Data Science

Kick-off Projekt „Big Data Analytics“

DHBW Heidenheim, WI 2022, 5. Semester

Dr. Ralf Höchenberger, ERGO Group AG

07.01.2025

Begriffsdefinition (vgl. Dhar 2013, Leek 2014) und Abgrenzung

Data Science ist

- *ein **interdisziplinäres Wissenschaftsfeld**, welches*
- ***wissenschaftlich fundierte Methoden**, Prozesse, Algorithmen und Systeme*
- *zur **Extraktion von Erkenntnissen, Mustern und Schlüssen** sowohl aus strukturierten als auch unstrukturierten **Daten** ermöglicht*

Verwandte Begriffe:

- ***Big Data Analytics***
- ***Business Intelligence***
- ***Knowledge Discovery in Databases (KDD)***
- ***Data Mining***
- ***Machine Learning***
- ***Künstliche Intelligenz***

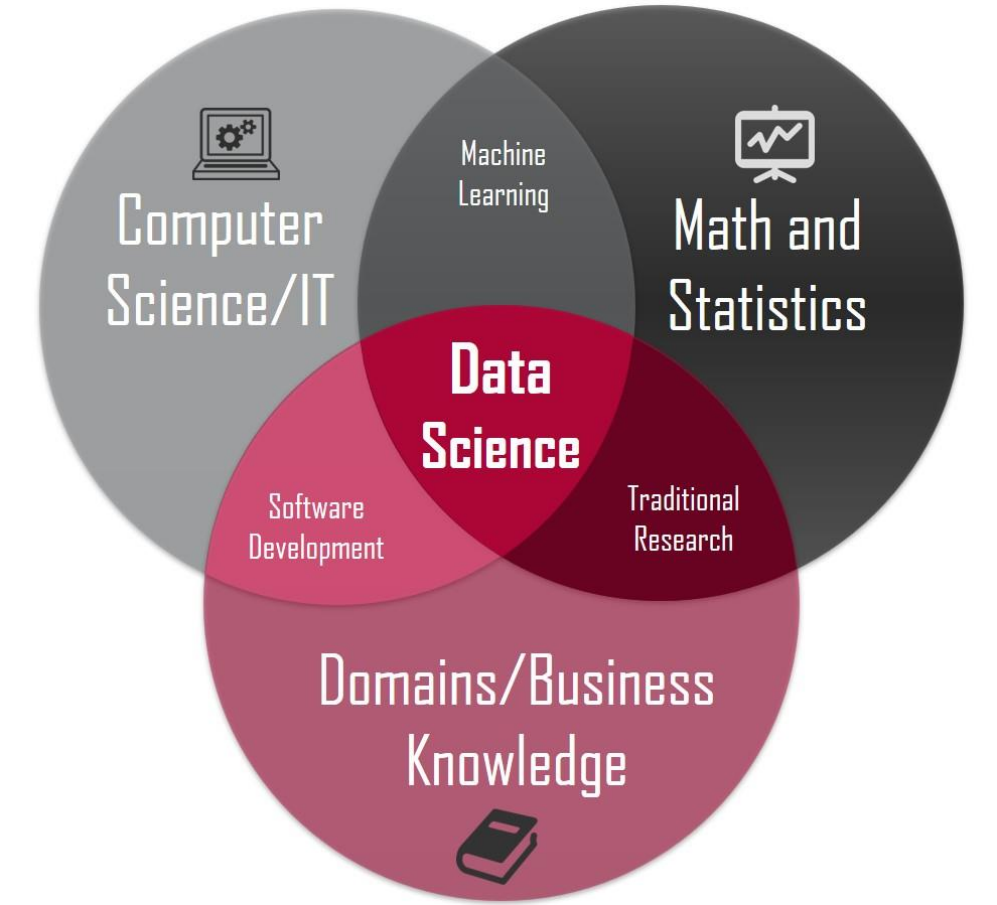
Was ist Data Science?

Warum Data Science?

Methoden

Bewertung

Data Science Venn Diagramm (Conway 2013)



Job-Profil Data Scientist (glassdoor.de):

Data scientists utilize their analytical, statistical, and programming skills to collect, analyze, and interpret large data sets…

Data scientists commonly have a … degree in statistics, math, computer science, or economics…

Data scientists have a wide range of technical competencies including: statistics and machine learning, coding languages, databases, machine learning, and reporting technologies…

Harvard Business Review 2012:

"Data Science: The Sexiest Job of the 21st Century"

Was ist Data Science?

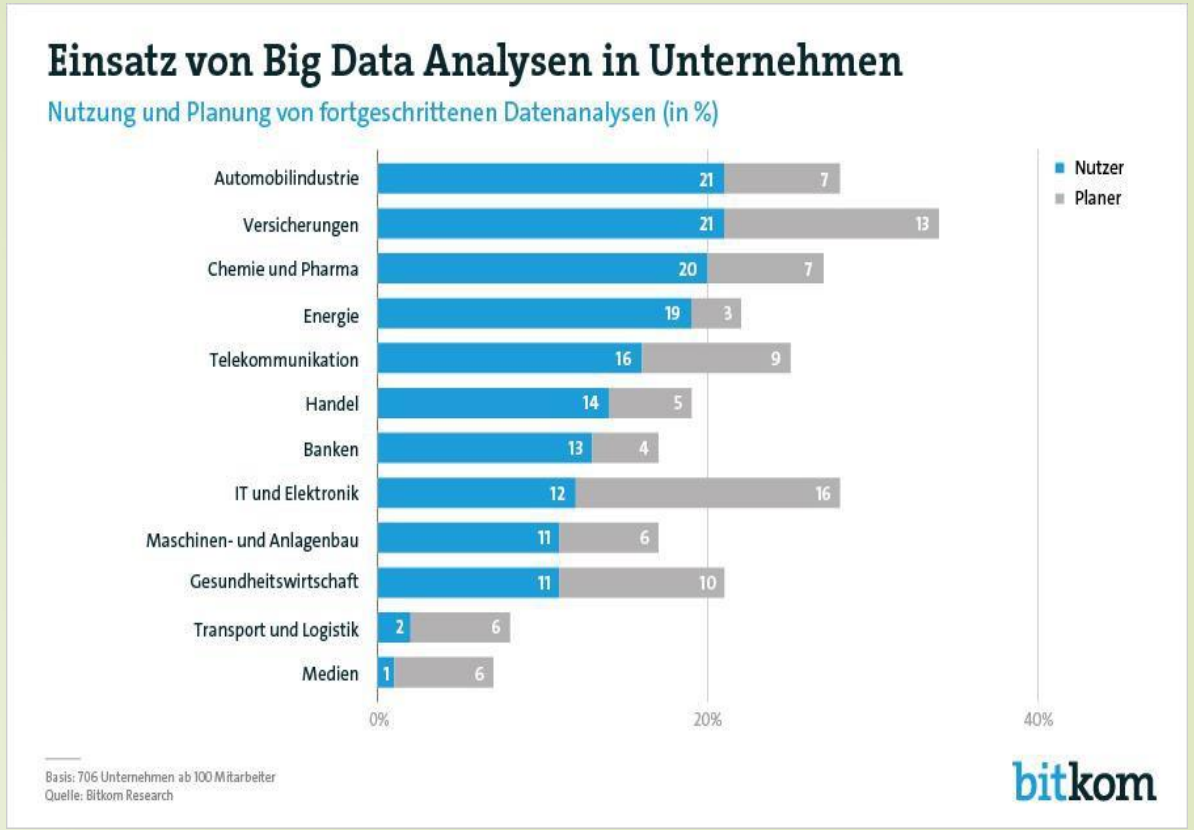
Warum Data Science?

Methoden

Bewertung

Typische Fragestellungen in der Praxis

- **Handel**: Kaufverhalten von Kunden
- **Finanzwesen**: Risikoanalyse bei Kreditvergabe
- **Produktion**: Vorhersage von Maschinenausfällen
- **Medizin**: Diagnostik, Mustererkennung in medizinischen Bildern
- **Umwelt/Gesellschaft**: Wettervorhersage, Pandemievorhersage



Was ist Data Science?

Warum Data Science?

Methoden

Bewertung

Data Science: Prozessschritte

- Eine Data Science-Studie besteht aus vier Schritten (Ng & Soo 2018):

1. Datenaufbereitung
- Vereinheitlichung unterschiedlicher Datenformate und –typen
- Auswahl relevanter Variablen unter vielen
- Umgang mit fehlenden Daten

Analysequalität zentral abhängig von Datenqualität!

2. Auswahl des Algorithmus (Machine Learning als Teilgebiet der KI)
- Versteckte Muster in den Daten entdecken („Unüberwachtes Lernen")
- Vorhersagen basierend auf bekannten Mustern („Überwachtes Lernen")
- Vorhersagen basierend auf neuen Daten verbessern („Bestärkendes Lernen")

Der Algorithmus muss zu der jeweils zu untersuchenden Fragestellung passen!

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Data Science: Prozessschritte

- Eine Data Science-Studie besteht aus vier Schritten (Ng & Soo 2018):

3. Abstimmung der Modellparameter für den gewählten Algorithmus
- Überanpassung („overfit"): Zufall wird als Muster fehlinterpretiert
- Unteranpassung („underfit"): Ignoranz offensichtlicher Strukturen

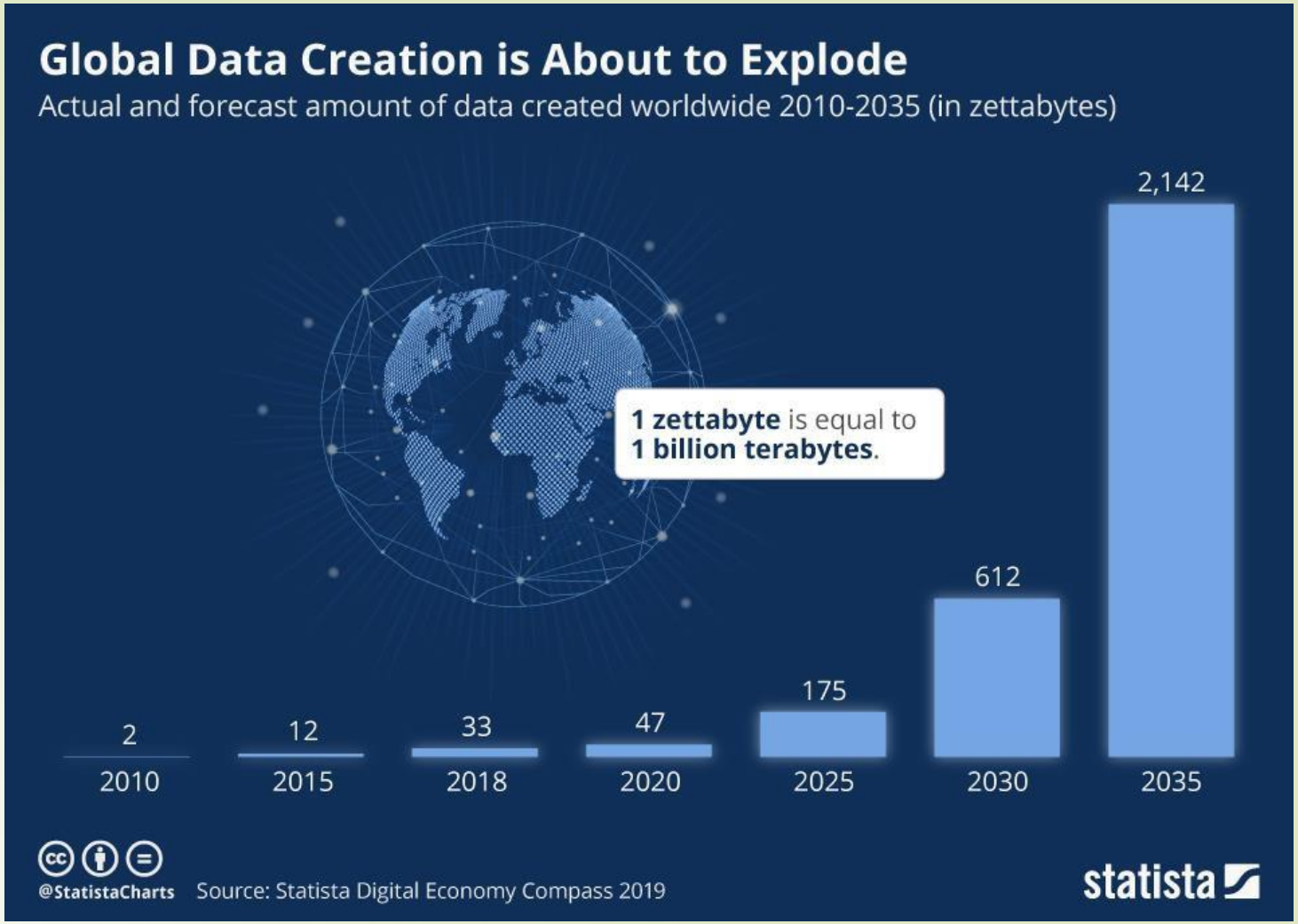
Balance zwischen dem Aufspüren wichtiger Trends und dem Ignorieren kleinerer Schwankungen!

4. Evaluation der Ergebnisse (Bewertungsmetriken)
- Prozentsatz korrekter Vorhersagen, Wahrheitsmatrix, Quadratisches Mittel
- Validierung: Prüfung der Genauigkeit an Hand neuer Daten

Modelle mit schlechter Bewertungsmetrik sind nutzlos!

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Aktuelle Trends



Was ist Data Science?

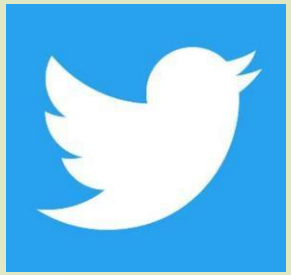
Warum Data Science?

Methoden

Bewertung

Aktuelle Trends

- 100 Mrd. Nachrichten pro Tag
- 2 Mrd. aktive Nutzer monatlich



- 500 Mio. Tweets pro Tag
- 125 Mio. Nutzer täglich
- über 140.000 Tweets pro Sekunde bei besonderen Ereignissen

- Über 200 Mio. Reviews zu über 2 Mio. Unternehmen
- Textform, teils veraltet oder gefälscht

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

„Big Data“

Die vier V's von Big Data (Gartner 2011, IBM 2012):

- **Volume**: Riesige Datenmengen
- **Variety:** Vielfältige Datenmengen
- **Velocity:** Hohe Geschwindigkeit von Datenflüssen
- **Veracity:** Verschiedene Stufen von Unsicherheit in den Daten

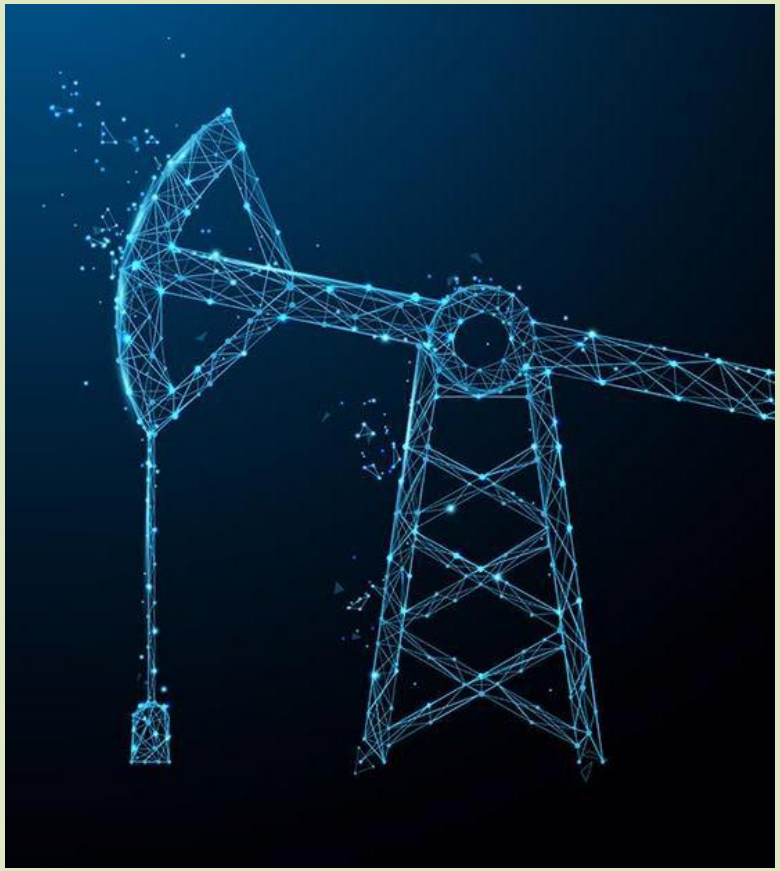
Daten liegen außerdem immer häufiger in unstrukturierter Form vor!

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Nutzen von Big Data

- **„Information is the oil of the 21st century, and analytics is the combustion engine"**
  – Peter Sondergaard, Senior Vice President, Gartner Research
- Traditionelle Analysemethoden und Softwaresysteme stoßen mit Big Data an ihre Grenzen

Wettbewerbsvorteil beruht auf der schnellen und effizienten Verwertung von Daten!



Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

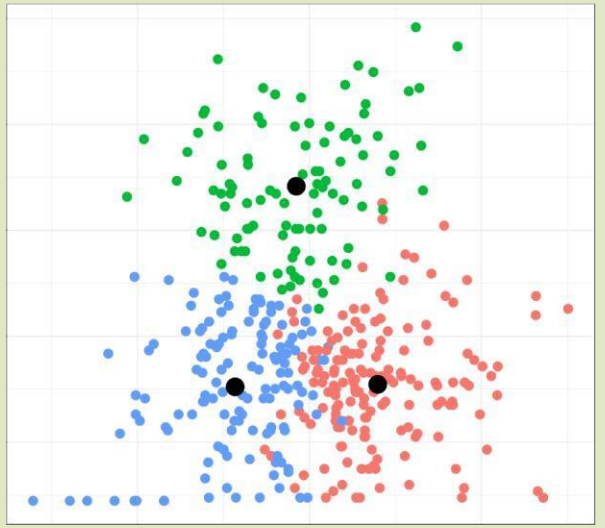
Methoden und Techniken des Data Science

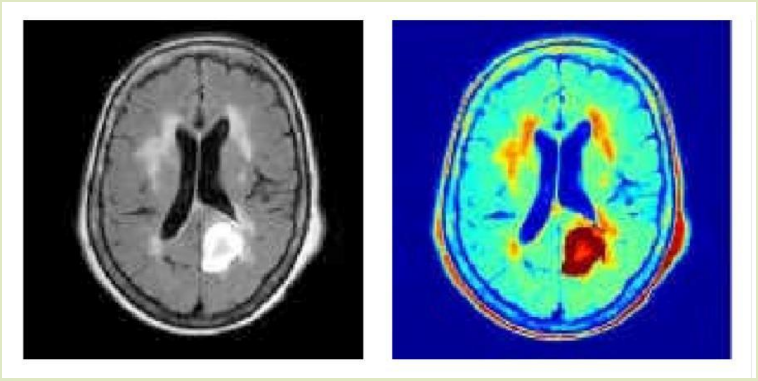
Methodenklasse	Algorithmus
Unüberwachtes Lernen	**k-Means-Clustering**
	Hauptkomponentenanalyse
	Assoziationsanalyse
	Soziale Netzwerkanalyse
Überwachtes Lernen	**Klassifikation/Regressionsanalyse**
	k-nächste Nachbarn
	Support-Vektor-Maschine
	Entscheidungsbaum
	Random Forests
	Neuronale Netze
Bestärkendes Lernen	A/B-Test, vielarmige Banditen

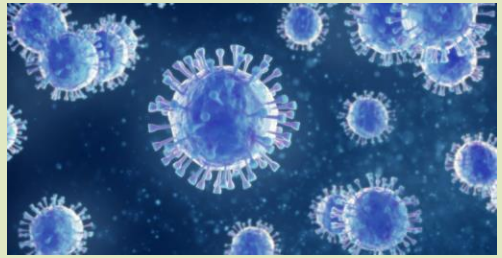
Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

k-Means-Clustering

- Aufteilung von Datenpunkten (z.B. Kunden) in möglichst **homogene Gruppen**
- **Gruppen** sind a priori **nicht bekannt**
- Data Scientist muss jedoch **Anzahl an Gruppen k vorab definieren**
- **Vorgehen:**
  - weise jeden Datenpunkt einem vorläufigen Cluster-Mittelpunkt zu
  - korrigiere den Cluster-Mittelpunkt an Hand der zugewiesenen Punkte
  - wiederhole dies solange, bis es keine Änderung mehr gibt
- **Anwendungsbeispiele:**
  - Kundensegmentierung (z.B. basierend auf demografischen Merkmalen)
  - Tumorerkennung in MRT-Bildern, Genom-Analyse
  - Erkennung von Pandemien







Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Assoziationsanalyse

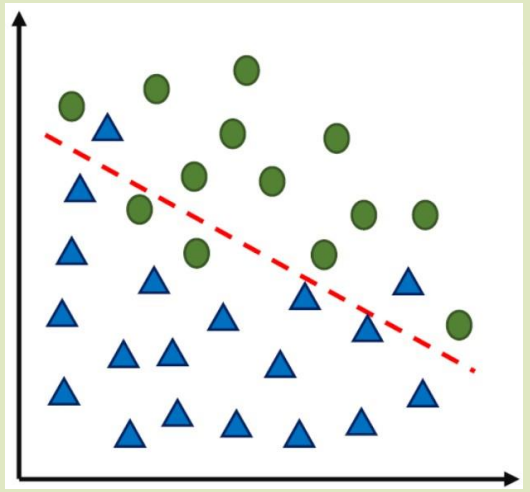
- Deckt auf, ob und wie Objekte mit anderen assoziiert sind, z.B: **Werden bestimmte Produkte gehäuft mit bestimmten anderen Produkten gekauft?**
- Vorgehen mittels **Assoziationsregeln**: Support, Konfidenz, Lift
- **Anwendungsbeispiele:**
  - Werbemaßnahmen (zugeschnittene Werbung, Produkte im selben Regal, kombinierte Produkte
  - Verbesserte Diagnostik und Behandlung von Patienten mit komorbiden Symptomen

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Klassifikation/Regression

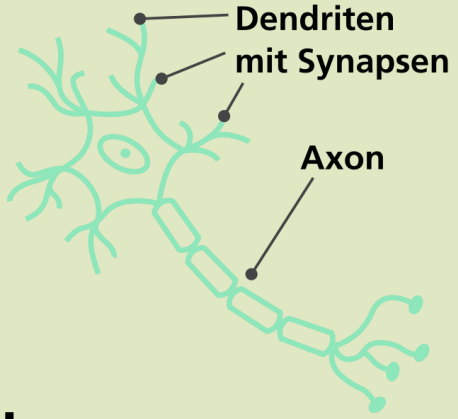
- Ableitung von **Vorhersagen zu bekannten Mustern**
- Vorgehen:
  - Algorithmus **lernt** aus vorgegebenen Mustern in **Trainingsdaten**
  - Algorithmus liefert dann **Prognosen für neue Datenpunkte**
- Input: **Prädiktorvariablen**
- Output:
  - **Binäre**/kategoriable Zielvariable: **Klassifikation**
  - **Diskrete/stetige** Zielvariable: **Regression**

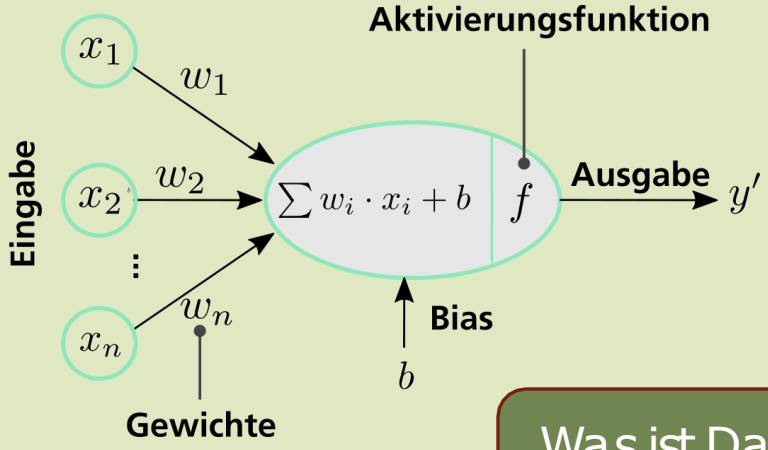
- **Anwendungsbeispiele:**
  - Entscheidung über Kreditvergabe
  - Diagnostik medizinischer Bilder



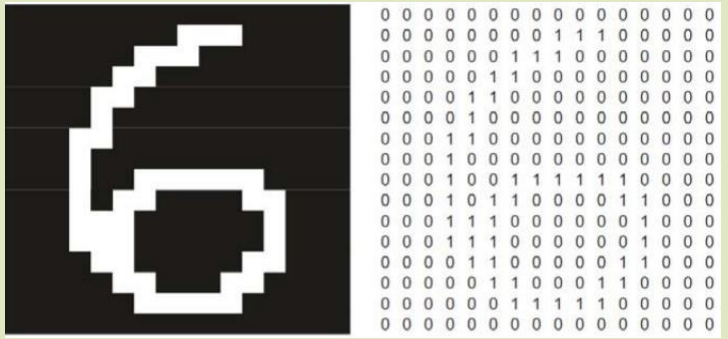
Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Neuronale Netze

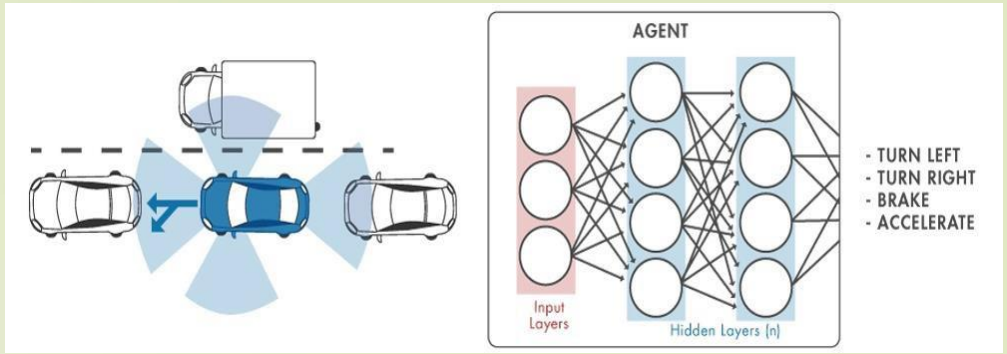




- Neuronen des Netzes sind in **Schichten** (*layers)* angeordnet (**input layer, hidden layer, output layer**)
- Output einer Schicht ist jeweils Input für die nächste Schicht
- **Aktivierungsregeln** bestimmen, wann Neuronen aktiv werden (Schwellenwerte)



- **Anwendungsbeispiele:**
  - Sprach-, Bild- und Texterkennung
  - Autonomes Fahren



Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Data Science Software

- Programmiersprachen: R, Python

- Softwarepakete: RapidMiner, Spark, Hadoop

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Wirtschaftliches und gesellschaftliches Potenzial

- Realisierung innovativer Produkt- und Geschäftsideen
- 1:1-Marketing bei Kundensegmentierung
- Verbesserte Entscheidungsfindung
- Reduktion von Bearbeitungszeiten
- Verbesserung von Diagnose und Behandlung in der Medizin
- Präventive Maßnahmen bei einschneidenden Ereignissen

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Methodische Risiken

- **Probleme bei sehr seltenen Ereignissen**

  Geheimdienst: Bei 800 Mio. Dokumenten täglich in Deutschland erkennt ein Data Science-Algorithmus nicht-verdächtige Dokumente mit 99,99% Sicherheit

  → ca. 80.000 Dokumente würden pro Tag **fälschlicherweise** als verdächtig eingestuft werden

- **Korrelation vs. Kausalität!**

  Existierende Zusammenhänge deuten nicht zwingend auf einen Ursache-Wirkungs-Mechanismus hin, Gefahr der Fehlinterpretation

- **„Black-Box"-Charakter**

  Die **Entscheidungsfindung** vieler Modelle ist **nicht nachvollziehbar**, dadurch potenziell fehlende Akzeptanz

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Soziale und ethische Risiken

- **„Filterblase“ (Pariser 2011)**

  Durch personalisiertes Internet werden Nutzer potenziell in einer Informationsblase intellektuell isoliert

  ➡ negative Folgen für Diskurs in der Zivilgesellschaft

- **„Echokammer“ (Sunstein 2001)**

  Spiraleffekt aus **einseitiger Mediennutzung** und **polarisierender Gesellschaft** (unnatürliche Verstärkung der „Meinung unter Gleichgesinnten“)

- **Datenschutz**

  Viele der verwendete Daten sind personenbezogen (**„gläserner Mensch“**)

Was ist Data Science?
Warum Data Science?
Methoden
Bewertung

Zusammenfassung

- Data Science ist eine **interdisziplinäre Wissenschaft** (u.a. Informatik, Mathematik, Wirtschaftswissenschaften), die sich mit der **Extraktion von Wissen aus Daten** beschäftigt
- Digitalisierung und **Big Data**-Phänomen führen zu einer **Bedeutungszunahme von Data Science und verknüpften Berufsfeldern**
- **Methoden** der Data Science sind im Wesentlich in **unüberwachtes Lernen** („Welche Muster sind in meinen Daten enthalten?“) und **überwachtes Lernen** („Welche Prognosen können basierend auf bekannten Mustern abgeleitet werden?“) zu unterscheiden
- Hohe **ökonomische und gesellschaftliche Potenziale** stehen **methodischen und ethischen Risiken** gegenüber, der Data Scientist trägt daher **enorme Verantwortung** („nicht alles was man tun kann, sollte getan werden“)

Was ist Data Science?

Warum Data Science?

Methoden

Bewertung