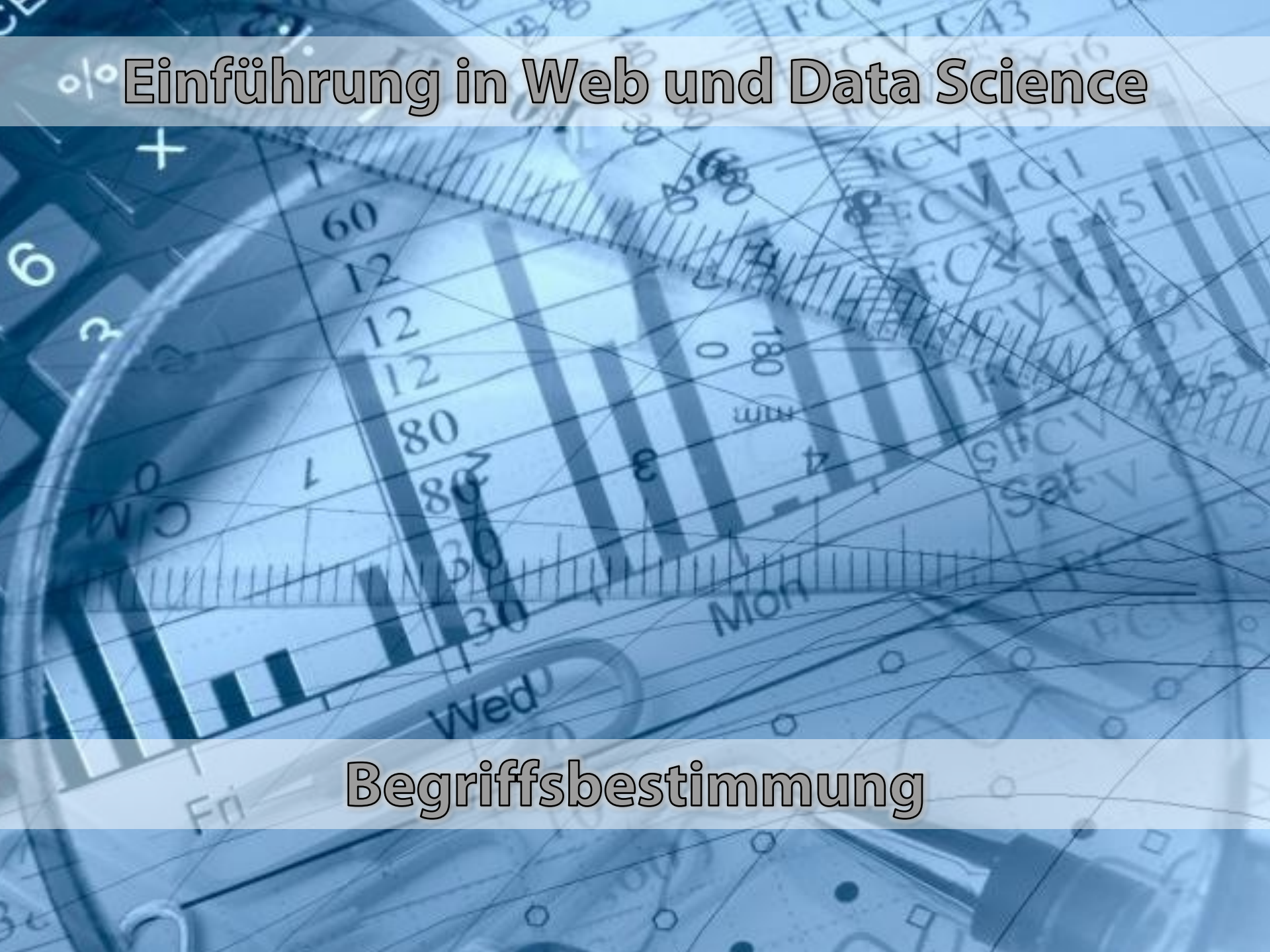
Einführung in Web- und Data-Science

Prof. Dr. Ralf Möller

Universität zu Lübeck

Institut für Informationssysteme

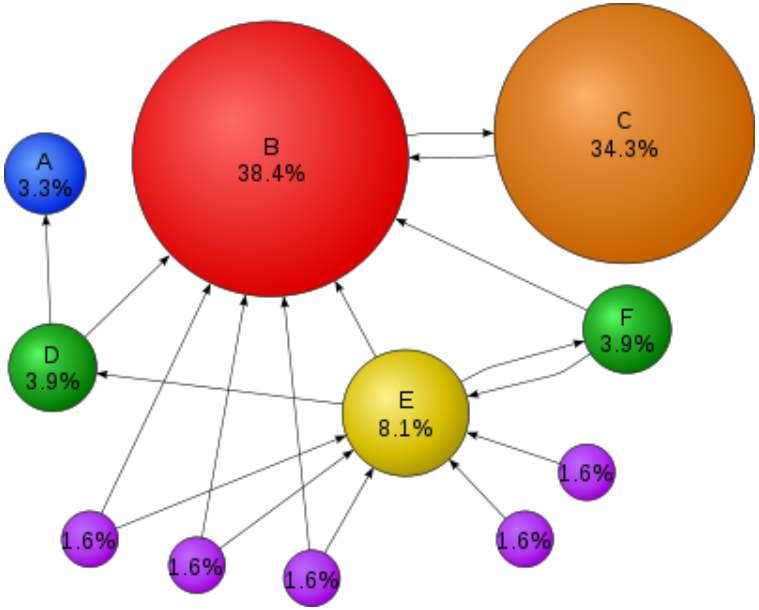
Einführung in Web und Data Science



Begriffsbestimmung

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)



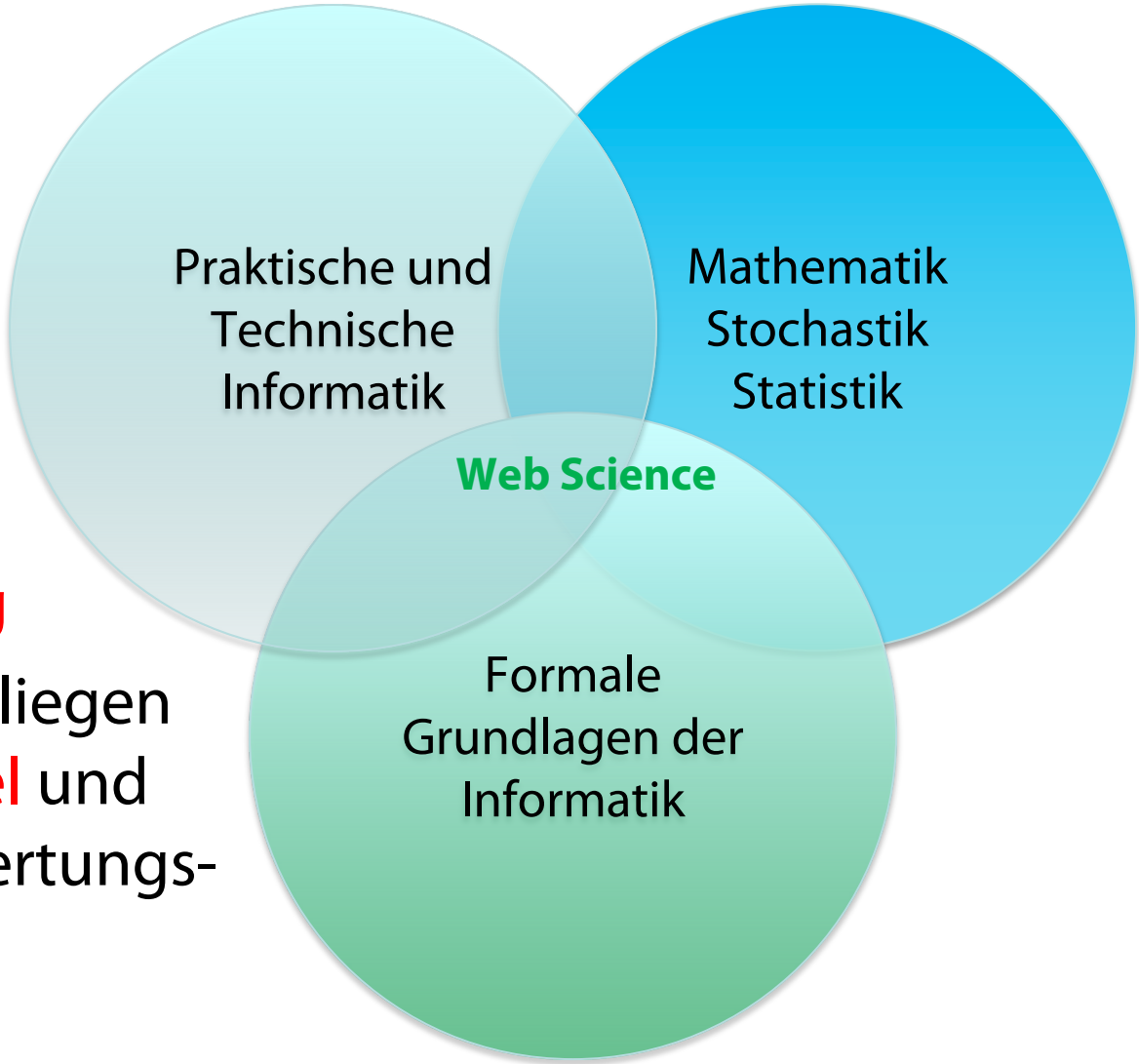
Zufallssurfer-Modell:

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

[Wikipedia]

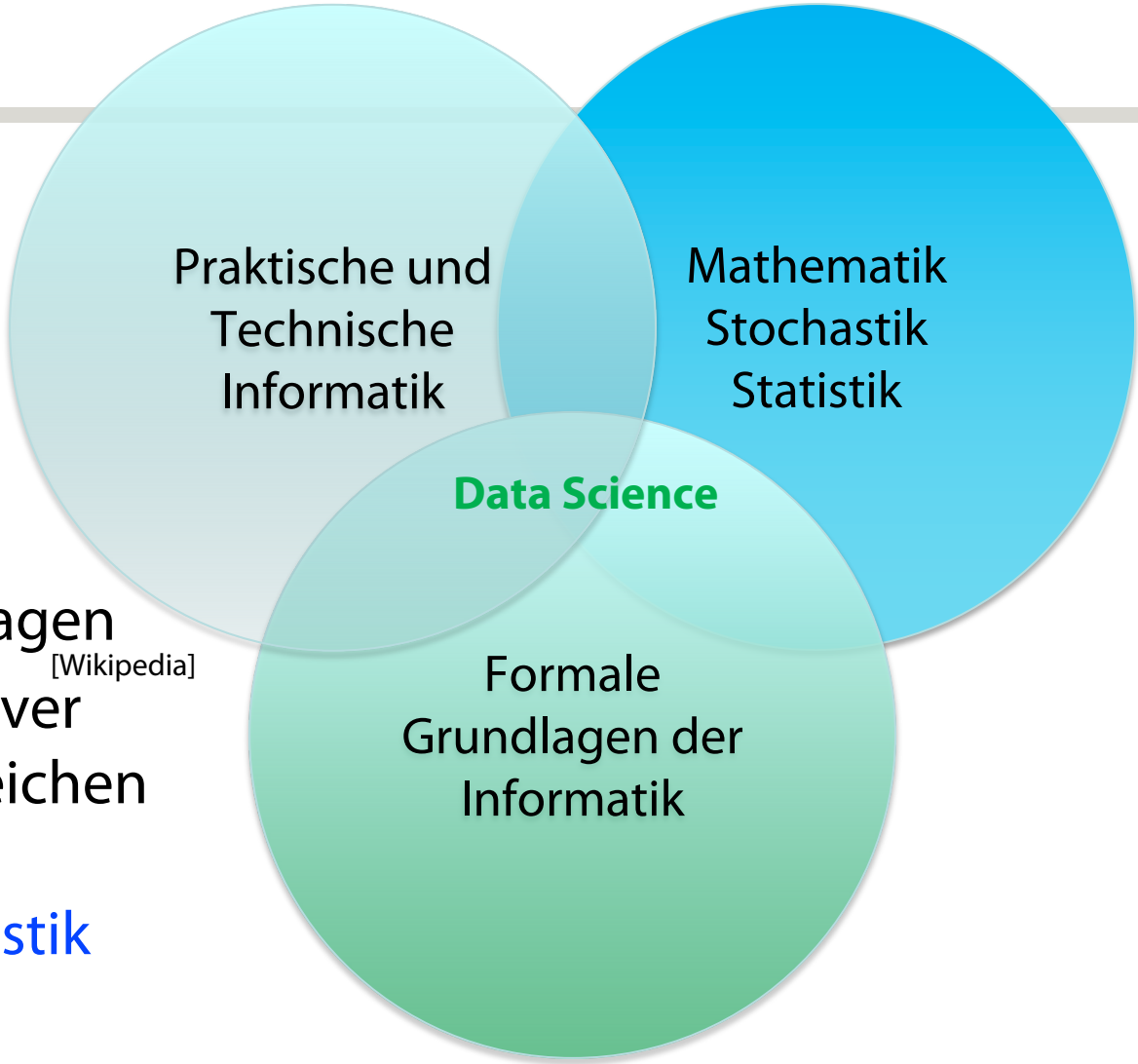
Web Science

- Graphstrukturen extrem groß
- Verfahren zur Lösung von Entscheidungs-problemen extrem aufwendig
- Graphdaten unterliegen ständigem Wandel und so auch die Auswertungs-ergebnisse



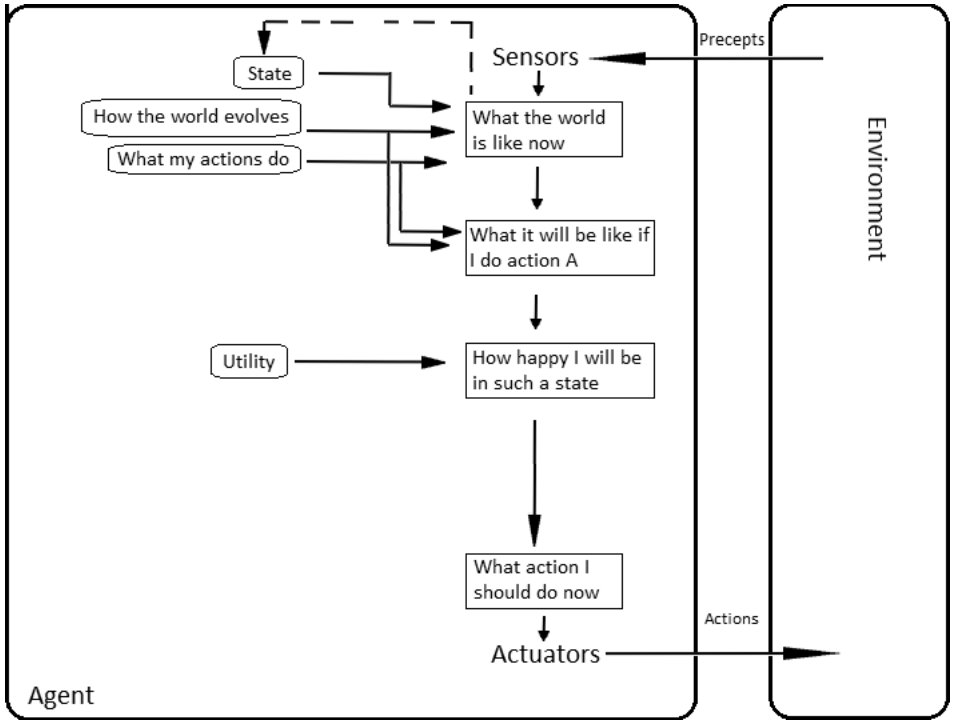
Data Science

- Extraktion von Wissen aus Daten (u.a. Graphdaten)
- Begriff schon vor 50 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?

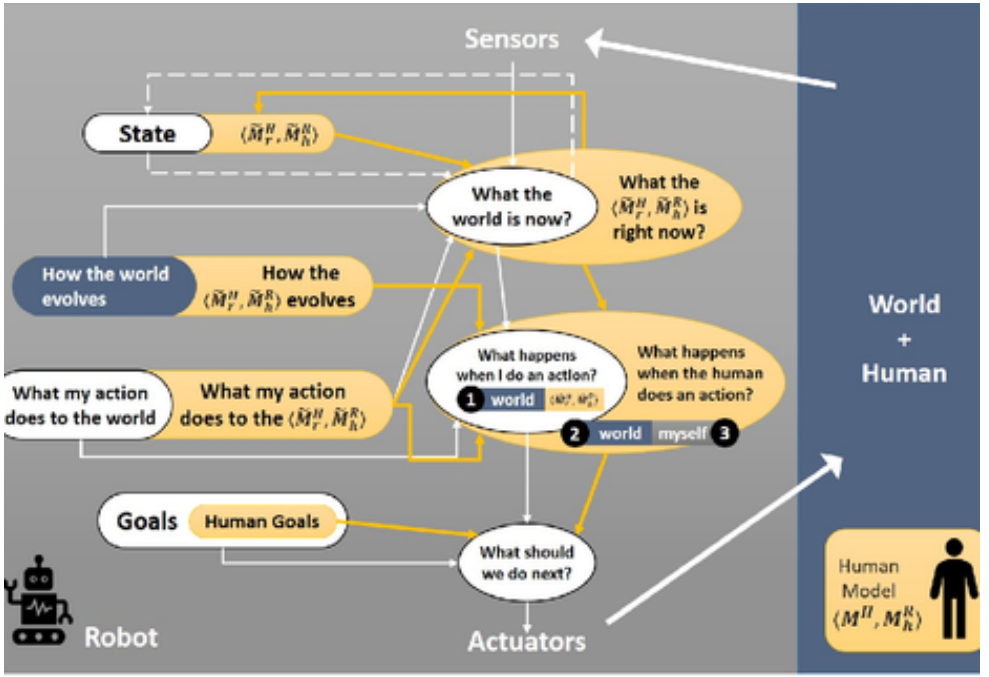
- Wissenschaft der intelligenten Systeme
- Agenten
  - Haben/bilden Ziele
  - Sensoren/Aktoren
  - Handlungsplanung
  - Lernen zur Laufzeit
- Mechanismen
  - Globale Kooperation von Agenten zur Erreichung eines gemeinsamen Ziels



- 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
  - Erwartungs-konformität
  - Beweisbar-nützliche Agenten



https://www.uni-luebeck.de/forschung/themen/informatiktechnik/artificial-intelligence.html

Data Models vs. Algorithmic Models

Data Modeling

$Y \leftarrow F(X, \text{random noise, parameters})$

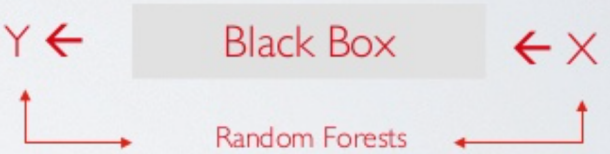
We understand the world ?

How well 'my data model' works

Linear Regression
Logistic Regression
Known Distributions
Confidence Intervals
Predictor Variables & Goodness of Fit

VS.

Algorithmic Modeling



We don't understand the world ?

The world produces data in a black-box

Random Forests, SVM
Unknown Multivariate Distributions
Iterative
Predictive Accuracy

Data Science
Statistics

Machine Learning

Data Science London

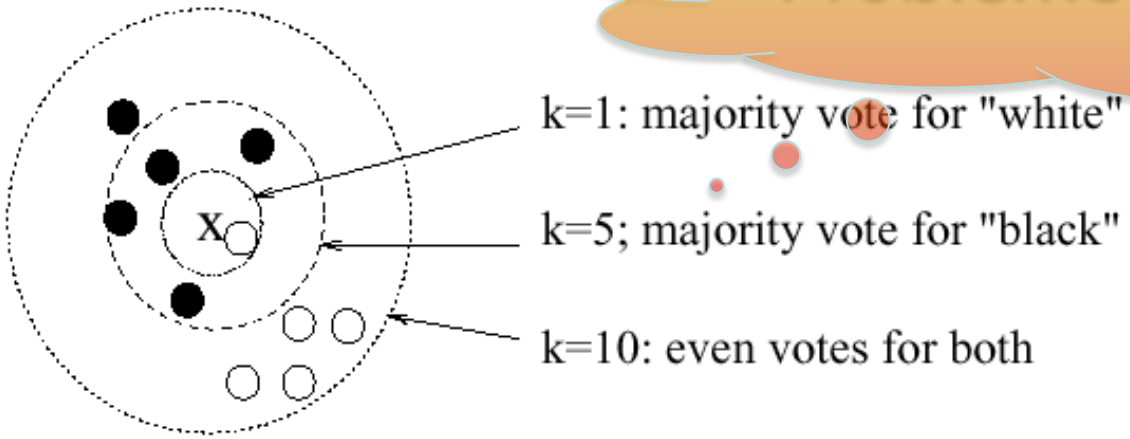
Web- und Data Science: Herausforderungen

- 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), Farbe ∈ { white, black }
- 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**

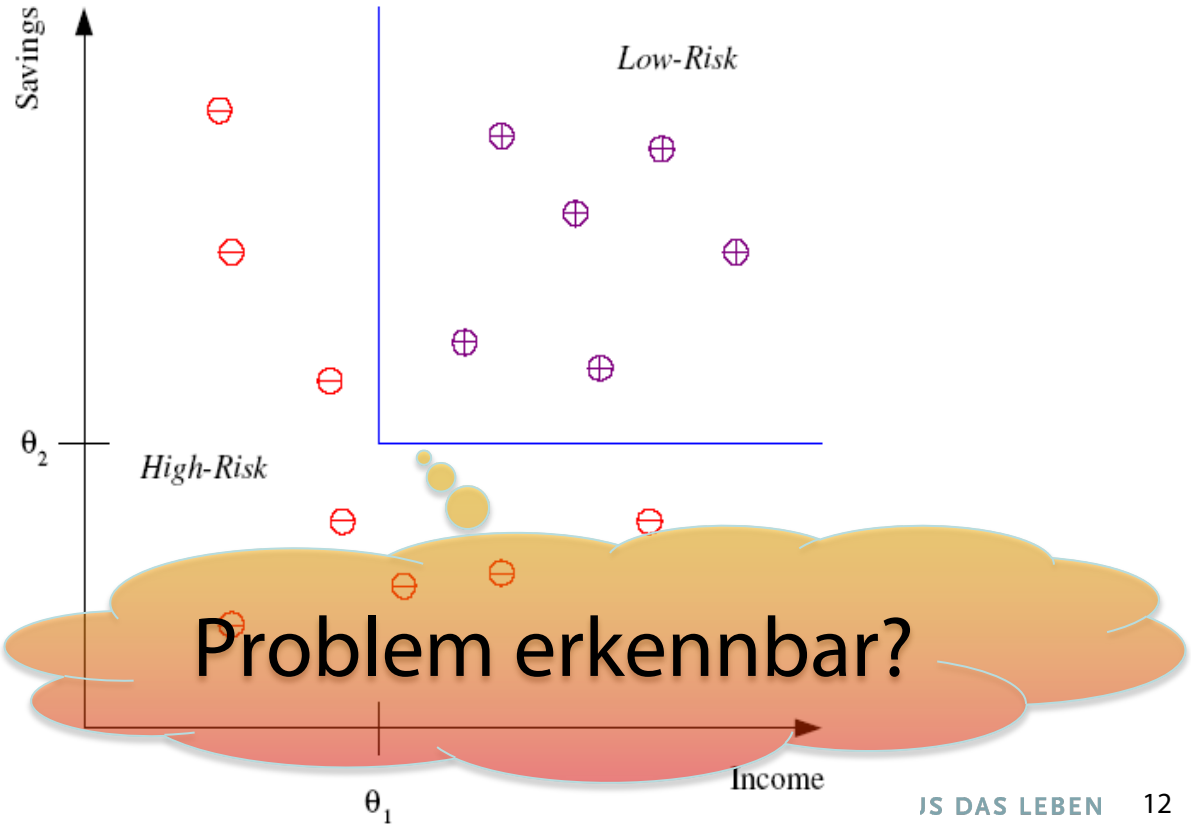
Probleme mit kNN

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

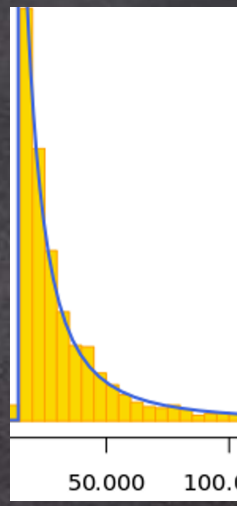
- Klassifikation basierend auf den Daten

Modellbasierte Anfragebeantwortung

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



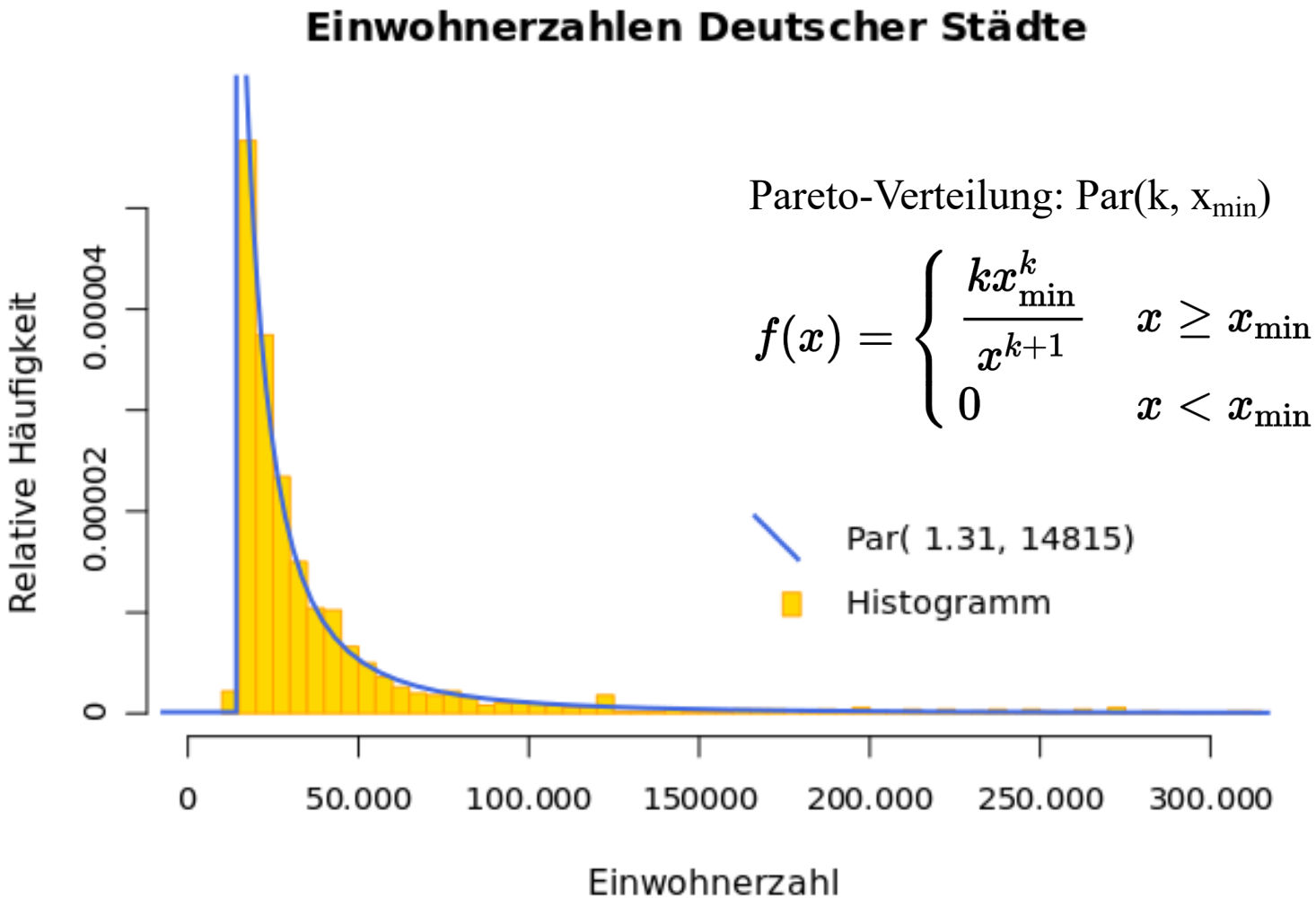
Aufgabe



- Anzahl von Städten mit bestimmten Einwohnerzahlen schätzen

- Daten: Liste von Einwohnerzahlen (auf 10000er 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"

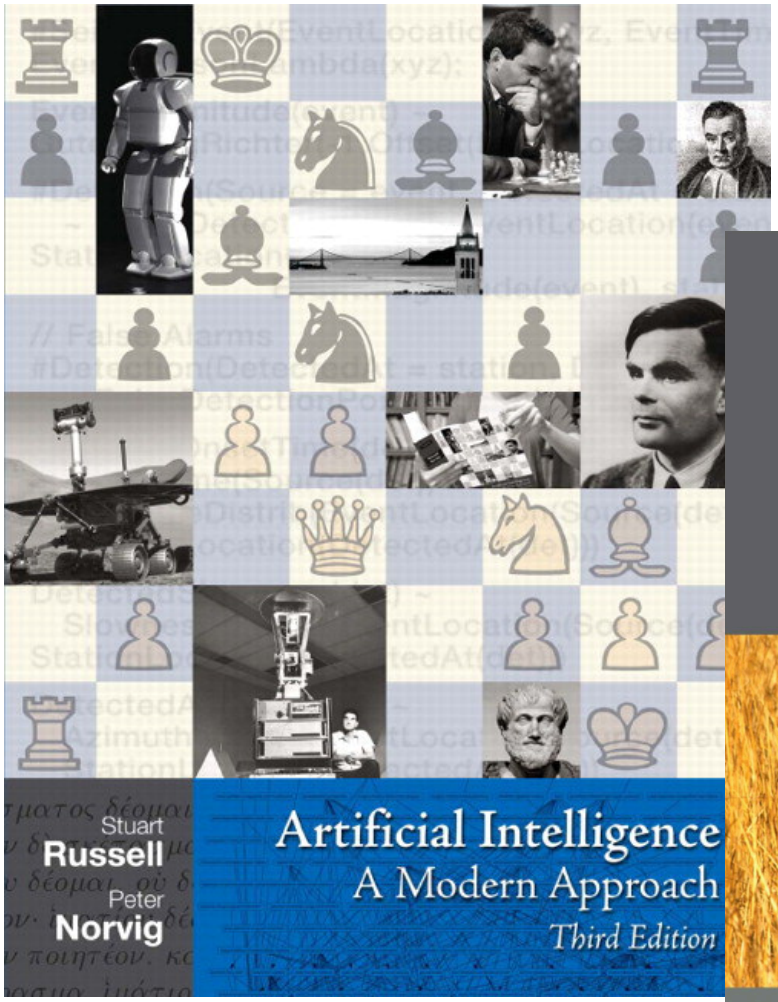


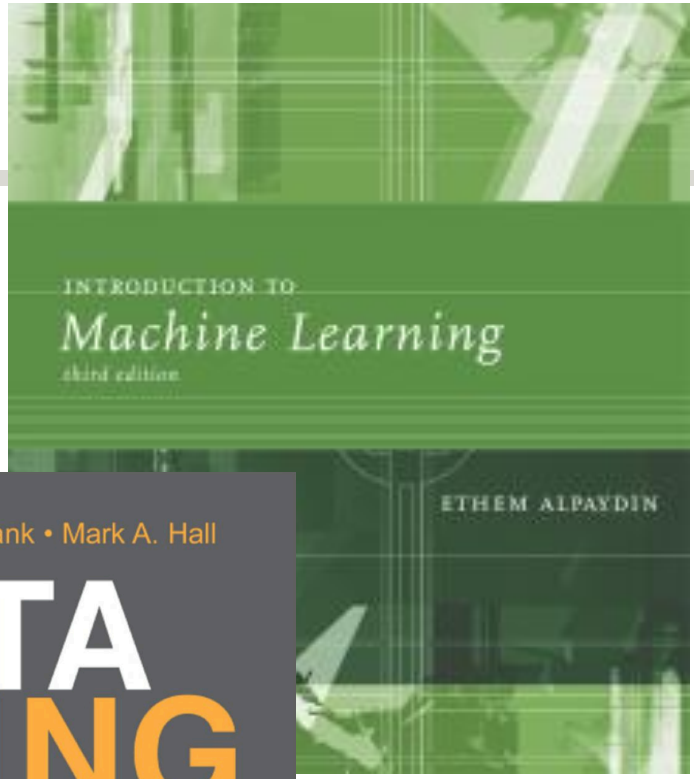
Pareto-Verteilung: Par(k, x_{min})

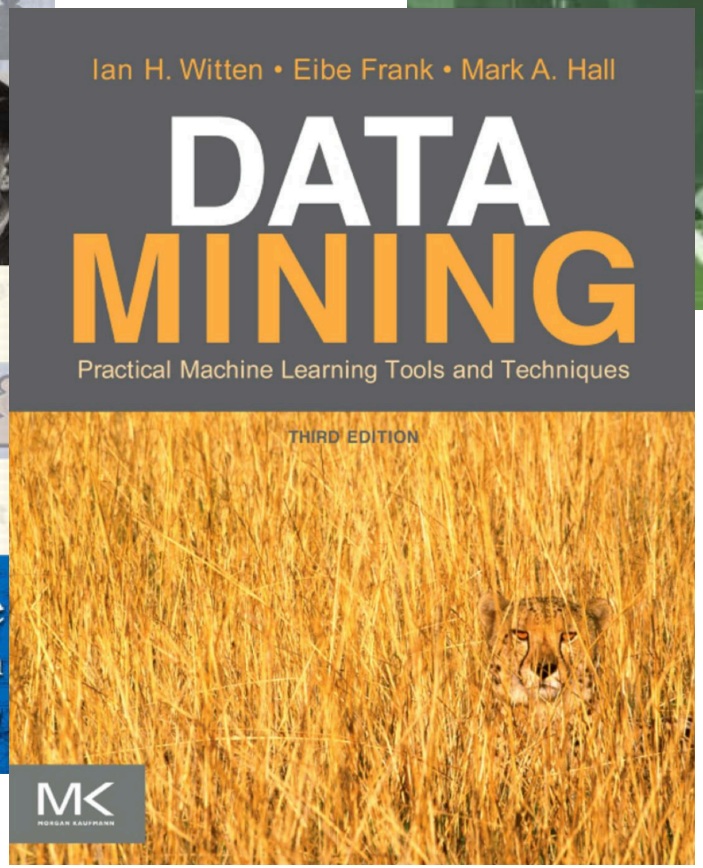
$$f(x) = \begin{cases} \frac{k x_{\min}^k}{x^{k+1}} & x \geq x_{\min} \\ 0 & x < x_{\min} \end{cases}$$

[Wikipedia]

Literatur

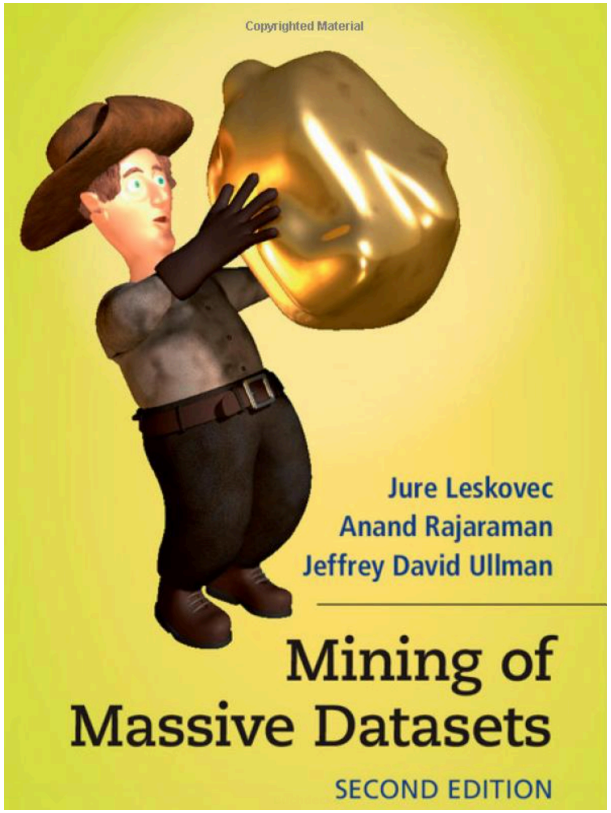

Stuart
Russell
Peter
Norvig
Artificial Intelligence
A Modern Approach
Third Edition


INTRODUCTION TO
Machine Learning
third edition
ETHEM ALPAYDIN


Ian H. Witten • Eibe Frank • Mark A. Hall
DATA
MINING
Practical Machine Learning Tools and Techniques
THIRD EDITION
MK

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