Image by: Bart van Dijk, unsplash.com

Data Mining

Prof. Dr. Andreas Hotho

Florian Buckermann

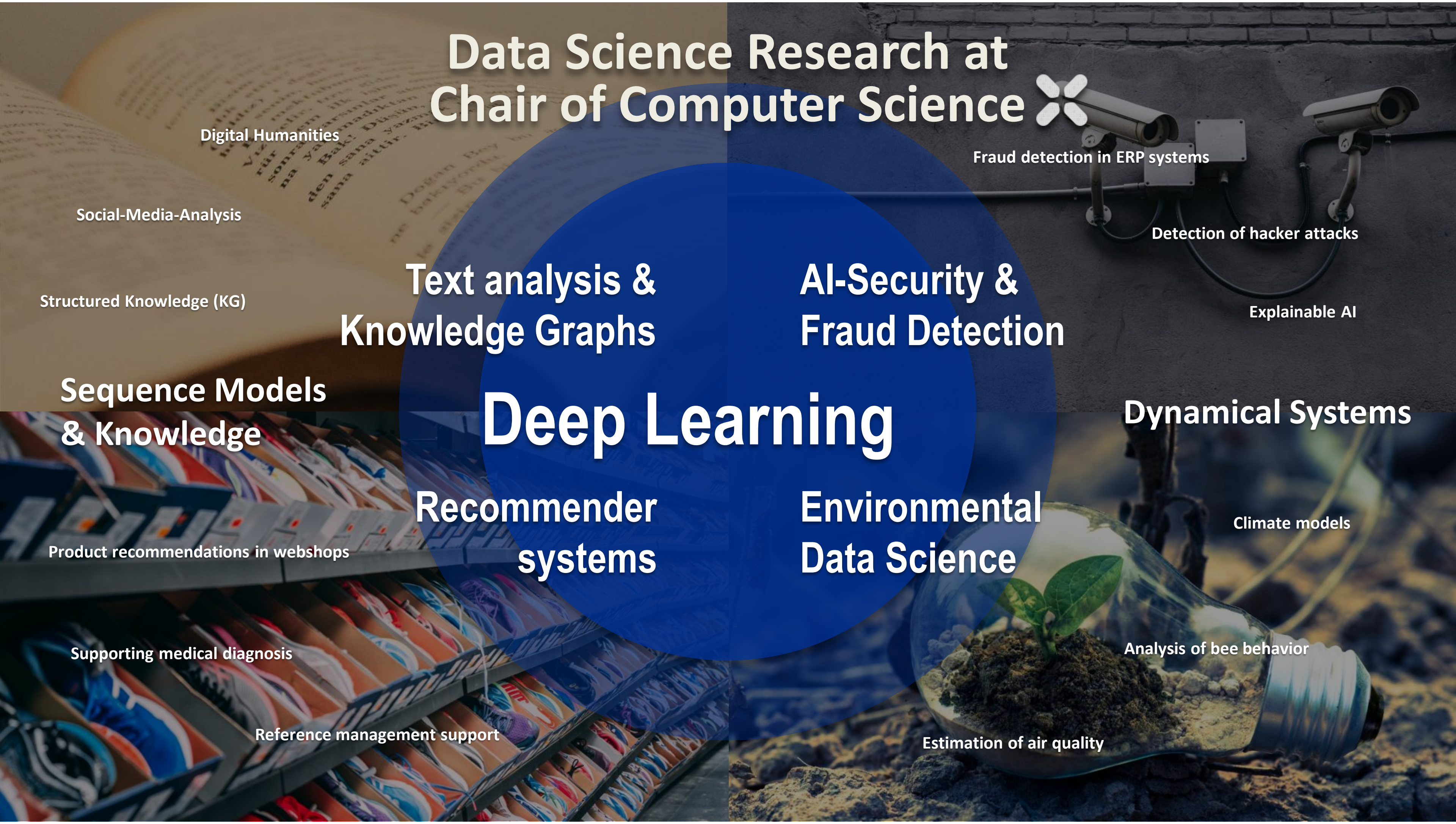
Image by Lukas Blazek, unsplash.com

Data Science

Prof. Dr. Andreas Hotho

Florian Buckermann

Data Science Research at Chair of Computer Science X



Digital Humanities

Social-Media-Analysis

Structured Knowledge (KG)

Sequence Models & Knowledge

Text analysis & Knowledge Graphs

Fraud detection in ERP systems

Detection of hacker attacks

Explainable AI

AI-Security & Fraud Detection

Deep Learning

Dynamical Systems

Recommender systems

Product recommendations in webshops

Supporting medical diagnosis

Reference management support

Environmental Data Science

Climate models

Analysis of bee behavior

Estimation of air quality

https://www.informatik.uni-wuerzburg.de/datascience/

Organizational matters: Dates

- **Lecture**
  - Start: May 2nd, 2022
  - Monday, 10:15 - 11:45 am, at the Turing-HS
- **Exercises**
  - Thursday, 08:15 - 09:45 in room SE I
  - Thursday, 14:15 - 15:45 in room ÜR II
  - Thursday, 16:15 - 17:45, in room SE III
  - is held as a supervised exercise (see slide 6)

- **Proof of performance**
  - **Examination (10-I-DM,** e.g. Bachelor of Computer Science**):** Expected 08.08.2022
  - **Oral examinations** (**10-I=DM**, e.g. Master of Computer Science or Master DH**):** tbd
  - Registration via WueStudy: 16.04. - 15.07.2022

Organizational matters

- **Contact for questions about the lecture:**
  - datamining@informatik.uni-wuerzburg.de
- **Consultation hours by arrangement:**
  - Prof. Dr. Andreas Hotho: hotho@informatik.uni-wuerzburg.de, room B112
  - Florian Buckermann: buckermann@informatik.uni-wuerzburg.de, Room B104
- **Information at WueCampus:**

  (https://wuecampus2.uni-wuerzburg.de/moodle/enrol/index.php?id=52903)

Following information can be found in WueCampus

- Current announcements (!)
- Slides
- Exercise sheets
- Dates (+ ZOOM links if applicable)

Organizational matters : Excercises

- Poll for excercises starts after the first lecture
- Supervised exercise
  - **Independent work on** the exercise sheet in small groups of 3-4 persons under the supervision of the assistant
  - **No repetition of** the lecture material in principle
  - **No presentation of** the sample solution
    (Sample solution will be uploaded in WueCampus).
- Necessary for this
  - Independent preperation **before the exercise**
  - Have the slides at hand (!)
  - Actively contribute to the exercise

Organizational matters

- **Exercice concept motivation**
  - Studying the algorithms actively is more productive
  - Your are encouraged to **recognize connections** in the content
  - You learn to think in a structured way and to work independently
  - You learn **teamwork**
  - You learn to explain your approach
  - **You actively train for the exam** ;-)
  - *„You have earned your degree in ... . Your personal strengths include initiative, willingness to communicate and cooperate, teamwork.“*

(Typical ad text)

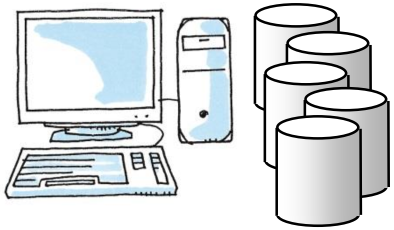
Motivation

Huge amounts of data are collected automatically.

Who else does the user know in the social network?

Which treatment is most appropriate for given symptoms? …

Which (new) series does the viewer want to see?



What associations exist between the goods bought in a supermarket?

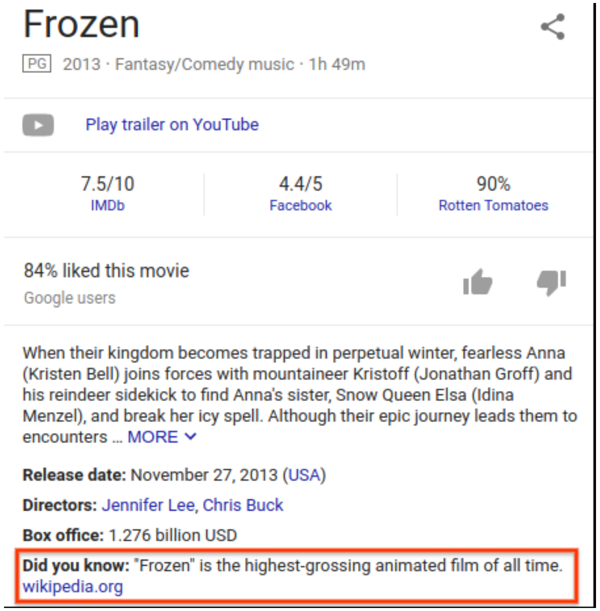
To which class does this star belong to?

➡ Non-trivial relationships
Analyses can no longer be performed manually.

Applications (examples)

Knowledge Discovery
Example: Extracting facts from Wikipedia

“There has been recent interest in providing fun facts. Obtaining such trivia at large scale is, however, non-trivial. […] we show how fun facts can be mined from superlative tables in Wikipedia.”*



Frozen

PG 2013 · Fantasy/Comedy music · 1h 49m

Play trailer on YouTube

7.5/10	4.4/5	90%
IMDb	Facebook	Rotten Tomatoes

84% liked this movie
Google users

When their kingdom becomes trapped in perpetual winter, fearless Anna (Kristen Bell) joins forces with mountaineer Kristoff (Jonathan Groff) and his reindeer sidekick to find Anna's sister, Snow Queen Elsa (Idina Menzel), and break her icy spell. Although their epic journey leads them to encounters ... MORE

Release date: November 27, 2013 (USA)

Directors: Jennifer Lee, Chris Buck

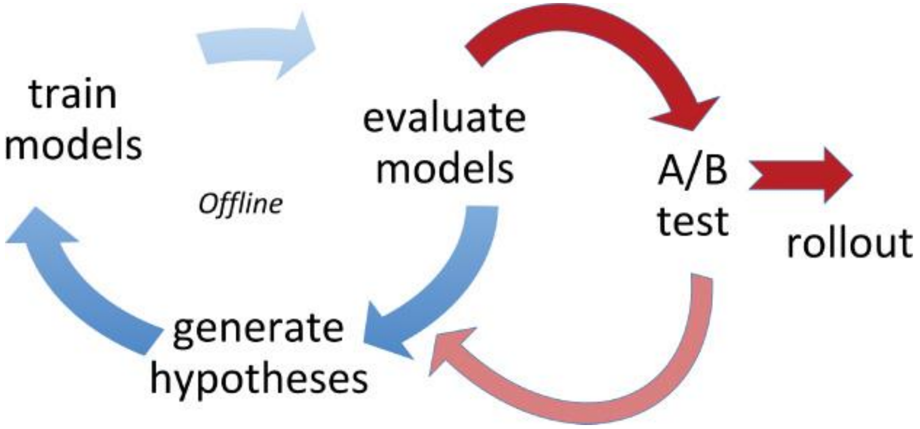
Box office: 1.276 billion USD

Did you know: "Frozen" is the highest-grossing animated film of all time. wikipedia.org

[*] Korn et al., Automatically Generating Interesting Facts from Wikipedia Tables, Proceedings of the International Conference on Management of Data, SIGMOD, 2019

Recommender Systems
Example 1: Video Streaming

- “Internet TV is about choice: what to watch, when to watch, and where to watch […] But humans are surprisingly bad at choosing between many options […]”*



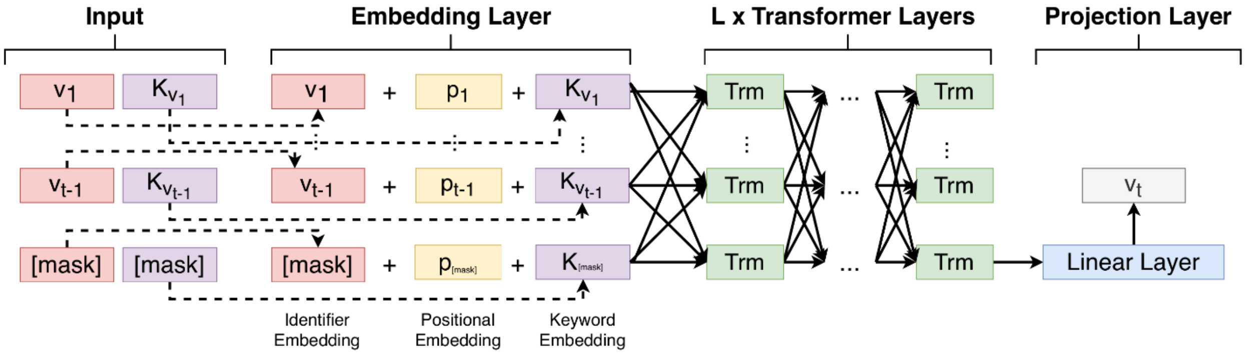
- Personalized Video Ranks (based on watch history)
- Trending Now (based on events, news, etc.)
- “Because You watched” (recommendation conditioned on a single title)

Procedure of algorithm development at Netflix, Inc.*

* The Netflix Recommender System: Algorithms, Business Value, and Innovation, ACM Transactions on Management Information Systems, Vol. 6, No. 4, 2016

Recommender Systems
Example 2: Personalization in Online Shops

Collaboration between the Data Science Chair and Adidas



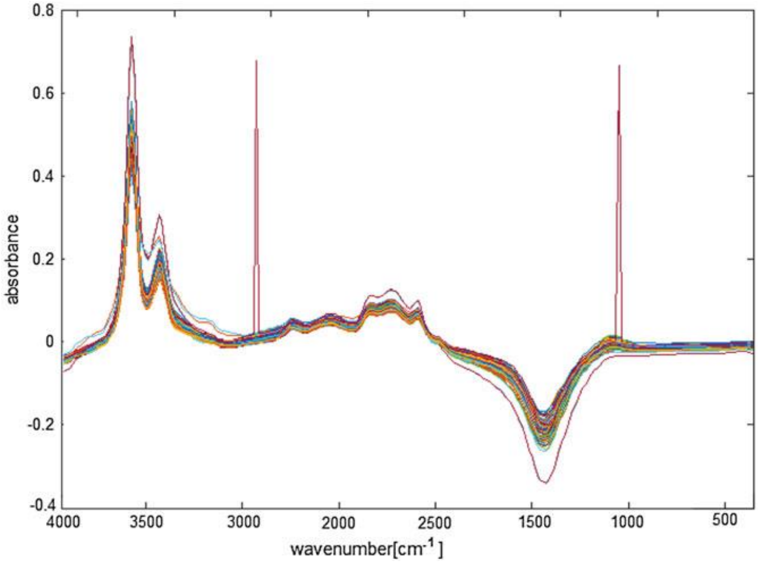
"[…] To include additional information into the new state-of-the-art-model BERT4Rec, we introduce KeBERT4Rec, a modification that allows to add keywords describing items (e.g. genres of a movie)."*

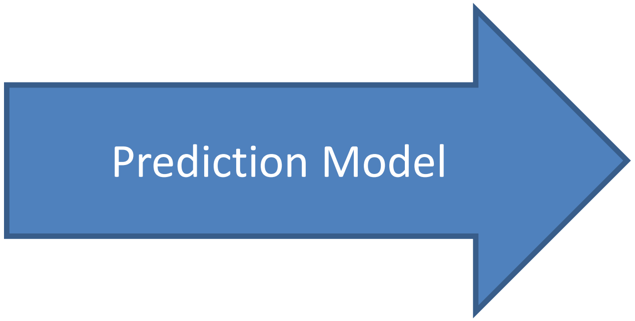
* Integrating Keywords into BERT4Rec for Sequential Recommendation, KI 2020

Analytical Chemistry
Example 1: Determining the alcohol content in wine

“[…] Estimating the concentration of chemical components of interest in a given product is difficult and challenging due to the collinearity between the spectral variables and the large number of variables to be treated.”*

Exemplary spectra of wine samples




Prediction Model

Alcohol Content

* One-dimensional convolutional neural networks for spectroscopic signal regression, Journal of Chemometrics, Volume 32, Number 5, 2018

Analytical Chemistry

Example 2: Determining process variables for production

Collaboration between the Data Science Chair and Knauf

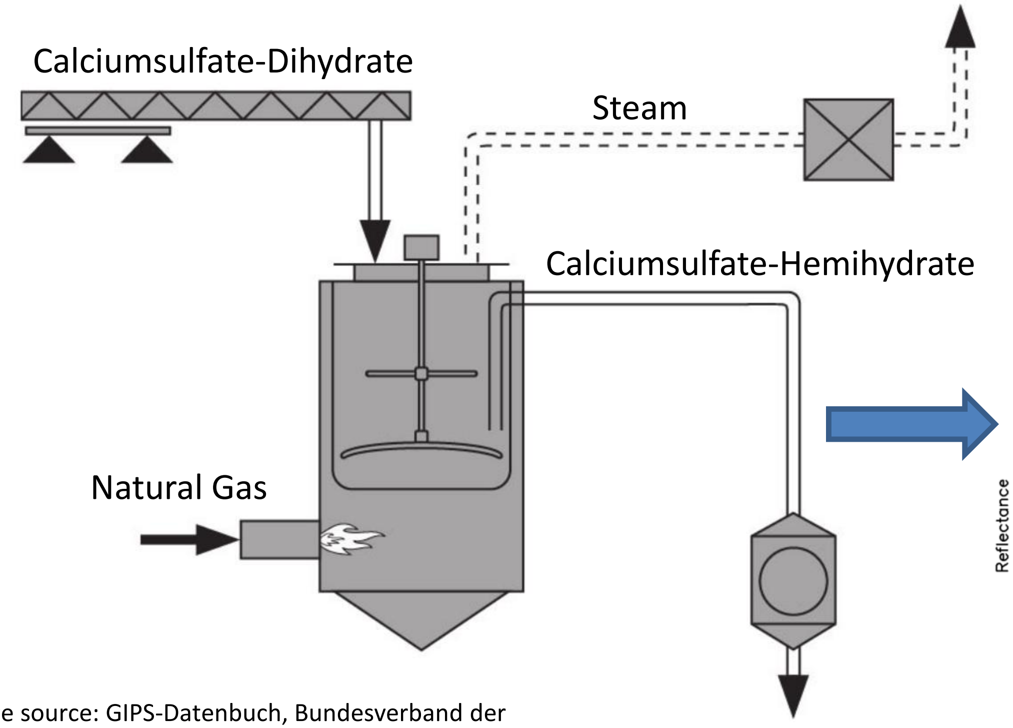


Image source: GIPS-Datenbuch, Bundesverband der Gipsindustrie e.V., 2013

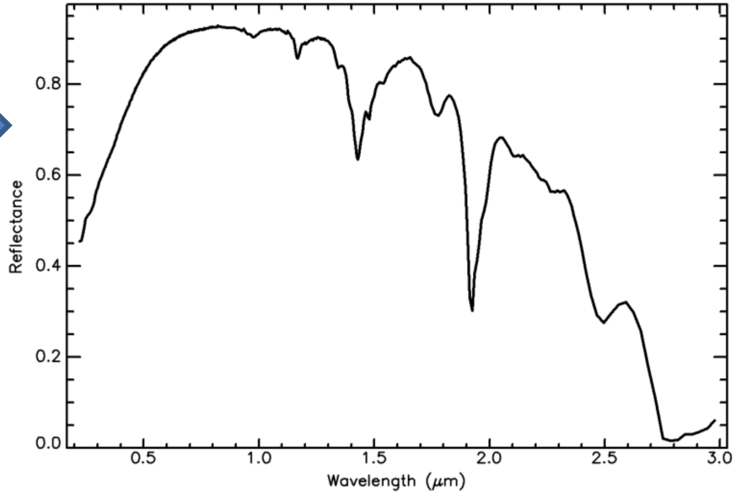


Image source: USGS Spectral Library

Content of the lecture

- Overview of KDD, Data Mining, Data Science
- Overview of the most **important tasks and algorithms** and their advantages and disadvantages
- Selection and use of an algorithm for a given application
- Development of own methods for a given application
- Questions from related topics like databases, web applications, etc.

Structure of the lecture

1. Basics
2. Clustering
3. Association Rules
4. Classification
5. Regression Analysis
6. Neural Networks (Introduction)

Literature

- Primary source for the structure of this lecture:

  Ester M., Sander J., "Knowledge Discovery in Databases: Techniken und Anwendungen", Springer, 2000.

- Data Science and Data Mining

  Grus J., Einführung in Data Science, O‘Reilly, 2019

  Grus J., Data Science from Scratch, O‘Reilly, 2019

  Aggarwal C., Data Mining – The Textbook, Springer, 2015

  Shmueli G., Data Mining for Business Analytics, Wiley, 2020

- Machine Learning and Statistics

  Bishop C., Pattern Recognition and Machine Learning, Springer, 2006

  Müller, A, Introduction to Machine Learning with Python, O‘Reilly, 2016

  Bruce P., Practical Statistics for Data Scientists, O‘Reilly, 2020

  James G., An Introduction to Statistical Learning, Springer, 2013

  Goodfellow I., Deep Learning, MIT Press, 2016

Literature

- Most books are available from the university network, e.g.

  Einführung in Data Science

  Data Science from Scratch

  Data Mining – The Textbook

- Some books are available for free, e.g.

  Pattern Recognition and Machine Learning

  Deep Learning

- A list of interesting books (including links) can be found under

  https://www.bibsonomy.org/tag/lecture:data-science

Differentiation of the lecture content from others, Special data types and applications

(not covered in detail in the lecture)

- Temporal Data Mining
  - Analysis of time-series data (time dependend characteristics)
  - Stock prices, inflation rates, blood pressure, precipitation, temperatures...
  - Examples: Trend analysis, event detection, sequential patterns,...
  - Tutorial for "Sequential User Behavior on the Web", http://sequenceanalysis.github.io/

- Spatial Data Mining
  - Analysis of spatial data
  - A given attribute has spatial dependencies (position and extent in a 2- or 3-dimensional space)
  - Geo Information Systems (GIS, Maps)

Further data types in data mining

(not covered in detail in the lecture)

- Images
  - Computer Vision
  - Information is encoded in in pixels which are spatially connected
  - Automatic and hierarchical extraction of features (e.g. with Convolutional Neural Networks)
- Web and Text Mining
  - Analysis of text and hypertext data
  - Adaptation of standard methods according to the properties of languages
  - Use of the link structures of the web
  - Extraction of special entities
- Graphs (social networks)
  - Social Network Analysis (SNA)
  - Metrics on graphs for a better insight
  - Special methods e.g. for clustering

Further methods in data mining

(not covered in detail in the lecture)

- Inductive logic programming
  - Given: Set of facts in a first-order predicate logical language
  - Wanted: first-order predicate logical rules, which hold true
  - Method: Search in the space of all possible rules

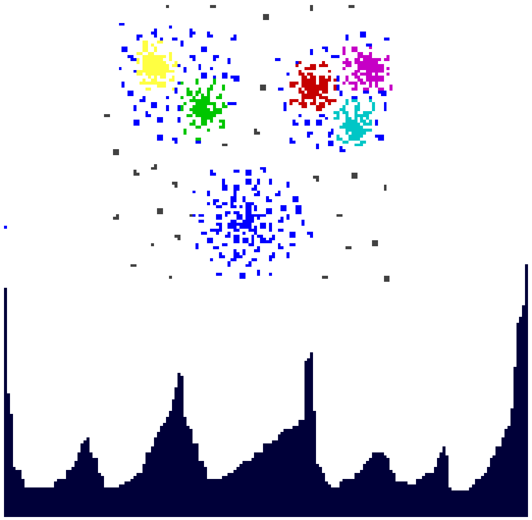
- Genetic Algorithms (Optimization)
  - General search method
  - Based on the principle of biological evolution
    individual, chromosome, gene, combination, mutation

Further methods in data mining

(not covered in detail in the lecture)

- Subgroup Discovery
  - Find subsets of the data set where a target attribute is distributed significantly differently than in the overall data set / the deviation of the distribution of a target concept is interesting

- Visualization of large amounts of data



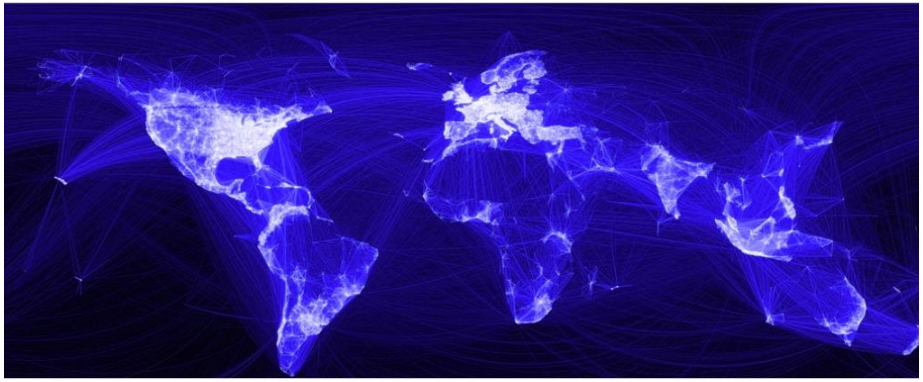


Figure: http://cdn.oreillystatic.com/en/assets/1/event/75/Visualizing%20Geo%20Data%20Presentation.pdf

Further courses

Other courses from the Data Science chair	
Lectures	Data Mining (BA) Information Retrieval (MA) Machine Learning for NLP (MA) NLP and Text Mining (MA) Interactive Artificial Intelligence (BA, given as KI I)
Practical Course	Machine Learning for Time Series Analysis (MA) Natural Language Processing (MA)
Seminar	Selected Topics of Machine Learning (BA/MA)

Further courses

Other courses related to Data Mining, Data Science, Machine Learning and AI	
Lectures	Time Series Analysis and Forecasting (Dr. Bauer) Programmierung mit Neuronalen Netzen (Prof Dr. Puppe) Künstliche Intelligenz I + II (Prof. Dr. Puppe) Wissensbasierte Systeme (Prof. Dr. Puppe) Machine Learning for Complex Networks (Prof. Dr. Scholtes) Statistical Network Analysis (Prof. Dr. Scholtes)
Practical Course	Aktuelle Trends in der Künstlichen Intelligenz (Prof. Dr. Puppe) Machine Learning for Complex Networks (Prof. Dr. Scholtes)
Seminar	Modellierung Intelligenter Systeme (Prof. Dr. Frank Puppe) Graph Neural Networks (Prof. Dr. Scholtes) Statistical Network Analysis (Prof. Dr. Scholtes) Data, AI, and Society (Dr. Wegner)

More courses will likely follow soon from our
Center for Artificial Intelligence and Data Science (CAIDAS)
At the JMU Würzburg