Contents

List of Figures **2**

1 Data mining **3**
1. What is Data Mining? 3
2. Evolution of data mining 3
3. Data mining process 4
4. Data mining techniques 6

Bibliography **7**

List of Figures

1.1 Evolution of data mining 4
1.2 Data mining processes 4
1.3 Different forms of data pre-processing 5

Chapter 1

Data mining

1. What is Data Mining?

In his article *Data mining: past, present and future*, Frans Coenen defines data mining as: "[...] a set of mechanisms and techniques, realized in software, to extract *hidden* information from data."[2]

It is essential to note the importance of the word *hidden* in the last definition. Data mining is not a simple search in a knowledge data base, but rather an ensemble of techniques to gain insightful conclusions and patterns from a given set of data.[2]

Charu C. Aggarwal in his book *Data mining - The textbook* goes further to include other subprocesses than extracting hdden information. He defines data mining as "the study of collecting, cleaning, processing, analyzing, and gaining useful insights from data"[1]. In this regard it is the same as Knowledge Discovery in Data, and many consider use the terms interchangeably [3].

2. Evolution of data mining

The history of data mining is rooted with the advent of data storage and processing by computers in the 1960s [6]. Statisticians and economists already started using terms like data fishing or data dredging, and Lovell pointed that data mining is "a research paradigm that masquerades under a variety of aliases", referring to these terms used by applied econometricians[5].

The introduction of relational databases in the 1980s and the exponential growth in size of data sets, opened the doors for new techniques to gather, store and analyse data.[3][6].

It is in the start of the 1990S that the term **Data Mining** was introduced in the database community and was popular among business and press communities. The rapid growth of data and databases contributed to the growth and evolution of databases. Especially four areas shaped the field to its current state. These areas are artificial intelligence, machine learning, statistics and databases. Statistics provided techniques to study relations existing within data. Artificial intelligence introduced the human thinking processes to statistical models. Machine learning provided computers with the ability to learn without explicit programming and Databases assured the archiving and organization of data [6].

The figure 1.1 shows the evolution of data mining with the technology trends that relied heavily on it [7].

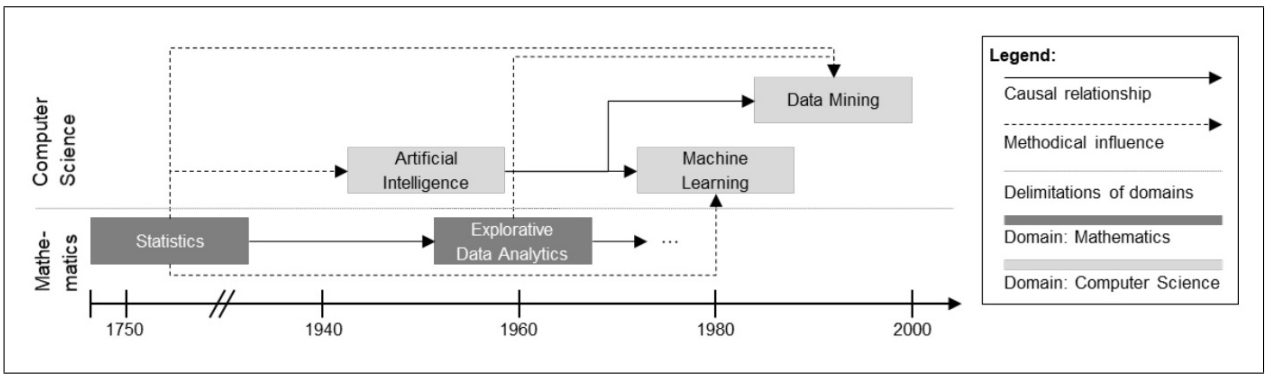

Figure 1.1: Evolution of data mining

3. Data mining process

Data mining is an iterative process with different phases from data collection to analytical processing. The figure 1.2 presents the data mining pipeline [1].

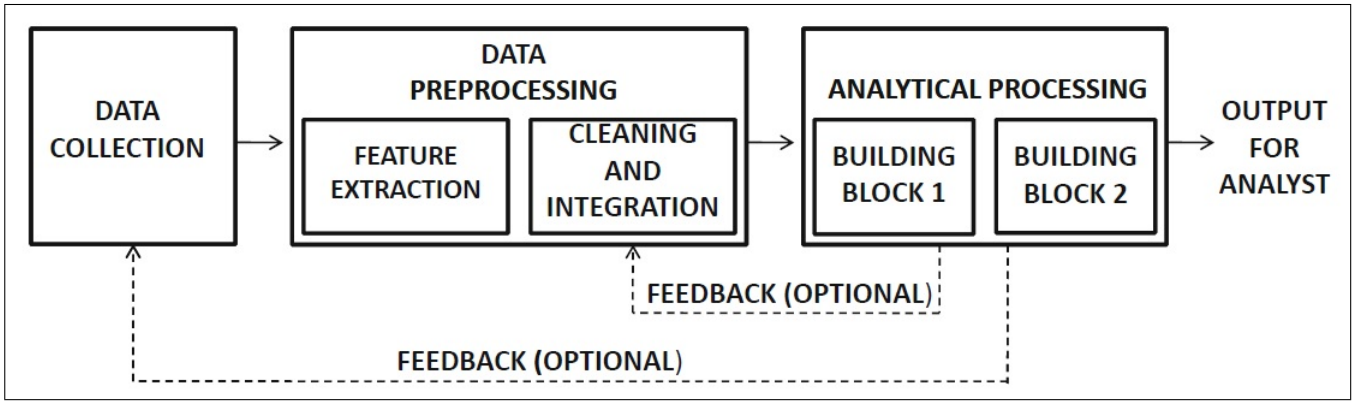

Figure 1.2: Data mining processes

3.1. Data collection

Data collection is an important phase because the choices made in the methods to collect and store data will have an impact on the data pre-processing phase. It is an application-specific phase, with different possible ways to collect data, like surveys, web crawling and network sensors among possible methods [1].

3.2. Data pre-processing

The main objective of this phase is to ensure that the data has the quality needed to satisfy the intended application. Different aspects are to be respected to ensure a good quality of data, such as accuracy, completeness, consistency, timeliness, believability, and interpretability [4].

The main two steps of data pre-processing are feature extraction, data cleaning and transformation.

1. **Feature extraction** The analyst will have to start working with a huge size of raw data. It is essential to define what features are relevant to the application and objective at hand, to extract the pertinent and meaningful data [1].

2. **Data cleaning**

   After defining the right features to work with, the data extracted may have erroneous values, missing entries, outliers, inconsistencies, etc... It is necessary to clean the data, in the sens of dropping wrong entries (data reduction), integrating multiple sources (data integration), identifying outliers, resolving inconsistencies, using statistica methods to estimate missing values, to have a an *accurate*, *complete* and *consistent* data set.

3. **Data transformation**

   In some cases, it is preferable to transform the data set attributes to another type of attributes that is ore amenable to analysis and algorithmic processing, like partitioning the attribute age into ranges [1].

The figure 1.3 summarizes the different forms of data pre-processing [4].

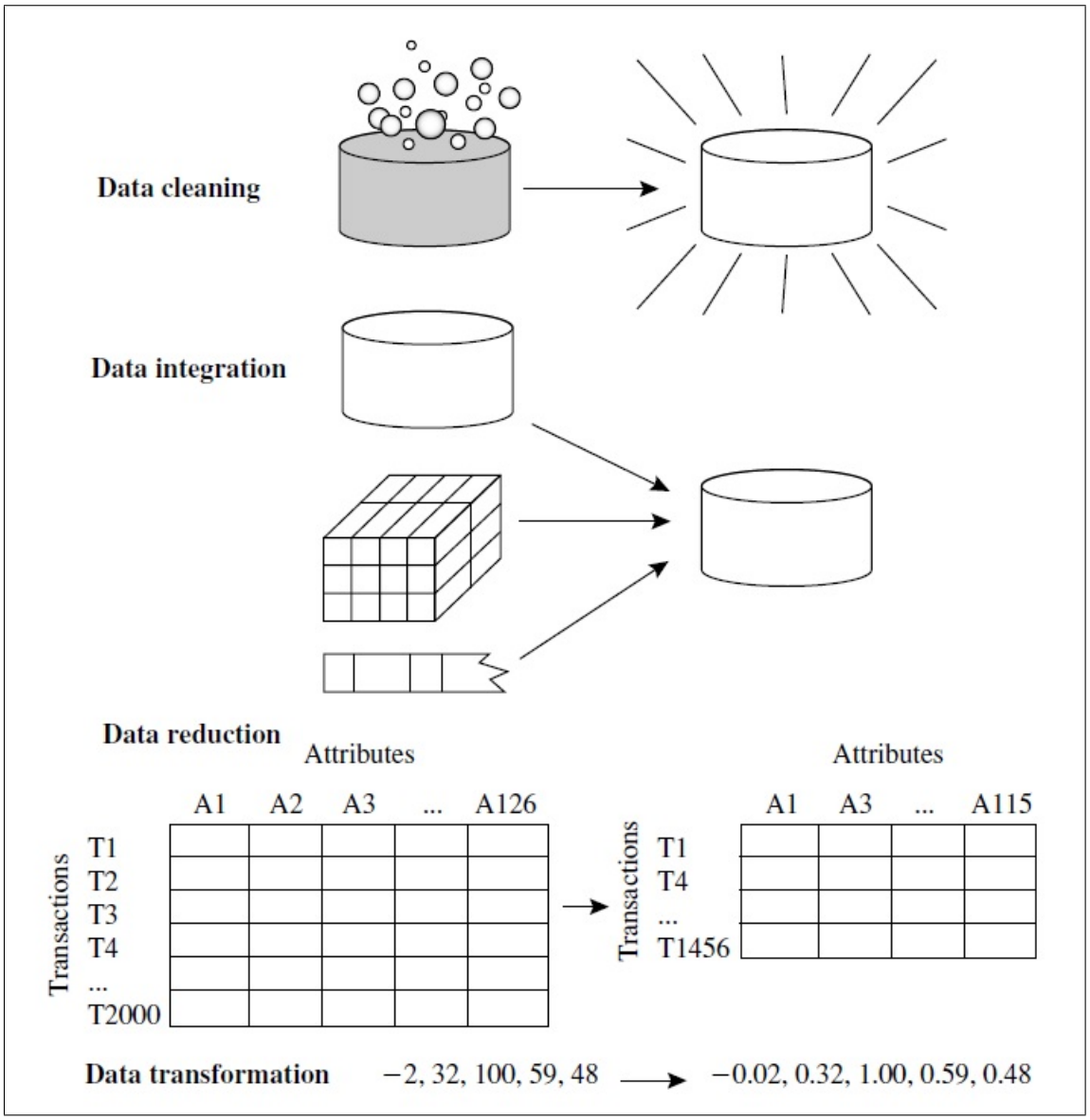

Figure 1.3: Different forms of data pre-processing

3.3. Analytical processing

The most important phase of the data mining process. After using different techniques to gather and prepare the data, different algorithms/techniques are applied construct models specific to

the application and to extract meaningful insights, recurrent relations and patterns [1].

4. Data mining techniques

4.1. Association pattern mining

4.2. Data clustering

4.3. Outlier detection

4.4. Data classification

Bibliography

[1] Charu C. Aggarwal. *Data mining - The textbook.* New York, USA: Springer., 2015.

[2] Frans Coenen. "Data mining: past, present and future". In: *The Knowledge Engineering Review* 26:1 (2011), pp. 25–29.

[3] Jiawei Han, Micheline Kamber, and Jian Pei. "1 - Introduction". In: *Data Mining (Third Edition)*. Ed. by Jiawei Han, Micheline Kamber, and Jian Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 1–38. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00001-0. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000010.

[4] Jiawei Han, Micheline Kamber, and Jian Pei. "3 - Data Preprocessing". In: *Data Mining (Third Edition)*. Ed. by Jiawei Han, Micheline Kamber, and Jian Pei. Third Edition. The Morgan Kaufmann Series in Data Management Systems. Boston: Morgan Kaufmann, 2012, pp. 83–124. ISBN: 978-0-12-381479-1. DOI: https://doi.org/10.1016/B978-0-12-381479-1.00003-4. URL: https://www.sciencedirect.com/science/article/pii/B9780123814791000034.

[5] Michael C. Lovell. "Data Mining". In: *The Review of Economics and Statistics* 65.1 (1983), pp. 1–12. ISSN: 00346535, 15309142. URL: http://www.jstor.org/stable/1924403 (visited on 12/29/2022).

[6] Majid Ramzan and Majid Ahmad. "Evolution of data mining: An overview". In: *2014 Conference on IT in Business, Industry and Government (CSIBIG)*. 2014, pp. 1–4. DOI: 10.1109/CSIBIG.2014.7056947.

[7] Günther Schuh et al. "Data Mining Definitions and Applications for the Management of Production Complexity". In: *Procedia CIRP* 81 (2019). 52nd CIRP Conference on Manufacturing Systems (CMS), Ljubljana, Slovenia, June 12-14, 2019, pp. 874–879. ISSN: 2212-8271. DOI: https://doi.org/10.1016/j.procir.2019.03.217. URL: https://www.sciencedirect.com/science/article/pii/S2212827119305220.