
Einführung in Web- und Data-Science

Begriffsbestimmungen

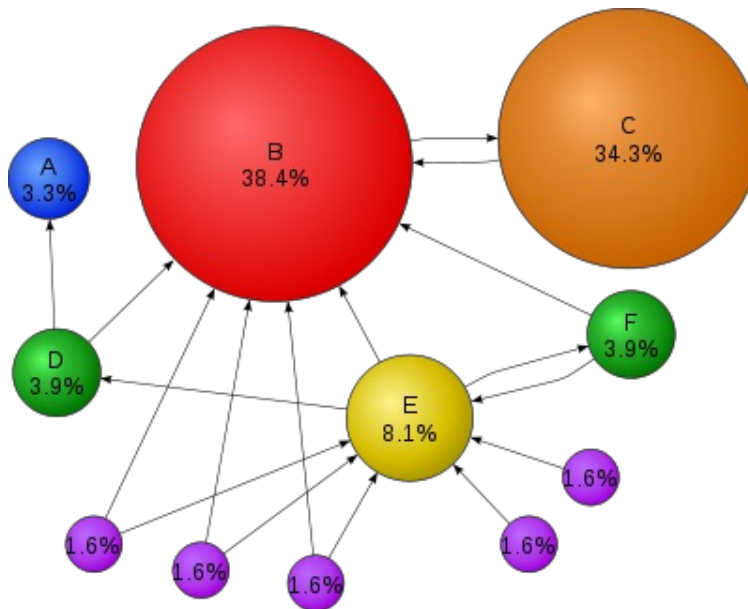
Dr. Marcel Gehrke

Universität zu Lübeck

Institut für Informationssysteme

Web und Data Science

- Web Science
 - Analyse von Strukturen im Web (Mensch und Computer)
 - Formalisierung durch große Graphstrukturen und entsprechende Entscheidungsprobleme über Graphen
 - Beispiel: Pagerank (Bewertung von Webseiten)



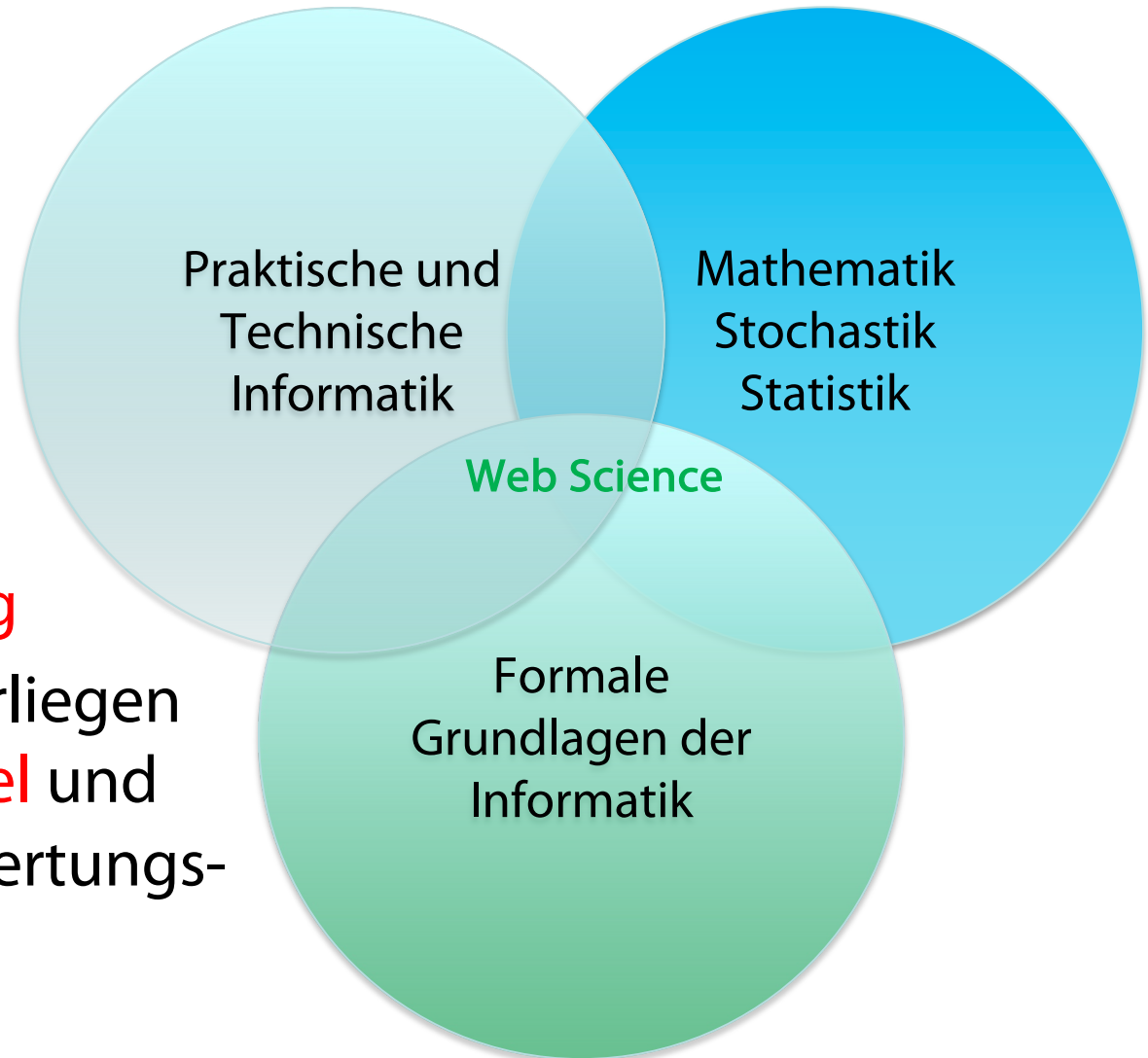
Zufallssurfer-Modell:

Größe der Kreise in etwa proportional der relativen Häufigkeit, mit der sich ein Surfer auf einer Seite befindet

[Wikipedia]

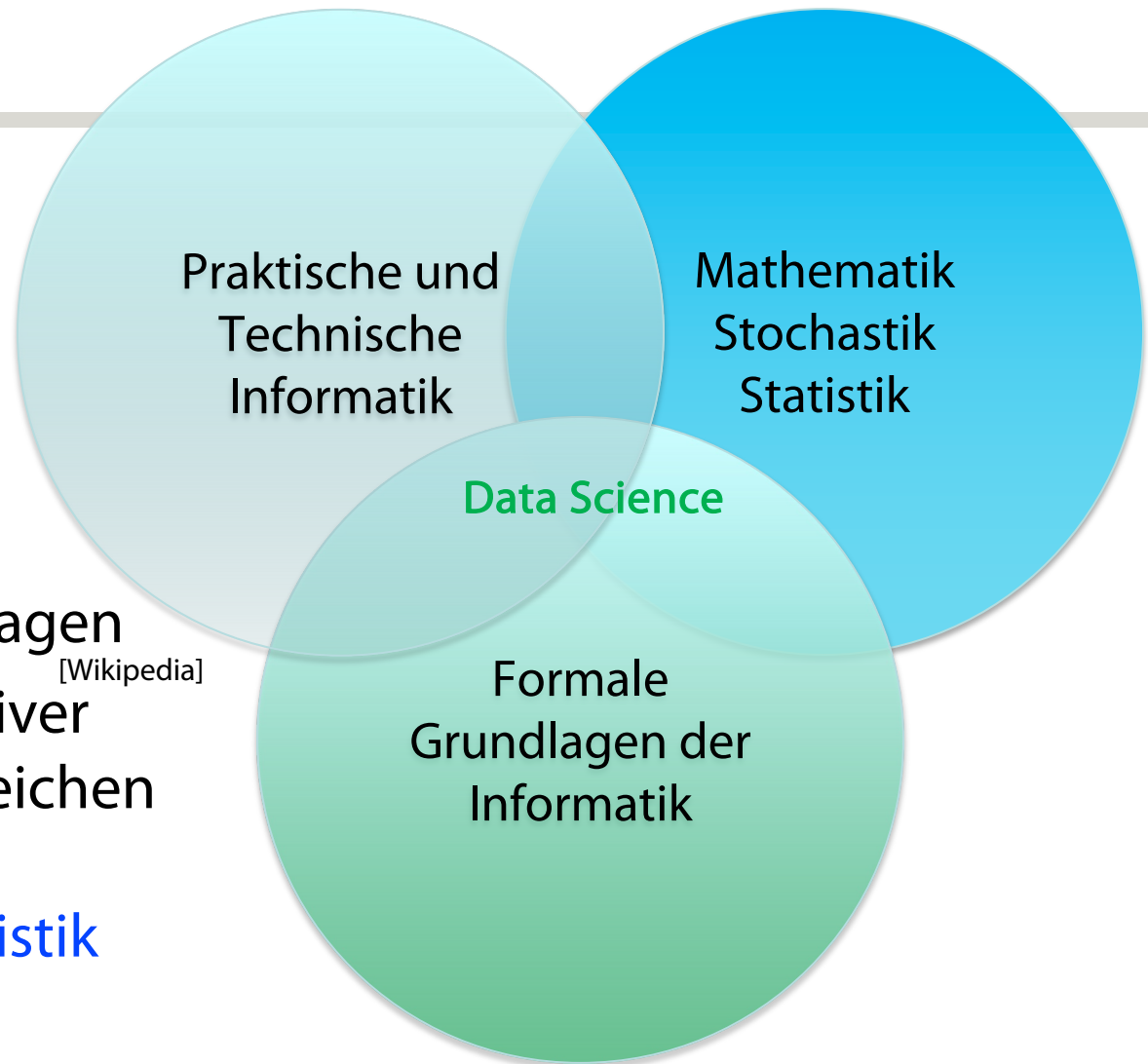
Web Science

- Graphstrukturen extrem **groß**
- Verfahren zur Lösung von Entscheidungsproblemen extrem **aufwendig**
- Graphdaten unterliegen ständigem **Wandel** und so auch die Auswertungsergebnisse



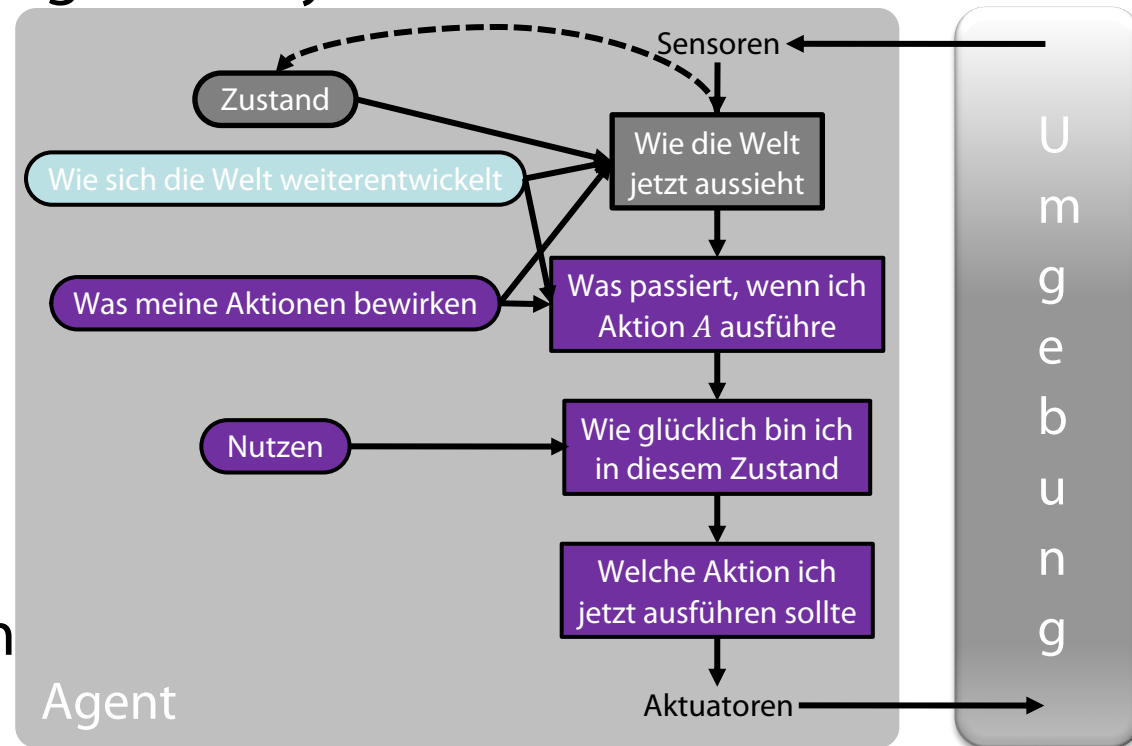
Data Science

- Extraktion von Wissen aus Daten (u.a. Graphdaten)
- Begriff schon vor 60 Jahren für Informatik vorgeschlagen [Wikipedia]
- Entwicklung innovativer Konzepte in den Bereichen **Logik, Datenbanken und Stochastik / Statistik** (Datenanalyse und Wissensentdeckung)
- Verwendung von **LADS und Analysis**



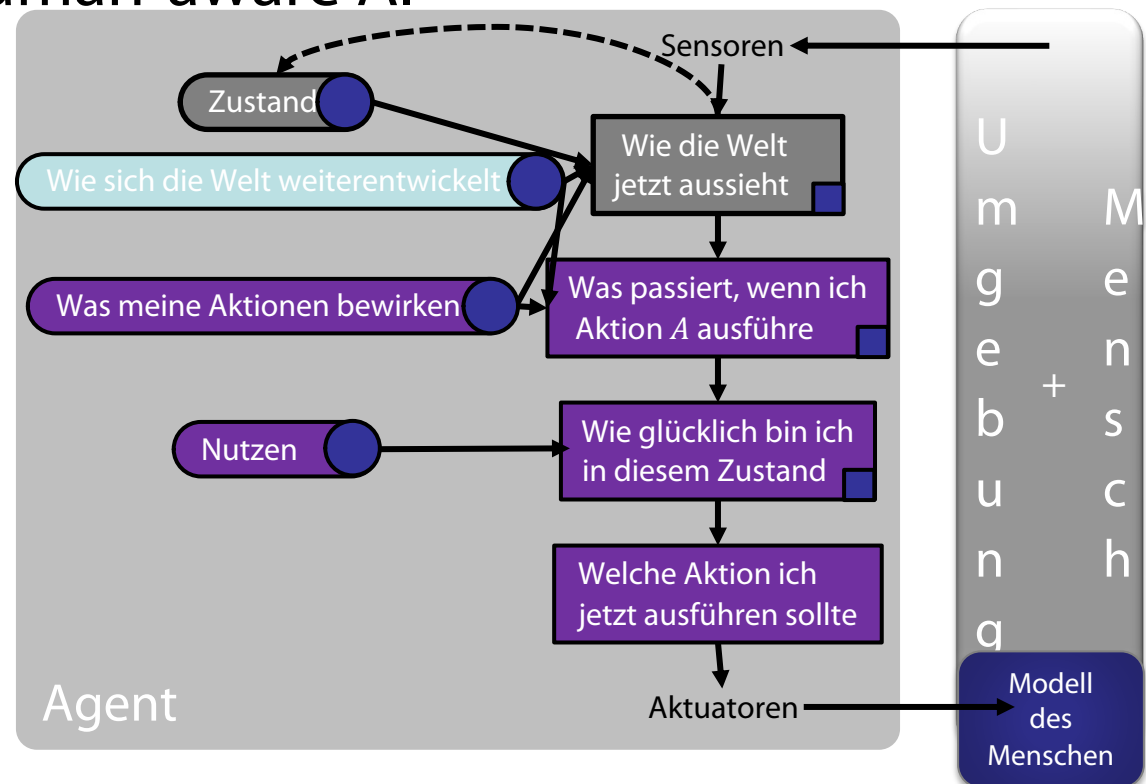
... und was ist mit Künstlicher Intelligenz?

- Wissenschaft der intelligenten Systeme
- Agenten
 - Haben/bilden Ziele
 - Sensoren/Aktoren
 - Handlungsplanung
 - Lernen zur Laufzeit
- Mechanismen
 - Globale Kooperation von Agenten zur Erreichung eines gemeinsamen Ziels
- Agenten interagieren mit Menschen (and anderen Agenten)
 - Ziele der Agenten beeinflussbar



Wissenschaft Künstliche Intelligenz

- Nur algorithmische Modellierung für Agenten?
 - Transparenz, Erklärungsfähigkeit
- Neue Aspekte: Human-aware AI
 - Erwartungskonformität
 - Beweisbar-nützliche Agenten



Daten Modelle vs. Algorithmische Modelle

Datenmodellierung

vs.

Algorithmische Modellierung

$$Y \leftarrow F(X, \text{zufälliges Rauschen}, \text{Parameter})$$



Wir verstehen die Welt ?

Wie gut funktioniert meine Datenmodellierung?

~~Statistiker, Datenanalyst, Data Miner~~

Lineare Regression

Logistische Regression

Bekannte Wahrscheinlichkeiten

Konfidenz Intervalle

Vorhersage von Variablen & Anpassungsgüte

Statistiker, Data Scientist
Maschinelles Lernen

Wir verstehen die Welt nicht ?

Die Welt produziert Daten in einer Black Box

~~Data Scientist~~

~~Maschinelles Lernen, X~~

Random Forrest, SVM

Unbekannte multivariate Verteilung

Iterativ

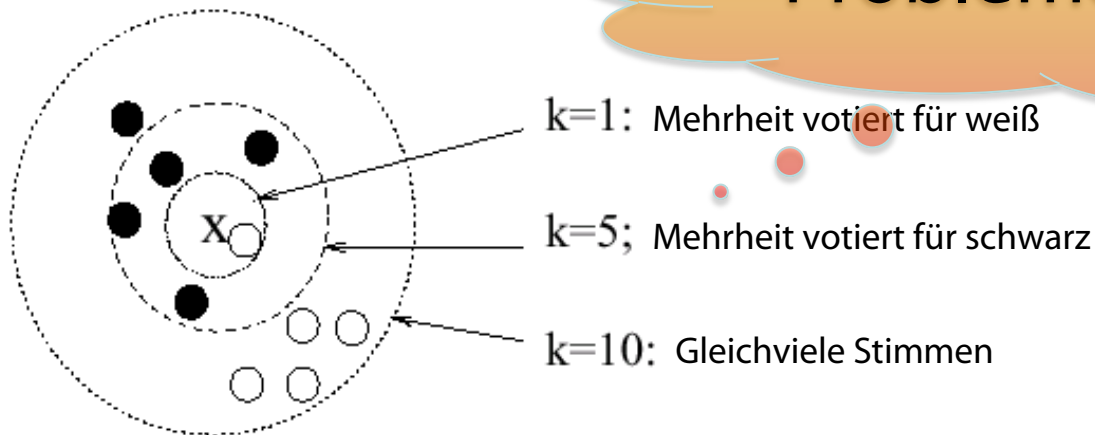
Vorhersagewahrscheinlichkeit

Web- und Data Science: Herausforderungen

- Große Datenbestände
 - Speicher und Zugriffstechnologie
- Starker Zuwachs an Daten, hohe Dynamik
 - Hohe Datenraten und Echtzeitanforderungen
- Heterogene Datenbestände
 - Verteiltes Datenmanagement
 - Datenintegration

Instanzbasierte Anfragebeantwortung

- Annahme: Gegeben viele Datenpunkte
 - Beispielmerkmale: (x, y, Farbe) , $\text{Farbe} \in \{\text{weiß}, \text{schwarz}\}$
- Anfrage: Datenpunkt ohne Wert für bestimmtes Merkmal
 - Beispiel: Merkmal Farbe ohne Wert
- Anfragebeantwortung (Klassifikation des Anfragepunkts):
Mehrheitsvotum der k-nächsten Nachbarn (kNN-Verfahren)
Probleme erkennbar?



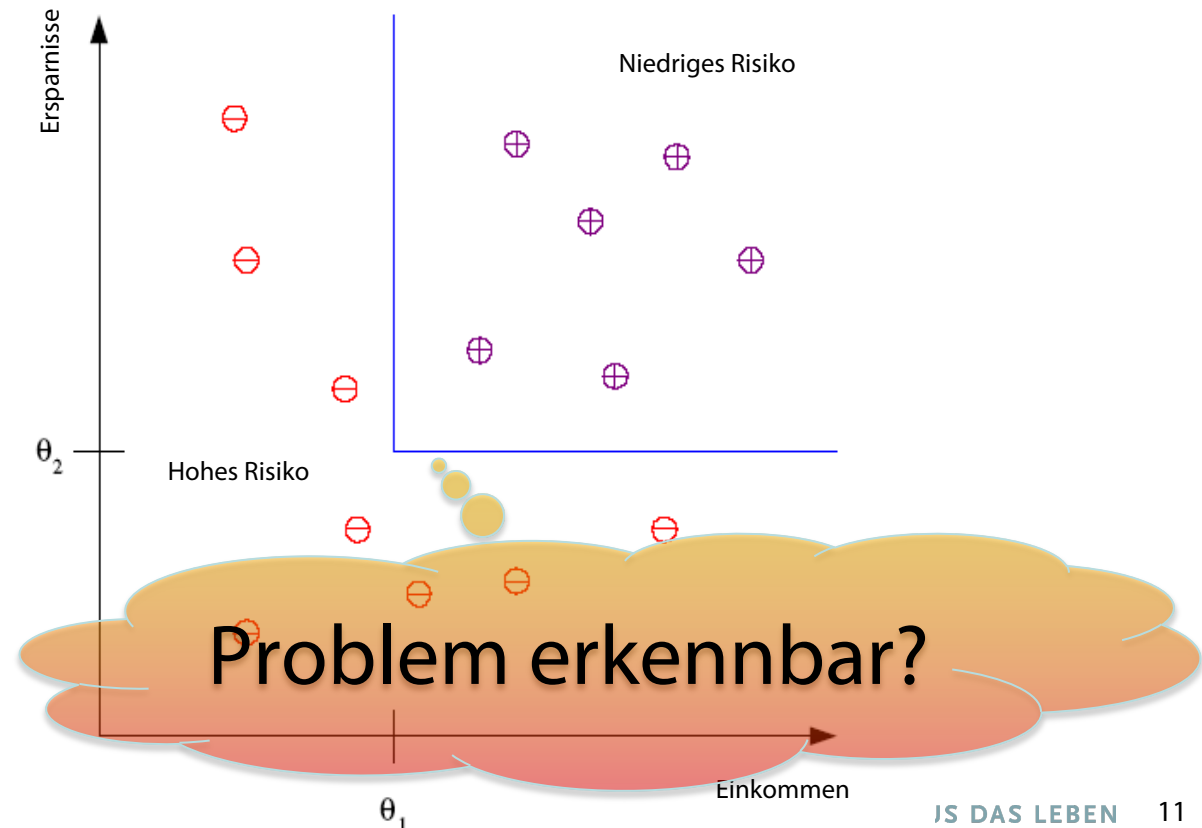
Fix, E., Hodges, J.L. Discriminatory analysis, nonparametric discrimination:
Consistency properties. Technical Report 4, USAF School of Aviation
Medicine, Randolph Field, Texas, 1951

Probleme mit kNN

- Klassifikationsergebnis stark von k abhängig
- Hoher Speicherbedarf
- Effizienter Zugriff auf "Nachbarn" erfordert weitere Maßnahmen (noch mehr Speicherbedarf)
- Klassifikation basierend auf den Daten

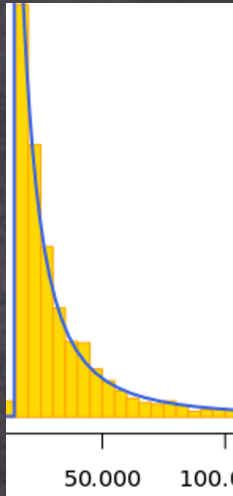
Modellbasierte Anfragebeantwortung

- Repräsentation der Daten durch Parameter eines Modells
 - Wenn $(\text{Einkommen} > \theta_1 \wedge \text{Ersparnisse} > \theta_2)$, dann kreditwürdig ($\oplus$), sonst nicht ($\ominus$)
- Nur 2 Parameter nötig: (θ_1, θ_2)
- Modell fordert geringen Speicher



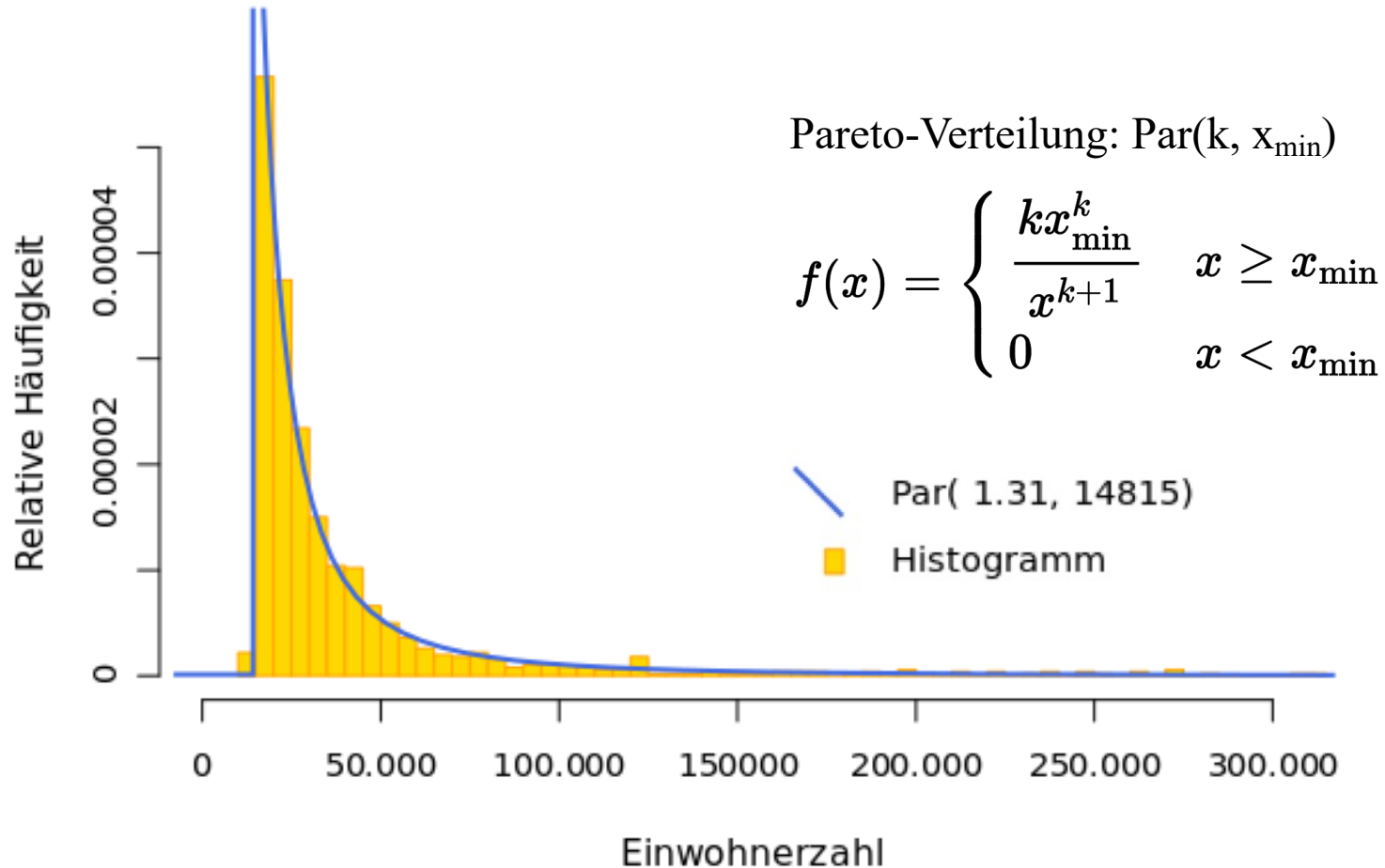
Aufgabe

- Anzahl von Städten mit bestimmten Einwohnerzahlen schätzen
- Daten: Liste von Einwohnerzahlen (auf 5000er gerundet)
- Explizites Modell: Zählerfeld aufbauen
 - Unvollständigkeit der Daten
- Implizites Modell: Potenzgesetz $y = ax^b$
 - a und b bestimmen (a positive, b negativ)
 - Aufwendiges Optimierungsproblem lösen

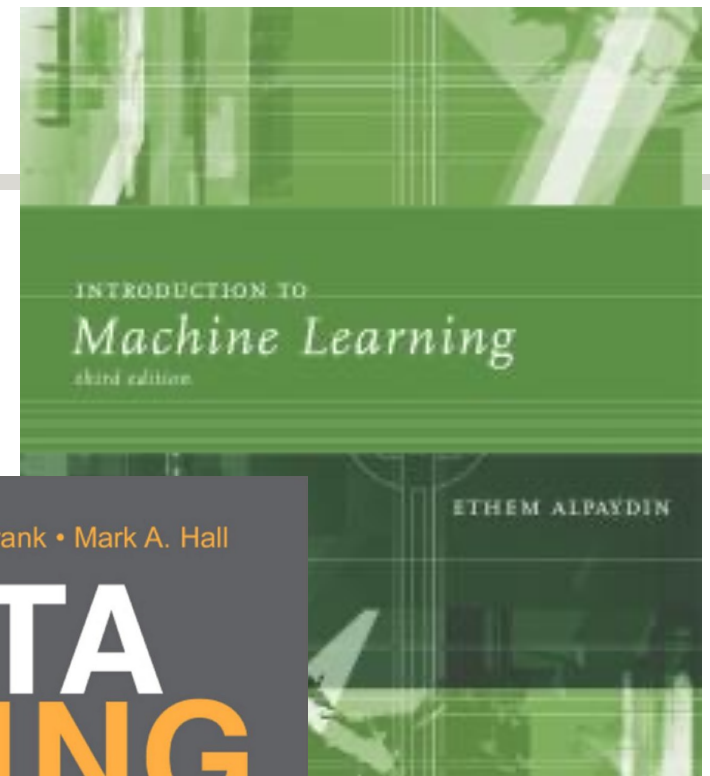
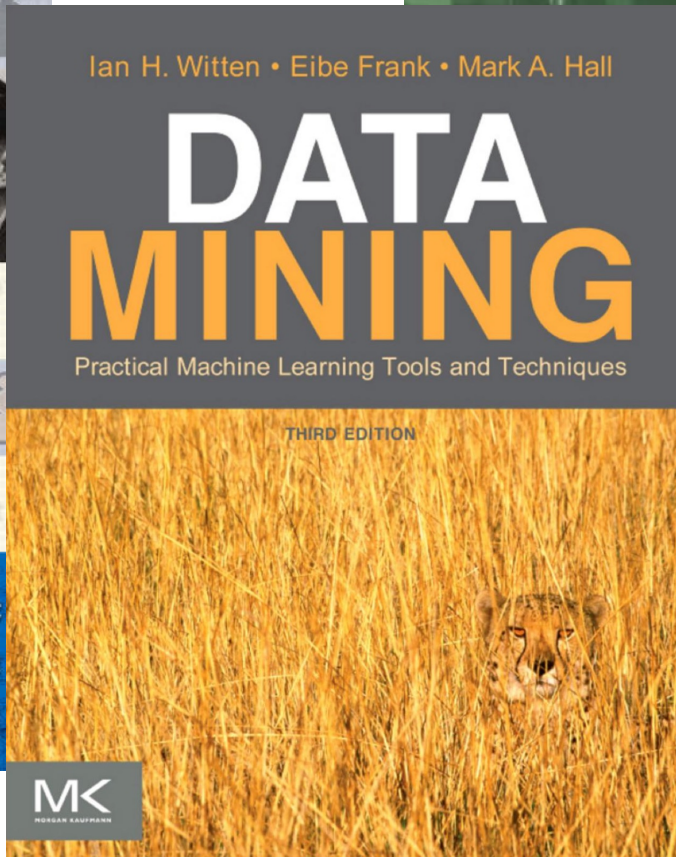
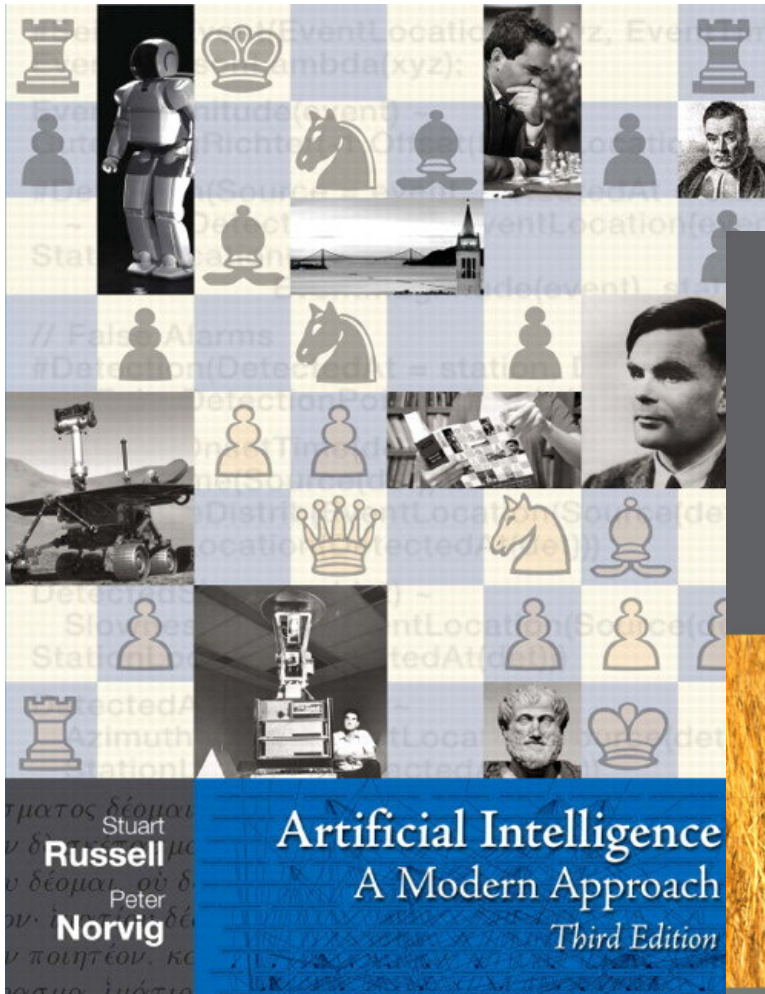


Begriff der "Verteilung"

Einwohnerzahlen Deutscher Städte



Literatur



Literatur

- Stuart Russell, Peter Norvig, **Artificial Intelligence – A Modern Approach**, Pearson, 2009 (oder 2003er Ed.)
- Ian H. Witten, Eibe Frank, Mark A. Hall, **Data Mining: Practical Machine Learning Tools and Techniques**, Morgan Kaufmann, 2011
- Ethem Alpaydin, **Introduction to Machine Learning**, 3rd Ed., MIT Press, 2014
- Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, **Mining of Massive Datasets**, 2nd Ed., Cambridge University Press, 2014
- Viele zusätzliche Bücher, Präsentationen, und Videos im Web

