Lecture notes

Knowledge Discovery in Databases II

Winter Semester 2012/2013

Lecture 1: Introduction & Overview

Lectures: PD Dr Matthias Schubert, Dr. Eirini Ntoutsi
Tutorials: Erich Schubert

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

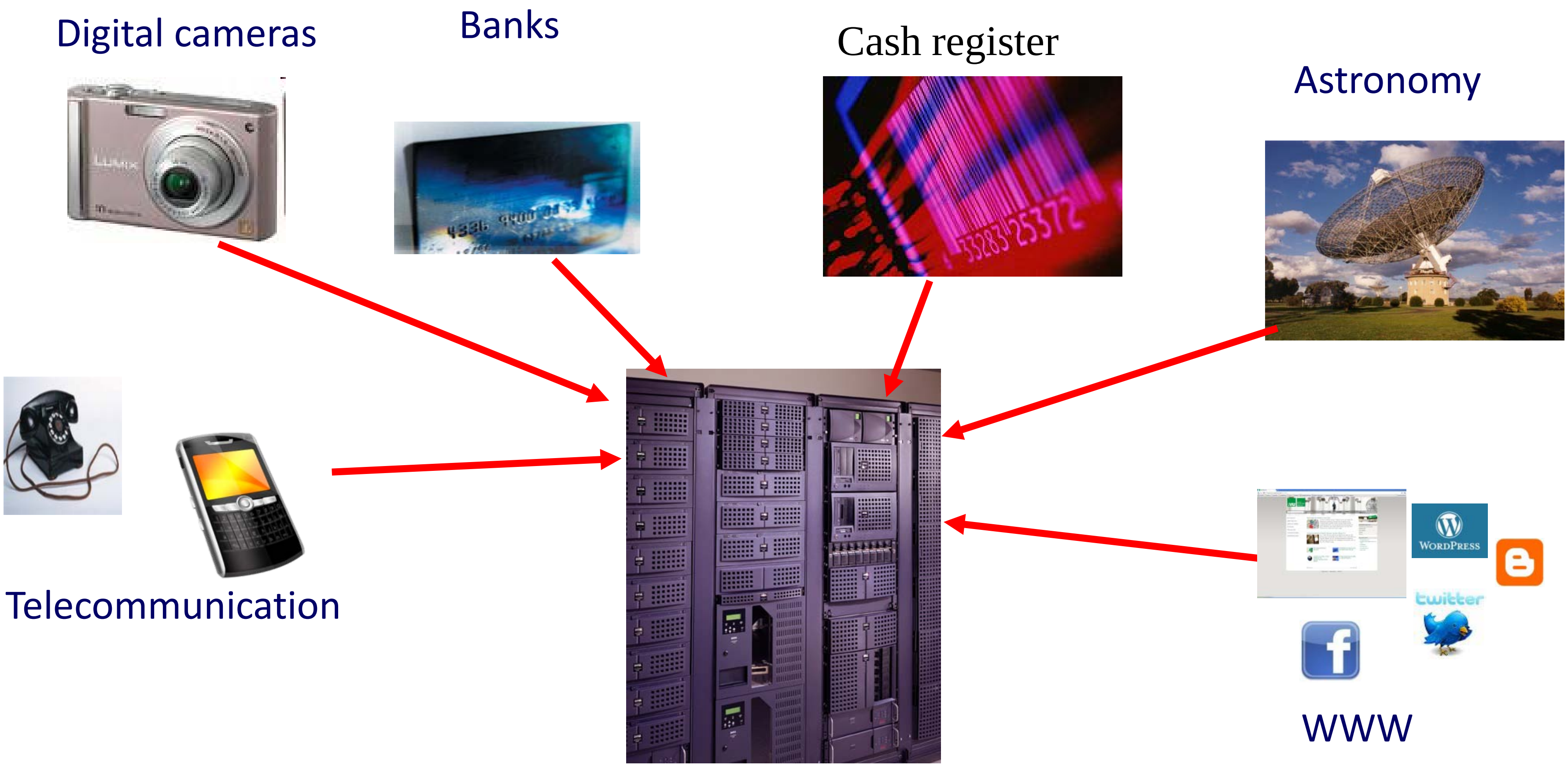
Class information

- **Class schedule**
  - Lectures: Tuesday, 14:00-17:00, Room 109 (Richard-Wagner-Str. 10)
  - Exercises: Thursday, 14:00-16:00, Room 061 (Oettingenstr. 67)
    16:00-18:00, Room 061 (Oettingenstr. 67)
  - Please regularly check the website for updates and other important information (lecture slides, tutorial slides)
    - http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_II_(KDD_II)
- **Exam**
  - You must register in the following url: https://uniworx.ifi.lmu.de/?action=uniworxCourseWelcome&id=91
  - The exam would be based in the material discussed in the class plus the tutorials. The notes are just auxiliary.
- **Grade:**
  - Final exam at the end of the term (written exam, 90 min, 6 ECTS credits)

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks (or overview of KDD I)
- KDD II contents
- Resources
- Things you should know
- Homework/tutorial

Motivation



- Huge amounts of data are collected nowadays from different application domains
- Is not feasible to analyze all these data manually → KDD

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks (or overview of KDD I)
- KDD II contents
- Resources
- Things you should know
- Homework/tutorial

What is KDD?

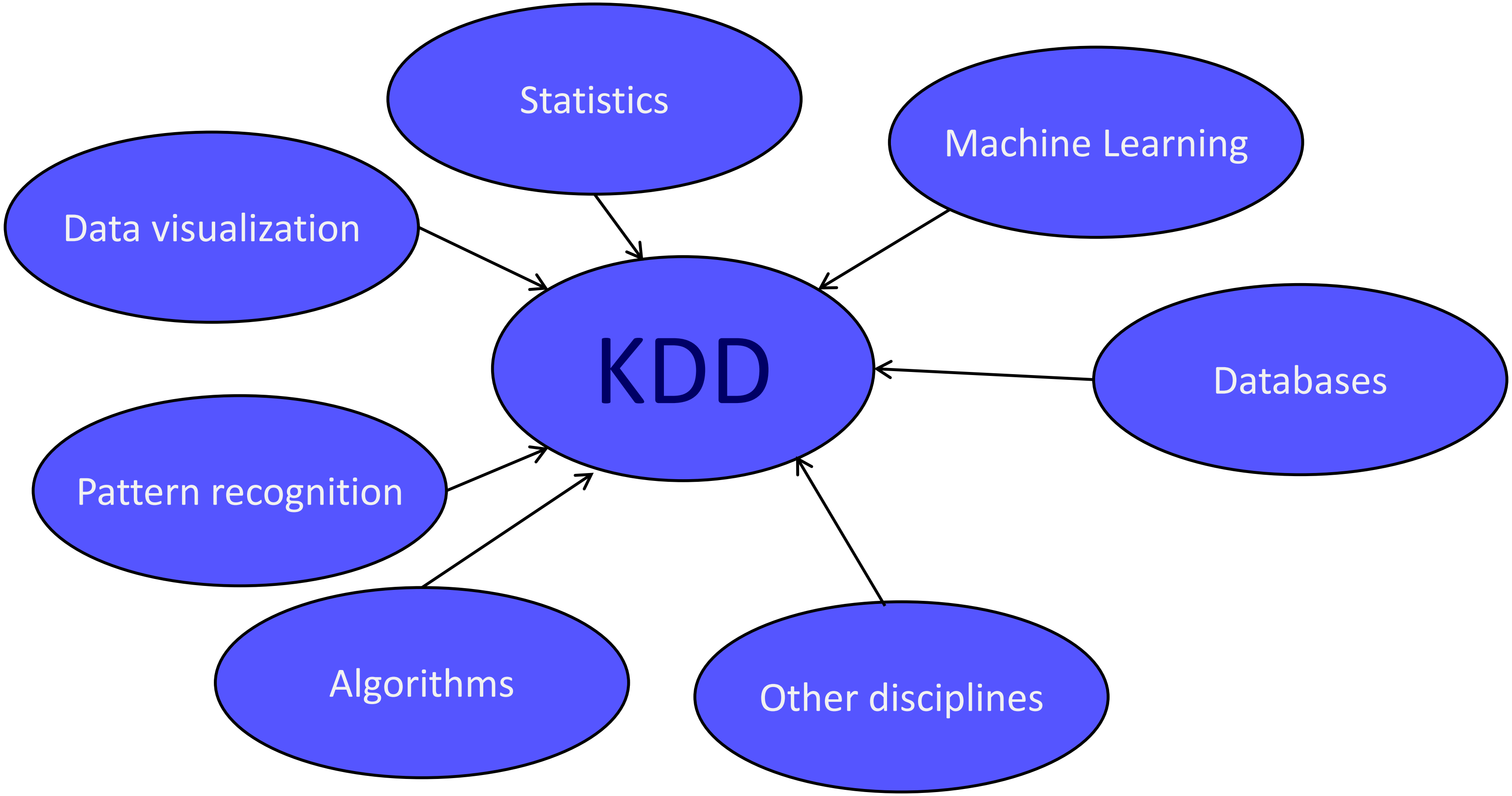
Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

[Fayyad, Piatetsky-Shapiro, and Smyth 1996]

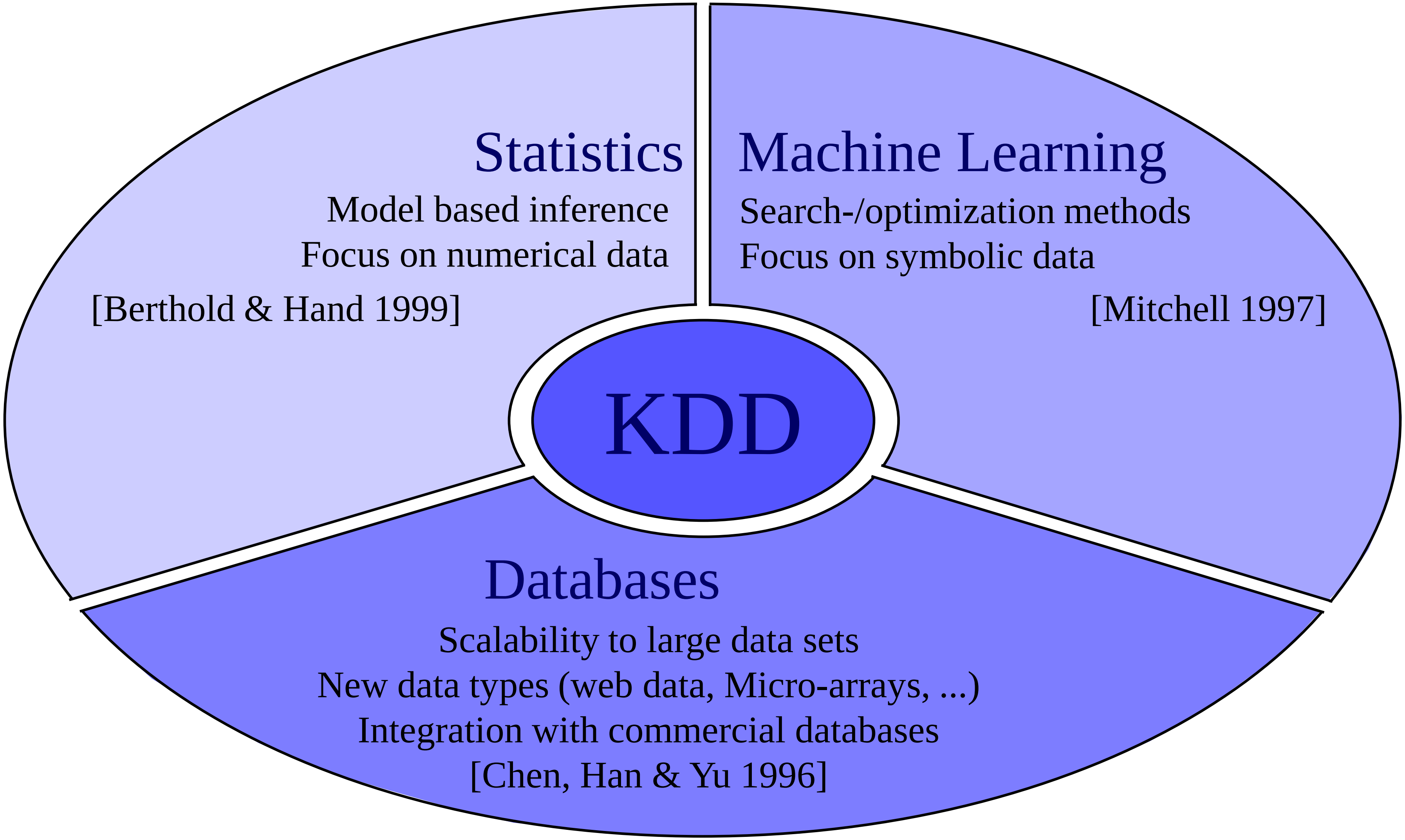
Remarks:

- *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

The interdisciplinary nature of KDD

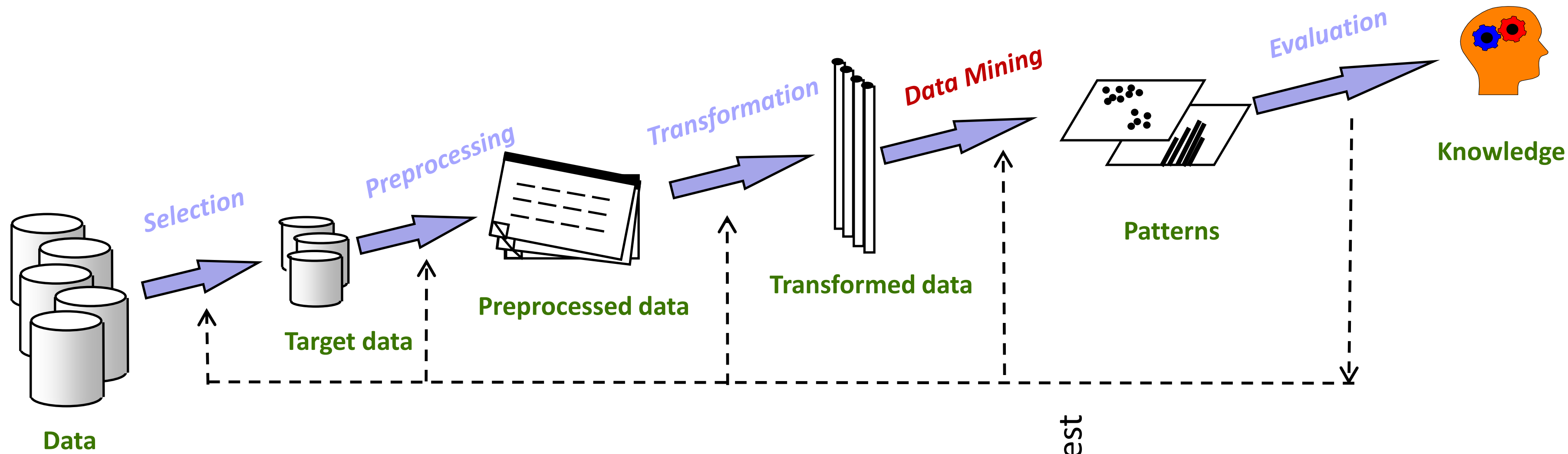


The interdisciplinary nature of KDD



The KDD process

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



Selection:
- Select a relevant dataset or focus on a subset of a dataset
- File / DB

Preprocessing/Cleaning:
- Integration of data from different data sources
- Noise removal
- Missing values

Transformation:
- Select useful features
- Feature transformation/ discretisation
- Dimensionality reduction

Data Mining:
- Search for patterns of interest

Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the models

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks (or overview of KDD I)
- KDD II contents
- Resources
- Things you should know
- Homework/tutorial

Supervised vs Unsupervised learning

There are two different ways of learning from data:

- **Supervised learning:**
  - Learns to predict output from input.
  - The output/ class labels is predefined, e.g. in a loan application it might be «yes» or «no».
  - A set of labeled examples (training set) is provided as input to the learning model. The goal of the model is to extract some kind of «rules» for labeling future data.
  - e.g., Classification, Regression, Outlier detection
- **Unsupervised learning:**
  - Discover groups of similar objects within the data
  - Rely on the characteristics/ features of the data
  - There is no a priori knowledge about the partitioning of the data.
  - e.g., Clustering, Association rules, Outlier detection

The majority of the methods operate on the so called feature vectors, i.e., vectors of numerical features.
However, there are numerous methods that work on other type of data like text, sets, graphs …

Main DM tasks

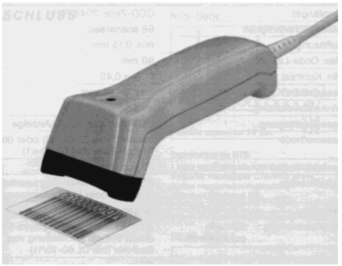
- Frequent Itemsets & Association Rules Mining
  - Apriori, …
- Clustering
  - Partitioning, Hierarchical, Density-based, Grid-based, …
- Classification
  - Decision trees, Nearest-neighbors classifiers, Support Vector Machines, Bayesian classifiers, …
- Outlier detection
- Regression

Also relevant to DMs,

- Data Warehousing
- Performance issues

Frequent Itemsets Mining & Association Rules

- 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?

The parable of the beer and diapers:
http://www.theregister.co.uk/2006/08/15/beer_diapers/

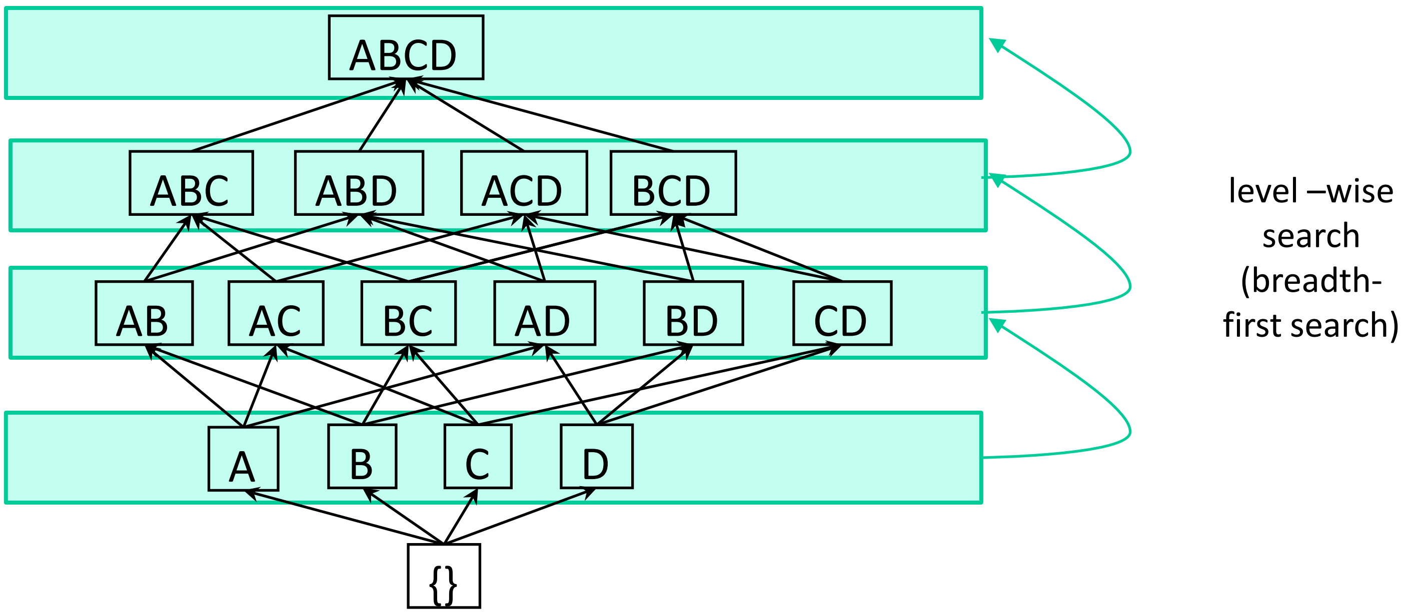
- Applications:
  - Improving store layout
  - Sales campaigns
  - Cross-marketing
  - Advertising

… its not only about market basket data

- Market basket analysis
  - Items are the products
  - Transactions are the products bought by a customer during a supermarket visit
  - Example: Buy(X, “Diapers”) $\rightarrow$ Buy(X, “Beer”) [0.5%, 60%]
- Similarly in an online shop, e.g. Amazon
  - Example: Buy(X, “Computer”) $\rightarrow$ Buy(X, “MS office”) [50%, 80%]
- University library
  - Items are the books
  - Transactions are the books borrowed by a student during the semester
- University
  - Items are the courses
  - Transactions are the courses that are chosen by a student
  - Example: Major (X, “CS”) $\wedge$ Course(X, “DB”) $\rightarrow$ grade(X, “A”) [1%, 75%]
- … and many other applications.
- Also, frequent pattern mining is fundamental in other DM tasks.

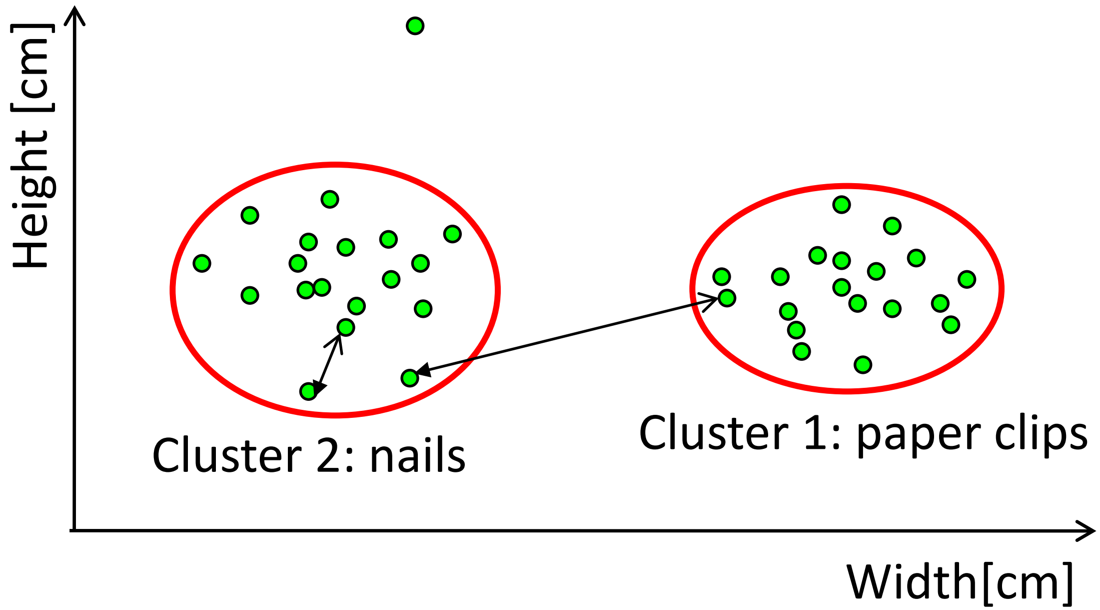
Apriori algorithm [Agrawal & Srikant @VLDB'94]

- First, frequent 1-itemsets are determined, then frequent 2-itemsets and so on



- Method:
  - Initially, scan DB once to get frequent 1-itemset
  - Generate length (k+1) candidate itemsets from length k frequent itemsets
  - Test the candidates against DB (one scan) *Downward closure property, minSupport threshold*
  - Terminate when no frequent or candidate set can be generated

Clustering

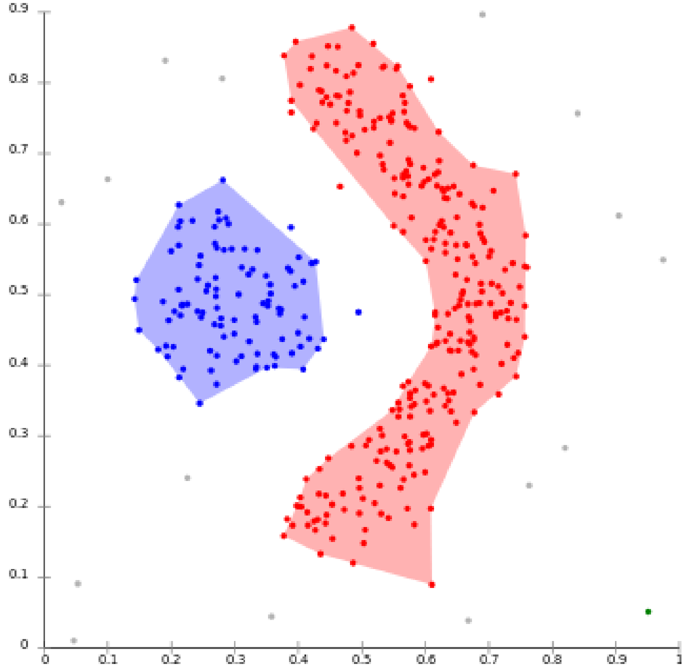


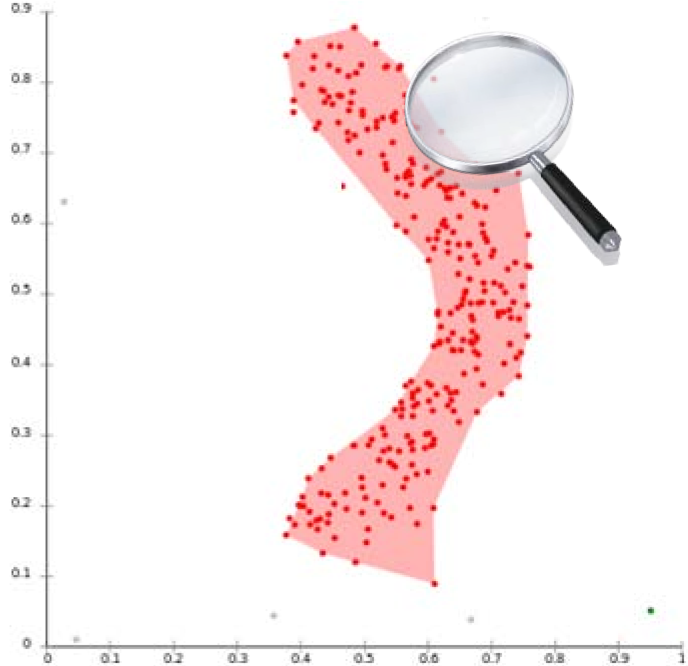
Clustering can be defined as the decomposition of a set of objects into subsets of similar objects (the so called *clusters*)

Goal: Group objects into groups so that the objects belonging in the same group are similar (high intra-class similarity), whereas objects in different groups are different (low inter-class similarity)

Why clustering?

- Clustering is widely used as:
  - As a stand-alone tool to get insight into data distribution
  - As a preprocessing step for other algorithms





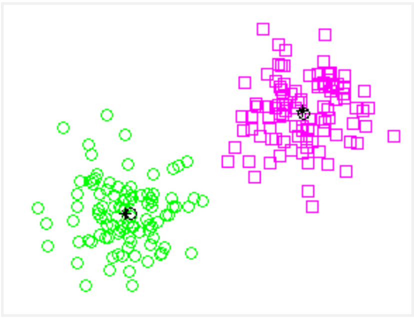
http://en.wikipedia.org/wiki/Cluster_analysis

Example applications

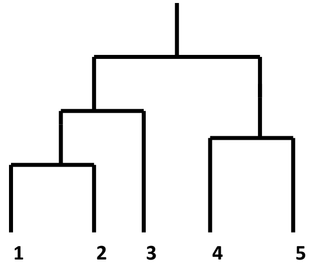
- Marketing:
  - Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- Telecommunications:
  - Build user profiles based on usage and demographics and define profile specific tariffs and offers
- Land use:
  - Identification of areas of similar land use in an earth observation database
- City-planning:
  - Identifying groups of houses according to their house type, value, and geographical location
- Bioinformatics:
  - Cluster similar proteins together (similarity wrt chemical structure and/or functionality etc)
- Web:
  - Cluster users based on their browsing behavior
  - Cluster pages based on their content (e.g. News aggregators)

Major clustering methods I

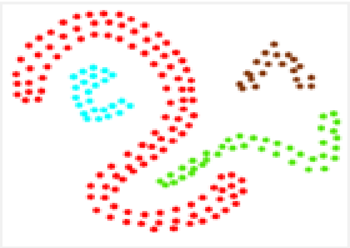
- Partitioning approach:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-Means, k-medoids, CLARANS



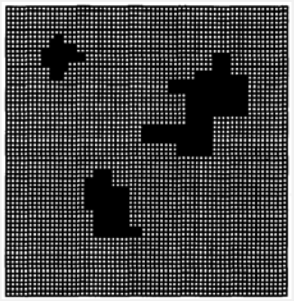
- Hierarchical approach:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, ROCK, CHAMELEON



- Density-based approach:
  - Based on connectivity and density functions
  - Typical methods: DBSCAN, OPTICS, DenClue



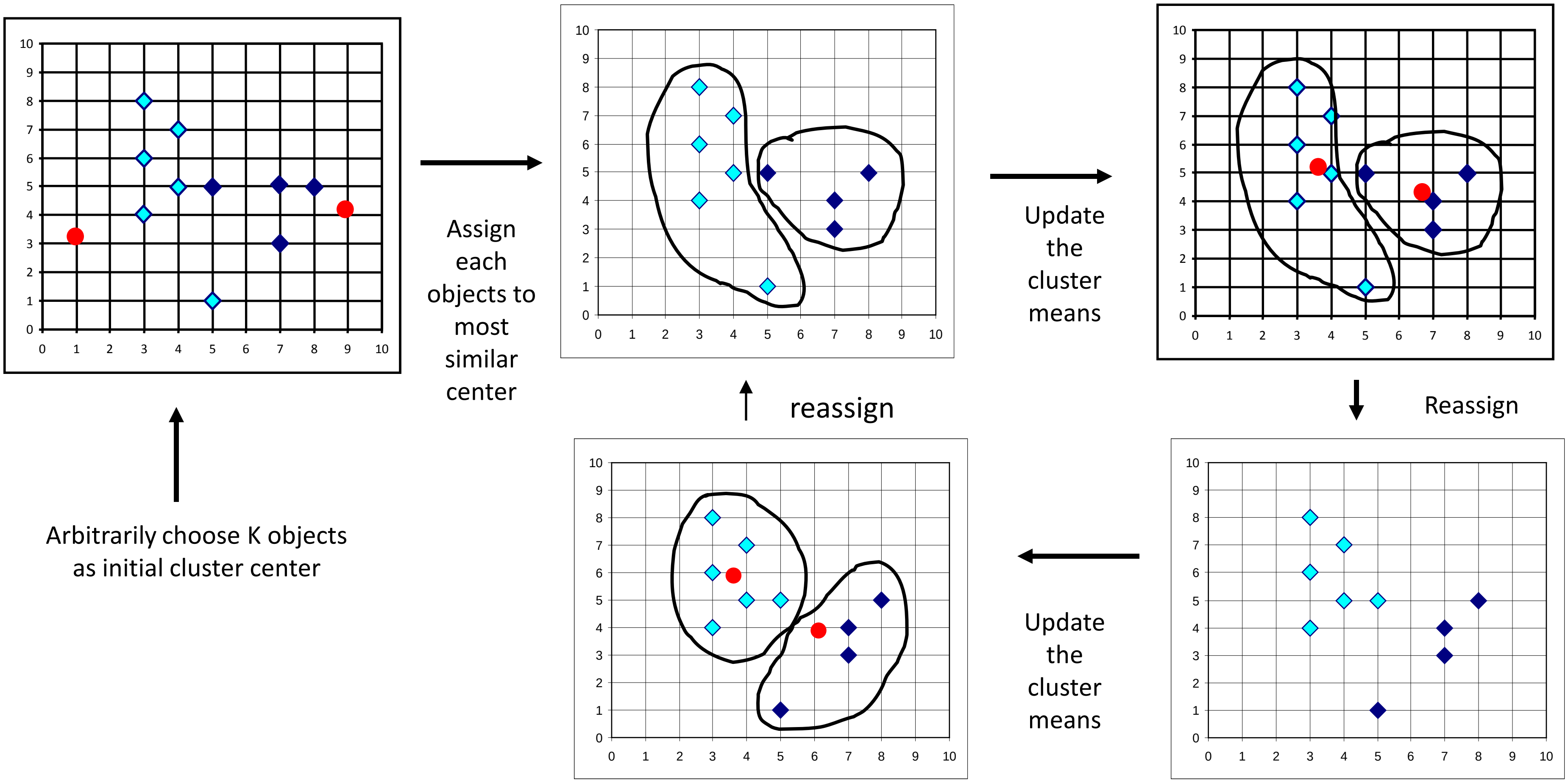
Major clustering methods II



- Grid-based approach:
  - based on a multiple-level granularity structure
  - Typical methods: STING, WaveCluster, CLIQUE
- Model-based:
  - 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
- Frequent pattern-based:
  - Based on the analysis of frequent patterns
  - Typical methods: pCluster
- User-guided or constraint-based:
  - Clustering by considering user-specified or application-specific constraints
  - Typical methods: COD (obstacles), constrained clustering

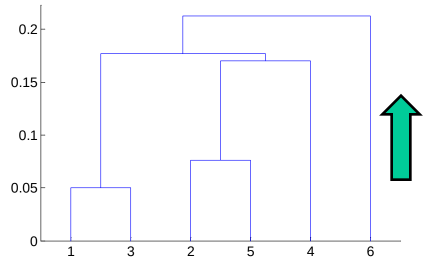
k-Means example

- k=2

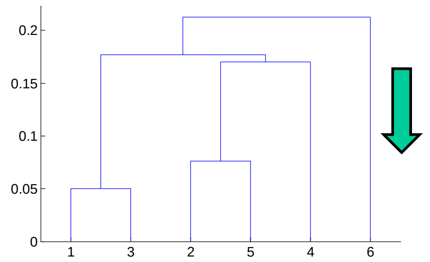


Hierarchical clustering methods

- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or *k* clusters) left
    - e.g., AGNES



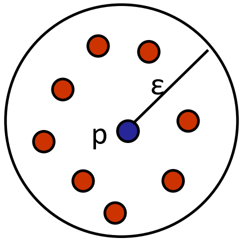
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are *k* clusters)
    - e.g., DIANA

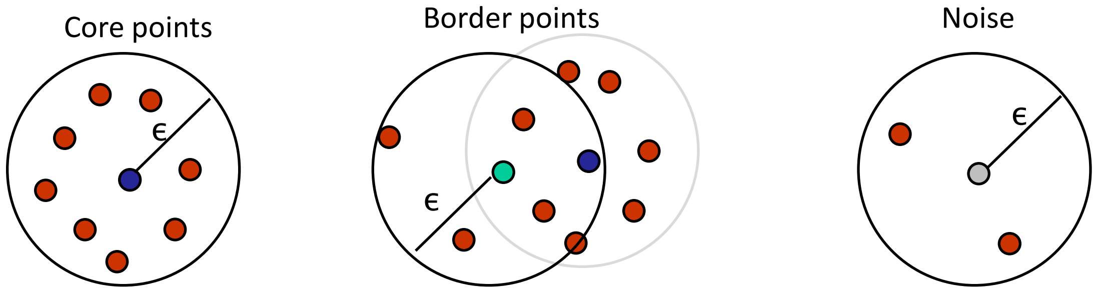


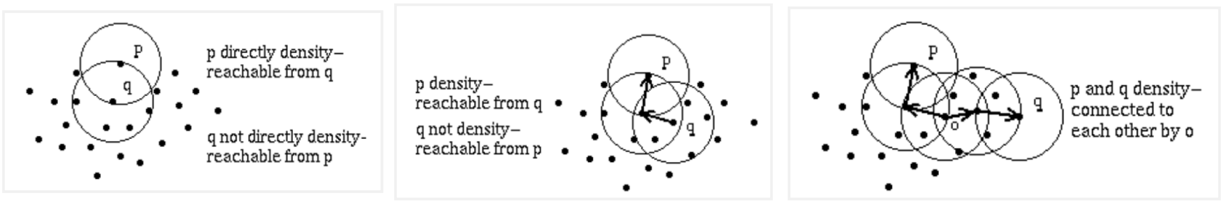
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Different ways to define similarity between clusters (single link, complete link, group average, centroid distance, …)
  - Merge or split one cluster at a time

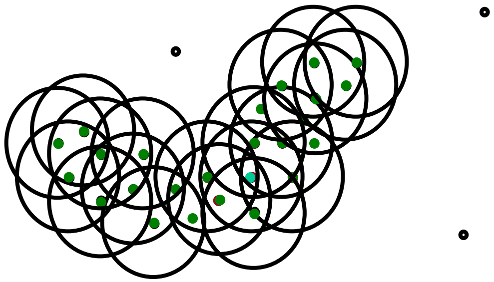
DBSCAN (Ester et al, KDD'96)

- Two parameters*:*
  - Eps (or ε): Maximum radius of the neighbourhood
  - MinPts: Minimum number of points in an Eps-neighbourhood of that point
- Eps-neighborhood of a point p in D
  - $N_{Eps}(p)$: {q belongs to D | dist(p,q) <= Eps}



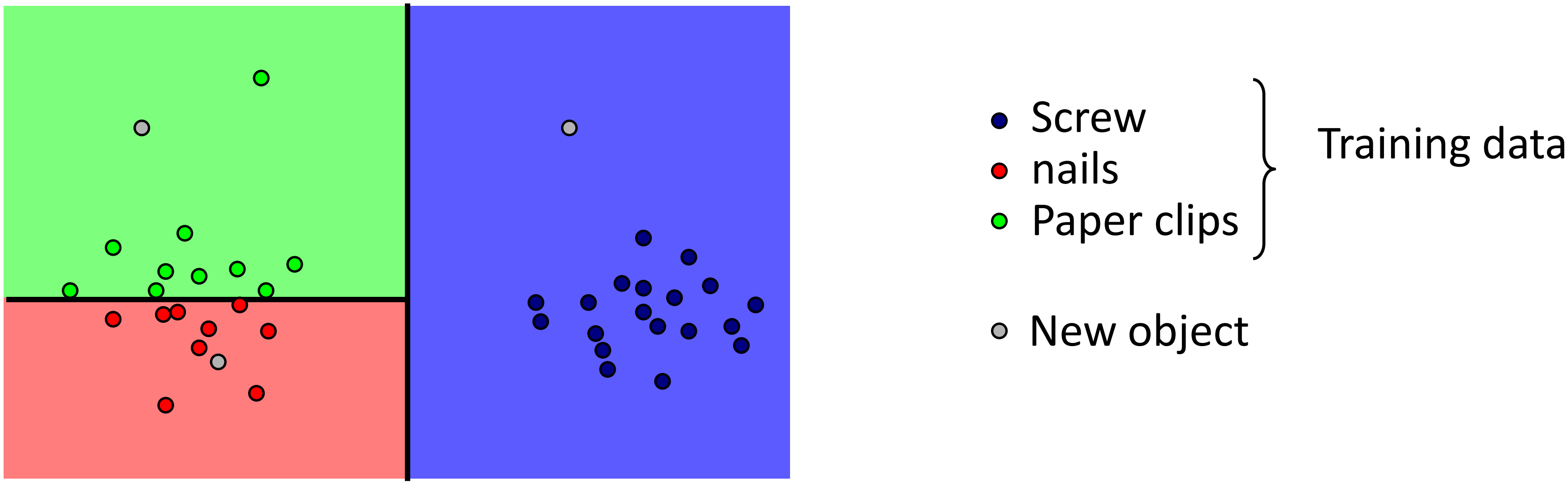






A cluster is a maximal set of density-connected points

Classification



Task:

Learn from the already classified training data, the rules to classify new objects based on their characteristics.

The result attribute (class variable) is nominal (categorical)

A simple classifier

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

A simple classifier:

if Alter > 50 **then** Risk= low;

if Alter $\leq$ 50 **and** Autotyp=LKW **then** Risk=low;

if Alter $\leq$ 50 **and** Autotyp $\neq$ LKW **then** Risk = high.

Applications

- Credit approval
  - Classify bank loan applications as e.g. safe or risky.
- Fraud detection
  - e.g., in credit cards
- Churn prediction
  - E.g., in telecommunication companies
- Target marketing
  - Is the customer a potential buyer for a new computer?
- Medical diagnosis
- Character recognition
- …

Overview of the classification process

- **Model construction:** describing a set of predetermined classes
  - The set of tuples used for model construction is training set
  - Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute
  - The model is represented as classification rules, decision trees, or mathematical formulae
- **Model evaluation:** estimate accuracy of the model
  - The set of tuples used for model evaluation is test set
  - The class label of each tuple/sample in the test set is known in advance
  - The known label of test sample is compared with the classified result from the model
    - Accuracy rate is the percentage of test set samples that are correctly classified by the model
  - Test set is independent of training set, otherwise over-fitting will occur
- **Model usage:** for classifying future or unknown objects
  - If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known

predefined class values

Class attribute: tenured={yes, no}

Training set

NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

known class label attribute

Test set

NAME	RANK	YEARS	TENURED	PREDICTED
Maria	Assistant Prof	3	no	no
John	Associate Prof	7	yes	no
Franz	Professor	3	yes	yes

known class label attribute

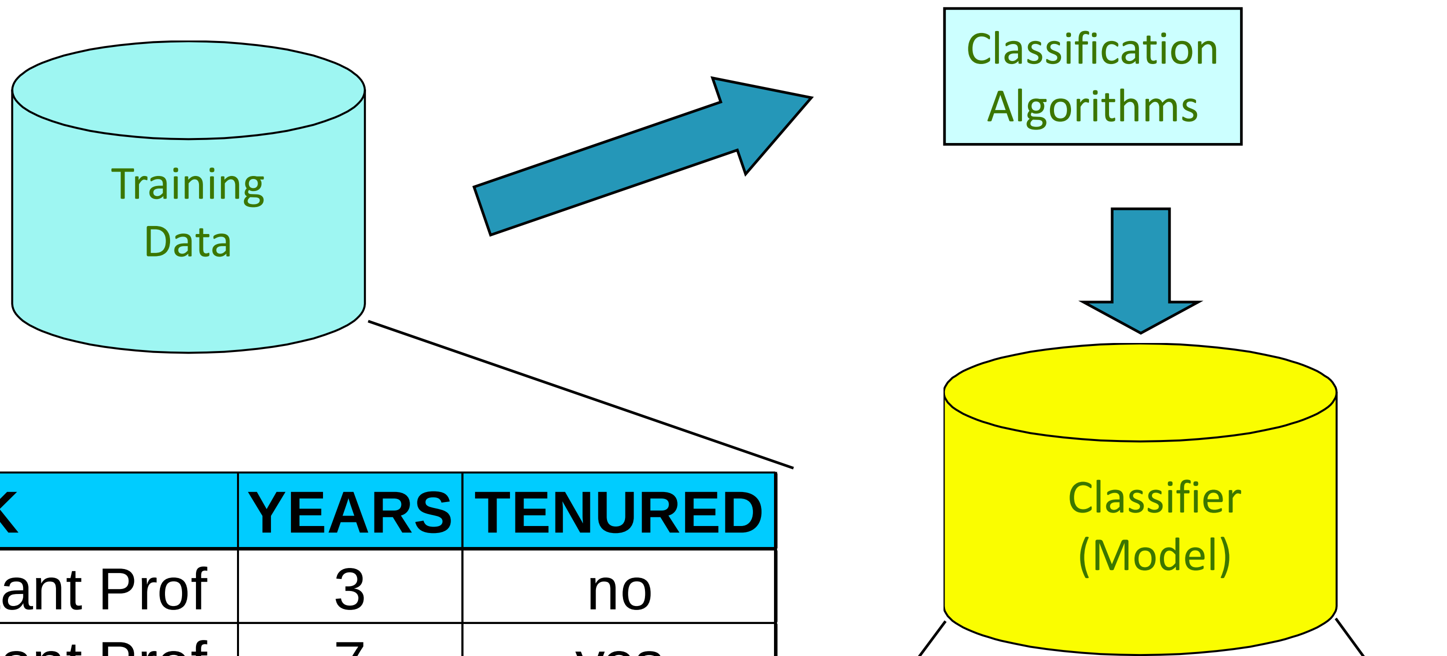
predicted class value by the model

NAME	RANK	YEARS	TENURED	PREDICTED
Jeff	Professor	4	?	yes
Patrick	Associate Prof	8	?	yes
Maria	Associate Prof	2	?	no

unknown class label attribute

predicted class value by the model

Model construction



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Attributes

Class attribute

IF rank = 'professor' OR years > 6
THEN tenured = 'yes'

IF (rank!='professor') AND (years < 6)
THEN tenured = 'no'

Model evaluation

Testing
Data

Classifier
(Model)

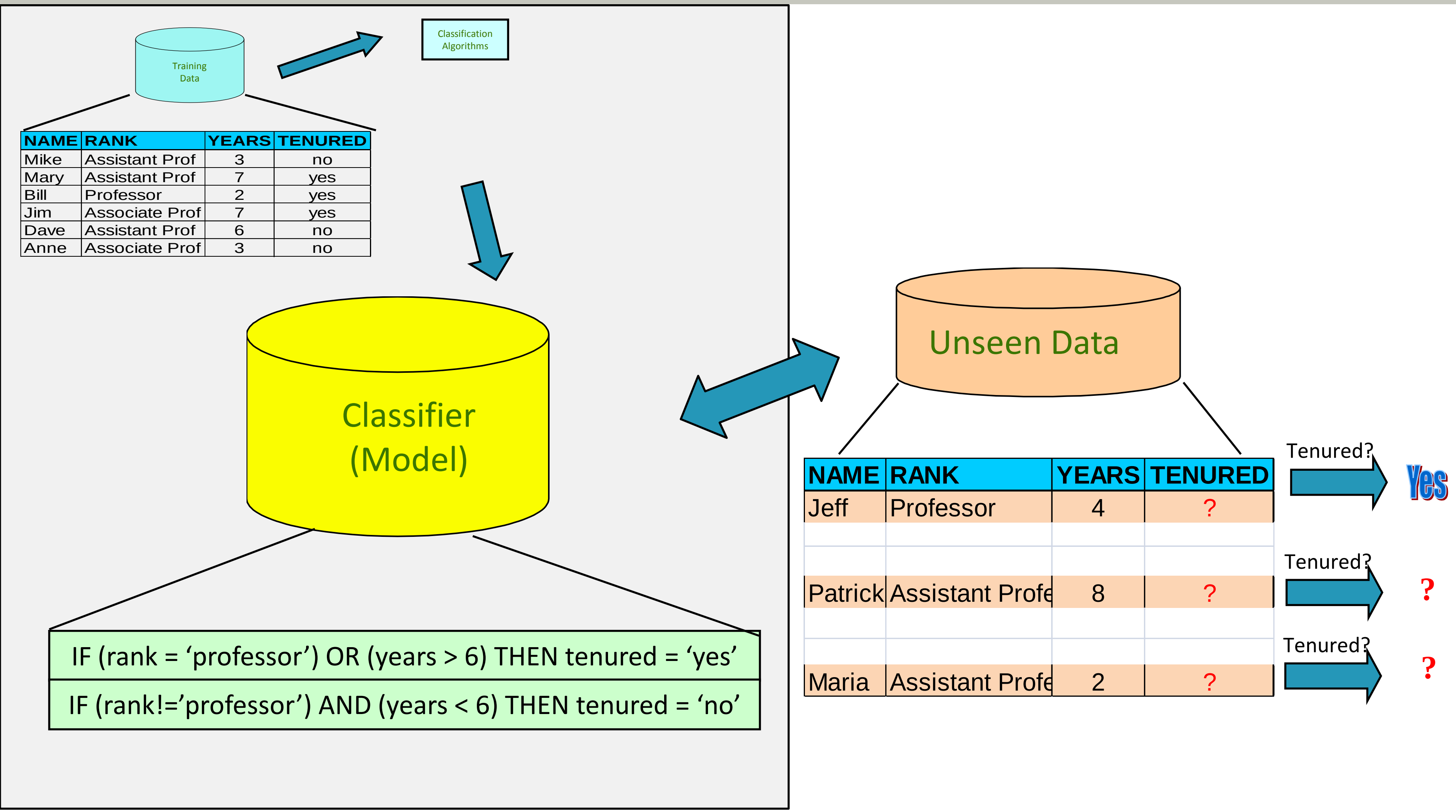
NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

IF rank = ‘professor’ OR years > 6
THEN tenured = ‘yes’

IF (rank!=’professor’) AND (years < 6)
THEN tenured = ‘no’

Classifier quality
Is it acceptable?

Model usage for prediction



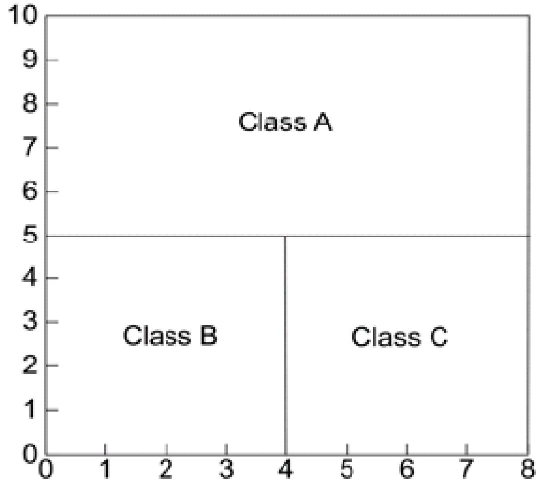
NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

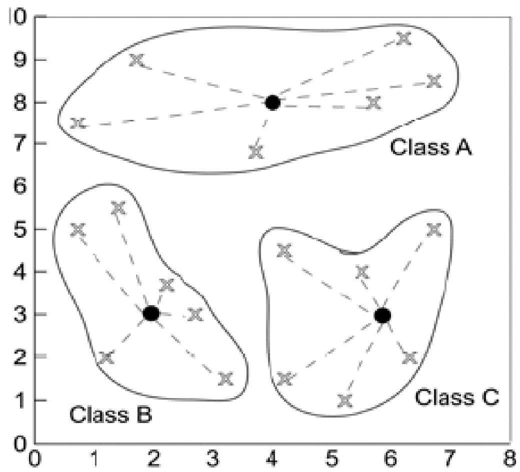
IF (rank = 'professor') OR (years > 6) THEN tenured = 'yes'
IF (rank!='professor') AND (years < 6) THEN tenured = 'no'

NAME	RANK	YEARS	TENURED
Jeff	Professor	4	?
Patrick	Assistant Profe	8	?
Maria	Assistant Profe	2	?

Classification techniques

- Statistical methods
  - Bayesian classifiers etc
- Partitioning methods
  - Decision trees etc
- Similarity based methods
  - K-Nearest Neighbors etc

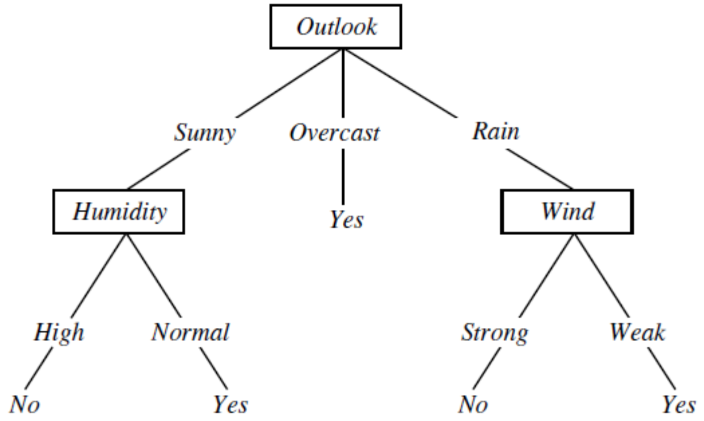




Decision tree classifiers

- A partition-based method

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No



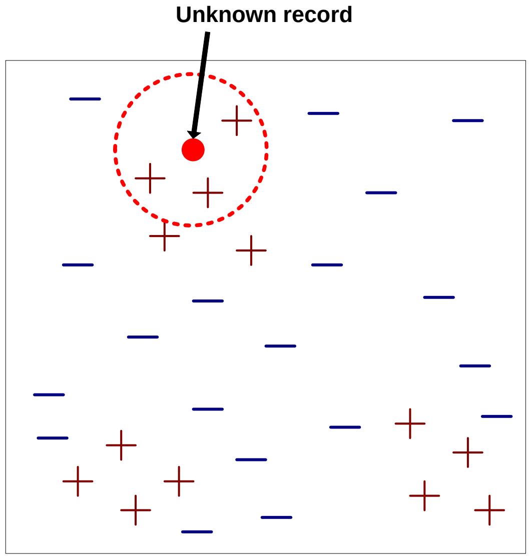
- Selecting the best attribute for splitting
- Avoiding overfitting

Naïve Bayes classifiers

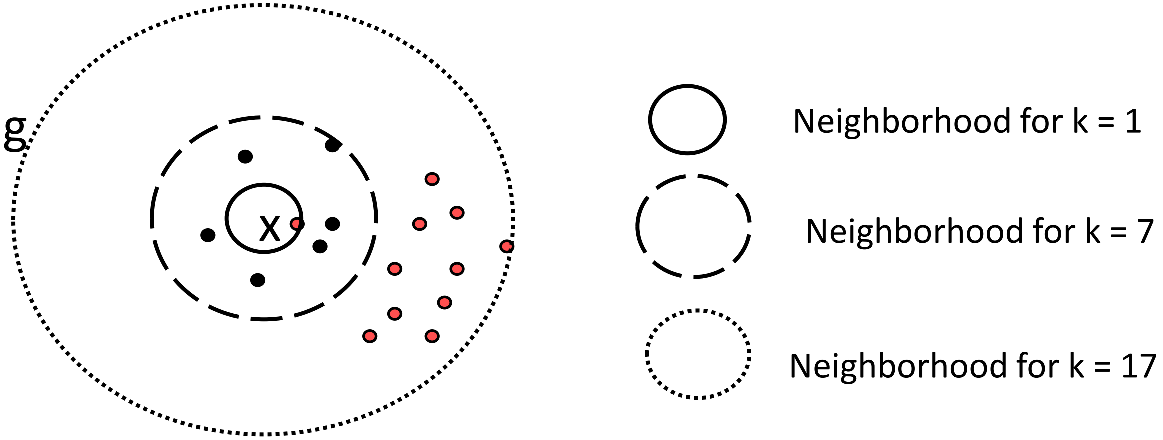
- A statistical method
- Maximum likelihood classification $c = \arg\max_{c \in C} P(c \mid X)$
- Bayes Rule $c = \arg\max_{c \in C} \frac{P(X \mid c)P(c)}{P(X)} = \arg\max_{c \in C} P(X \mid c)P(c)$
- Independency assumption: $\mathrm{P(X \mid c) = P(A_1 A_2 \ldots A_n \mid c)} = \prod P(A_i \mid c)$
- Estimating:
  - P(c)
  - $P(A_i \mid c)$
- Dealing with 0 probabilities

kNN classifiers

- A similarity-based method
- Learning from your neighbors
- Lazy learner



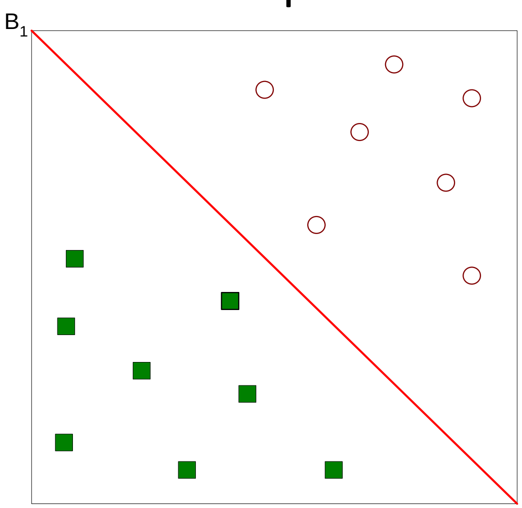
- Distance function
- # of neighbors (k)
- Voting
  - Majority voting
  - Weighted voting



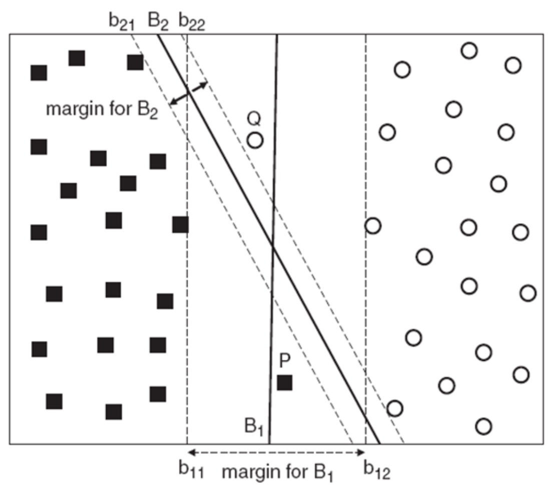
Support Vector machines

- A statistical method
- Maximizes the margin of the decision boundary

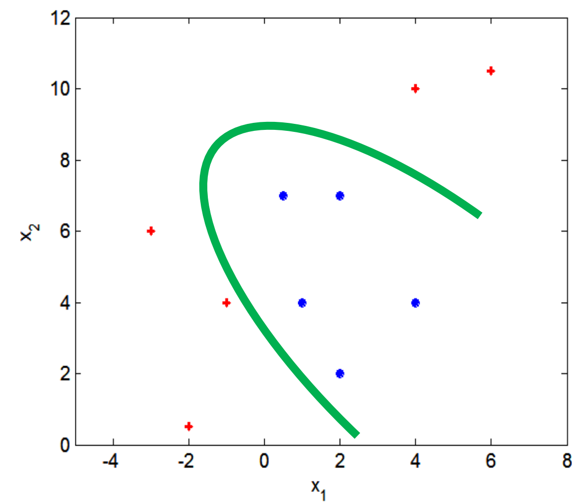
Linear separable



Linear nonseparable

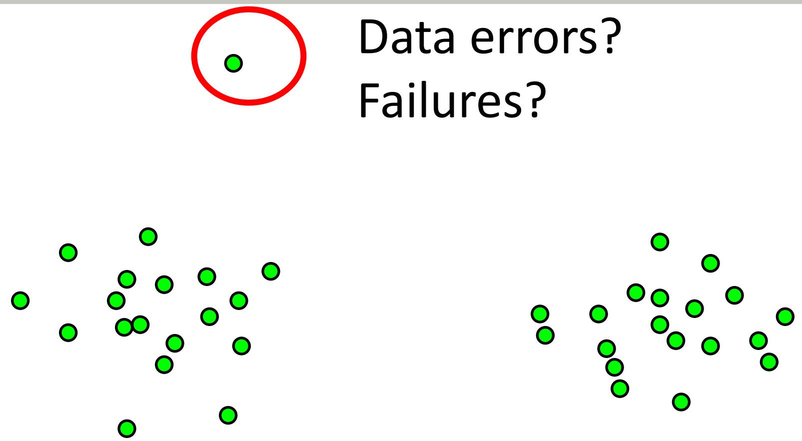


Non linear



- Kernel functions

Outlier detection



Data errors?
Failures?

- Goal: find objects that are considerably different from most other objects or unusual or in some way inconsistent with other objects
- Outliers / anomalous objects / exceptions
- Anomaly detection/ Outlier detection / Exception mining
- It is used either as a
  – Standalone task (anomalies are the focus)
  – Preprocessing task (to improve data quality)

Applications

- Fraud detection
  - Purchasing behavior of a credit card owner usually changes when the card is stolen
  - Abnormal buying patterns can characterize credit card abuse
- Medicine
  - Unusual symptoms or test results may indicate potential health problems of a patient
  - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
- Public health
  - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
  - Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc
- Sport statistics
- …

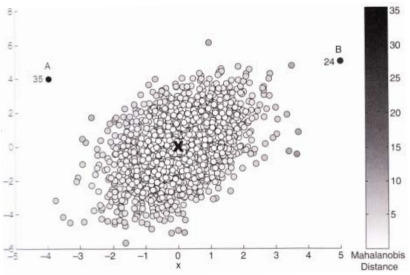
Application

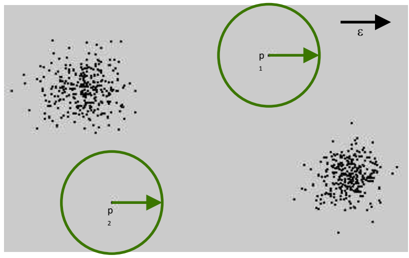
- Analysis of the SAT.1-Ran-Soccer-Database (Season 1998/99)
  - 375 players
  - Primary attributes: Name, #games, #goals, playing position (goalkeeper, defense, midfield, offense),
  - Derived attribute: Goals per game
  - Outlier analysis (playing position, #games, #goals)
- Result: Top 5 outliers

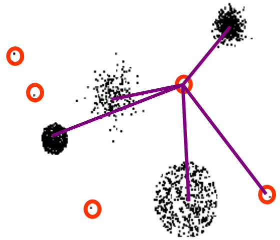
Rank	Name	# games	#goals	position	Explanation
1	Michael Preetz	34	23	Offense	Top scorer overall
2	Michael Schjönberg	15	6	Defense	Top scoring defense player
3	Hans-Jörg Butt	34	7	Goalkeeper	Goalkeeper with the most goals
4	Ulf Kirsten	31	19	Offense	2nd scorer overall
5	Giovanne Elber	21	13	Offense	High #goals/per game

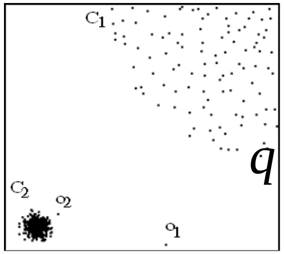
Outlier detection schemes

- General steps
  - Build a profile of the “normal” behavior (patterns or summary statistics for the overall population)
  - Use the “normal” profile to detect anomalies (Anomalies are observations whose characteristics differ significantly from the normal profile)
- Types of anomaly detection schemes
  - Model-based
    - a model is created for the data, and objects are evaluated w.r.t. how well they fit the model
  - Distance-based
    - Judge a point based on the distance(s) to its neighbors
  - Density-based
    - The relative density of a point compared to its neighbors is computed as an outlier score
  - Clustering-based
    - Objects that do not strongly belong to any cluster

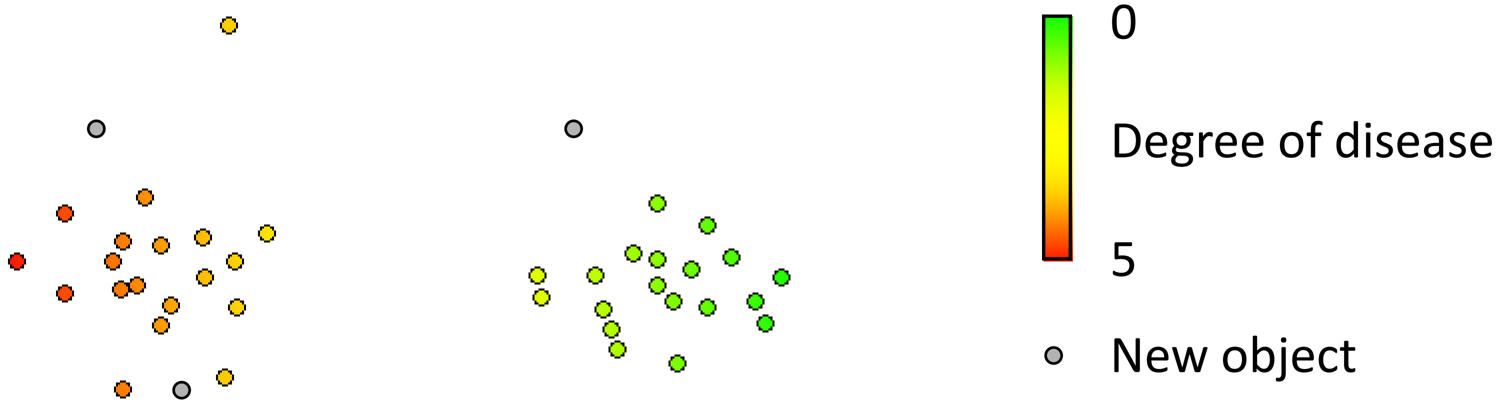








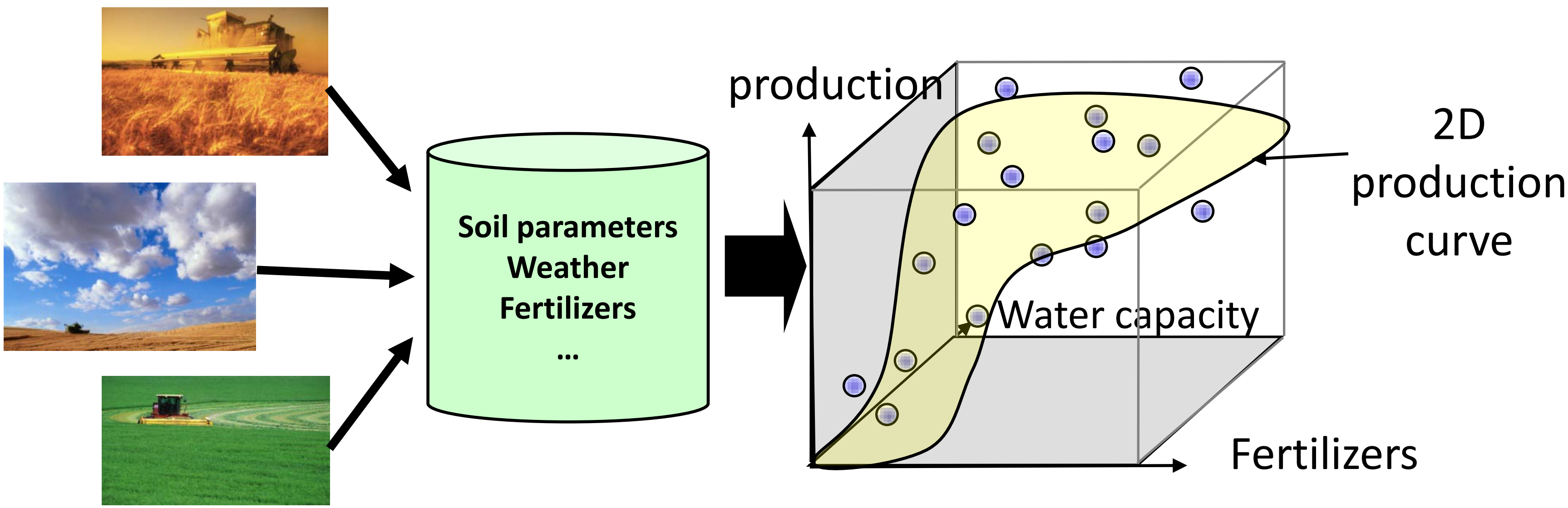
Regression



<u>Task:</u>

Similar to classification, but the feature-result to be learned is a *metric*

Application: Precision farming



- Create a production curve depending on multiple parameters like soil characteristics, weather, used fertilizers.
- Only the appropriate amount of fertilizers given the environmental settings (soil, weather) will result in maximum yield.
- Controlling the effects of over-fertilization on the environment is also important

Main DM tasks

- Its difficult to summarize KDD I in a single lecture
- For more information, refer to the KDD I course material
  - http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)_12
- For KDD II, it is important to understand the basics of KDD I
- But, necessary algorithms and techniques that will be used during KDD II will be discussed again

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks (or overview of KDD I)
- KDD II contents
- Resources
- Things you should know
- Homework/tutorial

Overview of KDD II

- Introduction

Part 1: High dimensional data

- Feature selection
- Feature reduction and distance learning
- Subspace clustering

Part 2: Structured data

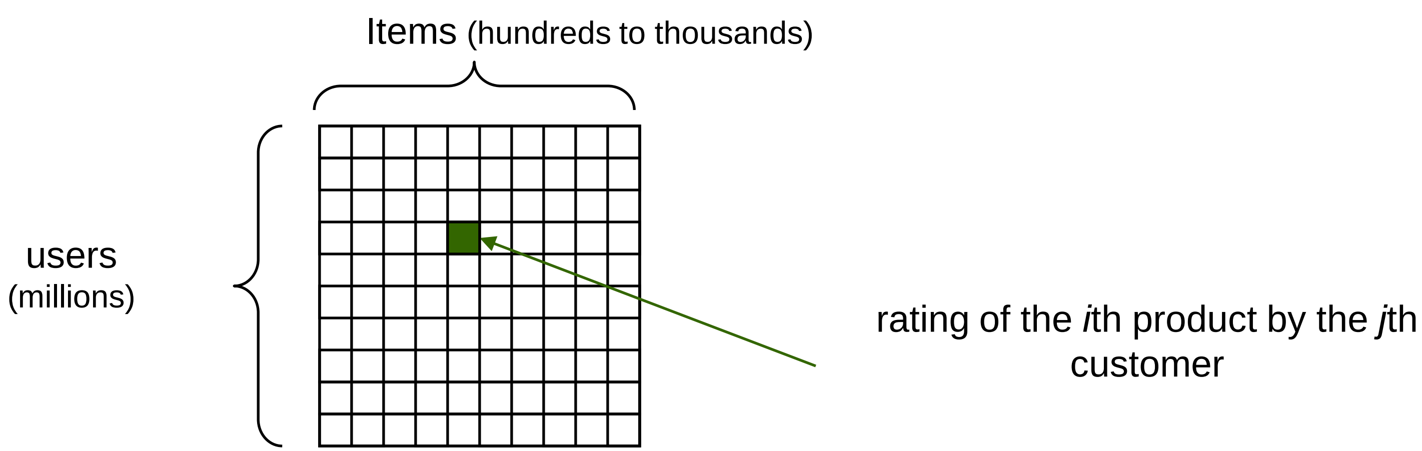
- Ensemble learning and multi-view mining
- Multi-Instance mining
- Graph mining

Part 3: Big data

- Distributed Data Mining & Privacy
- Clustering over data streams
- Classification over data streams

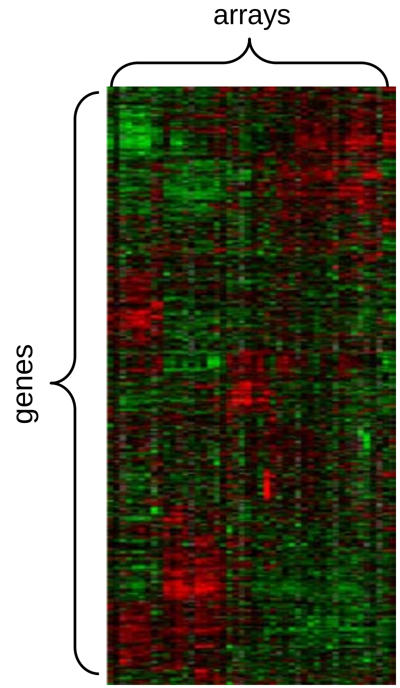
Part 1: High dimensional data - I

- Real data are high dimensional, i.e., described by many dimensions/ features
- Example: Collaborative filtering data
  - User ratings for given items (movies, videos,…)
  - Usually in the form of a data matrix



Part 1: High dimensional data - II

- Example: Micro array data
  - Measure gene expression
  - Often tens of thousands of genes (features)
  - Only tens of hundreds of samples



- Example application: Text
  - Single words (unigrams) or word combinations (n-grams) as features → huge amount of features
  - A document consists of a lot of words also
  - And, different documents are described through different words

Part 1: High dimensional data - III

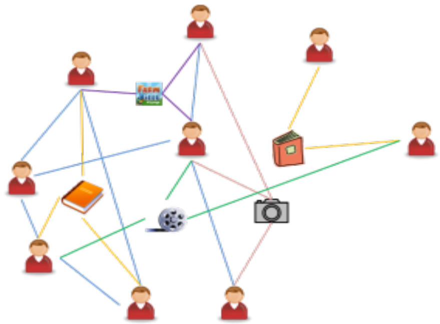
- Challenges due to high dimensionality (some of)
  - Distance functions (for clustering, outlier detection, …) loose their discriminative power in such data spaces
    - Solutions: Feature selection, Global dimensionality reduction techniques, Subspace clustering techniques
  - Different features might be relevant for different patterns (e.g., clusters)
    - Solutions: Feature selection, Subspace clustering techniques

- Our focus for this course
  - Feature selection
  - Dimensionality reduction
  - Distance learning
  - Subspace clustering

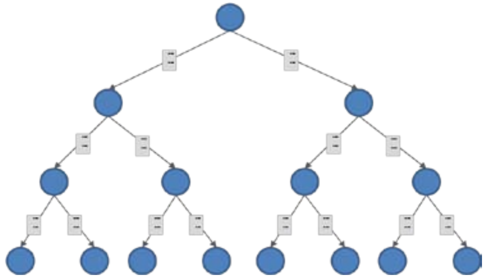
Part 2: Structured data I

- Usually, we assume that there is no relationship between the different data instances in a dataset. But, real data are far more complex

- Examples of structured data
  - Graph data: objects (nodes) are connected to each other via directed/ indirected edges
    - e.g., social networks (Twitter graph, Facebook graph), co-author network (DPLP), protein data

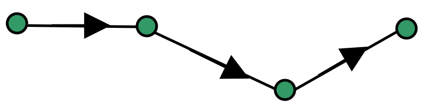


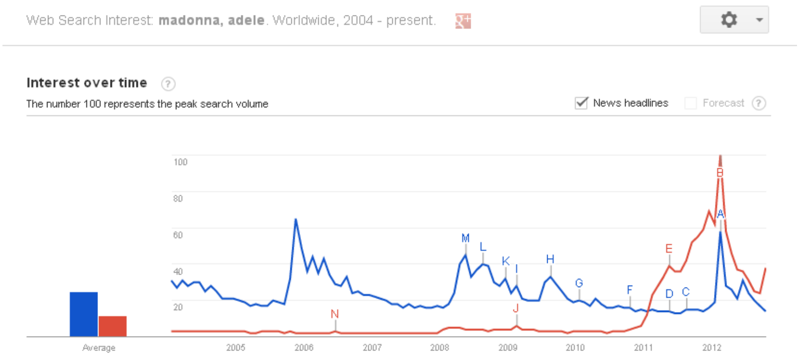
  - Tree structure data:
    - e.g., XML documents, sensor networks
    - Special cases of graphs



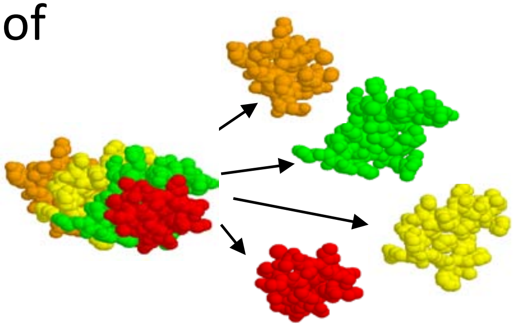
Part 2: Structured data II

- Examples of structured data (cont')
  - Sequences:
    - e.g., videos, audio, time series, trajectories



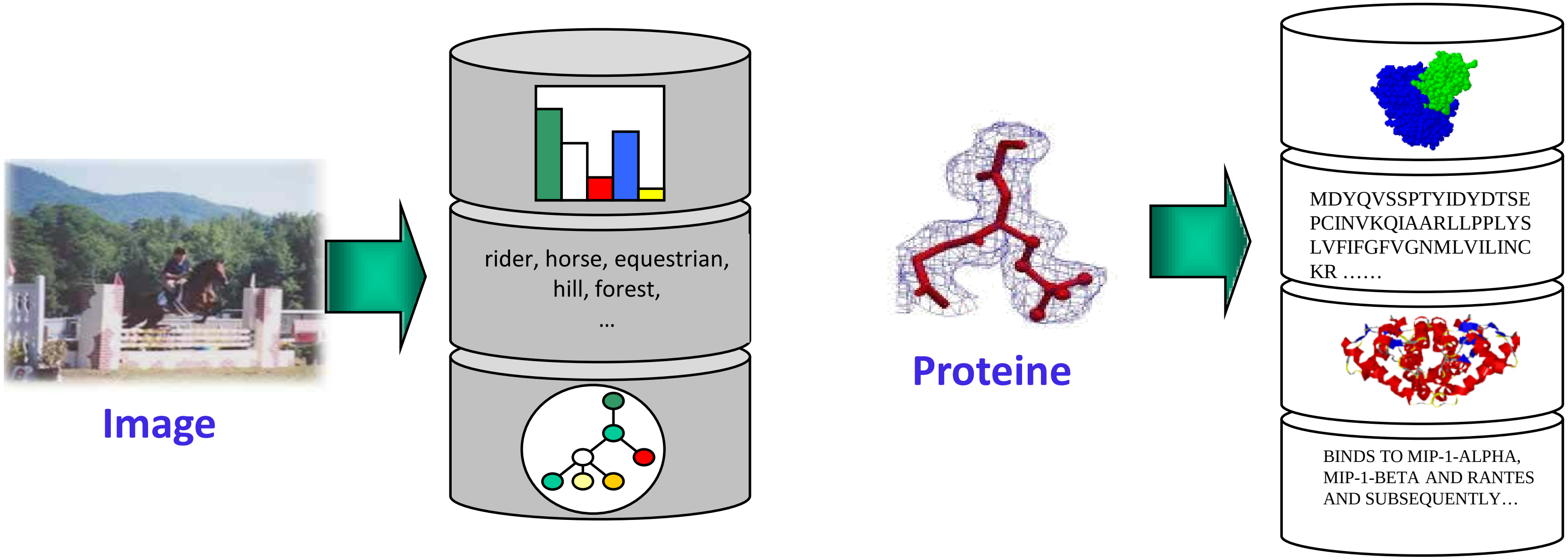


  - Multi-instance objects: each example is a set or bag of instances
    - e.g., a protein consists of amino-acids. The different amino-acids are the instances, and the protein itself is the example.



Part 2: Structured data III

- Examples of structured data (cont')
  - Multi-view objects: An object might be described by a variety of semantically different features.
    - e.g., an image might be described by a color distribution and texture description
    - e.g., proteins are characterized by an amino acid sequence, a secondary structure and a 3D representation



Part 2: Structured data III

- Challenges due to structured data (some of)
  - How to choose an effective data representation?
  - How to combine the different aspects of the data?
  - How to define patterns that encapsulate these different aspects of the data
    - Solutions: simpler descriptions, new similarity measures, new DM algorithms

- Our focus for this course
  - Multi-instance data mining
  - Multi-view data mining
  - Graph-mining
  - Link-mining

Part 3: Big data - I

- Due to the advances in hardware/software and due to the widespread usage of WWW, huge amounts of data are accumulated nowadays

- Example: Telecommunication companies
  - Call records/ sms/ mms/WWW usage/ GPS data

- Example: WWW
  - New posts/ tweets/ videos (content in general)

- Example: Facebook
  - New users/ connections between users
  - New items (e.g., videos, images) / links to these items (e.g. like, tag)

Part 3: Big data - II

- Example: Linkedin
  - New users/ companies/ pages
  - New connections between these users/ companies/ pages
  - New interactions (e.g., recommend, endorse)
- Example: environmental monitoring projects
  - Sensors spread all over the world broadcasting measurements about temperature, humidity, pollution …
- Example: scientific experiments like in CERN
  - "CERN experiments generating one petabyte of 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," he said.
    - This still means CERN is storing 25PB of data every year – the same as 1,000 years' worth of DVD quality video – which can then be analysed and interrogated by scientists looking for clues to the structure and make-up of the universe.
    - http://www.v3.co.uk/v3-uk/news/2081263/cern-experiments-generating-petabyte

Part 3: Big data - III

- Challenges due to big data (some of)
  - Finding an appropriate infrastructure w.r.t. efficiency, privacy, accuracy (a single computer might be not enough, data arriving at a rapid rate, no need to store all the details)
    - Solutions: parallel databases, distributed databases, data streams, cloud computing
  - Efficient mining issues
    - Parallel mining, privacy preserving mining, stream mining

- Our focus for this course
  - Parallel mining
  - Distributed data mining
  - Privacy preserving data mining
  - Stream mining

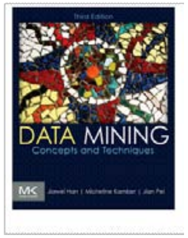
A not so funny video on privacy:
http://www.greektube.org/content/view/187432/2/

Outline

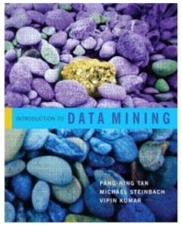
- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks (or overview of KDD I)
- KDD II contents
- Resources
- Things you should know
- Homework/tutorial

Recommended Reference Books

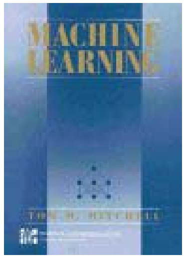
- Han J., Kamber M., Pei J. (English)
  Data Mining: Concepts and Techniques
  3rd ed., Morgan Kaufmann, 2011



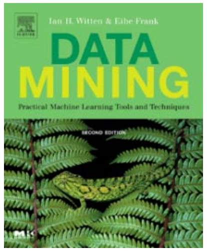
- Tan P.-N., Steinbach M., Kumar V. (English)
  Introduction to Data Mining
  Addison-Wesley, 2006



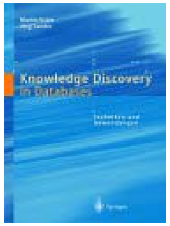
- Mitchell T. M. (English)
  Machine Learning
  McGraw-Hill, 1997



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



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



More references

- 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

Resources online

- *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 /DM 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)

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks (or overview of KDD I)
- KDD II contents
- Resources
- Things you should know
- Homework/tutorial

Things you should know!!!

- KDD definition
- KDD process
- DM step
- Supervised vs Unsupervised learning
- Main DM tasks

Homework/Tutorial

- No tutorial this week!!!
- **Homework:** Think of some real world applications of KDD.
  - What type of patterns would make sense for each application?
  - How the discovered patterns are exploited?

- **Suggested reading:**
  - U. Fayyad, et al. (1996), “From Knowledge Discovery to Data Mining: An Overview” Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press