Knowledge Discovery in Databases II

Winter Term 2014/2015

Chapter 1: Introduction and Outlook

Lectures : PD Dr Matthias Schubert
Tutorials: Markus Mauder, Sebastian Hollizeck

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_II_(KDD_II)

Organisatorisches

- **Times and Locations**
  - Lectures: Tuesday, 14:00-17:00, room A120 (Main building)
  - Tutorial: Thursday, 16:00-18:00, room U127 (Oettingenstr. 67)
    ??? Thursday, 18:00-20:00, room U127 (Oettingenstr. 67)???
  - All information and recent news can be found at:
    http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_II_(KDD_II)
- **Written Exam**
  - Registration for the written exam:
    https://uniworx.ifi.lmu.de/?action=uniworxCourseWelcome&id=344
  - The exam takes 90 min and accounts forr 6 ECTS points

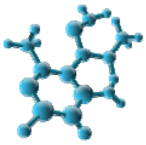
Chapter overview

- Knowledge Discovery in Databases, Big Data and Data Science
- Basic Data Mining Tasks (Recap KDD I)
- Topics of KDD II
- Literature and supplementary materials

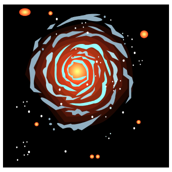
Knowledge Discovery

Large Amounts of data in multiple applications

connection data

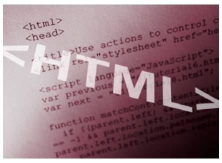


molecule process data



telescope data

payment data



Web data/ click streams

Manual analysis is infeasible

⇨ **Knowledge Discovery in Datenbanken und Data Mining**

Aims: - descriptive patterns: Explains the characteristics and behavior of observed data (explicit description)

- predictive methods and functions: predict the behavior of new data (unknown patterns and behaviors)

Important: Found patterns don‘t have to apply in 100 % of the cases.

Definition KDD

[Fayyad, Piatetsky-Shapiro & Smyth 1996]

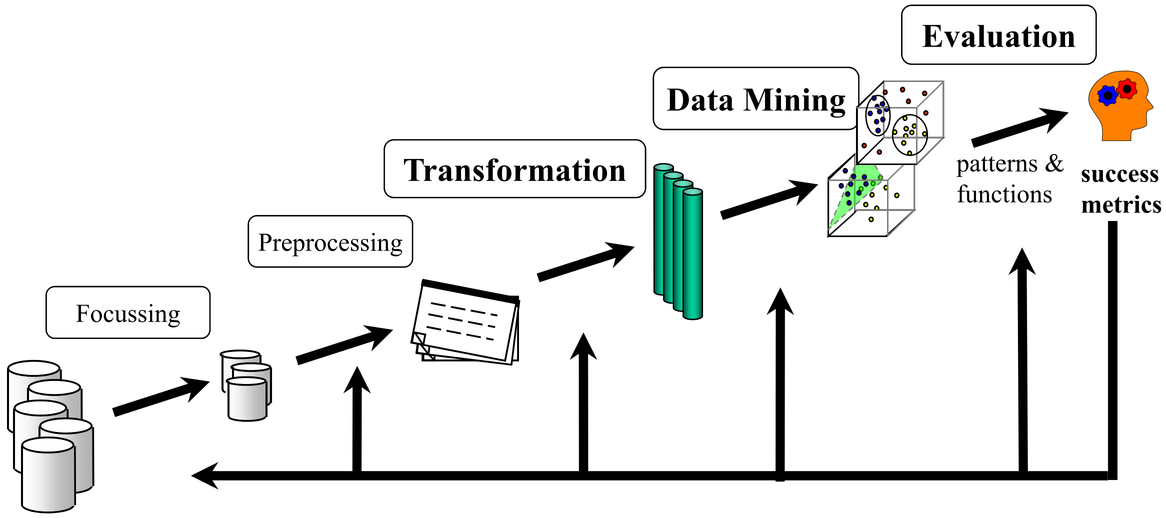
Knowledge Discovery in Databases (*KDD*) is the process of (semi-) automatically extracting pattern from databases which are

- *valid*
- *formerly unknown*
- *potentially useful*

remarks:

- (*semi-*) *automatically: in contrast to manual analysis*
- *valid*: from a statistic point of view
- *formerly unknown*: not explicitly known
- *potentially useful*: for a given application

Knowledge Discovery Prozesse



- KDD is a successive processing chain
- KDD process is a iterative process which can jump back to any previous step for recalibration until the wanted result is achieved

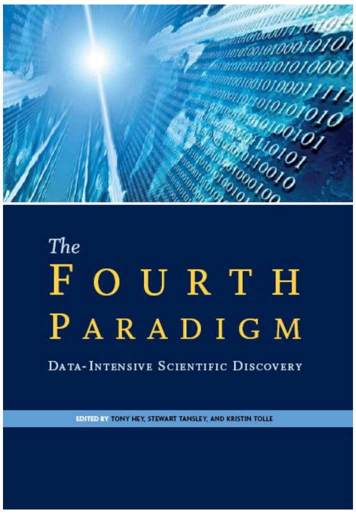
  => It is very important to a clearly defined goal

The Fourth Paradigm

1. 1000 years ago: experimental science
   - Describes natural phenomena
2. Last hundred years: theoretical science
   - Newton‘s laws, Maxwell-equations, …
3. Last decades: computer aided science
   - Simulation of complex phenomena
4. Present day: data intensive sciences
   - unites theories, experiments and simulations

“Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets.”

-The Fourth Paradigm - Microsoft


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Big Data

New challenges for scientist in all areas: Handling huge amounts of data

Excerpt from the McKinsey Report *“Big data: The next frontier for innovation, competition, and productivity“*, June 2011:

Capturing the value of Big data:

- 300 billion USD potential value for the north American health system per year
- 250 billion Euro potential value for the public sector in Europe per year
- 600 billion USD potential value through the use for location based services

“The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings.”

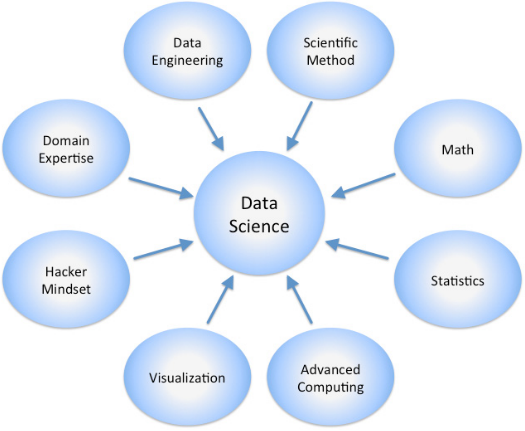
-www.mckinsey.com/mgi/publications/big_data/

Data Science

- Science of managing and analyzing data to generate knowledge
- Very similar to KDD

Differences to KDD:

1. Data Science is broader in its topics. (result representation, actions..)
2. Integrates all scienctifc directions being concerned with data analyses and knowledge representation.
3. New computational paradigms and hardware systems.



Wrap up: Many sciences worked on the topics for last decades.

Data Science can be seen as an umbrella comprising all of these areas.

Aspects of data

The four V‘s

- Volume: amount of objects and feature dimensions

  => very large volumes, scaling problems

- Variety: heterogenity of the data, complexity, structure

  => integration of various data sources, structured data and networks Strukturierte Daten, Netzwerke

- Velocity: change of data over the time

  => knowledge aging, change over time, periodic patterns

- Veracity: Uncertainty and data quality

  => measuring uncertainties, wrong observations, incomplete data

Chapter overview

- Knowledge Discovery in Databases, Big Data and Data Science
- Basic Data Mining Tasks (Recap KDD I)
- Topics of KDD II
- Literature and supplementary materials

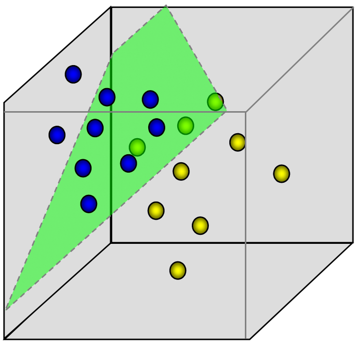
Inhalte KDD I

- Clustering

  partitioning, agglomerative, density-based Clustering etc.
- Outlier Detection
- Classification

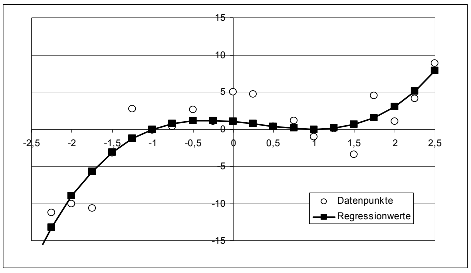
  NN-classification, Bayesian classifiers, SVMs, decision trees
- Assoziation rule mining and frequent pattern mining
- Regression

Classifikation and Regression



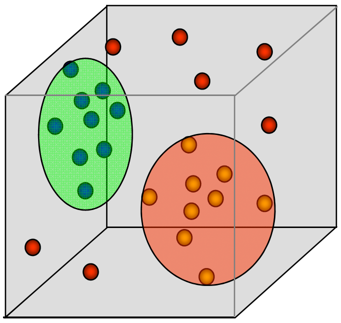
ClassifiCation: (supervised learning)
Example-based learning to distiguish a predefined class set.
⇒ Predict the class of new objects
⇒ describe the characteristics of a class

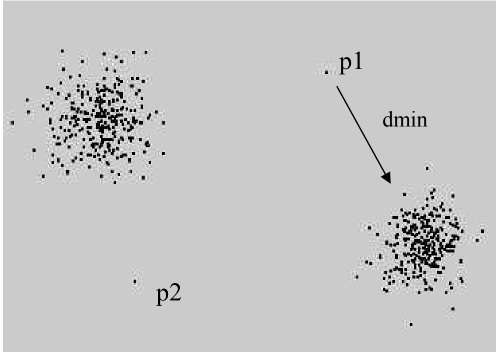
Regression: (supervised learning)
Example-based learning of functions mapping objects to real values:
⇒ determine the value for a new object
⇒ describe the connection between description space and prediction space



Clustering and Outlier Detection

Clustering: (Unsupervised Learning)
Find groups of objects(Cluster) in a data set where:
⇒ objects within a cluster are similar
⇒ objects from different clusters are dissimilar





Outlier Detection:
Find objects which cannot be explained by the known mechanisms in a data set:
⇒ outlier learning (supervised)
⇒ find outliers (via distances) (Unsupervised)

Assoziation Rule and Frequent Patterns

TransaktionsID	Items
2000	A,B,C
1000	A,C
4000	A,D
5000	B,E,F

Assoziation Rules:
Find rules over the elements in a transaction database.
(transaction = subset of the complete set of known items)

⇒ find items which often occurs together in the same transaction

⇒ derive rules on frequent item sets of the following syntax:
If ***A*** *is element of tansaction* ***T*** *then* ***B*** *is contained in* ***T*** *as well with a likelihood of mit* ***x%****.*

Chapter overview

- Knowledge Discovery in Databases, Big Data and Data Science
- Basic Data Mining Tasks (Recap KDD I)
- Topics of KDD II
- Literature and supplementary materials

Chapters overview of KDD II

1. Introduction

Part I: Volume

2. High-Dimensional Data
3. Large Object Cardinalities

Part II: Variety

4. Multi view Data and Ensembles
5. Multi-Instance Data
6. Graph Data

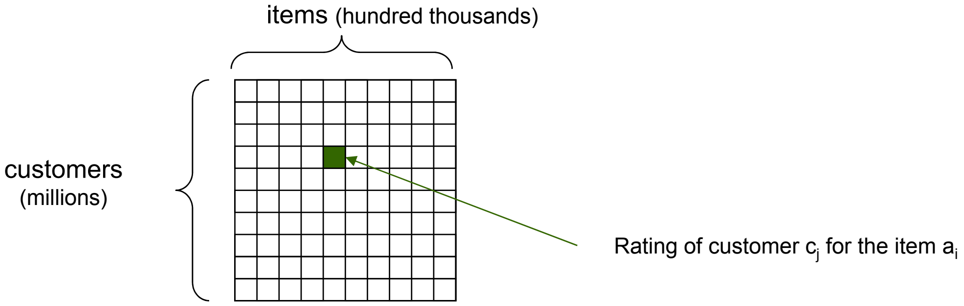
Part III: Velocity

7. Streams

Part I: Volume (1)

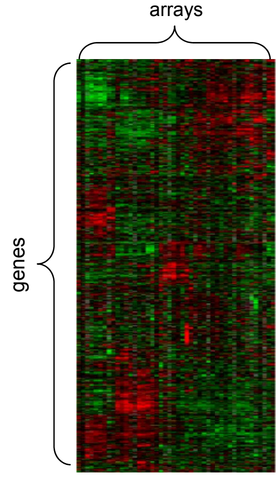
Large dimensionality and/or large object cardinalties

- Real data collections often employ large amounts of features
- Example: Collaborative Filtering
  - User ratings (Filme, Songs, …)
  - Matrix describing the relation between users and items


items (hundred thousands)
customers (millions)
Rating of customer c_j for the item a_i

Part I: Volume (2)

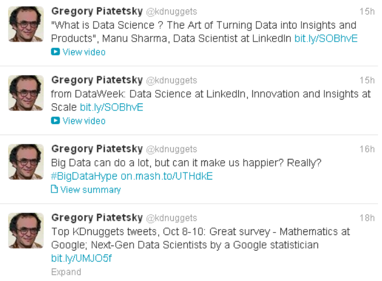
- example: micro array data
  - measures gene expressions
  - often thousands for genes (features)
  - but only 10-100 patients



- examle: text
  - Sinlge words (unigrams) or combined terms (n-grams) as features

    → large numbers of potential attributes

  - represent text documents as vector of word counts



Gregory Piatetsky @kdnuggets 15h
"What is Data Science ? The Art of Turning Data into Insights and Products", Manu Sharma, Data Scientist at LinkedIn bit.ly/SOBhvE
View video

Gregory Piatetsky @kdnuggets 15h
from DataWeek: Data Science at LinkedIn, Innovation and Insights at Scale bit.ly/SOBhvE
View video

Gregory Piatetsky @kdnuggets 16h
Big Data can do a lot, but can it make us happier? Really? #BigDataHype on.mash.to/UTHdkE
View summary

Gregory Piatetsky @kdnuggets 18h
Top KDnuggets tweets, Oct 8-10: Great survey - Mathematics at Google; Next-Gen Data Scientists by a Google statistician bit.ly/UMJO5f
Expand

Part I: Volume

- Recent development of new hardware, infrastructure and services facilities the generation of huge data collections
- example: telecommunication providers
  - Connection data
  - Location data (transmission towers, WLAN Routers)
  - IP connections/ network traffic

- example: WWW
  - Pages, tweets, posts, videos…
- example : Social networks
  - users/ links / groups
  - posts/likes (e.g. for images, text video) / external links

Part I: Volume (3)

- Challenges when mining high-dimensional data
  - Distance measures (for clustering, outlier detection, …) loose their disriminativeness in high dimensions (Curse of Dimensionality)
  - Patterns might occur in different subspaces and projections

    (each pattern might be only observable in certain subspaces)
- Challenges for mining large object cardinalities
  - Avoid quadratic runtime or decouple algorithm complexity and cardinality
  - Employ modern hardware architectures
  - Condsider privacy in distributed settings

Part I: Volume(4)

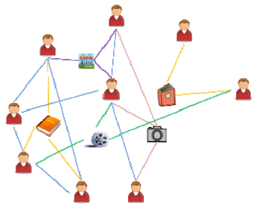
Approaches in the course:

- Feature selection
- Feature reduction
- Metric learning
- Subspace Clustering
- Sampling and Micro-Clustering
- Parallel Data Mining
- Distributed Mining and Privacy

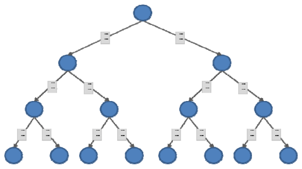
Part II: Variety (1)

- Basic method: object = feature vector

  but: data objects yield complex structures
- **examples**:
  - Graph data: Objects (nodes) have relations (edges) between each other
    - Social networks (e.g. Facebook graph)
    - Co-Authorship Graph (DBLP)
    - Protein interaction networks

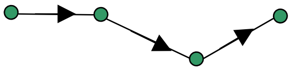


  - Tree structures objects
    - ○ XML documents
    - ○ Sensor networks

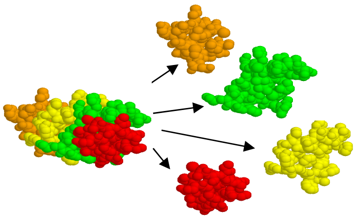


Part II: Variety (2)

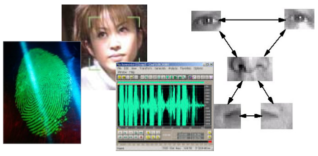
- Further types of structures objects
  - Sequence data:
    - ○ Video data, audio data, time series
    - ○ Trajectories, behavioral pattern



  - Multi-instance objects:
    - ○ Teams, local image descriptors (e.g. SIFT)
    - ○ Multiple measurements, spatial conformations of molecules



  - Multiview objects:
    - ○ Describe images content as combination of form, color and gradient features
    - ○ Describe proteins by primary, secondary and tertiary structure descriptions



Part II: Variety (3)

- Challenges in structured data collections
  - integration of multiple views, similarity measures and models
  - defining structural similarity
  - New pattern and functions are needed to express knowledge

- In the course:
  - Multi-Instance Data Mining
  - Multi-View Data Mining
  - Link-mining
  - Graph-mining

Part III: Velocity

- Example: Environmental monitoring
  - Wireless sensors measure temperature, humidity, polution..
  - Cameras constanty take pictures of public places or sites in nature

- Example: Large Scale Physical Experiments at CERN
  - Experiments generate a petabyete data every second
  - “We don’t store all the data as that would be impractical. Instead, from the collisions we run, we only keep the few pieces that are of interest, the rare events that occur, which our filters spot and send on over the network,”.
  - CERN stores 25PB of selected data each year which is the equivalent of 1000 years of video data in DVD quality.
  - The data is analyzed for hints of the structure of the universe.

http://www.v3.co.uk/v3-uk/news/2081263/cern-experiments-generating-petabyte

Part III: Velocity

- Challenges in volatile data
  - The complete history of data is often not available
  - Knowledge is changing over time
  - Generation of new information limits the time frame for analyzing data

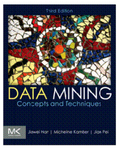
- Topics within the course
  - Data streams, knowledge aging, concept drift, cluster evolution
  - Clustering data streams
  - Classification in data streams

Chapter overview

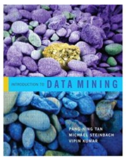
- Knowledge Discovery in Databases, Big Data and Data Science
- Basic Data Mining Tasks (Recap KDD I)
- Topics of KDD II
- Literature and supplementary materials

Literature

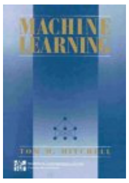
- Han J., Kamber M., Pei J. (English)
  Data Mining: Concepts and Techniques
  3rd ed., Morgan Kaufmann, 2011



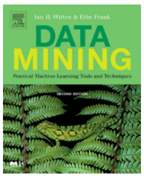
- Tan P.-N., Steinbach M., Kumar V. (English)
  Introduction to Data Mining
  Addison-Wesley, 2006



- Mitchell T. M. (English)
  Machine Learning
  McGraw-Hill, 1997



- Witten I. H., Frank E. (English)
  Data Mining: Practical Machine Learning Tools and Techniques
  Morgan Kaufmann Publishers, 2005


Ian H. Witten & Eibe Frank
DATA MINING
Practical Machine Learning Tools and Techniques

- Ester M., Sander J. (German)
  Knowledge Discovery in Databases: Techniken und Anwendungen
  Springer Verlag, September 2000

Further book titles

- C. M. Bishop, „*Pattern Recognition and Machine Learning*“, Springer 2007.
- S. Chakrabarti, „ *Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data*”, Morgan Kaufmann, 2002.
- R. O. Duda, P. E. Hart, and D. G. Stork, „*Pattern Classification*“, 2ed., Wiley-Inter-science, 2001.
- D. J. Hand, H. Mannila, and P. Smyth, „*Principles of Data Mining*“, MIT Press, 2001.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth: ``*Knowledge discovery and data mining: Towards a unifying framework*‘‘, in: Proc. 2nd ACM Int. Conf. on Knowledge Discovery and Data Mining (KDD), Portland, OR, 1996

Online Resourcen

- *Mining of Massive Datasets* book by Anand Rajaraman and Jeffrey D. Ullman
  - http://infolab.stanford.edu/~ullman/mmds.html
- *Machine Learning* class by Andrew Ng, Stanford
  - http://ml-class.org/
- *Introduction to Databases* class by Jennifer Widom, Stanford
  - http://www.db-class.org/course/auth/welcome
- Kdnuggets: Data Mining and Analytics resources
  - http://www.kdnuggets.com/

KDD /Data Mining tools

- Several options for either commercial or free/ open source tools
  - Check an up to date list at: http://www.kdnuggets.com/software/suites.html
- Commercial tools offered by major vendors
  - e.g., IBM, Microsoft, Oracle …
- Free/ open source tools

Weka

R

SciPy + NumPy

Elki

Environment for
DeveLoping
KDD-Applications
Supported by Index-Structures

Orange

Rapid Miner (free, commercial versions)