

Knowledge Discovery in Databases II

Winter Term 2015/2016

Chapter 1: Introduction and outlook

Lectures : Dr Eirini Ntoutsis, PD Dr Matthias Schubert
Tutorials: PD Dr Matthias Schubert

Script © 2015 Eirini Ntoutsis, Matthias Schubert, Arthur Zimek

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

- **Time and location**

- Lectures: Tuesday, 09:00-12:00, room C006 (Luisenstr. 37 (c))
- Tutorial: Thursday, 16:00-18:00, room 220 (Amalienstr. 73A)
- Tutorial: Friday, 16:00-18:00, room 220 (Amalienstr. 73A)

- All information and news can be found at:

[http://www.dbs.ifi.lmu.de/cms/Knowledge Discovery in Databases II \(KDD II\)](http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_II_(KDD_II))

- **Exam**

- Written exam, 90 min
- 6 ECTS points
- Registration for the written exam:
<https://uniworx.ifi.lmu.de/?action=uniworxCourseWelcome&id=344>

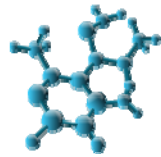
- Knowledge Discovery in Databases, Big Data and Data Science
- Basic Data Mining Tasks (Recap KDD I)
- Topics of KDD II
- Literature and supplementary materials

- Large amounts of data in multiple applications

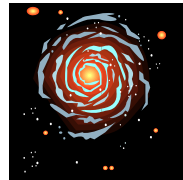
"Drowning in data, yet starving for knowledge. "
<http://www.kdnuggets.com/news/2007/n06/3i.html>



connection data



molecule
process data



telescope data

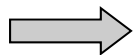


transaction data



Web data/
click streams

- Manual analysis is infeasible



Knowledge Discovery in Databases and Data Mining

Goals

- Descriptive modeling: Explains the characteristics and behavior of observed data
- Predictive modeling: Predicts the behavior of new data based on some model

Important: The extracted models/patterns don't have to apply to 100 % of the cases.

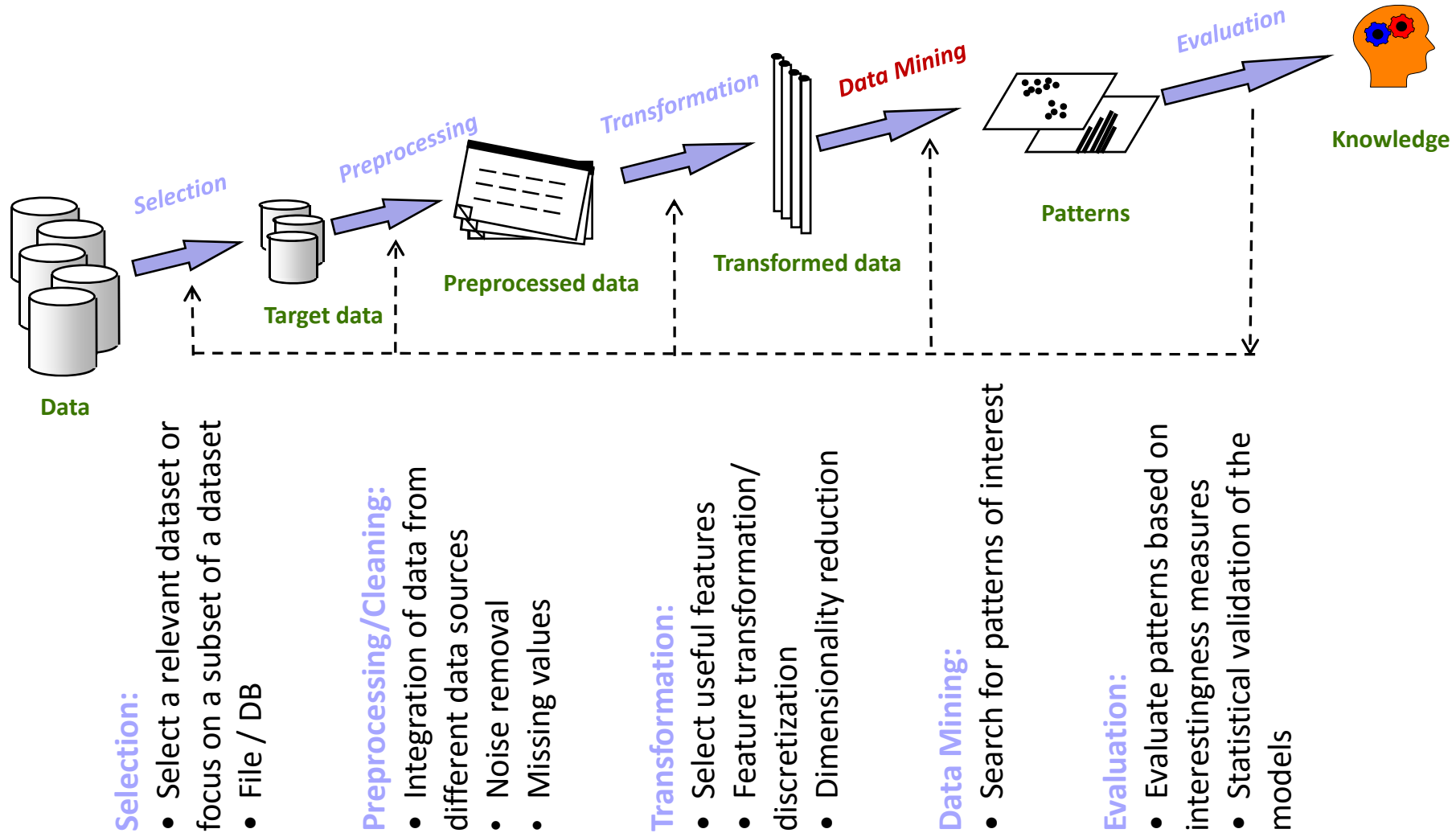
*Knowledge Discovery in Databases (KDD) is the **nontrivial process** of identifying **valid, novel, potentially useful, and ultimately understandable patterns in data.***

[Fayyad, Piatetsky-Shapiro, and Smyth 1996]

Remarks:

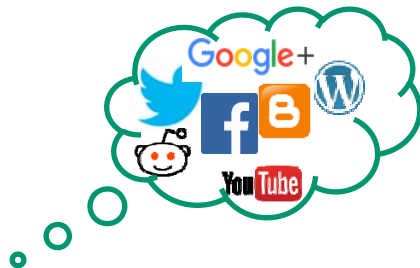
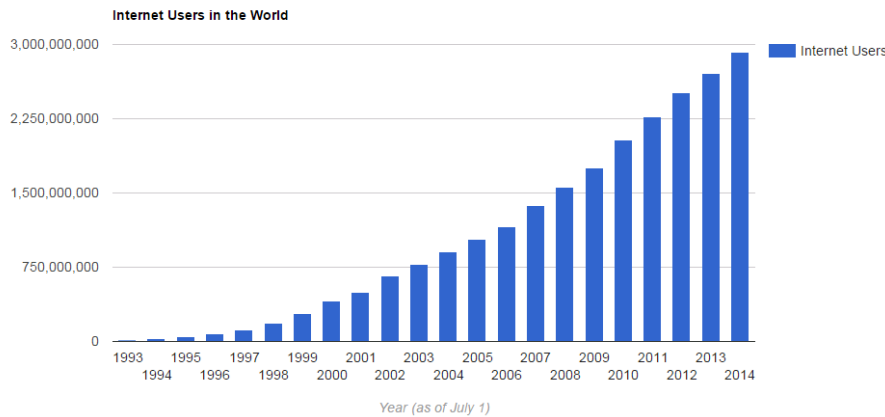
- *nontrivial*: it is not just the avg
- *valid*: to a certain degree the discovered patterns should also hold for new, previously unseen problem instances.
- *novel*: at least to the system and preferable to the user
- *potentially useful*: they should lead to some benefit to the user or task
- *ultimately understandable*: the end user should be able to interpret the patterns either immediately or after some postprocessing

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



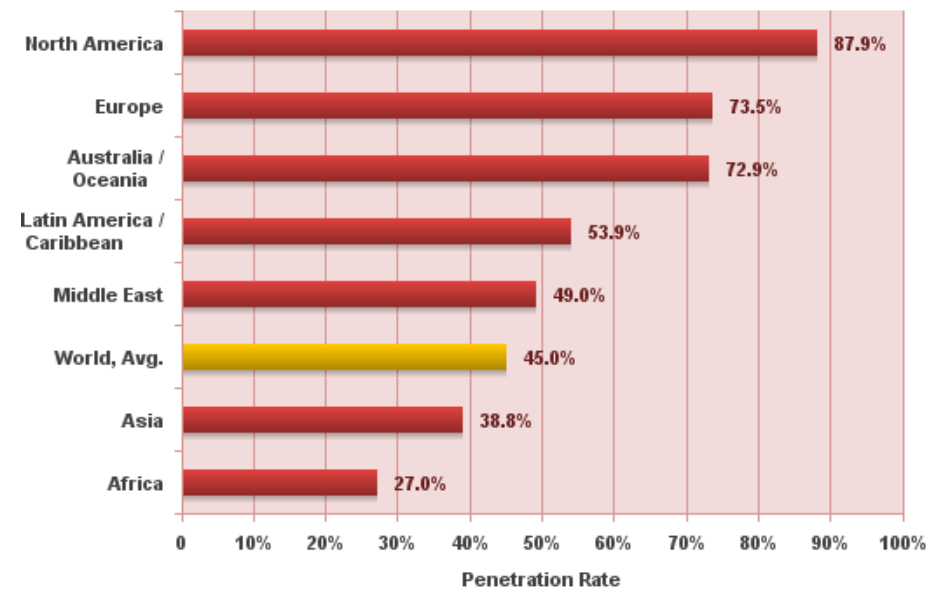
- Internet
- Internet of things
- Data intensive science
- Big data
- Data science
- ...

- Internet users (source: <http://www.internetlivestats.com/internet-users/>)



Web 2.0: A world of opinions

World Internet Penetration Rates by Geographic Regions - 2015 Q2



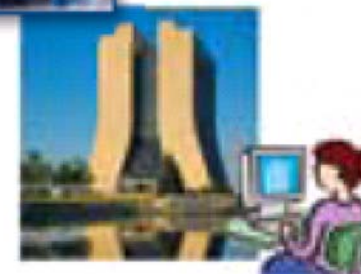
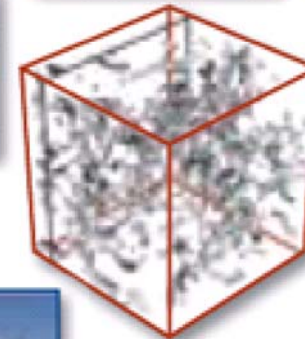
Source: Internet World Stats - www.internetworldststs.com/stats.htm
 Penetration Rates are based on a world population of 7,260,621,118 and 3,270,490,584 estimated Internet users on June 30, 2015.
 Copyright © 2015, Miniwatts Marketing Group

Science Paradigms

- Thousand years ago:
science was **empirical**
describing natural phenomena
- Last few hundred years:
theoretical branch
using models, generalizations
- Last few decades:
a **computational** branch
simulating complex phenomena
- Today: **data exploration** (eScience)
unify theory, experiment, and simulation
 - Data captured by instruments or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files using data management and statistics



$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a^2}$$



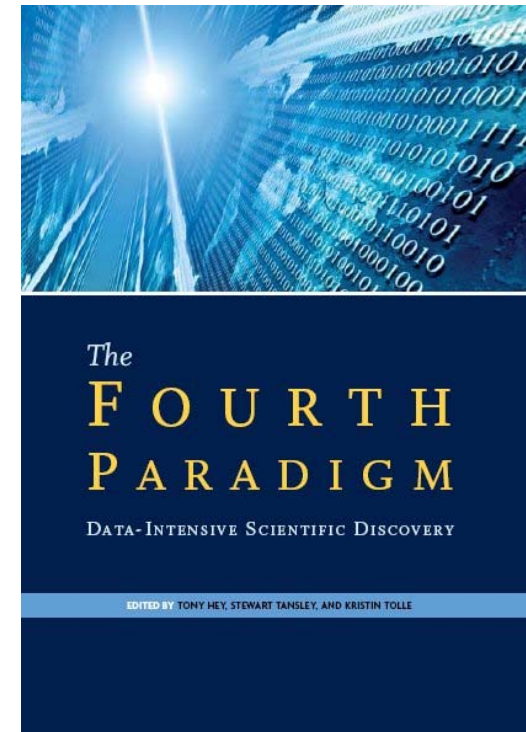
Slide from: http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt

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

-The Fourth Paradigm – Microsoft

Examples of e-science applications:

- Earth and environment
- Health and wellbeing
 - E.g., The Human Genome Project (HGP)
- Citizen science
- Scholarly communication
- Basic science
 - E.g., CERN



“Big data is a broad term for datasets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, and information privacy.”

Source: https://en.wikipedia.org/wiki/Big_data

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

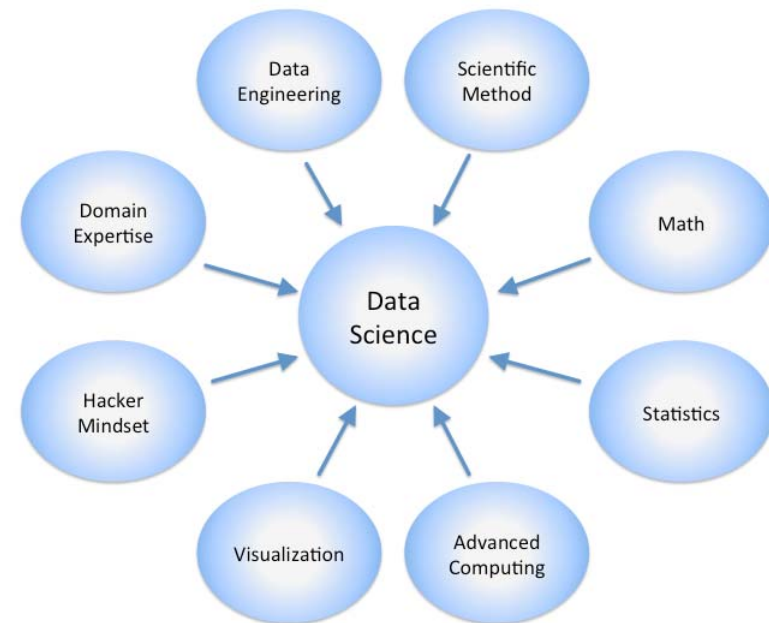
Source: McKinsey Report *“Big data: The next frontier for innovation, competition, and productivity”*, June 2011:

Data Scientist: The sexiest job of the 21st century:

“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.”

Source: <http://tinyurl.com/cplxu6p>

- Science of managing and analyzing data to generate knowledge
- Very similar to KDD, but
 - Data Science is broader in its topics. (result representation, actions..)
 - Integrates all scientific directions being concerned with data analyses and knowledge representation.
 - 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.

Modern data are characterized by:

- **Volume:** amount of objects and features/dimensions
=> very large volumes, scaling problems
- **Velocity:** change of data over the time
=> knowledge aging, change over time, periodic patterns
- **Variety:** heterogeneity of the data, complexity, structure
=> integration of various data sources, structured data and networks
- **Veracity:** Uncertainty and data quality
=> measuring uncertainties, wrong observations, incomplete data

- Knowledge Discovery in Databases, Big Data and Data Science

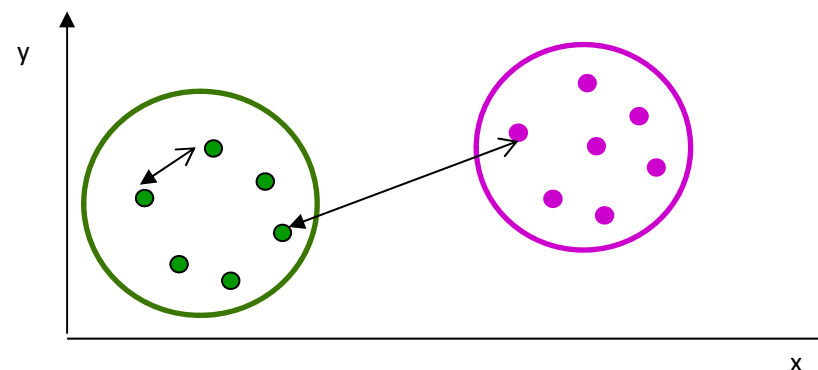
- Basic Data Mining Tasks (Recap KDD I)

- Topics of KDD II

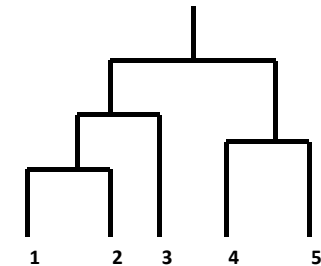
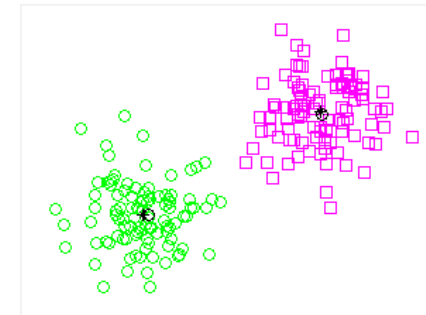
- Literature and supplementary materials

- Clustering
 - partitioning, agglomerative, density-based, grid-based
- Classification
 - NN-classification, Bayesian classifiers, SVMs, decision trees
- Association rule mining and frequent pattern mining
 - Apriori, FP-growth, FI, MFI, CFI
- Regression
- Outlier Detection

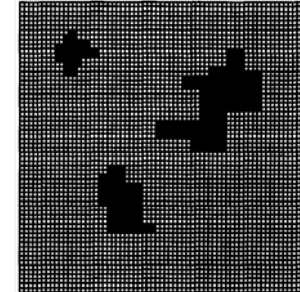
- **Goal:** Group objects into groups so that the objects belonging in the same group are similar (high intra-cluster similarity), whereas objects in different groups are different (low inter-cluster similarity)
- Similarity/ distance function
- Unsupervised learning
- A good clustering method will produce high quality clusters with
 - high intra-cluster similarity
 - low inter-cluster similarity



- Partitioning clustering:
 - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
 - Typical methods: k-means, k-medoids, CLARANS
- Hierarchical clustering:
 - Create a hierarchical decomposition of the set of data (or objects) using some criterion
 - Typical methods: Diana, Agnes, BIRCH, ROCK, CHAMELEON
- Density-based clustering:
 - Based on connectivity and density functions
 - Typical methods: DBSCAN, OPTICS



- Grid-based clustering:
 - based on a multiple-level granularity structure
 - Typical methods: STING, CLIQUE
- Model-based clustering:
 - A model is hypothesized for each of the clusters and tries to find the best fit of that model to each other
 - Typical methods: EM, SOM, COBWEB
- User-guided or constraint-based clustering:
 - Clustering by considering user-specified or application-specific constraints
 - Typical methods: COD (obstacles), constrained clustering



Given:

- a dataset of instances $D=\{t_1,t_2,\dots,t_n\}$ and
- a set of classes $C=\{c_1,\dots,c_k\}$

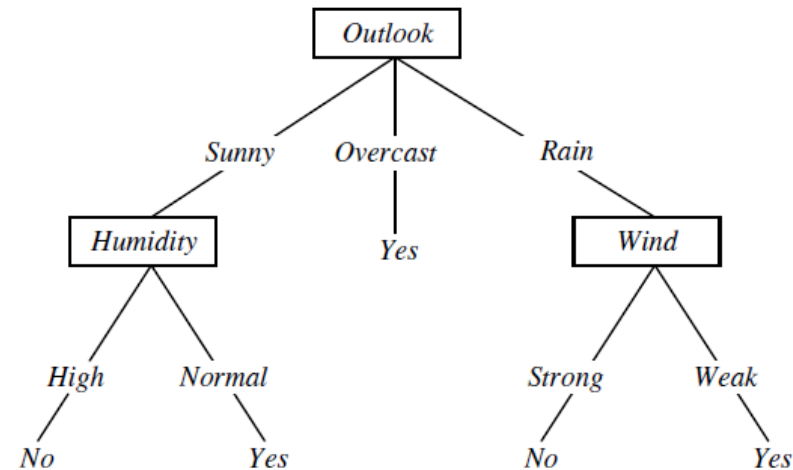
the classification problem is to define a mapping $f:D\rightarrow C$ where each instance t_i in D is assigned to one class c_j .

| | ID | Alter | Autotyp | Risk |
|--------------|-----------|--------------|----------------|-------------|
| Training set | 1 | 23 | Familie | high |
| | 2 | 17 | Sport | high |
| | 3 | 43 | Sport | high |
| | 4 | 68 | Familie | low |
| | 5 | 32 | LKW | low |

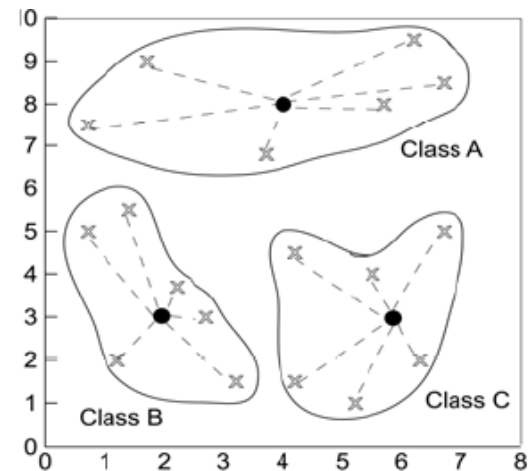
A simple classifier:

- if $\text{Alter} > 50$ then Risk= low;
- if $\text{Alter} \leq 50$ and $\text{Autotyp}=\text{LKW}$ then Risk=low;
- if $\text{Alter} \leq 50$ and $\text{Autotyp} \neq \text{LKW}$ then Risk = high.

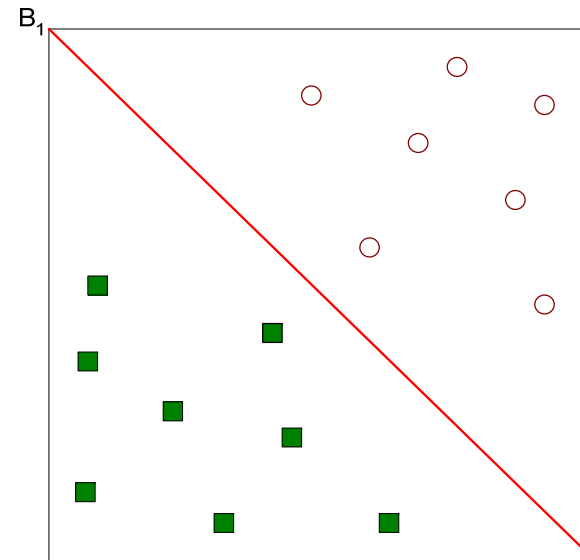
- Decision trees/ Partitioning



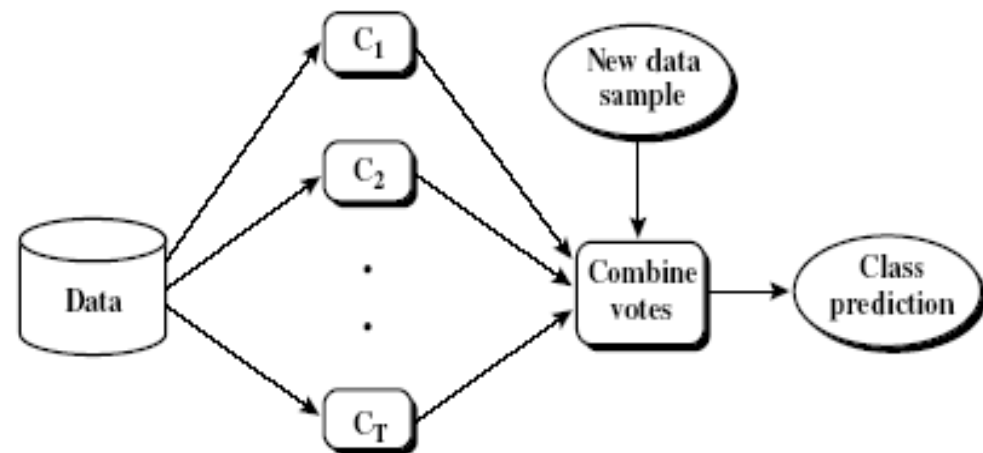
- Nearest Neighbors/ Lazy learners



- SVM

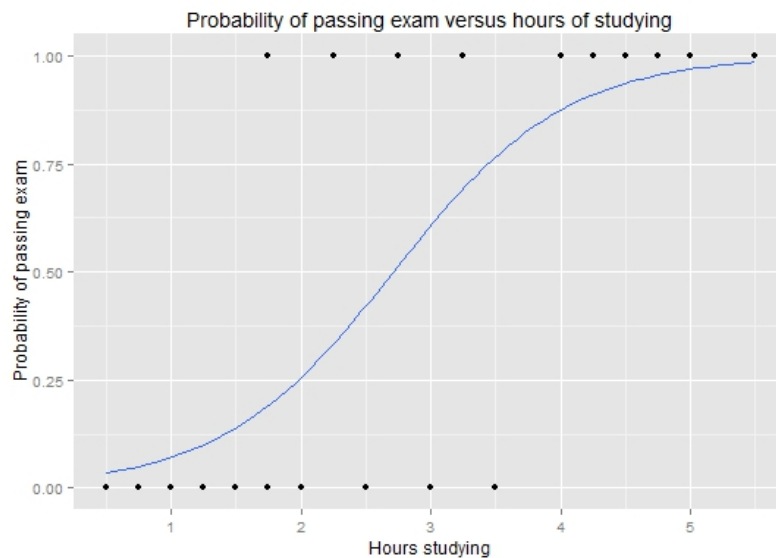


- Ensembles

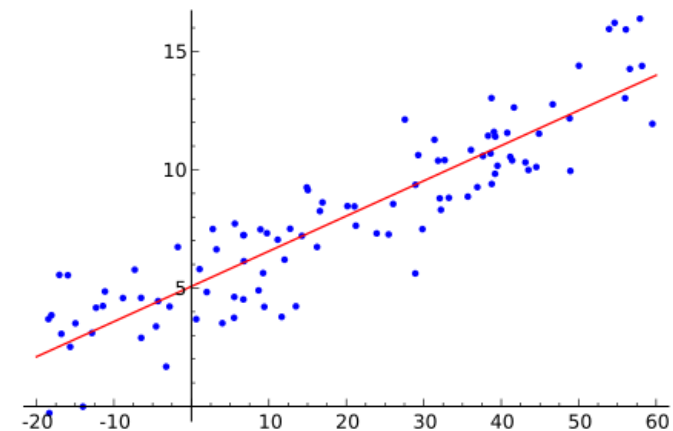


- Mapping objects to real values:
 - ⇒ determine the value for a new object
 - ⇒ describe the connection between description space and prediction space
- Supervised learning task

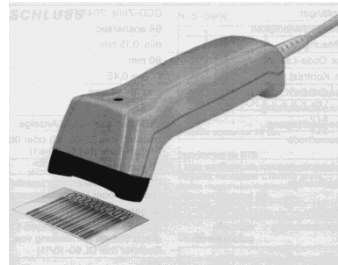
Logistic regression



Linear regression



- Frequent patterns are patterns that appear frequently in a dataset.
 - Patterns: items, substructures, subsequences ...
- Typical example: Market basket analysis



Customer transactions

| Tid | Transaction items |
|-----|----------------------------------|
| 1 | Butter, Bread, Milk, Sugar |
| 2 | Butter, Flour, Milk, Sugar |
| 3 | Butter, Eggs, Milk, Salt |
| 4 | Eggs |
| 5 | Butter, Flour, Milk, Salt, Sugar |

- We want to know: What products were often purchased together?

- e.g.: beer and diapers?



- Applications:

- Improving store layout
- Sales campaigns
- Cross-marketing
- Advertising

The parable of the beer and diapers:

http://www.theregister.co.uk/2006/08/15/beer_diapers/

- **Problem 1:** Frequent Itemsets Mining (FIM)
- Given:
 - A set of items I
 - A transactions database DB over I
 - A *minSupport* threshold s
- Goal: Find all frequent itemsets in DB , i.e.:
- $\{X \subseteq I \mid support(X) \geq s\}$

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

Support of 1-Itemsets:

(A): 75%, (B), (C): 50%, (D), (E), (F): 25%,

Support of 2-Itemsets:

(A, C): 50%,

(A, B), (A, D), (B, C), (B, E), (B, F), (E, F): 25%

- Popular methods: Apriori, FPGrowth

- **Problem 2: Association Rules Mining**
- Given:
 - A set of items I
 - A transactions database DB over I
 - A *minSupport* threshold s and a *minConfidence* threshold c
- Goal: Find all association rules $X \rightarrow Y$ in DB w.r.t. minimum support s and minimum confidence c , i.e.:
- $\{X \rightarrow Y \mid support(X \cup Y) \geq s, confidence(X \rightarrow Y) \geq c\}$
- These rules are called strong.

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

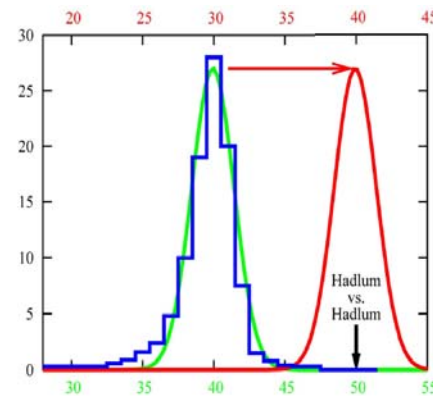
Association rules:

$A \Rightarrow C$ (Support = 50%, Confidence= 66.6%)

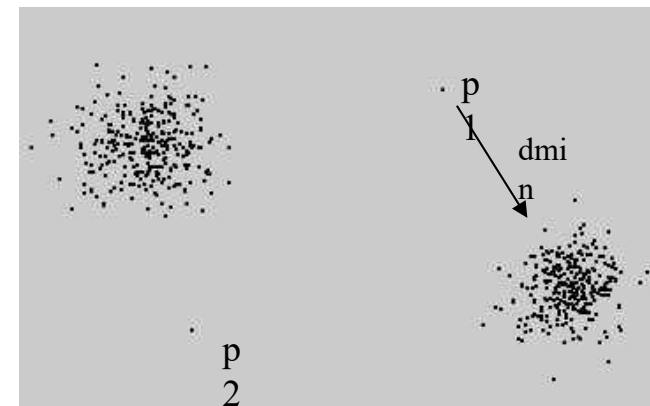
$C \Rightarrow A$ (Support = 50%, Confidence= 100%)

- Goal: find objects that are considerably different from most other objects or unusual or in some way inconsistent with other objects

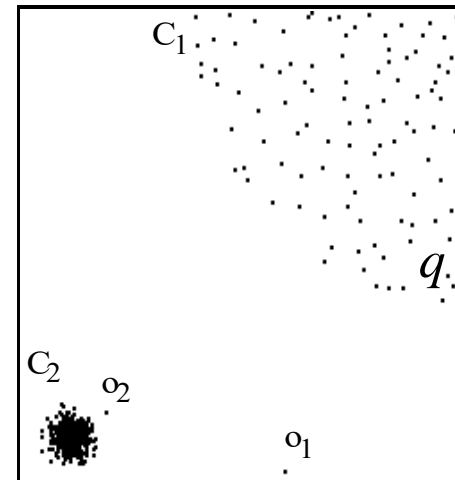
- Statistical approaches



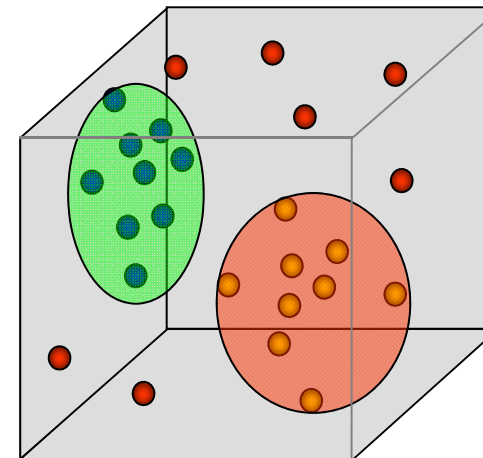
- Distance-based approaches



- Density-based approaches



- Clustering-based approaches



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

1. Introduction

Part I: Volume

2. High-dimensional data

3. Large object cardinalities

Part II: Velocity

4. Data streams

Part III: Variety

5. Multi view Data and Ensembles

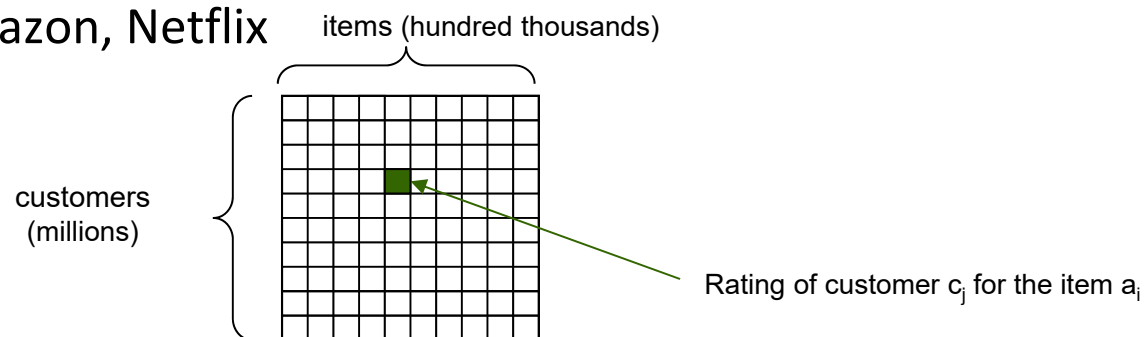
6. Multi-Instance Data

7. Graph Data

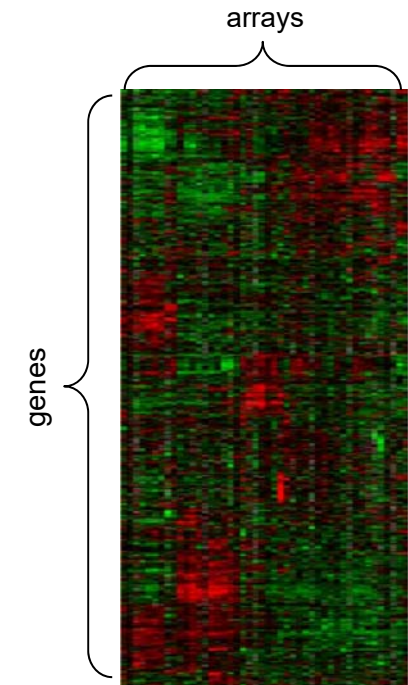
Large object *cardinalities* and/ or large *dimensionality*

- Real applications generate huge amounts of instances (objects)
- We keep storing every single detail about our instances (features)

- High cardinality-High dimensionality example: Recommendation systems
 - High cardinality (users)
 - High dimensionality (items like movies, songs etc)
 - Matrix describing the relation between users and items
 - E.g., Amazon, Netflix



- Low cardinality-High dimensionality example: micro array data
 - measures gene expressions
 - often thousands of genes (features)
 - but only 10-100 patients
- High cardinality-High/Low dimensionality example: text
 - Single words (unigrams) or combined terms (n-grams) as features
 - large numbers of potential attributes
 - represent text documents as vector of word counts




- 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



- Challenges when mining high-dimensional data
 - Distance measures (for clustering, outlier detection, ...) lose their discriminativeness in high dimensions (Curse of Dimensionality)
 - Patterns might occur in different subspaces and projections (each pattern might be only observable in certain subspaces)
- Challenges when mining large object cardinalities
 - Avoid quadratic runtime or decouple algorithm complexity and cardinality
 - Employ modern hardware architectures
 - Consider privacy in distributed settings

Topics within the course:

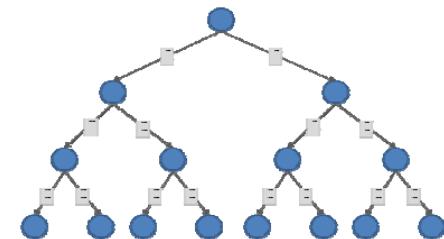
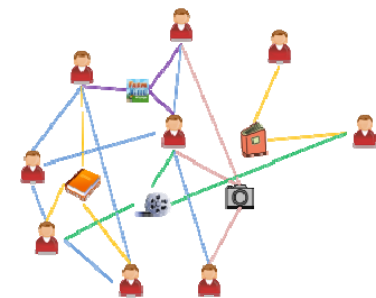
- Feature selection
- Feature reduction
- Metric learning
- Subspace Clustering
- Sampling and Micro-Clustering
- Stream clustering/ classification
- Parallel Data Mining
- Distributed Mining and Privacy

- Example: Environmental monitoring
 - Wireless sensors measure temperature, humidity, pollution..
 - Cameras constantly take pictures of public places or sites in nature
- Example: Large Scale Physical Experiments at CERN 
 - Experiments generate a petabyte 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.

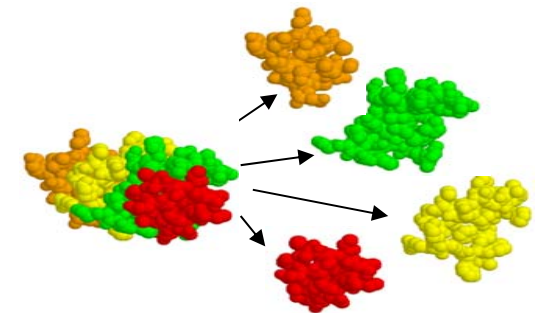
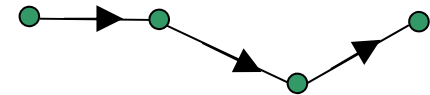
<http://www.v3.co.uk/v3-uk/news/2081263/cern-experiments-generating-petabyte>

- Challenges in volatile data
 - Only 1 look at the data: complete history of data is often not available
 - “Ta panta rhei” Heraclitus: Data is changing over time, same for extracted data mining models
 - Respond fast: Generation of new information limits the time frame for analyzing data
- Topics within the course
 - Data streams, data/knowledge aging, concept drift, evolution
 - Clustering data streams
 - Classification in data streams

- 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



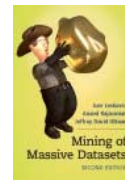
- Further types of structures objects
 - Sequence data:
 - Video data, audio data, time series
 - Trajectories, behavioral pattern
 - Multi-instance objects:
 - Teams, local image descriptors (e.g. SIFT)
 - Multiple measurements, spatial conformations of molecules
 - Multiview objects:
 - Describe images content as combination of form, color and gradient features
 - Describe proteins by primary, secondary and tertiary structure descriptions



- 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

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

- 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
- Leskovec J, Rajaraman A., Ullman J.
Mining of Massive Datasets
Cambridge University Press, 2014
- Ester M., Sander J. (German)
Knowledge Discovery in Databases: Techniken und Anwendungen
Springer Verlag, September 2000



- 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

- *Mining Massive Datasets* class by Jure Lescovec, Anand Rajaraman and Jeffrey D. Ullman
 - <https://www.coursera.org/course/mmds>
- *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/>

- 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



SciPy + NumPy



Orange



Rapid Miner (free, commercial versions)

