DATABASE SYSTEMS GROUP

Ludwig-Maximilians-Universität München
Institut für Informatik
Lehr- und Forschungseinheit für Datenbanksysteme

LMU

Lecture notes

Knowledge Discovery in Databases

Summer Semester 2012

Lecture 1: Introduction

Lecture: Dr. Eirini Ntoutsi
Exercises: Erich Schubert

http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)

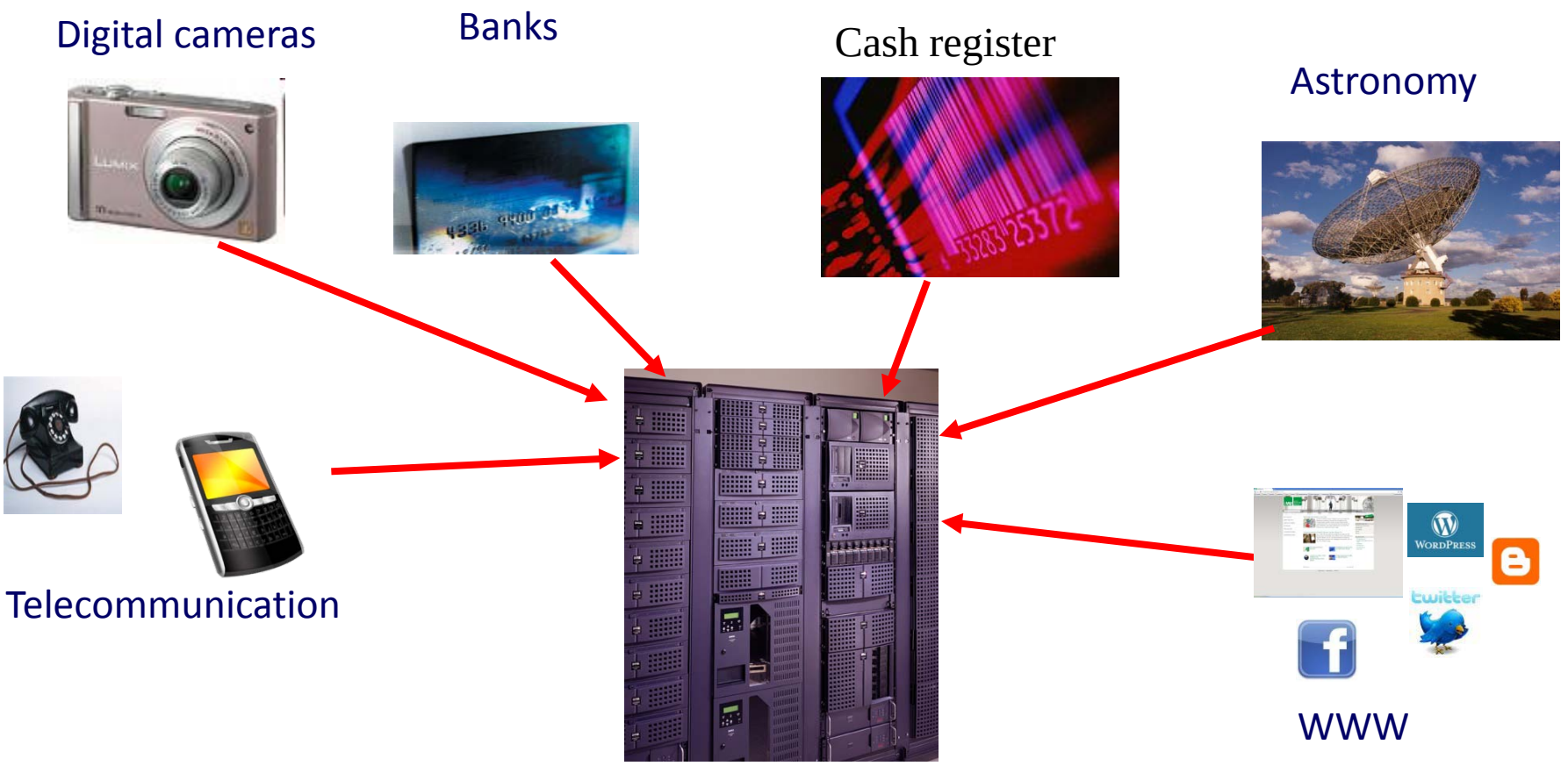
Class information

- **Class schedule**
  - Lectures: Tuesday, 9:00-12:00, Room B U 101 (Oettingenstr. 67)
  - Exercises: Thursday, 14:00-16:00, Room 057 (Oettingenstr. 67)
    16:00-18:00, Room B U 101 (Oettingenstr. 67)
  - Office hours:
    - Eirini: Thursday, 13:00-15:00, Room F 112 (Oettingenstr. 67)
    - Erich: ++++
- **Exam**
  - You must register in the following url:
    http://www.dbs.ifi.lmu.de/cms/Knowledge_Discovery_in_Databases_I_(KDD_I)
  - The exam would be based in the material discussed in the class plus the exercises. The notes are just auxiliary.
- **Grade:**
  - Final exam at the end of the term

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks
- What's next?
- Resources
- Homework/tutorial

Motivation



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

From data to knowledge

Data	Methods	Knowledge
 Call records	Outlier Detection	Detect fraud cases
 Bank transactions	Classification	Customer credibility for loan applications
 Customer transactions from supermarkets/ online stores	Association rules	Which products people tend to buy together
 Images Catalogs	Classification	What is the class of a star?

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks
- What's next?
- Resources
- Homework/tutorial

What is KDD?

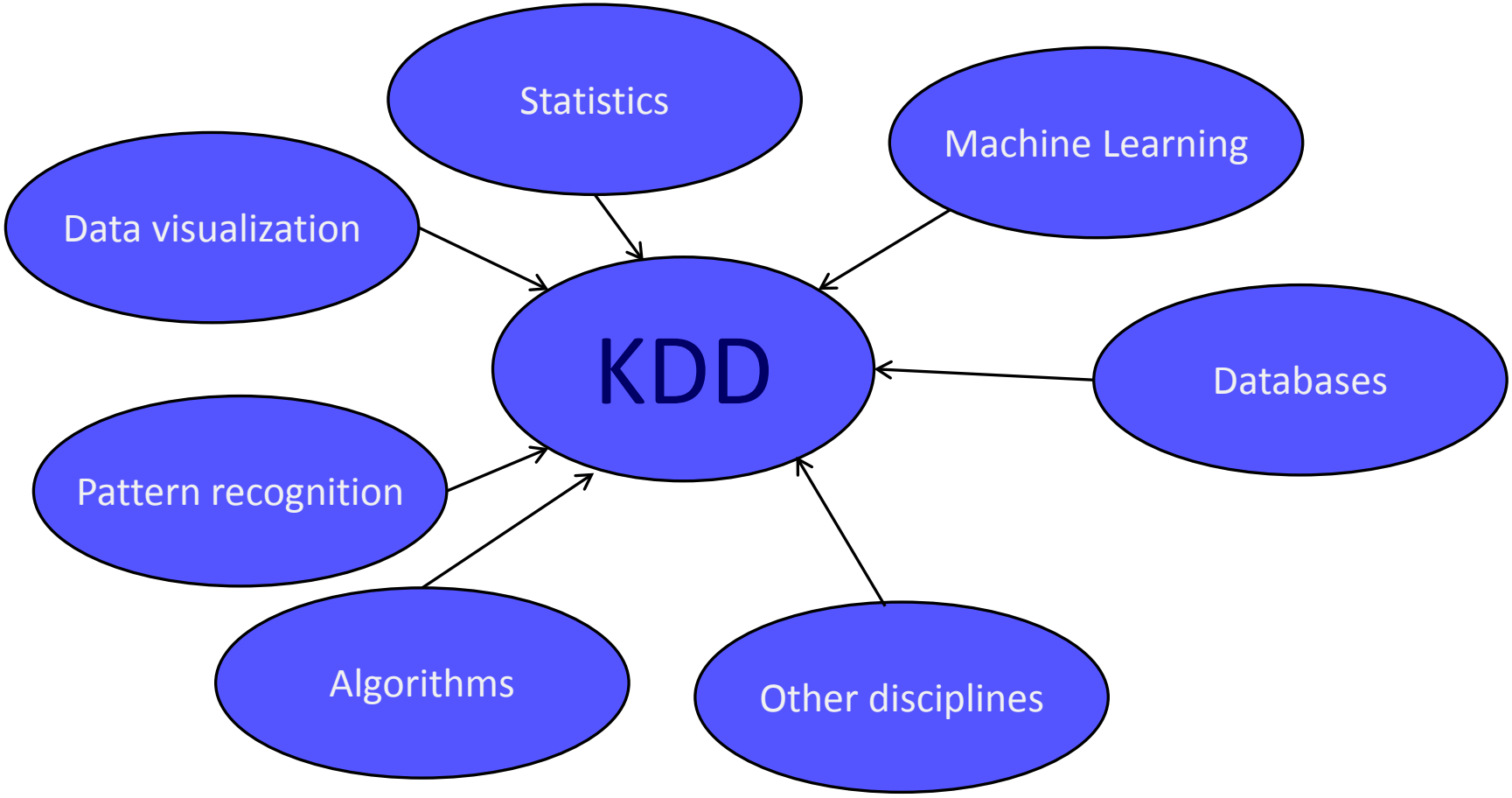
Knowledge Discovery in Databases (KDD) is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.

[Fayyad, Piatetsky-Shapiro, and Smyth 1996]

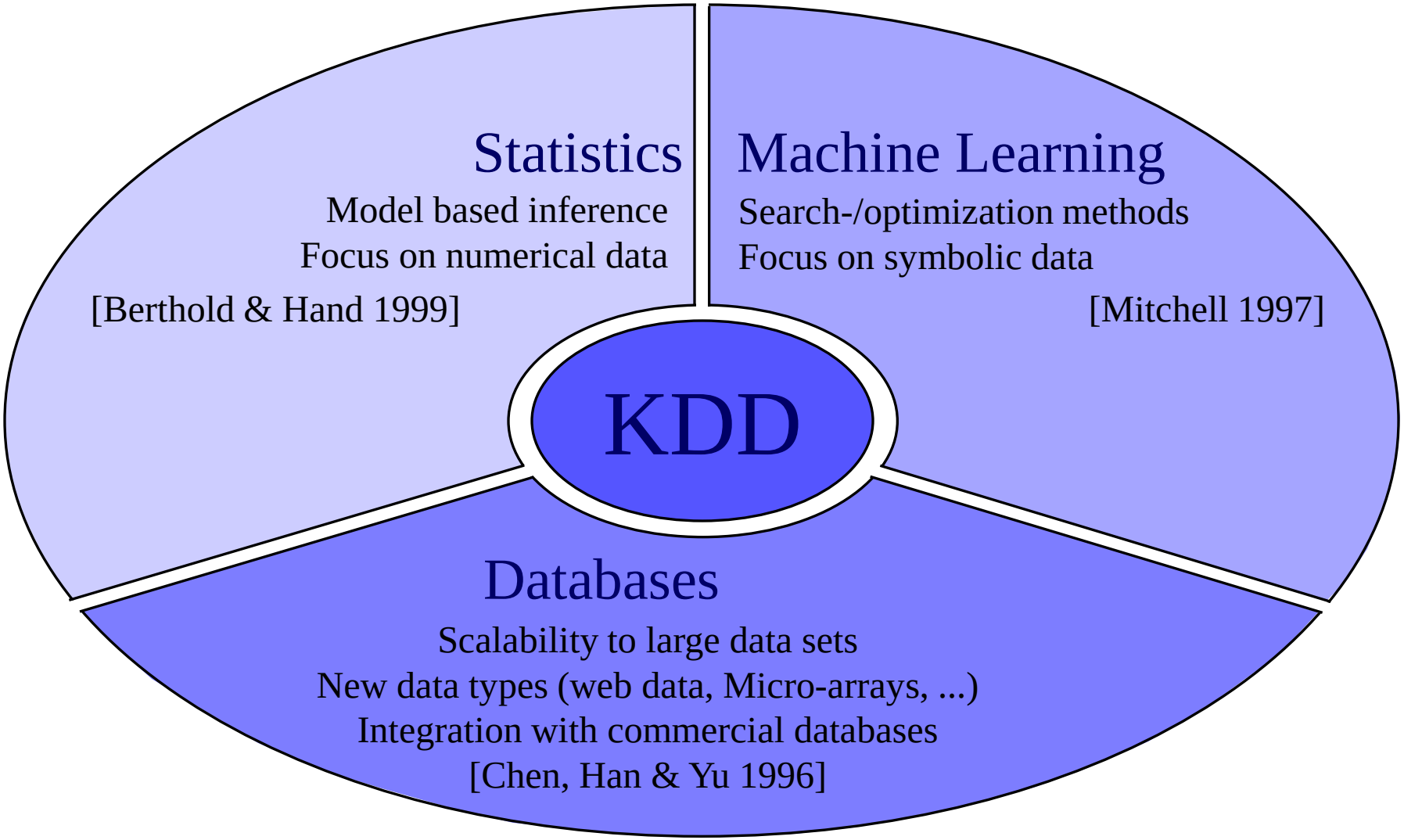
Remarks:

- *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

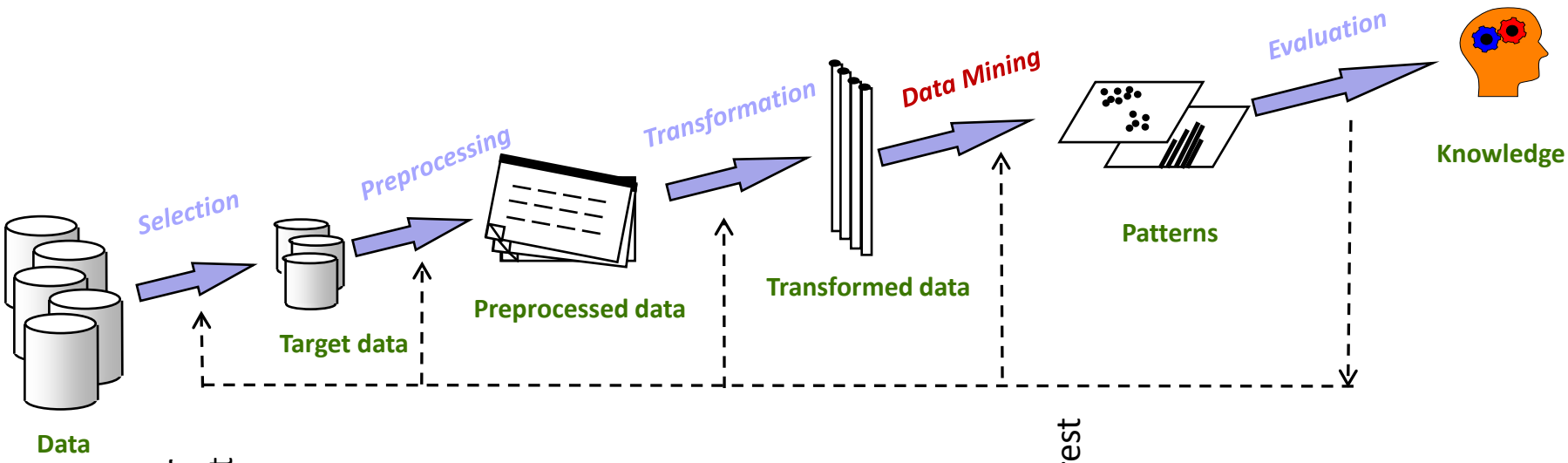


The interdisciplinary nature of KDD



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
- What's next?
- Resources
- Homework/tutorial

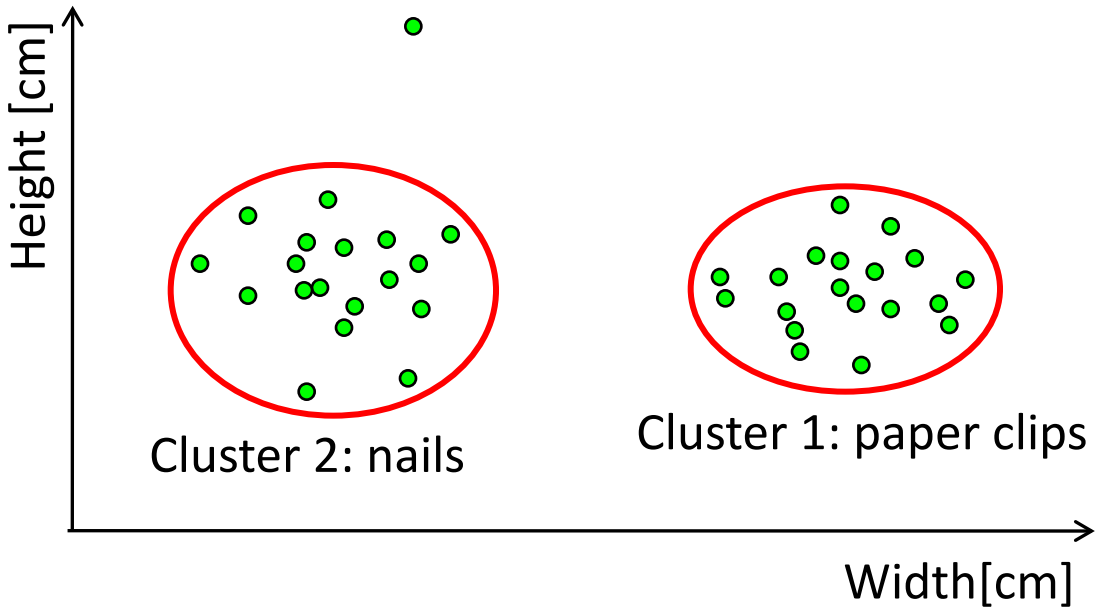
Supervised vs Unsupervised learning

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, Outlier detection, Association rules

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 …

Clustering

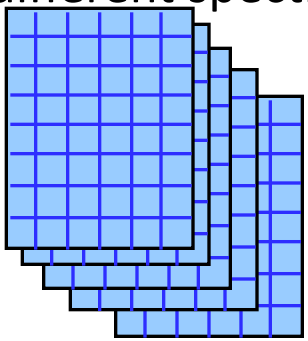


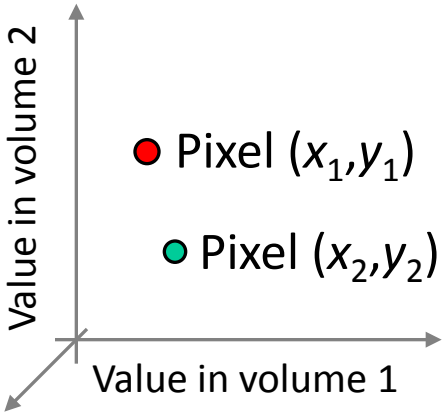
Clustering can be defined as the decomposition of a set of objects into subsets of similar objects (the so called clusters)

Ideas: The different clusters represent different classes of objects; the number of the classes and their meaning is not known in advance

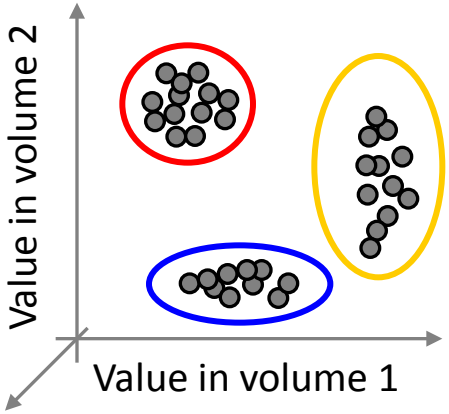
Application: Thematic maps

Recording the Earth surface in 5 different spectra



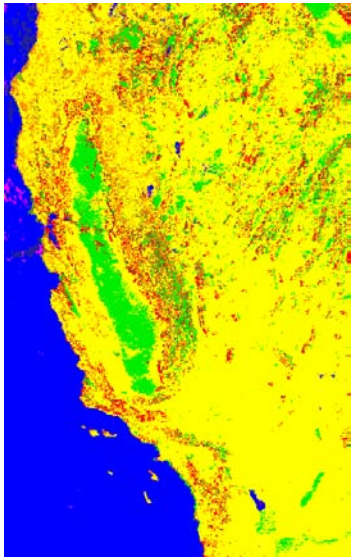


Cluster analysis

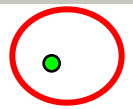


Back-transformation in *xy*-coordinates

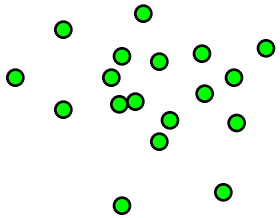
Color coding
Cluster membership

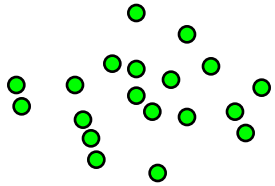


Outlier detection



Data errors?
Failures?





Outlier Detection is defined as:
Identification of non-typical data

Ideas: Outliers might indicate

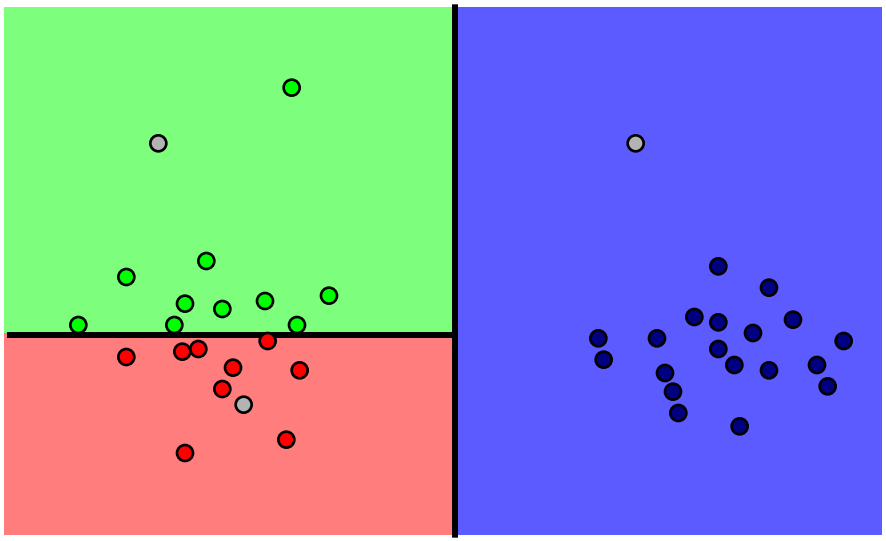
- Possible abuse of
  - Credit cards
  - Telecommunication
- Data errors

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

Classification



- Screw
- nails
- Paper clips

Training data

- New object

Task:

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

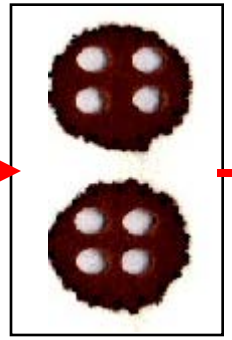
The result attribute (class variable) is nominal (categorical)

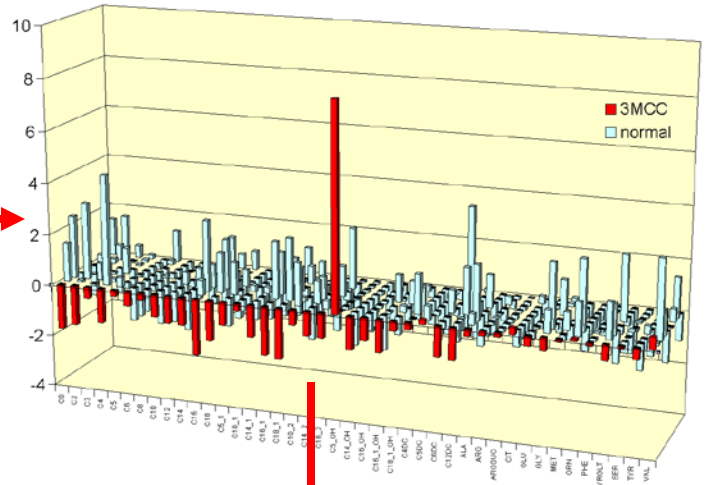
Application: Newborn screening

Blood samples of newborns

Mass spectrometry

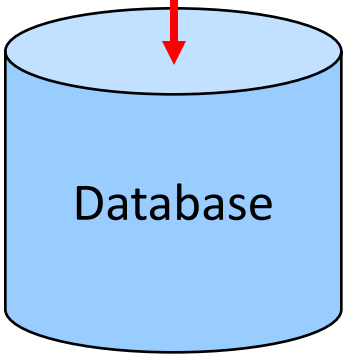
Metabolite spectrum



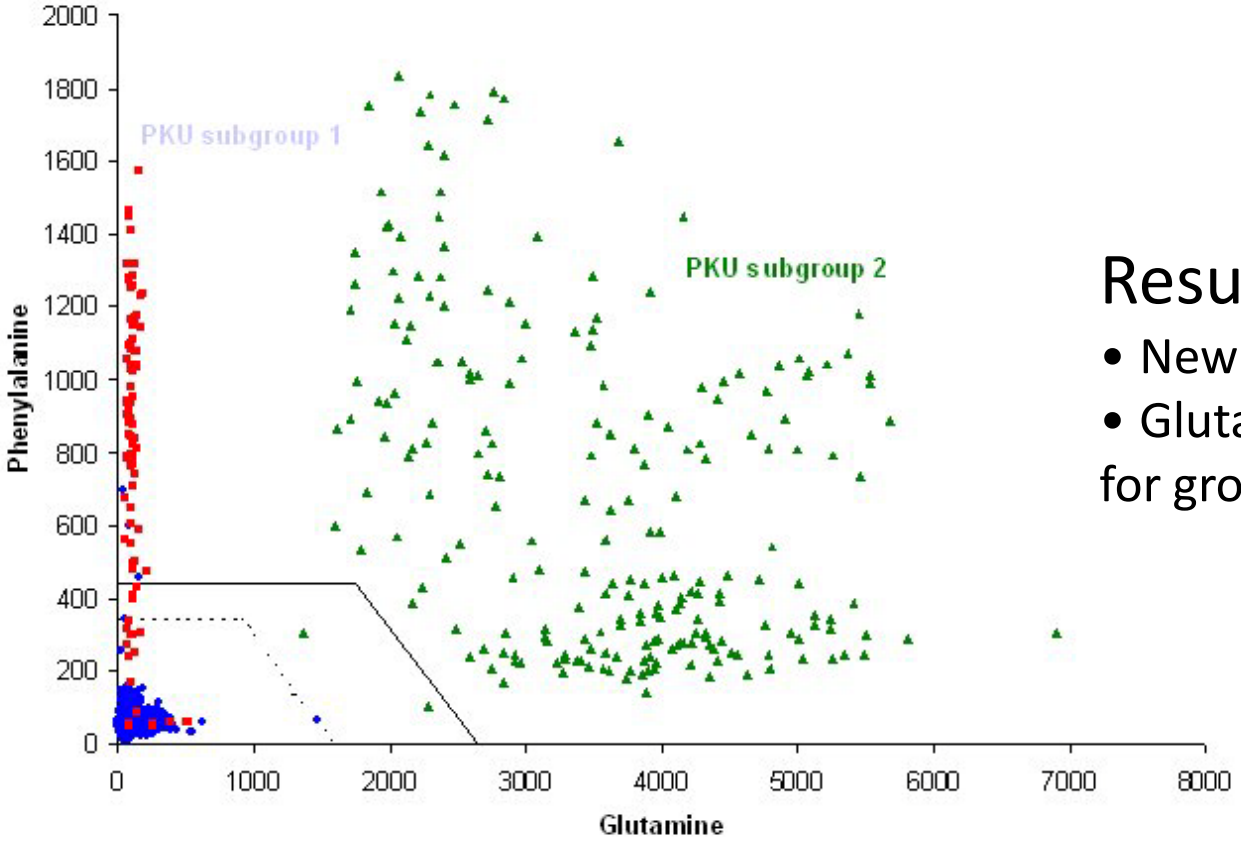


14 analysed amino acids:

alanine
arginine
argininosuccinate
citrulline
glutamate
glycine
methionine
phenylalanine
pyroglutamate
serine
tyrosine
valine
leucine+isoleucine
ornitine



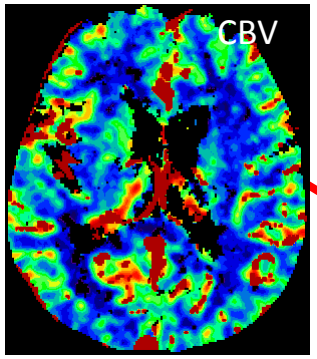
Application: Newborn screening

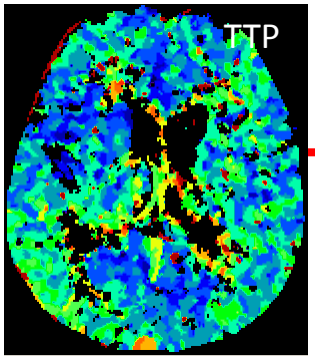


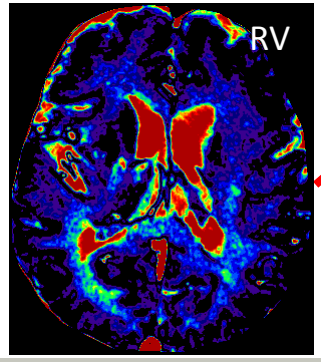
Result:

- New diagnostic tests
- Glutamine is a new marker for group differentiation

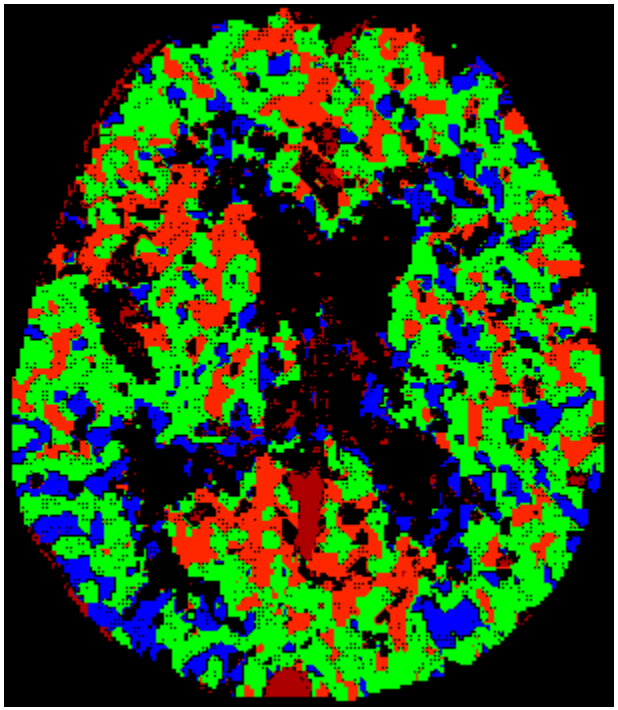
Application: Tissue classification







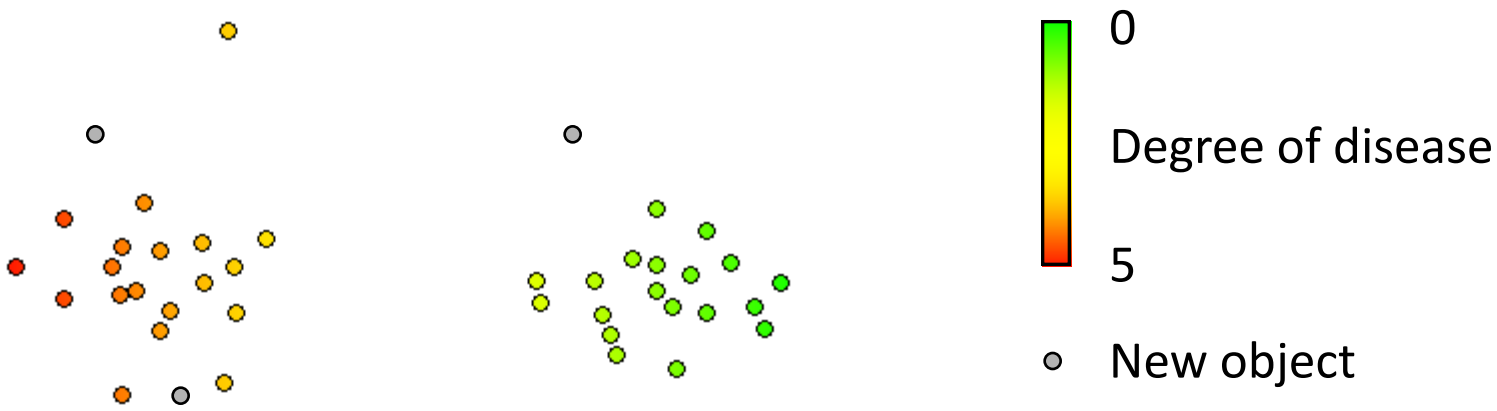
- Black: ventricle+ background
- Blue: tissue 1
- Green: tissue 2
- Red: tissue 3
- Dark red: large vessels



	Blue	Green	Red
TTP (s)	**20.5**	**18.5**	**16.5**
CBV (ml/100g)	**3.0**	**3.1**	**3.6**
CBF (ml/100g/min)	**18**	**21**	**28**
RV	**30**	**23**	**21**

Result: Classification of cerebral tissues is possible based on functional parameters obtained from dynamic CT

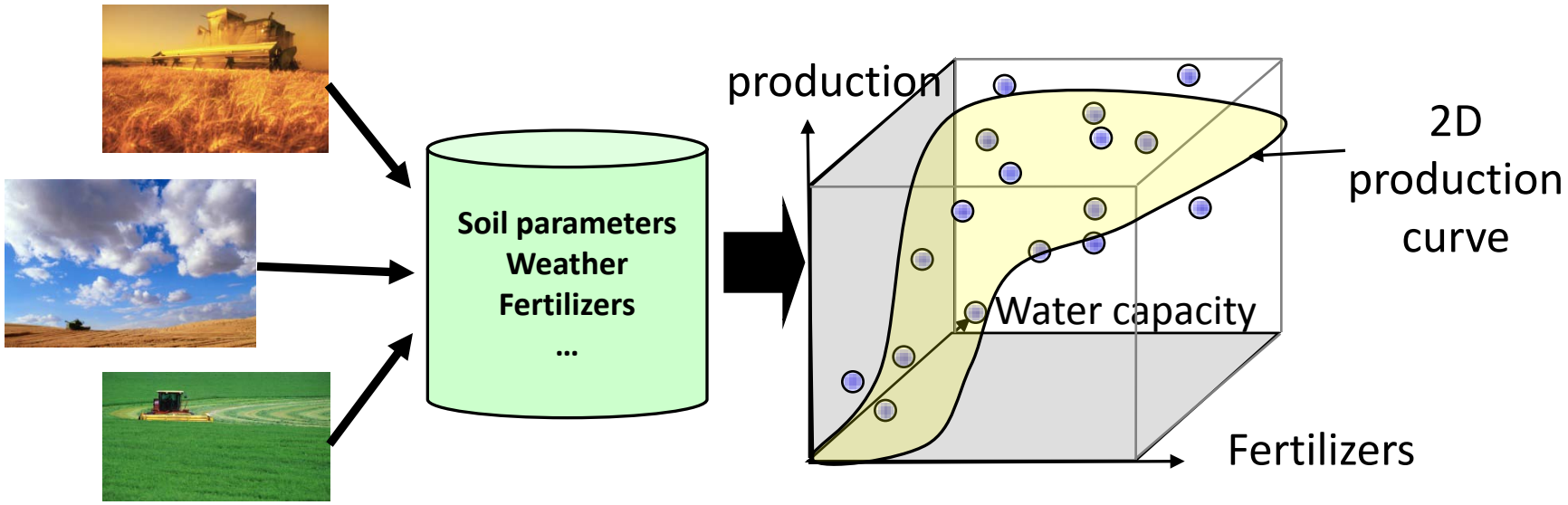
Regression



Task:

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

Association rules

a,**b,c,d**,e
b,c,d
a,**b,c,d**
a,**b,c,d**,e
a,c,e,f
d,c,e,f
a,**b,c,d**,f

In 5 out of 7 cases
(~71 %)
b,c,d appear together

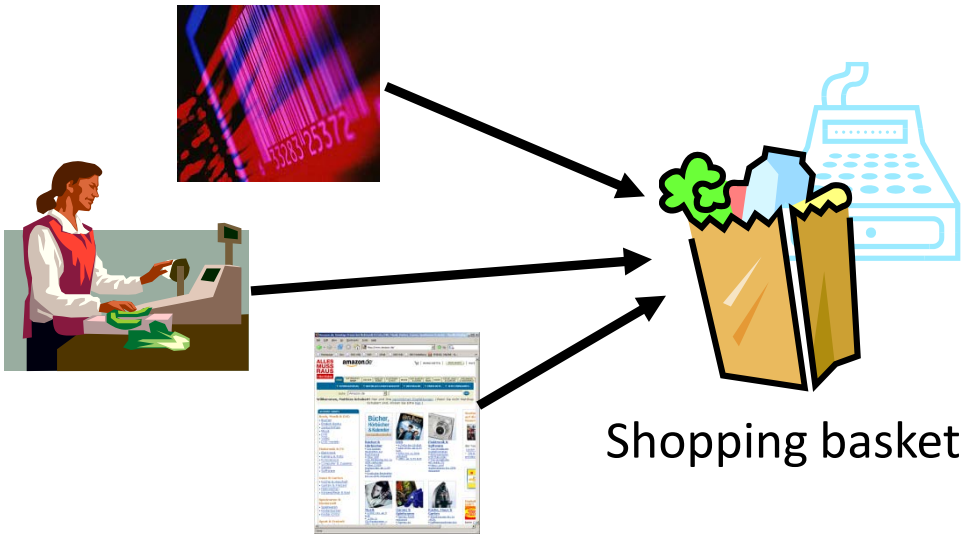
In 5 out of 5 cases (100 %)
it holds that:
Iff **b,c,** then **d** also exists.

Task:

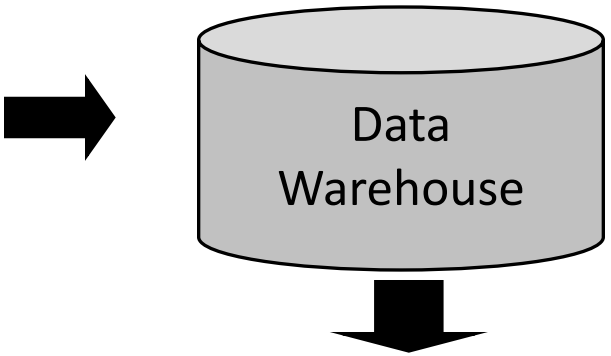
Find all rules in the database, in the following form:

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



Association rules

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 Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks
- What's next?
- Resources
- Homework/tutorial

Overview of the lectures
(current planning)

1. Introduction
2. Feature spaces
3. Association Rules
4. Classification
5. Regression
6. Clustering
6. Outlier Detection
7. DB support for efficient DM
8. Summary and Outlook: KDD II and Machine Learning and Data Mining

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks
- What's next?
- Resources
- Homework/tutorial

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

Orange

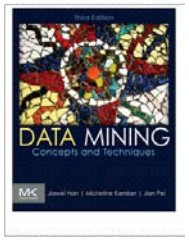
Rapid Miner (free, commercial versions)

Textbook and Recommended Reference Books

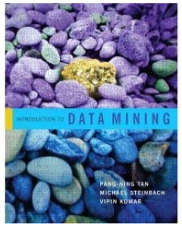
Textbook (German):

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

Recommended reference books (English):

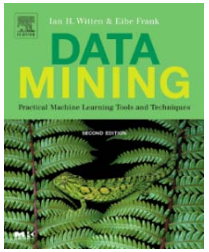


- Han J., Kamber M., Pei J.
  Data Mining: Concepts and Techniques
  3rd ed., Morgan Kaufmann, 2011



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

- 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

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/

Outline

- Why Knowledge Discovery in Databases (KDD)?
- What is KDD and Data Mining (DM)?
- Main DM tasks
- What's next?
- Resources
- Homework/tutorial

Things you should know!!!

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

Homework/Tutorial

- No tutorial this week!!!
- **Homework:** Think of some real world applications (from daily life also…) that you find suitable for KDD.
  - Why?
  - What type of patterns would you look for?

- **Suggested reading:**
  - U. Fayyad, et al. (1995), "From Knowledge Discovery to Data Mining: An Overview," Advances in Knowledge Discovery and Data Mining, U. Fayyad et al. (Eds.), AAAI/MIT Press