

# 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.

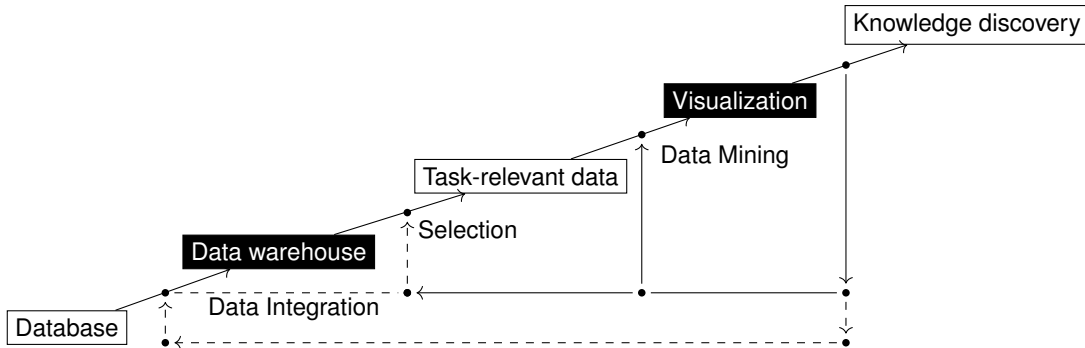
### Watch out: Is everything **data mining**?

- 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.



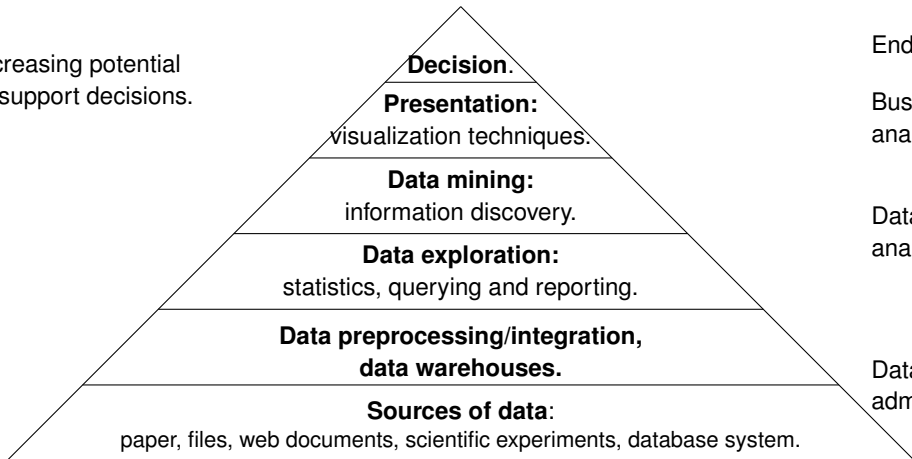
## Example: A Web-mining Framework

### 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.



End user.

Business  
analyst.

Data  
analyst.

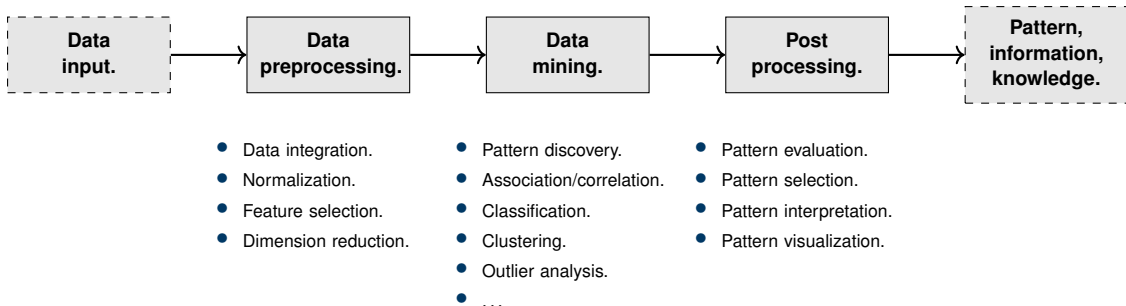
Database  
administration.

## 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.

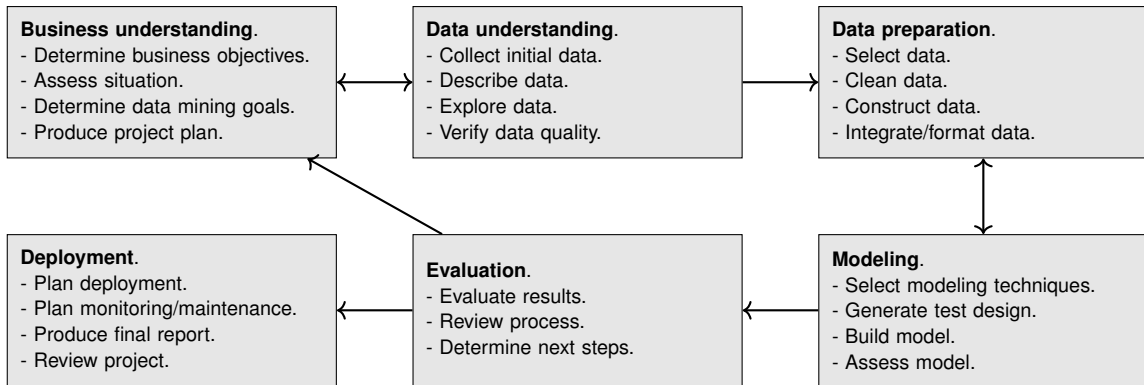


## Example: Medical Data Mining

- **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:**



## 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 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 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 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 (time, location ...).
- Some may not be representative, may be transient ...

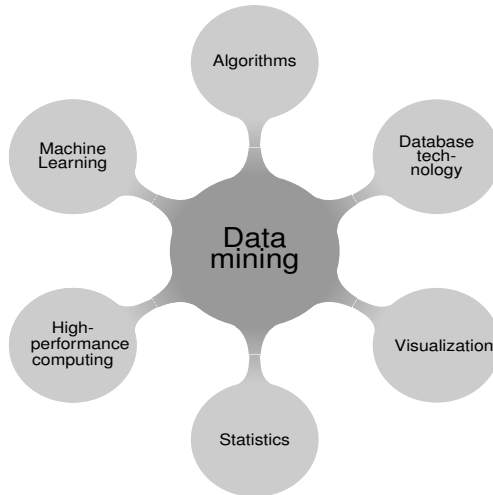
### Evaluation of mined knowledge → directly mine only interesting knowledge?

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

## Chapter I: What Technologies are Used?

- 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: 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.**

## Chapter I: What Kinds of Applications are Targeted?

- 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.



## Applications of Data Mining

### **Web-page analysis:**

From web-page classification, clustering to PageRank and HITS algorithms.  
HITS stands for Hyperlink-Induced Topic Search.

### **Collaborative analysis and recommender systems.**

### **Basket-data analysis for targeted marketing.**

### **Biological and medical data analysis:**

Classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis.

### **Data mining and software engineering:**

E.g. IEEE Computer, Aug. 2009 issue.

### **From major dedicated data mining systems/tools:**

E.g. SAS, MS SQL-Server Analysis Manager, Oracle Data-Mining Tools.

## Chapter I: Major Issues in 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.

## 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.

## Chapter I: A Brief History 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 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.



## Where to Find References? DBLP, CiteSeer, Google (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](http://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), AAI, IJCAI, COLT (Learning Theory), CVPR, NIPS, etc.
- Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.

## Where to Find References? DBLP, CiteSeer, Google (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.

## Chapter I: Summary.

- 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.**

## 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 and trend analysis, etc.

### **Data-mining technologies and applications.**

### **Major issues in data mining.**

## References (I)

- S. Chakrabarti: *Mining the Web: Statistical Analysis of Hypertext and Semi-Structured Data*. Morgan Kaufmann, 2002.
- T. Dasu and T. Johnson: *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, 2003.
- R. Duda, P. E. Hart and D. Stork: *Pattern Classification*. 2ed., Wiley-Interscience, 2000.
- U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy: *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press, 1996.
- U. Fayyad, G. Grinstein and A. Wierse: *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, 2001.
- J. Han, M. Kamber and J. Pei: *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3rd ed., 2012.
- D. Hand, H. Mannila and P. Smyth: *Principles of Data Mining*. MIT Press, 2001.
- T. Hastie, R. Tibshirani and J. Friedman: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer-Verlag, 2009.

## References (II)

- B. Liu: *Web Data Mining*. Springer 2006.
- T. M. Mitchell: *Machine Learning*. McGraw Hill, 1997.
- G. Piatetsky-Shapiro and W. Frawley: *Knowledge Discovery in Databases*. AAAI/MIT Press, 1991.
- P.-N. Tan, M. Steinbach and V. Kumar: *Introduction to Data Mining*. Wiley, 2005.
- S. M. Weiss and N. Indurkha: *Predictive Data Mining*. Morgan Kaufmann, 1998.
- I. H. Witten, E. Frank and M. A. Hall: *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, 3rd ed. 2011.
- C. Shearer: *The CRISP-DM Model: The New Blueprint for Data Mining*. Journal of Data Warehousing, vol. 5, no. 4, pp. 13-22.
- T. Xie, S. Thummalapenta, D. Lo and C. Liu: *Data Mining for Software Engineering*. IEEE Computer, August 2009, pp. 55-62.
- R. Hyndman and G. Athanasopoulos: *Forecasting: Principles and Practice*. 2nd ed. Monash University, Australia, April 2018.

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

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