

GRUNDLAGEN

Joern Ploennigs · AI4SC

ÜBERBLICK

© Midjourney: Application of ML in Construction



ZIELSETZUNG

- Praktische Kenntnisse und Fähigkeiten in der Anwendung von Data Science zur Analyse von Daten
- Verständnis und Kenntnisse zu den Grundlagen von einfachen Modellen des Maschinellen Lernens

© Midjourney: Saving the environment with machine learning



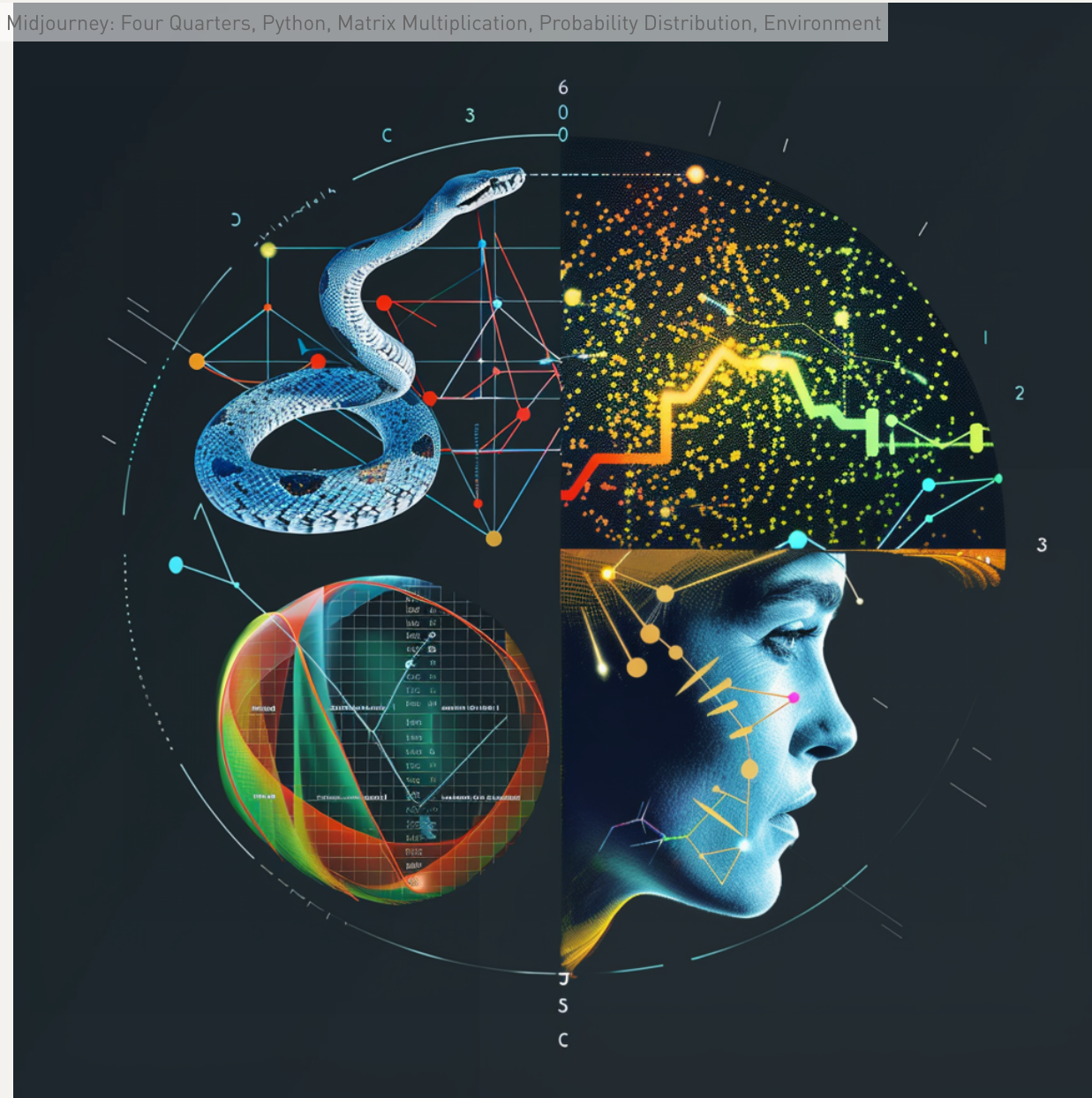
VORAUSSETZUNG

Grundlagenkenntnisse in:

- Python
- Lineare Algebra (Vektoren, Matrizen, Gleichungssysteme)
- Statistik (Zufallszahlen, Verteilungen)
- Herausforderungen der Domäne und Gesellschaft (Klimawandel, Bauwirtschaft, Technische Mechanik)

Keine umfangreichen Kenntnisse erforderlich, da ML vor allem das Anwenden von existierenden Methoden ist

© Midjourney: Four Quarters, Python, Matrix Multiplication, Probability Distribution, Environment



GRUNDLAGEN

- Programmierung
- Lineare Algebra
- Statistik
- Data cleaning

SUPERVISED LEARNING

- Regression
- Klassifikation

UNSUPERVISED LEARNING

- Clustering
- Dimensionsreduktion

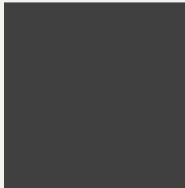
NEURONAL & REINFORCEMENT LEARNING

- Neuronale Netzwerke
- Reinforcement Learning

HERAUSFORDERUNGEN

- Skalierbarkeit
- Datenschutz
- Fairness
- Ethik

GRUNDLAGEN



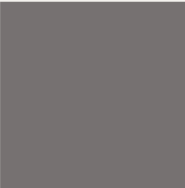
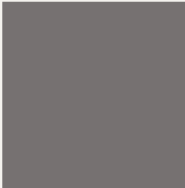
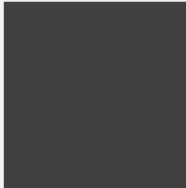
SUPERVISED LEARNING



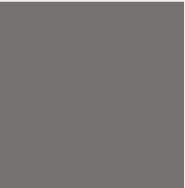
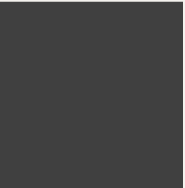
UNSUPERVISED LEARNING



NEURONAL & REINFORCEMENT LEARNING



HERAUSFORDERUNGEN



Verständnis für
Maschinelles
Lernen



VORLESUNG

Die Vorlesung erklärt
das zugrundeliegende
theoretische und
technische Wissen



ÜBUNG

Die Übungen
vermitteln praktische
Fähigkeiten zur
Anwendung an
typischen Datensätzen
aus der Praxis

Studienbegleitendes Portal für Präsenzlehre der Universität Rostock

What are you looking for?

Courses

44620 Vorlesung/Übung: Grundlagen Maschinellen Lernens (SS 2024)

Overview Administration Forum Participants **Files** Schedule Information Wiki Calendar Blubber More ...

Vorlesung/Übung: Grundlagen Maschinellen Lernens

Type	Name ^	Size	Downloads	Author	Date	Actions
	01b_Anwendungen.pdf	7.2 MB	0	Plönnigs, Jörn	Now	
	01c_Modelltypen.pdf	27.5 MB	0	Plönnigs, Jörn	Now	

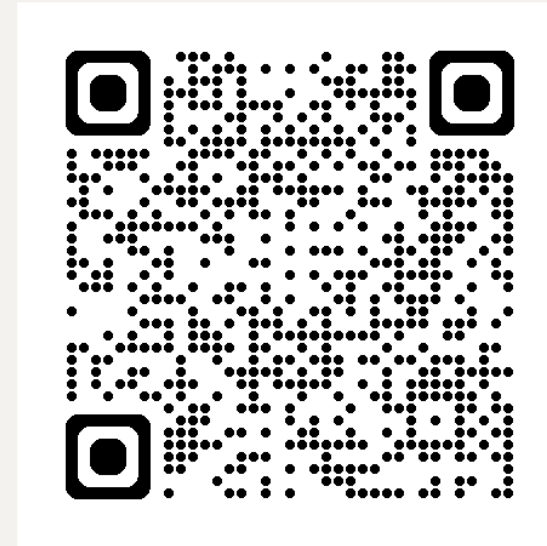
Download Move Copy Delete New folder Add document

Upload files

Add files via drag and drop to this area.

Views

- Folder view
- All documents



<https://studip.uni-rostock.de>

AI4SUSTAINABLECONSTRUCTION

Einführung

Grundlagen

Grundlagen

Computerhardware

Softwarearchitekturen

Programmiersprachen

Wissenspyramide

Quiz - Wissenspyramide

Datentypen

Operatoren

Quizz: Boolesche Operationen

Verzweigungen

Schleifen

Modularisierung

Funktionen

Objekte

Rekursive Funktionen

Programmablauf

Exceptions

Unit-Tests

Debugging

Pakete

Datenbanken

Umgang mit Dateien

Einführung

In diesem Kurs steht die Aus- und Weiterbildung von Informatikfähigkeiten speziell für den Anwendungsbereich der Bau- und Umweltingenieurwissenschaften im Fokus. Durch die Digitalisierung benötigen Ingenieure und andere technisch Interessierte in diesen Branchen immer häufiger grundlegende Fähigkeiten in der Programmierung, speziell im Bereich des Datenmanagements, -verarbeitung und -analyse. In den Bau- und Umweltingenieurwissenschaften handelt es sich hier z.B. um Datensätze aus Sensornetzwerken, Planungsprozessen oder Gebäudedatenbanken. Sehr gebräuchlich für die Arbeit mit solchen Daten ist die Programmiersprache Python mit ihren vielfältigen Programmbibliotheken. Das Bildungsangebot führt daher in die Grundlagen der Programmierung und den Umgang mit Datenbanken und Analysetools in Python ein. Es werden dabei zusätzlich theoretische Informatik-Grundlagen sowie fachspezifische Entwicklungsaspekte an verschiedenen Beispielen vermittelt (z.B. der Umgang mit räumlichen Daten).

GRUNDLAGEN

MotivationComputer und ArchitekturenProgrammierung und DatentypenVerzweigungen und Schleifen

MOELLIERUNG

Fehler und DebuggingObjektorientierung u. SoftwareentwurfFunktionen und Rekursion

DATENBANKEN

DatenhaltungDatenbankenDatenbankentwurfTrends und KICAD und BIM im 2. Semester

Der Kurs gliedert sich in drei Teile:

1. Grundlagen:

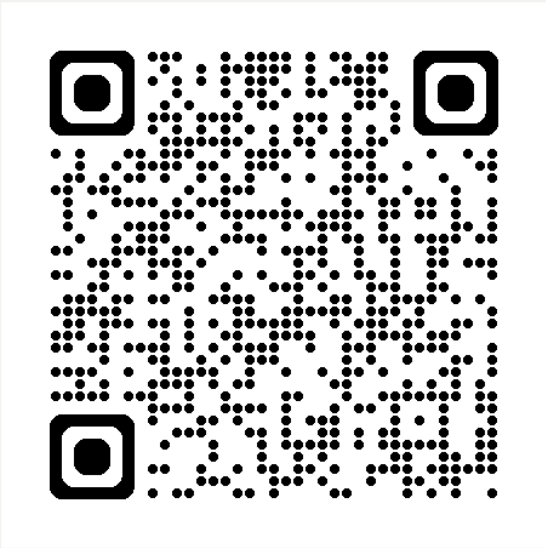
Bietet eine Einführung in Hard- und Software von Computern und in die Programmierung in Python.

2. Modellierung:

Thematisiert die Modellierung von Problemen mit Funktionen und Objekten und die Lösung von Problemen durch Debugging

3. Datenbanken:

Gibt einen Überblick in das Management kleiner und größerer Daten durch Dateien und Datenbanken.



<https://ml-book.ai4sc-lectures.auf.uni-rostock.de>

File Edit View Run Kernel Nbgrader Tabs Settings Help

+

+

+

+

Filter files by name

/ Lectures /

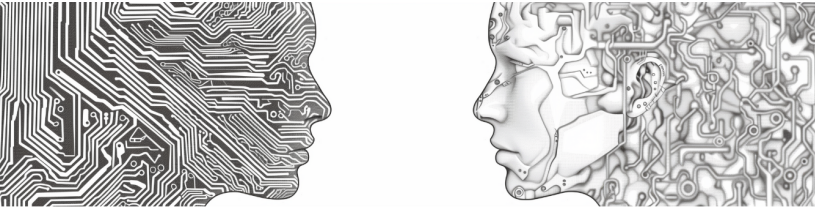
Name	Last Modified	File Size
ideas	yesterday	
images	21 hours ago	
quadro	2 minutes ago	
00_Grundi...	19 hours ago	18.4 KB
01_Jupyter...	17 hours ago	8.9 KB
✓ 02_Python...	17 hours ago	90.1 KB
03_Numpy...	17 hours ago	59.4 KB
04_Pandas...	17 hours ago	85 KB
05_Descri...	17 hours ago	129.9 KB
06_Descri...	17 hours ago	6 MB
07_Data_C...	17 hours ago	10.1 MB
08_Visual_...	17 hours ago	9.4 MB
Definitone...	yesterday	4.8 KB
header.html	21 hours ago	598 B
rise.css	19 hours ago	802 B

Launcher

00_Grundlagen.ipynb

Notebook Python 3 (ipykernel)

Grundlagen



Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last.

— Stephen Hawking

Machine Learning

Maschinelles Lernen (ML) ist ein Teilbereich der Künstlichen Intelligenz, der sich auf die Entwicklung von Algorithmen und Techniken konzentriert, die es Computern ermöglichen, aus Erfahrung zu lernen, ohne explizit programmiert zu werden. ML-Algorithmen analysieren Daten, erkennen Muster und treffen aufgrund dieser Muster Vorhersagen oder treffen Entscheidungen. Während KI ein breiteres Konzept ist, das sich auf die Idee von künstlicher Intelligenz im Allgemeinen bezieht, konzentriert sich maschinelles Lernen speziell darauf, wie Computer lernen können, aus Daten zu lernen und Probleme zu lösen. Wir definieren es als:

Definition: Maschinelles-Lernen

Maschinelles Lernen (ML) ist ein Teilgebiet der Künstlichen Intelligenz, das sich mit der Entwicklung und Untersuchung statistischer Algorithmen befasst, die aus Daten lernen und auf ungesehene Daten verallgemeinern und so Aufgaben ohne explizite Anweisungen ausführen können.



<https://ml-lab.ai4sc-lectures.auf.uni-rostock.de>



 jupyterhub Login

Sign In

Username:

Password:

Sign In

[Sign up](#) to create a new user.

Auf der Webseite auf "Sign up" klicken

Sign Up

Username:

nur_mit_kleinbuchstaben

Password:

Confirm password:

Create User

[Login](#) with an existing user.

Nutze nur **Kleinbuchstaben** im Nutzernamen

Write new message

To
Prof. Dr.-Ing. habil. Jörn Plönnigs

Search user Add more recipients

Subject
[44620: Grundlagen Maschinellen Lernens]

Message

Normal ... A- A+ [Icons]

Hi,

mein Nutzernamen ist: nur_mit_kleinbuchstaben

MfG,

Max Musterl

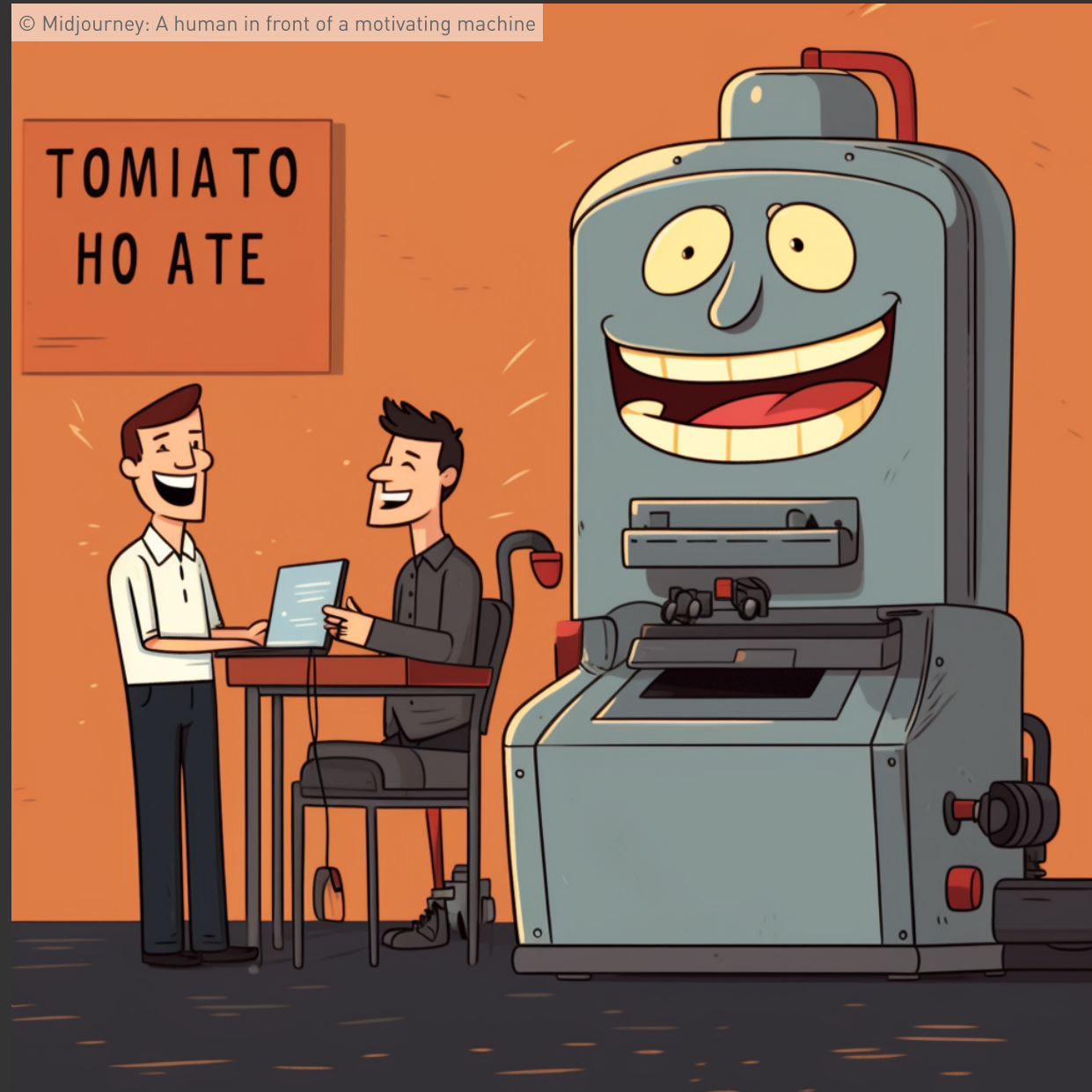
Attachments Labels Options

Submit Cancel

Schickt mir eine Nachricht mit eurem Nutzernamen

MOTIVATION

© Midjourney: A human in front of a motivating machine

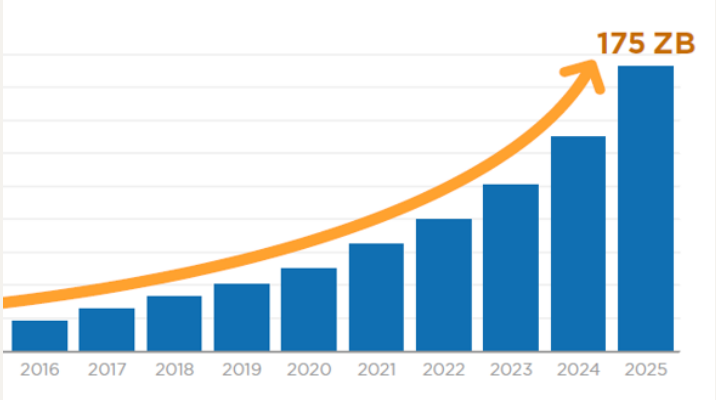


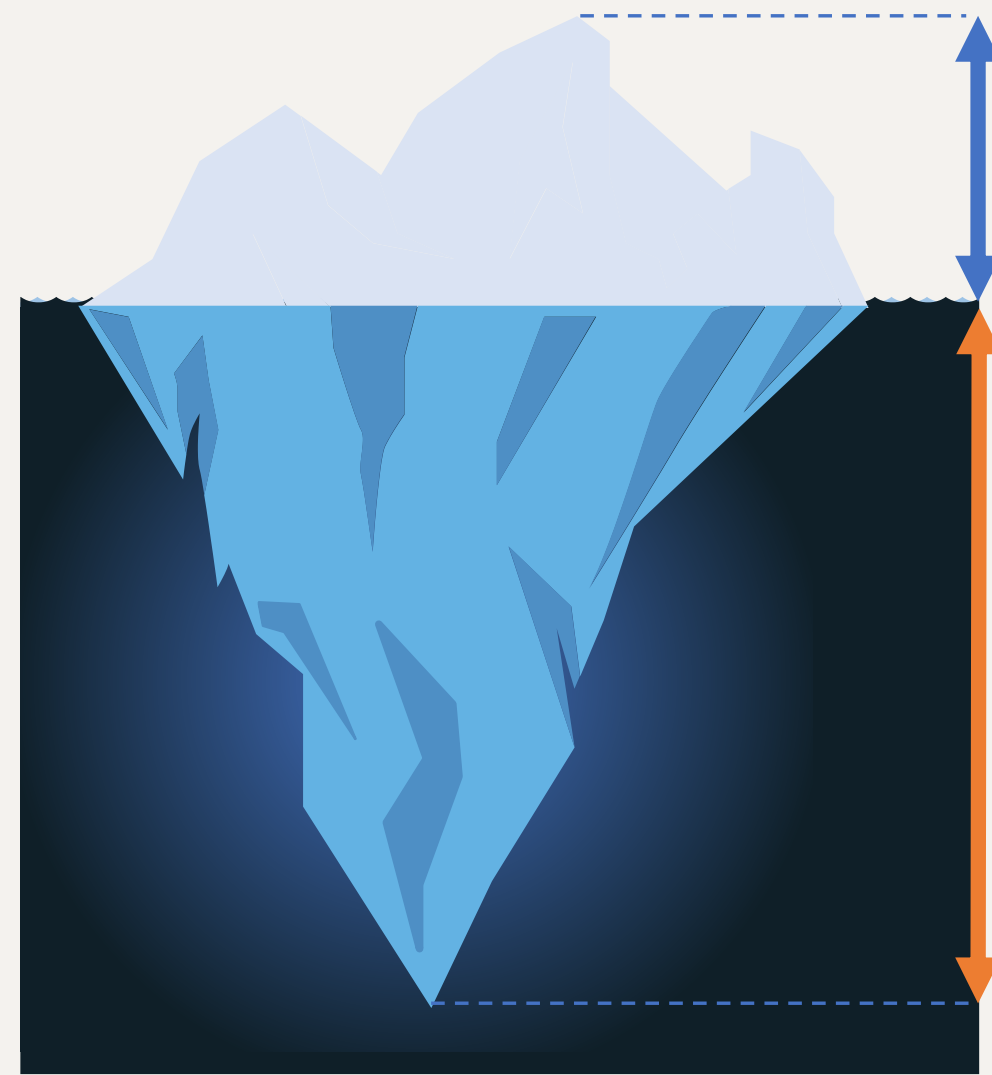
DATEN WACHSTUM

Die Größe der erzeugten Daten wächst zunehmend.

Sie werden in Zukunft vor sehr vielen Daten konfrontiert werden.

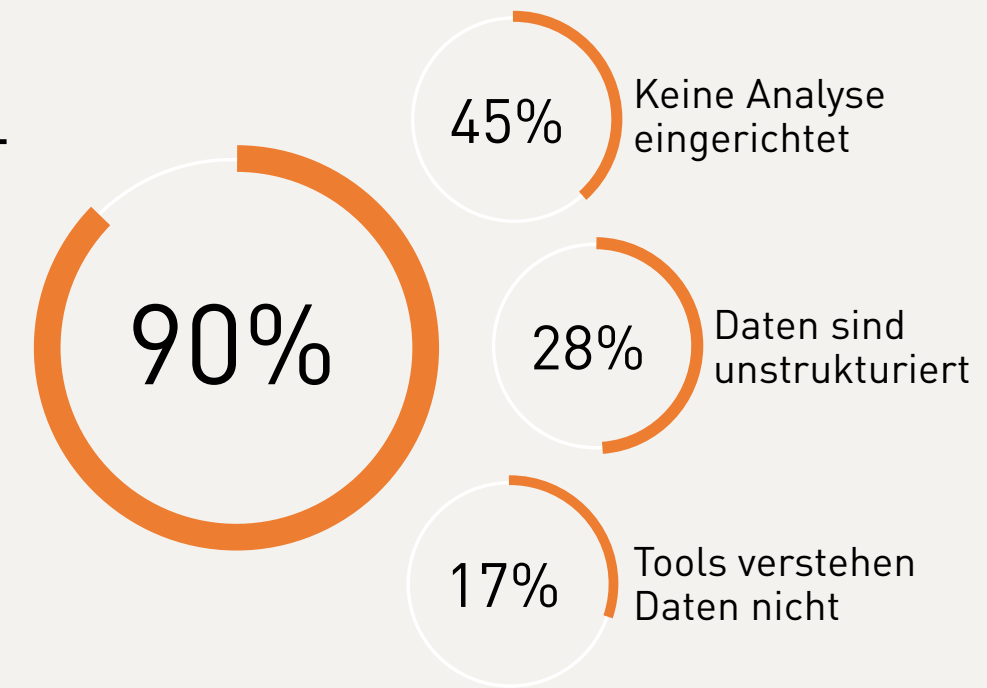
Computer müssen Ihnen helfen, diese Daten effizient zu analysieren.





BENUTZT

UNGENUTZT



HÖRSAALFRAGE

Wie viele Sensoren hat ein Supermarkt?

© Midjourney: An alien supermarket filled with exotic food



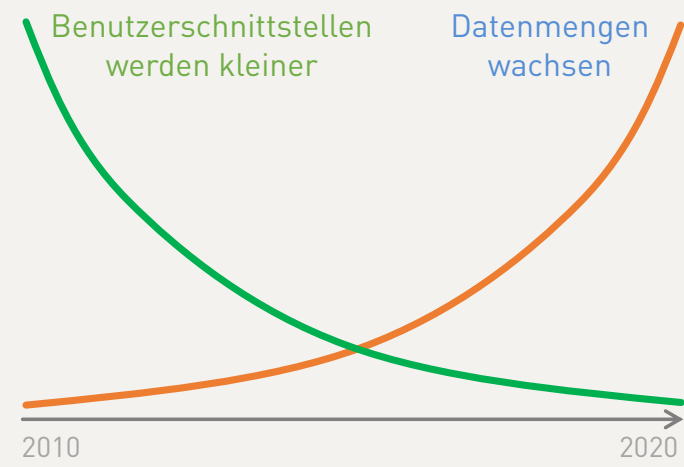
BEISPIEL – TESCO IRLAND

- Die Supermarktkette Tesco hat sehr früh in Sensoren und Monitoringsysteme der Supermärkte mit dem Ziel Energie einzusparen investiert
- Die gesammelten Datenmengen wurden so schnell so groß, dass niemand sie manuell analysieren konnte
- Durch maschinelle Lernmodelle konnte IBM Research die Daten analysieren und Tesco helfen, 20% des Kühlbedarfs einzusparen



DER BEDARF AN KI

- Die Digitalisierung erzeugt immer mehr Daten
- Gleichzeitig werden unsere Bildschirme mobil und kleiner
- KI-Algorithmen sind der wichtige Schritt um die Daten auf das Wesentliche zusammenzufassen



38% der Treibhausgase
aus Baugewerbe

78% der Firmen haben
Fachkräftemangel

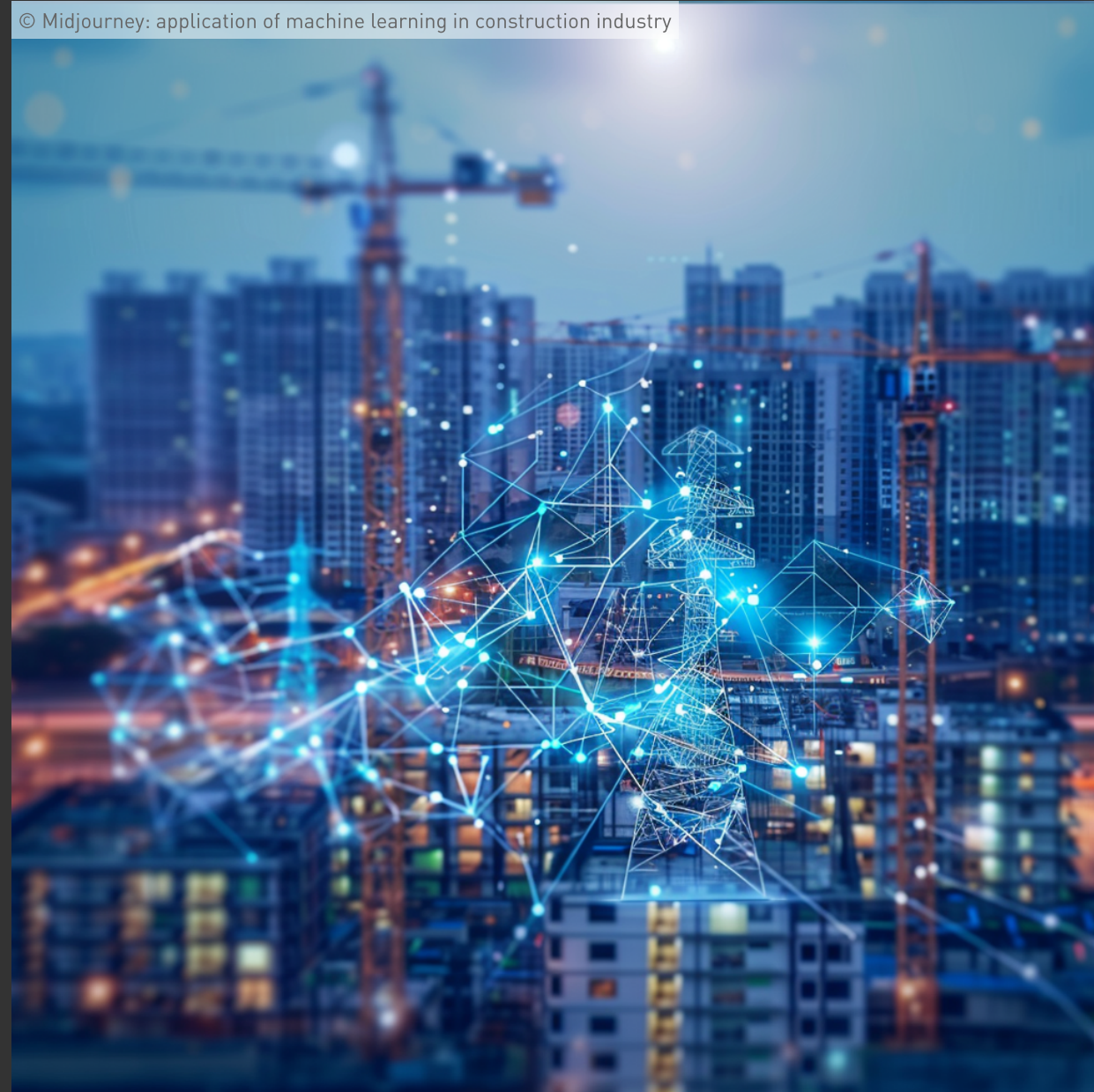
6% Der Firmen nutzen
Digitale Modelle



Digitalisierung
im Baugewerbe
ist dringend
notwendig

ANWENDUNG

© Midjourney: application of machine learning in construction industry



HÖRSAALFRAGE

Wo finden wir ML im Alltag?

© Midjourney: Every day life with machine learning



Onlinesuche

- Suchmaschinen Ranking
- Karten (Verkehrsvorhersage)
- Videoempfehlungen



Soziale Medien

- Freundschaftsempfehlungen
- Nachrichtenfilter
- Onlinespiele



Einkaufen

- Werbung
- Produktempfehlungen
- Betrugserkennung
- Kreditrisikobewertung



Kommunikation

- Spamfilter
- Spracherkennung
- Übersetzungen
- ChatGPT



Wo wird spezifische KI im Bauwesen Anwendung finden?

Routine Verwaltungsaufgaben

- Ticketmanagement
- Zeitplanung
- Raumplanung
- Assetmanagement
- Beschaffung
- Schriftverkehr



Roboter

- Sicherheitsüberwachung
- Reinigung
- Transport
- Vermessung



Routine Monitoring und Wartung

- Datenanalyse und Reporting
- Anomalieerkennung
- Verbrauchsvorhersage
- Regelungsoptimierung
- Predictive Maintenance
- Handlungsanweisungen (Reparaturen)



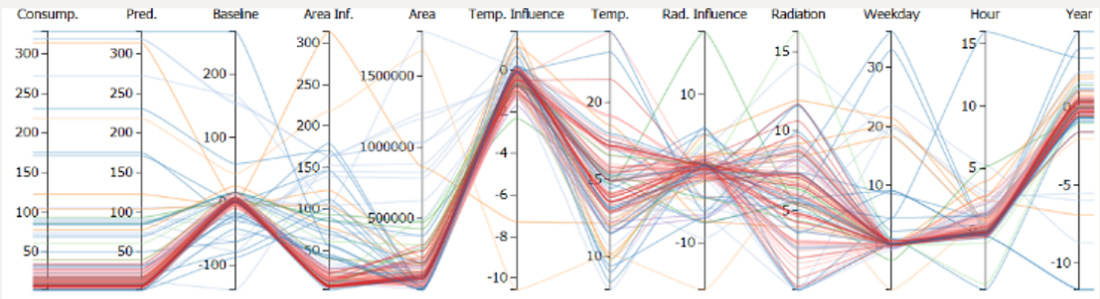
Kreative Aufgaben

- Entwurf
- Planung



GEBÄUDEENERGIEMANAGEMENT: DAS SKALIERUNGSPROBLEM

- Für multinationale Unternehmen erfolgt das Gebäudemanagement in globalem Maßstab
- Ein Gebäude liefert oft mehrere Tausend Datenpunkte
- Die Lösungen müssen für mehrere Standorte und weltweit verteilte Gebäude skalierbar sein
- Große Datenmengen, die integriert und verarbeitet werden müssen
- Die Fähigkeit, unstrukturierte Daten zu analysieren
- Die Analyse von Daten an der Schnittstelle ist entscheidend



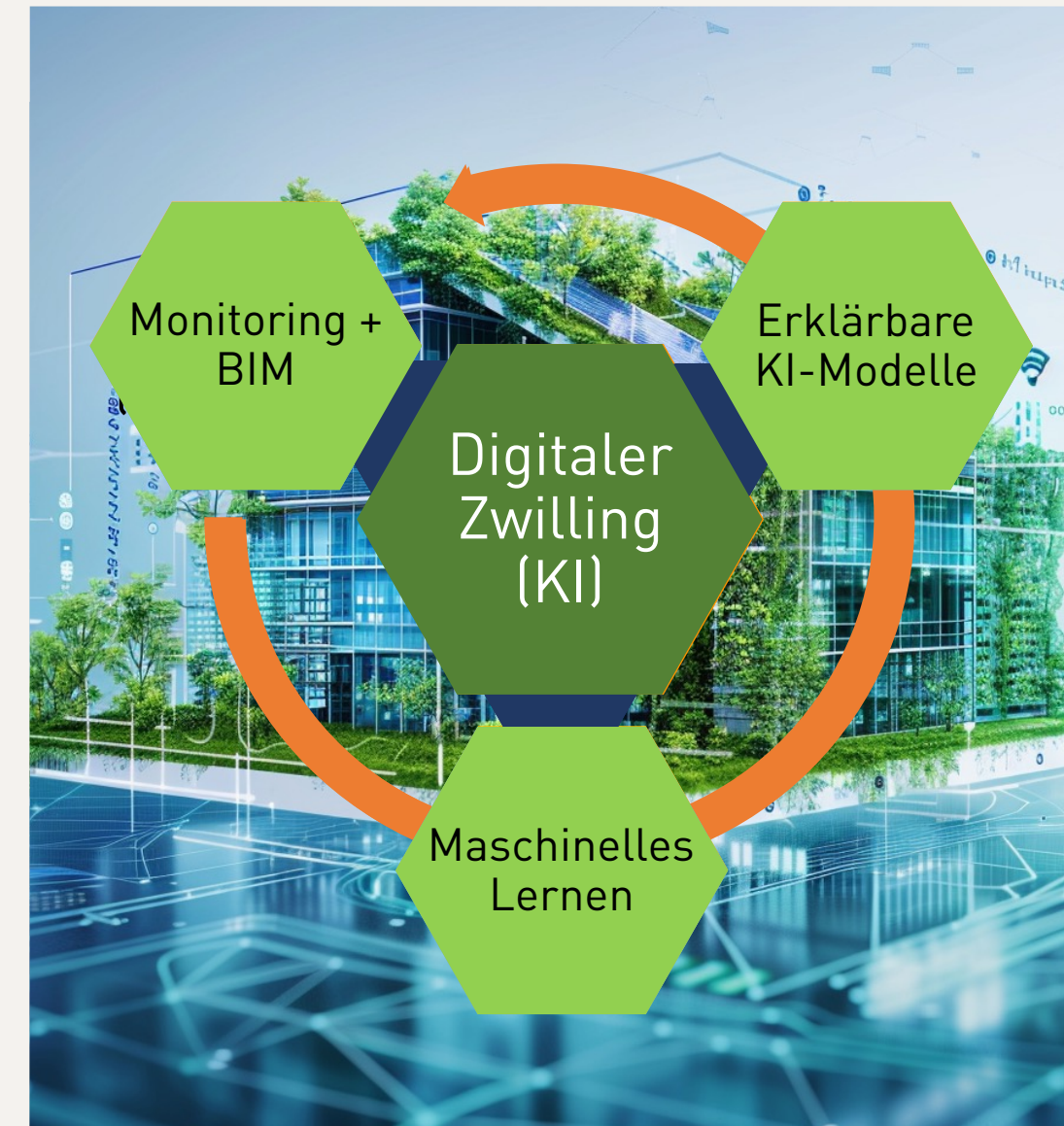
Site ID	Consump.	Pred.	Baseline	Area Inf.	Area	Temp. Influence	Temp.	Rad. Influence	Radiation	Weekday	Hour	Year	
766	327.6	327.6	266.2	55.50	1.226e+4	1.137	9.033	5.594	2.842e-14	4.461	15.97	3.166	Shopping Center/Shopping
45	318.6	318.6	153.0	164.6	1.200e+6	-0.4459	17.18	1.897	5.069	19.85	3.151	-5.957	Food Processing
10	318.2	318.2	156.1	163.8	1.030e+6	0.8632	13.93	-9.399	7.091	23.61	5.221	-8.666	Manufacturing
716	313.6	313.6	-0.9890	313.6	8.603e+5	-7.296	12.18	1.706	9.222	20.34	2.391	-6.890	Grocer/Market
55	271.9	271.9	165.8	104.3	6.251e+5	-1.132	13.00	-1.330	2.842e-14	19.76	2.733	-5.509	Primary/Secondary School
786	230.6	230.6	59.02	170.6	3.173e+4	0.6703	8.369	1.256	2.606	0.000	0.02876	3.785	Business Services
14	218.0	218.0	43.08	218.0	1.676e+6	-10.56	10.14	-16.30	6.971	-7.748	0.8709	-2.019	Commercial Real Estate
654	175.2	175.2	-27.80	179.1	3.360e+4	0.1915	16.98	-3.773	5.762	33.04	-0.001393	4.858	Bank/Financial Services
718	171.2	171.2	-0.6577	146.2	2.814e+4	-0.7418	13.15	0.02082	8.181	36.09	0.3268	0.1731	Corporate Office
755	122.7	122.7	-7.626	122.7	2.260e+5	1.569	9.844	4.438	11.93	21.43	0.9377	1.856	Other Light Industrial

DEFINITION DIGITALER ZWILLINGE

ISO/IEC 30173: Ein digitales Modell das:

- mit einer Zielentität über eine *Datenverbindung* mit einer geeigneten Synchronisationsrate verbunden ist.
- das eine *ständige Konvergenz* zwischen den physischen und digitalen Modellen ermöglicht.
- eine *integrierte Ansicht* über den *gesamten Lebenszyklus* bietet.

Er erfüllt *zweckgebunden* einige oder alle Fähigkeiten der Verbindung, Integration, Analyse, Vorhersage, Visualisierung, Optimierung, Kollaboration, usw.



VORGEIEN



HÖRSAALFRAGE

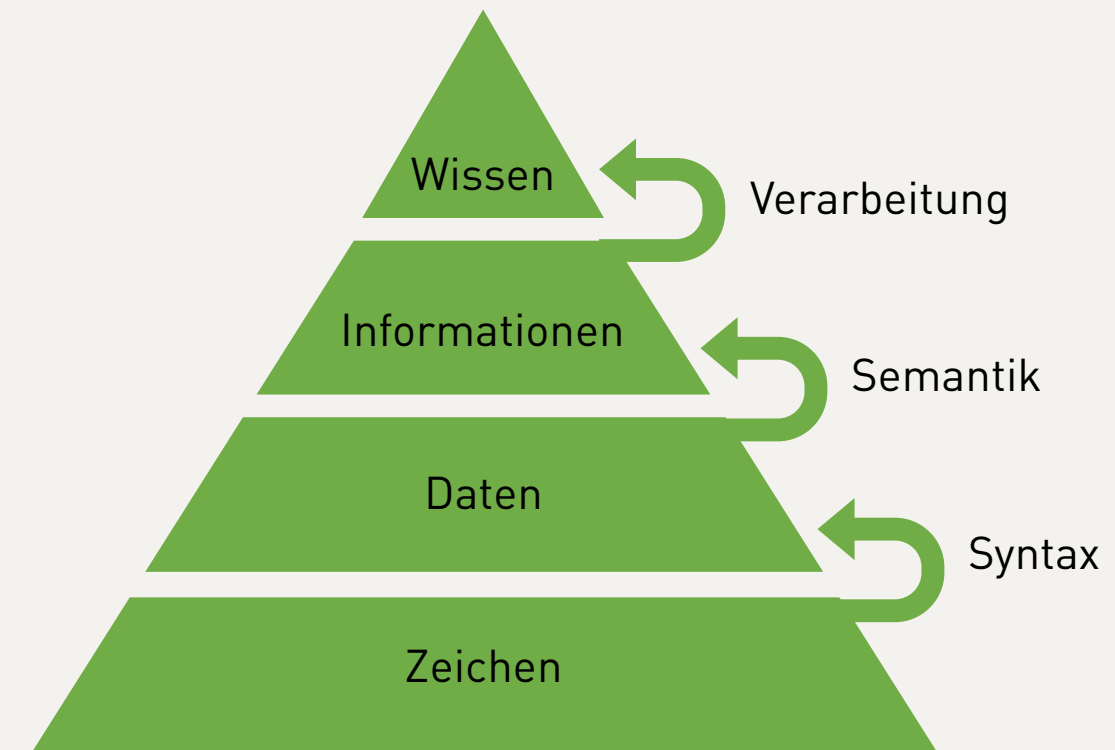
Welche Ebenen hat die Wissenspyramide?

© Midjourney: A pyramid in the style of Hieronymus Bosch



WISSENSPYRAMIDE

Die Wissenspyramide ist ein Modell zur Darstellung der Entstehung von Wissen. Die vier Elementtypen: Zeichen, Daten, Informationen und Wissen werden pyramidenförmig als 4 Ebenen dargestellt, wobei die Zeichen die Basis und das Wissen die Spitze der Pyramide bilden.

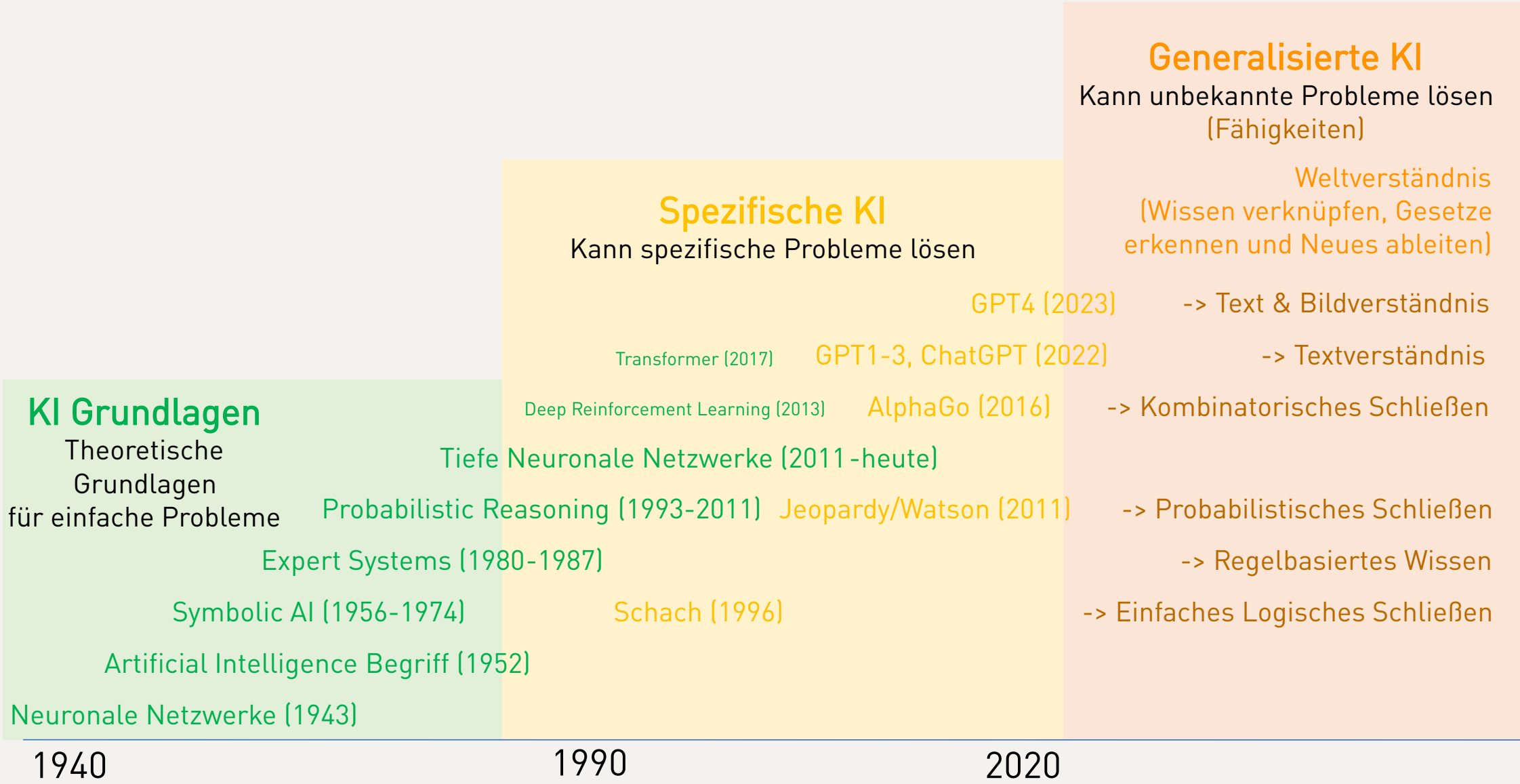


Definition: Maschinelles Lernen

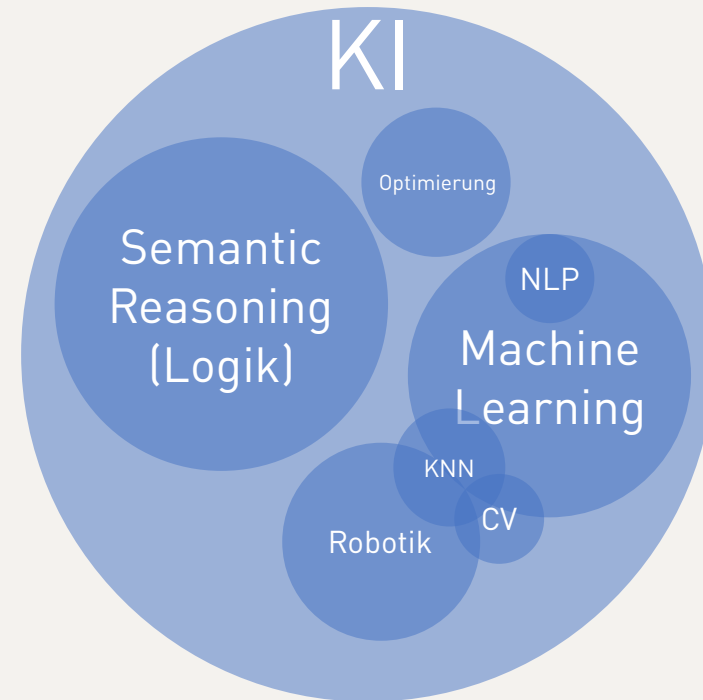
Maschinelles Lernen (ML) ist ein Teilgebiet der Künstlichen Intelligenz, das sich mit der Entwicklung und Untersuchung statistischer Algorithmen befasst, die aus Daten lernen und auf ungesehene Daten verallgemeinern und so Aufgaben ohne explizite Anweisungen ausführen können.

Definition: Künstliche Intelligenz

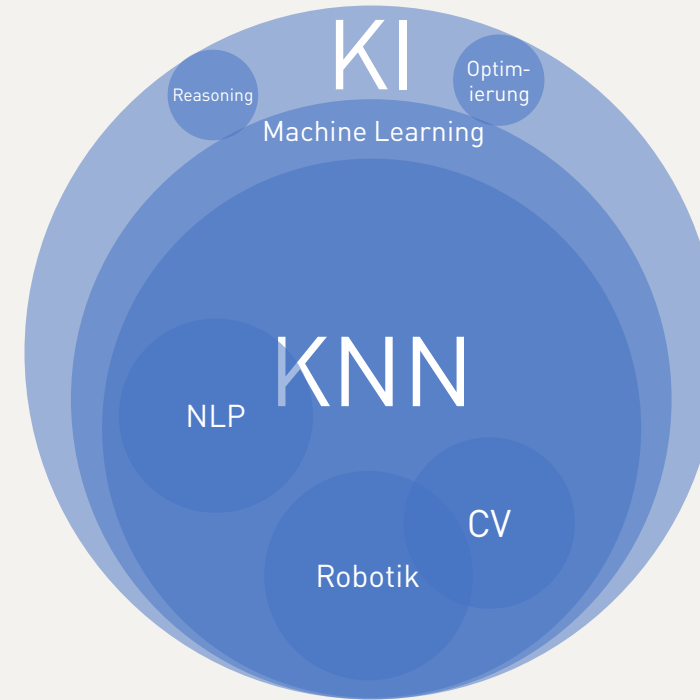
Künstliche Intelligenz (KI) ist ein Forschungsgebiet in der Informatik, das Methoden und Software entwickelt und untersucht, die es Maschinen ermöglichen, ihre Umgebung wahrzunehmen, zu verstehen, zu lernen und daraus Handlungen oder Entscheidungen abzuleiten, um definierte Ziele zu erreichen. Solche Maschinen werden als KI-Systeme bezeichnet.



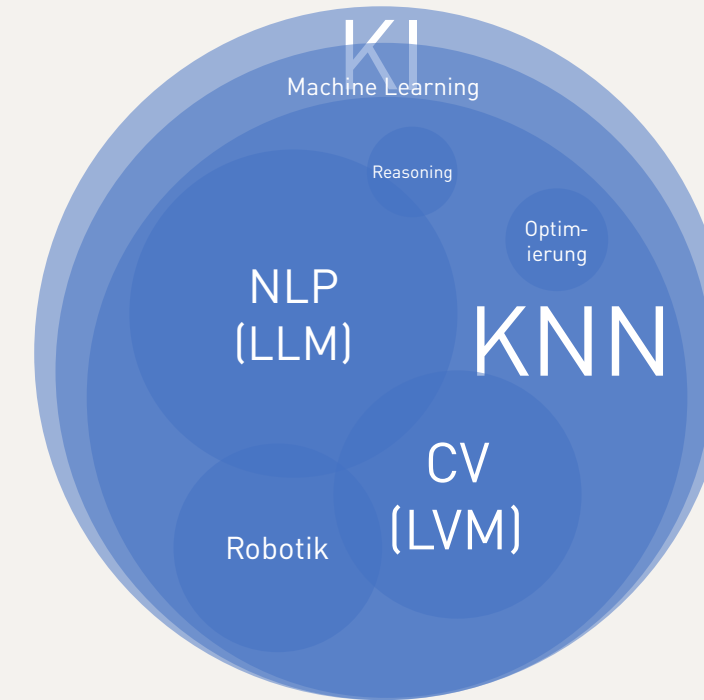
Semantic and Machine
Learning Age (2012)



Neural Network Age (2018)



Large Foundation Model Age (2022)
(a.k.a. Generative Models)



KNN – Künstliche Neuronale Netzwerke

CV – Computer Vision (Bildverarbeitung) und Large Vision Models (LVM)

NLP – Natural Language Processing (Sprachverarbeitung) und Large Language Models (LLM)

HÖRSAALFRAGE

Welche Schritte muss man beim Maschinellen Lernen durchlaufen?

© Midjourney: Process diagram



VORGEHENSWEISE IM MASCHINELLEN LERNEN



Typische Vorgehensweise beim Maschinellen Lernen

DATENEXPLORATION

In diesem Schritt werden die Daten eingehend untersucht, um ein besseres Verständnis für ihre Struktur, Eigenschaften und Muster zu erhalten. Häufige Aufgaben umfassen die Analyse von Verteilungen, Korrelationen zwischen Variablen, das Identifizieren von Ausreißern und das Visualisieren der Daten mittels Diagramme oder Grafiken. Ziel ist es, Einblicke in die Daten zu gewinnen, die für die Modellierung relevant sind, sowie potenzielle Probleme oder Herausforderungen zu identifizieren, die angegangen werden müssen.

VORGEHENSWEISE IM MASCHINELLEN LERNEN



Typische Vorgehensweise beim Maschinellen Lernen

DATENBEREINIGUNG

In diesem Schritt werden die Daten bereinigt und vorverarbeitet, um sicherzustellen, dass sie für das Modelltraining geeignet sind. Dies beinhaltet Aufgaben wie das Entfernen fehlender oder unvollständiger Daten, das Behandeln von Ausreißern, das Skalieren von Merkmalen und das Codieren kategorialer Variablen. Ziel ist es, qualitativ hochwertige Daten bereitzustellen, die frei von Störungen oder Verzerrungen sind und eine reibungslose Modellierung ermöglichen.

VORGEHENSWEISE IM MASCHINELLEN LERNEN



Typische Vorgehensweise beim Maschinellen Lernen

MODELLAUSWAHL

In diesem Schritt werden geeignete Modelle ausgewählt, die für das spezifische Problem am besten geeignet sind, sowie relevante Merkmale, die zur Vorhersage beitragen. Dies kann durch die Bewertung verschiedener Modelle und Merkmalskombinationen anhand von Leistungsmetriken oder durch den Einsatz von Techniken wie Feature Importance oder Cross-Validation erfolgen. Ziel ist es, das beste Modell und die am besten geeigneten Merkmale auszuwählen, um genaue und robuste Vorhersagen zu ermöglichen.

VORGEHENSWEISE IM MASCHINELLEN LERNEN



Typische Vorgehensweise beim Maschinellen Lernen

MODELLTRAINING

In diesem Schritt wird das ausgewählte Modell auf den Trainingsdaten trainiert, um die Beziehung zwischen den Eingangsmerkmalen und den Zielvariablen zu lernen. Das Modell wird durch Anpassen seiner Parameter an die Trainingsdaten verbessert, wobei ein Optimierungsalgorithmus wie Gradientenabstieg verwendet wird. Ziel ist es, ein Modell zu entwickeln, das die Trainingsdaten gut generalisiert und in der Lage ist, genaue Vorhersagen für neue, unbekannte Daten zu treffen.

VORGEHENSWEISE IM MASCHINELLEN LERNEN



Typische Vorgehensweise beim Maschinellen Lernen

MODELLVALIDIERUNG

Nach dem Training wird das Modell mit Validierungs- oder Testdaten bewertet, um seine Leistung zu überprüfen und zu sehen, wie gut es auf neue Daten generalisiert. Die Vorhersagen des Modells werden mit den tatsächlichen Werten verglichen, um die Genauigkeit und Zuverlässigkeit des Modells zu überprüfen. Ziel ist es, sicherzustellen, dass das Modell robust ist und konsistente Leistung über verschiedene Datensätze und Umgebungen bietet.

VORGEHENSWEISE IM MASCHINELLEN LERNEN



Typische Vorgehensweise beim Maschinellen Lernen

MODELLANWENDUNG

Nach erfolgreicher Validierung wird das trainierte Modell in eine Produktionsumgebung implementiert, wo es zur Vorhersage auf Echtzeitdaten angewendet wird. Dies kann die Integration des Modells in Softwareanwendungen, Webdienste oder andere Systeme umfassen. Ziel ist es, das Modell in einer Umgebung bereitzustellen, in der es praktische Anwendungen hat und kontinuierlich genutzt werden kann.

HÖRSAALFRAGE

Welcher Schritt ist der Zeitaufwändigste?

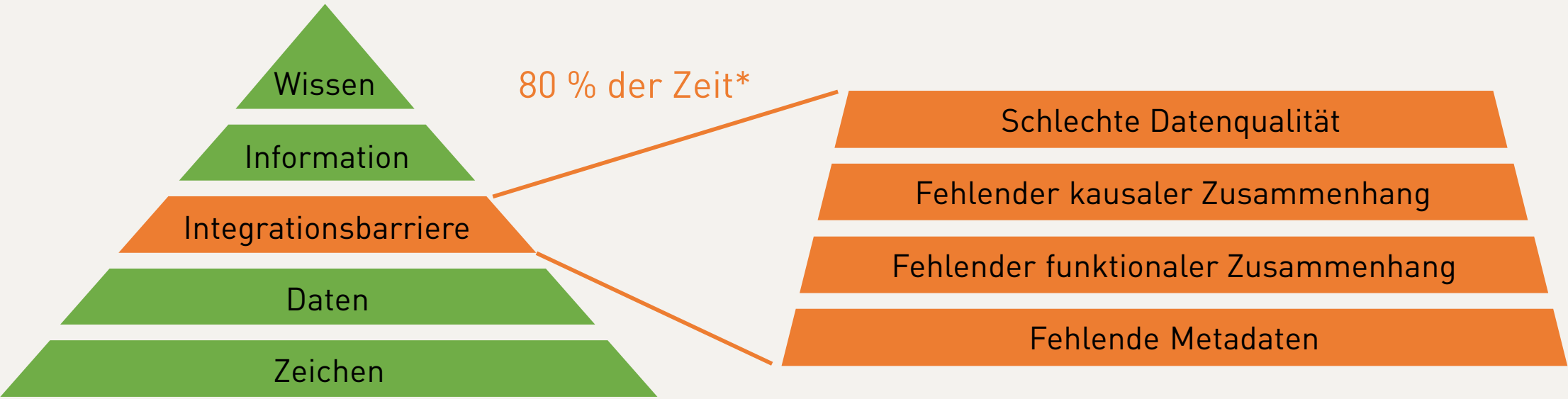
© Midjourney: Process diagram

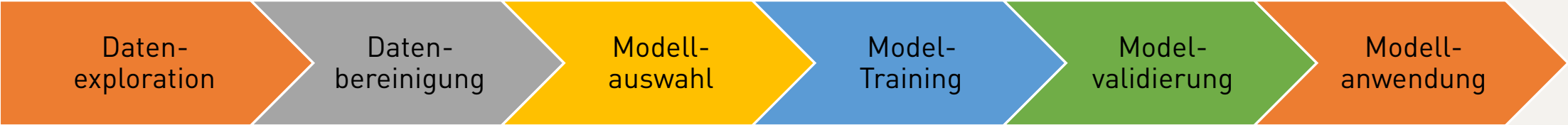


DIE HERAUSFORDERUNG VON ML

Datenaufbereitung ist der zeitaufwändigste (80%) und wenig herausforderndste langweiligste (76%) Schritt bei der Datenanalyse.¹

Die digitale Transformation besteht zu 80 % aus Menschen, zu 20 % aus Technologie: Domänenwissen ist von entscheidender Bedeutung, und daher der Bedarf an Experten, die das Geschäft verstehen und über das notwendige Wissen verfügen, um die richtigen Fragen zu stellen und die Antworten zu kontextualisieren.²





Typische Vorgehensweise beim Maschinellen Lernen

Die Datenanalyse in der Praxis geschieht meist nicht sequentiell, sondern man iteriert meist agil zu der richtigen Lösung. Das umfasst mehrmaliges Bereinigen der Daten, Überdenken der Modellauswahl sowie mehrmaliges Modelltraining und Modellvalidierung, bevor das richtige Modell gefunden wird.



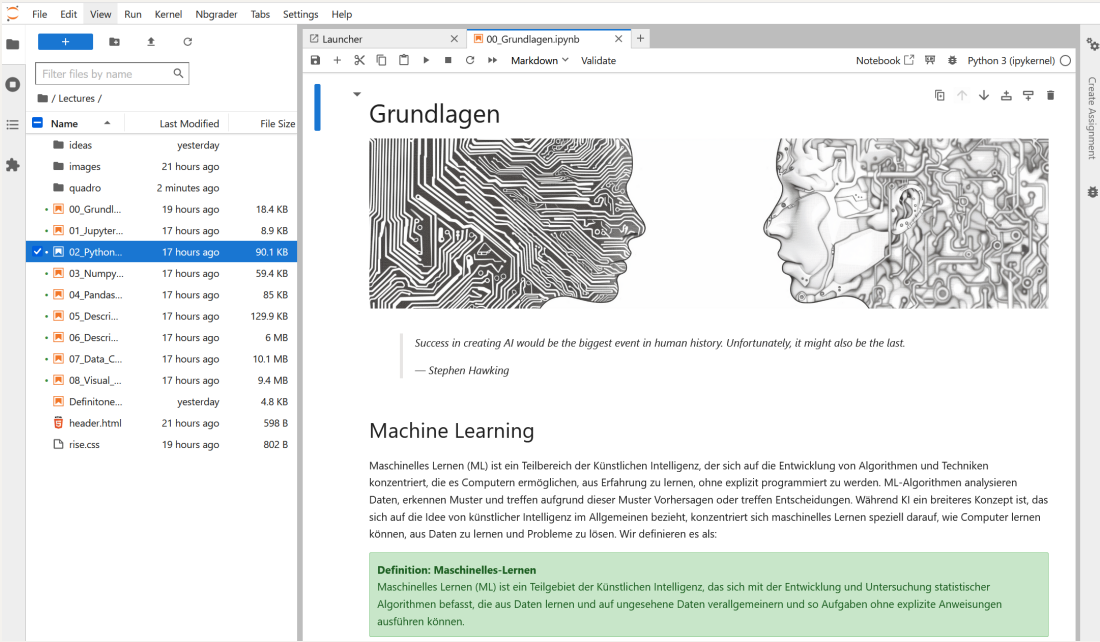
Beispielhafter Ablauf in relativer Zeit mit den Farben für die Schritte oben

TOOLS



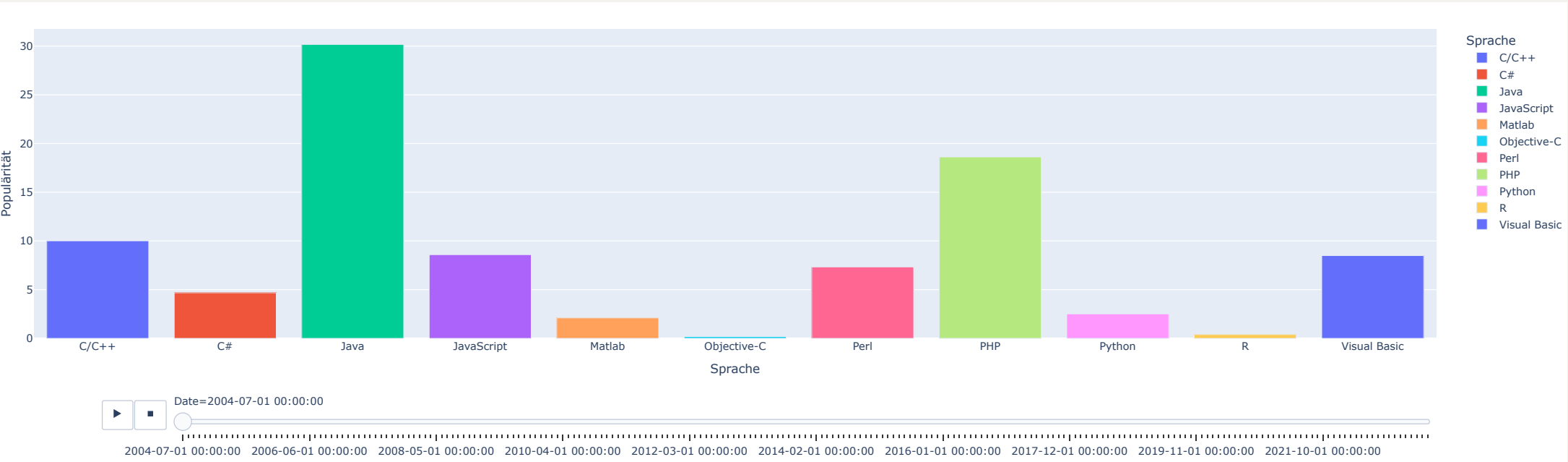
JUPYTER NOTEBOOKS

- Interaktive Entwicklungsumgebung zum schrittweisen Ausführen von Code
 - Agile, iterative Entwicklung von ML-Modellen
- Speichert die Ausgaben von Code wie Textausgaben und Visualisierungen
 - Ergebnisse sichtbar und belegt
- Ablauf der ML-Modellgenerierung ist belegt und kann wiederholt werden
 - Ablauf wiederholbar
- Kann um Texte und Bilder in Markdown erweitert werden
 - Erklärungen helfen Verständnis
- Kann dadurch einfach geteilt und gemeinsam bearbeitet werden
 - Einfache Kommunikation



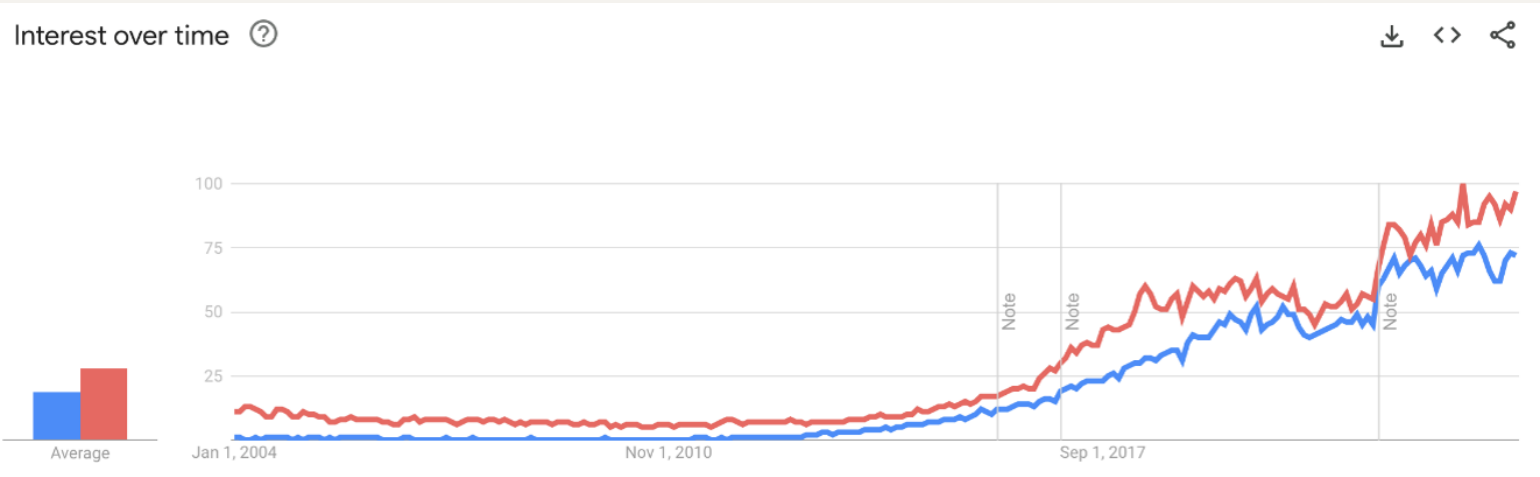
ENTWICKLUNG VON PROGRAMMIERSPRACHEN

Python hat in den letzten Jahren sehr stark an Popularität gewonnen.





Python wird in Data Science zum Maschinellen Lernen sehr oft verwendet. Sie hat sehr viele Bibliotheken zur Analyse von Daten. Heutzutage gibt es in Python Bibliotheken für fast jedes Anwendungsgebiet.



PYTHON

Entworfen: 1991

Auf Lesbarkeit ausgelegt

The Zen of Python (PEP 20):

- Beautiful is better than ugly
- Explicit is better than implicit
- Simple is better than complex
- Complex is better than complicated
- Readability counts

© A python programming a computer





Numpy ist die führende Bibliothek in Python für Vektor- und Matrixberechnungen. Daten und Berechnungen werden dabei nicht in Python sondern mit optimierten C-Objekten und Operationen ausgeführt. Es ist besonders gut für rein numerische Analysen wie lineare Algebra geeignet.



scikit-learn ist eine der wichtigsten Bibliotheken für traditionelles maschinelles Lernen in Python ohne spezifischen Fokus auf Neuronale Netzwerke. Es basiert intern auf Numpy und SciPy, welches viele Algorithmen für lineare Algebra implementiert.



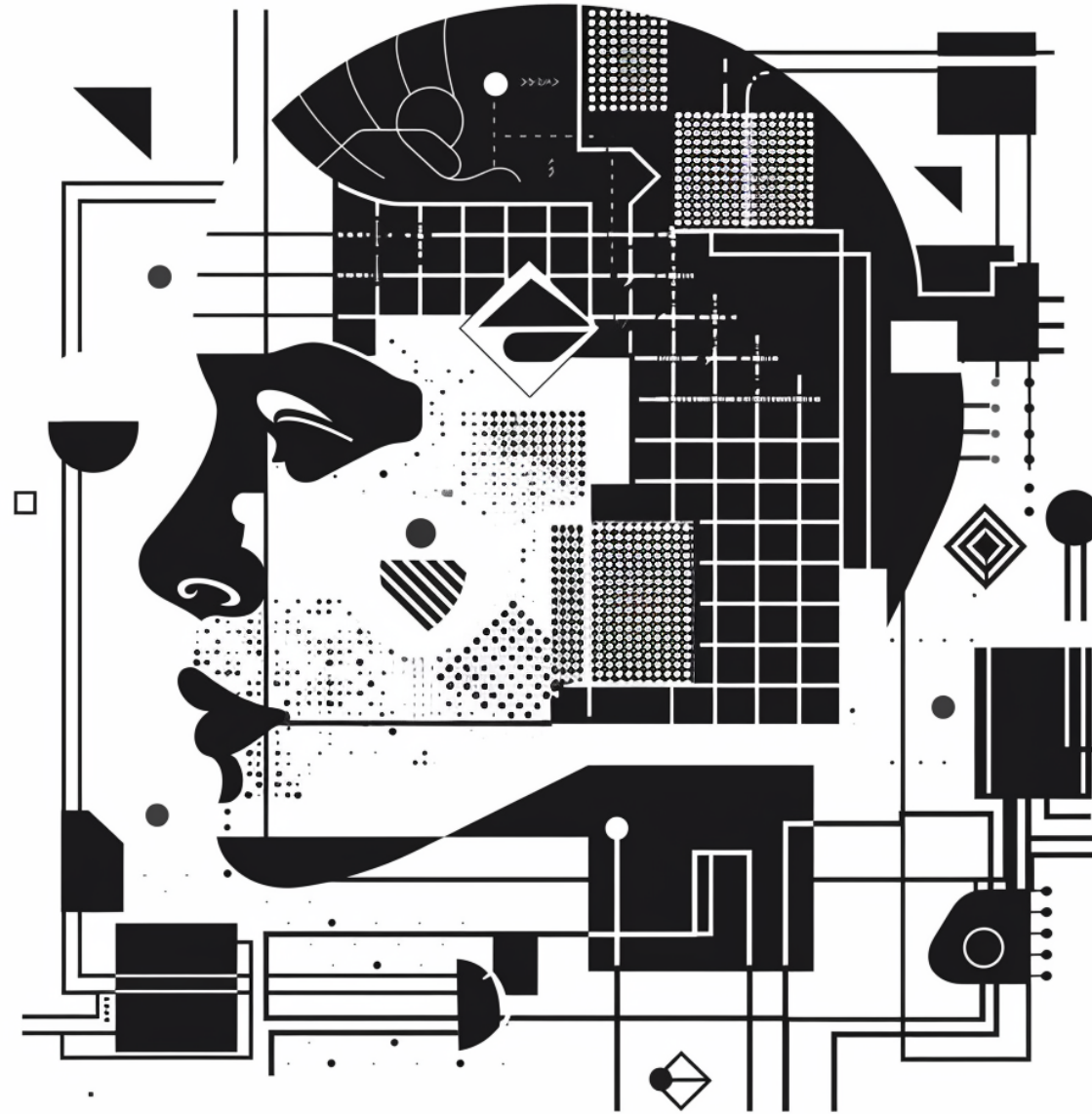
Pandas ist die wichtigste Python-Bibliothek für die Arbeit mit Tabellen (Dataframes). Pandas nutzt intern Numpy (Vektoren und Matrizen) fügt aber Tabellenkonzepte hinzu (Spaltennamen, Zeilenindizes, Aggregationsoperatoren, etc.).



Plotly ist eine Bibliothek zur Erstellung interaktiver Diagramme, die sehr einfach zu benutzen ist. Es können damit einfache Diagramme erzeugt werden bis zu ganzen Benutzeroberflächen. Es nutzt modern Webtechnologien (JavaScript) zur Visualisierung.

KLASSEN VON ML-MODELLEN

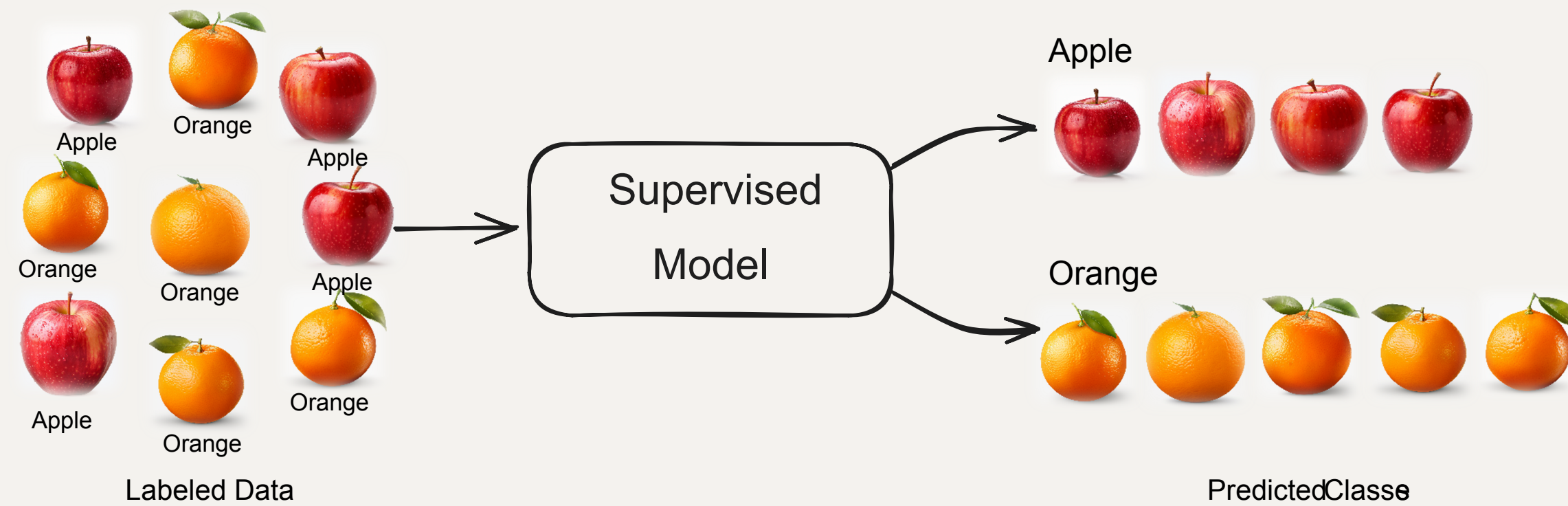
© Midjourney: Shapes of machine learning



ÜBERWACHTES LERNEN, EN. SUPERVISED LEARNING

Die Modelle benötigen vorher gelabelte Beispieldaten, die vorgeben, wie das Ergebnis aussehen soll. Beim Labeln weist man den Daten die zu erlernenden Zielwerte oder Kategorien zu. Das ML-Modell lernt daraus, welche Eingaben zu welchen Zielwerten oder Klassen gehören.

- Gut, wenn man die Zielwerte oder Klassen kennt und gelabelte Daten zum Trainieren hat.



Supervised Learning

Die Modelle benötigen *vorher* gelabelten Beispieldaten, die vorgeben, wie das Ergebnis aussehen soll.

Klassifikation

Bei dieser Aufgabe werden Eingabedaten in vordefinierte Klassen oder Kategorien eingeteilt.

Regression

Hierbei werden Zusammenhänge zwischen Eingabe- und Ausgabedaten modelliert, um numerische Werte vorherzusagen.

Spam-Erkennung

Bilderkennung

Diagnose

Wettervorhersage

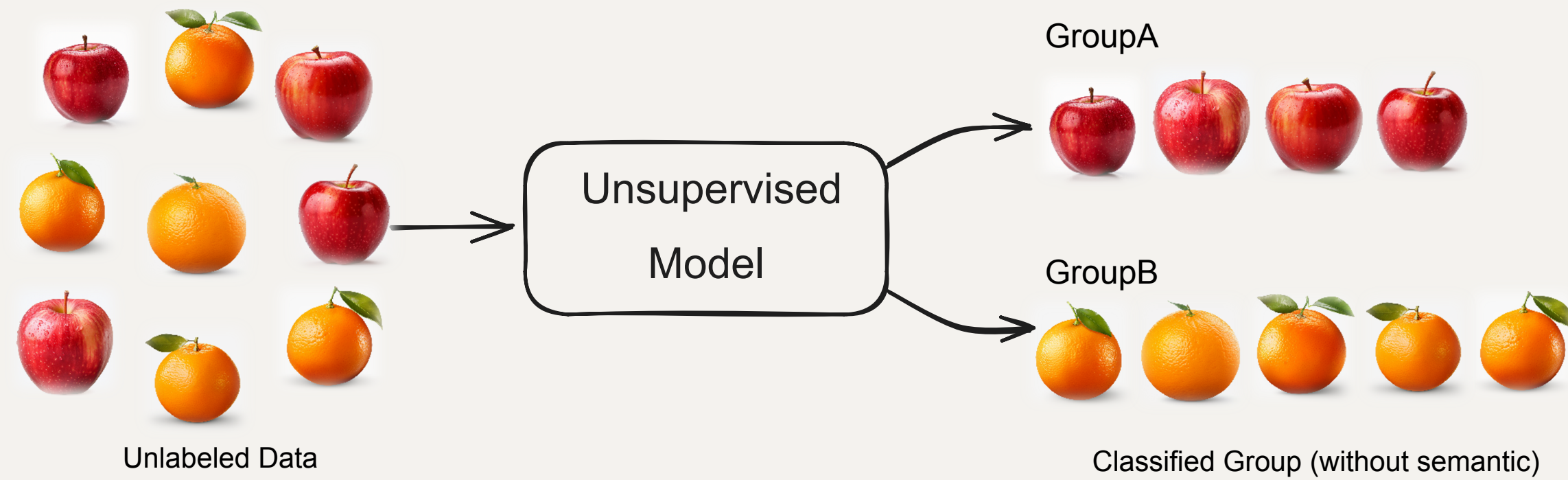
Kostenprognose

Sensordatenvorhersage

UNÜBERWACHTES LERNEN, EN. UNSUPERVISED LEARNING

Die Modelle benötigen keine gelabelten Beispieldaten zum Lernen, sondern identifizieren Strukturen in den Daten. Das ML-Modell lernt selbst die Elemente zu unterscheiden, kennt aber nicht den semantischen Namen der Cluster.

- Gut wenn man die Kategorien nicht kennt oder keine gelabelten Daten hat.



Unsupervised Learning

Die Modelle benötigen *keine* gelabelten Beispieldaten zum Lernen sondern identifizieren Strukturen in den Daten.

Clustering

Diese Aufgabe zielt darauf ab, ähnliche Datenpunkte zu gruppieren, ohne vorab definierte Klassen anzugeben.

Dimensionsreduktion

Dabei wird die Dimensionen der Daten reduziert, um die Komplexität zu verringern und wichtige Merkmale zu extrahieren.

Kaufempfehlungen

Ressourcenzuweisung

Anomalieerkennung

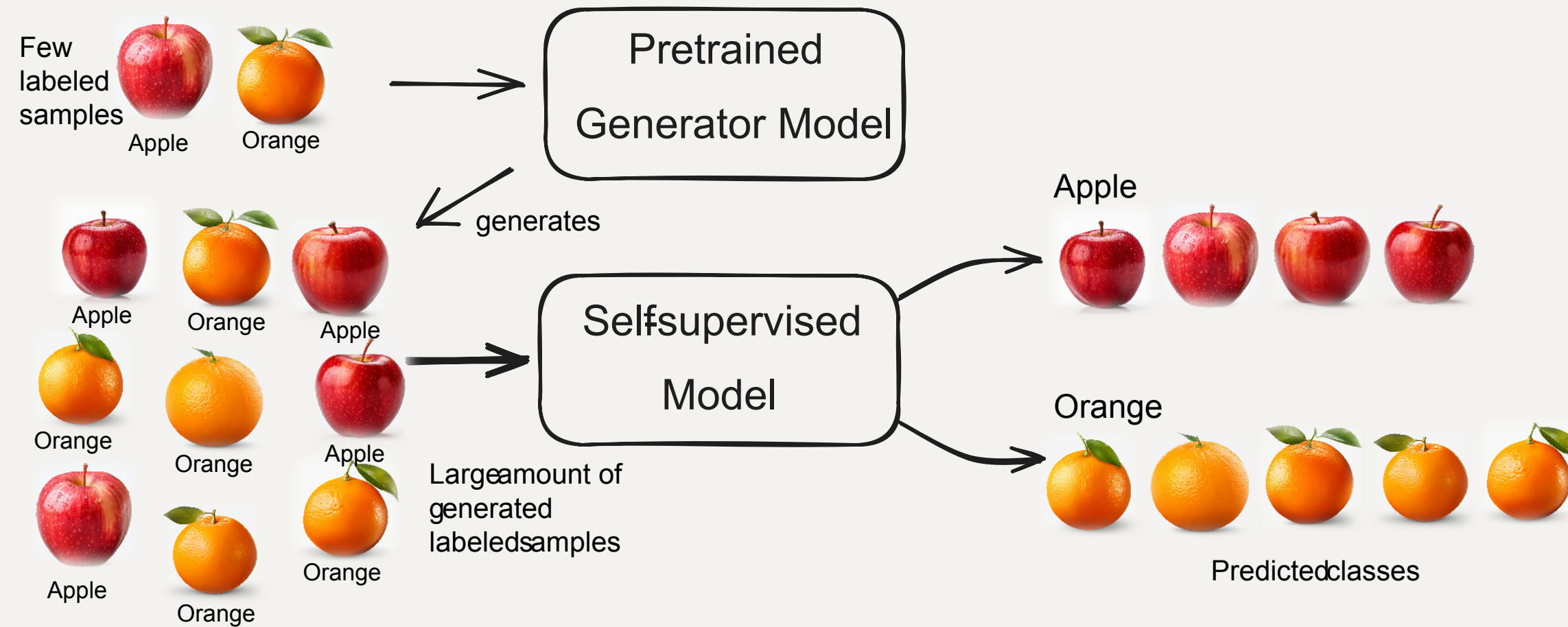
Sensordatenreduktion

Visualisierung

SELBSTÜBERWACHTES LERNEN, EN. SELF-SUPERVISED LEARNING

Beim selbstüberwachten Lernen werden die Trainingsziele automatisch aus den vorhandenen Daten selbst abgeleitet, indem zum Beispiel Daten weggelassen (maskiert) werden, variiert werden, oder durch andere Modelle erzeugt werden. Dadurch lassen sich große Mengen ungelabelter Daten zum Vortraining nutzen, bevor ein Modell gegebenenfalls mit gelabelten Daten nachtrainiert wird.

- Gut, wenn große Mengen ungelabelter Daten für automatisch abgeleitete Trainingsziele verfügbar sind.



Self-supervised Learning

Die Modelle *erzeugen selbst* gelabelte Trainingsdaten mit einem vortrainierten, internen Generatormodell.

Generierung (Vorhersage)

Bei dieser Aufgabe werden Ausgangsdaten aus einer kleinen Menge von Eingangsdaten (Prompt) erzeugt.

Optimierung

Hierbei werden optimale Lösungsstrategien für Probleme gesucht.

Textgenerierung

Fragenbeantworten

Bilderzeugung

Spiele (z.B. Go, Minecraft)

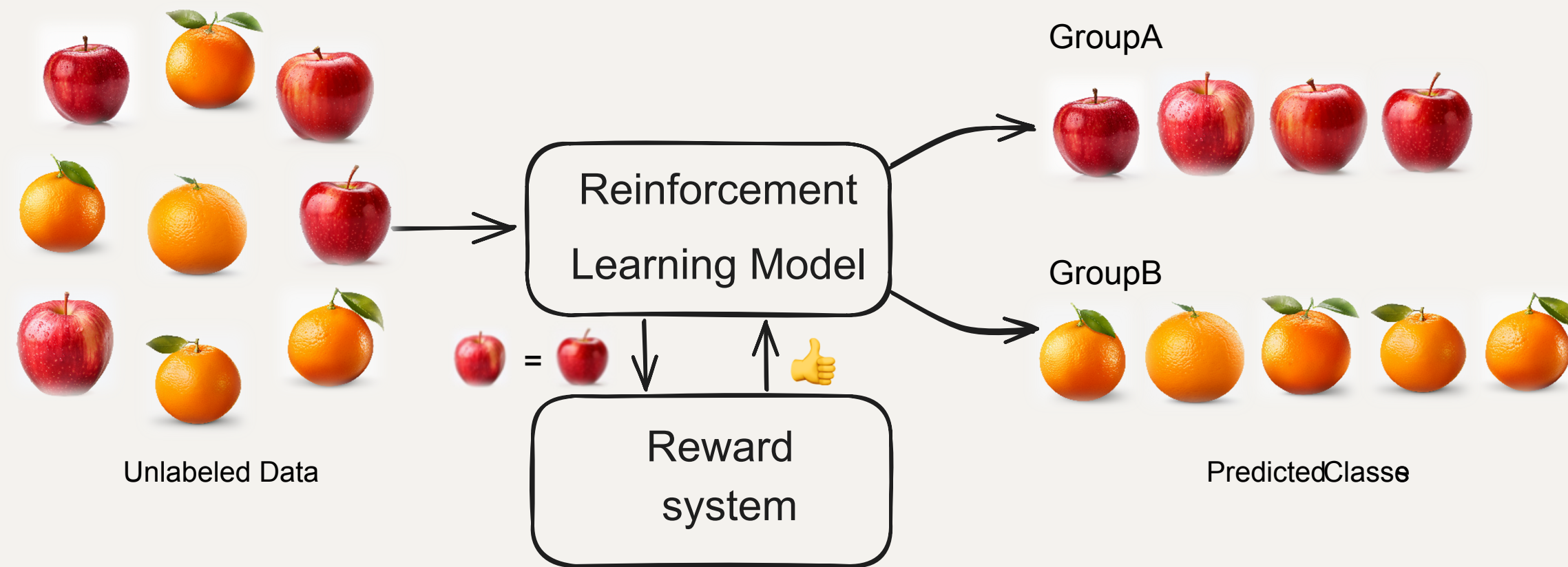
Spracherkennung

Logistikprobleme

VERSTÄRKENDES LERNEN, EN. REINFORCEMENT LEARNING

Die Modelle werden durch Belohnung trainiert, die angibt, ob eine Entscheidung oder Handlung gut oder schlecht war. Das Modell lernt aus dieser Rückmeldung, welche Entscheidungen in einer Situation tendenziell besser oder schlechter sind.

- Gut, wenn man die optimale Entscheidung nicht direkt kennt, aber bewerten (und belohnen) kann, welche Handlungen tendenziell besser oder schlechter sind.



Reinforcement Learning

Die Modelle werden durch Belohnung trainiert, die angibt ob ein Ergebnis gut oder schlecht ist.

Modellfrei

Das Modell lernt direkt aus der Erfahrung, indem es Aktionen in der realen Umgebung ausprobiert und die resultierenden Belohnungen beobachtet.

Modellbasiert

Das Modell lernt intern ein Modell der Umgebung und Belohnungsstruktur und kann so vorhersagen, welche Aktionen sich lohnen, ohne alle aktiv ausprobieren zu müssen.

Monte-Carlo-Simulation

Logistikoptimierung

Energiemanagement

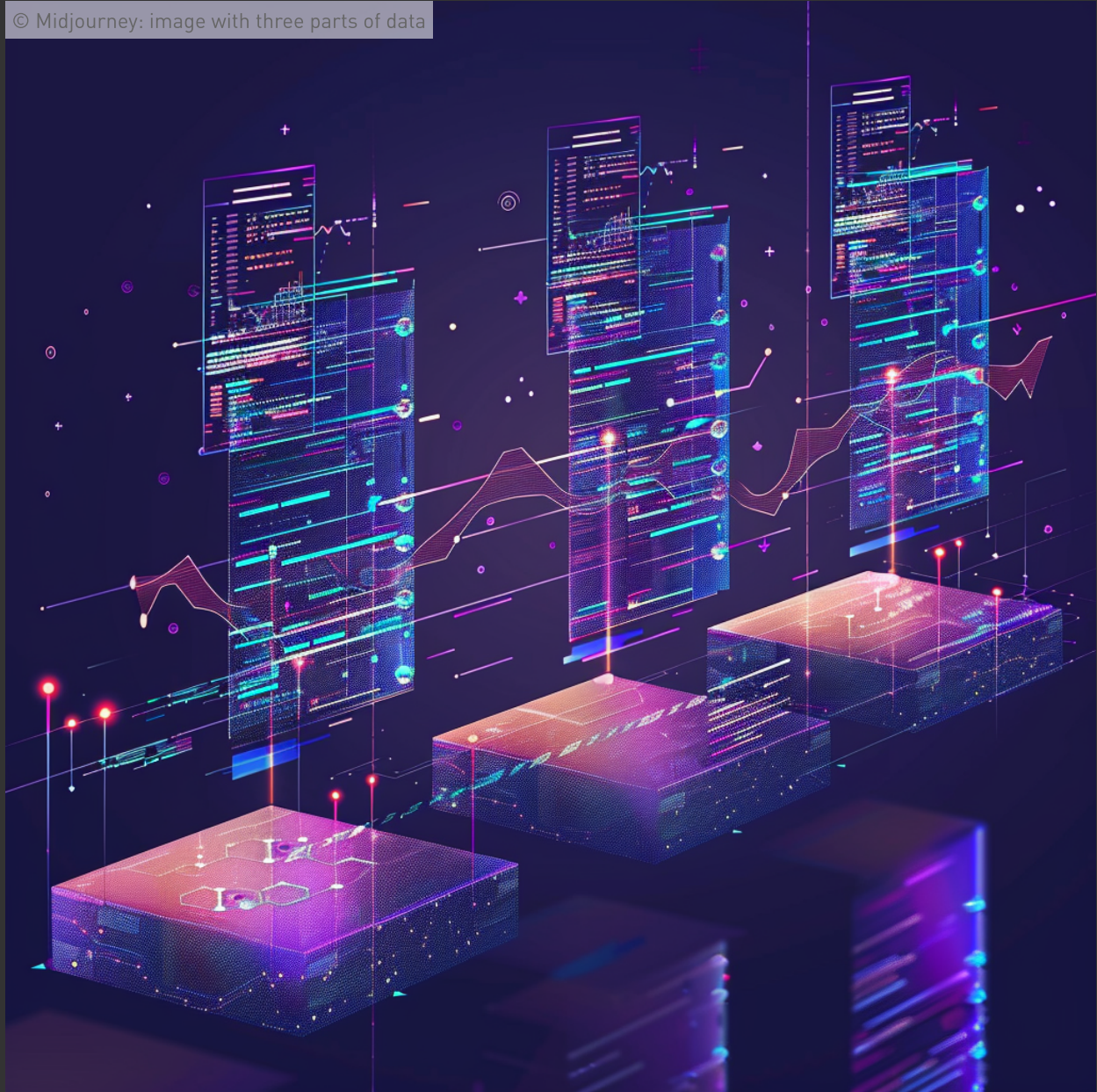
Robotersteuerung

Computerspiele

Modellbasierte Regelung

DATENKLASSEN

© Midjourney: image with three parts of data



Strukturiert

Die Struktur der Daten ist syntaktisch und semantisch wohl definiert.

Tabellen

Diese Daten sind in tabellarischer Form organisiert. (CSV-Dateien, SQL-Datenbanktabellen oder Excel-Tabellen)

- Statistiken, Baustoffkataloge, Projektdaten

Zeitreihen

Diese Daten sind zeitlich geordnet und bestehen aus einer Abfolge von Zeitpunkten und den jeweiligen Werten.

- Aktienkurse, Wetterdaten, Energiedaten, Sensordaten

Jede Zeile enthält einen zusammenhängenden Datensatz → (Kontext)

	A	B	C	D	E	F
1	Datum	Wert	Skalierungsfakor	Skalierter Wert	Wertänderung	
2	2021-01-02 02:00:01Z	5763.06	1000	5763060	0	
3	2021-01-02 03:00:00Z	5763.06	1000	5763060	0	
4	2021-01-02 04:00:01Z	5763.07	1000	5763070	10	
5	2021-01-02 05:00:01Z	5763.23	1000	5763230	160	
6	2021-01-02 06:00:00Z	5763.40	1000	5763400	170	
7	2021-01-02 07:00:01Z	5763.57	1000	5763570	170	
8	2021-01-02 08:00:00Z	5763.74	1000	5763740	170	
9	2021-01-02 09:00:00Z	5763.91	1000	5763910	170	
10	2021-01-02 10:00:00Z	5764.08	1000	5764080	170	
11	2021-01-02 11:00:00Z	5764.23	1000	5764230	150	
12	2021-01-02 12:00:01Z	5764.23	1000	5764230	0	
13	2021-01-02 13:00:00Z	5764.23	1000	5764230	0	
14	2021-01-02 14:00:00Z	5764.23	1000	5764230	0	
15	2021-01-02 15:00:00Z	5764.23	1000	5764230	0	
16	2021-01-02 16:00:00Z	5764.23	1000	5764230	0	
17	2021-01-02 17:00:03Z	5764.23	1000	5764230	0	
18	2021-01-02 18:00:00Z	5764.23	1000	5764230	0	
19	2021-01-02 19:00:01Z	5764.23	1000	5764230	0	
20	2021-01-02 20:00:01Z	5764.23	1000	5764230	0	

← Spalten haben Namen (Semantik)

↑
Jede Zeile hat einen Zeitstempel, deshalb ist das auch eine Zeitreihe

↑
Werte in Spalte haben festen Datentyp (Syntax)

Unstrukturiert

Die Struktur der Daten ist meist zwar syntaktisch definiert aber der semantische Inhalt ist nicht klar ableitbar.

Text

Textdaten sind unstrukturierte, da die Struktur komplex und die Bedeutung sehr Kontextspezifisch ist.

- E-Mails, Dokumente, Internettexte, Bauberichten, Vorschriften

Bilder

Bilder werden typischerweise in Form von Pixelwerte als mehrdimensionale Matrix gerastert.

- Fotos, Mängelbelege, Diagramme

Videos

Videos werden als Bildsequenz interpretiert.

- Filme, Robotervision, Überwachungskameras

Audios

Diese Daten enthalten Klangwellenformen und können auch als Zeitreihe gesehen werden.

- 1.Telefonate, Musik, Audioüberwachung, Straßenlärmmonitoring

Texte werden im Computer als Zeichenfolge gespeichert (Kontext). Der Computer erkennt zwar Worte (Syntax), aber nicht deren Bedeutung (Semantik).

>Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.



Bilder werden im Computer als Pixelmatrix gespeichert (Kontext). Die Farbe jedes Pixels wird dabei z.B. als RGB-Zahl gespeichert (Syntax). Aber wir kennen damit noch nicht die Semantik.

Semi-Strukturiert

Die Daten enthalten sowohl wohldefinierte strukturierte Elemente als auch unstrukturierte Elemente.

CAD/BIM

CAD-Daten können als semi-strukturierte Daten betrachtet werden, da die Basisstruktur (z.B. IFC) bekannt ist, aber unstrukturierte Elemente wie Text enthält.

- Bauplänen, Ansichten

XML/JSON

XML/JSON-Dateien haben zwar eine Objektstruktur, sie ist aber flexibler sind als tabellarische Daten.

- Webseiten in HTML
- Internetabfragen

```
{
  "name": "Georg",
  "alter": 47,
  "verheiratet": false,
  "beruf": null,
  "kinder": [
    {
      "name": "Lukas",
      "alter": 19,
      "schulabschluss": "Gymnasium"
    },
    {
      "name": "Lisa",
      "beschreibung": "Rote Haare mit Zopf."
    }
  ]
}
```

JSON wird im Computer als Zeichenfolge gespeichert und verbindet verschiedene Daten in einem Objekt (Kontext). Die Objekte folgen zwar einer gewissen Struktur (Syntax) können aber davon abweichen und unstrukturierte Elemente, wie Text enthalten.

```
#14290= IFCQUANTITYAREA('GrossSideArea',$,$,13.5);
#14292= IFCQUANTITYAREA('NetSideArea',$,$,13.5);
#14294= IFCQUANTITYVOLUME('GrossVolume',$,$,2.376);
#14296= IFCQUANTITYVOLUME('NetVolume',$,$,2.376);
#14298= IFCQUANTITYAREA('GrossFootprintArea',$,$,0.88);
#14300= IFCELEMENTQUANTITY('39WRingk_9CBgT2(yur)QGel',#29,'BaseQuantities',$,' ',
(#14284,#14286,#14288,#14290,#14292,#14294,#14296,#14298));
#14305= IFCRELDEFINESBYPROPERTIES('3pBHJ4_rT0Hf0lp_Fbrd9j',#29,$,$,(#14122),#14300);
#14307= IFCMATERIALLAYERSETUSAGE(#14105, .,AXIS2, .,NEGATIVE, .,0.1);
#14308= IFCDIRECTION((0,.-1,0.));
#14312= IFCCARTESIANPOINT((12.,7.,0.));
#14316= IFCAXIS2PLACEMENT3D(#14312,#52,#14308);
#14320= IFCLOCALPLACEMENT(#2946,#14316);
#14323= IFCWALLSTANDARDCASE('2Jxn8iNNv7LFvRbi4z0_Vd',#29,'Ward-019',$,$,
$,#14320,#14393,'35C73C96-4BA7-4334-AG-A1-102EF027E980');
#14342= IFCCARTESIANPOINT((0, .,0.));
```

IFC-Dateien (CAD/BIM) sind Textdateien, aus denen ein Objektgraph aufgebaut wird (Kontext). Es gibt eine definierte Syntax und die Semantik der IFC-Tags selbst, aber keine Semantik für enthaltene Bezeichner (z.B. Raumtypen).

Strukturiert

Die Struktur der Daten ist syntaktisch und semantisch wohl definiert.

Tabellen

Diese Daten sind in tabellarischer Form organisiert. (CSV-Dateien, SQL-Datenbanktabellen oder Excel-Tabellen)

- Statistiken, Baustoffkataloge, Projektdaten

Zeitreihen

Diese Daten sind zeitlich geordnet und bestehen aus einer Abfolge von Zeitpunkten und den jeweiligen Werten.

- Aktienkurse, Wetterdaten, Energiedaten, Sensordaten

Text

Textdaten sind unstrukturierte, da die Struktur komplex und die Bedeutung sehr Kontextspezifisch ist.

- E-Mails, Dokumente, Internettexte, Bauberichten, Vorschriften

Bilder

Bilder werden typischerweise in Form von Pixelwerte als mehrdimensionale Matrix gerastert.

- Fotos, Mängelbelege, Diagramme

Videos

Videos werden als Bildsequenz interpretiert.

- Filme, Robotervision, Überwachungskameras

Audios

Diese Daten enthalten Klangwellenformen und können auch als Zeitreihe gesehen werden.

1.Telefonate, Musik, Audioüberwachung, Straßenlärmmonitoring

CAD/BIM

CAD-Daten können als semi-strukturierte Daten betrachtet werden, da die Basisstruktur (z.B. IFC) bekannt ist, aber unstrukturierte Elemente wie Text enthält.

- Bauplänen, Ansichten

Semi-Strukturiert

Die Daten enthalten sowohl wohldefinierte strukturierte Elemente als auch unstrukturierte Elemente.

XML/JSON

XML/JSON-Dateien haben zwar eine Objektstruktur, sie ist aber flexibler sind als tabellarische Daten.

- Webseiten in HTML
- Internetabfragen

HÖRSAALFRAGE

Welche Datentypen gibt es in Programmiersprachen?

© Midjourney: data types

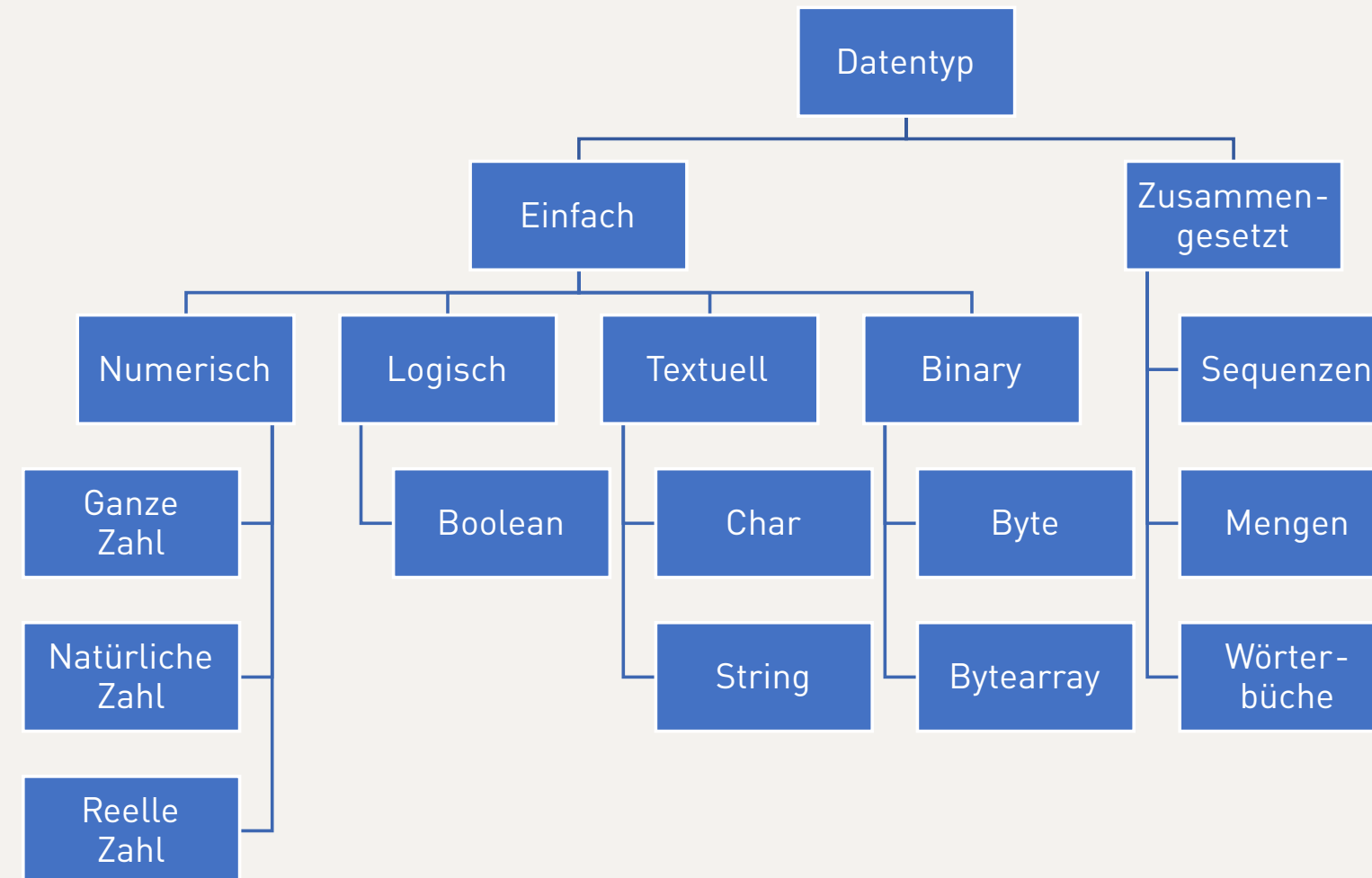


DATENTYPEN



Definition: Datentyp

Der Datentyp gibt an, von welcher Art die Daten sind, die mit ihm beschrieben werden (Datenvereinbarung), und welche Operationen auf diesen ausgeführt werden können.



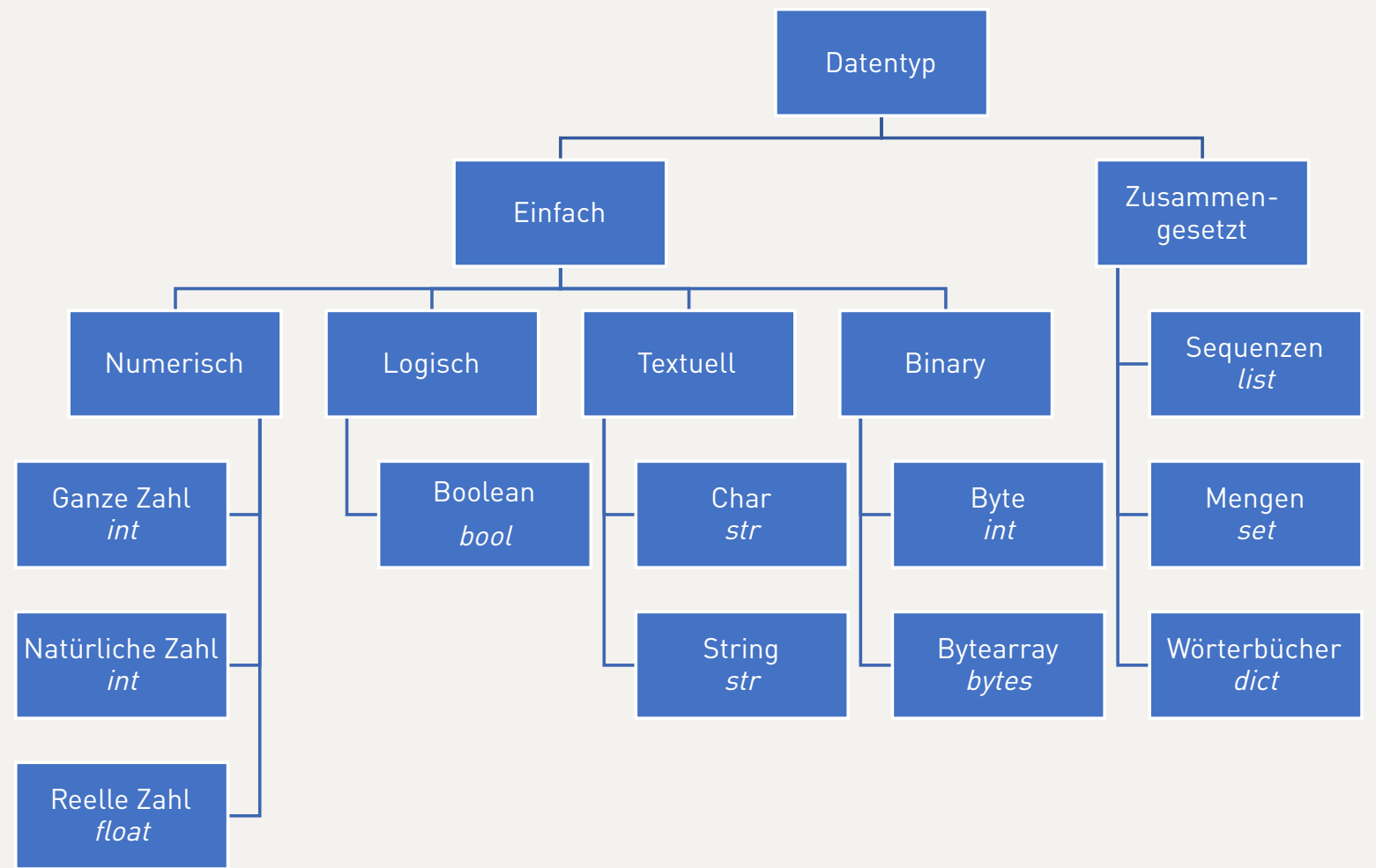
HÖRSAALFRAGE

Welche Datentypen gibt es in Python?

© Midjourney: A python programming a robot

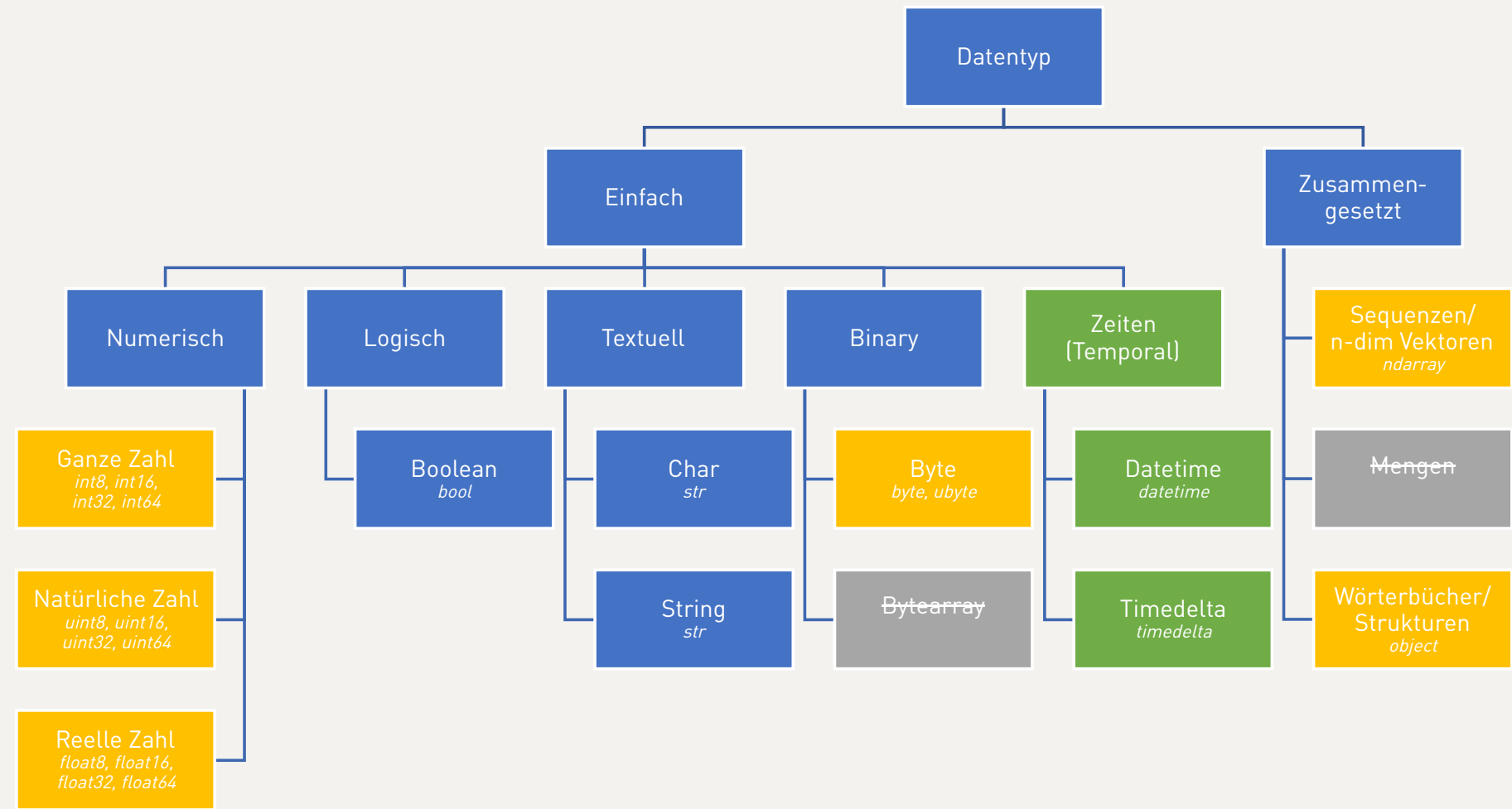


Python vereinfacht die Datentypen auf einige wenige Typen und weist die Typen dynamisch zu. Das ist einfach zu benutzen, aber bei vielen Berechnungen nicht sehr performant.



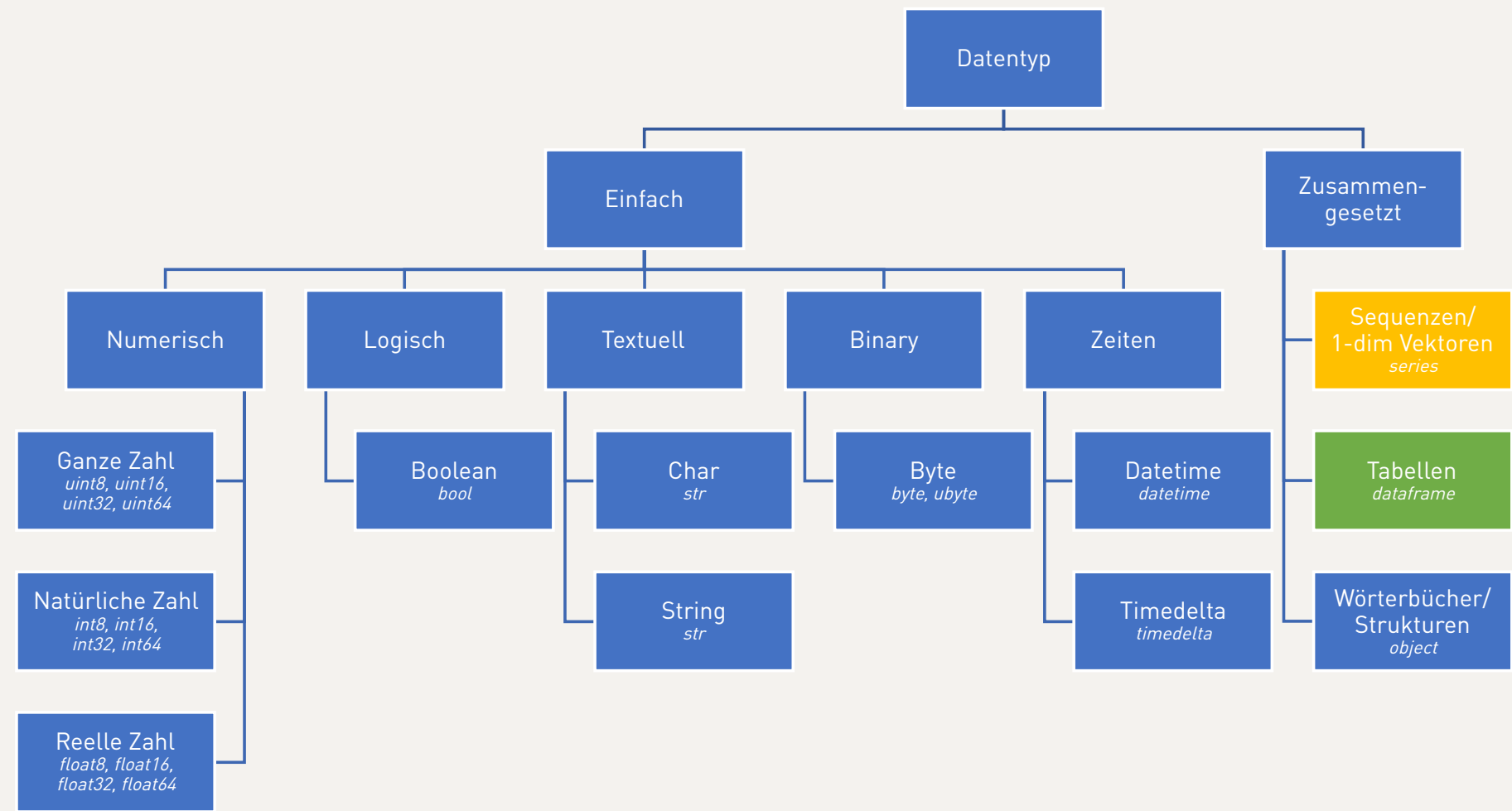
DATENTYPEN IN NUMPY

Numpy führt explizite Datentypen ein, damit Berechnungen deutlich performanter durchgeführt werden können, dies wird erzielt durch optimierte Speicherdarstellung, Vermeidung von Konvertierungen und effizienten Rechnungen in C.



DATENTYPEN IN PANDAS

Pandas nutzt die Basisdatentypen von Numpy führt allerdings Serien und Tabellen (Dataframe) ein, damit Berechnungen deutlich performanter durchgeführt werden können, dies wird erzielt durch optimierte Speicherdarstellung, Vermeidung von Konvertierungen und effizienten Rechnungen in C.



QUESTIONS