

# 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 kinds 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 kinds 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.

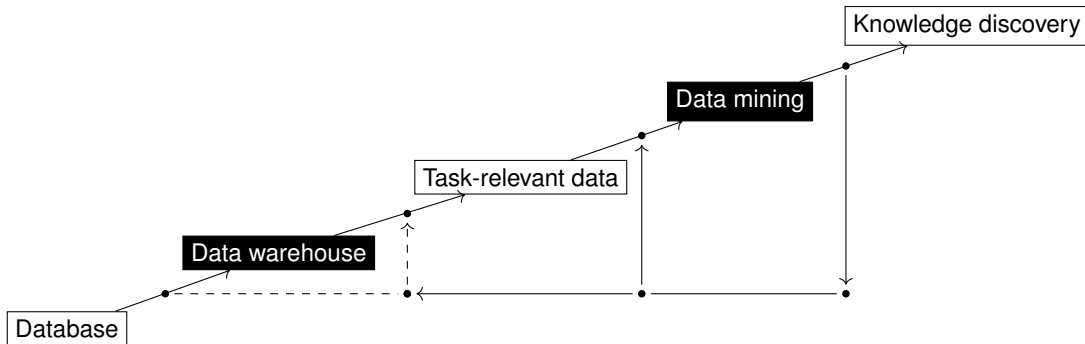
### Watch out: Is everything **data mining**?

- 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.



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

Ask them now, or again, drop me a line:  
✉ [luciano.melodia@fau.de](mailto:luciano.melodia@fau.de).