Data Science Methodology

D

Matthias Pohl
https://orcid.org/0000-0002-6241-7675
Otto von Guericke University, Germany

Christian Haertel
Otto von Guericke University, Germany

Daniel Staegemann
https://orcid.org/0000-0001-9957-1003
Otto von Guericke University, Germany

Klaus Turowski
Otto von Guericke University, Germany

INTRODUCTION

Data Science (DS) represents a relatively new, emerging field of science (Cao, 2017; Pohl et al., 2018). In summary, this discipline aims to extract knowledge and value from data using structured methods and techniques (I. Martinez et al., 2021; L. S. Martinez, 2017). The knowledge can be desirable in companies and other organizations to achieve improvements in performance. In order to profit from the promising advantages of DS as an organization, associated projects require successful completion. Therefore, process models for the project management are needed to conduct goal-oriented DS and provide further support in execution (Saltz & Shamshurin, 2015).

Up to 80 percent of data-intensive projects are not completed or do not meet the set goals, which emphasizes the need of a supporting project structure (Kelly & Kaskade, 2013; VentureBeat, 2019). A wide range of process models were developed from different application directions after the appearance of the first approaches (e.g., KDD, CRISP-DM). Handling data requires a high level of attention in such projects and is always a challenge (Chapman et al., 2000; Fayyad et al., 1996b; Saltz et al., 2017). This chapter is intended to provide an overview of DS process models that have emerged and to highlight project-specific features, such as activities, roles, and documents, without addressing specific elements of project management. The so-called process models are widely considered as methodologies and could be led to a reference model in future. The designation as a process model or framework is used equivalently in the presentations at hand. The common structure of the process models can certainly be termed methodology.

Other concepts have been established in the context of DS. In the last decade, Big Data has become a ubiquitous term. In the meantime, it is no longer exclusively mentioned in information technology, but can also be found in other fields such as sociology, medicine, biology, management and economics (de Mauro et al., 2016). Big Data is described as fast-growing data volumes that are too large for classic data processing systems and therefore require new technologies (Provost & Fawcett, 2013). The National Institute of Standard Technology even relates the term directly to DS (NIST Big Data Public Working Group, 2019).

DOI: 10.4018/978-1-7998-9220-5.ch070

In the course of the terminological development of data science, it became clear that there is a conceptual proximity to data mining (Martinez, 2017). It is therefore imperative that this term is addressed here. Fayyad et al. described data mining as a sub-step of the process, in which algorithms of data analysis and discovery are applied to identify certain patterns in data (Fayyad et al., 1996a, 1996b). Shearer took a different view when he presented his data mining process model, which also assigns the phases of business understanding and data preparation to the data mining process (Shearer, 2000).

In the age of digitalization, artificial intelligence (AI) is a widespread term. However, different interpretations make it difficult to break AI down to a common definition. As descriptions are discussed since the 1960s, the scientific field is concerned with teaching computers to do things that humans are currently better at (Rich, 1983). At this point, a definition of AI should not be discussed, only the connection to DS is stated.

Concrete applications can be found in the field of machine learning (ML), which is characterized as a subfield of AI (Ho et al., 2007). ML is concerned with the development of algorithms and techniques to create computer systems that can improve themselves through experience (Ho et al., 2007). These algorithms use large amounts of data for pattern recognition and effective learning to train the machine to be able to make autonomous decisions (Helm et al., 2020).

The representing terms and related models will be integrated at the presentation in the following sections.

BACKGROUND

For years, the individual experiences of data scientists were held only in inner circles of research institutions or companies. In the 1990s, the first attempts were made to bring together the experiences in DS process frameworks. The Knowledge Discovery in Databases (KDD) and the Cross-Industry Standard Process for Data Mining (CRISP-DM) models have evolved from these aspirations and are already widely used in science and industry. Further developments have been made with new experiences gained from the developed data landscape and led to new approaches. The following overview of data science frameworks (Table 1) is the result of a reproduced literature analysis (Martinez et al., 2021). It is not excluded that further derivatives of these approaches will be developed and applied in companies or research institutions for individual use.

Instead of describing each process model in detail, the similarities and differences should be highlighted. Within projects, the sequential process stages and activities are of primary interest. The corresponding activities require certain skills of the participants, so that a further comparison of the intended roles is made. For the documentation and traceability of the project results, a final view is taken of the artifacts to be produced in the methodological frameworks.

Stages and Activities

First, an overview of the phases or super-activities of all DS process models included is to be given (see Table 2). Accordingly, only the methods are mentioned that contain such a task structure at all. The Agile Data Science Lifecyle and MIDST frameworks do not integrate such phase models and are excluded from the overview. CRISP-DM and RAMSYS are listed here in one column because they contain the same stages. The methodological phases assigned in the further comparison are already shown in the first columns.