Outline

- Why to study Data Mining?
- Why do we need Data Mining?
- What is Knowledge Discovery in Databases (KDD) and Data Mining?
- Main data mining tasks
- What's next

What is KDD

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

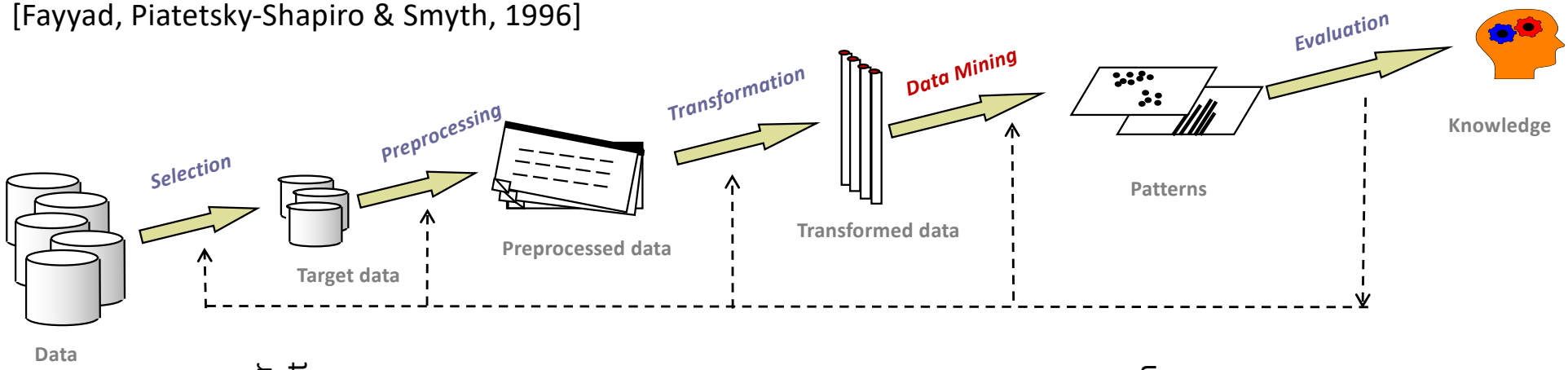
[Fayyad, Piatetsky-Shapiro, and Smyth 1996]

Remarks:

- *valid*: the discovered patterns should also hold for new, previously unseen problem instances.
- *novel*: at least to the system and preferably 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 post-processing

The KDD process and the Data Mining step

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



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

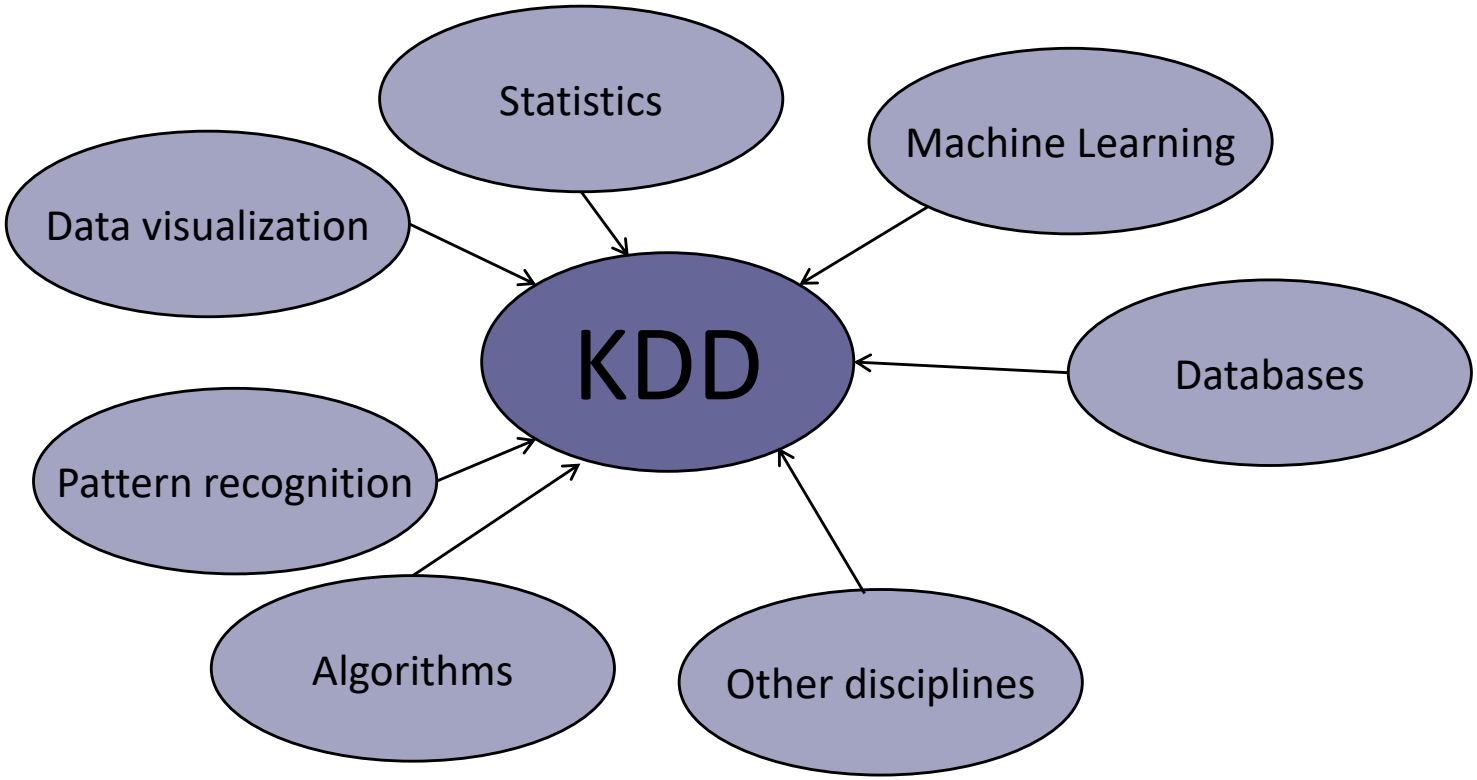
Preprocessing/Cleaning:
- Integration of data from different data sources
- Noise removal
- Missing values

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

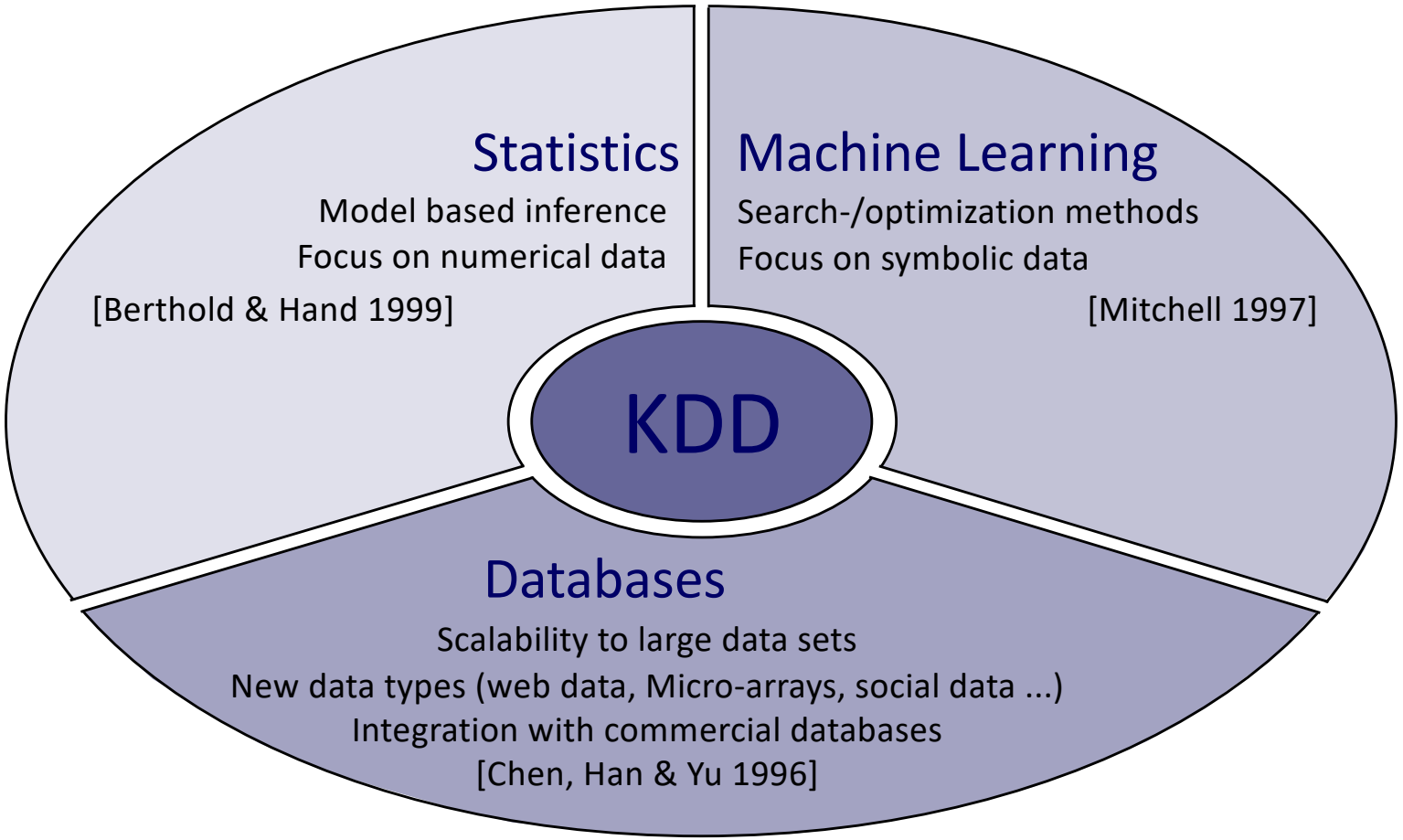
Data Mining:
- Search for patterns of interest

Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the Models
- Visualization
- Descriptive Statistics

The interdisciplinary nature of KDD 1/2



The interdisciplinary nature of KDD 2/2



Outline

- Why to study Data Mining?
- Why we need Data Mining?
- What is Knowledge Discovery in Databases (KDD) and Data Mining?
- Main data mining tasks
- What's next

Supervised vs Unsupervised learning

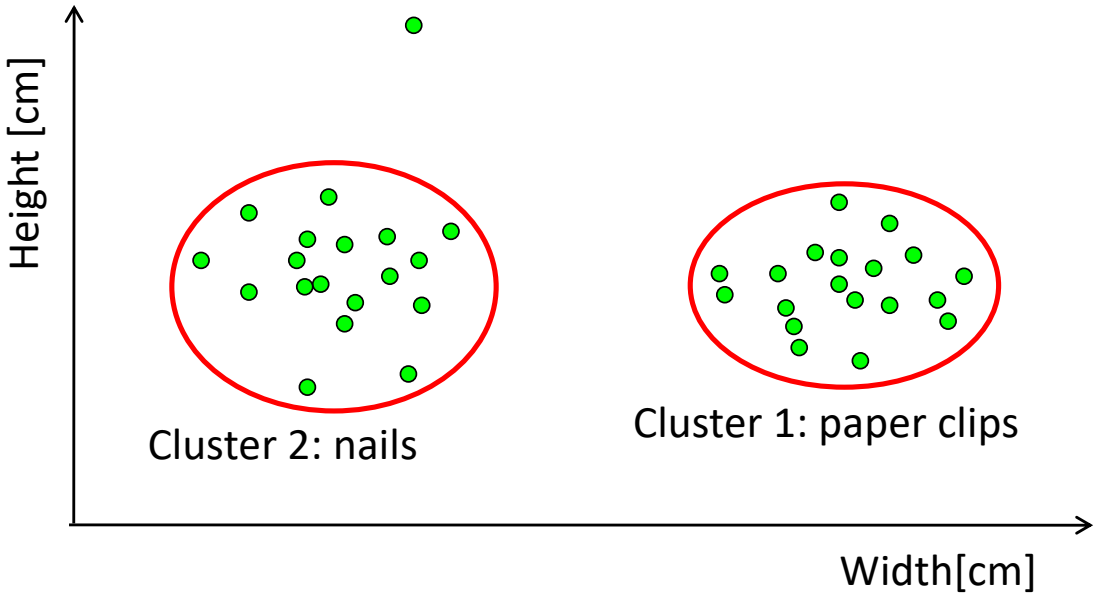
There are two different ways of learning from data:

- **Unsupervised learning/ Descriptive:**
  - 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, Outlier detection, Association rules
- **Supervised learning/ Predictive:**
  - 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
- The majority of the methods operate on the so called feature vectors, i.e., vectors of numerical features. There are numerous methods though that work on other type of data like text, sets, graphs …

Clustering definition

- Clustering can be defined as the decomposition of a set of objects into subsets of similar objects (the so called clusters)
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- The different clusters represent different classes of objects; the number of the classes and their meaning is *not known* in advance.

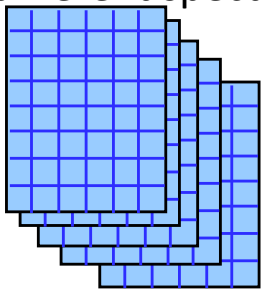
Clustering: an example

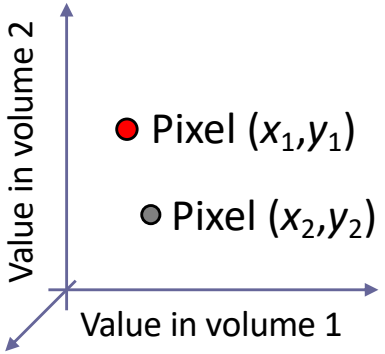


- Each point described in terms of its height and width
- No information on the actual classes (nails, paper clips) is available to the clustering algorithm.

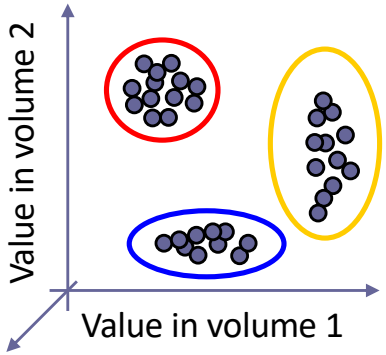
Application: Thematic maps

Recording the Earth surface
in 5 different spectra



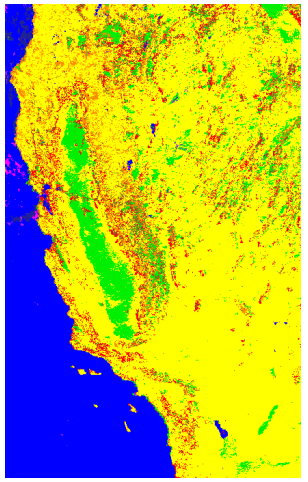


Cluster analysis



Back-transformation
in *xy*-coordinates

Color coding
Cluster membership



Clustering applications 1/2

Application: Market Segmentation

- Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
- Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
    - E.g., age, income, education, family status, ….
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering applications 2/2

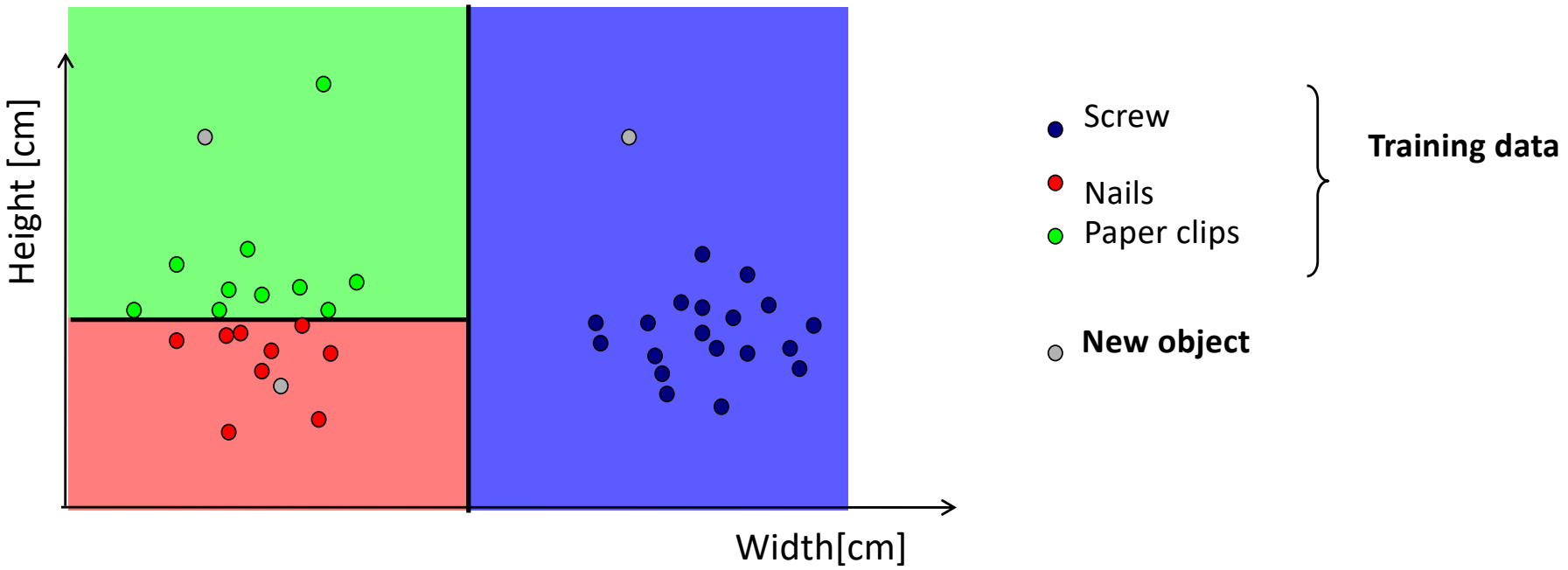
Application: Document clustering

- Find groups of documents (topics) that are similar to each other based on the important terms appearing in them.
- Approach:
  - Identify important terms in each document.
  - Form a similarity measure between documents.
  - Cluster based on the similarity measure.
- Gain:
  - Help the end user to navigate in the collection of documents (based on the extracted clusters).
  - Utilize the clusters to relate a new document or search term to clustered documents.
- Check for example, Google News.

Classification definition

- Given a collection of records (training set)
  - Each record contains a set of attributes, one of the attributes is the class attribute.
    - The class variable is nominal (categorical), e.g, {“fraud”,”normal”}, {“yes”, “no”}
- Find a model for class attribute as a function of the values of other attributes.
  - The goal is to learn from the already labeled training data, the ”rules” to classify new objects based on their attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A test set is used to determine the accuracy of the model. Usually, the given data set is divided into training and test, with training set used to build the model and test set used to validate it.

Classification: an example



- The goal is to learn a mapping from the “height, width space” to the class space (nails, screw,paper clips)
- For the new objects, the result of the classification is one of the class labels {nails, screw,paper clips}

Classification applications 1/3

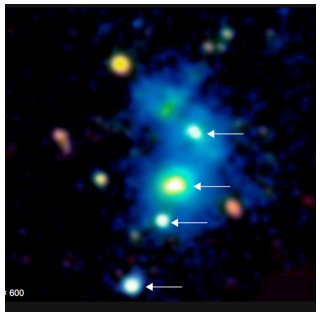
- Application: Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Classification applications 2/3

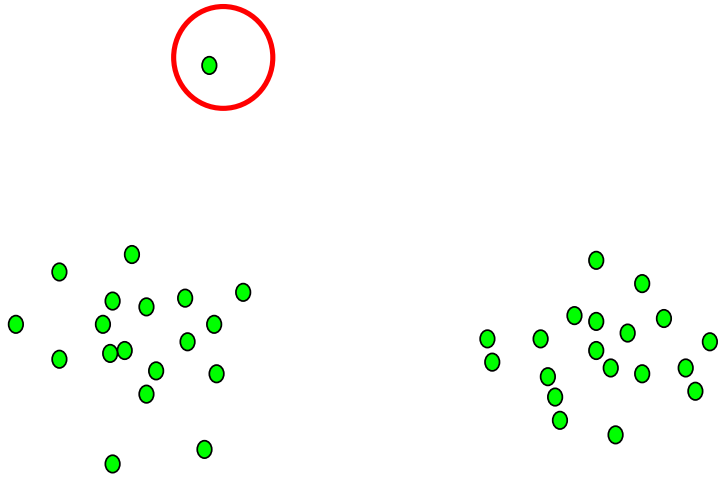
- Application: Churn prediction in telco
  - Goal: Predict whether a customer is likely to be lost to a competitor
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal (class attribute).
    - Find a model for customer loyalty
    - Use this model to predict churn and organize possible retain strategies.

Classification applications 3/3

- Application: Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!



Outlier detection



- Outlier detection is defined as identification of non-typical data
- Outliers might indicate
  - possible abuse of credit cards, mobile phones
  - data errors
  - device failures

Application

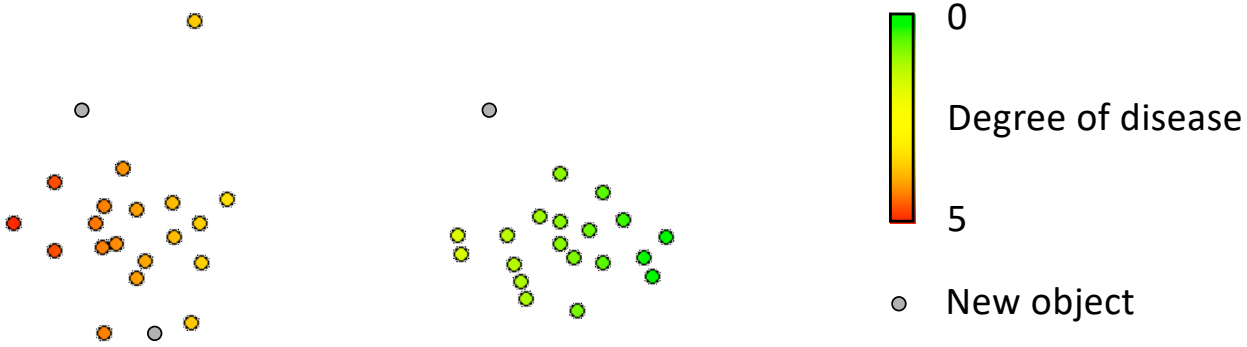
- Analysis of the SAT.1-Ran-Soccer-Database (Season 1998/99)
  - 375 players
  - Primary attributes: Name, #games, #goals, playing position (goalkeeper, defense, midfield, offense),
  - Derived attribute: Goals per game
  - Outlier analysis (playing position, #games, #goals)
- Result: Top 5 outliers

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

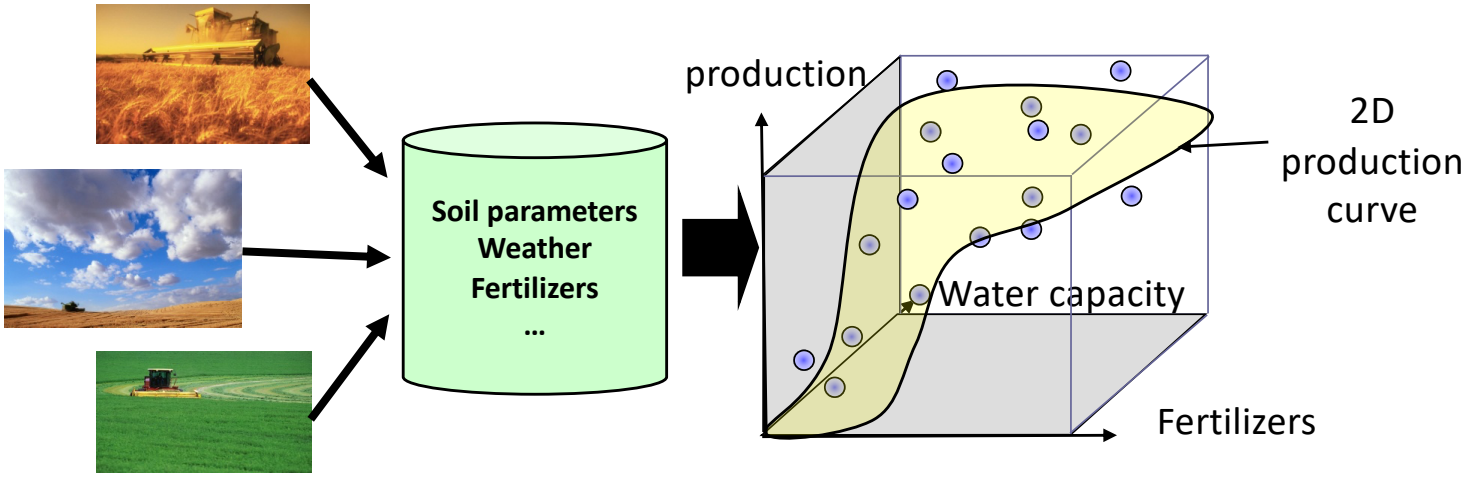
Note: “Outliers” is not necessarily a negative term.

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
  - Similar to classification, but the feature-result to be learned is continuous

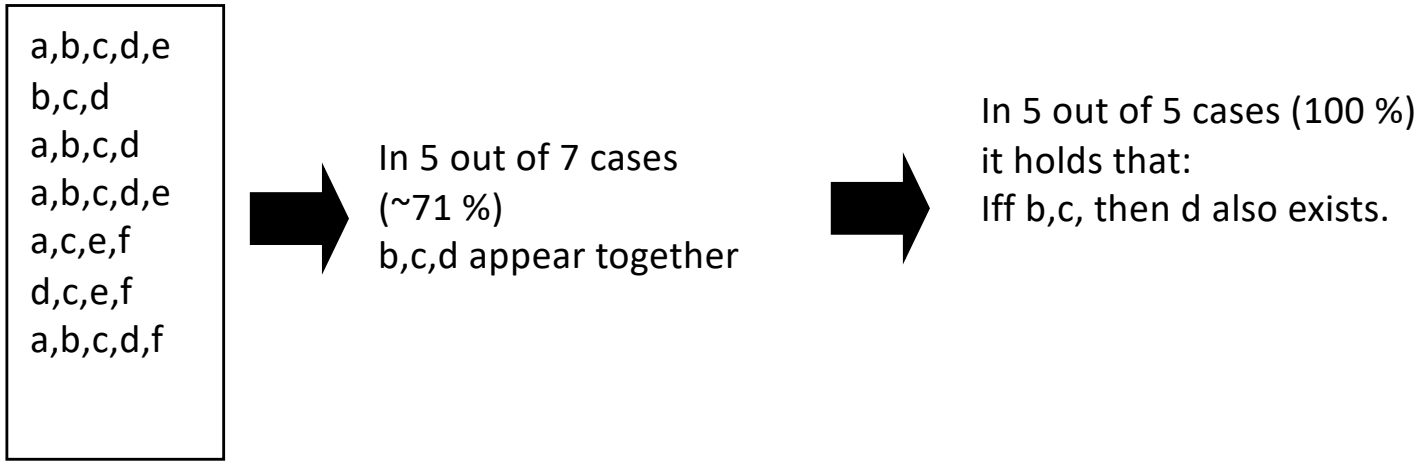


Application: Precision farming



- 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

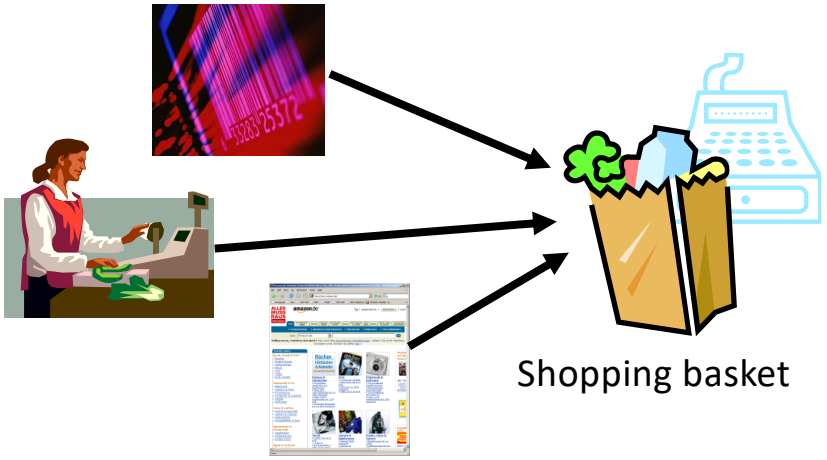
Association rules



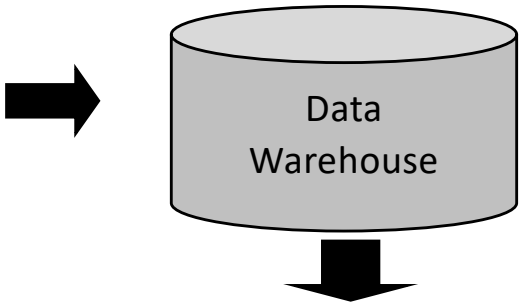
- Task: Find all rules in the database, in the following form:

  If *x, y, z* are contained in a set M, then *t* is also contained in M with a probability of at least X%.

Application: Market basket analysis



Possible generalizations:
- Paprika-Chips ➡Snacks
- Enrichment of customer data



- Result:
  - Frequently purchased items together may be better to be positioned close to each other: E.g. since diapers are often purchased together with beers => Place beer in the way from diapers to the checkout
  - Generate recommendations for customers with similar baskets:
    => e.g. Customers that bought „Star Wars“, might be also interested in „The lord of the rings “.

Outline

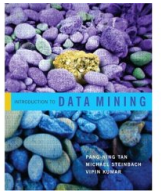
- Why to study Data Mining?
- Why we need Data Mining?
- What is Knowledge Discovery in Databases (KDD) and Data Mining?
- Main data mining tasks
- What's next

Overview of the lectures (current planning)

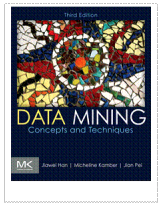
1. Introduction
2. Feature spaces
3. Association Rules
4. Classification
5. Clustering
6. Outlier Detection

Textbook and recommended readings

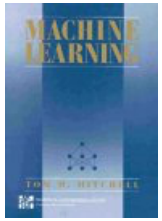
- Textbook:
  - Tan P.-N., Steinbach M., Kumar V., *Introduction to Data Mining,* Addison-Wesley, 2006



- Recommended readings
  - Han J., Kamber M., Pei J., *Data Mining: Concepts and Techniques,* 3rd ed., Morgan Kaufmann, 2011



  - Mitchell T. M., *Machine Learning,* McGraw-Hill, 1997



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

Online resources

- *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/

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)

Things you should know from this lecture

- KDD definition
- KDD process
- DM step
- Supervised vs Unsupervised learning
- Main DM tasks
  - Clustering
  - Classification
  - Regression
  - Association rules mining
  - Outlier detection

Acknowledgement

- The slides are mainly based on the Data Mining course slides of **Eirini Ntoutsi** (Associate professor at the Faculty of Electrical Engineering and Computer Science, Leibniz Universitaet Hannover) and material from the following sources:
  - KDD I lecture at LMU Munich (Johannes Aßfalg, Christian Böhm, Karsten Borgwardt, Martin Ester, Eshref Januzaj, Karin Kailing, Peer Kröger, Eirini Ntoutsi, Jörg Sander, Matthias Schubert, Arthur Zimek, Andreas Züfle)
  - Introduction to Data Mining book slides at http://www-users.cs.umn.edu/~kumar/dmbook/
  - Pedro Domingos Machine Lecture course slides at the University of Washington
  - Machine Learning book by T. Mitchel slides at http://www.cs.cmu.edu/~tom/mlbook-chapter-slides.html
  - Old Data Mining course slides at LUH by Prof. Udo Lipeck