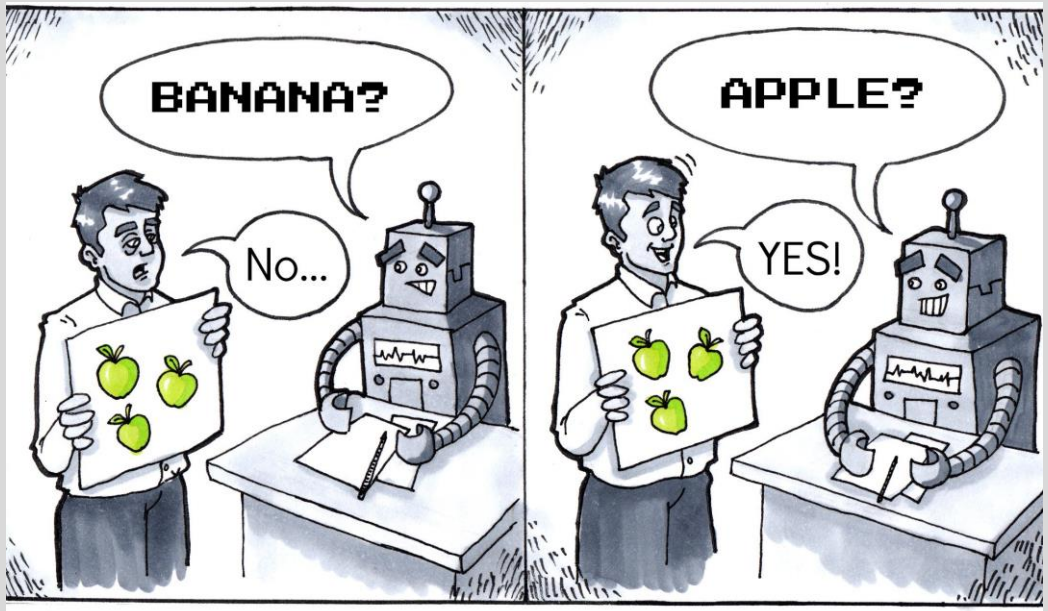
Kassenloser Einkaufsladen

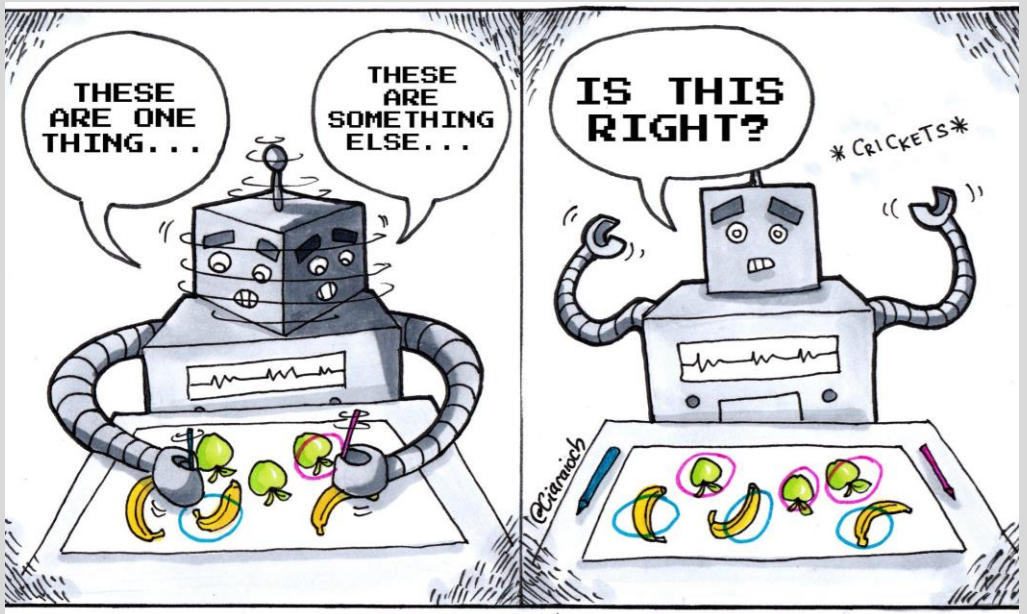
Agenda

1. Motivation
2. Einführung in Machine Learning
3. Aktuelle Ergebnisse und Trends aus der Forschung
4. Herausforderungen für die Praxis und Lösungsansätze
5. Fazit

Einführung in Machine Learning

Machine Learning: *Ein künstliches System lernt aus Beispielen und kann diese nach Beendigung der Lernphase verallgemeinern. Dazu bauen Algorithmen beim maschinellen Lernen ein statistisches Modell auf, das auf Trainingsdaten beruht*. (Wikipedia)





https://twitter.com/athena_schools/status/1063013435779223553

Einführung in Machine Learning

Populärste Variante: Supervised Learning (überwachtes Lernen)

•Task definieren

- Spam-Email-Erkennung
  - Eingabe: Email
  - Ausgabe: Spam ja/nein
- Hauspreis-Vorhersage
  - Eingabe: Lage, Fläche, …
  - Ausgabe: Preis
- Übersetzung
- Spracherkennung
- …

•Datensatz annotieren

- Manuelle Erstellung eines Datensatzes mit konkreter Eingabe und erwarteter Ausgabe
- Spam-Email-Erkennung
  - „*… who is hotter…*“: Spam (ja)
  - „*…bestätige Auftrag…*“: Spam (nein)
- Hauspreis-Vorhersage
  - (Hannover, 150m²): 450 TEUR
  - (Hamburg, 40 m²): 80 TEUR

•Modell definieren

- Architektur und Parameter für Vorhersage-Modell definieren
- Lineare Modelle
- Neuronale Netzwerke
- ...

•Modell trainieren

- Parameterwerte des Modells finden, so dass das Modell für Eingaben in den Trainingsdaten möglichst korrekte Ausgaben erzeugt
- Lösung des Optimierungsproblems durch Gradient Descent

•Modell nutzen

- Mit trainiertem Modell Ausgabe zu Eingabe vorhersagen
- Spam-Email-Erkennung
- „..unser Angebot…“: Spam?
- Hauspreis-Vorhersage
- (München, 100m²): Preis?

Agenda

1. Motivation
2. Einführung in Machine Learning
3. Aktuelle Ergebnisse und Trends aus der Forschung
4. Herausforderungen für die Praxis und Lösungsansätze
5. Fazit

Aktuelle Trends und Ergebnisse aus der Forschung

Supervised Learning erfordert i. d. R. *viele annotierte* Daten zum Training!

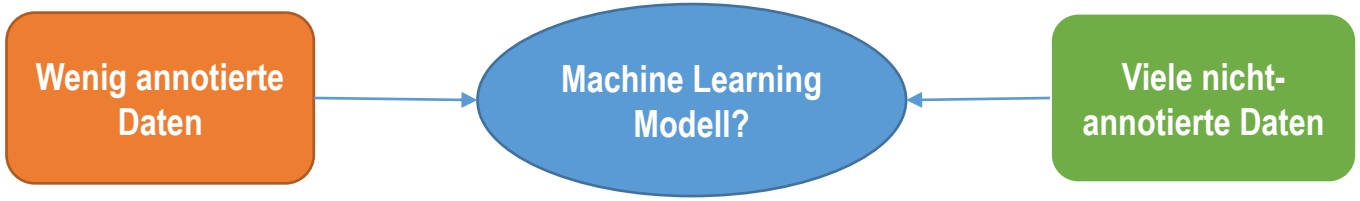
- **MNIST**: 70.000 annotierte Bilder
- **ImageNet**: mehr als 14 Mio. annotierte Bilder mit über 21.000 Kategorien
- **SQuAD**: über 100.000 Frage-Antwort-Paare
- **Gesichtserkennung von Baidu**: über 1 Mrd. Bilder

Das Annotieren von Daten muss durch Menschen erfolgen und ist aufwändig!

…but if unrefined it cannot really be used…

https://medium.com/@randhirhebbar/data-is-the-new-oil-but-are-we-making-the-most-of-it-e636fa30e9ce

Es stehen oft viele nicht-annotierte Daten zur Verfügung: Wie können diese genutzt werden?



Aktuelle Trends und Ergebnisse aus der Forschung

Self-Supervised Learning: *… Es handelt sich um eine Art autonomes Lernen…, bei dem keine durch Menschen im Voraus klassifizierten Beispieldaten benötigt werden*. (Wikipedia)

Analogie: Lernendes Kind

(1) Kind schaut sich selbständig viele Tierbücher an
und lernt dadurch Tiere zu unterscheiden (Self-Supervised Learning)

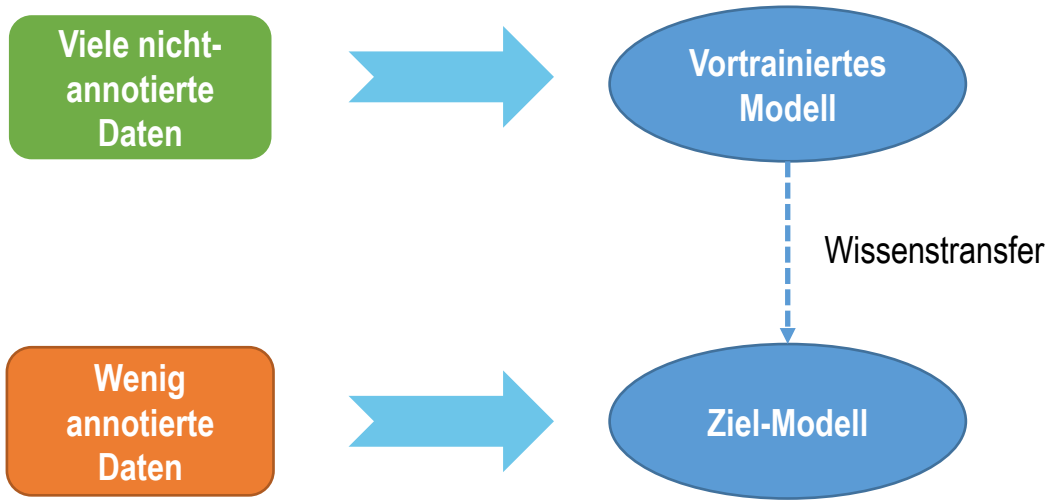
Wissenstransfer

(2) Mutter nennt dem Kind die Tiernamen zu wenigen
Beispielbildern (Supervised Learning)

Aktuelle Trends und Ergebnisse aus der Forschung

Semi-Supervised Learning = Self-Supervised Learning + Supervised Learning

(1) Modell vortrainieren mit Proxy-Task
(Self-Supervised Learning)

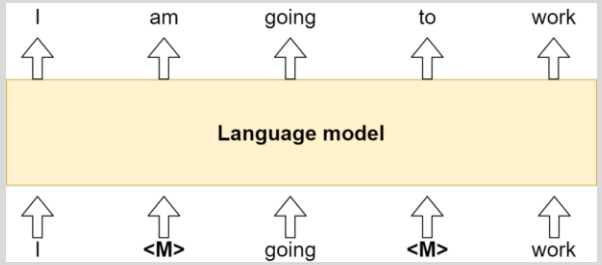


(2) Modell verfeinern mit Ziel-Task
(Supervised Learning)

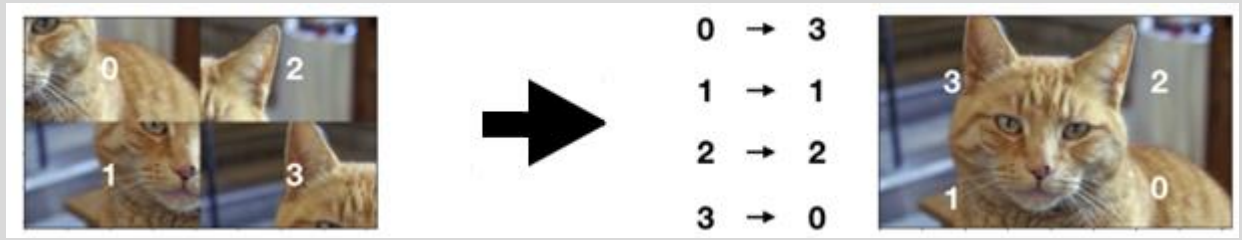
Aktuelle Trends und Ergebnisse aus der Forschung

Proxy-Tasks im Self-Supervised Learning: erzeuge automatisch (künstlich) annotierte Trainingsdaten aus den Rohdaten

- Textverarbeitung: Wortlücken füllen



- Bildverarbeitung: Lösen von Puzzles



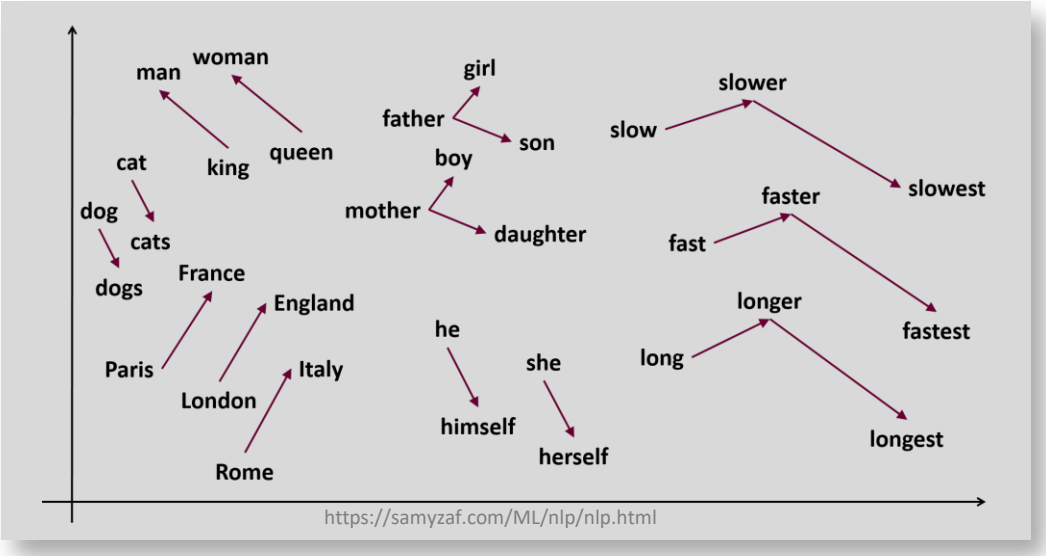
https://medium.com/analytics-vidhya/what-is-self-supervised-learning-in-computer-vision-a-simple-introduction-def3302d883d

Self-Supervised Learning ist erstmal nur ein Vorschritt, um sinnvolle Aufgabenstellungen mit wenig annotierten Daten lösen zu können!

Aktuelle Trends und Ergebnisse aus der Forschung

Warum ist Self-Supervised-Learning im Bereich Textverarbeitung sinnvoll?

- Es gibt sehr viele rohe Textdaten (Wikipedia, News, Gesetze, Dokumente, …)
- Modell baut allgemeines Wissen über die Welt und gesunden Menschenverstand auf
- Modell lernt (semantische) Ähnlichkeiten zwischen Wörtern:
  z. B. *„Alexa, mach das Licht im Wohnzimmer an“* vs. *„Alexa, schalte die Lampe in der Stube ein“*
- Vortrainierte Modelle können mit wenig Trainingsdaten an sinnvolle Tasks angepasst werden
- Populärer Vertreter: BERT von Google aus 2018 (Devlin et al., 2019)



Aktuelle Trends und Ergebnisse aus der Forschung

SQuAD-Benchmark

- Question Answering Datensatz von Stanford Universität aus 2016
- Über 100.000 Frage-Antwort-Paare in Wikipedia-Artikeln
- Neue schwierigere Version 2.0 in 2018,
  nachdem erste Version gelöst wurde
- Grundlage für diverse Anwendungen:
  - Passende Antworten in Suchmaschinen finden
  - Automatische Beantwortung von Fragen in Chatbots
  - Extraktion von Informationen aus Dokumenten

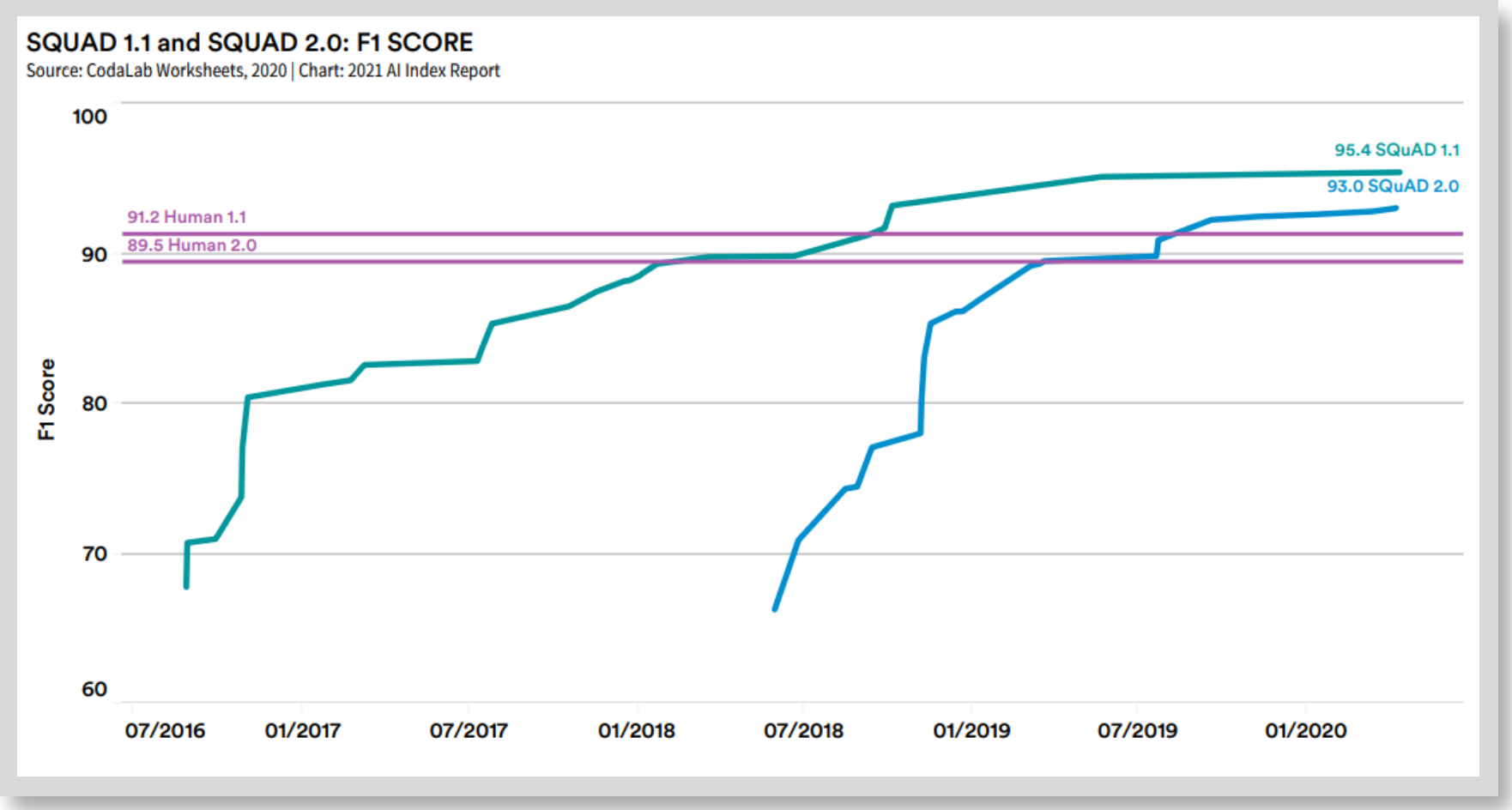
The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

Aktuelle Trends und Ergebnisse aus der Forschung

SQuAD 1.1 nach ca. 2 Jahren und SQuAD 2.0 nach ca. 1 Jahr Forschung mit Hilfe von Self-Supervised-Learning gelöst…



Aktuelle Trends und Ergebnisse aus der Forschung

SuperGLUE-Benchmark

- Benchmark für acht schwierige Tasks im Bereich Textverständnis
- 2019 veröffentlicht, nachdem der Vorgänger Benchmark (GLUE) gelöst wurde
- Zwei Beispiel-Tasks:

MultiRC

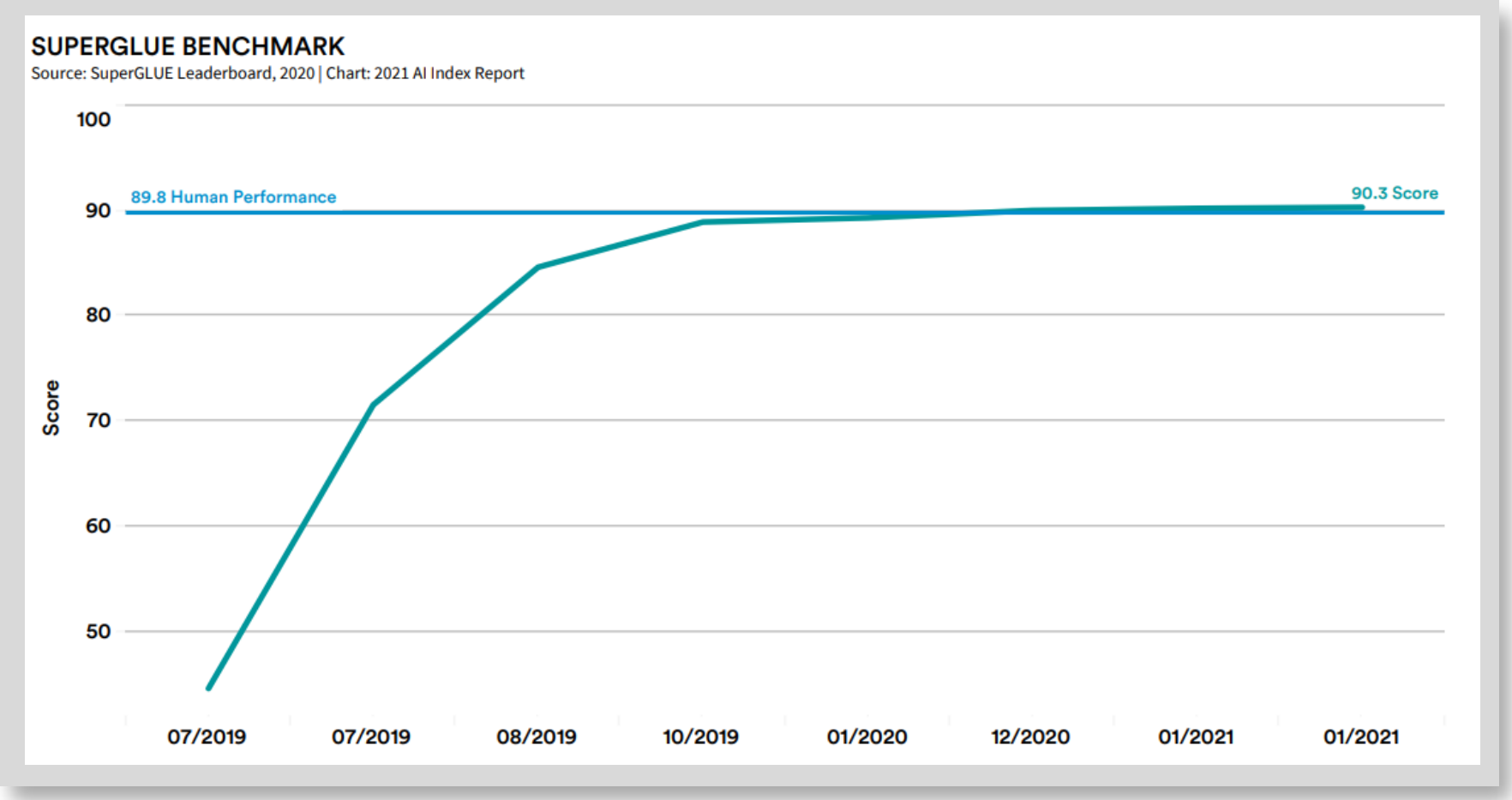
Paragraph: *Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week*
Question: *Did Susan's sick friend recover?* **Candidate answers:** *Yes, she recovered* (T), *No* (F), *Yes* (T), *No, she didn't recover* (F), *Yes, she was at Susan's party* (T)

RTE

Text: *Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.*
Hypothesis: *Christopher Reeve had an accident.* **Entailment:** `False`

Aktuelle Trends und Ergebnisse aus der Forschung

Der SuperGLUE-Benchmark wurde nach ca. 1,5 Jahren Forschung mit Hilfe von Self-Supervised-Learning gelöst…



Aktuelle Trends und Ergebnisse aus der Forschung

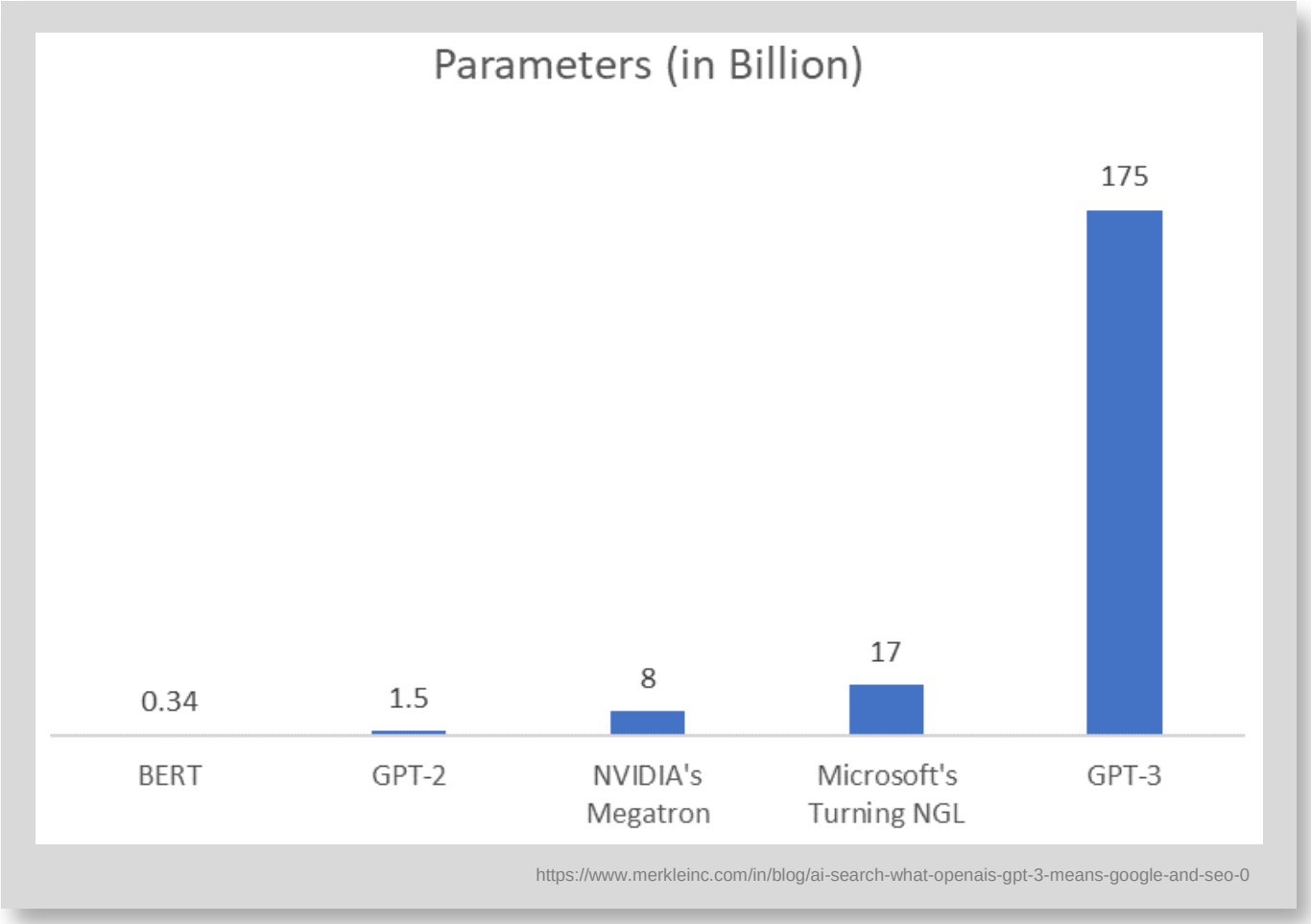
Zwischenfazit

- Self-Supervised Learning ermöglicht einem Modell vollständig autonom „allgemeines Wissen“ in unstrukturierten Daten aufzubauen
- Damit können leistungsfähige Modelle mit *relativ wenig* annotierten Daten für eine Ziel-Aufgabe trainiert werden
- Self-Supervised Learning hat zum rasanten Fortschritt in der KI-Forschung beigetragen

Menschen können jedoch aus sehr wenig Beispielen lernen. Können wir das auch in einem KI-System imitieren?

Aktuelle Trends und Ergebnisse aus der Forschung

Was passiert, wenn man noch größere Modelle mit noch mehr Daten mit Self-Supervised-Learning trainiert?



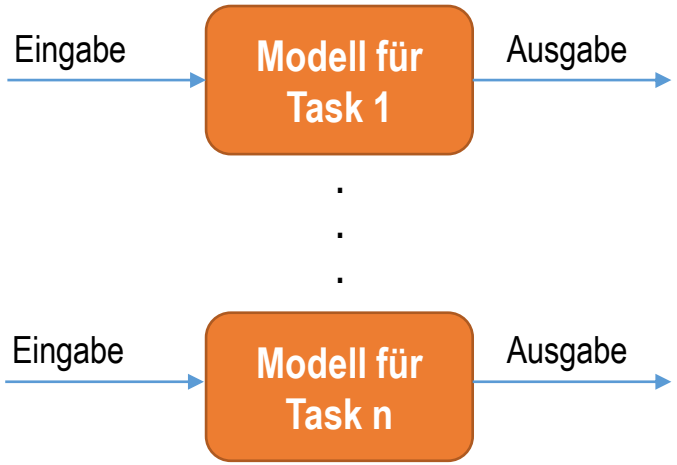
https://www.merkleinc.com/in/blog/ai-search-what-openais-gpt-3-means-google-and-seo-0

Aktuelle Trends und Ergebnisse aus der Forschung

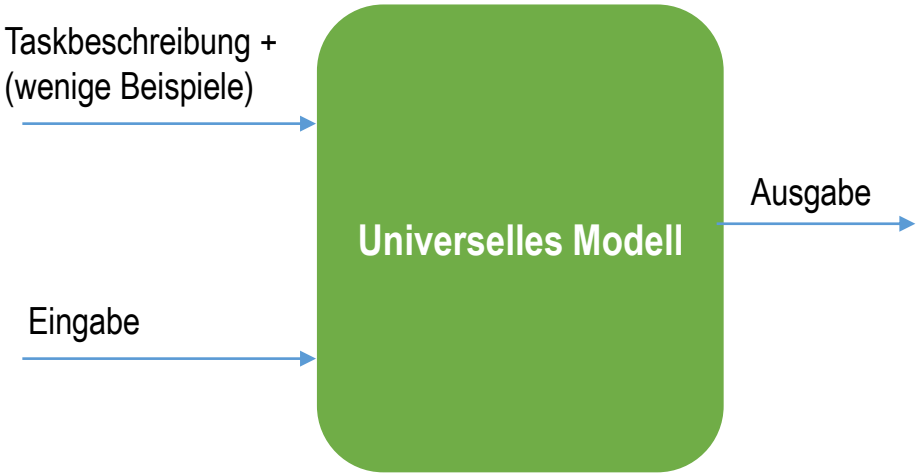
Neues Paradigma durch immer größere Modelle: Few/One/Zero-Shot Learning

- Supervised Learning mit ***sehr wenig*** Trainingsdaten
- Universelle Modelle werden zunehmend möglich (Brown et al. 2020)

Aktuelles Paradigma:

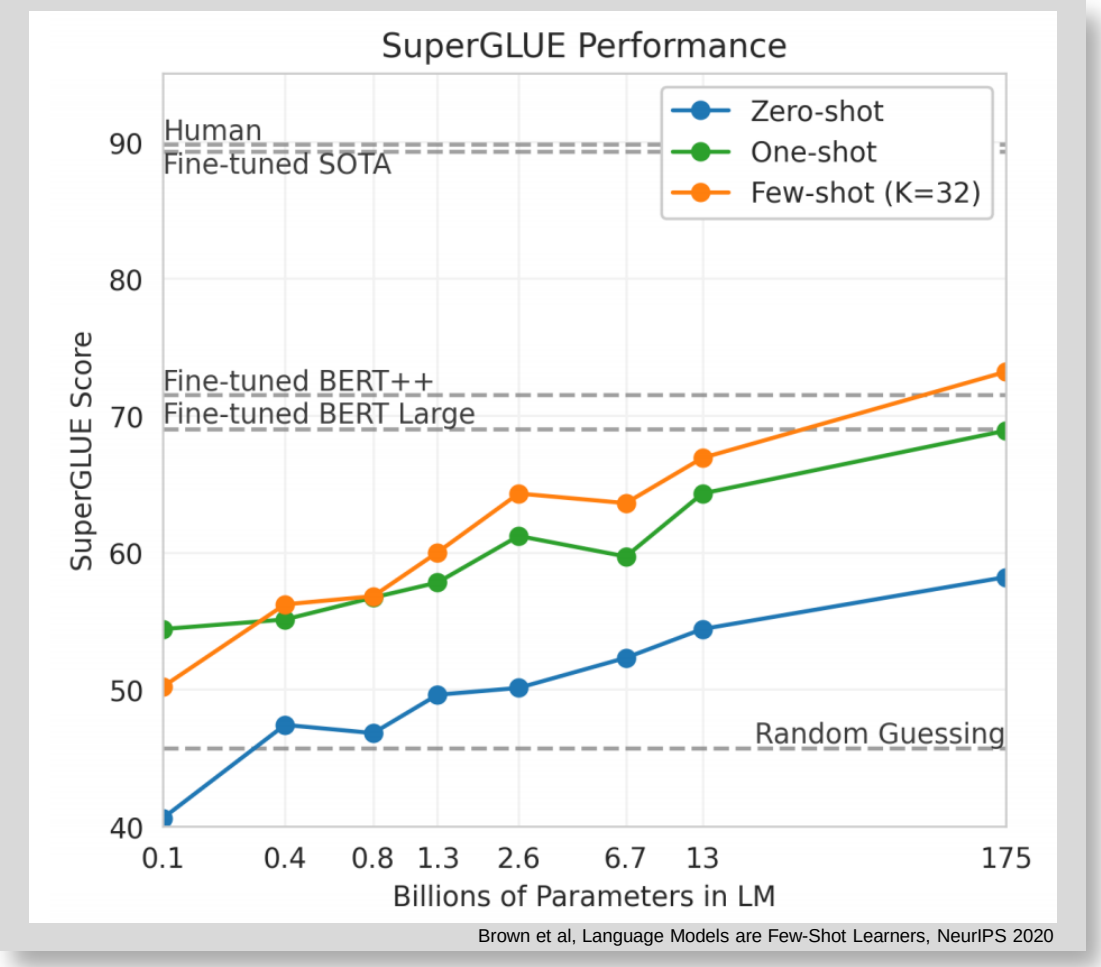


Mögliches neues Paradigma:



Aktuelle Trends und Ergebnisse aus der Forschung

GPT-3 (175 Milliarden Parameter, vortrainiert auf 570 GB Text) erzielt mit sehr wenig Trainingsdaten bereits gute Ergebnisse



Brown et al, Language Models are Few-Shot Learners, NeurIPS 2020

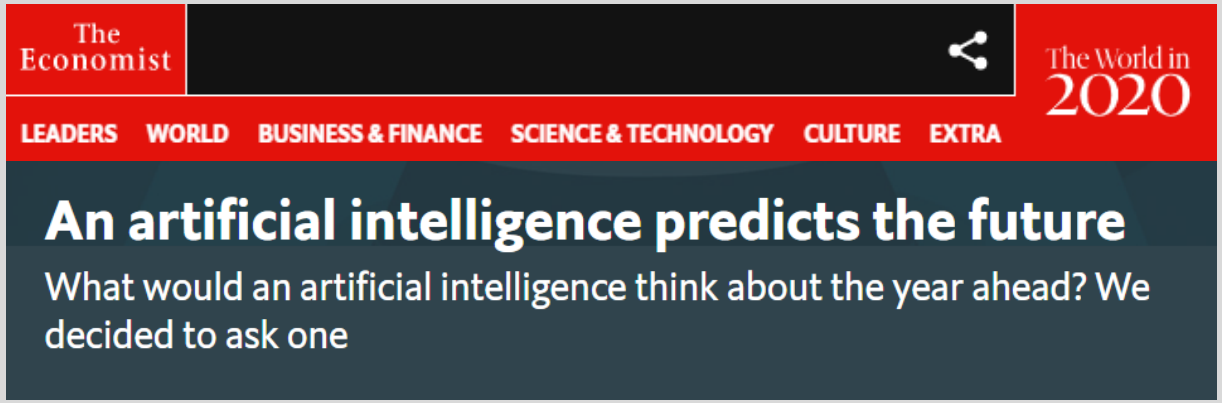
Aktuelle Trends und Ergebnisse aus der Forschung

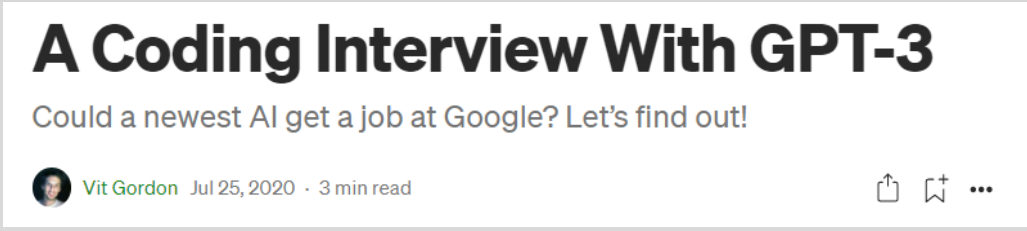
Interessante (Spaß-) Anwendungen mit GPT-3

OpenAI GPT-3 Demos

- 3.1 1. OpenAI GPT-3 becomes an Author
- 3.2 2. OpenAI GPT-3, the Blogger
- 3.3 3. OpenAI GPT-3 can steal jobs from Software Engineer?
- 3.4 4. Writing Realistic Business Memos
- 3.5 5. Testing OpenAI GPT-3 with a Turing Test
- 3.6 6. Generating Real Like Faux Interviews
- 3.7 7. GPT-3 Generating Cooking Recipies
- 3.8 8. GPT-3 Changes the Tone of the Sentence
- 3.9 9. GPT-3 Cracking Jokes
- 3.10 10. English to Regex Conversion with GPT-3
- 3.11 11. Creating Website Mockups with GPT-3
- 3.12 12. Generating Use Cases of Objects
- 3.13 13. Autoplotter – Creating Plots automatically with GPT-3
- 3.14 14. Learn From Any One with GPT-3
- 3.15 15. GPT-3 can Explain Codes
- 3.16 16. GPT-3 the Quiz Master
- 3.17 17. GPT-3 Converts English to Latex
- 3.18 18. GPT-3 Making Intelligent Analogies
- 3.19 19. GPT-3 Converts English into Linux Commands
- 3.20 20. GPT-3 generates Machine Learning Model
- 3.21 21. GPT-3 generates Faces

https://machinelearningknowledge.ai/openai-gpt-3-demos-to-convince-you-that-ai-threat-is-real-or-is-it/





Agenda

1. Motivation
2. Einführung in Machine Learning
3. Aktuelle Ergebnisse und Trends aus der Forschung
4. Herausforderungen für die Praxis und Lösungsansätze
5. Fazit

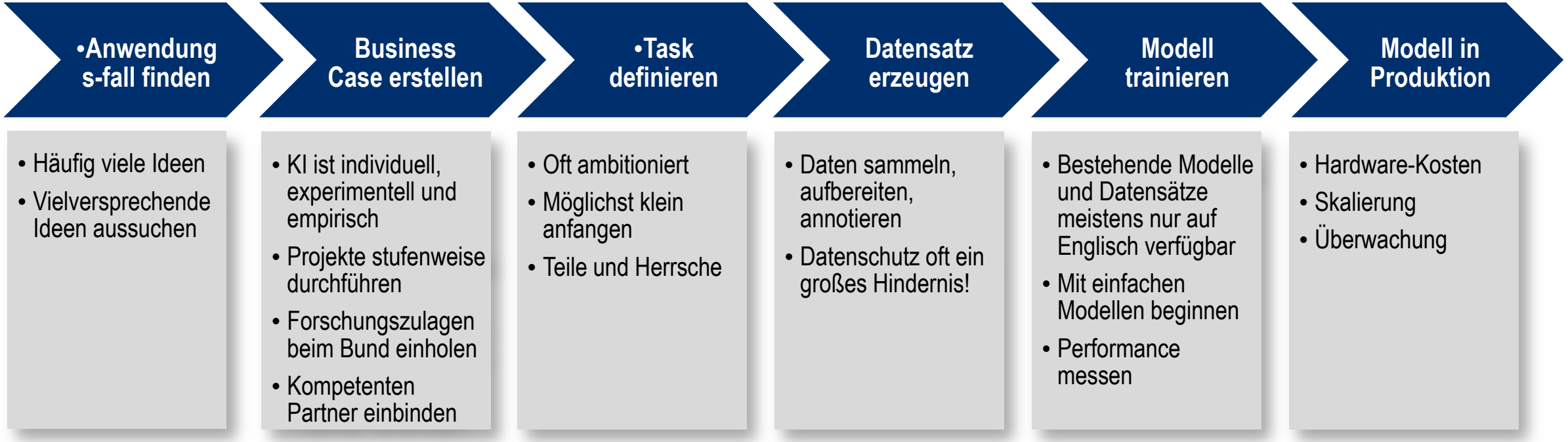
Herausforderungen für die Praxis und Lösungsansätze

Mögliche Einsatzszenarien von KI in unserer Branche

- Kundenkommunikation
    - Kundenanfragen (teil-)automatisch bearbeiten
    - Strukturierte Informationen aus Dokumenten und Nachrichten extrahieren
    - Meinungsbild von Unternehmen auf Social Media ermitteln
- Produktion
    - Vorhersagen für die Produktionsplanung und –steuerung
    - Predictive Maintenance
- Versand
    - Anschriftenerkennung
    - Tourenoptimierung
- Komplexität von Prozessen effizient handhaben und reduzieren
    - Teilautomatische Parametrisierung von Anwendungen
    - Automatische Extraktion von verarbeitungsrelevanten Daten

Herausforderungen für die Praxis und Lösungsansätze

Möglicher Ablauf und Herausforderungen bei KI-Projekten in der Praxis



Agenda

1. Motivation
2. Einführung in Machine Learning
3. Aktuelle Ergebnisse und Trends aus der Forschung
4. Herausforderungen für die Praxis und Lösungsansätze
5. Fazit

Fazit

- Der Fortschritt in der KI-Forschung schreitet rasant voran
- Es entstehen immer mehr praxistaugliche KI-Anwendungen
- Moderne KI-Methoden haben viel Potenzial die Flut an unstrukturierten Daten zu strukturieren
- Einsatz von KI-Methoden in der Praxis ist mit einigen Hürden verbunden
- KI wird für die Wettbewerbsfähigkeit von Unternehmen zunehmend wichtiger

Vielen Dank

Arthur Brack
Geschäftsführer

SET GmbH

Rühmkorffstraße 5
30163 Hannover

E-Mail: arthur.brack@set.de

LinkedIn: https://de.linkedin.com/in/arthur-brack
Twitter: https://twitter.com/ArthurBrack
Publikationen: https://www.semanticscholar.org/author/Arthur-Brack/32617107

DOXNET®
THE DOCUMENT X-PERTS NETWORK

KONTAKT & WEITERE INFORMATIONEN

Wir freuen uns
über Ihr Interesse

The Document X-perts Network e.V.

Grüner Weg 22
35578 Wetzlar
Germany

http://www.doxnet.de