3. DATA MINING AND KNOWLEDGE DISCOVERY

Author: Dejan Mirčetić

As discussed in Chapter 2, the concept of extracting the data and knowledge generation from it is closely related to the **Data Mining techniques**. In contemporary businesses, Data Mining techniques have a high impact on the overall performance of the companies, since operational, tactical and strategic decisions are made based on the input information gained from the Data Mining process. In Chapter 1, it was noted that the world contains over 97 zettabytes of data, with single databases often reaching sizes in the terabytes. In most organisations, data is doubling every two years. This data is stored across various platforms and in different formats, including structured, unustructured and semi-structure data (see Chapter 1). It is estimated that up to 90% of business data exists in an unstructured format. Additionally, a significant portion of this data may containt errors due to inappropriate storage or formating, or manual errors during data collection. As a result, not all data held by organisations is necessarily accurate or reliable. Nevertheless, this vast amount of data holds valuable strategic information for companies. However, when faced with such a large volume of complex data types, how can they be effectively 'mined' to extract the meaningful insights they contain? The answer lies in Data Mining, which serves to inrease revenues and to reduce costs by rapidly and automatically extracting useful knowledge na business insights from massive datasets.

Data Mining emerged from the necessity of efficiently extracting valuable information, requiring techniques that focus on identifying understandable patterns that can be interpreted as useful or interesting knowledge. Thus, **Data Mining is an iterative and interactive process** aimed at discovering valid, novel, useful, and understandabule knowledge (patterns, models, rules, etc.) in massive database (Behera et al., 2019). The main objective of Data Mining is to **reveal critical insights** that **support decision-making** within a business organisation.

The upcoming sub-chapters will address how data is 'mined' for Business Analytics and Knolwedge Discovery.

3.1. What is Data Mining?

Before defining Data Mining, it's important to place this term in context with the terms it's commonly associated with, connected to, and often mistakenly equated with. Non-experts frequently mix up the terms Data Mining and Big Data technology. However, these are two separate concepts. **Big Data describes** extremely large and complex **datasets** that require specialized software applications for processing. On the other hand, **Data Mining goes a step beyond**, it involves analyzing such vast amounts of data to uncover hidden rules and patterns that may not be readily apparent.

Data Mining is a broad term encompassing various analytical techniques, including statistics, artificial intelligence, and machine learning. These methods are used to sift through vast amounts of data stored in an organization's databases or online repositories. The primary goal is to uncover patterns within the dataset. **Business Analytics (BA)** refers to the compehensive process of utilizing skills, technologies, established practices, and algorithms associtaed with Data Mining. Hence, **Data Mining commonly serves as the backend of the BA function**, while the frontend of BA function consists of executive reporting metrics and collated information presented in a format that enables managers to make informed business decisions. When using Data Mining, the BA professionals act like a 'data detective' (Lee, 2013), analyzing data to better describe and understand an organistion's present and past situation (descriptive analytics), predict future outcomes (predictive analytics) and take effective action (prescriptive analytics).

Data Mining is a core component of the **Knowledge Discovery in Databases (KDD)** process, but it is just one step in the overall process. The Data Mining aspect of the KDD process focuses on using algorithms to extract and identify patterns from data. In the broader KDD process, these mined patterns are evaluated and potentially interpreted to determine which patterns may be considered new "knowledge" (Behera et al., 2019). Defined in this manner, with Data Mining as a backend and Knolwedge Discovery as a sort of frontend of BA, they represent, along with business data, its key pillars as outlined in Chapter 2.

Data Mining employes a variety of algorithmes to mine huge datasets, identifying patterns that can yield valuable business insights. Data Mining is a tool, not a magical solution. It doesn't passively observe your database and alert you to interesting patterns. Understanding your business, your data, and analytical methods remains crucial. Data Mining helps business analysts uncover patterns and relationships in data but doesn't determine the

value of these patterns for organizations. Therefore, Data Mining does not replace skilled business analysts, instead, it provides them with a powerful new tool to improve their work.

Data Mining involves the computational process of identifying trends, rules, hidden patterns, and other valuable information by analyzing large datasets. Data Mining adopt its technique from many research areas, including statistics, machine learning, database systems, visualization, neural networks, etc. It is the process of extracting actionable knowledge from diverse data sources sorted in various formats. Data Mining has become increasingly relevant in recent years due to advancements in data storage technologies (Big Data), Artificial Intelligence (AI), and Robotic Process Automation (RPA).

The **Knowledge Discovery in Databases (KDD)** process involves utilizing the database, including necessary selection, pre-processing, sub-sampling, and transformations, to apply Data Mining algorithms to identify patterns (Behera et al., 2019). It also includes evaluating the results of Data Mining. The common standard for describing the steps of Knowledge Discovery process is CRISP-DM (Cross-Industry Standard Process for Data Mining), which is shown in Figure 3.1.

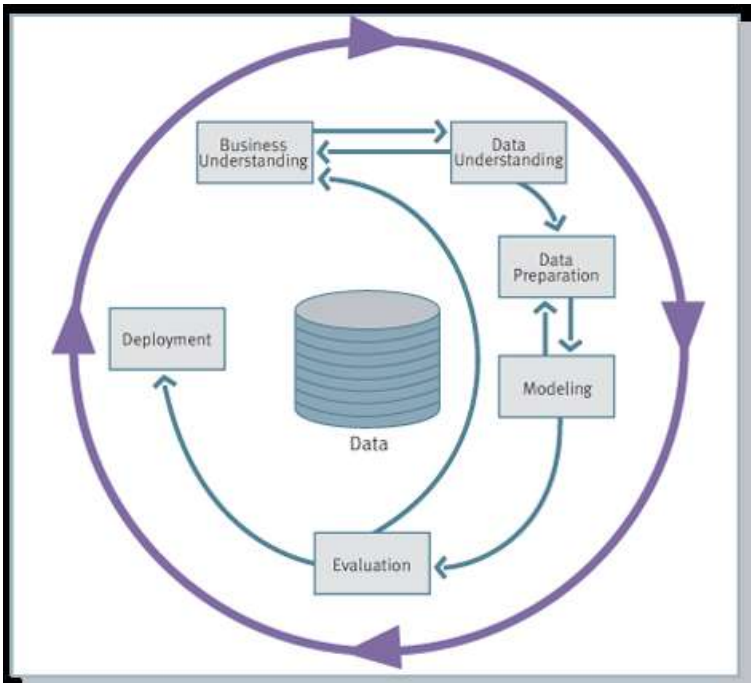

Figure 3.1 Cross-Industry Standard Process for Data Mining (CRISP-DM)

Source: Rahman et al. (2016).

In Figure 3.1, the first phase is business understanding, which involves comprehending the goals to be achieved and conducting detailed fact-finding about resources and assumptions.

The second phase, data understanding, explores various descriptive data characteristics. The third phase, data preparation, is the most challenging and time-consuming part of the KDD process, aiming to select relevant data and format it appropriately for analysis. This phase includes activities like data selection, filtering, transformation, and integration. The fourth phase, modeling, entails applying analytical methods and selecting the most suitable algorithms. This phase also involves verifying the model's quality through testing and cross-validation. The fifth phase, evaluation, involves interpreting and assessing the discovered knowledge (Rahman et al., 2016).

3.2. Knowledge Discovery in Logistics and Supply Chain Management

In the context of SC and logistics, Data Mining is different from other general-purpose applications. The reason is related to the already mentioned and discussed diversity of data sources and structure of business data which pose significant challenges for engineers when designing appropriate modeling solutions. The process of generating the knowledge from the data can be summarized in Figure 3.2.

Based on Figure 3.2, the knowledge discovery and generation from the data is highly dependent on the knowledge source and can be divided into the **judgmental** and **statistical**. Both of these directions have their merits, but the procedure of extracting the knowledge from the source is fundamentally different. To apply a more formal approach for Data Mining, i.e., the statistical approach, the key foundation is the existence of quantitative data. In supply chains, there is a large number of sectors, locations, and transactions which generate data flows which can be used for Data Mining and extracting useful feedback. On the other hand, in SC and logistics, there are also a lot of data sources which are not quantitative, and therefore not subjected to formal quantitative procedures. Rather this data is traditionally subjected to expert judgemental panels (Delphi method) and decisions are made based on experts' experience, knowledge and authority.

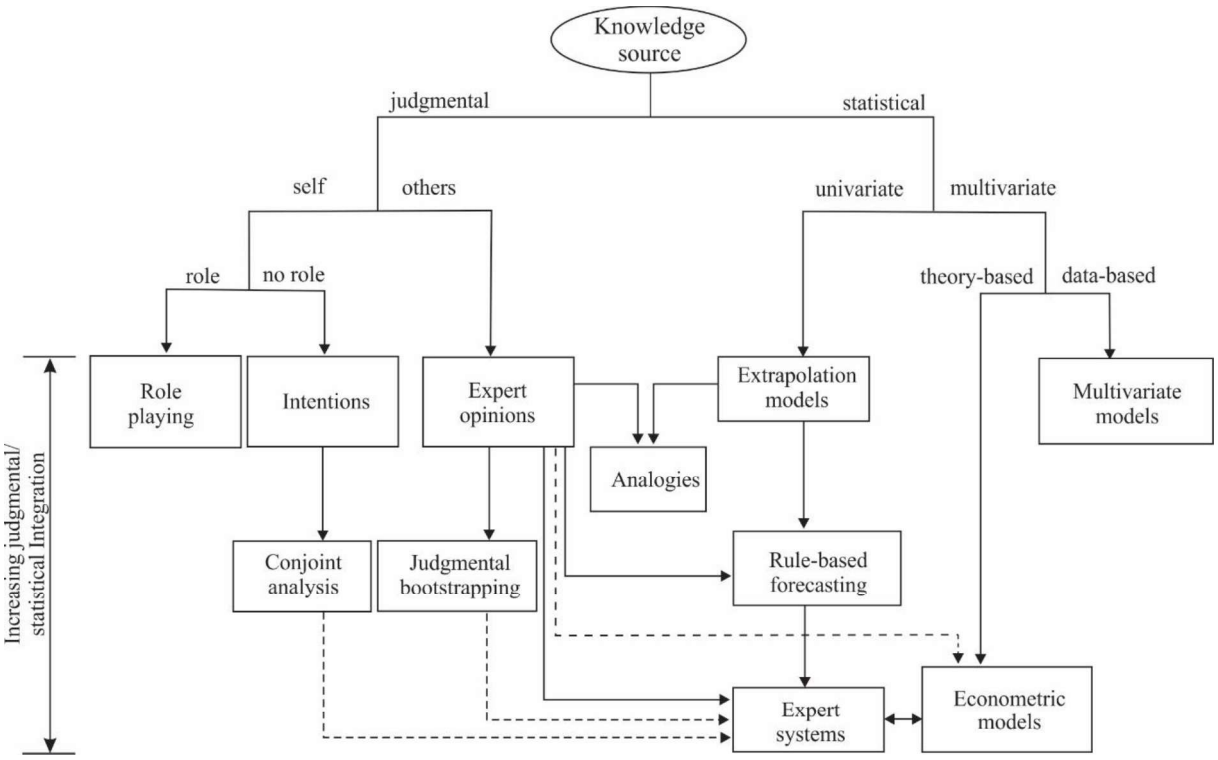

Figure 3.2 Principles of knowledge extraction from the data

Source:

In this book, emphasis will be put on the **quantitative/mathematical techniques**, although we will briefly discuss some of the methodologies for capturing expert knowledge in structured frames, i.e. we will discuss the ES and its applications & examples.

3.3. Delphi's approach to judgmental Knowledge Creation

Valuable insights from seasoned professionals in the supply chain and logistics domain often go unrecognized. These insights should be shared with less experienced professionals in the field. The Delphi method is one approach for capturing and disseminating expert knowledge. According to Steurer (2011), the Delphi method, named after the oracle of Delphi in ancient Greece, which was initially used to consult on various public and personal matters in ancient Greece, in the 1950s, it evolved into a technique where experts replaced the oracle to reach consensus among a group of experts in some field. „Project Delphi", funded by the U.S. Air Force, was the first project using this method to forecast technological developments. Since then, the Delphi method has evolved and improved, finding applications across various scientific disciplines. The Delphi method was developed to achieve a reliable expert concensus, oftne serving as a stand-in for empirical evidence when such evidence is lacking. That is, the

Delphi technique is an iterative process where experts anonymously provide judgments on a particular issue, aiming to gather consensus and dissent along with their justifications. It is a highly structured group communication process where experts assess uncertain and incomplete knowledge (Naisola-Ruiter, 2022). According to Paivarinta et al. (2011), the Delphi method, among others, is extensively utilized in information systems research. It's employed to choose IS projects, prioritize software development project risks, define IS project requirements, pinpoint key issues in IS management, create a framework for knowledge manipulation activities, comprehend the roles and extents of knowledge management systems in organizations, and examine IS research on offshoring.

The Delphi method has become a standard practice for quantifying the outcomes of group elicitation processes. It is utilized across various disciplines to forecast trends, prioritize research areas, assess potential impacts of different policy choices, establish performance indicators, and develop clinical guidelines, among other applications. **The Delphi techniques is also used in SC and logistics area**. For example, it is highly recommended as an instrument for supply risk identification and assessment, for various kind of evaluation in logistics processes, for determined logistics best practices, for strategic decision-making and policy development, for mapping future SCM practices and for logistics foreacasting. The four key characteristics or basic principles of the Delphi method are:

- Iterative and multistage process (and data collection as well);
- Participant feedback (controlled in some level) with the opportunity for participants to revise their answers;
- Statistical determination of group response; and
- Certain degree of anonymity.

A typical Delphi process involves presenting a series of questions over multiple rounds. Panelists, selected for their expertise and knowledge, respond anonymously. Each round is followed by feedback on the aggregated responses, allowing participants to see how their answers compare to those of the entire panel. Panelists can then adjust their answers and provide rationales for any changes in subsequent rounds. This iterative process continues until consensus is reached or a predetermined number of rounds is completed.

3.3.1. Steps to conduct the Delphi method

The Delphi method is a structured approach that entails collecting expert insights and opinions to achieve a consensus on a particular topic. The process typically involves four main steps (Figure 3.3).

Figure 3.3 Steps to carry out the Delphi method

Source:

Step 1 - Defining the objectives: The initial step involves defining the goals and scope of the Delphi study. This includes identifying the specific questions or topics requiring expert input and outlining the key issues to be discussed. This foundational step ensures the study maintains focus and relevance throughout.

Step 2 – Selection of experts: Choosing the correct group of experts is essential for the success and effectiveness of the Delphi technique. These experts should have pertinent knowledge, skills, and background concerning the subject being studied. The group should be varied to offer a broad spectrum of viewpoints. The number of participants can differ based on the study's size and intricacy, but it's generally advised to include at least 10-15 experts.

Step 3 – Elaboration and launching of questionnaires: This step involves developing questionnaires for gathering input from experts. Typically, the initial questionnaire is open-ended to allow experts to freely share their opinions without influence. In Round 1, experts receive the open-ended questionnaire and independently provide insights, predictions, or suggestions related to the study's objectives. In Round 2, the facilitator summarizes and anonymizes the responses from Round 1 to create a more focused questionnaire for the next round. If it is necessary additional rounds can be conducted to refine opinions based on achieved consensus levels, continuing until a predefined consensus is reached or the facilitator decides to end the process.

Step 4 – Use the results: After completing the Delphi process and achieving a consensus, the results are analyzed and utilized for decision-making, forecasting, policy development, or

other purposes outlined in the study's goals. The anonymity of Delphi studies helps ensure that the final results are impartial and reflect the combined expertise of the experts involved.

3.4. Quantitative Data Mining approach for Knowledge Discovery

As demonstrated in Figure 3.2, quantitative **Data Mining is heavily backed on formal mathematical tools, more directly statistical ones**. We could argue that any statistical operation on the data could be regarded as a quantitative analysis. The main purpose of these operations is to extract the real patterns from the data and generate useful insights in the observed process (if the case of analysis in business or supply chains). This is not a straightforward task since there is a significant mismatch between statistical/mathematical theory assumptions and the distributions & patterns which are present in the real business data. The majority of business analyses in practice have a major flow based on that mismatch. It is of vital importance when applying quantitative methods that data fulfils the theoretical mathematic assumptions constrained by the observed model, in order to treat the model results as valid and potentially make decisions based on them.

As discussed in previous section, the Data Mining methods form a core component of the KDD process and are used repatedly whitin it. Data Mining is a multidisciplinary technique, with its final analytical methods grounded in mathematics. Statistics, particularly, plays a vital role in data analysis during the data preparation phase, forming the foundation for several data mining methods. Data Mining involves the use of efficient algorithms to uncover expected or believed patterns. As illustrated in Figure 3.4, **Data Mining tasks** can be classified into five categories: clustering, classification, regression, association rules, and generalization. **Clustering** aims to group database objects so that objects within a cluster are similar, while those in different clusters are dissimilar. **Classification** involves learning a function that assigns attribute values to predefined classes. **Regression**, a statistical method, estimates relationships among variables and is commonly used for prediction and forecasting, overlapping significantly with machine learning. **Association rules** are used to describe strong relationships within transaction processes, such as "when A and B then C". **Generalization** seeks to express a large amount of data as compactly as possible (Su, 2016). The main techniques used in Data Mining are: classification rules or decision trees, regression, clustering, genetic algorithms, agent-based modeling, etc.

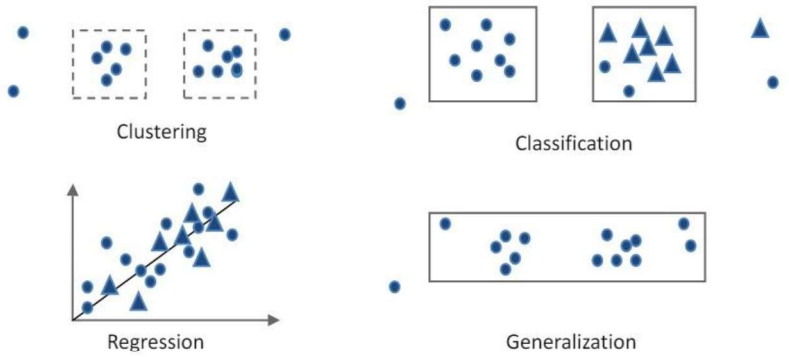

Figure 3.4 Tasks of Data Mining

Source: Su (2016).

The knowledge could be extracted from the processed quantitative data. The **KDD process comprises nine steps** as depicted in Figure 3.5. It is important to note that Data Mining is conducted on transformed data, with non-relevant information already excluded from the original dataset. The patterns discovered through this process are then interpreted and evaluated within a specific context to acquire knowledge that can aid in decision-making.

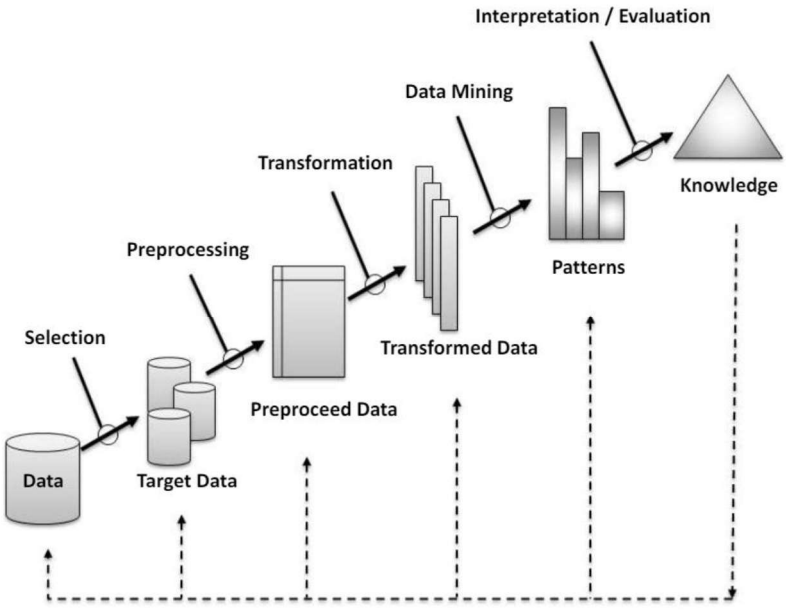

Figure 3.5 Steps that compose the KDD process

Source: Fayyad et al. (1996).

As it previously stated, the common standard for describing the steps of KDD process is CRISP-DM, that represent a leading industrial model. With regard to the current publications and

research reviewed by Su (2016), the typical Data Mining techniques and applications in supply chains include:

- **Decision Trees:** Solving suppliers problems that can be reduced to, e.g. for each decision, a set of possible outcomes, together with an assessment of the likelihood of each outcome occurring
- **Regression:** Forecasting and estimating customer demand for a new product
- **Association Rule:** Identifying the cause roots of product fauiler, optimizing the manufacturing capacity and enabling the condition-based maintenance
- **Genetic Algorithm:** Evaluating the improved hypothese of operating VMI in an uncertain demand environment
- **Clustering Algorithms:** With k-Mean algorithm to categorizing the returned commodities in order to improve the manufacturing processes quality or assigning customers in different segments based on their demographics and purchase behaviours.
- **Multi Agent Data Mining System:** Supporting production planning decisions based on the analysis of historical demand for products

The knowledge extracted by Data Mining is typically stored and presented using **Experts Systems (ES)**. An ES is a sophisticated knowledge system designed to mimic human expertise in various application areas. Olson and Courtney (1992) define ES as „a computer program within a specific domain, involving a certain amount of Artificial Intelligence to emulate human thinking in order to arrive to the same conclusions as a human expert would". An ES component is ideal to assist a decision-maker in an area where expertise is required (Turban, 1995). Essentially, an ES transfers expertise from an expert (or other source) to the computer. It can either support decision-makers or completely replace them, and it is the most widely applied and commercially successful artificial intelligence technology (Turban et al., 2007). One of the justifications for building an ES is to provide expert knowledge to a large number of users (Kock, 2005). According to Turban et al. (2007), ESs are considered to be part of Decision Support System (DSS), that could be characterized as a computer-based information system that combines models and data in an attempt to solve semi-structured and unstructured problems with extensive user involvement (Turban et al., 2007; Mirčetić et al., 2016). The next chapter discusses Machine Learning (ML), which refers to computers' ability to learn and represent knowledge from input data. ML can be seen as a bridge between the

results produced by Data Mining and the Business Intelligence tools used to present executive reporting metrics in a format that enables managers to make informed business decisions.

REFERENCES

1. Behera, P. C., Dash, C. & Mohapatra, S. (2019). Data Mining and Knowledge Discovery (KDD). International Journal of Research and Analytical Reviews, 6(1), pp. 101-106.
2. Fayyad, U. M., Patetsky-Shapiro, G. & Smith, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 17, pp. 37-34.
3. Kock, E. D. (2005) Decentralising the Codification of Rules in a Decision Support Expert Knowledge Base (MSc thesis). University of Pretoria; 2005. Available from: http://repository.up.ac.za/handle/2263/22959
4. Lee, P. M. (2013). Use of Data Mining in Business Analytics to Support Business Competitiveness. Review of Business Information Systems, 17(2), pp. 53-58.
5. Mirčetić, D., Ralević, N., Nikoličić, S., Maslarić, M. & Stojanović, Đ. (2016). Expert system models for forecasting forklifts engagement in a warehouse loading operation: A case study. Promet-Traffic&Transportation, 28(4), pp. 393-401.
6. Naisola-Ruiter, V. (2022). The Delphi technique: a tutorial. Research in Hospitality Management, 12(1), pp. 91-97.
7. Olson, D. L., Courtney, J. F. & Courtney, J. F. (1992). Decision support models and expert systems. New York: Macmillan, USA.
8. Paivarinta, T., Pekkola, S. & Moe, C. E. (2011). Grounding Theory from Delphi Studies. In: Proceedings of the 32nd International Conference on Information Systems (ICIS 2011): Research Methods and Philoshophy, 4-7 December 2011, Shanghai, China.
9. Rahman, F. A., Shamsuddin, S. M., Hassan, S. & Haris, N. A. (2016). A Review of KDD-Data Mining Framework and Its Application in Logistics and Transportation. International Journal of Supply Chain Management, 2(1), pp. 1-9.
10. Steurer, J. (2011). The Delphi method: an efficient procedure to generate knowledge. Skeletal Radiol, 40, pp. 959-961.

11. Su, W. (2016). Knowledge Discovery in Supply Chain Transaction Data by Applying Data Farming (Master thesis). Technical University of Dortmund, Dortmund, DE.
12. Turban, E. (1995). Decision support and expert systems Management support systems. Prentice-Hall, Inc. New York, USA.
13. Turban, E., Rainer, R. K. & Potter, R. E. (2007). Introduction to Information Systems: Supporting and Transforming Business. John Wiley & Sons, Inc. USA.