2. Introduction

Knowledge Discovery in Databases

Dominik Probst, Dominik.probst@fau.de
Chair of Computer Science 6 (Data Management), Friedrich-Alexander-University Erlangen-Nürnberg
Summer semester 2024

Outline

1. Why data mining?
2. What is data mining?
3. A Multidimensional View of Data-Mining
4. What kind of data can be mined?
5. What kind of patterns can be mined?
6. What technologies are used?
7. What kind of applications are targeted?
8. Major issues in data mining
9. Summary

Why data mining?

Why Data Mining? (I)

The explosive growth of data: from terabytes to petabytes and more.

- Data collection and availability:
  - Automated data collection tools.
  - Database systems.
  - World wide web.
  - Computerized society.
  - Digitization.
- Major sources of abundant data:
  - Business: web, e-commerce, transactions, stocks . . .
  - Science: remote sensing, bioinformatics, scientific simulation . . .
  - Society: news, digital cameras, social media . . .
- The era of **big data** (as inflationary used buzzword).

Why Data Mining? (II)

The initial situation:

- We are drowning in data
- We are starving for knowledge

The basic idea behind data mining:

- We can analyze the data to satisfy our hunger for knowledge

Evolution of Sciences (I)

- Before 1600, era of **empirical science**.
- 1600 – 1950s, rise of **theoretical science**.
  - Each discipline has grown a theoretical component.
  - Theoretical models often motivate experiments and generalize our understanding.
- 1950 – 1990s, rise of **computational science**.
  - Over the last 50 years most disciplines have grown a third, computational branch.
    - E.g. empirical, theoretical, and computational ecology.
    - E.g. physics, linguistics or biology.
  - Computational science traditionally meant simulation.
  - It grew out of our inability to describe reality by closed-form mathematical models.

Evolution of Sciences (II)

- 1990—now, rise of **data science**.
  - The flood of data from new instruments and modern simulations.
  - The ability to economically store and manage petabytes of data.
  - The internet makes all these archives world wide accessible.
  - Scientific *information management*,
    acquisition,
    organization,
    query, and
    visualization scale almost linearly with amount of data.
  - **Data mining** is a major new challenge!
- For further reading:
  Jim Gray and Alex Szaly: *The World Wide Telescope: An Archetype for Online Science*,
  Communications of the ACM 45(11): 50-54, 2002.

Evolution of Sciences (III)

- 1960s: Data collection, database creation,
  integrated management systems (IMS), and
  network database management systems (DBMS).
- 1970s: Relational data model, relational DBMS implementation (RDBMS).
- 1980s: RDBMS products, database creation,
  advanced data models (extended relational, object oriented, deductive etc.),
  application-oriented DBMS (spatial, scientific, engineering etc.).
- 1990s: Data mining, data warehousing, multimedia databases, web databases.
- 2000s: Stream data management and mining,
  data mining and applications,
  web technology (XML, data integration), and global information systems.

What is data mining?

What is Data Mining?

Data mining or knowledge discovery from data:

- Extraction of interesting (**non-trivial, implicit, previously unknown and potentially useful**) patterns from huge amounts of data.
- Is **data mining** a misnomer?

Alternative names:

- Knowledge discovery/mining in databases (KDD).
- Knowledge extraction.
- Data/pattern analysis.
- Data archeology/dredging.
- Information harvesting.
- Business intelligence.

Examples: Is everything Data Mining?

Considered to be data mining:

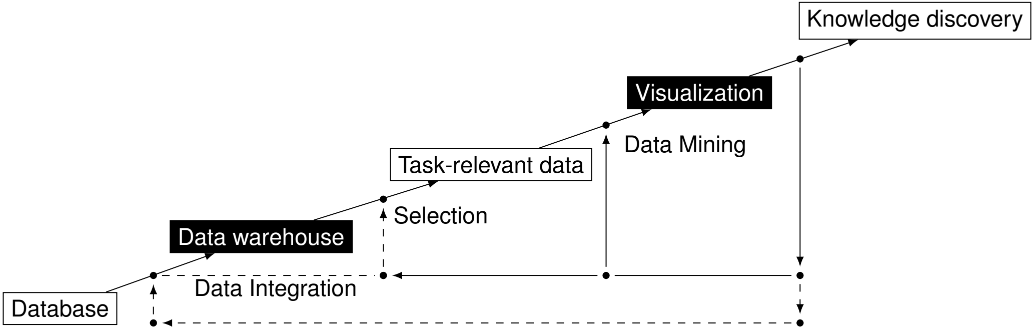
- Analysis of customer behavior for user-related advertising.
- Analysis of payment histories for fraud detection.
- Analysis of infection behavior for better understanding of a pandemic.

NOT considered to be data mining:

- Simple search for females in a customer database.
- Simple join of two database tables.
- Simple deductive database validating a new tuple with regards to predefined constraints.

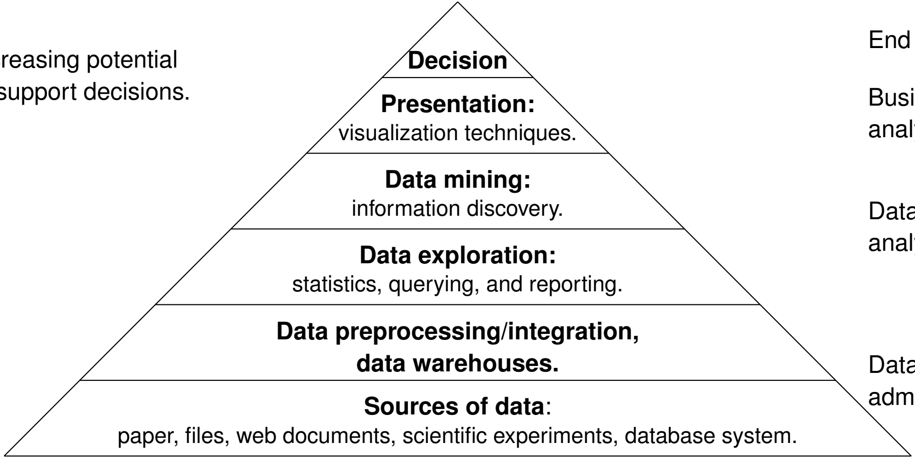
Data Mining in the Database-Systems Community

- **Knowledge discovery pipeline** is a typical view from the database-systems and data-warehousing community.
- Data mining plays an essential role in the knowledge-discovery process.



Data Mining in the Business Community

Increasing potential
to support decisions.



Decision

Presentation:
visualization techniques.

Data mining:
information discovery.

Data exploration:
statistics, querying, and reporting.

**Data preprocessing/integration,
data warehouses.**

Sources of data:
paper, files, web documents, scientific experiments, database system.

End user.

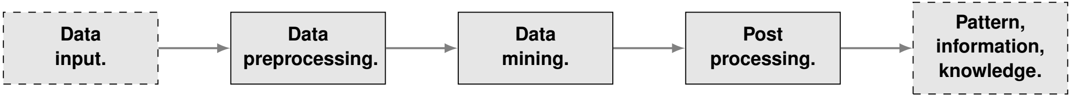
Business
analyst.

Data
analyst.

Database
administration.

Data Mining in the Machine Learning and Statistics Community

Machine-learning and statistics communities usually classify data mining as the central part of their pipeline:



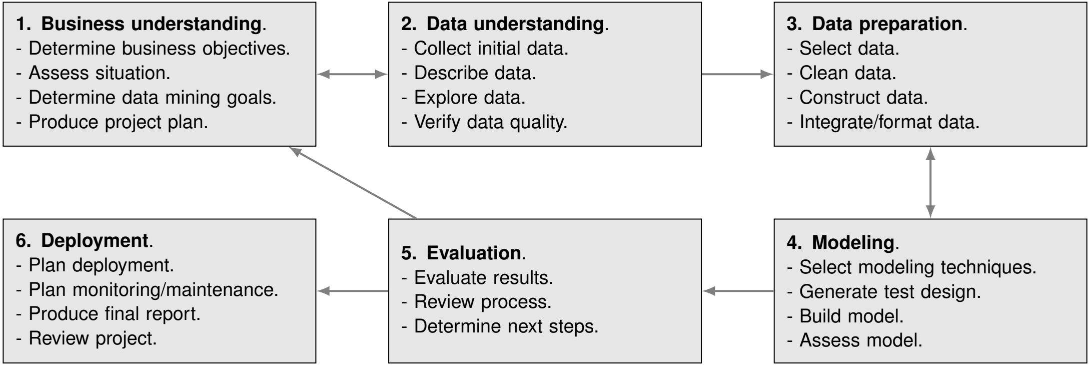
- Data integration.
- Normalization.
- Feature selection.
- Dimension reduction.

- Pattern discovery.
- Association/correlation.
- Classification.
- Clustering.
- Outlier analysis.
- ...

- Pattern evaluation.
- Pattern selection.
- Pattern interpretation.
- Pattern visualization.

The Data Mining Process: CRISP-DM

- **CRoss-Industry Standard Process for Data Mining**:



1. Business understanding.
- Determine business objectives.
- Assess situation.
- Determine data mining goals.
- Produce project plan.

2. Data understanding.
- Collect initial data.
- Describe data.
- Explore data.
- Verify data quality.

3. Data preparation.
- Select data.
- Clean data.
- Construct data.
- Integrate/format data.

4. Modeling.
- Select modeling techniques.
- Generate test design.
- Build model.
- Assess model.

5. Evaluation.
- Evaluate results.
- Review process.
- Determine next steps.

6. Deployment.
- Plan deployment.
- Plan monitoring/maintenance.
- Produce final report.
- Review project.

A Multidimensional View of Data-Mining

A Multidimensional View of Data Mining

Data mining projects can be described in four dimensions:

- **What data is available?**:
  Data can exist in a wide variety of forms and must therefore be treated differently in data mining.
- **What patterns are searched for?**:
  Various functions in data mining can be used to detect different patterns.
- **What technologies are used?**:
  The technologies used can vary greatly in data mining.
- **What is the actual target application?**:
  The actual target application also differs from case to case.

What kind of data can be mined?

What kind of data can be mined? (I)

- **Any kind of data as long as meaningful for the target application.**
- Most basic forms of data sources:
  - **Relational database:**
    Collection of tables, where the tables consist of a set of attributes and usually a large set of tuples.
  - **Data warehouse:**
    Repository of information collected from multiple sources, stored under a unified schema.
  - **Transactional database:**
    Captures transactions, such as customer purchases, flight bookings, or user clicks on a website.

What kind of data can be mined? (II)

Advanced data sets and advanced applications:

- Data streams and sensor data.
- Time series data, temporal data, sequence data (incl. biosequences).
- Structure data, graphs, social networks and multi-linked data.
- Object-relational databases.
- Heterogeneous databases and legacy databases.
- NoSQL databases.
- Spatial data and spatiotemporal data.
- Multimedia databases.
- Text databases.
- The world wide web.

What kind of patterns can be mined?

What kind of patterns can be mined?

- **Searching for the right patterns is important.**
- Which patterns can be mined depends on:
  - **Data mining function.**
    Different functions can reveal different patterns.
  - **Data set.**
    Some types of records contain special patterns that can be found only in them.
- Patterns do not always lead to useful information.
  $\longrightarrow$ Always validate whether the gained knowledge is interesting.

Data Mining Function: I. Generalization

Information integration and data warehouse construction:

- Data cleaning.
- Transformation.
- Integration.
- Multidimensional modeling.

Data cube technology:

- Characterization (contrast data characteristics).
  E.g. dry vs. wet regions from numerical humidity values.
- Discrimination.
- Generalization.
- Summarization/Aggregation.

Data Mining Function: II. Association and Correlation Analysis

Frequent patterns or item sets:
What items are frequently purchased together in your supermarket.

Association, correlation vs. causality:
A typical association rule: Diapers $\rightarrow$ Beer $[0.5\%, 75\%]$ (support, confidence).
Are strongly associated items also strongly correlated?

How to mine such patterns and rules efficiently in large datasets?
How to use such patterns for classification, clustering and other applications?

Data Mining Function: III. Classification

Classification and (class-)label prediction:
Construct models (functions) based on training examples.
Hence: "supervised".
Describe and distinguish classes or concepts for future prediction.
E.g. classify countries based on climate or classify cars based on gas mileage.
Classifying something means to predict unknown class labels.

Typical methods:
Decision trees, naive Bayesian classification, support-vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression . . .

Typical applications:
Credit-card-fraud detection, direct marketing, classifying stars, diseases, web pages . . .

Data Mining Function: IV. Cluster Analysis

Unsupervised learning: I.e. class labels are unknown.
Group data: I.e. cluster houses to find distribution patterns.

Principle:
Maximize intra class similarity and minimize inter class similarity.

What is **similarity?**

Data Mining Function: V. Outlier Analysis

Outlier: A data object that does not comply with the general behavior of the data.

Noise or exception?
One person’s garbage could be another person’s treasure.

Methods:
By-product of clustering or regression analysis.
Useful in fraud detection or rare-events analysis.

Time and Ordering: Sequential Pattern, Trend and Evolution Analysis

Sequence, trend, and evolution analysis.

- Trend, time-series, and deviation analysis.
  E.g., regression and value prediction (forecasting).
- Sequential-pattern mining.
  E.g. customers first buy a digital camera, then buy large SD memory cards.
- Periodicity analysis.
- Motifs and biological-sequence analysis.
  Approximate and consecutive motifs.
- Similarity-based analysis.
- Mining data streams.
  Ordered, time-varying, potentially infinite (unbounded).

Structure and Network Analysis

Graph mining:
Finding frequent subgraphs (e.g. chemical compounds), trees (XML), substructures (web fragments), information-network analysis.

Social networks:

- Social networks: Actors (objects, nodes) and relationships (edges).
  E.g., author networks in CS, terrorist networks.
- Multiple heterogeneous networks.
  A person could be in multiple information networks such as friends, family, classmates.
- Links carry a lot of semantical information: link mining.

Web mining:

- Web is a big information network: from PageRank to Google.
- Analysis of web information networks.
- Web community discovery, opinion mining, usage mining.

Evaluation of Knowledge

Is all mined knowledge interesting?

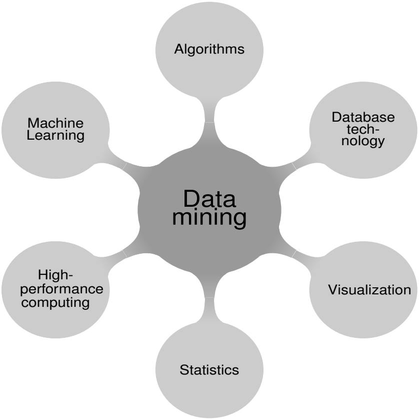
- One can mine tremendous amounts of "patterns" and knowledge.
- Some may fit only certain dimension space (e.g. time, location).
- Some may not be representative, may be transient.

Evaluation of mined knowledge → directly mine only interesting knowledge?

- Descriptive vs. predictive.
- Coverage.
- Typically vs. predictive.
- Accuracy.
- Timeliness.
- . . .

What technologies are used?

Data Mining: Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

Tremendous amount of data:

- Algorithms must be highly scalable to handle also terabytes of data.

High dimensionality of data:

- DNA microarrays may have tens of thousands of dimensions.
  Collections of microscopic DNA spots attached to a solid surface.

High complexity of data:

- Data streams and sensor data.
- Time-series data, temporal data, sequence data.
- Structure data, graphs, social networks, and multi-linked data.
- Heterogeneous databases and legacy databases.
- Spatial, spatiotemporal, multimedia, text, and web data.
- Software programs, scientific simulations.

New and sophisticated applications.

What kind of applications are targeted?

Applications of Data Mining (I)

- **Wherever there is data and more knowledge is desired, there are data mining applications.**
- Typical data mining applications:
    - **Business Intelligence**
      Provides historical, current, and predictive views of business operation.
    - **Web Search Engines**
      Need to decide which pages to index, which ones to index and how to rank them for search.
    - **Fraud detection**
      Possible fraud attempts automatically based on suspicious patterns in transactions.
    - **Predictive Maintenance**
      Evaluation of sensor data to maintain machines in time before a defect occurs.

Applications of Data Mining (II)

- Example research projects using data mining at FAU[1]:
  - **Prediction of product properties using data mining methods.**
    Prof. Dr.-Ing. Sandro Wartzack (Chair of Engineering Design)
  - **Combustion and fuel optimization for the utilization of residues in biomass furnaces.**
    Prof. Dr.-Ing. Jürgen Karl (Chair of Energy Process Engineering)
  - **CoralTrace – A new approach to understanding climate-induced reef crises.**
    Prof. Dr. Wolfgang Kießling (Chair of Palaeontology)
  - **Performance Analysis in Team Sports.**
    Prof. Dr. Björn Eskofier (Machine Learning and Data Analytics Lab)
  - **And many more.**
    Chair of computer science 6 (data management) has some projects related to data mining, too. More information will be given in the last lecture.

[1] Found in the FAU CRIS (Current research information system): `https://cris.fau.de/`

Major issues in data mining

Major Issues in Data Mining (I)

Mining methodology:

- Mining various and new kinds of knowledge.
- Mining knowledge in multi-dimensional space.
- Data mining: An interdisciplinary effort.
- Boosting the power of discovery in a networked environment.
- Handling noise, uncertainty, and incompleteness of data.
- Pattern evaluation and pattern- or constraint-guided mining.

User interaction:

- Interactive mining.
- Incorporation of background knowledge.
- Presentation and visualization of data mining results.

Major Issues in Data Mining (II)

Efficiency and scalability:

- Efficiency and scalability of data-mining algorithms.
- Parallel, distributed, stream and incremental mining methods.

Diversity of data types:

- Handling complex types of data.
- Mining dynamic, networked and global data repositories.

Data mining and society:

- Social impacts of data mining.
- Privacy-preserving data mining.
- Invisible data mining.

Summary

Summary

Data mining:
Discovering interesting patterns and knowledge from massive amounts of data.

A natural evolution of database technology:
In great demand, with wide applications.

KDD pipeline includes:
Data cleaning, data integration, data selection, transformation, data mining, pattern evaluation, and knowledge presentation.

Mining can be performed in a variety of data.
Data-mining functionalities:
Characterization, discrimination, association, classification, clustering, outlier analysis, and trend analysis.

Data-mining technologies and applications.
Major issues in data mining.

Any questions about this chapter?

Ask them now or ask them later in our forum:

StudOn Forum
🔗 https://www.studon.fau.de/frm5699567.html

Appendix

A Brief History of Data Mining Society

- **1989 IJCAI Workshop on Knowledge Discovery in Databases:**
  Knowledge Discovery in Databases (G. Piatetsky-Shapiro and W. Frawley, 1991).
- **1991-1994 Workshops on Knowledge Discovery in Databases:**
  Advances in Knowledge Discovery and Data Mining (U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy, 1996).
- **1995-1998 International Conferences on Knowledge Discovery in Databases and Data Mining (KDD'95-98):**
  Journal of Data Mining and Knowledge Discovery (1997).
- **ACM SIGKDD conferences since 1998 and SIGKDD Explorations.**
- **More conferences on data mining:**
  PAKDD (1997), PKDD (1997), SIAM-Data Mining (2001), (IEEE) ICDM (2001), etc.
- **Journal ACM Transactions on KDD starting in 2007**.

Conferences and Journals on Data Mining (I)

KDD Conferences:

- ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases and Data Mining (KDD).
- SIAM Data Mining Conf. (SDM).
- (IEEE) Int. Conf. on Data Mining (ICDM).
- European Conf. on Machine Learning and Principles and Practices of Knowledge Discovery and Data Mining (ECML-PKDD).
- Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD).
- Int. Conf. on Web Search and Data Mining (WSDM).

Conferences and Journals on Data Mining (II)

Other related conferences:

- DB conferences: ACM SIGMOD, VLDB, ICDE, EDBT, ICDT, . . .
- Web and IR conferences: WWW, SIGIR, WSDM, . . .
- ML conferences: ICML, NIPS, ICLR . . .
- PR conferences: CVPR, ICPR . . .

Journals:

- Data Mining and Knowledge Discovery (DAMI or DMKD).
- IEEE Trans. On Knowledge and Data Eng. (TKDE).
- KDD Explorations.
- ACM Trans. on KDD.

Starting points to Find References? (I)

Data mining and KDD (SIGKDD: CD-ROM):

- Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
- Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD.
- KDnuggets: `www.kdnuggets.com`.

Database systems (SIGMOD: ACM SIGMOD Anthology CD-ROM):

- Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA.
- Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.

AI & Machine Learning:

- Conferences: Machine learning (ML), AAAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
- Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

Starting points to Find References? (II)

Web and IR:

- Conferences: SIGIR, WWW, CIKM, etc.
- Journals: WWW: Internet and Web Information Systems.

Statistics:

- Conferences: Joint Stat. Meeting, etc.
- Journals: Annals of Statistics, etc.

Visualization:

- Conferences: CHI, ACM-SIGGraph, etc.
- Journals: IEEE Trans. Visualization and Computer Graphics, etc.