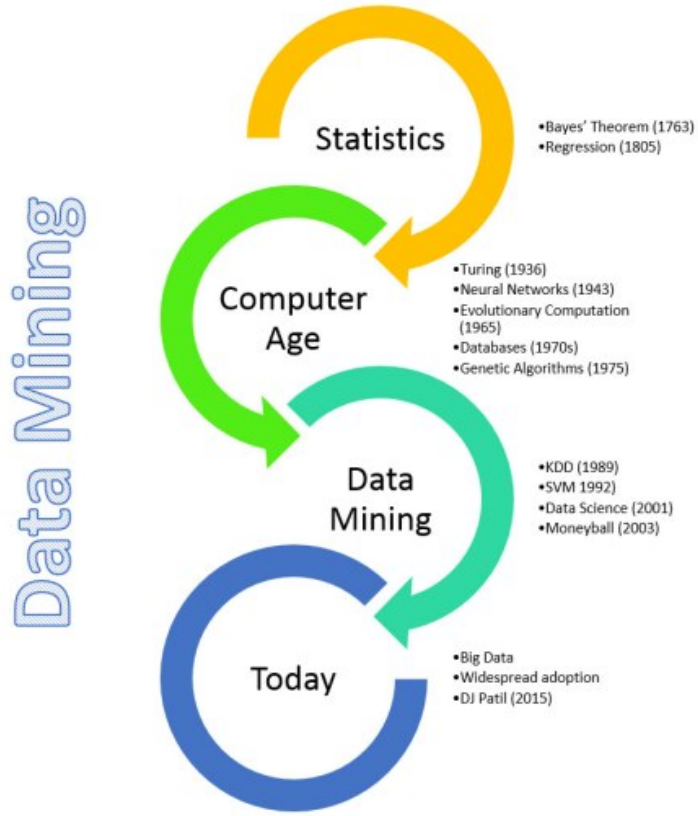
Introduction to Data Mining

Lecture, winter term 2018/2019, 5 CP

Thomas Hermann

1. What is Data Mining

Definition: "Data mining is the computational process of exploring and uncovering patterns in large data sets a.k.a. Big Data. It's a subfield of computer science which blends many techniques from statistics, data science, database theory and machine learning." (from: http://www.kdnuggets.com/2016/06/rayli-history-data-mining.html (http://www.kdnuggets.com/2016/06/rayli-history-data-mining.html))



(Illustration from kdnuggets.com (http://kdnuggets.com))

1.1. Goals

The goal of Datamining is the discovery of hidden structures and regularities in usually large data collections, and to make these insights applicable.

Alternative terms are 'Siftware' und 'Knowledge discovery in Data Bases' (KDD), (coined by Gregory Piatetsky-Shapiro)

It's a misnomer:

- It is not resulting in data (data are not mined) but of knowledge extraction.
- It plays with the metaphor of mining (iron mining, gold mining) where few nuggets are hidden in a huge amount of ore / dirt.

In the focus are coherences of empirical nature, which can only badly or not at all be derived from theoretical considerations, such as for instance the decision behaviour of customers.

Some Aspects

- uncover hidden regularities in data bases, make them perceivable and usable.
- Techniques to support the interpretation of large data collections
- Transform **data** into **information**
  - Derivation of rules
  - Diagnosis of dependencies
  - Recognition of trends
  - Prognosis

Related Procedures

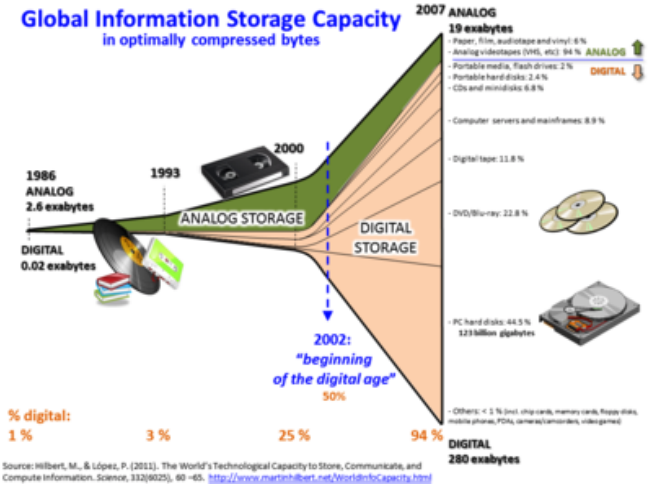
- **OLAP** (Online Analytical Processing): }
  - Organization of data collections into a multidimensional data cube
  - providing fast access modes for slices for interactive exploration of mostly simple coherences between dimensions
  - Typically as a client-server architecture.
- **Statistical Analysis**
  - are primarily *deductive*: a (whereever coming from) hypothesis is tested at hand of available data (e.g. computing the significance)
  - The goal of Data Mining is in contrast to that primarily **inductive**: at hand of data we aim at new good hypothesis or models on coherences between attributes.
- Statistical Methods serve as important tool to evaluate discovered patterns/rules/structures concerning their significance

Data Mining = EDA + CDA

1.2. Evolution of the field

The field of data mining established since the begin of the 90ies

- 1989: first KDD workshop at AAAi 89 (Piatetsky-Shapiro)
- 1991, 1993: further workshops
- 1995: first international conference at IJCAI 95
- 1996: NRW Forschungsverbund "Virtuelle Wissensfabrik"
- 1997: first journals: Datamining & Knowledge Discovery, and Intelligent Data Analysis
- 1997++: Machine Learning
- 2001: William S. Cleveland introduces the term *Data Science* in his paper "Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics"
- 2008++: Big Data
  - "Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time" (Snijders, C.; Matzat, U.; Reips, U.-D. (2012). "'Big Data': Big gaps of knowledge in the field of Internet". International Journal of Internet Science. 7: 1–5.)
  - famous 3 Vs: volume (amount of data), velocity (in/out speed of data), and variety (various data types and sources).
  - Due to massive growth of digital data over the computer age:



- about 2010++: Deep Learning

1.4. Neighboring Fields

There are tight connections of data mining / KDD with many neighboring disciplines:

- Pattern Recognition
- Data Analysis, Statistics
- Neural Networks / Deep Learning
- Maschine Learning
- Visualization, Computer Graphics
- Databases, knowledge-based systems
- Cognitive Science

1.5 Basic Cases for Datamining

- Whereever data are available in abundance and where the underlying laws are not sufficiently known (or applicable).

The majority of applications lies in business, particularly the analysis of customer behavior:

- **Segmentation of Customer types ("Kundensegmentierung")**: what types of customers can be discerned and what are their essential differences?
- **Churn-Analysis**: what factors let customers go to other providers? How can the customers' loyality be increased? *(churn = Butter-Rührfass, aufrühren)*
- **Fraud-Analysis**: what patterns (e.g. while using internet services or mobile phone services) allow provides to assume/conclude incorrect or even fraudulent behaviours of the users?
- **Target group analysis**: what target groups for an action (advertisement, bargain offer, survey, call for donations) is for the envisioned purpose the most promising?
- **Buying behaviour**: What factors influence the buying behaviour of my customers in a wished manner? What prognosis can I make?
- **Product marketing**: what factors make my product more attractive than those of the competitors?
- **Portfolio management**: intelligent selection of product or stocks in order to optimize a certain target function (e.g. interest, popularity etc.)

1.6. A typical Data Mining Example

Marketing optimization in an american catalogue company, historical example, approx 1997

- Order of Magnitude
  - 400 Mio USD annual turnover
  - several catalogue types
  - ca. 100 Mio catalogue sendings per year
  - more than 200 branches
- Business goals: more accurate and thus more economic Marketing (Pinpoint Marketing) by better utilization of customer information
  - What customers contribute mostly to the sales?
  - what helps to bind these customer groups better?
  - How do the customers' interests distribute on the different catalogues and special offers?
  - Comparison of marketing use between large catalogues and smaller specialized catalogues
  - Dependency of the buying pattern from the season
  - more accurate, target group specific criteria for catalogue mailing
  - overarching goal: return-of-investment (ROI) optimization

Data Basis:

- Customer address data base (ca. 10 GBytes)
- demographic customer data (ca. 10 GBytes)
- documented transactions (ca. 500 GBytes)
- results from sales campaigns (ca. 2 GBytes)

Involved Groups:

- IT
- Database Department
- Marketing
- Sales
- Upper Management

Project process

- Initialization by the Marketing Dept.
- Definition of Success criteria: resulting savings, increase of customer loyalty, etc.
- Setting up the technical requirements (computer configuration, database setup, required tools)
- tools: SAS Datamining tools, DB2, Oracle RDBMS under SQL
- Data Selection: Sales Transaction of the past 36 months, sorted by volume and product type
- Creation of a project dedicated data base
- Data preparation:
  - Filtering inconsistent data
  - Treating missing values
- Setup of forecast models
- Mainly applied Methods:
  - logistic regression = regression for discrete output variables, modelling the logit $\log(p(1)/(1-p(1))$ as linear combination of the input variables.
  - Neural networks
  - Decision trees
- Optimization of model parameters and comparison of models
- subsampling in order to accelerate the throughput:
  - 1 million training points, 0.5 million examples for validation and test.

Results:

- Identification of a customer segment of 30% which is responsible for the majority of the sales
- better tuning of offers for this customer segment
- better timing of sales campaigns for a more steady turnover
- reduction of costs for catalogue sendings and advertisement
- **Data preparation (Selection, Cleaning, Recoding, Fusing, etc) made 70-80% of the overall effort**

1.7 More Examples:

- Pharmacy / Chemistry: Searching for Similarities between chemical compounds
- Paper production:
  - Process modelling: finding relevant sensor values for best quality
  - describing process behaviour in form of rules
- **Environmental Research**: Change of vegetation, atmosphere (smog, ozone, climate), pollutant concentraion, seizmographic data
- Telematics: **Predicting inner city traffic** e.g. Cologne: 120 induction loops and data about events (concerts, sports events, etc.)

1.8. Important Questions

- How can we discern regularities from random fluctuations?
  - → **Statistics**
- What features are important? How to cope with the problems of high-dimensional data?
  - → Methods for **Dimensionality Reduction**
- How can we merge data in a meaningful way according to their similarity?
  - → **Clustering techniques**
- How can we perceptualize data
  - → **Visualization and Sonification techniques**
- How can we classify data?
  - → **Classification methods**
- How to model dependencies
  - → **Model extraction and machine learning**
- How can we forecast into the future?
  - → **time series analysis**
- How can we identify complex structures
  - → techniques for the induction of **decision trees**

1.9. Focus of this Lecture

- Understanding the underlying problems in high-dimensional data analysis
- Learning how central methods are rooted in mathematics
- Learning by doing: implement and investigate algorithms to become familiar
- Mastering the basics rather then knowing all details
  - NOT much interest in 'howto use existing data mining packages'
- Providing a solid basis for more advanced data science courses

[winter term 2018/2019: end of T1 (2018-10-10)]