CMPT 741 Data Mining

Martin Ester
Fall 2019

Introduction

Introduction

Contents of this Chapter

Big Data, Data Science, Data Mining

Relationship to other disciplines

Knowledge discovery process

Overview of the course

References

One Minute Survey

- What are the positive effects of Data Mining?
- What are the negative effects?

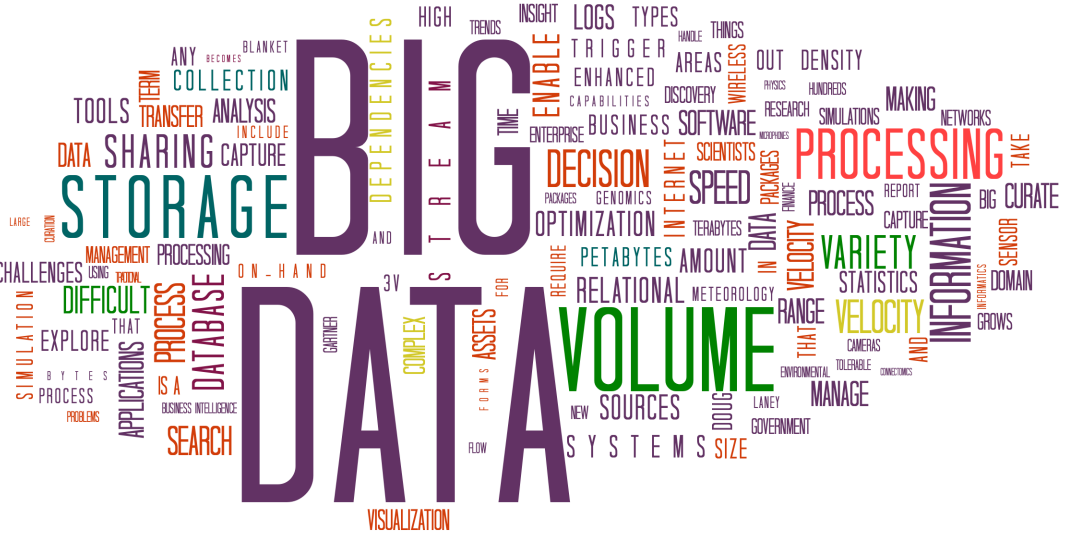
→ Discuss with your neighbor!

→ Share with the class!

Big Data, Data Science, Data Mining

Big Data

- Massive volume of both structured and unstructured data that is too large to be processed using traditional database and software techniques.



- Technology that an organization requires to store and analyze the large amounts of data.

Big Data, Data Science, Data Mining

Big Data Applications

Facebook

- Captures more than 1.5PB weblog data daily.
- Recommends friend and items.
- Makes targeted ads.

Amazon

- Collects more than 200TB of weblog data daily.
- Recommends items.
- Makes targeted ads.

Big Data, Data Science, Data Mining

Big Data Applications

Obama election campaign

- Voter's particular interests were recorded by door-to-door campaign members.
- Stored in campaign database.
- Designed emails from the local organizer to voters, each corresponding to a voter's favorite campaign issue.
- More effective and more economical ads targeting the precise demographic slices the Obama campaign was trying to reach.

Big Data, Data Science, Data Mining

Big Data Applications

Large Hadron Collider

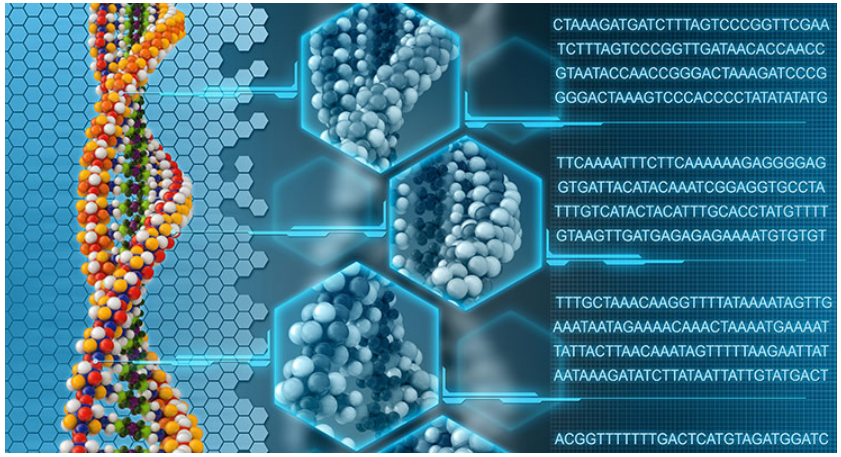
- For particle physics experiments, the largest experimental facility in the world.
- About 150 million sensors delivering data 40 million times per second.
- Nearly 600 million collisions per second.
- Detect the 100 collisions of interest per second.
- Detect and characterize new particles.

Big Data, Data Science, Data Mining

Big Data Applications

Genomics

- Sequencing one human genome produces ~1TB of data.
- Precision Medicine Initiative: 1 million volunteers provide EHRs, healthcare claims, biological samples (e.g. for DNA collection).



- Goal: understand relationship between genotype and phenotype for more precise, personalized diagnostics and treatment.

Big Data, Data Science, Data Mining

Data Science

What is it?

- Understanding the past data and predicting the future to gain actionable insight for an organization.

Is it just a new, fancy term for Statistics?

Data Scientist

- Asks the right questions.
- Manipulates data sets.
- Creates visualizations to communicate results.
- High-ranking professional with the training and curiosity to make discoveries in the world of big data.

Big Data, Data Science, Data Mining

Data Science

Requirements

- Not only technical skills,
- but also domain knowledge and communication skills.

Demand

- The shortage of data scientists is becoming a serious constraint in some sectors.
- The median salary of a junior level data scientist is $91,000, and those managing a team of ten or more data scientists earn base salaries of well over $250,000.

→ Harvard Business Review 2012: „the sexiest job of the 21st century“

Definition KDD

Knowledge discovery in databases (*KDD*) is the process of (semi-)automatic extraction of knowledge from databases which is

- *valid,*
- *previously unknown,* and
- *potentially useful*.

Remarks

- *(semi)-automatic*: different from manual analysis.
  Often, some user interaction is necessary.
- *valid*: in the statistical sense.
- *previsouly unknown*: not explicit, no „common sense knowledge“.
- *potentially useful*: for a given application.

Relationship to Other Disciplines

Contributions from Database Systems

- scalability for large datasets
- integration of data from different sources (data warehouses)
- novel datatypes (e.g. text and web data)

Contributions from Statistics

- probabilistic knowledge
- model-based inferences
- evaluation of knowledge

Contributions from Machine Learning

- different paradigms of learning
- supervised learning
- hypothesis spaces and search strategies

Relationship to Other Disciplines

Database Systems

+ discovery of implicit (not explicit) patterns
+ learning capabilities

Statistics

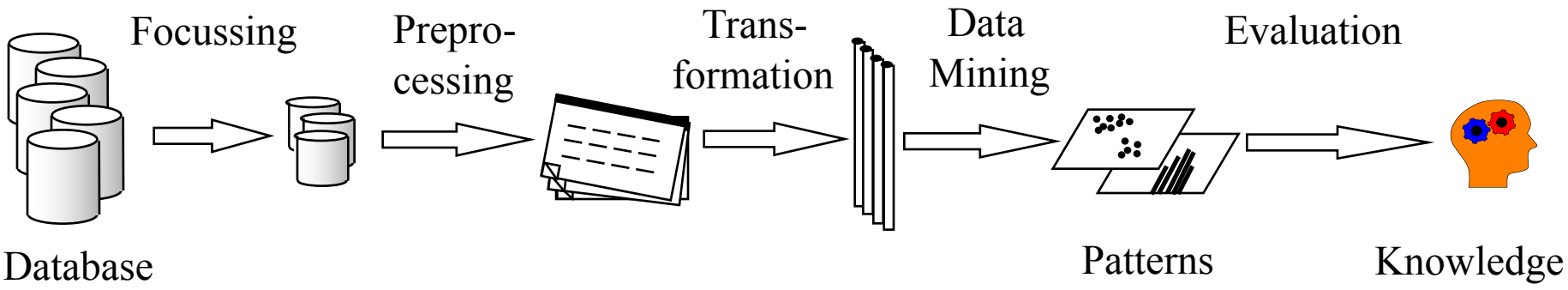
+ analysis of existing databases (not designed for task)
+ automatic generation of plausible hypotheses
+ efficient algorithms

Machine Learning

+ dealing with imperfect data
+ very large datasets
+ understandability of knowledge

KDD Process

KDD Process Model



iterative and
interactive process

Focussing

Understanding the application

- Ex.: make new telecommunication rates

Definition of the KDD goal

- Ex.: customer segmentation

Data acquisition

- Ex.: from operational billing DB

Data management

- file system or DBS?

Selection of relevant data

- Ex.: 100‘000 important customers with all calls in 2015

Preprocessing

Integration of data from different sources

- Simple conversion of attribute names (e.g. CNo → CustomerNumber)
- Use of domain knowledge for duplicate detection (e.g. spatial match based on ZIP codes)

Consistency check

- Test of application specific consistency constraints
- Resolution of inconsistencies

Completion

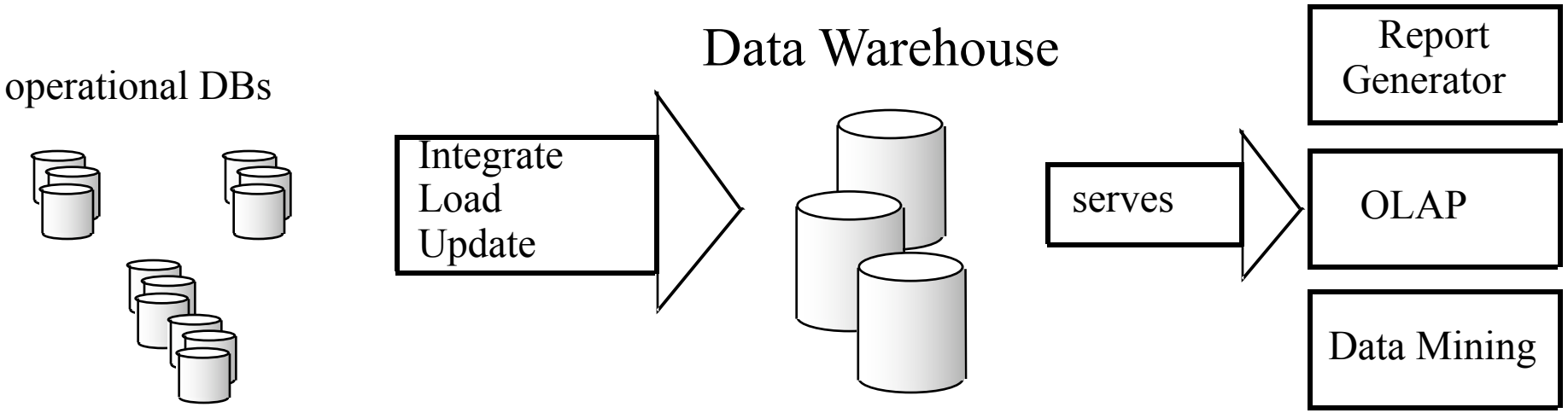
- Substitution of unknown attribute values by defaults
- Distribution of attribute values shall not be changed

preprocessing is often the most expensive KDD step

Preprocessing

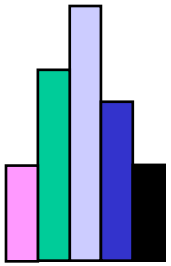
Data Warehouse

- persistent
- integrated collection of data
- from different sources
- for the purpose of analysis or decision support



Transformation

Discretization of numerical attributes



- Independent from the data mining task

  Ex.: partitioning of the attribute domain in equal-length intervals

- Specific for the data mining task

  Ex.: partitioning in intervals such that the information gain w.r.t. class membership is maximized

Generation of derived attributes

- Aggregation over sets of data records

  Ex.: from single call records to

  „Total minutes daytime / evening, weekday / weekend“

- Combination of several attributes

  Ex.: minutes change = total minutes 2018 – total minutes 2017

Transformation

Selection of attributes (features)

- *manual*
  if domain knowledge available on the attribute semantics and on the data mining task
- *automatic*
  bottom-up (starting from the empty set, add one attribute at a time)
  or top-down
  (starting from the set of all attributes, remove one attribute at a time)
  e.g. optimizing the discrimination between the different classes

⇒ too many attributes can lead to inefficient and ineffective data mining

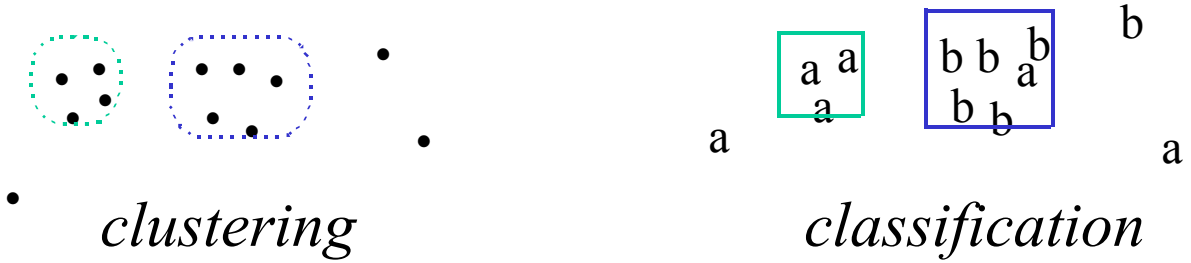
⇒ some transformations can be realized by OLAP-systems

Data Mining

Definition

Data Mining is the application of efficient algorithms that determine the patterns contained in a database.

Data mining tasks



clustering *classification*

A and B → C

association rules

other tasks: regression, outlier detection . . .

Applications

- Clustering

  customer segmentation, structuring sets of web documents,

  determining protein families and superfamilies

- Classification

  automatic credit check, automatic interpretation of astronomical images,

  prediction of protein function

- Association rules

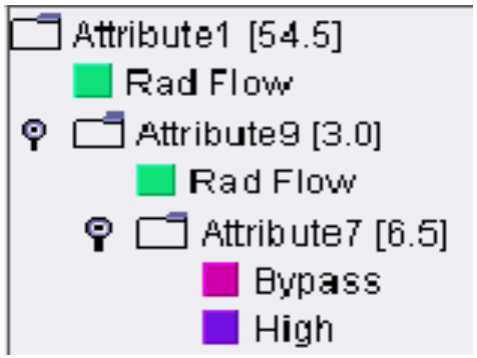
  redesign of supermarket layout, improving cross-selling,

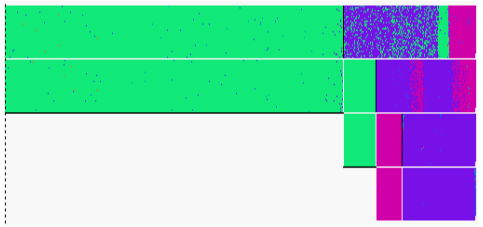
  improving the structure of a website

Evaluation

Procedure

- Presentation of discovered patterns supported by appropriate visualizations
- Evaluation of the patterns by the user
- If evaluation not satisfactory: repeat data mining with
  - Different parameters
  - Different methods
  - Different data





- If evaluation o.k.:

  Integration of discovered knowledge in the enterprise knowledge base

  Use of the new knowledge for future KDD-processes

Evaluation

Evaluation of discovered patterns

Interestingness

- Pattern already known?
- Pattern surprising?
- Pattern relevant for the application?

Predictive power

- How accurate is the pattern? (*confidence*)
- For how many cases does the pattern apply? (*support*)
- How well does the pattern generalize to unseen cases?

Overview of the Course

Prerequisites

Basic Algorithms

- algorithms,
- data structures,
- complexity,
- efficiency.

Basic Statistics

- means, standard deviation,
- probability, probability distributions,
- sampling.

Overview of the Course

Learning Outcomes

- Understanding of the main KDD concepts
- Knowledge of the most important data mining tasks and methods
- Practical skills of data mining
- Ability to select and implement data mining methods for a given application
- Foundation for research on new data mining methods

Overview of the Course

Outline

1. Introduction
2. Data preprocessing
3. Cluster analysis
4. Classification
5. Association rules and frequent pattern mining
6. Outlier detection
7. Graph mining and social network analysis
8. Recommender systems
9. Outlook

References

Textbook

- Aggarwal, C.: „*Data Mining: The Textbook*“, Springer, 2015.
- http://rd.springer.com/book/10.1007/978-3-319-14142-8

Further recommended books

- Han J., Kamber M., Pei J.: „*Data Mining: Concepts and Techniques*“, Morgan Kaufmann Publishers, 3rd ed., 2011.
- Leskovec J., Rajaraman A., Ullman J.D.: „*Mining of Massive Datasets*“, Cambridge University Press, 2nd ed., 2014.

Research articles

will be provided in class

References

Other resources

- KDNuggets: a very comprehensive resource of KDD software, companies, publications and more. http://www.kdnuggets.com/
- ACM SIGKDD: ACM's special interest group on Knowledge Discovery in Databases. http://www.acm.org/sigkdd/

Open source data mining tools resources

- WEKA (Java): http://sourceforge.net/projects/weka/
- R: https://www.r-project.org/
- Scikit-learn (Python): https://scikit-learn.org/stable/

Tentative Grading Scheme

Assignments

„Paper and pencil“ assignments 30%
Understand algorithms
Solve research-oriented problems
Prepare for the exam

Course project 30%
Data mining on a large real-life dataset
Clustering, classification, recommendation tasks
Groups of 2-3 students, working on the same tasks
Project report or poster presentation

Final exam 40%

One Minute Survey

- What is the main benefit of Data Mining?

☑ Higher sales
☑ Better service to customers
☑ Better medical services to patients
☑ More efficient government

- What is the greatest risk of Data Mining?

☑ Discrimination of groups
☑ Privacy violations
☑ Loss of jobs through automation