Chapter I: Introduction

Knowledge Discovery in Databases

Luciano Melodia M.A.
Evolutionary Data Management, Friedrich-Alexander University Erlangen-Nürnberg
Summer semester 2021

Chapter I: Introduction

This is our agenda for this lecture:

- **Why data mining?**
- What is data mining?
- A multi-dimensional view of data mining.
- What kind of data can be mined?
- What kinds of patterns can be mined?
- What technologies are used?
- What kinds of applications are targeted?
- Major issues in data mining.
- A brief history of data mining.
- Summary.

Why data mining?

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).

We are drowning in data, but starving for knowledge. Necessity is the mother of invention.
For data mining it is the automated analysis of massive data sets.

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.

Chapter I: What is data mining?

- Why data mining?
- **What is data mining?**
- A multi-dimensional view of data mining.
- What kind of data can be mined?
- What kinds of patterns can be mined?
- What technologies are used?
- What kinds of applications are targeted?
- Major issues in data mining.
- A brief history of data mining.
- Summary.

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.
- Data dredging.
- Information harvesting.
- Business intelligence.

Watch out: Is everything **data mining**?

- Simple search and query processing is considered not to be.
- Neither are deductive expert systems.

Knowledge discovery pipeline

- This is a typical view from a typical database-systems and data-warehousing community.
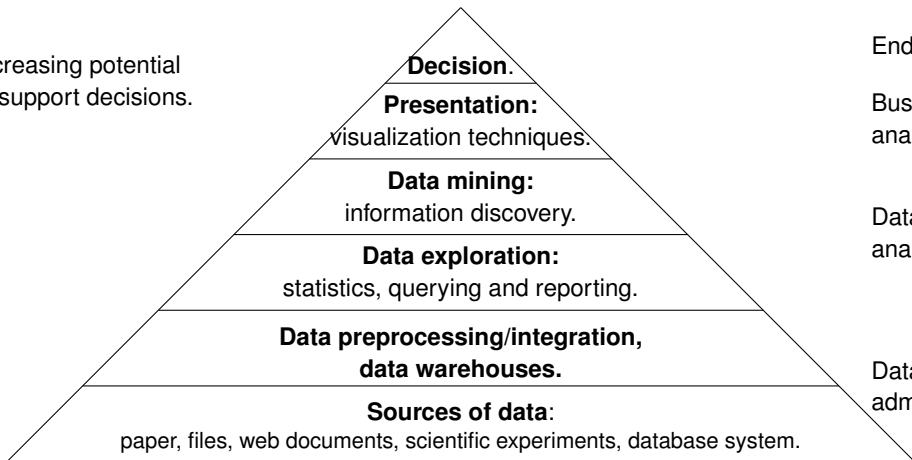- Data mining plays an essential role in the knowledge-discovery process.
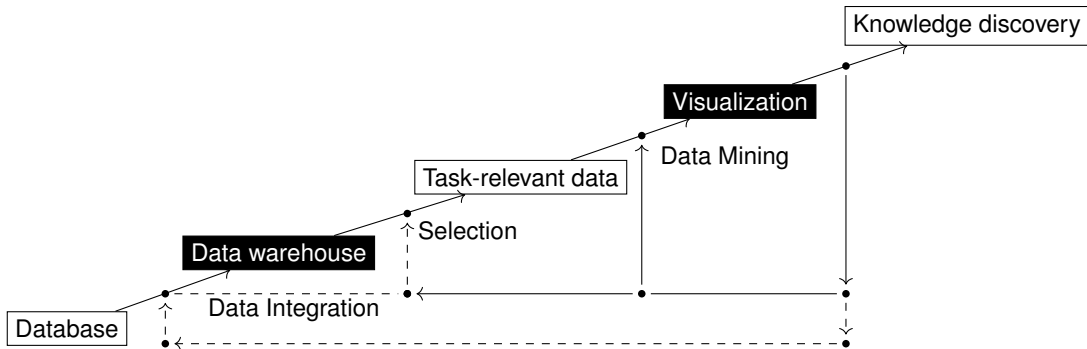
Example: a web-mining framework

Web mining usually involves:

- Data cleaning.
- Data integration from multiple sources.
- Warehousing the data.
- Data-cube construction.
- Data selection for data mining.
- Data mining.
- Presentation of the mining results.
- Patterns and knowledge to be used or stored in a knowledge base.

Data mining in business

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.

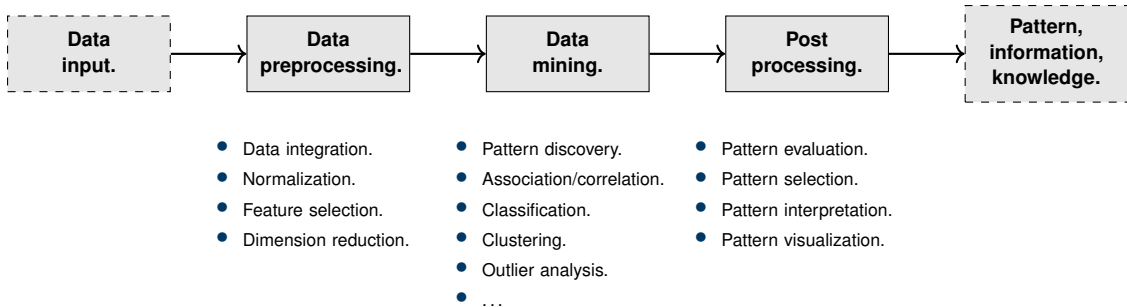End user.

Business analyst.

Data analyst.

Database administration.

Example: mining vs. data exploration

- Business intelligence view:
  - Warehouse, data cube or reporting.
  - But not much mining.
- Business objects vs. data mining tools.
- Supply chain example: tools.
- Data presentation.
- Exploration.

KDD pipeline: a typical view from machine learning and statistics

- This is a view from typical machine-learning and statistics communities.

```
┌ ─ ─ ─ ─ ─ ┐      ┌───────────┐      ┌───────────┐      ┌───────────┐      ┌ ─ ─ ─ ─ ─ ─ ┐
   Data           │   Data    │      │   Data    │      │   Post    │        Pattern,
   input.    ──→  │preprocessing.│ ─→ │  mining.  │ ─→   │processing.│ ─→   information,
                  │           │      │           │      │           │        knowledge.
└ ─ ─ ─ ─ ─ ┘      └───────────┘      └───────────┘      └───────────┘      └ ─ ─ ─ ─ ─ ─ ┘
```

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

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

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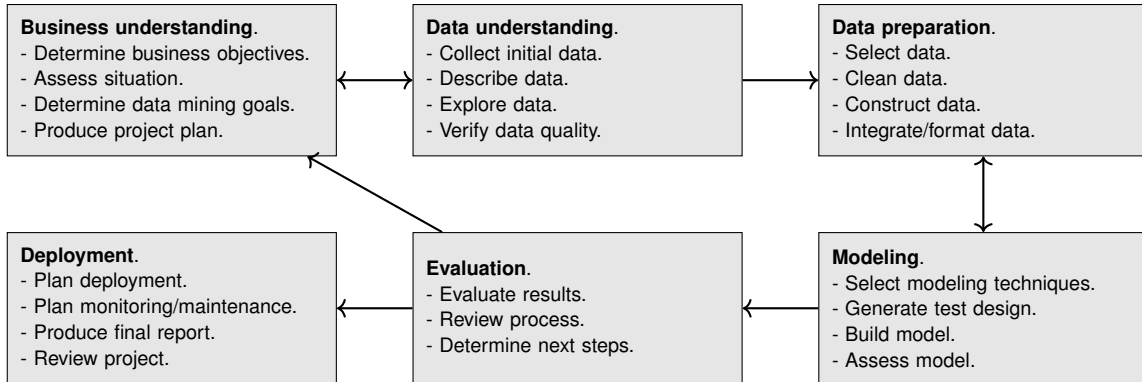- Pattern interpretation.
- Pattern visualization.

Example: medical data mining

- **Health care and medical data mining**:
  - Often adopted such a view in statistics and machine learning.
- **Preprocessing of data**:
  - Includes feature extraction and dimension reduction.
- **Classification and/or clustering processes.**
- **Post processing for presentation**.

CRISP-DM

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

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

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

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

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

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

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

Chapter I: A multi-dimensional view of data mining.

- Why data mining?
- What is data mining?
- **A multi-dimensional view of data mining.**
- What kind of data can be mined?
- What kinds of patterns can be mined?
- What technologies are used?
- What kinds of applications are targeted?
- Major issues in data mining.
- A brief history of data mining.
- Summary.

A multidimensional view of data mining

- **Data to be mined**:
  Database data (extended relational, object oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and Web, multi-media, graphs.

- **Knowledge to be mined (or data mining functions)**:
  - Characterization, discrimination, association, classification, clustering, outlier analysis, etc.
  - Descriptive vs. predictive data mining.
  - Multiple/integrated functions and mining at multiple levels.

- **Techniques utilized**:
  Database, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high performance computing, etc.

- **Applications adapted**:
  Retail, telecommunication, banking, fraud analysis, bio data mining, stock market analysis, text mining, web mining, etc.

Chapter I: What kind of data can be mined?

- Why data mining?
- What is data mining?
- A multi-dimensional view of data mining.
- **What kind of data can be mined?**
- What kinds of patterns can be mined?
- What technologies are used?
- What kinds of applications are targeted?
- Major issues in data mining.
- A brief history of data mining.
- Summary.

Data mining: on what kinds of data?

- Database oriented data sets and applications:
  - Relational database.
  - Data warehouse.
  - Transactional database.
- 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.

Chapter I: What kinds of patterns can be mined?

- Why data mining?
- What is data mining?
- A multi-dimensional view of data mining.
- What kind of data can be mined?
- **What kinds of patterns can be mined?**
- What technologies are used?
- What kinds of applications are targeted?
- Major issues in data mining.
- A brief history of data mining.
- Summary.

Data mining function: I. Generalization

Information integration and data warehous 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.
- Summary.

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 menas 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 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: friends, family, classmates …
- Links carry a lot of semantic 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 (time, location . . .).
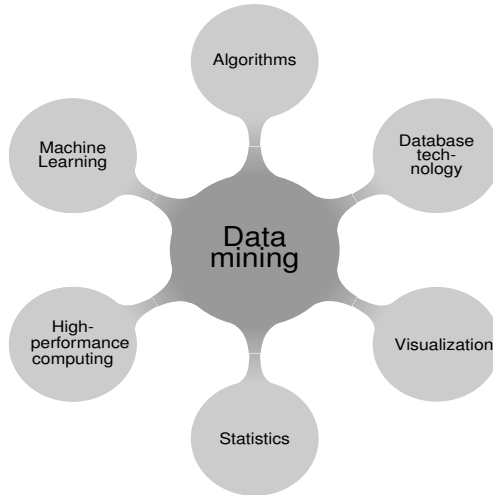- Some may not be representative, may be transient . . .

Evaluation of mined knowledge → directly mine only interesting knowledge?

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

Chapter I: What technologies are used?

Data mining: confluence of multiple disciplines

Thank you for your attention.
Any questions about the introduction?

Ask them now, or again, drop me a line:
✈ luciano.melodia@fau.de.