

# INF-KDDM: Knowledge Discovery and Data Mining

Winter Term 2019/20

## Lecture 1: Introduction

Lectures: Prof. Dr. Matthias Renz

Exercises: Christian Beth

## About the course

---

- Class schedule
  - Lectures: Monday 12:15 - 01:45 pm, CAP 3, HS 3, Tuesday 8:30 – 10:00 am, LMS2, R.Ü3.
  - Exercise Class: Wednesday 14:15 – 15:45, CAP 4 – R.715.
  - Homework Assignments: Assignments almost every week. Discussion in Exercise Class.
- Office hours:
  - Prof. Renz: Tuesdays, 1:00 - 2:00 pm, CAP 4, R708.
  - Assistants (Christian Beth) by appointment (email: [stu115337@mail.uni-kiel.de](mailto:stu115337@mail.uni-kiel.de)).
- Exam:
  - Final Exam: date/time TBA, in written form.
  - Exam will be based on the material discussed in the class with focus on the exercises.
  - No requirements for the admission of the final exam
- Grade
  - Final written exam at the end.

## Who am I?

- Diploma in Electrical Engineering

Munich University of Applied Sciences, Germany, 1997-2002.

- Diploma in Computer Science

Department of Computer Science, Ludwig-Maximilians University Munich, Germany, 2002.

- PhD in Computer Science

Department of Computer Science, Ludwig-Maximilians Universität (LMU) Munich, Germany, 2006.

- Habilitation in Computer Science

Department of Computer Science, Ludwig-Maximilians Universität (LMU) Munich, Germany, 2011.

- Acting chair of the Database Systems Group

Department of Computer Science, Ludwig-Maximilians Universität (LMU) Munich, Germany, 2015.

- Associate Professor

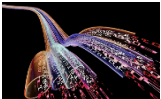
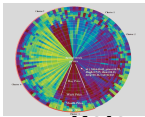



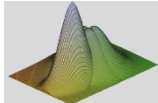
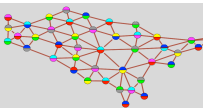

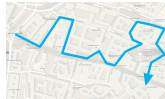
Department of Computational & Data Science, George Mason University, 01/15/2016 – 06/01/2018.

- Professor at CAU Kiel, 07/01/2018 – present.



# My research interests

## ■ My Portfolio:

<b>Applications</b> <ul style="list-style-type: none"><li>• Big Data Analytics</li><li>• Industrie4.0</li></ul> <b>Industry</b>				<ul style="list-style-type: none"><li>• eScience</li><li>• <b>Data-driven Science</b></li></ul> <b>Science</b>				<b>Environments</b> <ul style="list-style-type: none"><li>• Sensor Networks</li><li>• Embedded Systems</li><li>• Mobile Devices</li></ul> <div>HW</div> <ul style="list-style-type: none"><li>• Privacy</li><li>• Reliability</li></ul> <b>Meta</b>									
 Streams		 High Dimensional Data		 Text		 Enriched Geo		 Timeseries		 Uncertain		 Graphs		 Multimedia		 Spatiotemporal	
<b>Managing:</b> <ul style="list-style-type: none"><li>• Data Management</li><li>• Similarity Models</li><li>• Indexing</li></ul>				<b>Methods</b> <b>Searching:</b> <ul style="list-style-type: none"><li>• Similarity Search</li><li>• Query Processing</li></ul>				<b>Pattern Mining</b> <ul style="list-style-type: none"><li>• Clustering</li><li>• Outlier Detection</li><li>• Frequent Pattern Mining</li></ul>									
<b>Application-Oriented:</b> <ul style="list-style-type: none"><li>• SSDBM</li><li>• <b>DASFAA</b></li><li>• <b>SSTD</b></li><li>• <b>SIGSPATIAL</b></li></ul>				<b>Conferences (Top-10 Venues)</b> <b>Data Engineering and Mgmt.</b> <ul style="list-style-type: none"><li>• ICDE</li><li>• CIKM</li><li>• EDBT</li><li>....</li></ul>				<b>DB &amp; Data Mining Flag Ship Conf.</b> <ul style="list-style-type: none"><li>• <b>SIGMOD</b></li><li>• VLDB</li><li>• KDD</li><li>....</li></ul>									

## 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

## Why to study Data Mining – famous quotes\*

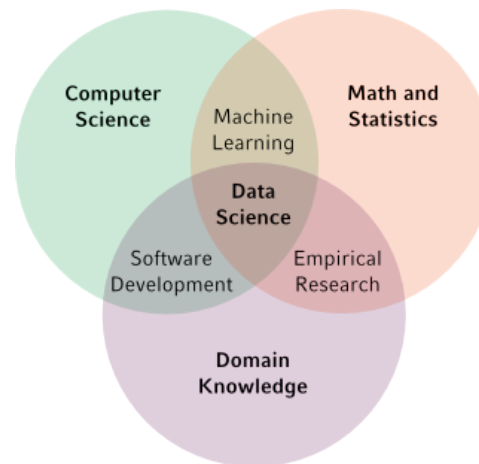
- **Data Mining** is often associated with **Machine Learning**. It is an area that has taken much of its inspiration and techniques from machine learning (and some, also, from statistics), but is put to different *ends*.
  - **Data Mining** is about using statistics as well as other programming methods to find patterns hidden in the data so that you can *explain* some phenomenon.
  - **Machine Learning** uses **Data Mining** techniques and other learning algorithms to build models of what is happening behind some data so that it can *predict* future outcomes.
  - “A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Microsoft)
  - “Machine learning is the next Internet” (Tony Tether, Director, DARPA)
  - “Machine learning is the hot new thing” (John Hennessy, President, Stanford)
  - “Machine learning is going to result in a real revolution” (Greg Papadopoulos, Former CTO, Sun)
  - “Machine learning today is one of the hottest aspects of computer science” (Steve Ballmer, CEO, Microsoft)
- \*Source: Pedro Domingos <http://courses.cs.washington.edu/courses/cse446/15sp/slides/intro.pdf>

## Why to study Data Mining - Data Scientist: The sexiest job of 21<sup>st</sup> century

*“If “sexy” means having rare qualities that are much in demand, data scientists are already there. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain. There simply aren’t a lot of people with their combination of scientific background and computational and analytical skills.”*

Source: Harvard Business Review. Data Scientist: The Sexiest Job of the 21st Century. October 2012 [link](#)

Key disciplines in Data Science:



## Data Mining – Data Science – Big Data – Machine Learning – Analytics ...

---

- New fancy words for knowledge discovery from data
  - Data mining, machine learning have been focusing on knowledge discovery from data for decades
  - Well defined set of tasks and solutions
- Big data and analytics are more business terms and ill-defined

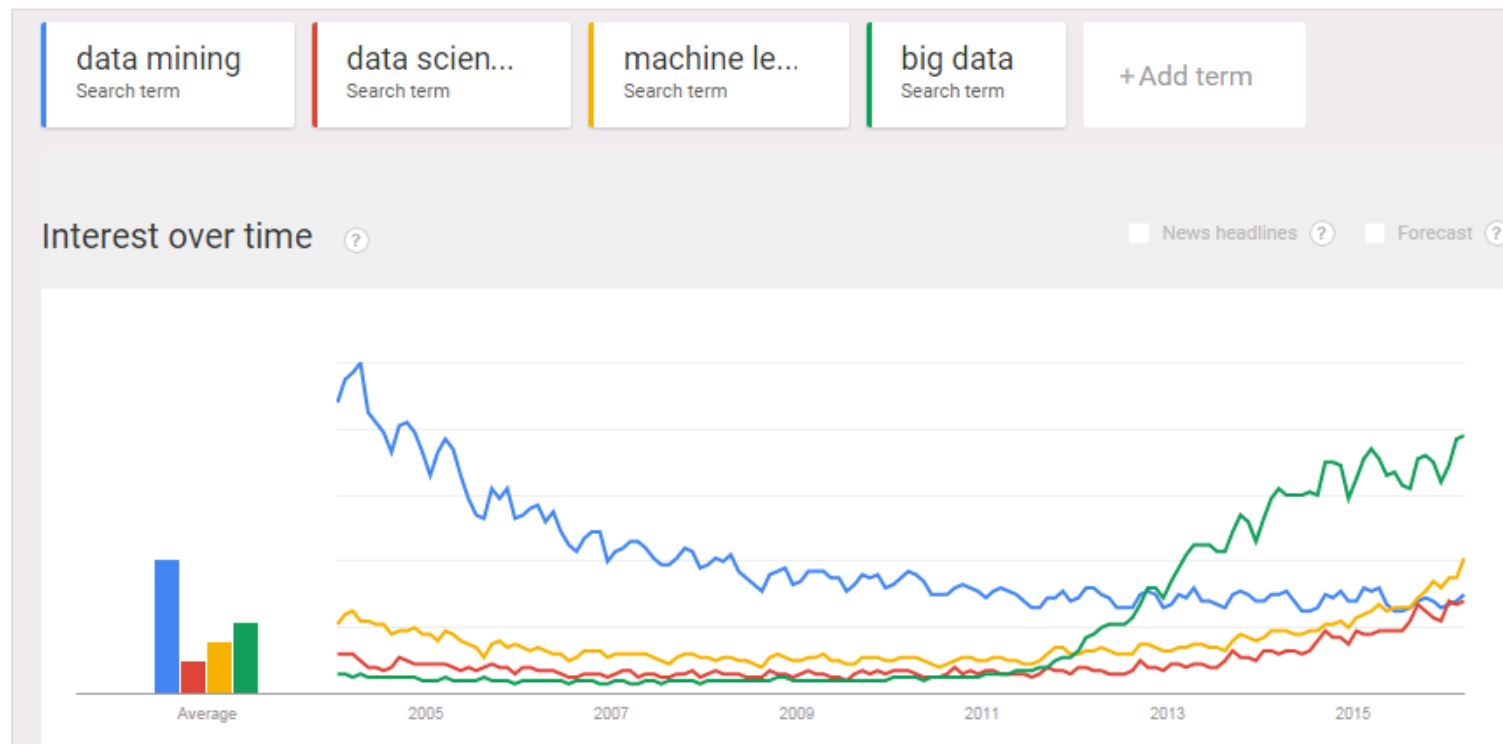
*“Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it.”*

Source: Dan Ariely, Duke University

- Though nowadays we have more data than ever and the infrastructure to deal with it
  - → more opportunities and challenges for data mining and machine learning



## Interest over time



Source: Google trends, query on 18.3.2016

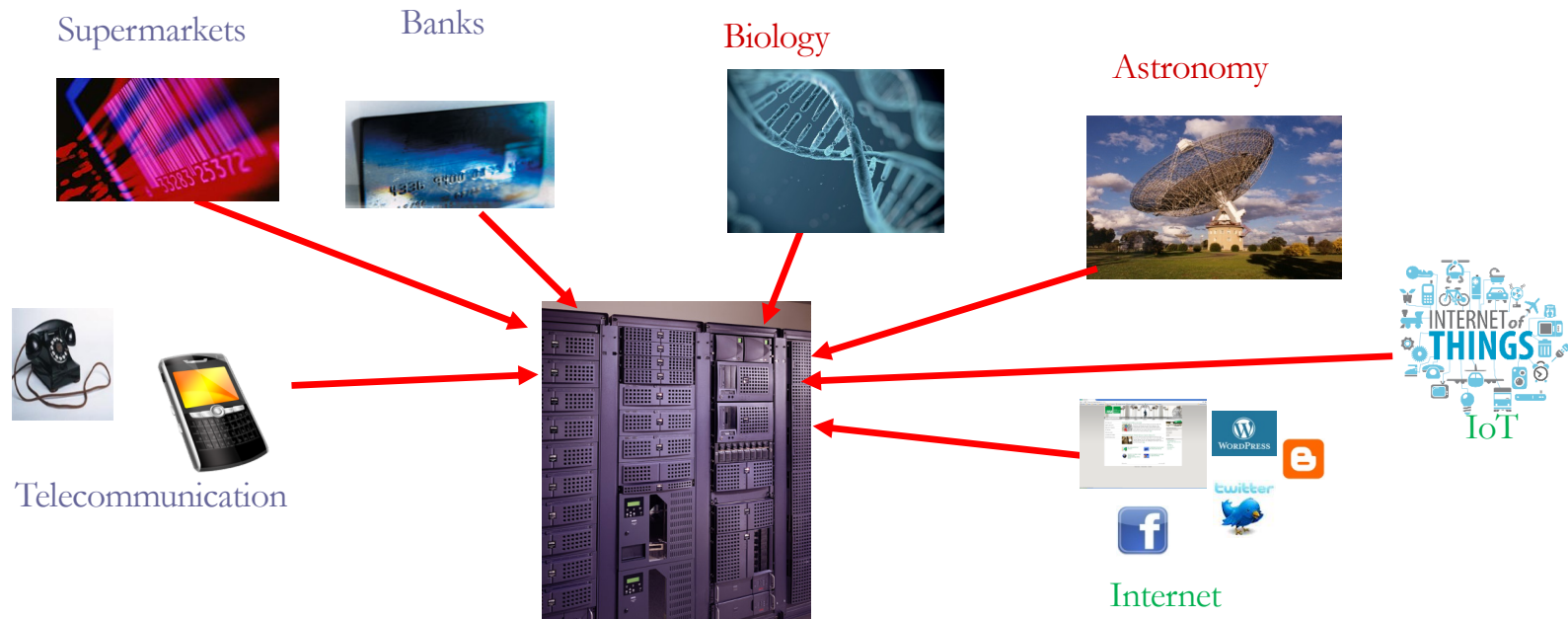
## Outline

---

- 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

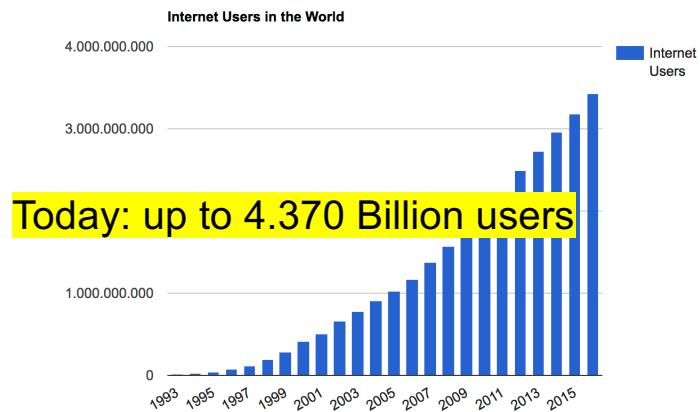
## Why we need Data Mining

- Huge amounts of data are collected nowadays from different application domains
- “We are drowning in information but starving for knowledge” John Naibett [link](#)
- The amount and the complexity of the collected data does not allow for manual analysis.



# Examples of data sources: The Internet

- 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

### Internet Users by Country (2016)

See also: [2015 Estimate](#) and [2014 Finalized](#)

#	Country	Internet Users (2016)	Penetration (% of Pop)	Population (2016)	Non-Users (internetless)	Users 1 Year Change (%)	Internet Users 1 Year Change	Population 1 Y Change
1	China	721,434,547	52.2 %	1,382,323,332	660,888,785	2.2 %	15,520,515	0.46 %
2	India	462,124,989	34.8 %	1,326,801,576	864,676,587	30.5 %	108,010,242	1.2 %
3	U.S.	286,942,362	88.5 %	324,118,787	37,176,425	1.1 %	3,229,955	0.73 %
4	Brazil	139,111,185	66.4 %	209,567,920	70,456,735	5.1 %	6,753,879	0.83 %
5	Japan	115,111,595	91.1 %	126,323,715	11,212,120	0.1 %	117,385	-0.2 %
6	Russia	102,258,256	71.3 %	143,439,832	41,181,576	0.3 %	330,067	-0.01 %
7	Nigeria	86,219,965	46.1 %	186,987,563	100,767,598	5 %	4,124,967	2.63 %
8	Germany	71,016,605	88 %	80,682,351	9,665,746	0.6 %	447,557	-0.01 %
9	U.K.	60,273,385	92.6 %	65,111,143	4,837,758	0.9 %	555,411	0.61 %
10	Mexico	58,016,997	45.1 %	128,632,004	70,615,007	2.1 %	1,182,988	1.27 %
11	France	55,860,330	86.4 %	64,668,129	8,807,799	1.4 %	758,852	0.42 %

#### Penetration Rate

Source: Internet World Stats - [www.internetworldstats.com/stats.htm](http://www.internetworldstats.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

## Examples of data sources: Internet of things

- The Internet of Things (IoT) is the network of physical objects or "things" embedded with electronics, software, sensors, and network connectivity, which enables these objects to collect and exchange data.

Source: [https://en.wikipedia.org/wiki/Internet\\_of\\_Things](https://en.wikipedia.org/wiki/Internet_of_Things)

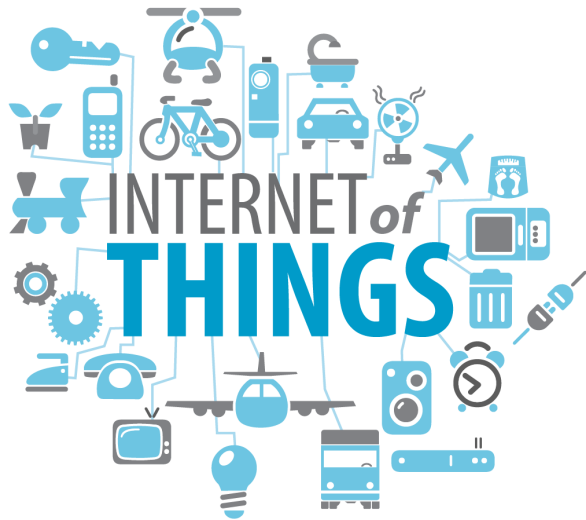


Image source: <http://tinyurl.com/prtfqxf>

During 2008, the number of things connected to the internet surpassed the number of people on earth... By 2020 there will be 50 billion ... vs 7.3 billion people (2015).

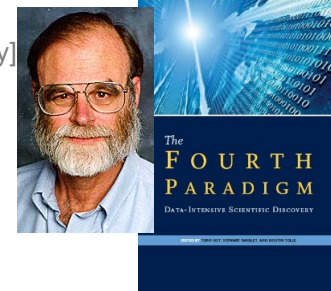
These things are everything, smartphones, tablets, refrigerators .... cattle.

Source: <http://blogs.cisco.com/diversity/the-internet-of-things-infographic>

## Examples of data sources: data intensive science

- The Fourth Paradigm:  
Age of data driven exploration  
→ Data Science
- Science Paradigms

[Comp. Science Pioneer Jim Gray]

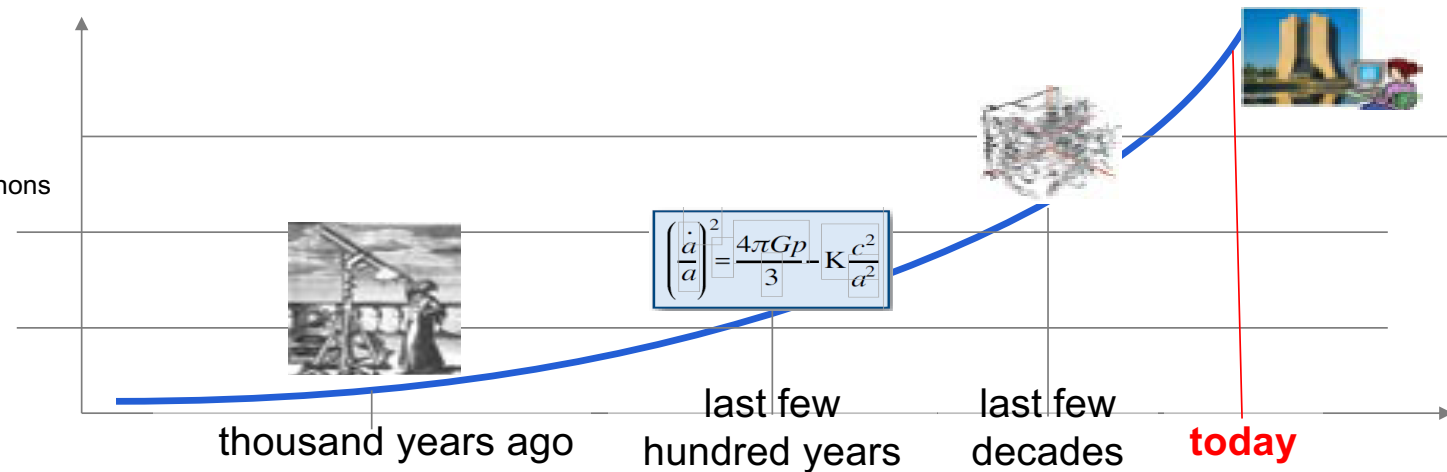


**Data driven exploration**  
→ Data Science

**Computer-driven –**  
Simulation complex phenomenons

**Theoretical –**  
Development of models

**Empirical -**  
Description of natural  
phenomenons

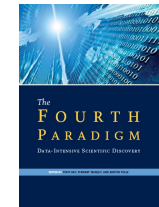


source:[http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb\\_escience.ppt](http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt)

## Examples of data sources: data intensive science

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

“*Modern science increasingly relies on integrated information technologies and computation to collect, process, and analyze complex data.*”







-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

Slide from:[http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb\\_escience.ppt](http://research.microsoft.com/en-us/um/people/gray/talks/nrc-cstb_escience.ppt)

## 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?
	Telescope images	Classification	What is the class of a star? E.g., early, intermediate or late formation



## 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

## What is KDD

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

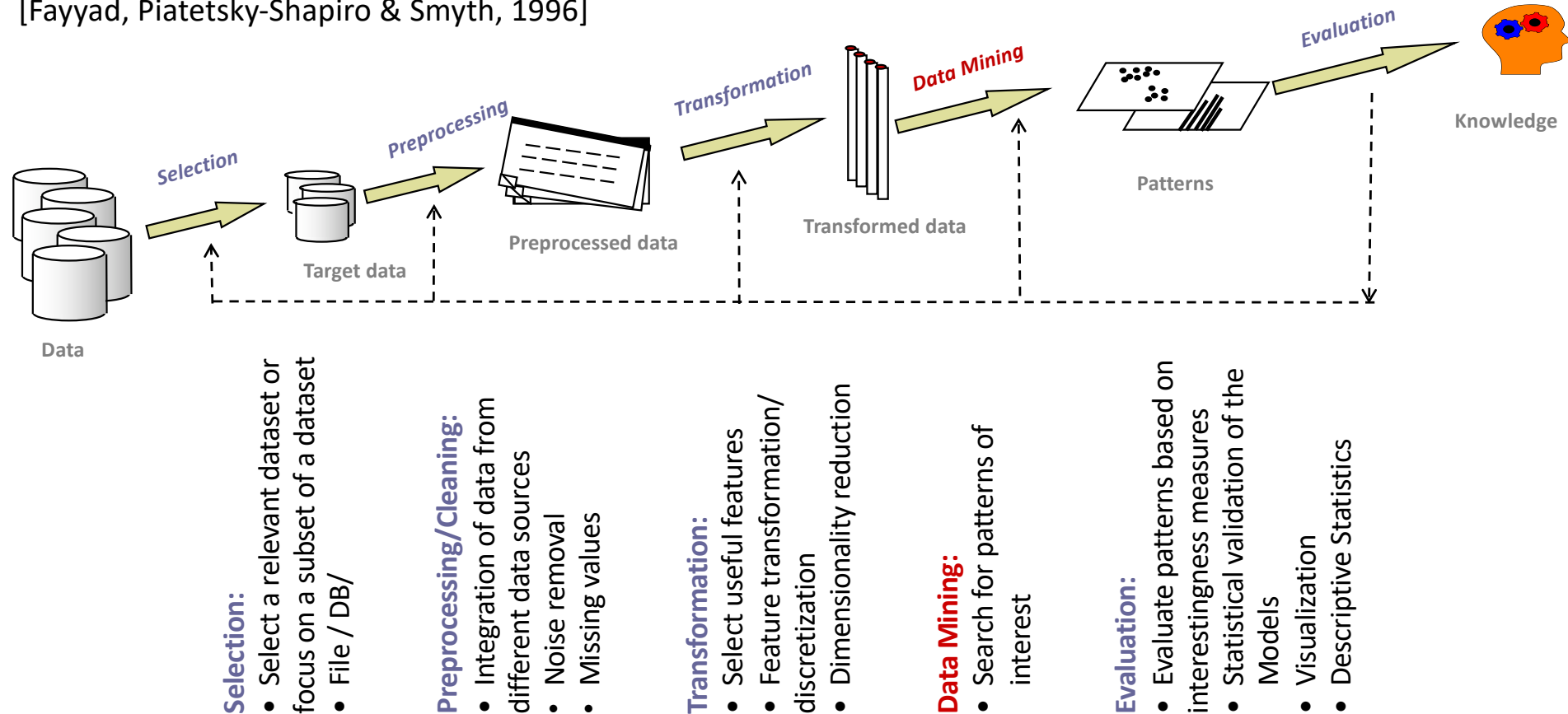
[Fayyad, Piatetsky-Shapiro, and Smyth 1996]

### Remarks:

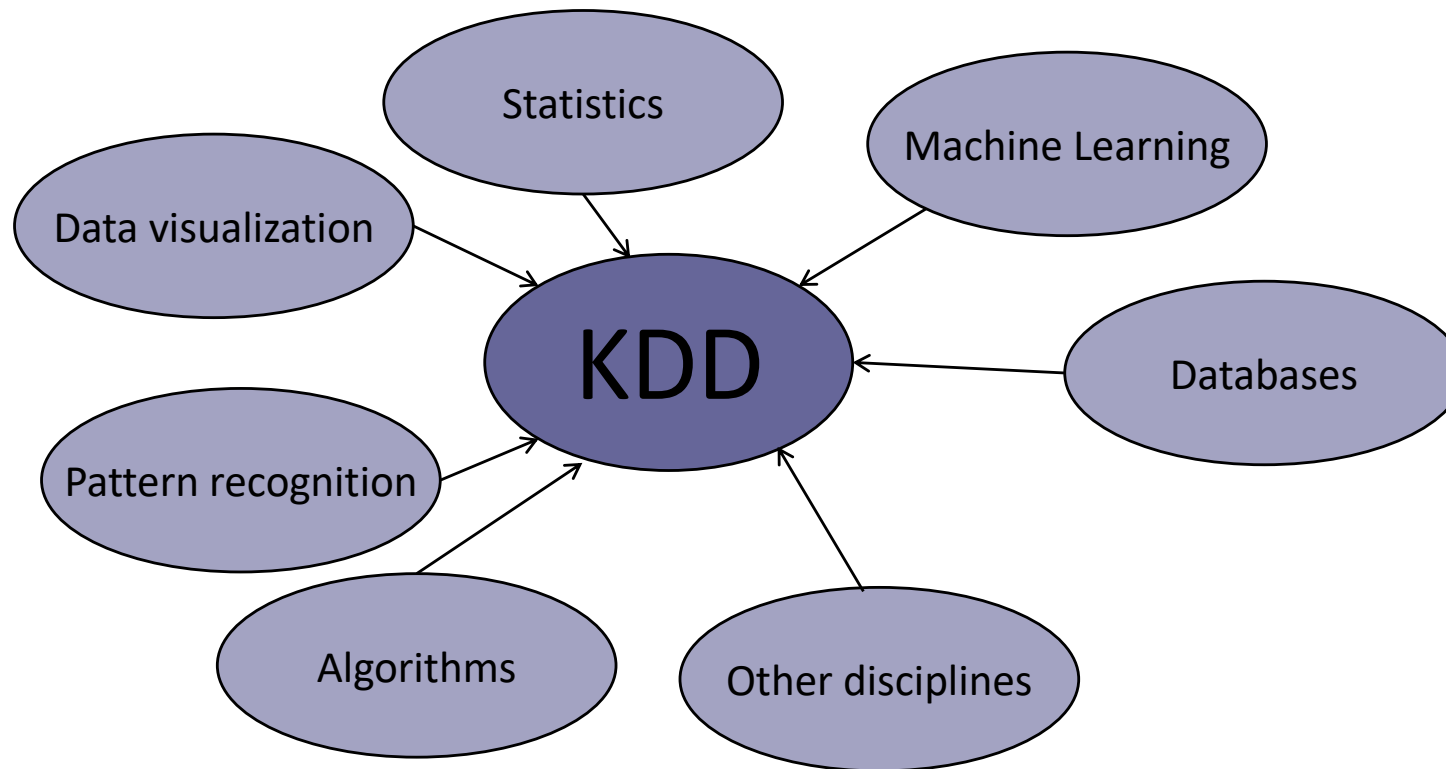
- *valid*: the discovered patterns should also hold for new, previously unseen problem instances.
- *novel*: at least to the system and preferably to the user
- *potentially useful*: they should lead to some benefit to the user or task
- *ultimately understandable*: the end user should be able to interpret the patterns either immediately or after some post-processing

# The KDD process and the Data Mining step

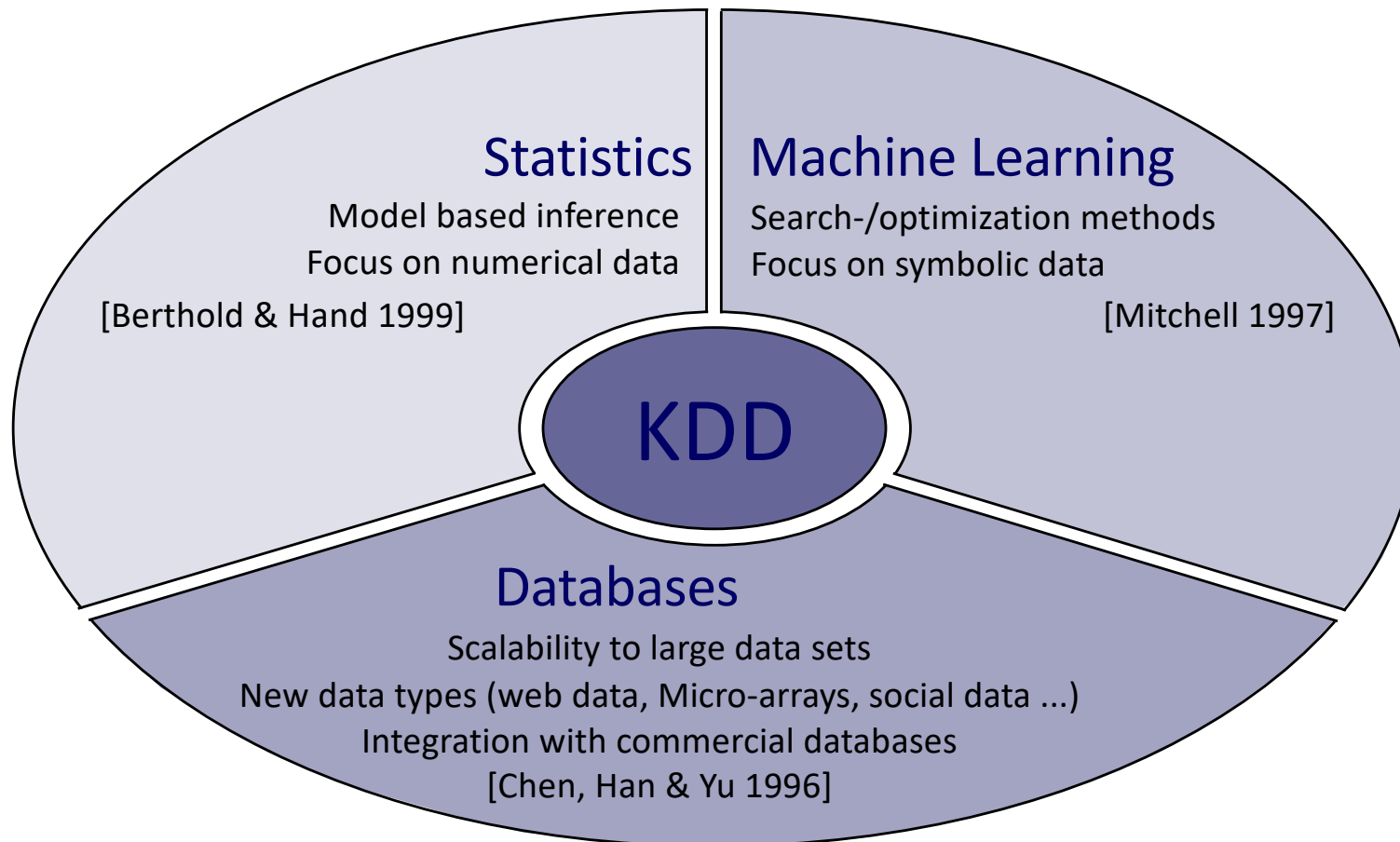
[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



## 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 ...

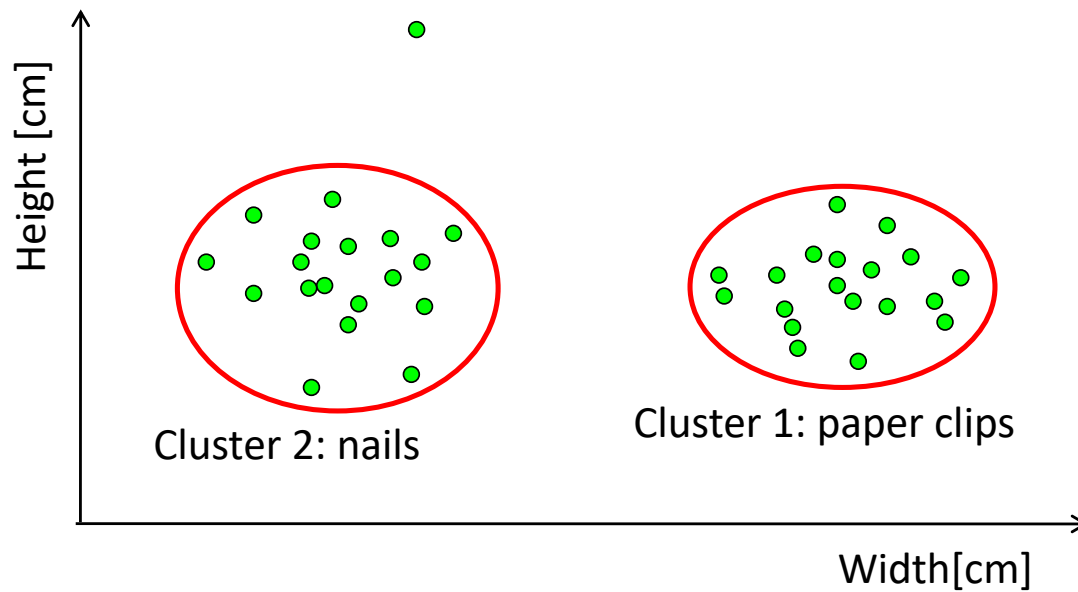
## Clustering definition

---

- 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.

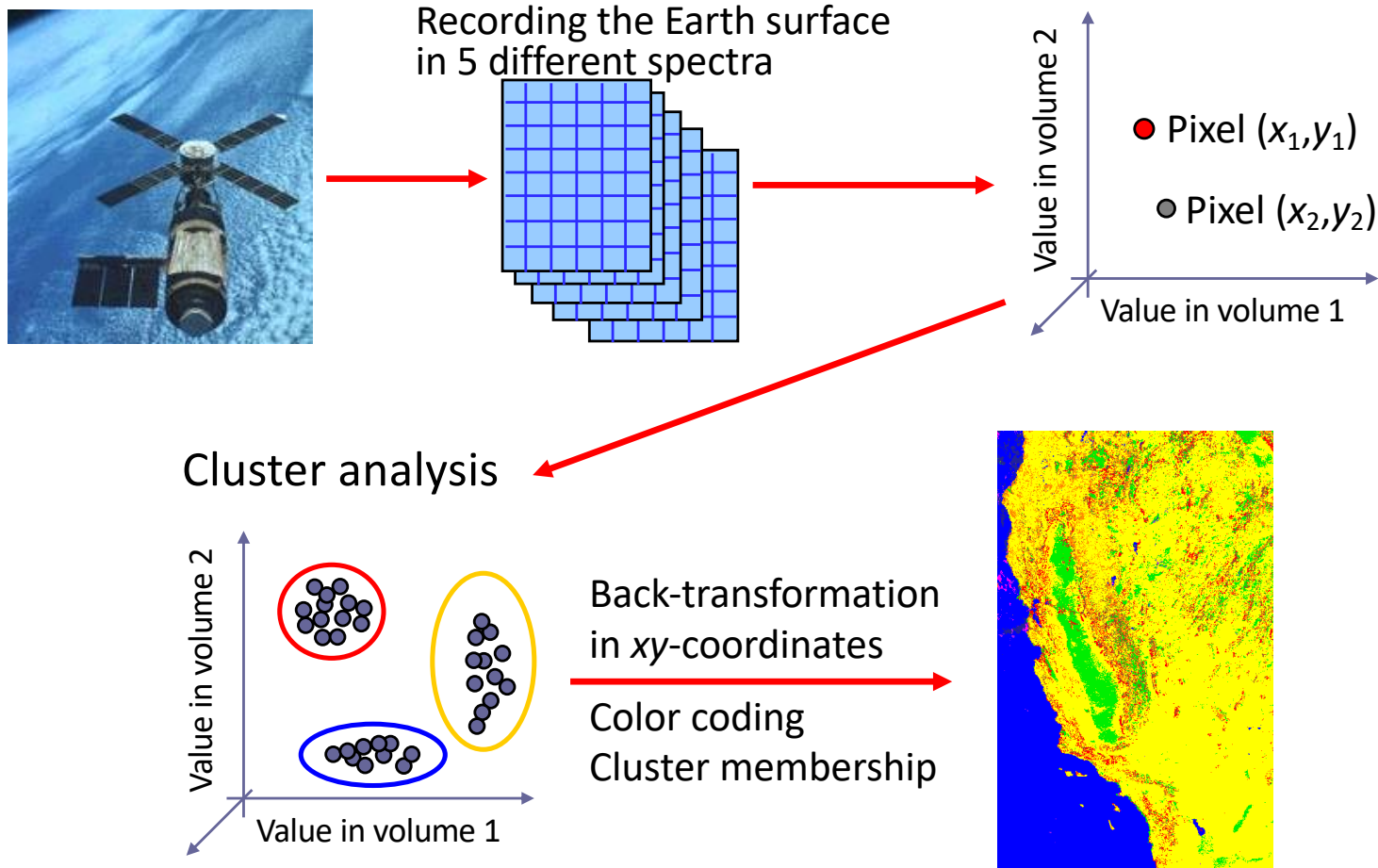


## Clustering: an example



- 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



## 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

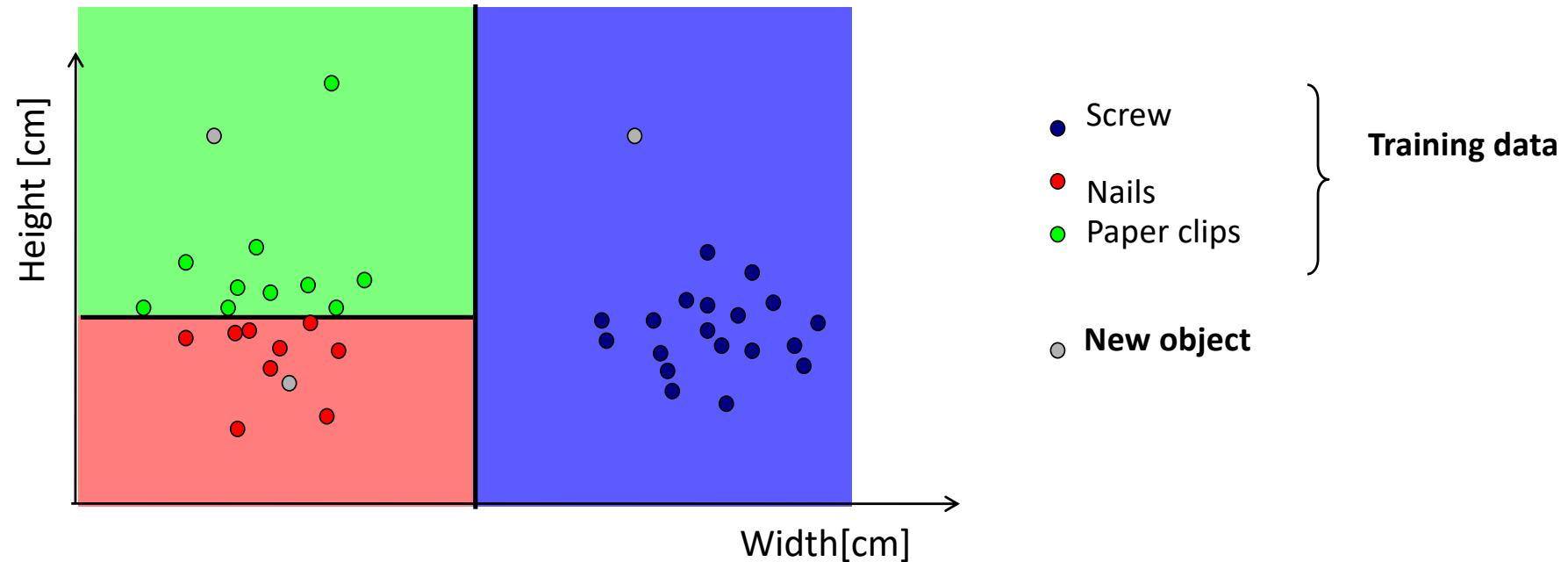
- Find groups of documents (topics) that are similar to each other based on the important terms appearing in them.
- Approach:
  - Identify important terms in each document.
  - Form a similarity measure between documents.
  - Cluster based on the similarity measure.
- Gain:
  - Help the end user to 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

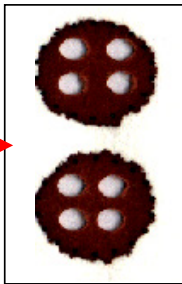
---

### ■ 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!

## Application: Newborn screening

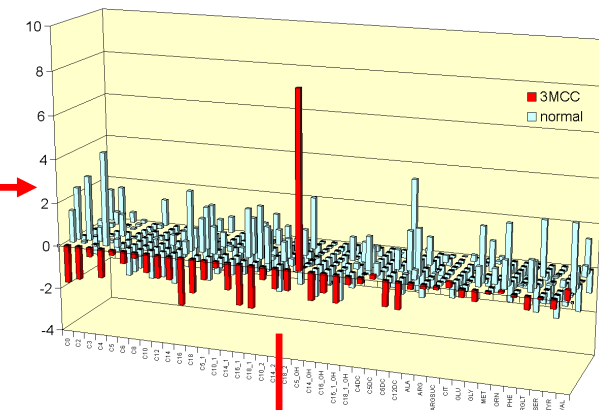
## Blood samples of newborns



## Mass spectrometry



## Metabolite spectrum



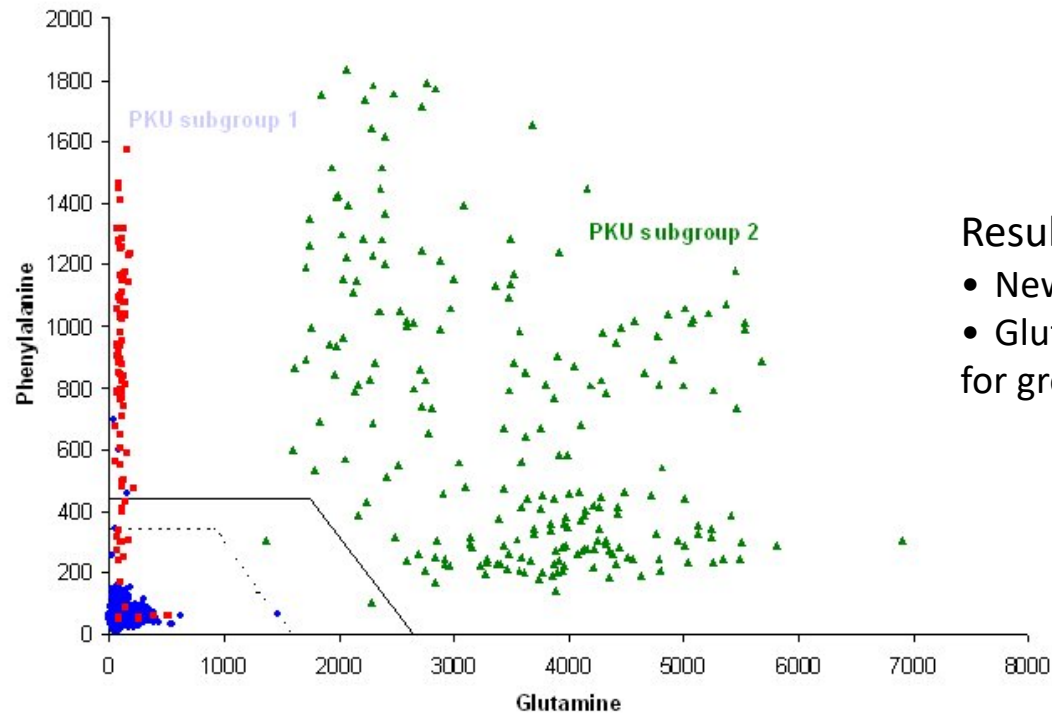
**14 analysed amino acids:**

alanine  
arginine  
argininosuccinate  
citrulline  
glutamate  
glycine  
methionine

- phenylalanine
- pyroglutamate
- serine
- tyrosine
- valine
- leucine+isoleucine
- ornitine

## Database

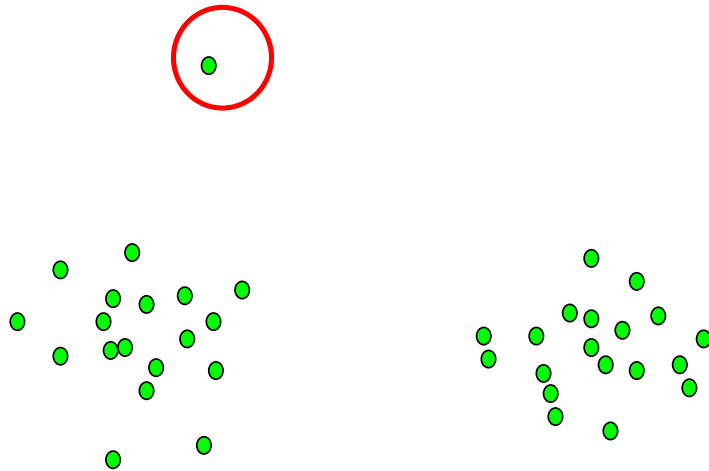
## Application: Newborn screening



### Result:

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

## 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

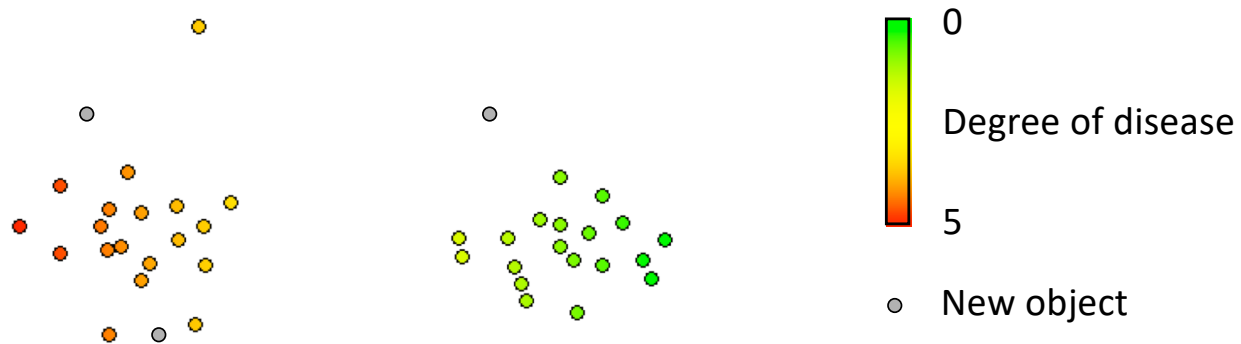
- Analysis of the SAT.1-Ran-Soccer-Database (Season 1998/99)
  - 375 players
  - Primary attributes: Name, #games, #goals, playing position (goalkeeper, defense, midfield, offense),
  - Derived attribute: Goals per game
  - Outlier analysis (playing position, #games, #goals)
- Result: Top 5 outliers

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	2 <sup>nd</sup> scorer overall
5	Giovanne Elber	21	13	Offense	High #goals/per game

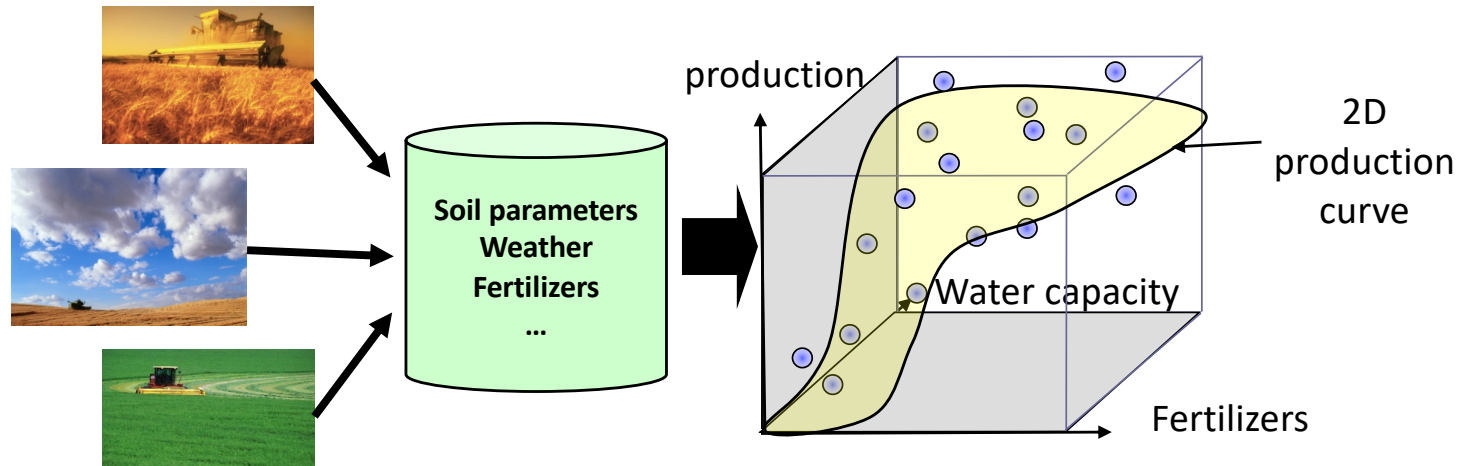
Note: “Outliers” is not necessarily a negative term.

## 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

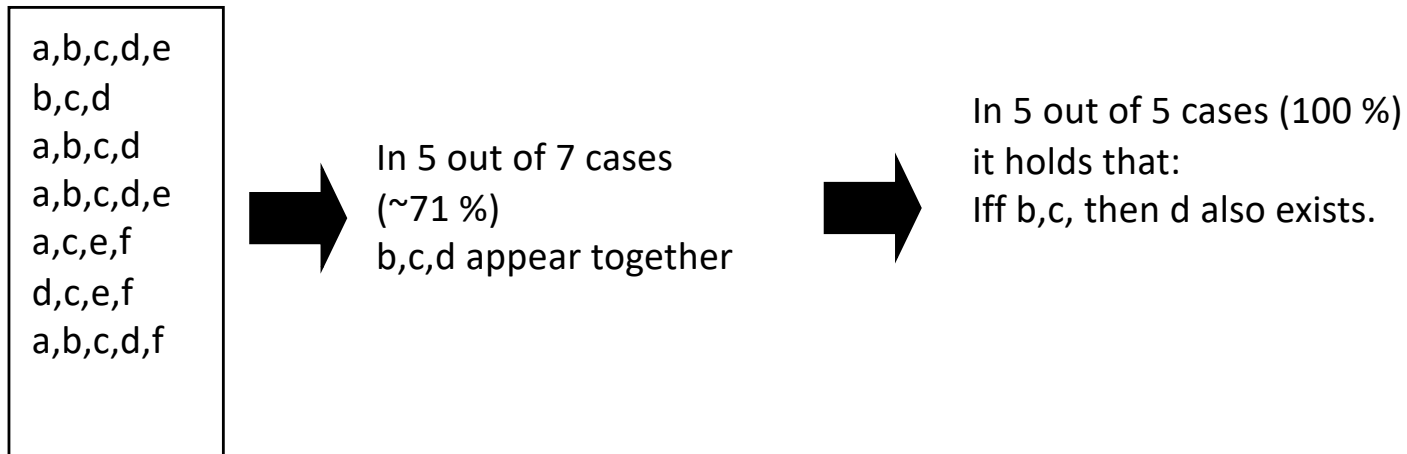


## Application: Precision farming



- 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

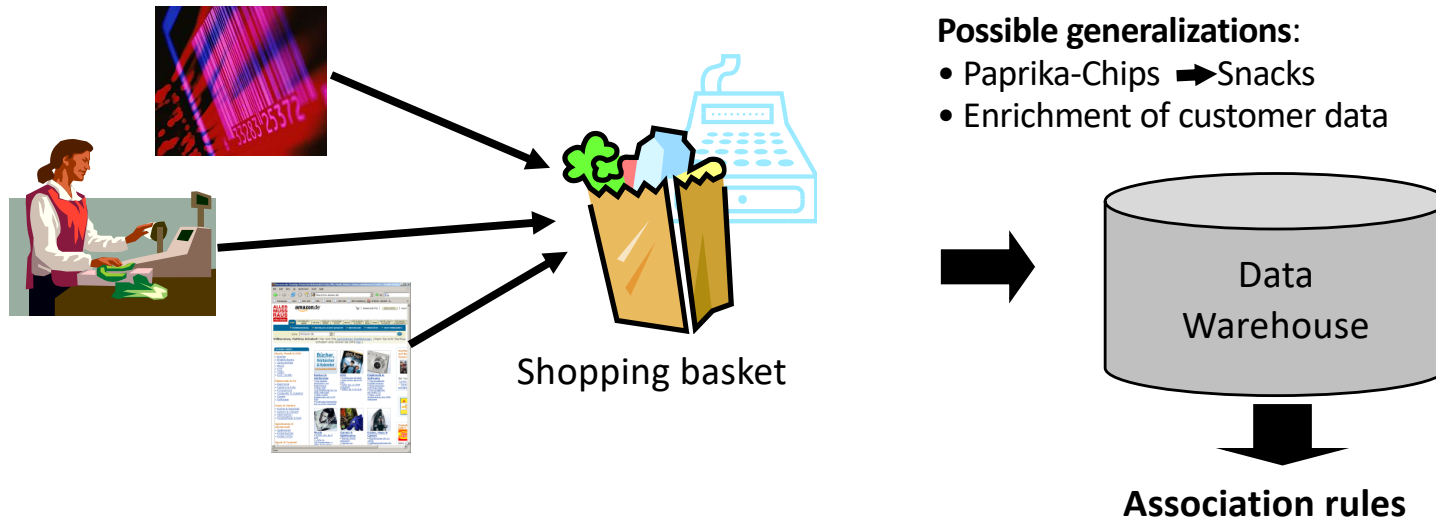


- Task: Find all rules in the database, in the following form:

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



### ■ 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

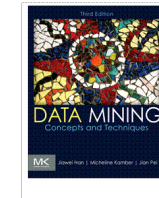
### ■ Textbook:

- 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



SciPy + NumPy



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