Knowledge Discovery in Databases II

Summer Semester 2018

Lecture 1: Introduction and outlook

Lectures : Prof. Dr. Peer Kröger, Yifeng Lu
Tutorials: Yifeng Lu

http://www.dbs.ifi.lmu.de/cms/studium_lehre/lehre_master/kdd218/

Course organization

- **Time and location**
  - Lectures: Wed, 09:00-11:30, room B U101 (Oettingenstr. 67)
  - Tutorials: Mon, 14:00-16:00, 16:00-18:00
    Tue, 14:00-16:00, 16:00-18:00

  - All information and news can be found at:
    http://www.dbs.ifi.lmu.de/cms/studium_lehre/lehre_master/kdd218/

- **Exam**
  - Written exam, 90 min
  - 6 ECTS points
  - Registration for the written exam through UniWorX (now possible)

Chapter overview

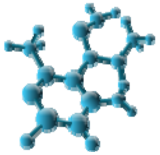
- Knowledge Discovery in Databases, Big Data and Data Science
- Data Mining with Vectorized Data (Recap KDD I)
- Topics of KDD II
- Literature and supplementary materials

Motivation

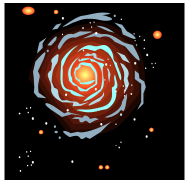
- Large amounts of data in multiple applications

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

connection data



molecule
process data



telescope data

transaction data

Web data/
click streams

. . .

- Manual analysis is infeasible

⇒ **Knowledge Discovery in Databases and Data Mining**

Goals

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

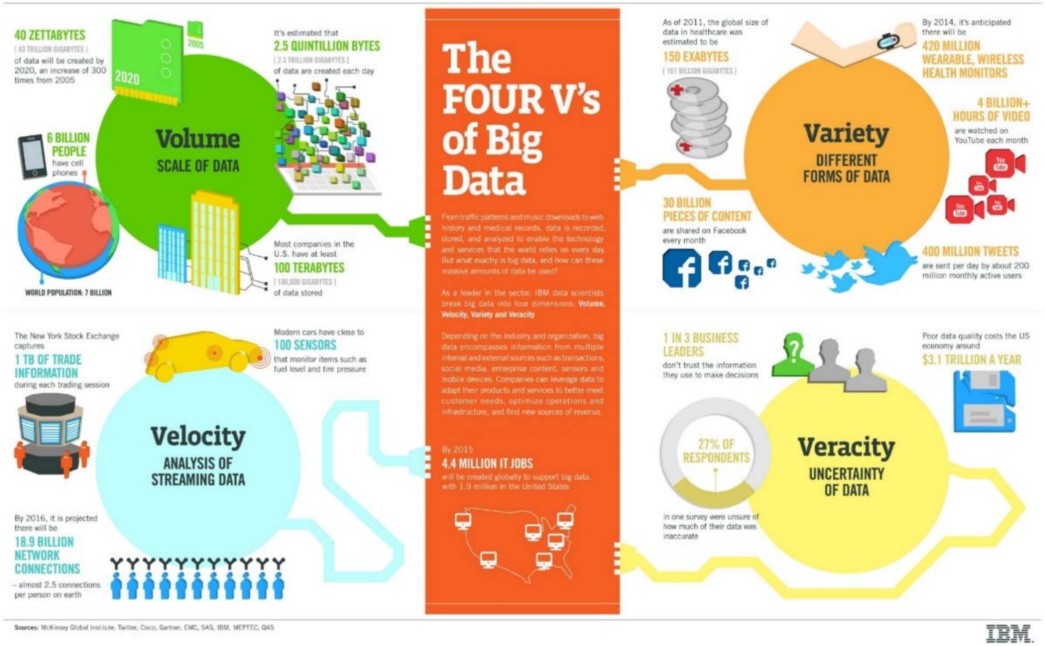
Important: The extracted models/patterns don't have to apply to 100 % of the cases.
WHY???

BuzzWord Bingo

- Big Data (McKinsey-Report 2011, …)
- Data Science
- Machine Learning und KI (AI)

BuzzWord Bingo

- Big Data (McKinsey-Report 2011, …)
  - BIG vs. VERY LARGE, some/many V's
  - Scalability/Throughput
  - Industry 4.0, Data Lake, ….
  - More a data Engineering task
  - …



BuzzWord Bingo

- Data Science

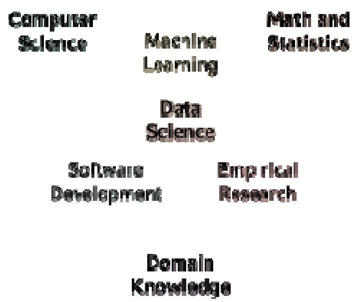
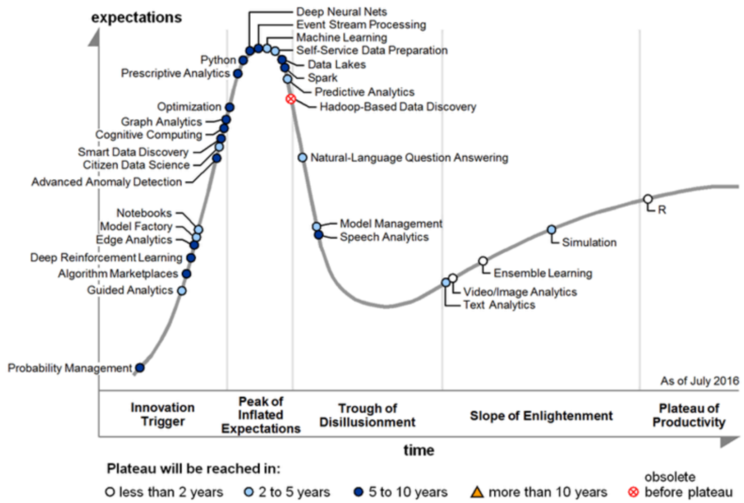


Figure 1. Hype Cycle for Data Science, 2016



Source: Gartner (July 2016)

- Often considered as a more general process to gain value from data

BuzzWord Bingo

- Machine Learning and KI (AI)

- AI: an extremely broad subject within CS (reasoning, problem solving, knowledge representation, planning, learning, natural language processing, perception, motion and manipulation, social intelligence, creativity, general intelligence

=> some major overlap to machine learning and data analytics

  - Learning in the AI context:
    - Deductive: use facts and rules to derive new facts with logic inference
    - From general to specific facts
    - Example:

      Facts: Kröger is German, all Germans have no sense of humor

      Derived fact: Kröger has no sense of humor

BuzzWord Bingo

- Machine Learning and KI (AI)

- ML: inductive learning
  - Learn general facts from single observations
  - Since we usually have not all possible observations, the derived rules are probably not 100% true
  - Example:

    Observations:
    Kröger is German, Kröger has no sense of humor
    Seidl is German, Seidl has no sense of humor
    Schubert is German, Schubert has no sense of humor

    Learned: Germans have no sense of humor

  - ML vs Data Mining: modelling vs. algorithmic approach

What is KDD?

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

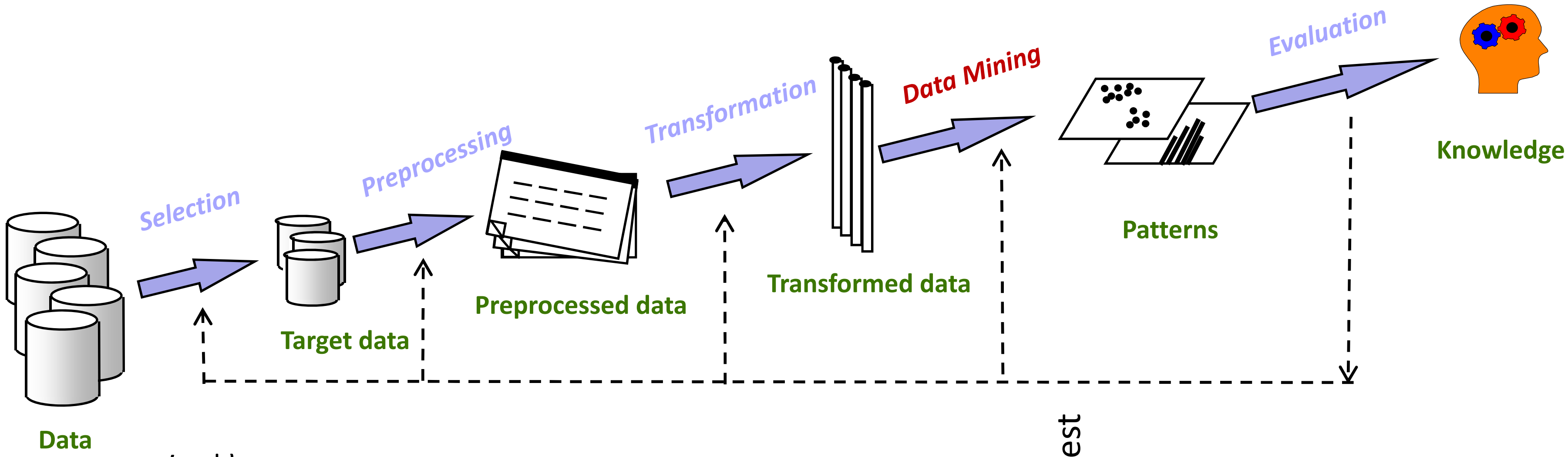
[Fayyad, Piatetsky-Shapiro, and Smyth 1996]

Remarks:

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

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/ discretization
- Dimensionality reduction

Data Mining:
- Search for patterns of interest

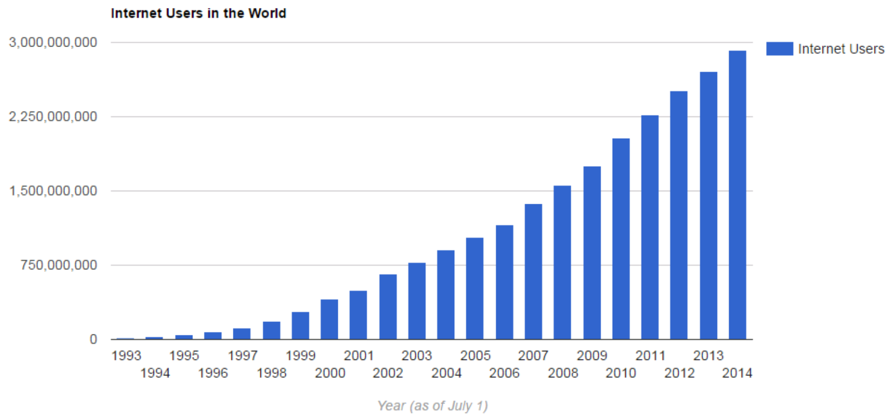
Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the models

KDD landscape today

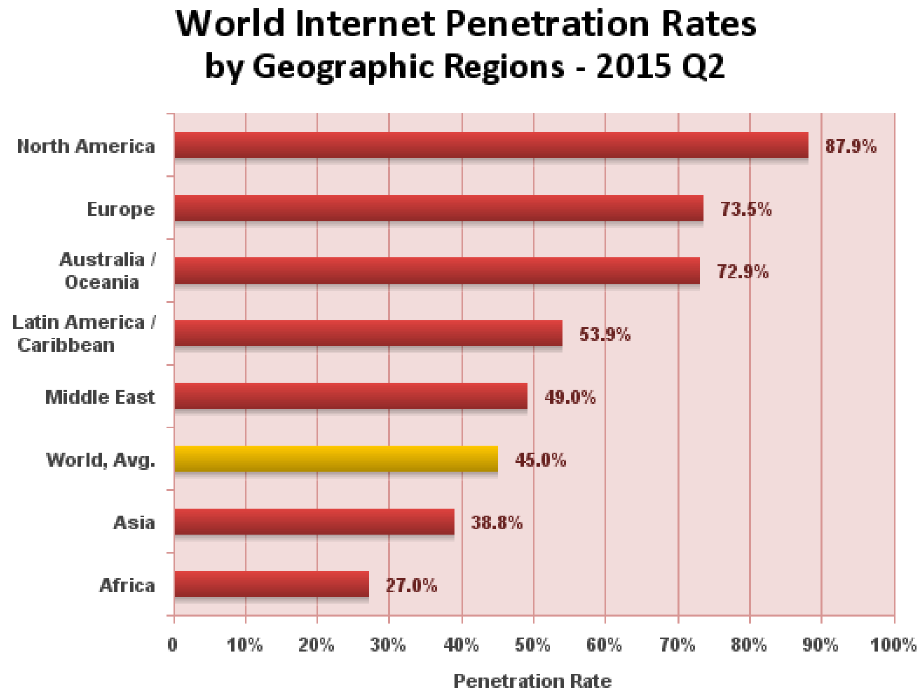
- Internet
- Internet of things
- Data intensive science / eScience
- Big data
- Data science
- …

Internet

- Internet users (Source: http://www.internetlivestats.com/internet-users/)



Web 2.0: A world of opinions



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

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

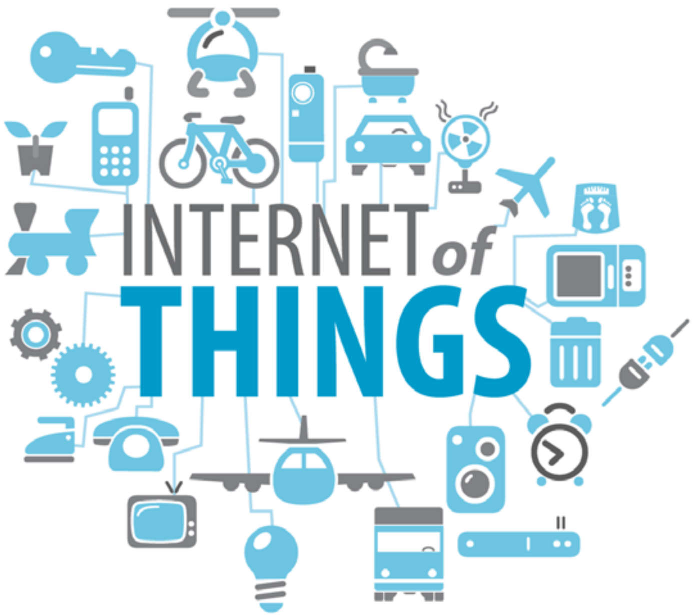


Image source:http://tinyurl.com/prtfqxf

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.

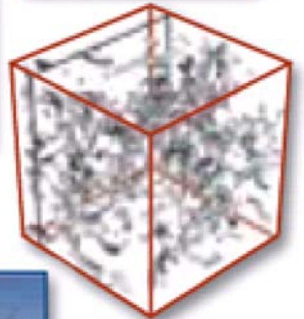
Source: http://blogs.cisco.com/diversity/the-internet-of-things-infographic

The Fourth Paradigm: Data Intensive Science 1/2

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

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



- Last few decades:
  a **computational** branch
  simulating complex phenomena

- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database/files using data management and statistics

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

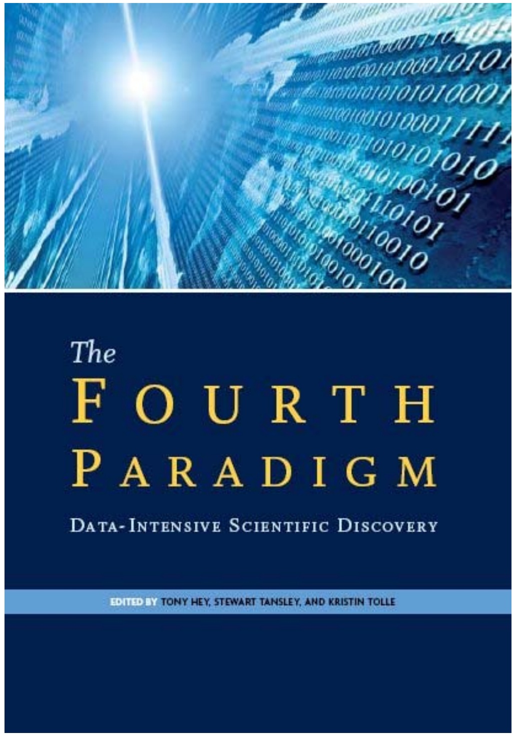
The Fourth Paradigm: Data Intensive Science 2/2

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

-The Fourth Paradigm – Microsoft

Examples of e-science applications:

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



Big Data

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

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

Capturing the value of big data:

- 300 billion USD potential value for the north American health system per year
- 250 billion Euro potential value for the public sector in Europe per year
- 600 billion USD potential value through the use for location based services

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

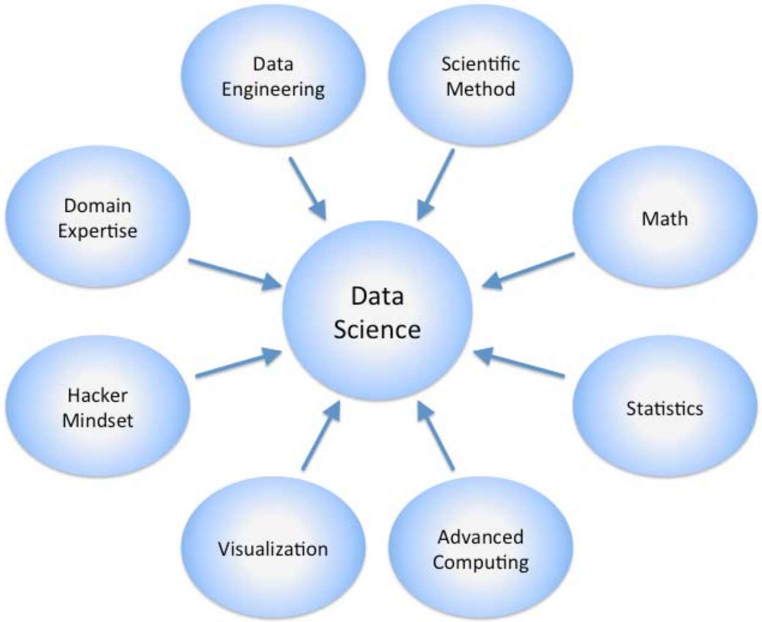
Data Scientist: The sexiest job of the 21st century:

"The United States alone faces a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts to analyze big data and make decisions based on their findings."

Source: http://tinyurl.com/cplxu6p

Data Science

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



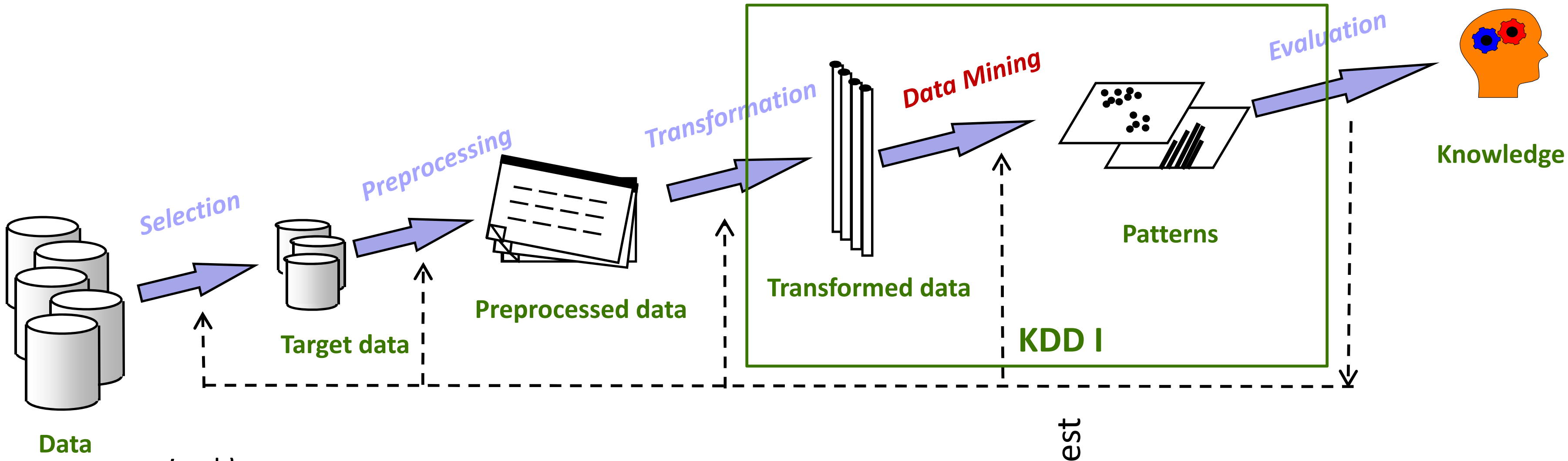
Wrap up: Many sciences worked on the topics for last decades. Data Science can be seen as an umbrella comprising all of these areas.

Chapter overview

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

The KDD process in KDD I

[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

KDD I topics

- Clustering

  partitioning, agglomerative, density-based, grid-based

- Classification

  NN-classification, Bayesian classifiers, SVMs, decision trees

- Assosiation rule mining and frequent pattern mining

  Apriori, FP-growth, FI, MFI, CFI

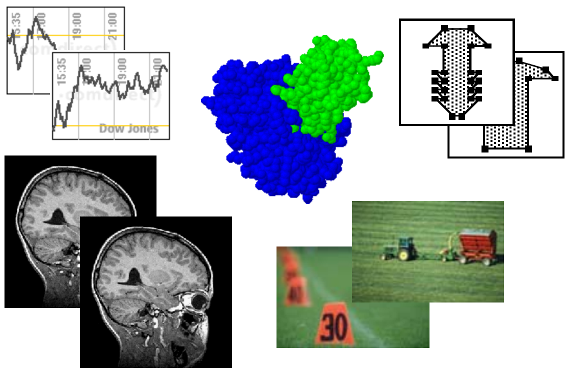
- Regression
- Outlier Detection

Most of the methods coverd by KDD I assume the data to be a set of

feature vectors

Feature Vectors/Feature Transformation

- Isn‘t this assumption to work with feature vectors extremely limiting?
  - Well …
- The concept of „Feature Transformation“ (Similarity modelling)
  - Extract characteristic (***numeric***) features from each object
  - Each object is represented as a high-dimensional (feature) vector
  - Characteristic features: similar vectors indicate similar objects



Data Space

Feature Transformation →

Histogramms

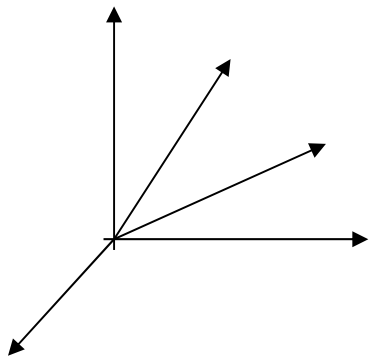
Moment Invariants

Covering

Sectoring

Fourier Transformation

…

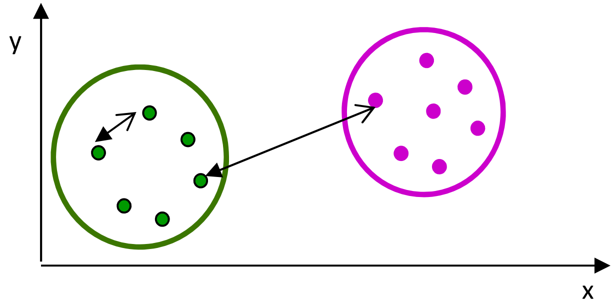


Feature Space

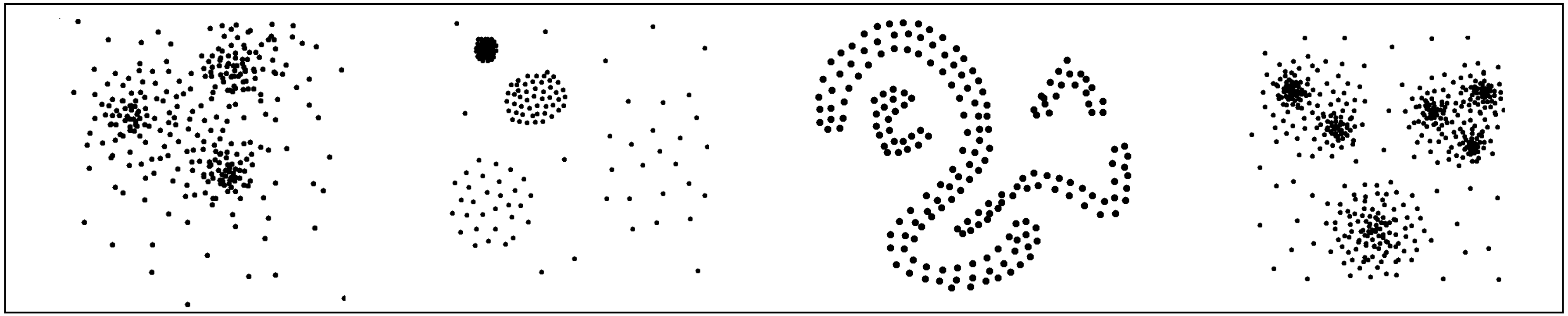
Clustering 1/3

- **Goal:**

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

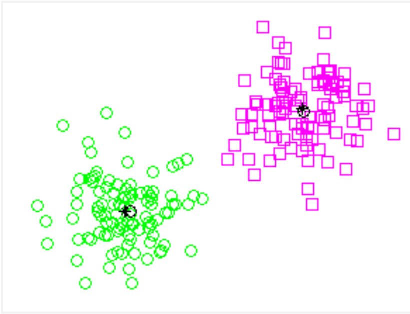


- Similarity/ distance function
- Unsupervised learning
- What is a good clustering ???

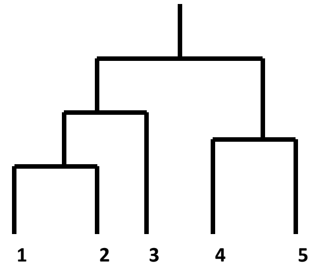


Clustering 2/3

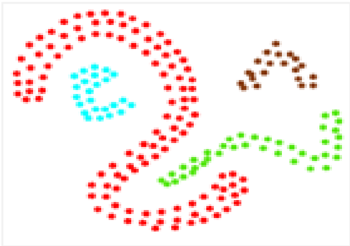
- Partitioning clustering:
  - Construct various partitions and then evaluate them by some criterion, e.g., minimizing the sum of square errors
  - Typical methods: k-means, k-medoids, CLARANS



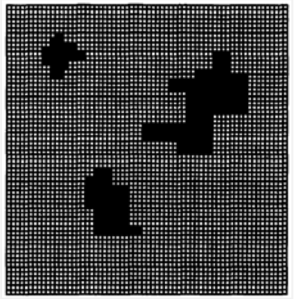
- Hierarchical clustering:
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion
  - Typical methods: Diana, Agnes, BIRCH, ROCK, CHAMELEON



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



Clustering 3/3



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

Classification 1/3

Given:

- a dataset of instances $D=\{t_1,t_2,\ldots,t_n\}$ (the ***training set***) and
- a set of classes $C=\{c_1,\ldots,c_k\}$

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

Training set D

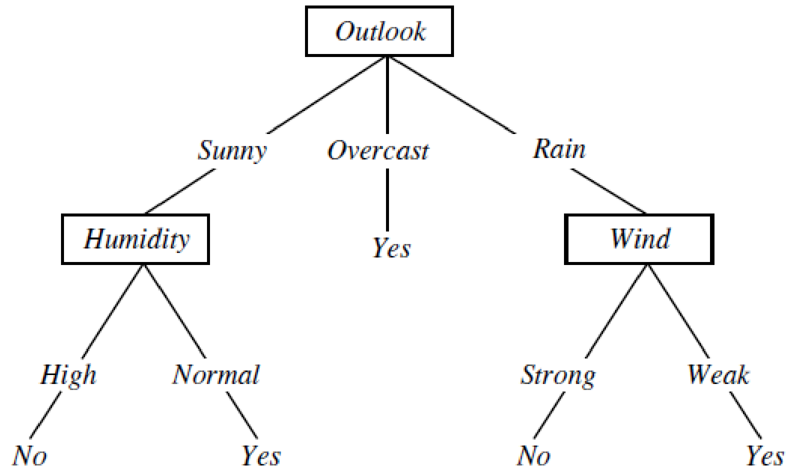
ID	Alter	Autotyp	Risk
1	23	Familie	high
2	17	Sport	high
3	43	Sport	high
4	68	Familie	low
5	32	LKW	low

A simple classifier:

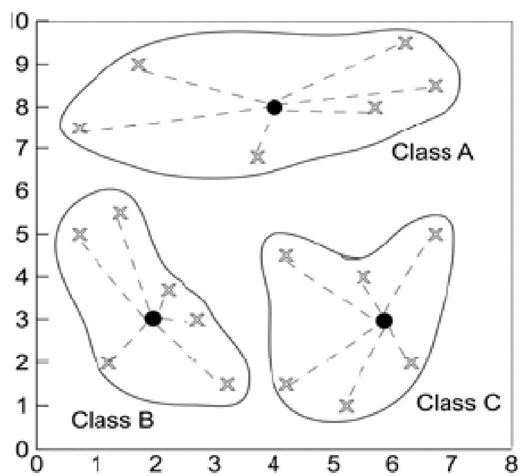
- if Alter > 50 then Risk= low;
- if Alter $\leq$ 50 and Autotyp=LKW then Risk=low;
- if Alter $\leq$ 50 and Autotyp $\neq$ LKW then Risk = high.

Classification 2/3

- Decision trees/ Partitioning
  - Partitioning along attributes
  - Purity measures (IG, Entropy)
  - Attribute independency

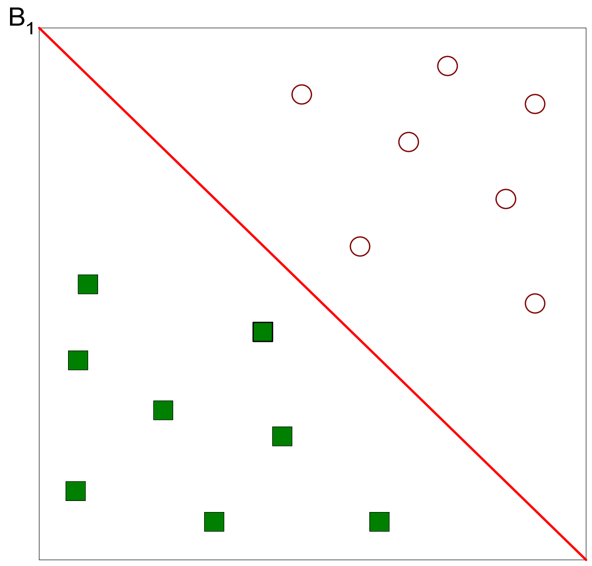


- Nearest Neighbors/ Lazy learners
  - What is the (k-th) nearest class?
  - Sensitive to outliers

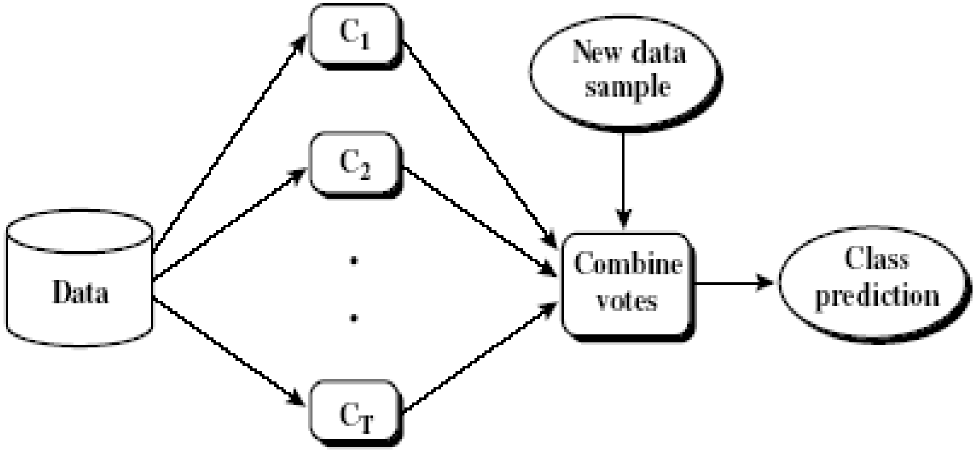


Classification 3/3

- SVM
  - Separation through hyperplane
  - Non-linearity through Kernel trick



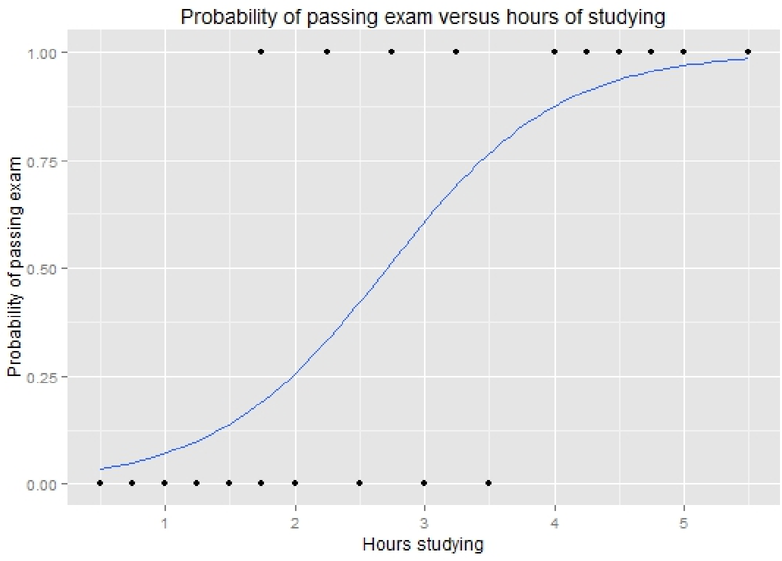
- Ensembles
  - Combination through
    e.g. majority voting



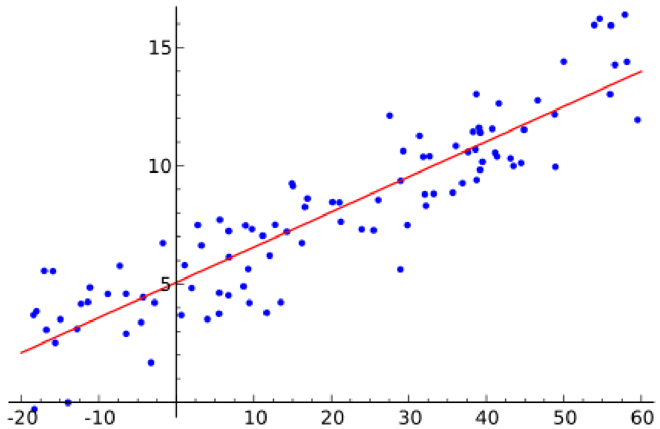
Regression

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

Logistic regression (binary outcome)

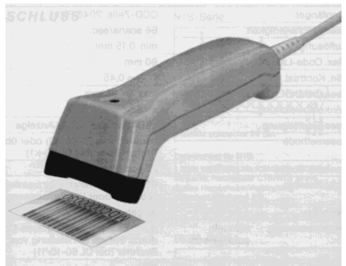


Linear regression (continuous outcome)



Association rules/ frequent patterns 1/3

- 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

Association rules/ frequent patterns 2/3

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

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

Support of 1-Itemsets:

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

Support of 2-Itemsets:

(A, C): 50%,

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

- Popular methods: Apriori, FPGrowth

Association rules/ frequent patterns 3/3

- **Problem 2: Association Rules Mining**
- Given:
  - A set of items *I*
  - A transactions database DB over *I*
  - A *minSupport* threshold *s* and a *minConfidence* threshold *c*

- Goal: Find all association rules $X \rightarrow Y$ in *DB* w.r.t. minimum support *s* and minimum
- confidence *c*, i.e.:
- $\{X \rightarrow Y \mid support(X \cup Y) \geq s, confidence(X \rightarrow Y) \geq c\}$
- These rules are called strong.

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

Association rules:

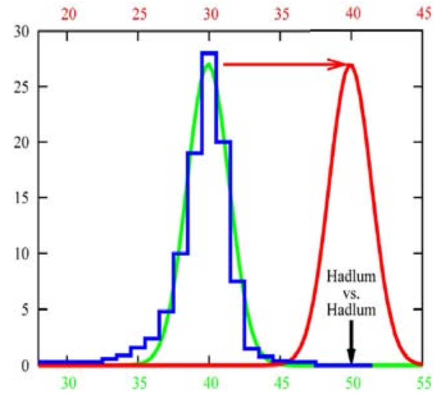
$A \Rightarrow C$ (Support = 50%, Confidence= 66.6%)

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

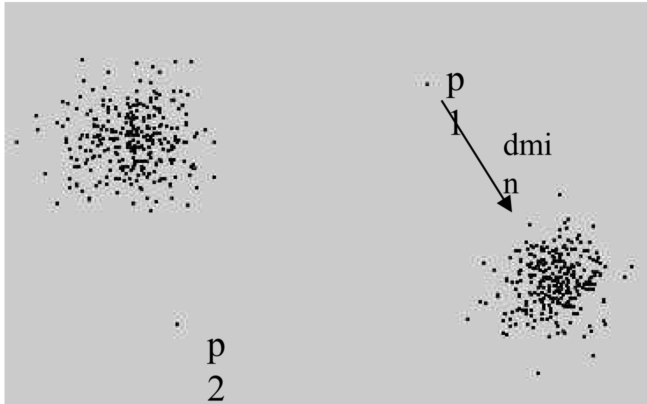
Outlier detection 1/2

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

- Statistical approaches
  - Keys:
    - Probabilistic models
    - Deviation from models

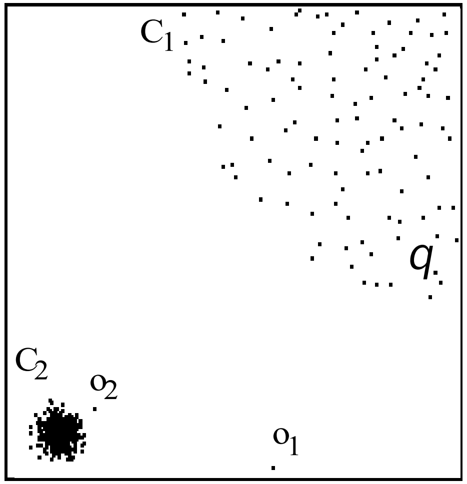


- Distance-based approaches
  - Keys:
    - Distance threshold
    - Exceeding threshold

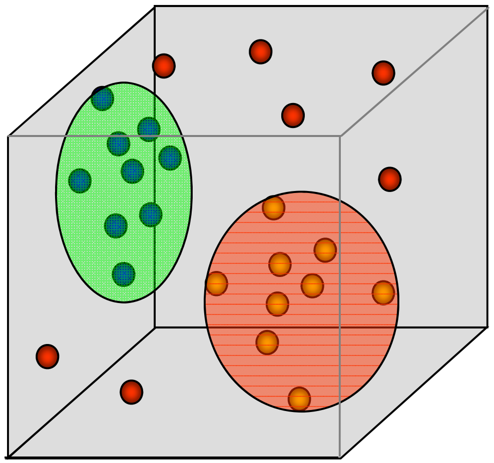


Outlier detection 2/2

- Density-based approaches
  - Keys:
    - Local density
    - Deviation from density



- Clustering-based approaches
  - Keys:
    - Clustering model
    - Missfit to model



KDD I Recap

- In KDD I, we focus on how to solve specific data mining tasks
- Observations:
  - Almost all methods work on feature vectors (only)
  - Similarity / Distance measures play a key role in various data mining tasks
    - Clustering, Classification, Prediction, etc.
    - However, only simple distance functions were introduced
- In real world, useful information hidden in data with different forms
  - Suitable Feature Transformation not easy to find
  - Feature Transformation is a simple model that might loose object semantics (compare: relational vs. object model, table vs. graphs, …)
- How to handle different types of data?
  - KDD II

Chapter overview

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

KDD I vs. KDD II

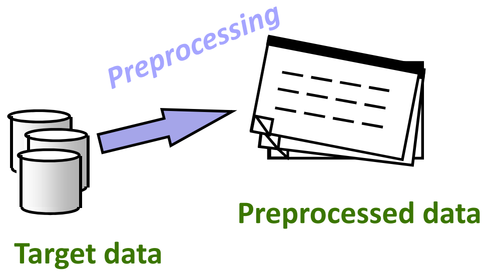
- Simple data types in KDD I
  - Vector Data
- KDD II: How to deal with different complex objects.
  - Graph
  - Text
  - High-dimensional
  - Time serious
  - Shapes
  - Spatial-temporal data
  - Multi-media data
  - Heterogeneous
  - ……

But Before We Start: Data Cleaning

- "Dirty" in Data:
  - Dummy Values, Absence of Data, Multipurpose Fields, Contradicting Data, etc.
- Steps in Data Cleaning
  - Parsing: locates and identifies individual data elements in raw data
  - Correcting: corrects parsed individual data components using sophisticated data algorithms
  - Standardizing: applies conversion routines to transform data into standard formats
  - Matching: Searching and matching records within and across data based on predefined rules
  - Consolidating: Merges data into one representation

Data Cleaning

- ...may take >60% of effort
- Integration of data from different sources
  - Mapping of attribute names (e.g. C_Nr → O_Id)
  - Joining different tables
    (e.g. Table1 = [C_Nr, Info1]
    and Table2 = [O_Id, Info2] ⇒
    JoinedTable = [O_Id, Info1, Info2])



- Elimination of inconsistencies
- Elimination of noise
- Computation of Missing Values (if necessary and possible)
  - Fill in missing values by some strategy (e.g. default value, average value, or application specific computations)
  - Uncertainty: Model each missing value by a (discrete) sample of possible values or a (continuous) distribution of possible values

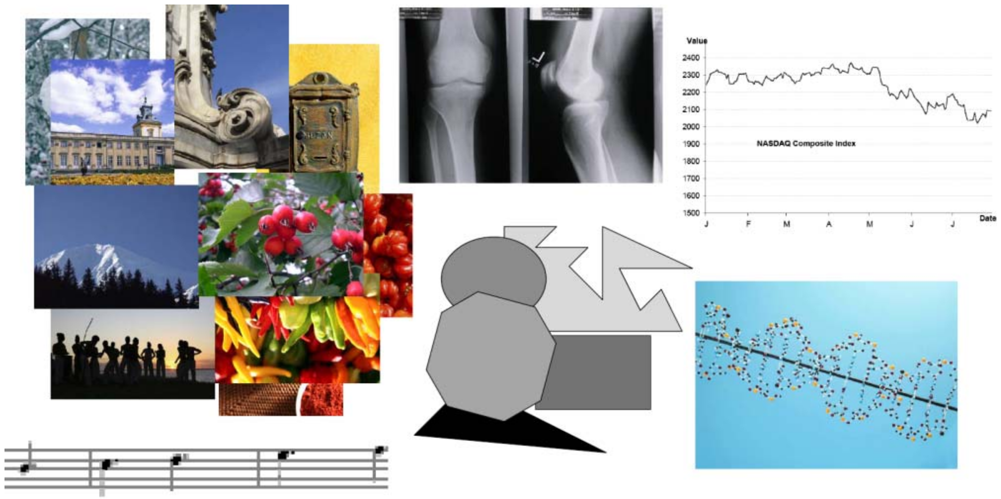
Data Cleaning (Example)

- Data Quality Mining with Association Rules
  - Association rule mining generates rules for all transactions with confidence level
  - For each transaction:
    - Determine transaction type
    - Generate all related association rules
    - Summing the confidence values of the rules it violates
  - Based on the score, user can decide whether to accept or reject the data

Association Rule	Confi-dence
Model: S-Class → Engine: Petrol	90%
Model: S-Class → Equip: AirCondTypeC	75%
Model: S-Class → Equip: AutoWindshWiper	75%
Model: S-Class → Equip: NavigSystemD	75%
⋮	⋮

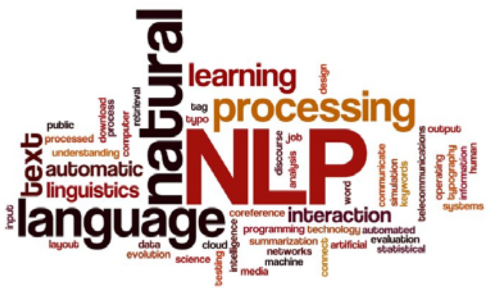
Complex Object - High-dimensional data

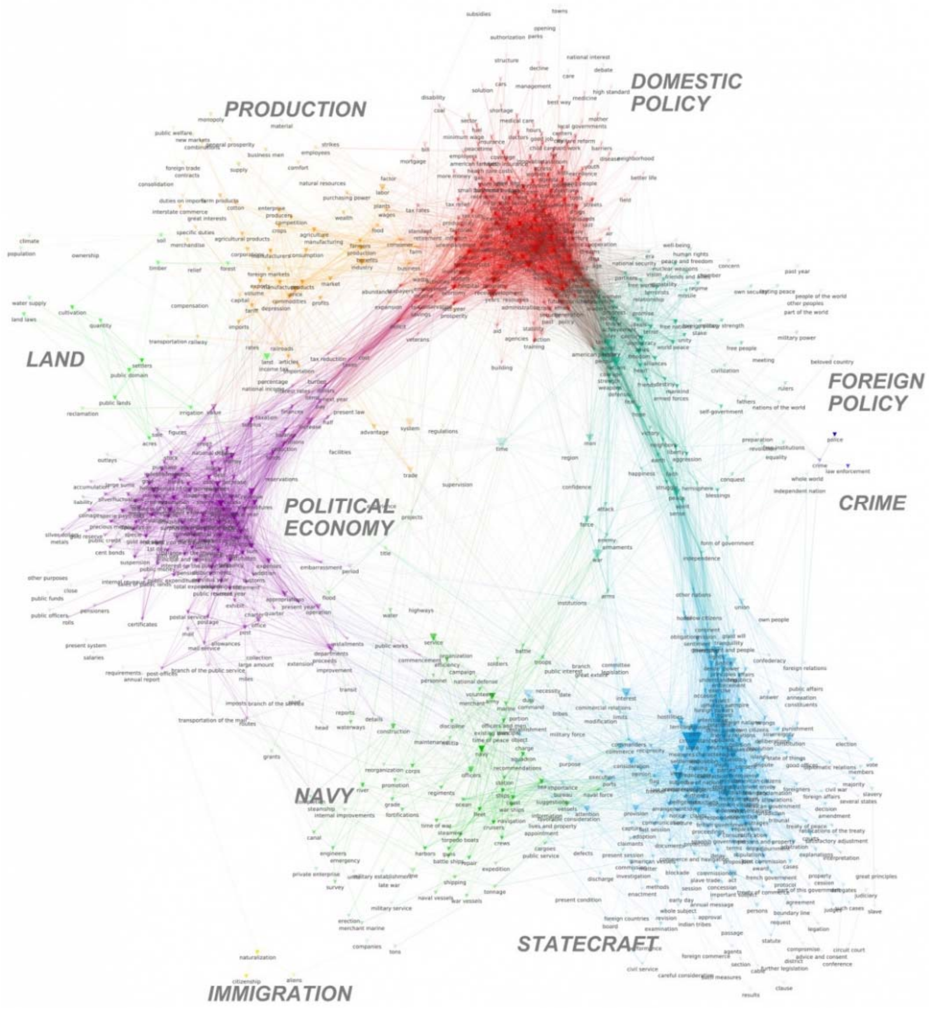
- New applications deal with high-dimensional data (business intelligence: customers, sensors; multimedia: images, videos; biology: genes, molecules)
- High-dimensional points are abstracted to feature vectors



Complex Object - Text

- Text: Sequence of Characters
  - Sentiment analysis
  - NLP
  - Books, static text corpi
  - Streams: Twitter, …





The global network structure of the SoU address, 1790–2014 [from: sciencenode.org]

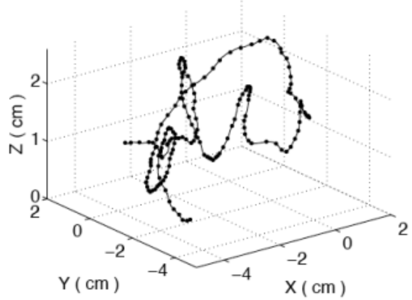
Complex Object – Sequence and Time Series Data

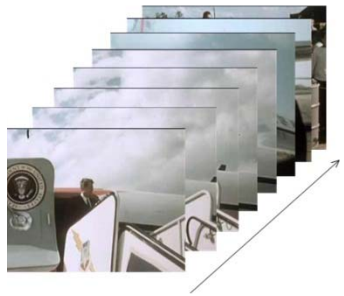
- Sequence: log of events happened in order

ATGAATTAGCTAAGGTTGTAGCTTATTTTCCATAGG
GTTTTGCTCCGGACCATCCGGTCGTGTAGCGCGATT
GACTTGCCGGGTTGTGTCCCCGTATCCAGGTCACGA
CCTCATGGGGAACTAGTGGCTGTCCGGCAGTATCCT
GGTACGCACCTCATGTGGTATGCGTGGCTGTTGGTC
CGTATATGGACCTATATATGGATCGAAGC

- Time series are a special type of sequences
  - Typically, values that are recorded over time
  - Index set I_n represents specific points in time
- Examples for **univariate time series**:
  - stock prices
  - audio data
  - temperature curves
  - ECG
  - amount of precipitation

- Examples for **multivariate time series**:
  - trajectories (spatial positions)
  - video data (e.g., color histograms)
  - combinations of sensor readings

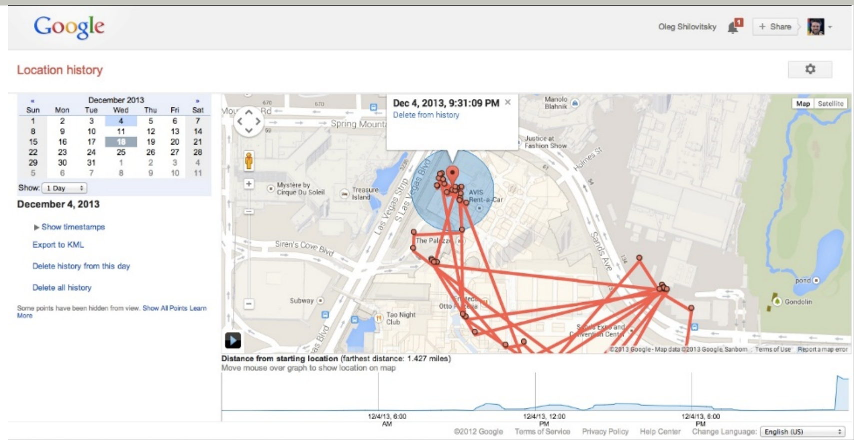


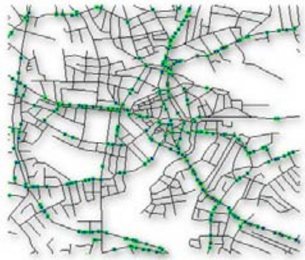


- Similarity models of time series are often based on sequence similarity models

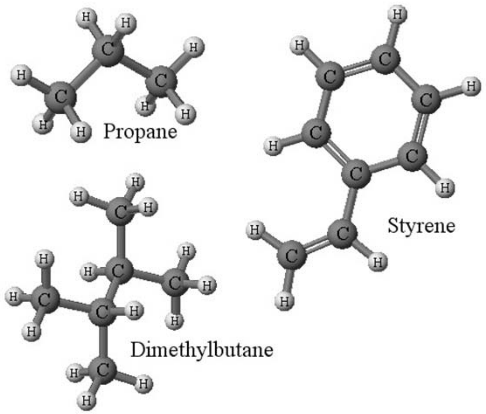
Complex Object - Spatial-temporal data

- Objects moving in space and time
- Location-based services
- Gestures
- …

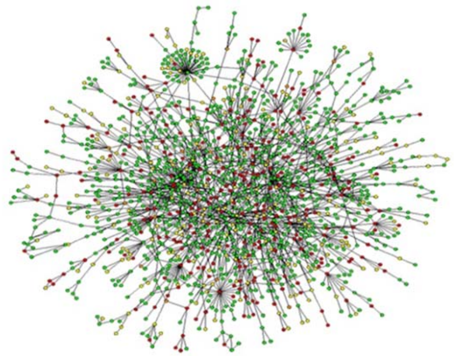




Complex Object - Graph



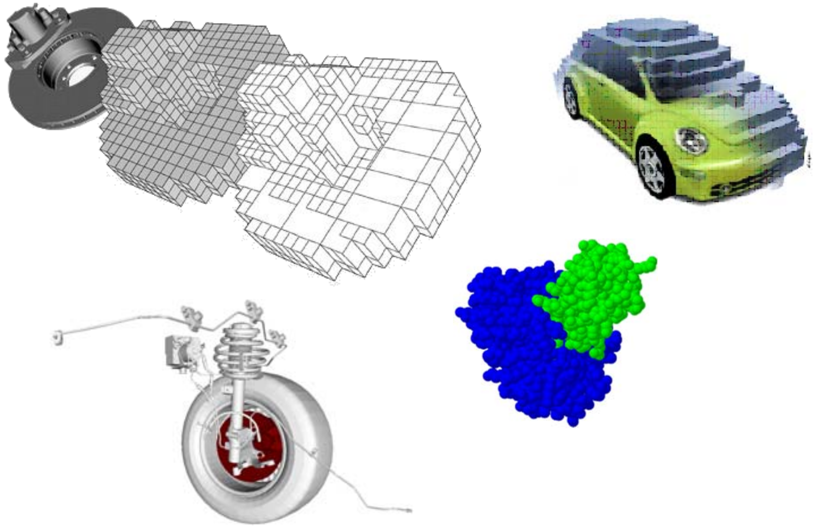
- Graphs, graphs everywhere!
  - Chemical data analysis, proteins
  - Biological pathways/networks
  - Program control flow, traffic flow, work flow analysis
  - XML, Web, social network analysis
- Graphs form a complex and expressive data type
  - Trees, lattices, sequences, and items are degenerated graphs
  - Different applications result in different kinds of graphs and tasks
    - Diversity of graphs and tasks → diversity of challenges
  - Complexity of algorithms: many problems are of high complexity (NP-complete or even P-SPACE!)

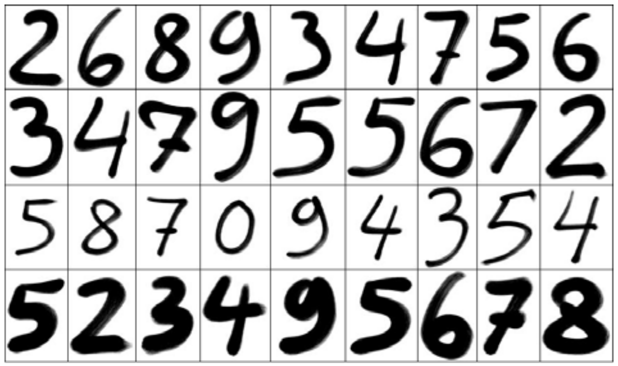


Complex Object - Shapes

- (Objects in) Images

- 2D/3D objects

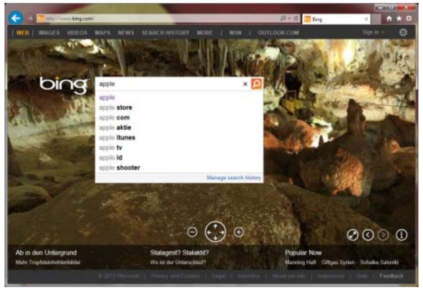


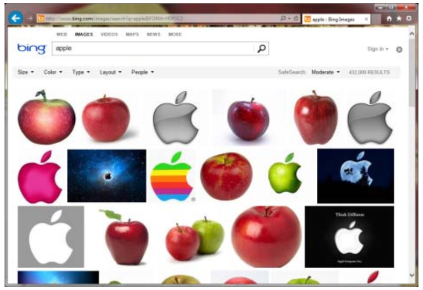


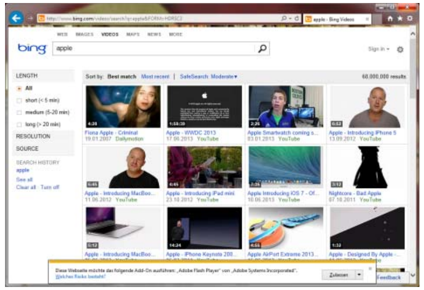
Complex Object - Multi-media data

- Rapid spread of multi-media data
- Nearly all device can generate and share multi-media data

http://www.bing.com/



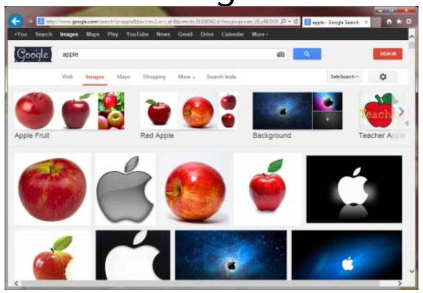


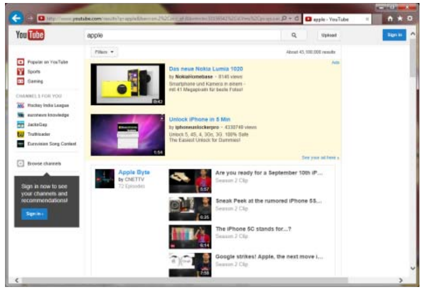


images

videos

http://www.google.com/





Chapter overview

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

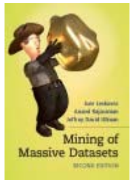
Literature

- 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

- Lescovec J, Rajaraman A., Ulman J.
  Mining of Massive Datasets
  Cambridge University Press, 2014



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

Further book titles

- 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

Online Resources

- *Mining Massive Datasets* class by Jure Lescovec, Anand Rajaraman and Jeffrey D. Ullman
  - https://www.coursera.org/course/mmds
- *Machine Learning* class by Andrew Ng, Stanford
  - http://ml-class.org/
- *Introduction to Databases* class by Jennifer Widom, Stanford
  - http://www.db-class.org/course/auth/welcome
- Kdnuggets: Data Mining and Analytics resources
  - http://www.kdnuggets.com/