

Marcel Gehrke

Data Understanding vs. Machine Training

Introduction

Organisation

Intellectics: The Science of AI

- Type: lecture + seminar (in English)
- Credits: 8CP
- Regular attendance at the seminars is required
- Examination information: presentation (15 minutes) + homework
- Topics and dates for the presentations will be coordinated during the seminar

Organisation

Elective Module

- Type: lecture (in English)
- Credits: 3CP
- Regular attendance at the lecture is recommended



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

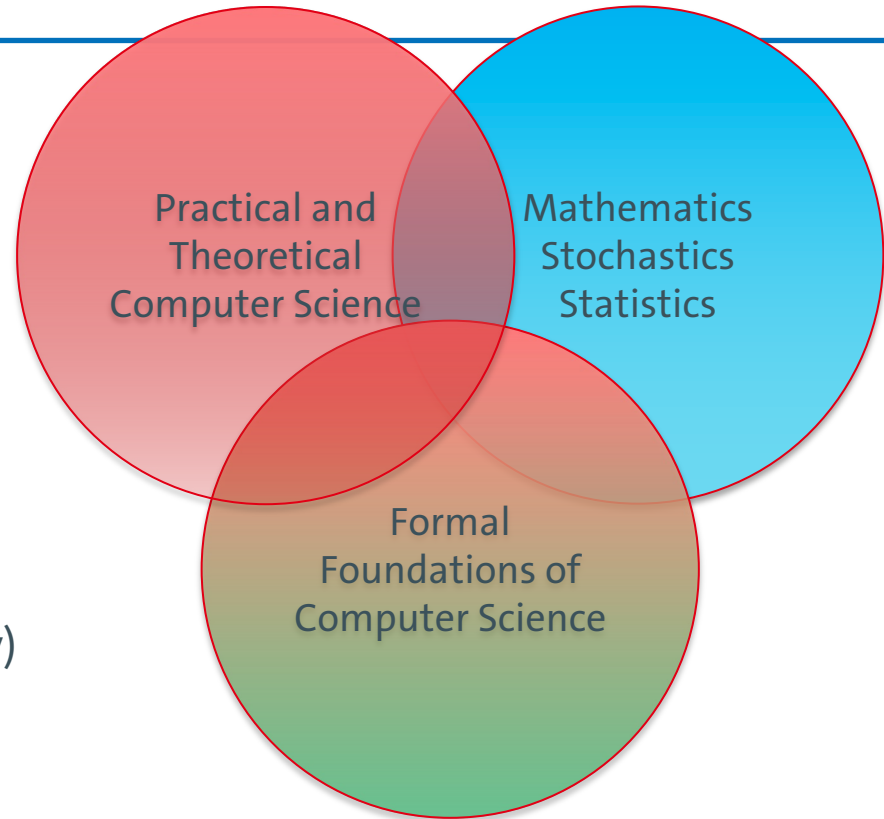


CHAI
Humanities-Centered AI

Introduction

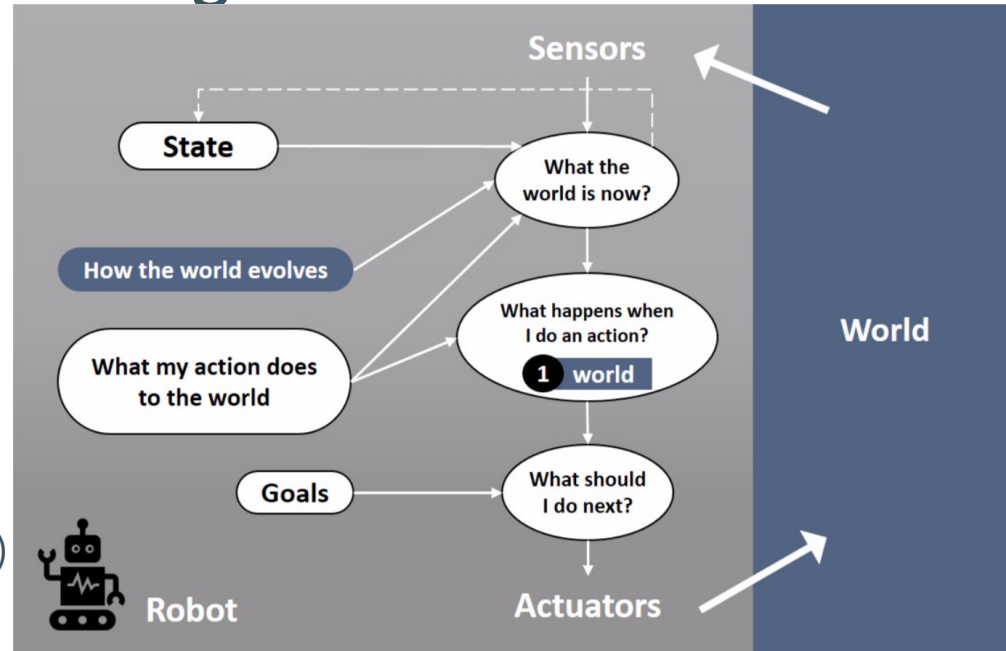
Data Science

- Extraction of knowledge from data (including graph data)
- Term proposed over 60 years ago for computer science
- Development of innovative concepts in the fields of logic, databases and stochastics / statistics (Data analysis and knowledge discovery)
- Based on mathematical foundations



... and what about Artificial Intelligence?

- Science of Intelligent Systems
- Agents
 - Have/form goals
 - Sensors/Actuators
 - Action planning
 - Online learning
- Mechanisms
 - Global cooperation of agents to achieve a common goal
- Agents interact with humans (and other agents)
 - Goals of the agents can be influenced



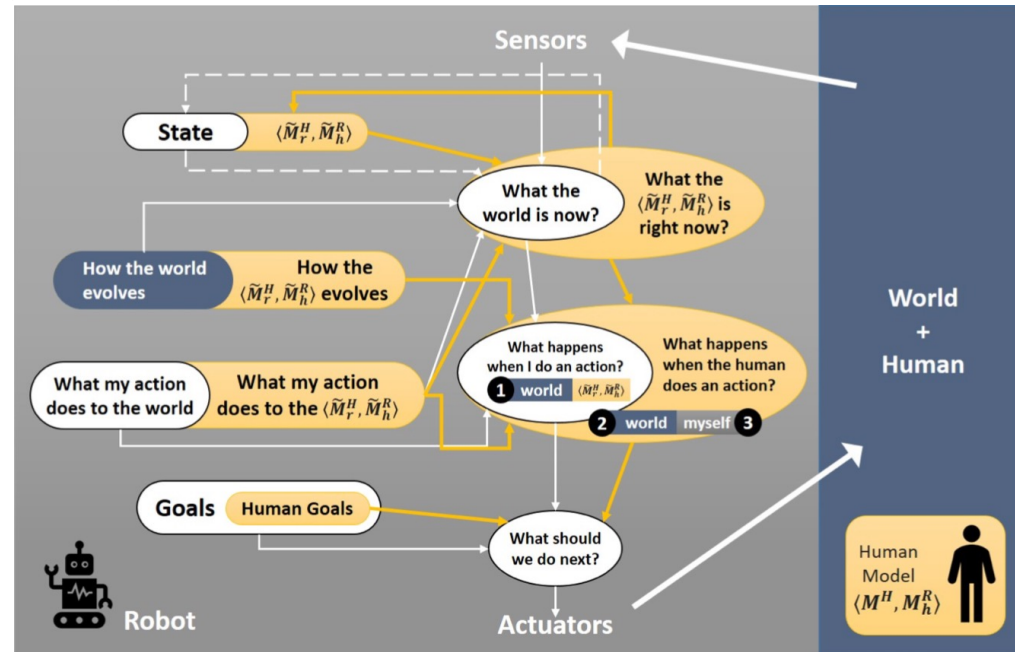
Science of Artificial Intelligence

Algorithmic modeling for agents only?

- Transparency, explainability

Rather new aspects:
Human-aware AI

- Conformity of expectations
- Proofable beneficial agents



Data Models vs. Algorithmic Models

Data Modeling

vs.

Algorithmic Modeling

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



We understand the world ?

Data Science
Statistics

We do not understand the world ?

How well 'my data model' works
~~Statisticians, Data Analysts, Data Miners~~
 Linear Regression
 Logistic Regression
 Known Distributions
 Confidence Intervals
 Predictor Variables & Goodness of Fit

Machine Learning

The world produces data in a black-box
~~Data Scientists~~
~~Machine Learning~~ X
 Random Forests, SVM
 Unknown Multivariate Distributions
 Iterative
 Predictive Accuracy

"Statistical Modelling: The Two Cultures" Leo Breiman, 2001

Data Science: Challenges

Large data sets

- Storage and access technology

Fast growth in data, high dynamics

- High data rates and real-time requirements

Heterogeneous data sets

- Distributed data management
- Data integration

Instance-based Query Answering

Assumption: Given many data points

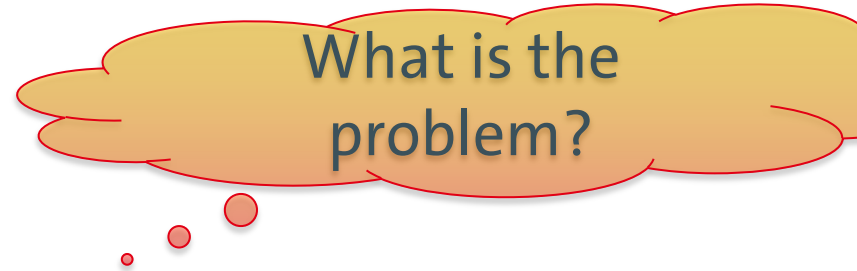
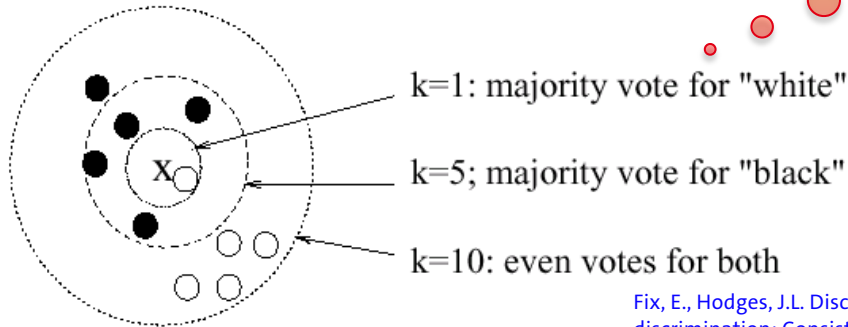
- Example features: (x, y, Colour) , $\text{Colour} \in \{\text{white}, \text{black}\}$

Query: Data point without a specific feature

- Example: Point which is neither black or white

Query answering (classification of the point queried):

Majority vote of the k -closest neighbours (kNN-approach)



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

Problems with kNN

- Classification result strongly dependent on k
- High memory requirements
- Efficient access to "neighbors" requires further measures (even more storage requirements)

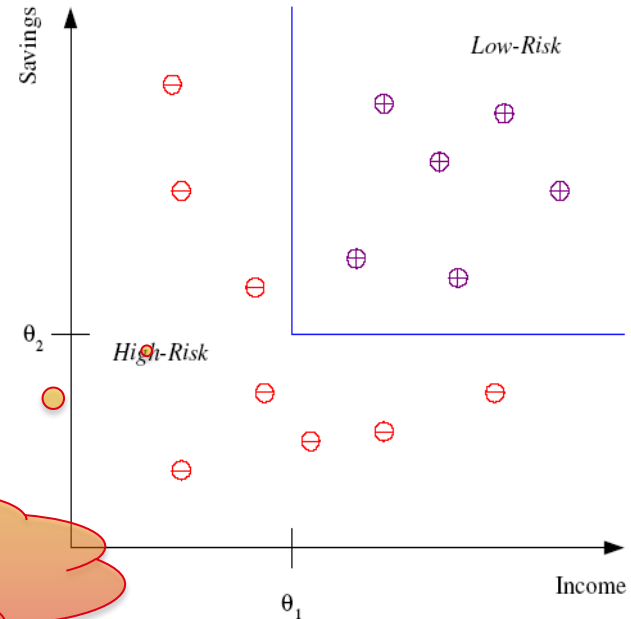
- Classification based on the data

Model-based Query Answering

Representation of the data by parameters of a model

- If $(\text{income} > \theta_1 \wedge \text{savings} > \theta_2)$, then creditworthy (\oplus), otherwise not (\ominus)

Only 2 parameters needed: (θ_1, θ_2)
model requires low memory to store



What is the problem?

Task

Estimate the number of cities with a certain number of inhabitants

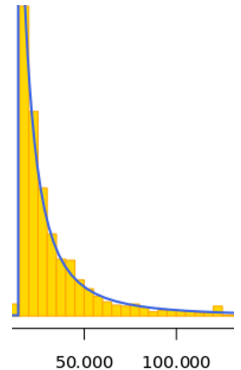
Data: List with number of inhabitants
(rounded to 5000)

Explicit model: Count all occurrences

- **Incompleteness** of data

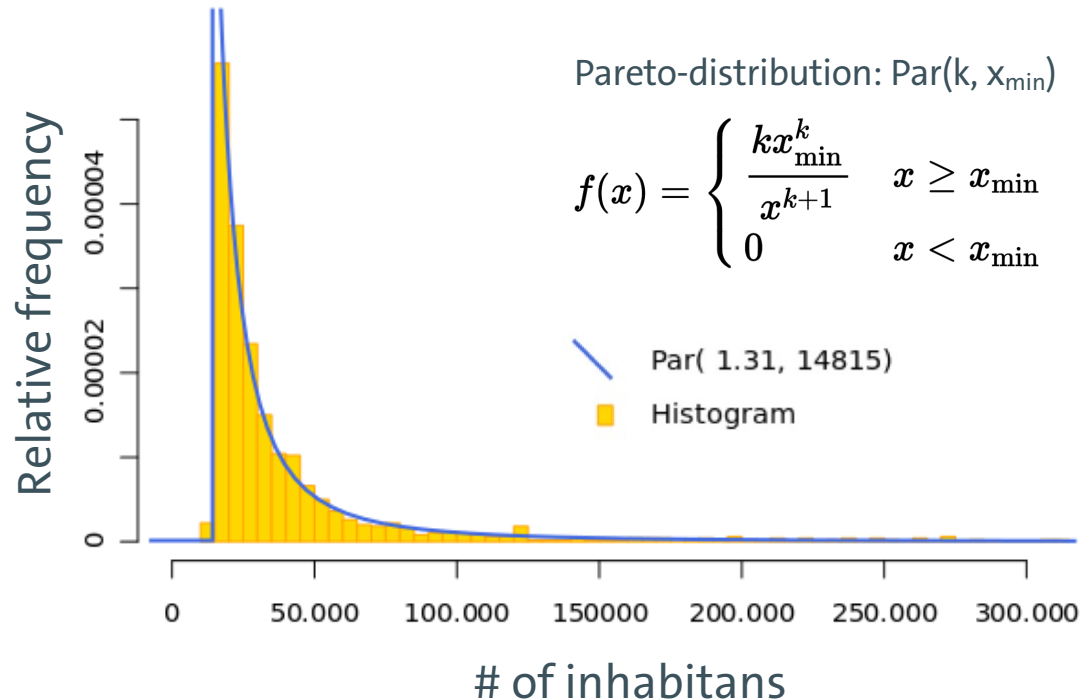
Implicit model: Use the potency law $y=ax^b$

- Determine a and b (a positive, b negative)
- Solve **complex** optimization problem



Concept of "Distribution"

Inhabitans of German cities



Literature

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

Many additional books, presentations, and videos on the web

