Digital Applications & Data Management

WS25/26

Dr. Jens Kohl

Roadmap Vorlesung

1. Einführung und Übersicht
2. Grundlagen Data Science
3. Vorgehen Data Science Use Case
4. Case Study Data Science
5. Grundlagen unüberwachtes Lernen
6. Grundlagen überwachtes Lernen (tabellarische Daten)
7. Case Study überwachtes Lernen (tabellarische Daten)
8. Grundlagen überwachtes Lernen (Bilddaten)
9. Case Study überwachtes Lernen und Transfer Learning (Bilddaten)
10. Grundlagen Generative AI
11. Generative AI mit Texten und Prompt Engineering
12. Agentic AI
13. Ausblick: Machine Learning in der Cloud und Reinforcement Learning

Vorlesung 1:
Einführung und Übersicht

Prüfungsleistung

- Abzugebender Leistungsumfang:
  - Implementierung von 2 Use Cases aus Kategorie Data Science, Künstliche Intelligenz oder GenAI.
  - Zusätzlich schriftliche Ausarbeitung von max. 12 Seiten (ohne Anhang & Quellen) je Use Case:
    - Vorgehensweise: Detaillieren & erklären eingesetzter Verfahren sowie deren Implementierung
    - Ergebnisse: Visualisierung Ergebnisse und Bewertung anhand Metriken
    - Reflektion und Ausblick

Template wird bereitgestellt

- Umsetzung:
  - Erarbeitung in Gruppen möglich.
  - Aber: jeder Student muß individuelle, bewertbare Leistung abgeben.

- Use Cases werden bereitgestellt, Sie können aber gern eigene Themen vorschlagen.

Digital Applications & Data Management

Was ist das? Was machen wir?? Wozu brauchen Sie das???

- Reduktion Data Science & künstliche Intelligenz auf für **Sie wesentliche Inhalte**
- Verstehen der Grundlagen, **des Impacts auf Sie** sowie praktisches Anwenden
- Sammeln von Hands-on Experience an relevanten Themen aus der Praxis
- Fragen, Fragen, Fragen! Zweiwöchentliche Sprechstunde

Digital Applications & Data Management

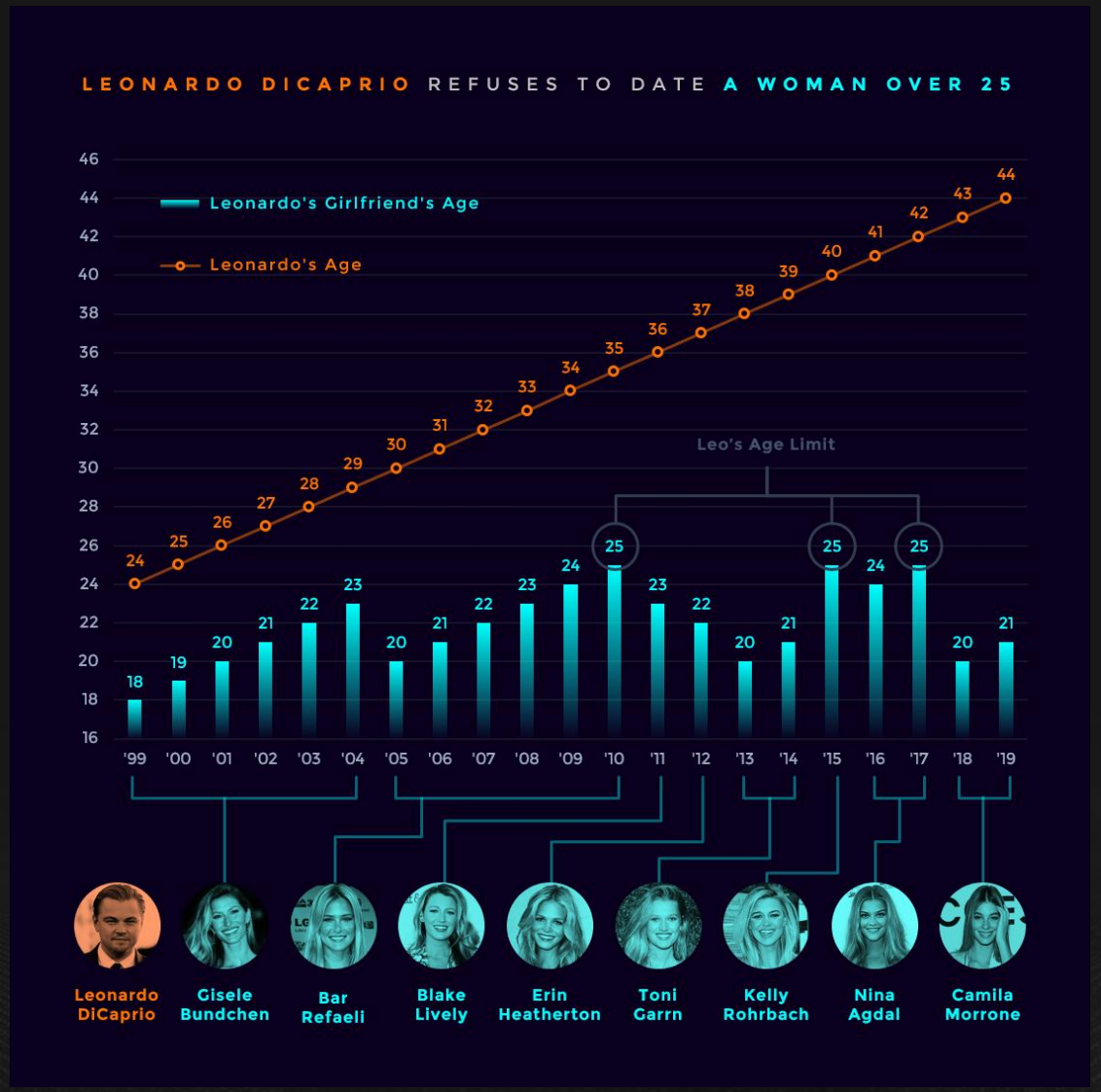
Was ist das? Was machen wir?? Wozu brauchen Sie das???

- Reduktion Data Science & künstliche Intelligenz auf für **Sie wesentliche Inhalte**
- Verstehen der Grundlagen, **des Impacts auf Sie** sowie praktisches Anwenden
- Sammeln von Hands-on Experience an relevanten Themen aus der Praxis
- Fragen, Fragen, Fragen!
- Zweiwöchentliche Sprechstunde

Fallbeispiele Data Science

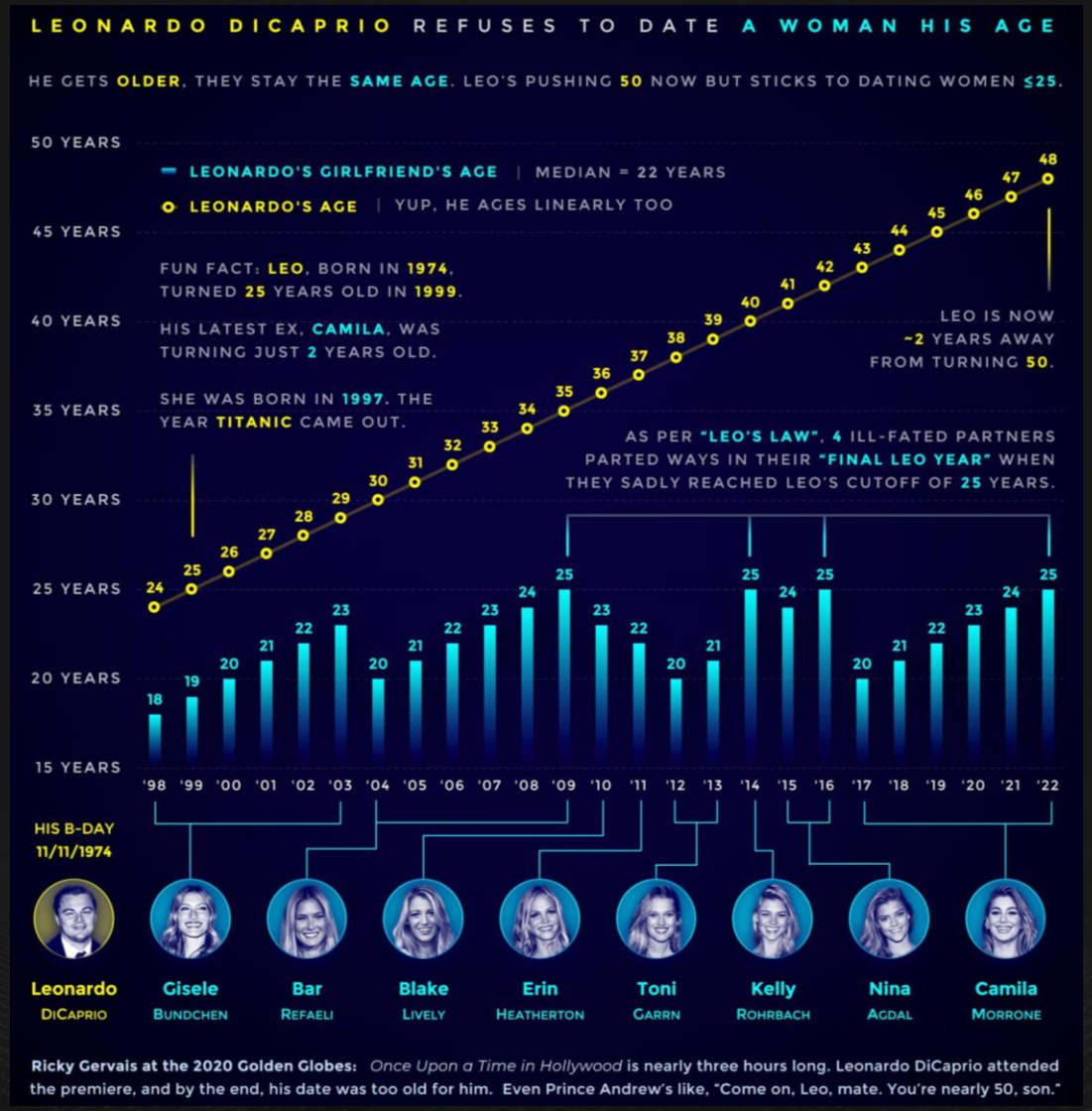
Ziel: Daten analysieren und daraus Erkenntnisse gewinnen und visualisieren für einfachen Wissenstransfer

Data Science



Quelle: https://www.reddit.com/r/dataisbeautiful/comments/azjti7/leonardo_dicaprio_refuses_to_date_a_woman_over_25/

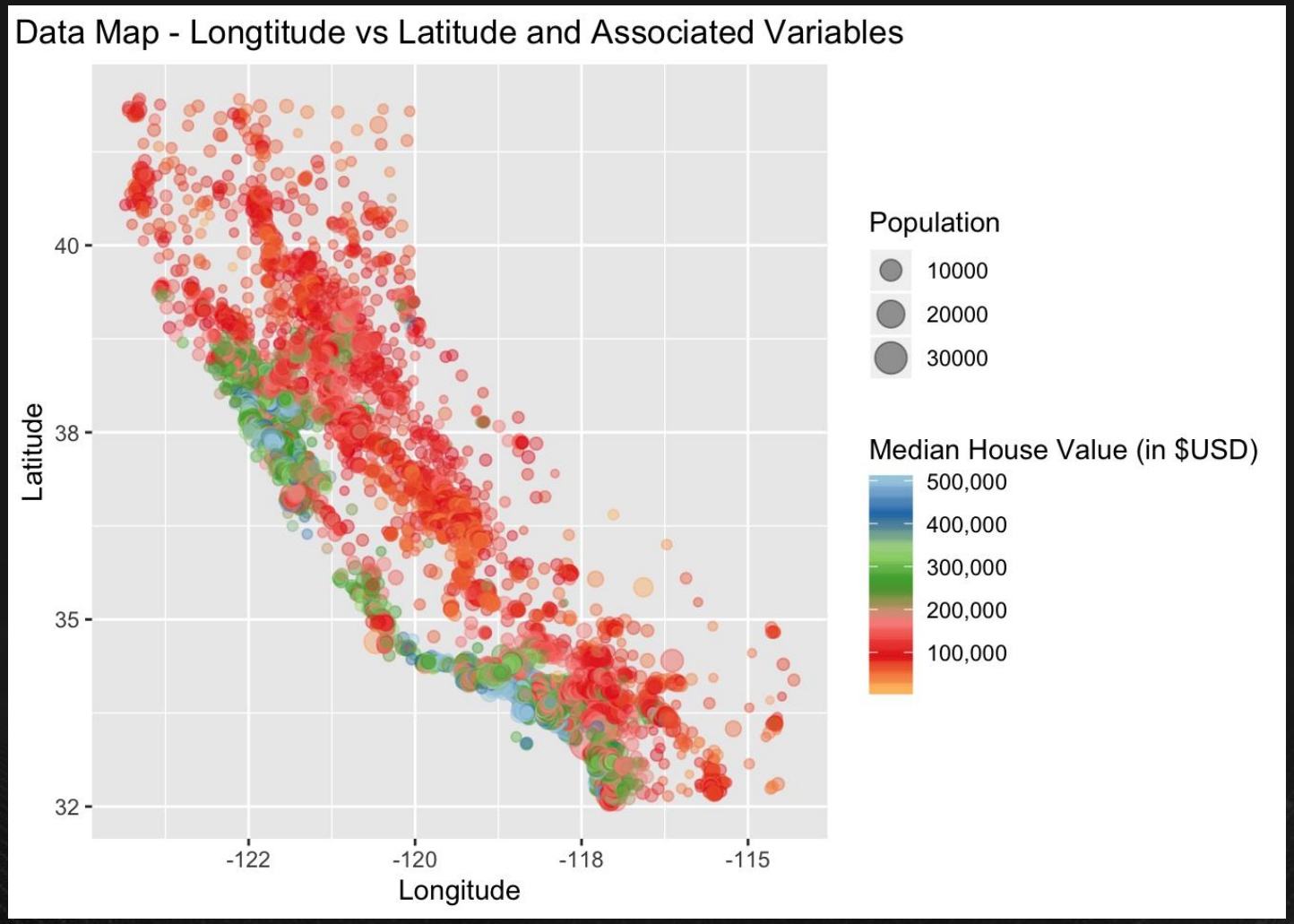
Data Science



Quelle: https://www.reddit.com/r/popculturechat/comments/x3x9qa/leos_law_by_utrustlittlebrother/?tl=de

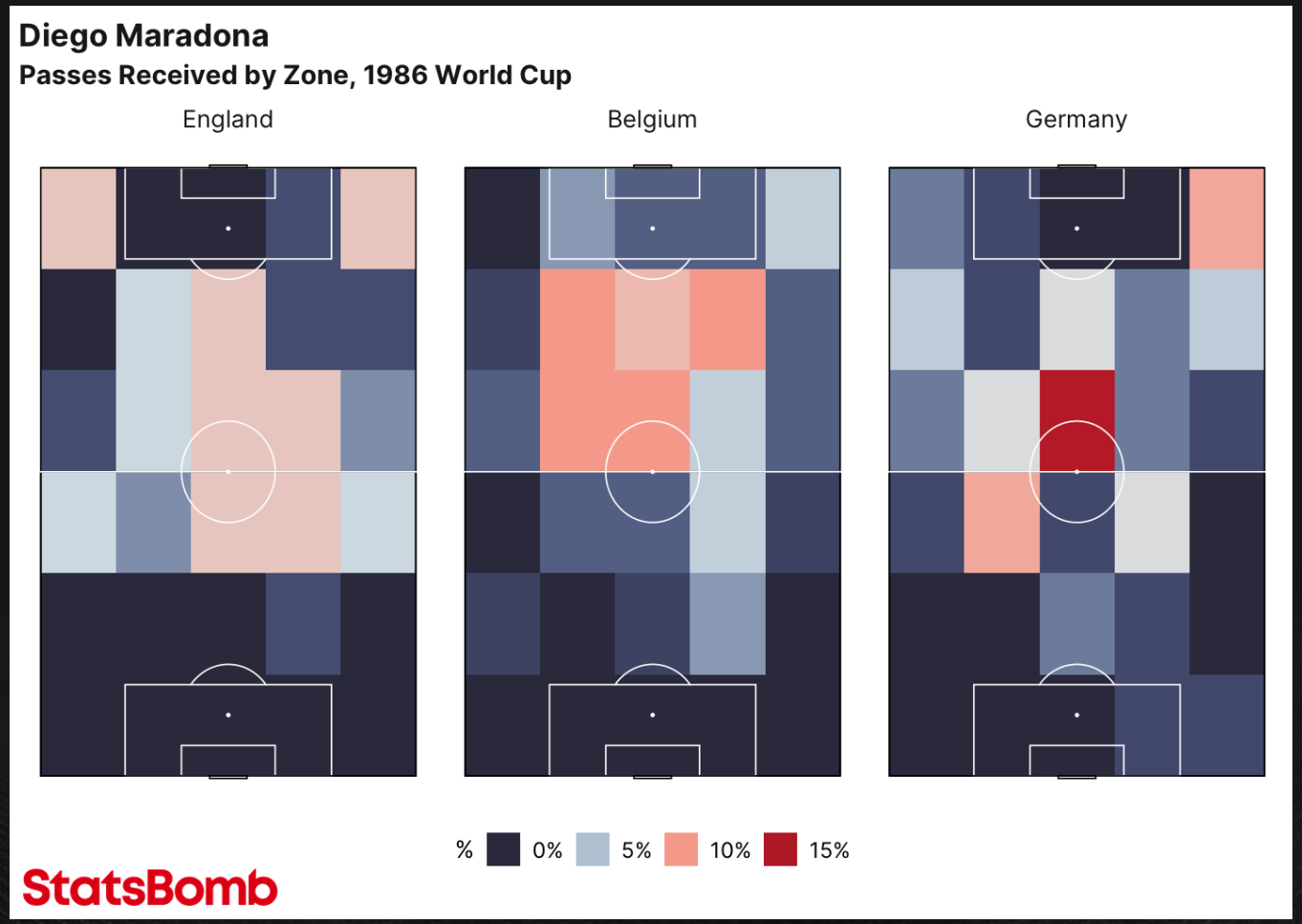
Data Science

California Housing prices



Data Science

Spiele- und Spieleranalysen im Sport



Quelle: https://www.hudl.com/blog/statsbomb-icons-diego-maradona

Fallbeispiele künstliche Intelligenz und Machine Learning

Ziel: aus Daten ein Modell lernen, das eigenständig Erkenntnisse gewinnt

Beispielhafte AI Use Cases

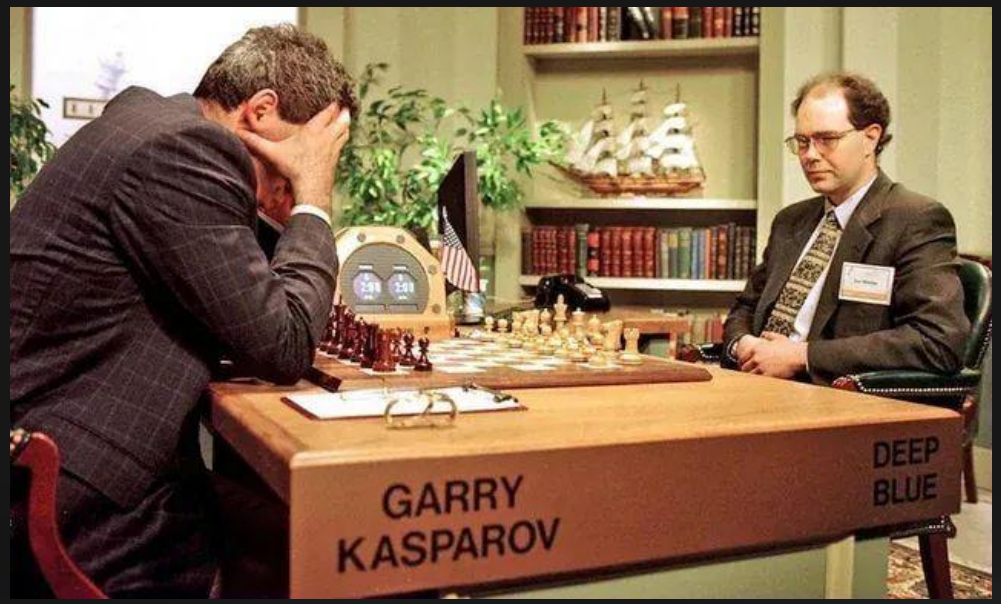
Unüberwachtes Lernen für Empfehlen Produkte

Beispielhafte AI Use Cases

Unüberwachtes Clustern von Kunden

Beispielhafte AI Use Cases

Überwachtes Lernen: Deep Blue (endlicher Zustandsraum)



Quellen:
linkes Bild: https://www.reddit.com/r/chess/comments/1mpslpf/kasparov_vs_deep_blue_1997/?tl=de
Rechtes Bild: https://theblogisright.com/wp-content/uploads/2011/02/848f5-fullscreencapture218201143319pm-bmp.jpg

Beispielhafte AI Use Cases

Reinforcement Learning kombiniert mit Convolutional Neural Networks (expontieller Zustandsraum)

Quelle: https://www.theguardian.com/technology/2016/mar/15/alphago-what-does-google-advanced-software-go-next

Beispielhafte AI Use Cases

DeepMind: Playing Atari with reinforcement learning

Quelle: https://youtu.be/Q70ulPJW3Gk

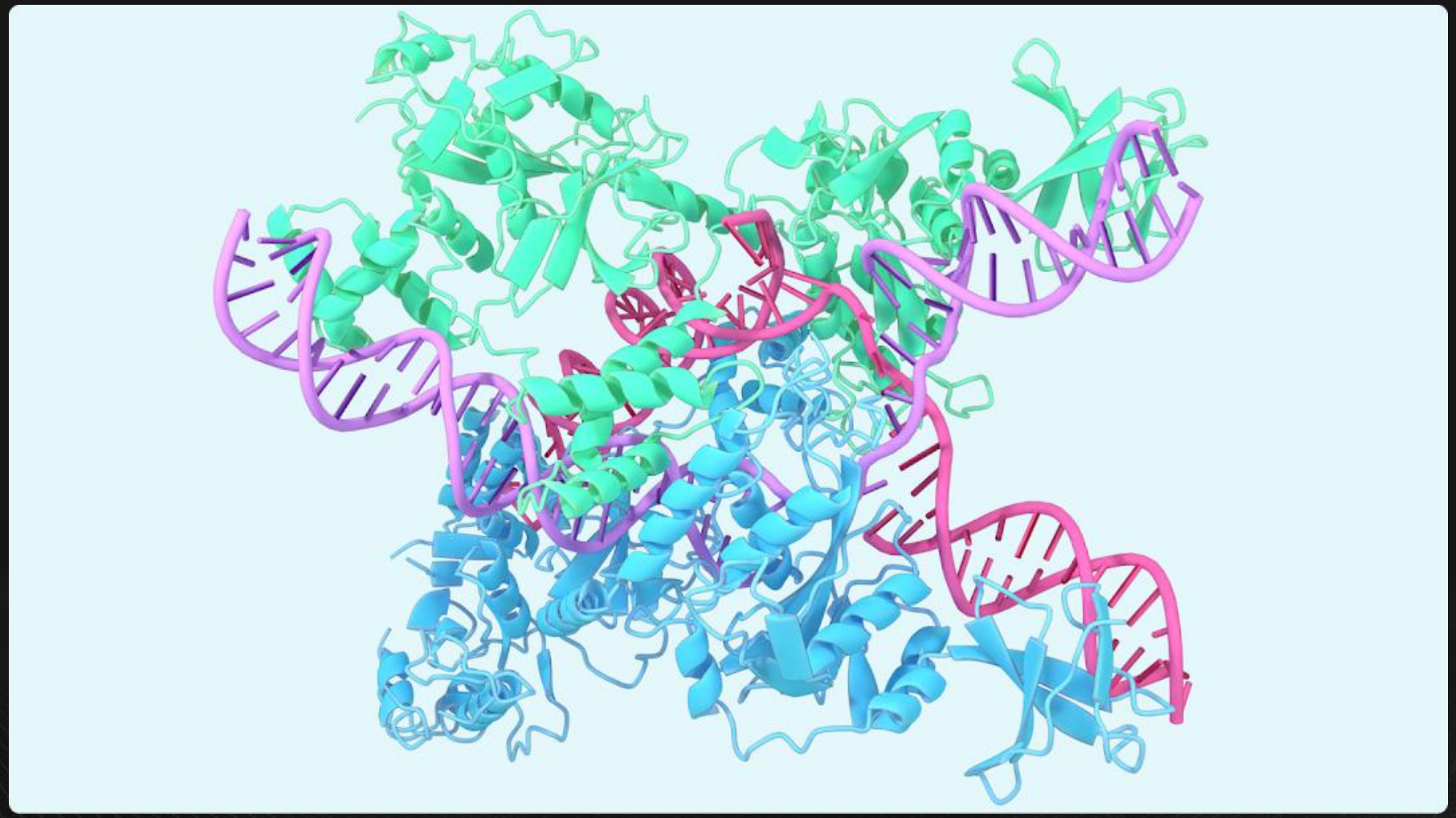
Beispielhafte AI Use Cases

Multi-agent reinforcement learning

Quelle: Screenshot: aus https://www.heise.de/hintergrund/Wie-die-DeepMind-KI-AlphaStar-Profispieler-in-StarCraft-2-besiegte-4308763.html
Veröffentlichung: Vinyals et al., „Grandmaster level in StarCraft II using multi-agent reinforcement learning", Nature, Vol. 575, Nr. 7782, pp. 350 – 354, 2019.

Beispielhafte AI Use Cases

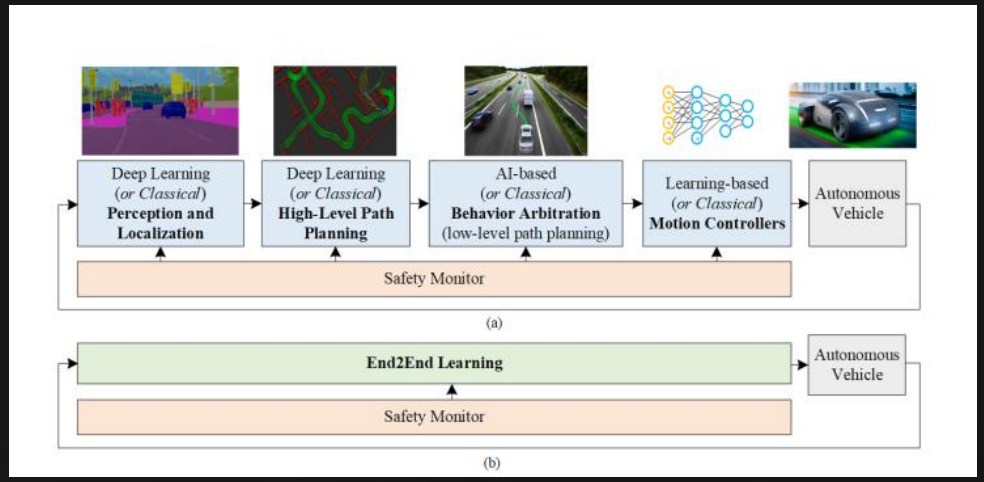
AlphaFold: Prädiktion dreidimensionaler Proteinstruktur



Quelle: https://deepmind.google/blog/a-glimpse-of-the-next-generation-of-alphafold/

Beispielhafte AI Use Cases

Autonomes Fahren



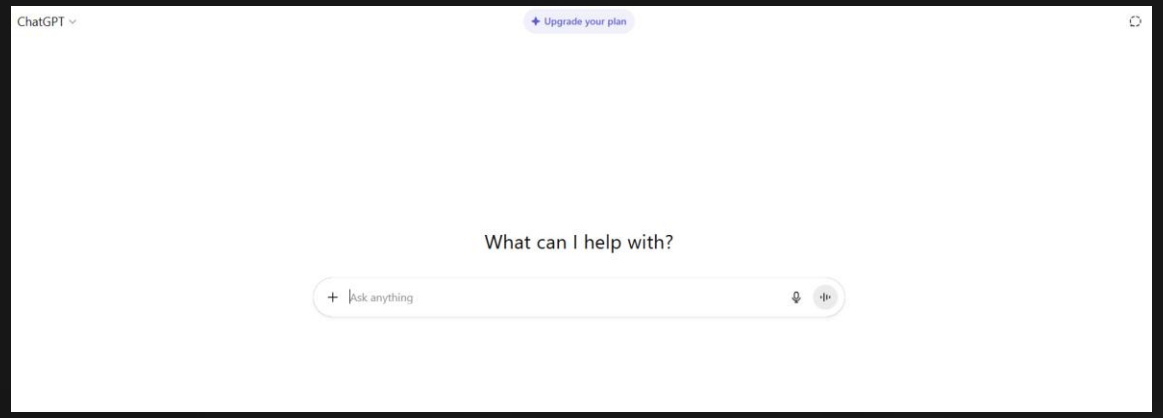
Quellen: https://waymo.com/blog/, https://www.tesla.com/de_de/fsd, https://arxiv.org/pdf/1910.07738

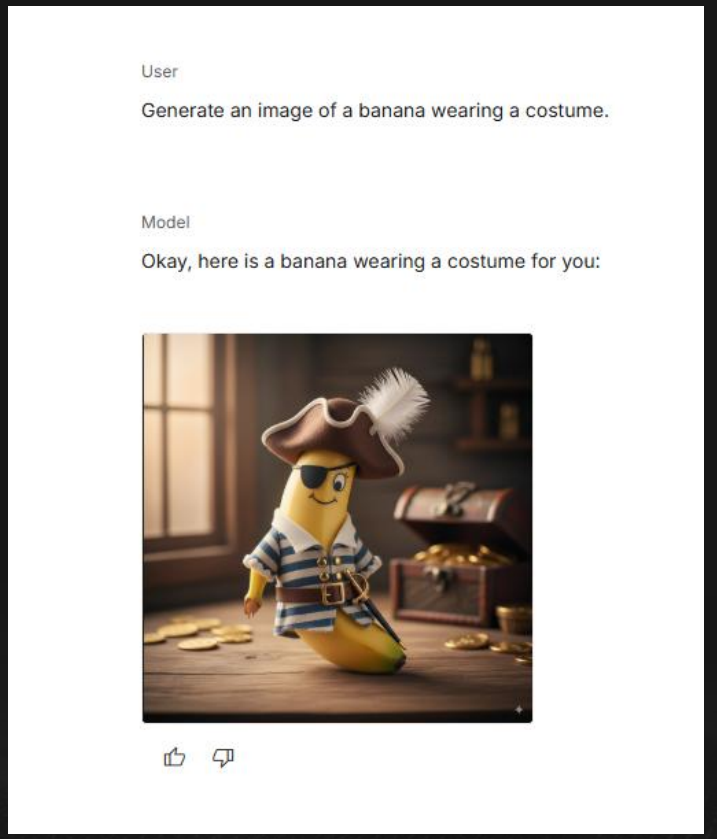
Fallbeispiele Generative AI

Ziel: aus Daten ein Modell lernen, das eigenständig Erkenntnisse gewinnt

Beispielhafte GenAI Use Cases

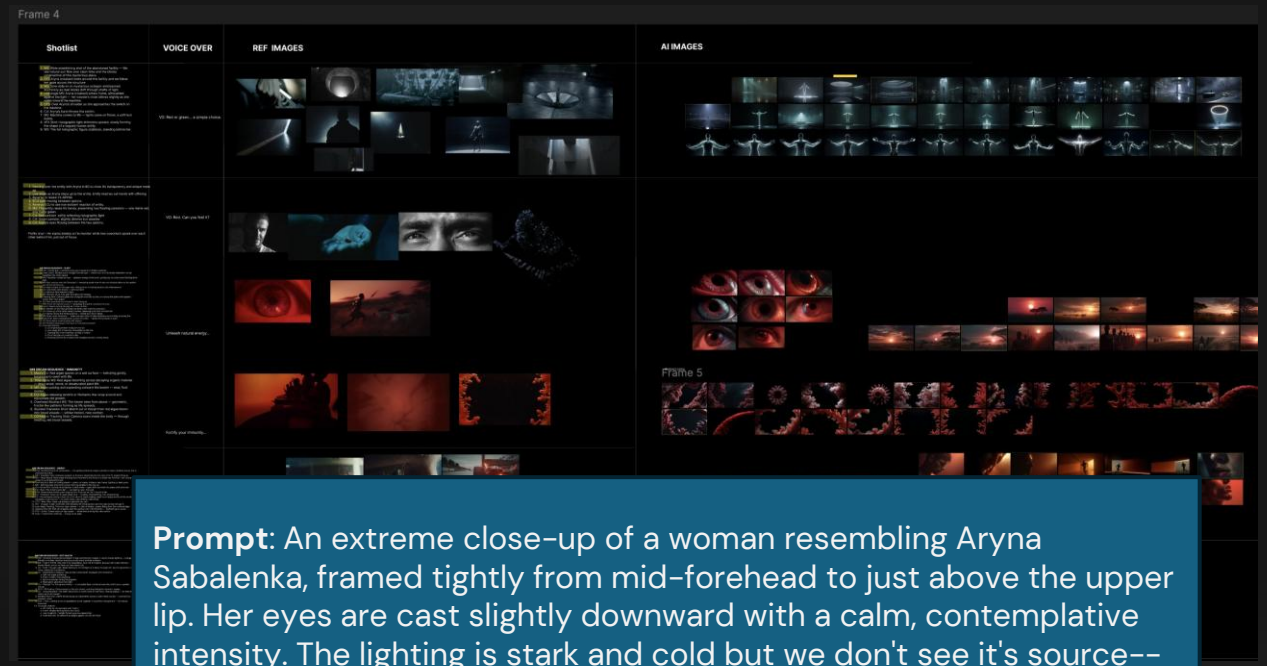
Generieren von Text und Bildern durch Generative AI



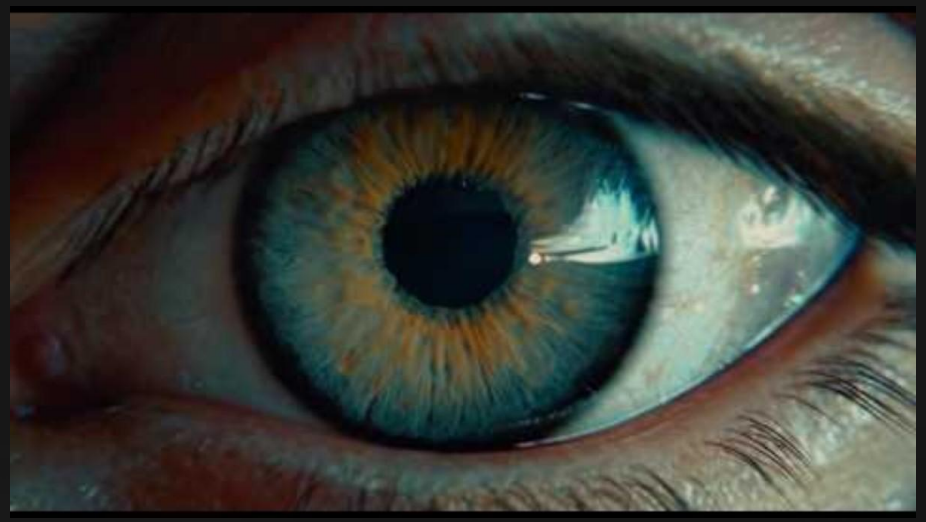


Beispielhafte GenAI Use Cases

Generieren von Videos durch Generative AI



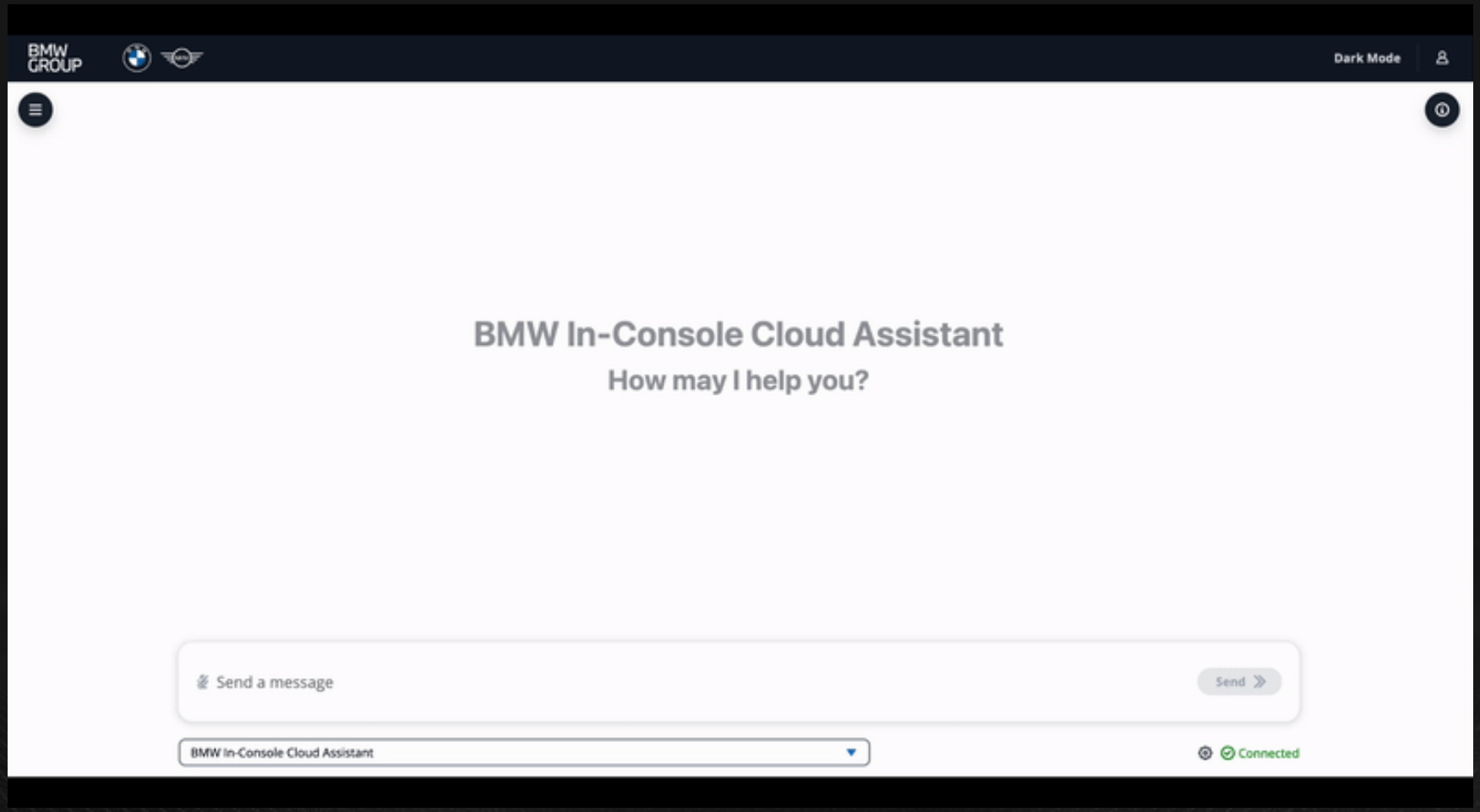
Prompt: An extreme close-up of a woman resembling Aryna Sabalenka, framed tightly from mid-forehead to just above the upper lip. Her eyes are cast slightly downward with a calm, contemplative intensity. The lighting is stark and cold but we don't see it's source--icy blue and softly diffused. Cool tones brush across her skin, revealing every pore and lash with subtle clarity. She looks straight at the camera. Her features are strong yet serene, the reflections in her irises hinting at a cold, metallic space beyond. The frame feels atmospheric, cinematic, and intimate--like a pivotal moment in a high-budget sci-fi film.



Quelle: https://x.com/PJaccetturo/status/1960600712251105650

Beispielhafte GenAI Use Cases

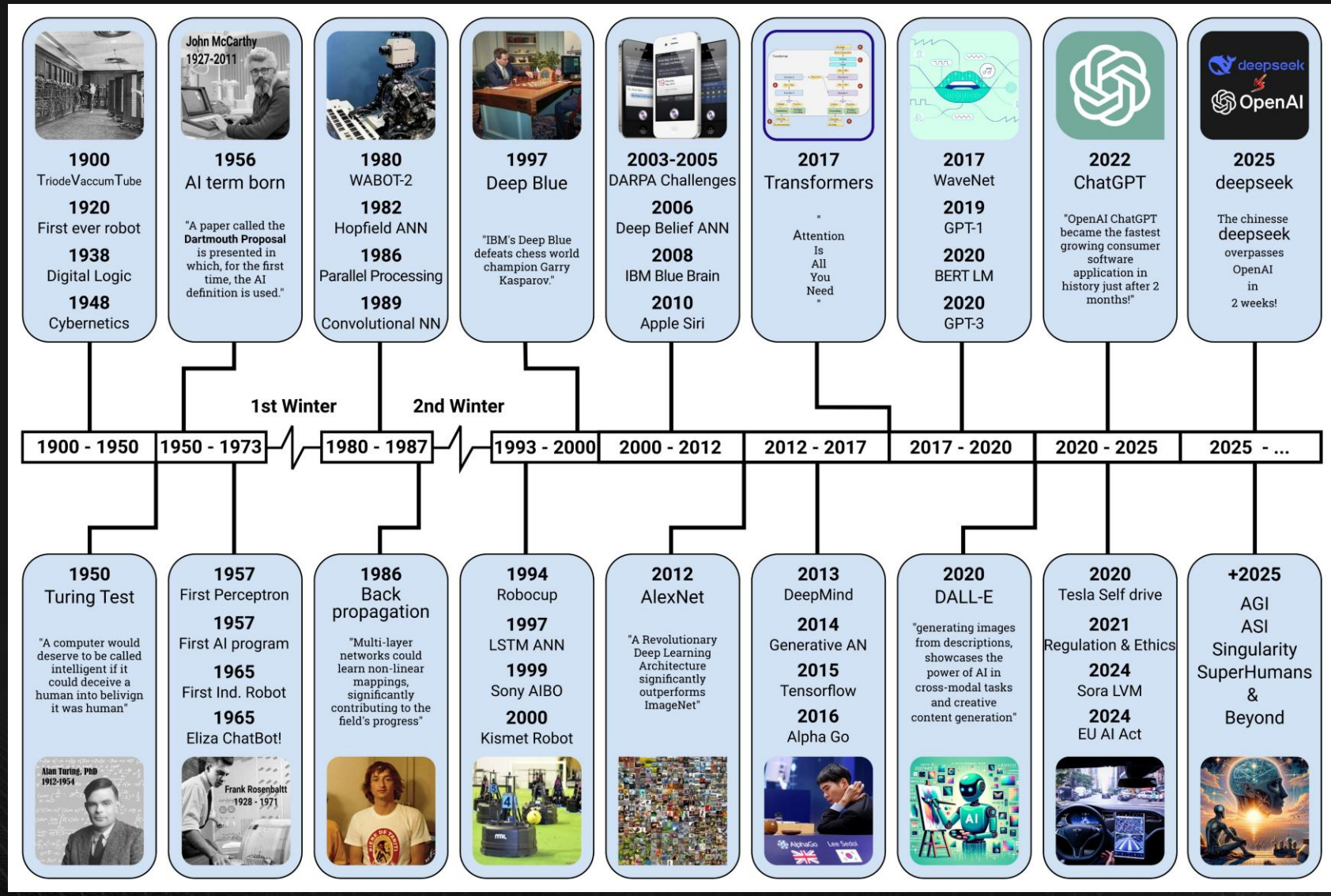
LLM-basierte Agenten



Quelle: https://aws.amazon.com/blogs/industries/bmw-group-develops-a-genai-assistant-to-accelerate-infrastructure-optimization-on-aws/

Übersicht Künstliche Intelligenz

Meilensteine



Mehr Details:
Wiki: Link
Paper: Link

Quelle: Link

Künstliche Intelligenz

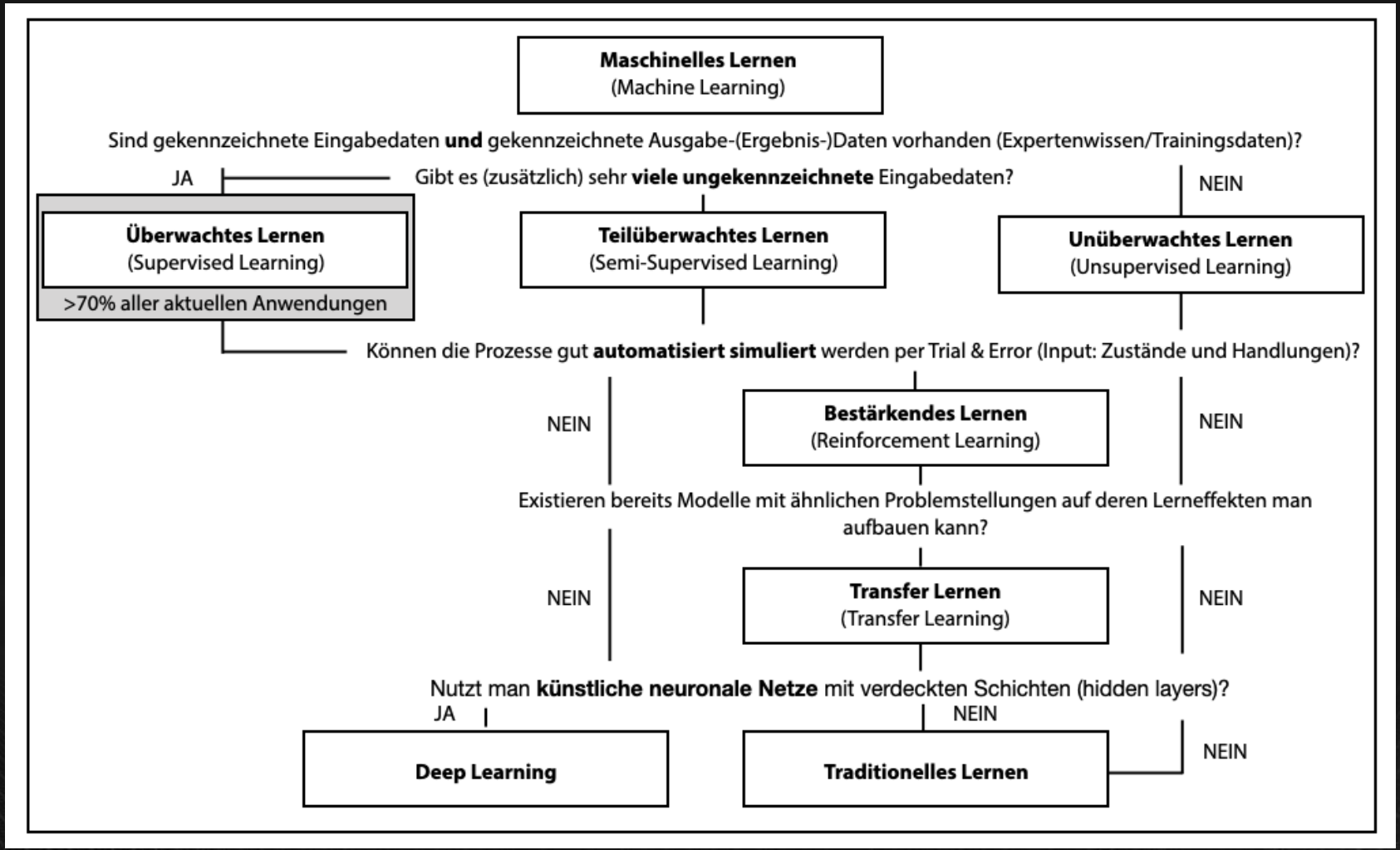
Vier Grundvoraussetzungen

Algorithmen	Rechenleistung	Menschen	Daten

Algorithmen

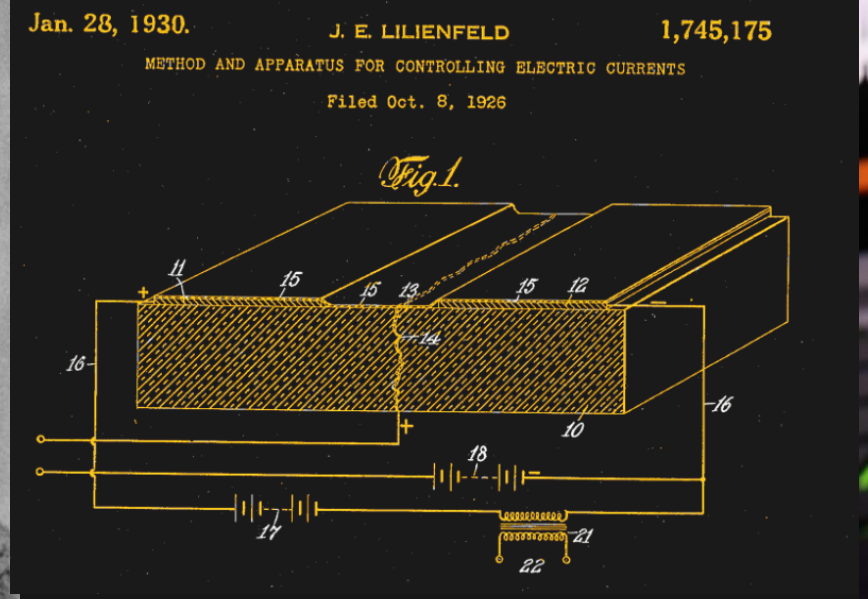
Algorithmen

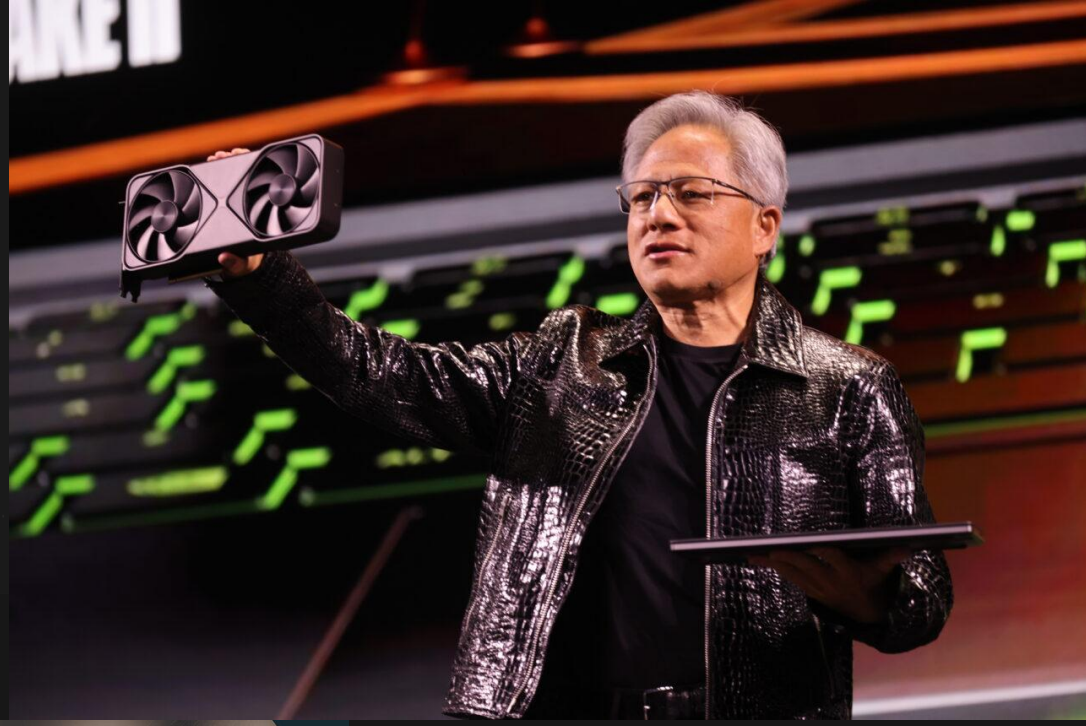
Übersicht wesentlicher Algorithmen

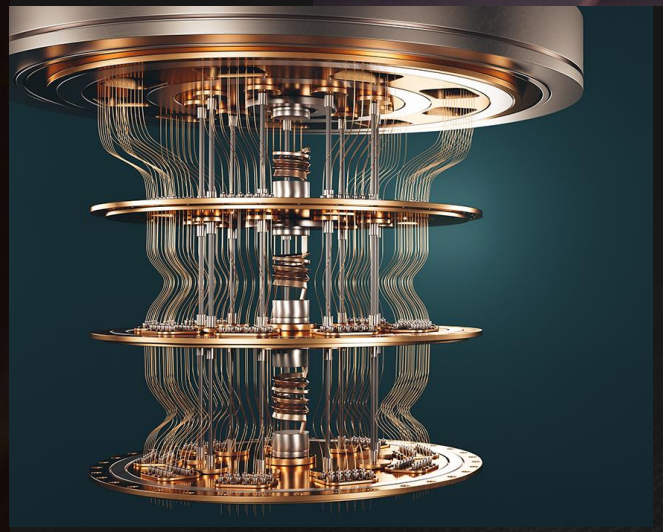


Quelle: Gröner, Heinecke: „Kollege KI"

Rechenleistung

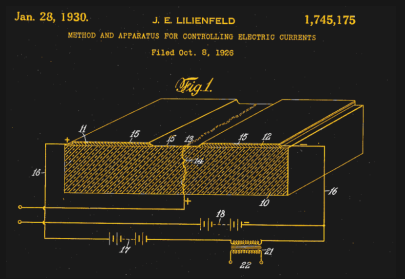




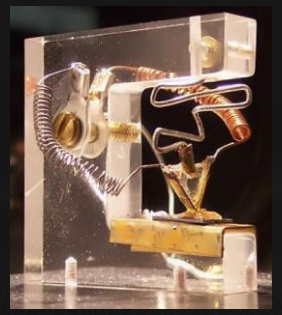


Quellen:
https://de.wikipedia.org/wiki/Julius_Edgar_Lilienfeld,
https://hackaday.com/2018/12/11/julius-lilienfeld-and-the-first-transistor/,
https://de.wikipedia.org/wiki/John_Bardeen,
https://qse.group/bill-gates-on-the-future-of-quantum-computing-are-we-just-3-5-years-away-3/
https://blogs.nvidia.com/blog/ces-2025-jensen-huang/

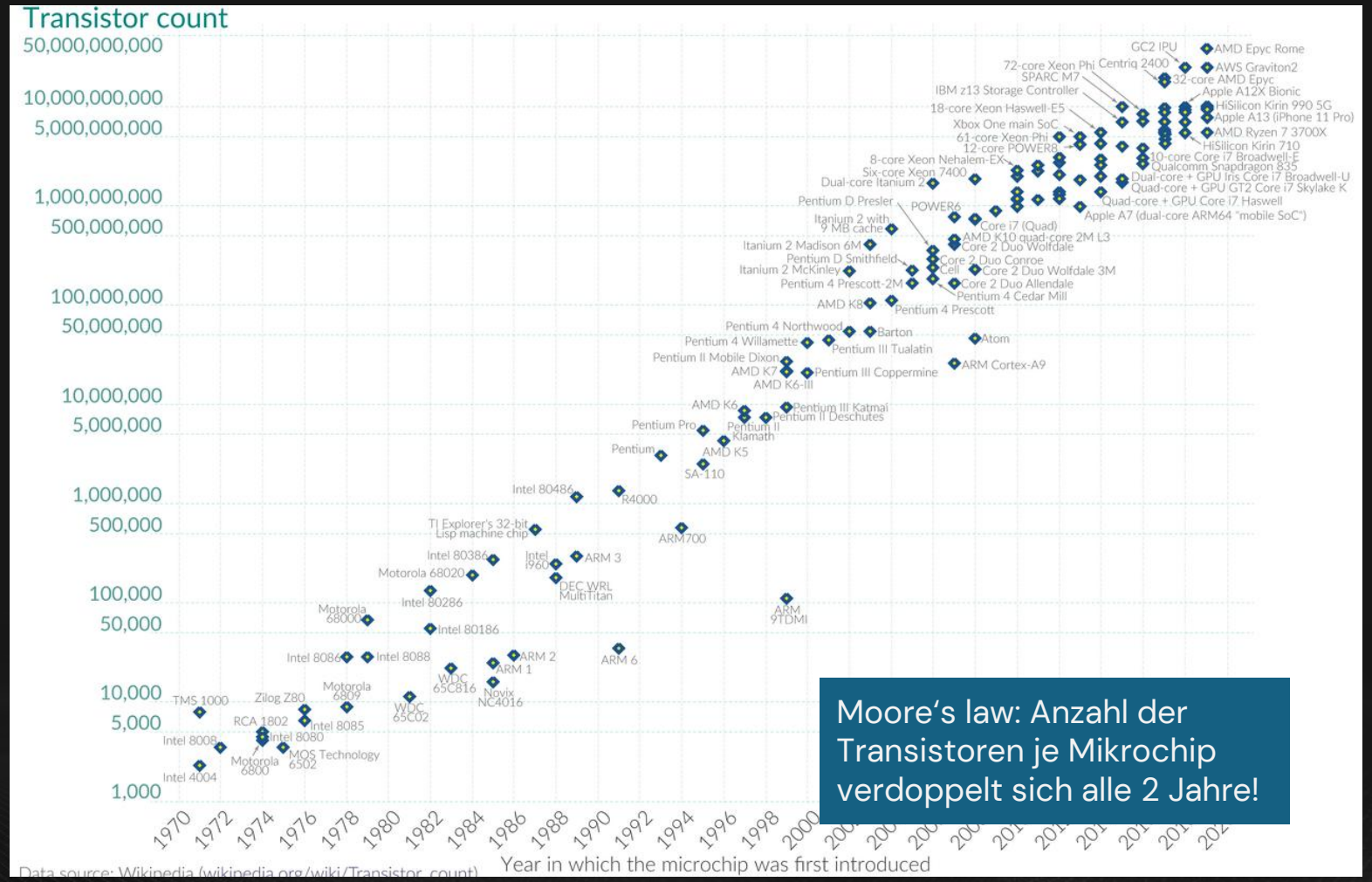
Rechenleistung: exponentielle Steigerung seit 1947



Lilienfeld Patentschrift „Method and apparatus for controlling electric currents", 1926



Nachbau des ersten Bipolartransistors (Shockley, Bardeen, Brattain 1947/48)



Moore's law: Anzahl der Transistoren je Mikrochip verdoppelt sich alle 2 Jahre!

Apple M4 Chip mit ca. 30 Mrd. Transistoren

Nvidia Blackwell mit ca. 208 Mrd. Transistoren für AI-Trainieren

Quellen:
https://hackaday.com/2018/12/11/julius-lilienfeld-and-the-first-transistor/,
https://de.wikipedia.org/wiki/John_Bardeen,
https://de.wikipedia.org/wiki/Mooresches_Gesetz,
https://www.heise.de/news/Apple-M4-TSMCs-FinFlex-hilft-den-Performance-Rechenkernen-9766547.html,
https://www.theinformation.com/articles/top-developers-want-nvidia-blackwell-chips-everyone-else

Menschen

Menschen

Von:

- Mangel an Investitionen
- Digitale Infrastruktur
- Veraltete Lehrpläne
- Fehlender Schwerpunkt auf Technologie und Datenanalyse
- Kultureller Widerstand
- Mangelnde interdisziplinäre Teamarbeit
- Regulatorik….

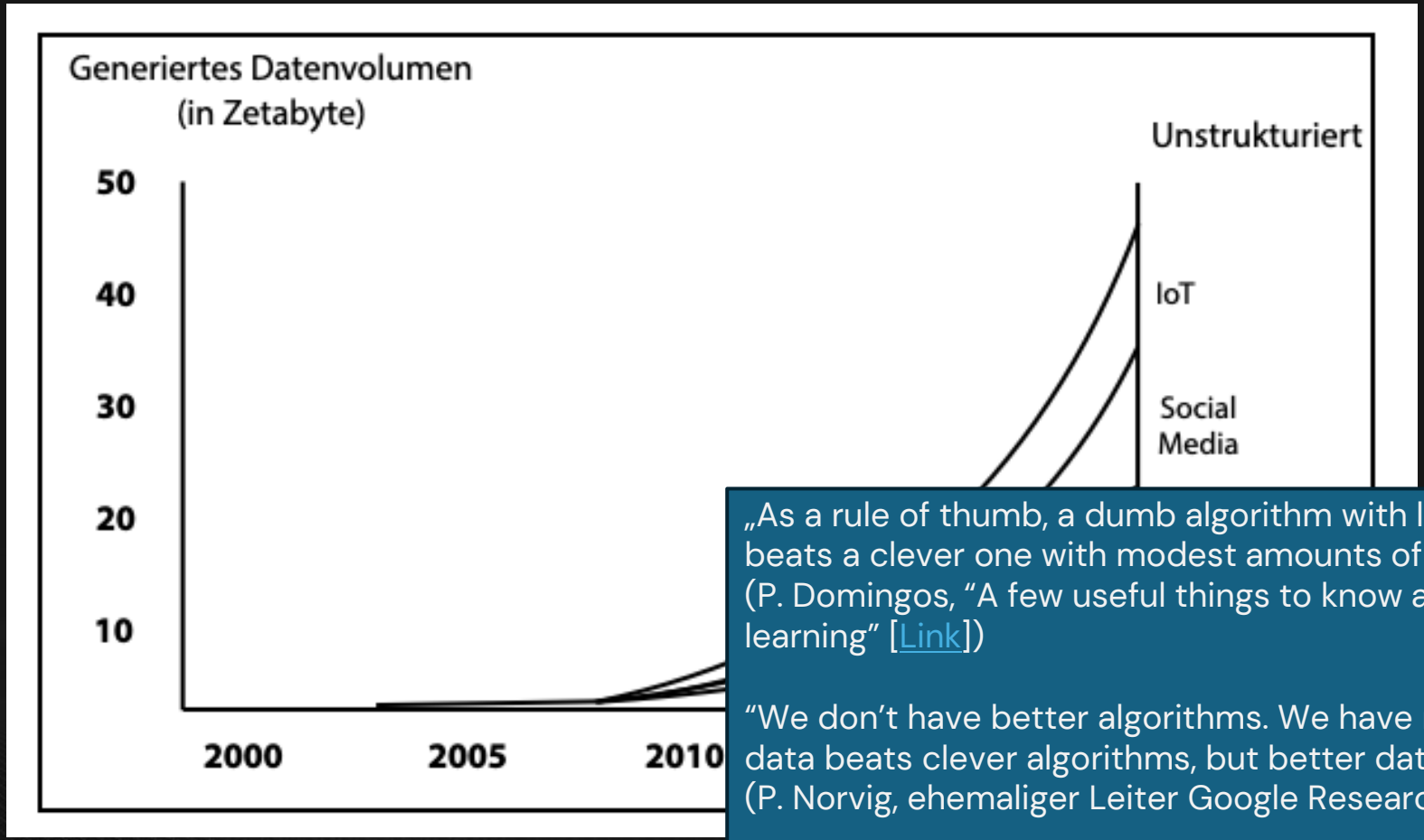
Zu:

- Investitionen in Infrastruktur, Technik, Leute, …
- Aufbau leistungsfähige Infrastruktur
- Aktualisierung und Integration in Lehrpläne
- ….
- Neues, chancenorientiertes Denken!
- Cross-funktionale Teams
- Das wird schwer….

Daten

Daten

Die verfügbare Datenmenge wächst seit Jahren exponentiell



„As a rule of thumb, a dumb algorithm with lots and lots of data beats a clever one with modest amounts of it"
(P. Domingos, "A few useful things to know about machine learning" [Link])

"We don't have better algorithms. We have more data. More data beats clever algorithms, but better data beats more data."
(P. Norvig, ehemaliger Leiter Google Research)

" Data is the new oil" Clive Humby (engl. Mathematiker)

Quelle: Gröner, Heinecke: „Kollege KI"

Daten

Datenbasierte Geschäftsmodelle

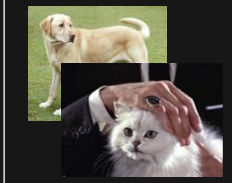
"Uber, the world's largest taxi company, owns no vehicles. Facebook, the world's most popular media owner, creates no content. Alibaba, the most valuable retailer, has no inventory. And Airbnb, the world's largest accommodation provider, owns no real estate. Something interesting is happening." Tom Goodwin (2015)

Geschäftsmodelle heutiger Tech-Firmen basieren auf der Sammlung, Verknüpfung und Auswertung von Daten.
Aufgrund deutlich geringerer Grenzkosten als bei physischen Produkten, ist Ausweitung auf Millionen User sehr einfach möglich.

Daten

Wie erhalten diese Firmen Ihre Daten?

Rohdaten



Datenerfassung und –management/ Extract, Transform, Load (ETL)

Discover, Update or Initiate (Trigger) → Ingest/ Store raw data → Process/ Enrich Data → Store prepared data

Veröffentlichte Daten/ System

Mainframes/ Server

Datenbanken, Data Warehouses

MySQL, ORACLE, mongoDB

Cloud Systeme

Spark, APACHE hadoop, Azure, aws, Google Cloud

Daten analysieren, Erkenntnisse generieren und visualisieren

Datenanfrage/ Query (SQL, Spark, Python, …)

Datenbereitstellung (Tabellen, CSV, JSON, …)

Inhalt der nächsten Vorlesungen ☺

Quellen: Workflow basierend auf Google Dataflow, abgerufen am 07.09.2020 unter: Link sowie Link
Bilder: Stockpictures sowie Homepages der jeweiligen Hersteller.

Datenbasierte Geschäftsmodelle

Business Canvas

Key Partners

- Who are our Key Partners?
- Who are our key suppliers?
- Which Key Resources are we acquiring from partners?
- Which Key Activities do partners perform?

Key Activities

- What Key Activities do our Value Propositions require?
- Our Distribution Channels?
- Customer Relationships?
- Revenue streams?

Key Resources

- What Key Resources do our Value Propositions require?
- Our Distribution Channels?
- Customer Relationships?
- Revenue Streams?

Value Proposition

- What value do we deliver to the customer?
- Which one of our customer's problems are we helping to solve?
- What bundles of products and services are we offering to each Customer Segment?
- Which customer needs are we satisfying?

Customer Relationships

- What type of relationship does each of our Customer Segments expect us to establish and maintain with them?
- Which ones have we established?
- How are they integrated with the rest of our business model?
- How costly are they?

Channels

- Through which Channels

Customer Segments

- For whom are we creating value?
- Who are our most important customers?

Cost Structure

- What are the most important costs inherent in our business model?
- Which Key Resources are most expensive?
- Which Key Activities are most expensive?

Revenue Streams

- For what value are our customers really willing to pay?
- For what do they currently pay?
- How are they currently paying? How would they prefer to pay?
- How much does each Revenue Stream contribute to overall revenues?

Fragen für die Einbindung von Daten:
- Wie funktioniert das datenbasierte Geschäftsmodell genau?
- Wie generiert die Firma (mit Daten) Value für den Kunden?
- Wie generiert die Firma (mit Daten) Einkommen?
- Welche Daten werden hierfür aus Ihrer Sicht benötigt?
- Welche Kriterien für Datenqualität sind wichtig?

→ Erster Schritt Richtung Durchführung Data Science/ AI Use Case

Business Canvas[1] ist Bestandteil der Lean Startup Methode[2] und wird häufig im Umfeld Startups eingesetzt.

Quelle: 1 Osterwalder, Pigneur & al. 2010 (Link Template)
2 Eric Ries, „The Lean Startup", 2011

Case study data-intensive distributed system am Beispiel Instagram

Was ist ein data-intensive distributed system?

Begriffsklärung

- Data-intensive: Systeme, die große Datenmengen speichern, verarbeiten und verwalten (Daten sind die Hauptlast des Systems, nicht Rechenleistung).
- Distributed: System besteht aus vielen einzelnen Komponenten an verschiedenen Orten, die miteinander über Nachrichten interagieren.

- Wir schauen uns so ein Gesamtsystem näher anhand des Fallbeispiels Instagram an.
- Dabei fokussieren wir auf das Design des Gesamtsystems und der Datenperspektive
- Wir gehen vor wie im „richtigen Leben“ solche Systeme designt werden, aber aufgrund Zeit machen wir keinen Deep-Dive.

In der Vorlesung befassen wir uns mit bereits vorhandenen Daten.
Es ist aber interessant zu sehen, wie solche großen Systeme entwickelt werden.

Case study Instagram

Schritt 1: Klärung Betrachtungsumfang

- **Funktionale Anforderungen:**
  - Bilder/ Videos hochladen, anschauen, liken und kommentieren
  - Anderen Usern folgen
  - Newsfeed
  - **Datenanalyse – und auswertung** (im Hintergrund)
  - Monetarisierung/ Werbung
  - Nicht im Fokus: User-/ Account-Management

- **Nicht-funktionale Anforderungen:**
  - Hohe Verfügbarkeit
  - Geringe Latenz (Wartezeit)
  - Hohe Verläßlichkeit (Daten gehen nicht verloren)

Case study Instagram

Schritt 2: Abschätzung Last – was muß das System können?

Speicherplatz

- Wie viele Daten kommen pro Tag/ Jahr zusammen?
- Wie lange sollen die Daten gespeichert werden?

Abschätzung:

- 1 Mrd. aktive Anwender *
- Upload 3 Photos pro Tag je User *
- 300 KB Größe je Bild *

= 10 MB je Sekunde = 900 GB pro Tag = 328 TB p.a.

Leselast

Wie viele Daten werden je Sekunde abgerufen?

Abschätzung:

- 1 Mrd. aktive Anwender *
- 10 Bilder pro Tag je User abgerufen *
- 300 KB Bildgröße

= 3 PB pro Tag und ~34 GB je Sekunde

Dies sind Annahmen, die initial getroffen werden und regelmäßig überprüft werden; darauf basierend wird das System angepaßt.

Case study Instagram

Schritt 3: Schnittstellen zu anderen Systemen und möglicher Aufbau

Programmierschnittstellen (API)

Daten speichern (POST API):

- Bilder hochladen: Funktion UploadImage(userID, Image, ImageCaption, ImageLocation, Tags, Comments)
- Bilder liken: LikeImage(UserID, PhotoID)
- Bilder kommentieren: Comment(UserID, photoID, comment)
- Anderem User folgen: FollowUser(UserID, OtherUserID)

Daten erhalten (GET API):

- Bilder anschauen: ViewImage(myUserId, PhotoID)
- News Feed: GetFeed(myUserID, FollowerID)
- Werbung ausspielen: GetAdverts(userID, LinkToSpot)

Datenstruktur

Name der Tabelle: User	
Primary Key	UserID: integer
	Name: string Email: string Location: string LastActive: time

Name der Tabelle: Photos	
Primary Key	PhotoID: integer
Foreign key	UserID: integer
	Caption: string Location: string Path: string Tags: string Comments: string

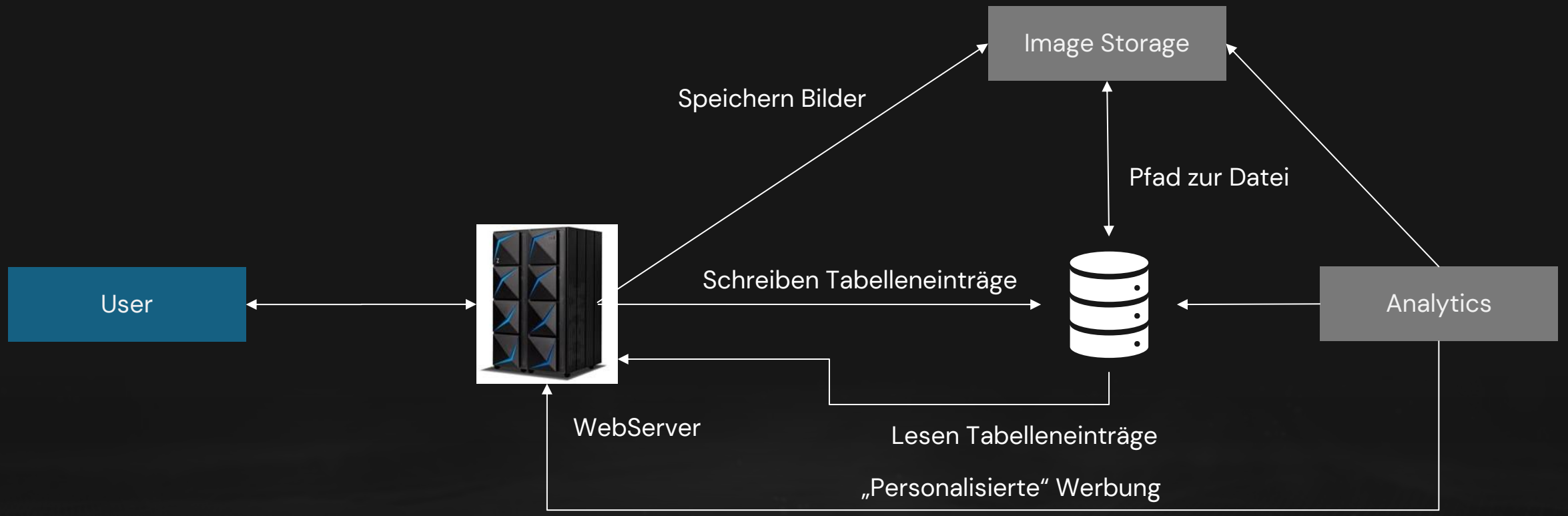
Name Tabelle: Follower	
Foreign key	UserID: integer
Foreign key	FollowerID: integer

Welche weiteren Daten wären interessant, um mit Werbung (oder anderen Diensten) Geld zu verdienen?

Schnittstellen und Datenstruktur werden im Laufe des Lebenszyklus kontinuierlich angepaßt.

Case study Instagram

Schritt 4: High Level-/ Grob-Design



Dies ist eine grundlegende Architektur, die aber nicht für eine hohe Anzahl User und Last geeignet ist.
Grund: wir haben verschiedene Flaschenhälse, die paralleles Abarbeiten Aufträge und Anfragen verhindert.

Case study Instagram

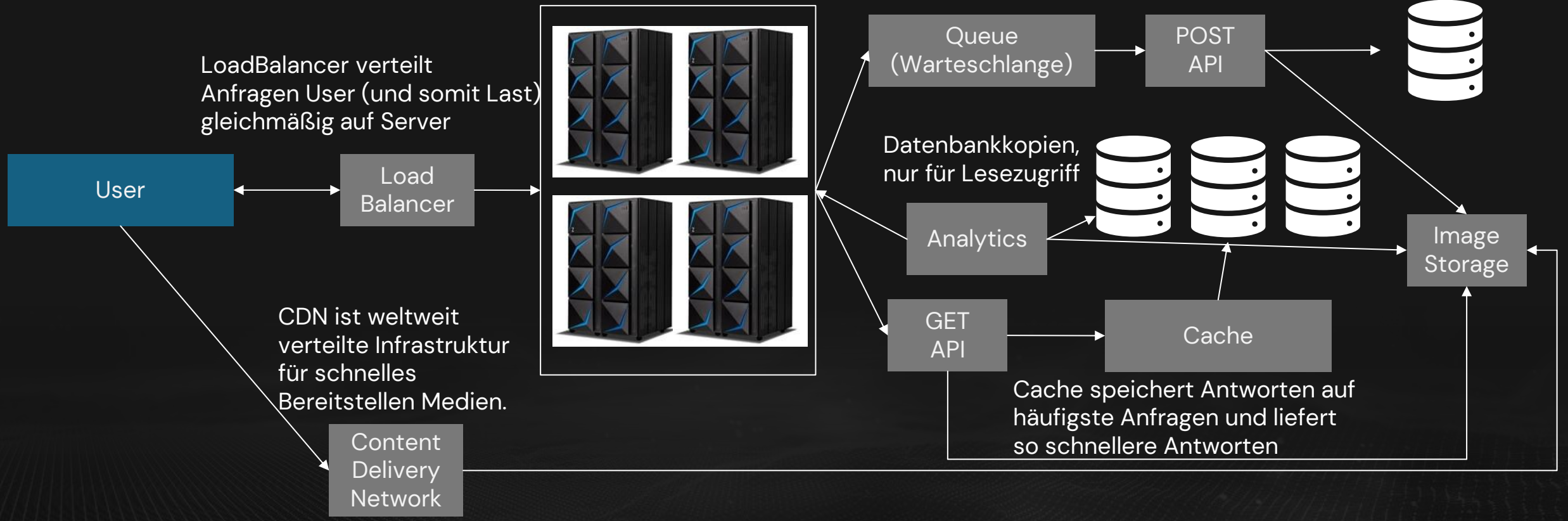
Schritt 5: Skalierbares Design

weltweit verteilte **WebServer** für geringe Latenz
bei der Abarbeitung User-Anfragen.
Anzahl wird automatisiert an Last angepasst.

Warteschlange ermöglicht
Entlasten beim Schreiben
(schreiben wenn geringere Last)

NoSQL-Datenbank
(schreiboptimiert)

LoadBalancer verteilt
Anfragen User (und somit Last)
gleichmäßig auf Server



Dies ist nur eine beispielhafte Lösung und unterscheidet sich je nach Fokus auf Verfügbarkeit, Latenz, Kosten, ...

Backup