Friedrich-Alexander-Universität
Technische Fakultät

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 2025

Outline

Why Data Mining?

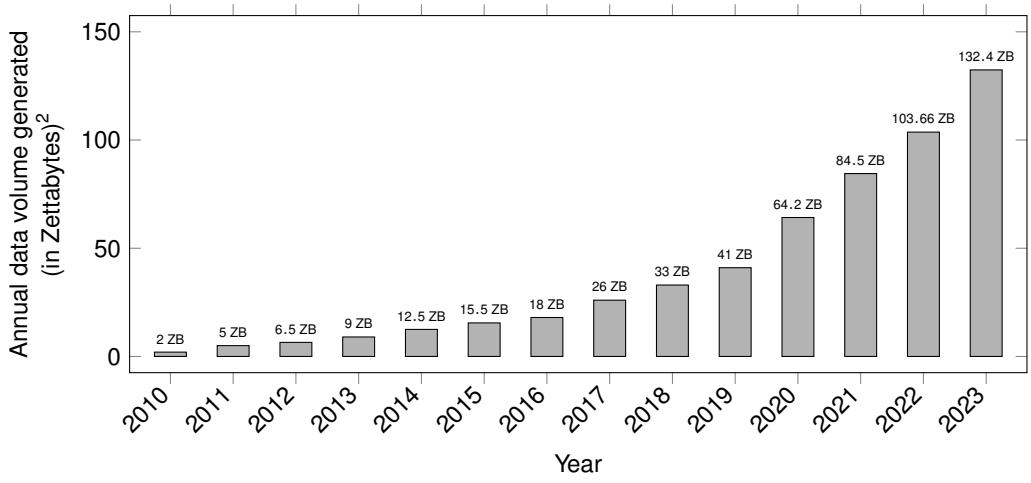
Why Data Mining? (I)

The explosive growth of data: from petabytes to exabytes to zettabytes[1] and beyond.

- 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).

[1] 1 Zettabyte = 1,000,000,000,000,000,000,000 Byte (21 Zeros!)

Why Data Mining? (II)



[2]Source: https://de.statista.com/statistik/daten/studie/267974/umfrage/prognose-zum-weltweit-generierten-datenvolumen/

Why Data Mining? (III)

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

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:

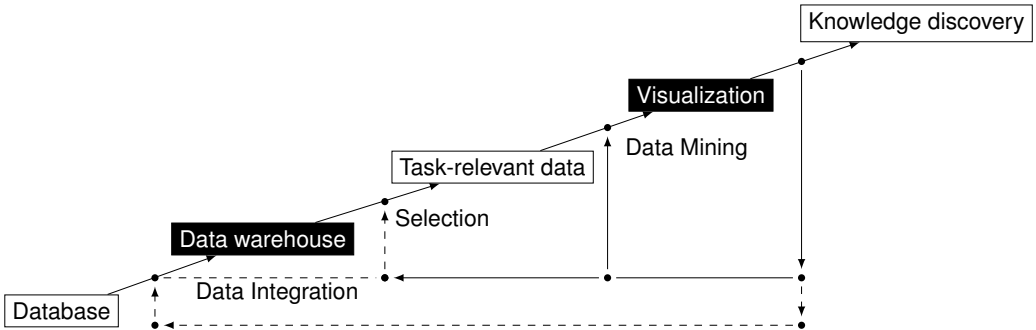
- Analysis of customer behavior for user-related advertising.
- Analysis of payment histories for fraud detection.
- Analysis of infection behavior for better understanding of a pandemic.

NOT considered to be data mining:

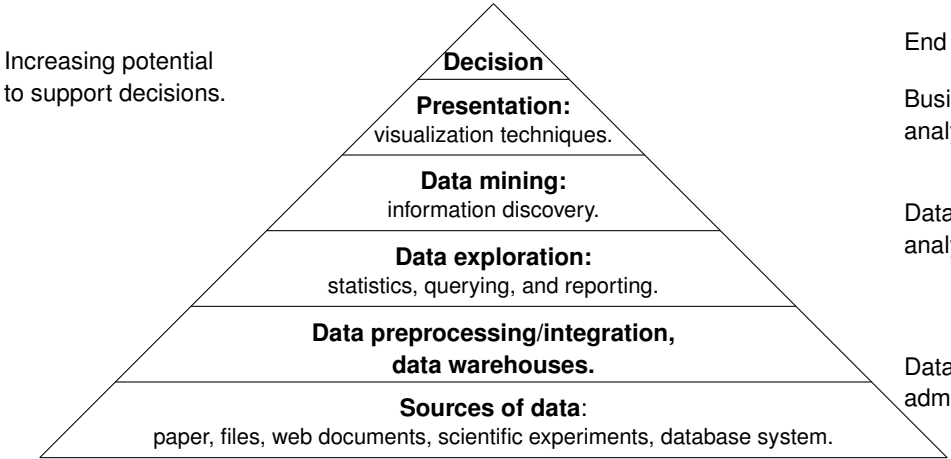
- Simple search for females in a customer database.
- Simple join of two database tables.
- Simple deductive database validating a new tuple with regards to predefined constraints.

Data Mining in the Database Community

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



Data Mining in the Business Community



Data Mining in the ML Community

The Machine Learning community usually classifies data mining as the central part of its pipeline:

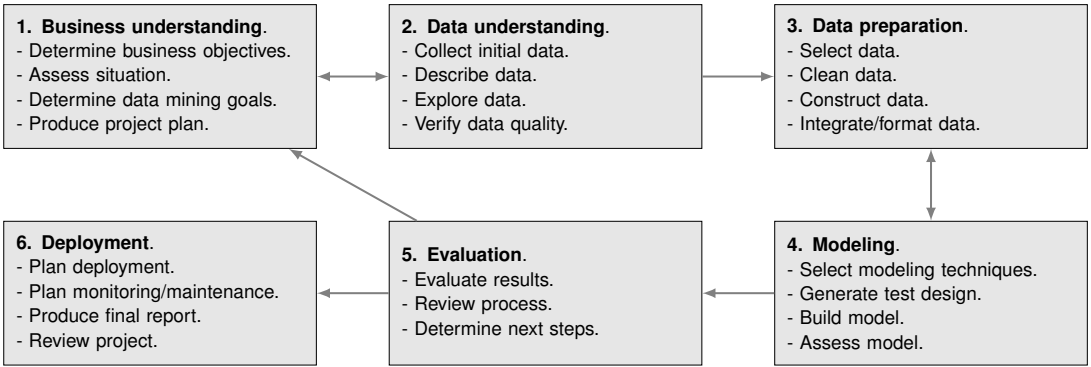
- Data integration.
- Normalization.
- Feature selection.
- Dimension reduction.

- 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**:



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.

A Multidimensional View of Data-Mining

What Data Is Available?

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.

A Multidimensional View of Data-Mining

What Patterns Are Searched For?

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.
  $\rightarrow$ Always validate whether the gained knowledge is interesting.

Data Mining Functions: Examples

Some will be covered in this lecture:

- Data Transformations $\rightarrow$ *Lectures 3, 4, and 5*
- Frequent Pattern Analysis $\rightarrow$ *Lecture 6*
- Classification $\rightarrow$ *Lecture 7*
- Clustering $\rightarrow$ *Lecture 8*
- Outlier Analysis $\rightarrow$ *Lecture 9*

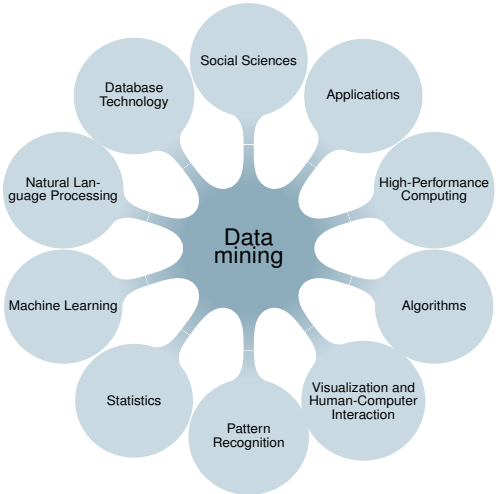
But where are many more. E.g.:

- Network Analysis
- Web Mining
- Time Series Analysis
- Neural Networks
- . . .

A Multidimensional View of Data-Mining

What Technologies Are Used?

Confluence of Multiple Disciplines



Why Confluence of Multiple Disciplines?

Each discipline contributes something different. E.g.:

- **Algorithms:**
  Basic algorithms to get started.
- **Machine Learning:**
  Specialized algorithms for learning from data.
- **High-Performance computing:**
  Parallel and distributed computing to handle large datasets.
- **Database Technologies:**
  Efficent storage and retrieval of data.
- **Etc.:**
  . . .

A Multidimensional View of Data-Mining

What Is the Actual Target Application?

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[3]:
  - **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.

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

Major Challenges in Data Mining

Major Challenges 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 Challenges 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

In this lecture:

- **Data Mining:**
  Discovering interesting patterns and knowledge from massive amounts of data.
- **Data Mining Process:**
  Every community focusses on different aspects.
- **Important Aspects:**
  What Data? What Patterns? What Technologies? What Applications?
- **Challenges:**
  A lot of things to consider.

Any questions about this chapter?

Ask them now or ask them later in our forum:

https://www.studon.fau.de/studon/goto.php?target=lcode_OLYeD79h

Appendix

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.

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.