2020

Introducing DASC-PM: A Data Science Process Model

Michael Schulz
NORDAKADEMIE University of Applied Sciences, michael.schulz@nordakademie.de

Uwe Neuhaus
Nordakademie University of Applied Sciences, uwe.neuhaus@nordakademie.e

Jens Kaufmann
Hochschule Niederrhein University of Applied Sciences, jens.kaufmann@hs-niederrhein.de

Daniel Badura
Valantic Business Analytics GmbH, daniel.badura@ba.valantic.com

Stephan Kuehnel
Martin Luther University Halle-Wittenberg, stephan.kuehnel@wiwi.uni-halle.de

See next page for additional authors

Introducing DASC-PM: A Data Science Process Model

Completed research paper

Michael Schulz, Uwe Neuhaus
NORDAKADEMIE University of Applied Sciences
Elmshorn, Germany
Email: {michael.schulz, uwe.neuhaus}@nordakademie.de

Jens Kaufmann
Hochschule Niederrhein University of Applied Sciences
Mönchengladbach, Germany
Email: jens.kaufmann@hs-niederrhein.de

Daniel Badura
valantic Business Analytics GmbH
Hamburg, Germany
Email: daniel.badura@ba.valantic.com

Stephan Kuehnel
Martin Luther University Halle-Wittenberg
Halle (Saale), Germany
Email: stephan.kuehnel@wiwi.uni-halle.de

Wolfgang Badewitz, David Dann, Simon Kloker
Karlsruhe Institute of Technology
Karlsruhe, Germany
Email: badewitz@fzi.de, {david.dann, simon.kloker}@kit.edu

Emal M. Alekozai
Robert Bosch GmbH
Stuttgart, Germany
Email: mohammademal.alekozai@de.bosch.com

Carsten Lanquillon
Heilbronn University of Applied Sciences
Heilbronn, Germany
Email: carsten.lanquillon@hs-heilbronn.de

Abstract

Data-driven disciplines like data mining and knowledge management already provide process-based frameworks for data analysis projects, such as the well-known cross-industry standard process for data mining (CRISP-DM) or knowledge discovery in databases (KDD). Although the domain of data science addresses a much broader problem space, i.e., also considers economic, social, and ecological impacts of data-driven projects, a corresponding domain-specific process model is still missing. Consequently, based on a total of four identified meta requirements and 17 corresponding requirements that were collected from experts of theory and practice, this contribution proposes the empirically grounded data science process model (DASC-PM)—a framework that maps a data science project as a four-step process model and contextualizes it among scientific procedures, various areas of application, IT infrastructures, and impacts. To illustrate the phase-oriented specification capabilities of the DASC-PM, we exemplarily present competence and role profiles for the analysis phase of a data science project.

Keywords Data Science, Process Model, Procedure Model, Competencies, Roles

1 Introduction

Traditional methods of data analysis have reached their limits due to several recent developments. On the one hand, available data are increasingly heterogeneous and unstructured (Hendler 2014). Large amounts of data are freely available in different formats, including in text, images, and video, whose analysis requires integration, interpretation, and sense-making (Alharthi et al. 2017; Dhar 2013; Oussous et al. 2017). The integration and analysis of such data rely on the combination of approaches from different disciplines, including social science, cognitive science, information science, and mathematics (Cao 2018). On the other hand, computers are increasingly being used as agents that interpret data automatically (Dhar 2013). When computers act as decision-makers, several new dynamics emerge that warrant consideration, ranging from the costs of wrong decisions to ethical and data protection issues (Cao 2018; Dhar 2013). Because those dynamics cannot be captured with traditional data analysis, their investigation justifies a new, distinct field of research: data science.

To date, data science has generated solutions to overcome some of those emerging limitations, including analysis procedures for high-dimensional data (Gao 2015), multimedia data of various formats (Gandomi and Haider 2015; Pouyanfar et al. 2018), and procedures that combine approaches from different disciplines for analysis purposes (Hu and Zhang 2017). At the same time, isolated solutions are also often developed for specific problems that primarily rely on mathematical, statistical, or technological approaches focused on the data-driven problem space (Cao 2018; Oussous et al. 2017; Pouyanfar et al. 2018). However, data science is more than simply experimenting with data-driven approaches, and a holistic, more project-oriented approach to the field can expand the data-driven perspective to include business and decision-making goals, as well as an understanding of the underlying problem. In turn, the approach can expand the problem space to include goal- and problem-driven aspects (Cao 2018).

To address the problem space holistically, it is essential to provide a framework that embeds the process of data science projects in its environment, comprising, e.g., academia, areas of application, IT infrastructures, associated requirements, and impacts. Although model-based frameworks developed for data analysis projects are already available in related disciplines, including knowledge management (Azevedo and Santos 2008; Piatetsky-Shapiro 1993) and data mining (Azevedo and Santos 2008; Wirth and Hipp 2000), no comprehensive framework specifically for data science projects has been developed, at least to the best of our knowledge and belief. Against that backdrop, the goal of our research is the development of a framework for data science projects by means of a process model. Our basis for doing so is knowledge of essential requirements imposed upon such a process model in both theory and practice. To that purpose, we formulated two research questions (RQ):

RQ1: *Which theoretical and practical requirements are imposed upon data science process models?*

RQ2: *How can a data science process model that is aligned with relevant theoretical and practical requirements be conceptualized?*

To answer our RQs, we conducted multiple structured surveys from April 2019 to February 2020, aimed at identifying theoretical and practical requirements placed upon data science models and, moreover, at developing an empirically grounded process model for data science projects. Data collection was conducted in a working group consisting of 22 experts, including 9 professors as well as 13 practitioners and scientists with relevant theoretical and practical experience in data science.

In the first round of surveys, we addressed the theoretical background of data science. We analyzed definitions of the term formulated thus far and surveyed what practitioners and scientists in the working group understand by data science. Based on the results, we condensed a new working definition in Section 2. We distinguish that definition from related terms, such as data mining and knowledge management. In the second round of surveys, we empirically identified 17 requirements of typical data science process models from theoretical and practical perspectives, which serve to answer RQ1. We present these requirements in Section 3.1 and subsume them under four meta-requirements (MR). In the third round of surveys, we sought well-known models from related disciplines. We ultimately pinpointed models from knowledge management, data mining, and data science, as presented in Section 3.2, all of which are widely implemented in research and practice. Building upon the working group's assessments, in Section 3.3 we evaluate those models with respect to the previously identified requirements. As our analysis in Section 3.3 shows that none of the related process models meets all the requirements, in Section 4, we present the main contribution of the paper—the data science process model (DASC-PM)—a framework that maps a data science project as a four-step process model and contextualizes it among scientific procedures, various areas of application, IT infrastructures, and impacts. The model addresses RQ2, was conceptualized by a core

team in the working group and revised during several rounds of improvement until all 22 participants agreed that the requirements had been sufficiently met. In Section 4.1, we discuss the conceptualization of the DASC-PM and its phases, followed by an elaboration of the model's analysis phase in terms of competence and role profiles in Section 4.2. The paper concludes in Section 5 with a summary of the major results, a discussion of the limitations, and an outlook on future research.

2 Theoretical Background of Data Science

Taken literally, data science is the study of data. Of course, real-world practice bears out the fact that science always involves data in one way or another. In the following, we shed some light on what data science truly entails and why it has emerged as a unifying scientific discipline specializing in the study of data.

The term *data science* was originally discussed in mathematical and statistical communities in the context of modern methods of data analysis (Cao 2017b; Donoho 2017; Naur 1974; Weihs and Ickstadt 2018). For example, Cao (2017) has described how Peter Naur (1968) argued that *data science*, or *datalogy*, was an appropriate term for the field that instead became better known as *computer science*. By contrast, Jeff Wu (2020) suggested in a public lecture that *data science* would be an adequate term for modern statistics. Today, the term has become increasingly common, extends far beyond the original focus of data-driven research in data mining and machine learning, and even refers to "the next generation of statistics" (Cao 2017a). With the advent of the big data era, data science has emerged as an important topic and gained considerable momentum, even ubiquity, in both business and academia. The trend can be observed, for example, not only in the increased number of courses in data science degree programs but also in the growing demand for experts to conduct data science projects in an array of businesses and areas of application. Even so, in scientific discourse, two important questions linger (Cao 2017b; Carmichael and Marron 2018; Donoho 2017; Patil 2011; van der Aalst and Damiani 2015; Weihs and Ickstadt 2018): 1) How does data science differ from disciplines of traditional data analysis?, and 2) Is it justifiable to delimit data science as an independent field of research?

Most authors agree that data science somehow involves converting data into insights by way of data analysis or analytics (Cao 2017b). However, aside from data science, there is a plethora of other names for the field and related ones that focus on analyzing data, including machine learning and data mining. Most people would probably regard the terms *data mining* and *data science* as synonyms. As an unsurprising consequence, the cross-industry standard process for data mining (CRISP-DM), the most used process model for data mining, was silently adopted as a data science process model. Still others regard data science as a superset comprising all of the various disciplines dealing with data analysis as well as any discipline focusing on collecting, accessing, storing, and processing data in general. Therefore, data science should be understood more broadly as a comprehensive discipline concerned with the study of data and converting data into insights. To provide a foundation for our process model for conducting data science projects, here we first provide our new working definition of *data science*. To that end, our working group conducted a rigid collection and synthesis of the key properties of data science projects and initiatives, which ultimately yielded eight major aspects that form the cornerstones of a concise yet comprehensive definition of *data science*.

To begin, the chief purpose of data science projects is *to gain knowledge about the data (1)* that have been selected as the basis of the analysis. Although our understanding of data science focuses on using *analysis methods (2)*, we do not restrict its meaning to the use of any specific category of methods or algorithms. Because it is important only that each result created has been systematically generated, we emphasize that data science initiatives need to follow a *scientific approach (3)*, which in data science is inevitably *interdisciplinary (4)* (van der Aalst 2016). Where the explanatory power and technical capabilities of traditional fields end, they need to branch out to other professions. That necessity is especially the case in data science, which requires, among other fundamentals, a thorough understanding of a certain application-specific domain, mathematical knowledge, and a solid technological background. In recent years, *big data* has operated as a widely accepted term for unstructured data in high volumes (Chen 2014). Although we agree that the trend toward big data has strengthened research in data science (Dhar 2013), we acknowledge that data science is not limited to that field and consider that all forms of *complex data (5)* to be valid (and necessary) sources for analysis. Because using technology is vital in processing complex data and essential to guaranteeing reproducibility, the procedures need to be at least *semiautomatic (6)* before they can be considered to represent data science. Because every aspect of the world has or could become datafied, data science can have massive influence on society, as illustrated by the Cambridge Analytica scandal. Therefore, the potential use and misuse of data and its *effects on society (7)* need to be considered.

Nonetheless, *utilizing data-borne insights (8)* in market-oriented forms is a major step in any data science project, especially if not done for purely academic purposes. The value of any data science project is determined by its results, its processes, and the overarching knowledge gained. Considering all of the above, we define *data science* as follows:

Data science is a field of interdisciplinary expertise in which scientific procedures are used to (semi)automatically generate insights from conceivably complex data leveraging existing or newly developed analysis methods. The knowledge gained is subsequently utilized, taking into account the effects on society.

Although our definition provides a thorough foundation for data science initiatives, it does not focus on the people involved—that is, data scientists. Recent advances in computer technology and easy access to vast amounts of data in the last few decades have given data analysis new momentum, while public awareness of the term *data scientist* rose dramatically when an article in *Harvard Business Review* named the profession of data scientist "the sexiest job of the 21st century" (Davenport and Patil 2012, p. 70). Simply put, data scientists are the people who work on data-driven projects. However, as clear from the comprehensive definition of *data science*, the number of professional competencies required is potentially overwhelming. Building upon the contributions of Conway (2020) and Davenport and Patil (2012), data scientists are characterized by competencies in mathematics, statistics, IT, application domains, strategy, and management. Depending upon the specific application, in-depth competencies in one or a subset of the areas mentioned may be necessary as well. If the individuals lack competencies in one of these key areas, then they cannot be treated as fully trained data scientists. Furthermore, data scientists should also be able to communicate with all stakeholders in a suitable language as a means to oversee the management of data science projects and strategically classify activities (Davenport and Patil 2012).

Generally speaking, however, it is unlikely that a single person can develop profound skills in all of the areas listed (Zschech et al. 2018). For that reason, data scientists need to have general knowledge in all areas but specialize in only one or a few of them. The process model proposed in this paper provides a convenient structured approach to arranging the roles and steps in a data science project.

3 Related Work

The examination of related work involves three steps. First, we discuss the requirements of data science process models that were identified as relevant by our working group. The requirements were collected via a survey, cover both scientific and practical aspects, and, thus, address RQ1. Next, the identified requirements were subsumed under four MR: *horizontal completeness* (**MR1**), *vertical completeness* (**MR2**), *guidance* (**MR3**), and *realities and impact* (**MR4**). *Horizontal completeness* refers to the completeness of the process model at the given level of abstraction and includes aspects such as domains of application, scientificity, elements of the data and their analysis, utility and usability, and required infrastructures. By contrast, *vertical completeness* refers to the model's scalability considering different levels of abstraction, team members and their roles, as well as all necessary information flows and terms. *Guidance* refers to the ability of a process model to provide targeted recommendations to support the process of the data science project and offer guidance in critical decision-making and documentation. Last, *realities and impact* refers to the consideration of realities as well as economic, ecological, and social impacts. Altogether, the four MRs encompass 17 requirements, hereafter labeled with "R" and discussed at length in Section 3.1. Second, we outline related process models that the working group classified as highly related to the topic. Third, we evaluate the related process models in terms of the requirements discussed in the first step.

3.1 Requirements of Data Science Process Models

In general, the quality of data science projects should be improved by using a process model. The completion of all steps, from project conception to the utilization of the knowledge gained, has to be considered and documented. In our case, we regard a process model for data science as being complete when it provides a process for not only analyzing (**R1.1**) but also gathering and handling data (**R1.2**) as well as utilizing (**R1.3**) and using (**R1.4**) results. In particular, the point at which insights are gained by applying analytical methods has to be recognized, and interpretations have to be supplemented by domain-specific knowledge (**R1.5**). That process ensures the reproducibility, reusability, and generalizability of the results, all of which are vital aspects of any scientific method (**R1.6**), albeit depending upon the domain and necessary infrastructure (**R1.7**).

Furthermore, the model has to be sufficiently scalable (**R2.1**) to support projects of different sizes. To that end, when developing a process model, determining the level of the abstraction (**R2.2**) of the

tasks contained is pivotal. If the chosen level of abstraction is too high, then few benefits result, all limited to the conceptual level. If the chosen level of abstraction is too low, however, then the model's generalizability, which is vital due to data science's various fields of application, and comprehensibility become more difficult, which in turn endangers the model's applicability. Beyond that, a high level of complexity can cause the conscious or unconscious omission of tasks and thus calls into question the general use of a standardized approach. By dividing the model into levels of different degrees of abstraction, the model's clarity is maintained, and assistance can be provided in resolving detailed questions. At lower levels of abstraction, modularization can also be useful. In concrete applications, the irrelevant components of the model can be ignored without significantly influencing the project process. However, such omissions also require suitable, documented justification in order to preserve the traceability of the generation of results. As an alternative to modularization, specialized variants of the process model can be created according to the domain under consideration and/or the analytical methods used. Suitable interfaces between the model's individual building blocks have to be defined at every level of abstraction and in every form of instantiation.

Large projects require a team of experts from different areas with complementary expertise whose cooperation can be supported by the process model (**R2.3**). For example, by using the model, the participants should be able to identify their own tasks and understand the tasks of others. Active exchange (**R2.4**) between the team members on a data science project should also be promoted by suitable recommendations for action, which can ensure that analytical procedures are applied correctly and in a targeted manner from the perspective of all groups involved. In that context, the process model should also provide a framework for a uniform understanding of terms (**R2.5**) in order to simplify communication between the different groups of people involved.

In scalable models, it is additionally necessary to distinguish project activities to be performed from qualitative requirements for project coordination and organization. It is also essential to depict special features of each phase and to provide clear recommendations for the further application of the project. To that end, it is necessary to provide clear directions (**R3.1**) for using the model in the context of a specific application and to provide guidance with critical decision-making (**R3.2**) and all documentation required (**R3.3**).

Because numerous analytical procedures can often be used in data science projects, the time needed to become familiar with new topics and to test and reject various analytical procedures also has to be considered. Although those tasks may not contribute directly to the project's success, they are necessary to the project's processes. That requirement is only one example of how a process model needs to account for realities (**R4.1**).

Last, because data science exerts economic, social, and ecological impacts, those dimensions also need to be considered in the process model (**R4.2**). However, that process can occur only in the context of the specific application domain.

3.2 Related Process Models

In the following, we briefly describe the four process models that our working group identified as being the most relevant following our survey.

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

The CRISP-DM provides "a framework for carrying out data mining projects which is independent of both the industry sector and the technology used" (Wirth and Hipp 2000, p. 29f). As its name suggests, the CRISP-DM focuses on data mining but explicitly adds nontechnical steps to that process (e.g., business understanding). In particular, the CRISP-DM's reference model emphasizes the iterative character of the data-mining process, which consists of the phases of business understanding, data understanding, data preparation, modeling, evaluation, and deployment. The first four phases involve bidirectional interactions and can be subdivided at multiple levels into generic tasks, specialized tasks, and process instances. The concretization of generic tasks to specialized tasks, in turn, is achieved by applying the context to the generic task. Other than the reference model, CRISP-DM also provides a user guide containing details about each phase and documents such as checklists, questionnaires, tips regarding tools and techniques, stepwise sequences, points at which decision-making should occur, and common pitfalls. Although an independent, remarkably well documented tool (Azevedo and Santos 2008), the CRISP-DM, particularly its reference model, remains rather compact and neatly conceptualized. In 2015, IBM published the "Analytics Solution Unified Method for Data Mining and Predictive Analytics" (ASUM-DM) (IBM 2016), often regarded as a refined version of CRISP-DM, despite being less specific and less popular.

3.2.2 Team Data Science Process (TDSP)

The team data science process (TDSP) is "an agile, iterative data science methodology" developed by Microsoft (2020). At the TDSP's core is a data science life cycle comprising the five stages of business understanding, data capture and understanding, modeling, deployment, and customer acceptance. Each stage consists of tasks, which in turn consist of steps, and artifacts that serve as deliverables are defined for each stage. The stages are profoundly interlinked except for the final stage of customer acceptance that concludes a project. TDSP also differentiates four roles—group manager, team lead, project lead, and data scientist—and their corresponding tasks, as well as provides a standardized project structure, makes recommendations about infrastructure and resources, and identifies tools and utilities.

3.2.3 Knowledge Discovery in Databases (KDD)

The knowledge discovery in databases (KDD) process model from Fayyad et al. (1996) describes the process of generating knowledge from data. In contrast to the CRISP-DM and TDSP, the KDD model almost exclusively focuses on the required technological and programming steps and pays only slight attention to business logic. It describes the process as involving five steps: selecting a subset of the target dataset, conducting preprocessing, transforming data, applying data-mining methods to identify patterns, and interpreting and evaluating them as a means to answer questions that could not be answered before the process. As a result, knowledge is created. According to Fayyad et al. (1996), any number of iterations of the KDD model can be executed between steps. Although the data-mining step has probably received the most coverage in the literature, the authors have stressed that the other steps are equally necessary in order to ensure the successful application of the KDD process.

3.2.4 Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA)

The sampling, exploring, modifying, modeling, and assessing (SEMMA) process was developed by the SAS Institute and, similar to the KDD model, consists of five stages (SAS 2020). In a sense, the SEMMA process can be regarded as a practical implementation of the KDD process, for it was indeed implemented in the SAS Enterprise Miner software (Azevedo and Santos 2008). In Step 1 (i.e., sampling), input data are identified, samples are taken, and datasets are divided into training, validation, and test datasets (SAS 2003). Step 2 (i.e., exploring) involves graphing data, generating descriptive statistics, identifying key variables, and performing association analyses (SAS 2003). In Step 3 (i.e., modifying), additional variables are considered or transformed, outliers are identified, missing values are replaced, and cluster analyses are performed (SAS 2003). In Stage 4 (i.e., modeling), a predictive model is fitted by using regression models, decision trees, and/or neural networks. In Step 5 (i.e., assessing), competing predictive models are compared (SAS 2003).

3.3 Evaluation of Related Process Models

In this section, we examine the extent to which the related process models fulfill the requirements discussed in Section 3.1. Table 1 offers an overview of the results of our investigation by the specific requirement and process model. Filled Harvey balls indicate that a requirement is addressed by the respective process model, half-filled ones that a requirement is at least mentioned, and empty ones that a requirement is neither mentioned nor addressed. As indicated in Table 1, none of the considered models fulfills all of the requirements of a comprehensive data science process model.

First, the KDD model, due to its distinct technological focus, almost wholly neglects the requirements of vertical completeness. The model provides valuable guidelines for processing data, as Fayyad et al. (1996) have demonstrated for various interdisciplinary areas of application. However, its strong technological focus leaves questions unanswered that inevitably arise from taking a holistic view on a data-mining project. Although some of these questions (e.g., regarding organizational structuring) are addressed in the CRISP-DM, the different roles in a data-mining team and their specific responsibilities, means of project documentation, and the overall economic, ecological, and social impact of the project remain unclear. Moreover, the SEMMA process shares all of those weaknesses and is even less complete and specific.

Even the CRISP-DM, though a firmly established model with strengths in applicability for many data-mining projects and hands-on material for practitioners, has several shortcomings. For one, its less pronounced differentiation of procedural steps requires, on the one hand, the groups involved to cooperate more closely and, on the other hand, does not allow the exact delimitation of tasks as is possible within the KDD model. For another, because the CRISP-DM was developed in the idustry, it is inherently limited to the business context and does not explicitly account for scientific requirements. Furthermore, although the CRISP-DM extends the KDD model with a deployment phase, the phase

remains rather undefined amid rapid rises in data-driven products and services. On top of that, its sections on technological issues are slightly outdated, as demonstrated in multiple papers introducing extensions such as stream analytics, cyber-physical production systems, and Industry 4.0 scenarios (Kalgotra and Sharda 2016; Huber et al. 2019). Last, the CRISP-DM neither provides insights into the roles and profiles of team members, nor does it offer any consideration of the economic, social, and ecological impacts of the project's data usage and insights.

(Meta-)Requirement	**KDD**	**SEMMA**	**CRISP-DM**	**TDSP**
MR1: Horizontal completeness				
• R1.1: Analysis	●	●	●	●
• R1.2: Data	●	●	●	●
• R1.3: Utilization	◑	○	◑	●
• R1.4: Usage	◑	○	○	◑
• R1.5: Domain	●	○	◑	◑
• R1.6: Scientific method	○	○	○	○
• R1.7: Infrastructure	○	○	◑	◑
MR2: Vertical completeness				
• R2.1: Scalability	○	○	◑	●
• R2.2: Different levels of abstraction	○	○	●	◑
• R2.3: Team roles and composition	○	○	○	◑
• R2.4: Defined information exchange	○	○	◑	◑
• R2.5: Defined terminology	○	○	◑	◑
MR3: Guidance				
• R3.1: Directedness	●	●	●	○
• R3.2: Guidance regarding decisions	◑	○	●	○
• R3.3: Guidance regarding documentation	○	○	◑	●
MR4: Reality and impact				
• R4.1: Account for realities	◑	◑	◑	◑
• R4.2: Economic, social, and ecological impact	○	○	○	○

Note. KDD = knowledge discovery in databases process; SEMMA = sampling, exploring, modifying, modeling, and assessing process; CRISP-DM = cross-industry standard process for data mining; TDSP= team data science process; MR_i = meta requirement *i*; $R_{i,j}$ = requirement *j* of meta-requirement *i*; ○ = $R_{i,j}$ is neither mentioned nor addressed; ◑ = $R_{i,j}$ is at least mentioned; ● = $R_{i,j}$ is addressed.

Table 1. (Meta-)Requirements of a data science process model.

The TDSP's most obvious merit is in providing a management structure, with standardized uses and tools, including a file structure for repositories, that support the management of data science projects. Beyond that, the artifacts and single steps in each stage of the life cycle can serve as a checklist to track progress. Indeed, that management-oriented view shapes the TDSP as a whole. However, IT roles deduced from the organizational hierarchy do not account for the variety of skills needed and persons engaged in data science projects. Also, it is strictly focused on programming and documentation, meaning that it neglects to incorporate all stakeholders and to address economic, social, and scientific aspects of data science projects. Apart from that, the TDSP is heavily intertwined with the cloud computing platform Microsoft Azure, although it can be deployed independently of any platform.

In all, it can be concluded that the process models discussed principally focus on data mining. As such, they are not at all or only weakly sufficient regarding vertical completeness, referring to the specification of roles and information exchange on teams. Finally, they often end with analysis or utilization but do not account for usage, let alone economic, social, or ecological impacts.

4 The Novel Data Science Process Model (DASC-PM)

Recognizing that none of the related, well-known process models could fulfill the 17 identified requirements placed upon process models for data science projects, our working group developed a novel data science process model called DASC-PM to address RQ2. The model was constructed with reference to March and Smith's (1995) methodology, which prescribes building models based on domain-specific knowledge and empirical findings such that they can represent new theories and/or phenomena in their various elements and connections. According to March and Smith (1995), such models are not primarily about truth but about usefulness. Therefore, we conceptualized the DASC-PM based on insights from the theoretical background in Section 2 and our requirements analysis in Section 3.

4.1 Model Conception and Phase Description

Any data science project is embedded in the *domain of its application*—that is, where the problem being analyzed benefits from the solution developed. In contrast to the specifications of the CRISP-DM and TDSP, the domain of application is not limited to a corporate context but can comprise elements related to medicine, natural sciences, and/or engineering, among others. In the domain, use cases that justify data analysis are identified, and a concrete project order formulated based on one or more use cases is processed as a data science project. If domain experts can formulate explicit requirements at that point of the project, then the requirements often represent domain-specific conditions that influence tasks to be completed in other phases. Therefore, the domain continuously needs to be taken into account.

The defined project order is processed in each project phase following a *scientific procedure* (i.e., compliance with guidelines for research integrity and good scientific practice). Such adherence ensures that the results are not only generated by utilizing up-to-date methods but also both comprehensible and reproducible. It also affords a thorough discussion of assumptions made and the limitations as a result. Although scientific procedures have not been treated as a key area in any existing process model, such procedures, especially compared with processes in engineering science, confer a cachet of thoroughness and logic, thereby increasing the potential to learn from past projects and recycle artifacts. For that reason, key findings should continue to be published or otherwise disseminated within the scientific community. Last, the required level of scientific knowledge has to be determined in consideration of the project's real-world circumstances and domain-related specifics.

The *phase of data provision* involves data preparation (i.e., data acquisition, integration, transformation, and storage), data management, and exploratory analysis to survey the data. Ultimately, the result of undertaking data provision is a data source suitable for analysis from a methodological and technical point of view. The *analysis phase* entails the application of existing methods or the development of novel methodological approaches. The identification of suitable methods can constitute a considerable challenge. The phase's artifact is an analytical result evaluated from both a methodological and a technical perspective.

In the *deployment phase*, the results of analysis have to be prepared such that they are suitable for their intended use, which can vary greatly depending on the specific project. For instance, the artifacts may consist of results made available to the addressees verbally or by means of technical reports. Models and even the analytical procedures may also constitute the results of a data science project. However, the subsequent utilization of such artifacts, which is addressed in the *utilization phase*, is seldom regarded as a principal part of data-driven projects to date and was thus not considered in any of the mentioned process models. Nevertheless, depending on the concrete form of utilization, monitoring usage may be necessary to ensure the model's continued suitability for the application and to gain insights for further developments.

All steps of a data science project depend on the underlying IT infrastructure, which is currently only addressed in the TDSP. The true extent of dependence, however, has to be assessed on a project-specific basis. Even if the ordering organization predetermines the use of specific hardware and software, both the IT infrastructure's limiting and enabling characteristics have to be taken into account in all phases of the project.

Figure 1 depicts the procedural phases in a simplified, concise manner. Solid arrows indicate a primary path in the DASC-PM, whereas dashed arrows indicate the possibility of a return to previous phases, which may become necessary if the intermediate results of the current phase are unsatisfactory. The degree of adaption in such a case can range from the minor adjustment of parameters to the complete revision of the phase. Although not visualized in Figure 1, the termination of a data science project should also be considered in each individual project phase. However, even if the goal defined in the project order is not achieved, that outcome does not necessarily mean that the project has failed completely. After all, the knowledge gained up to the time of termination can be used in the application's domain and in subsequent data science projects.

Each phase of the DASC-PM can be further divided into specific activities. Thus, the abstract phases of the model only appear in the form of white boxes at the highest level of abstraction. Although specific activities in the process phases, along with their interactions, were elaborated in our working group, limitations of space prevent us from further specifying them here.

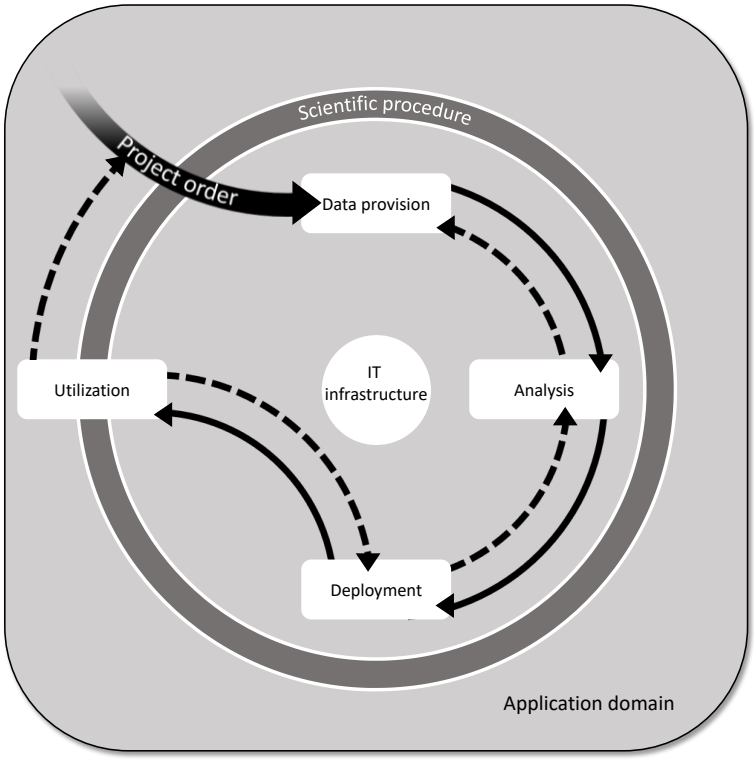

Figure 1: The novel data science process model (DASC-PM).

4.2 Example Specification of the Analysis Phase: Competencies and Roles

In a final in-depth round of interviews, we determined what kinds of competencies are needed in each phase of the model, with a goal to support the planning and staffing of a data science project. As mentioned in Section 2, data scientists are attributed six core competencies: mathematics and/or statistics, IT, the scope of application, communication, strategy, and management. As part of our survey, each member of the working group assessed the relevance of each competency. If their estimates, all recorded on numerical scales, deviated significantly from each other, then the reasons for the deviations were discussed within the working group, and group members were allowed to revise their estimates. After a consensus was reached, all estimates were combined into a single numerical value for each competency. The resulting numerical values, displayed in radar charts, provide a quick overview of the distribution of competencies for each phase. In a similar way, we assessed the degree to which each role in the project is typically involved in particular phases. The results, again illustrated in radar charts, characterize the involvement of the various roles in each phase of the project. As part of our survey, we collected and evaluated the competence and role profiles for all phases of the DASC-PM. Due to limited space, we are unable to present them in full. However, Figure 2 exemplarily illustrates the competence (Figure 2, left) and role profiles (Figure 2, right) of the DASC-PM's analysis phase.

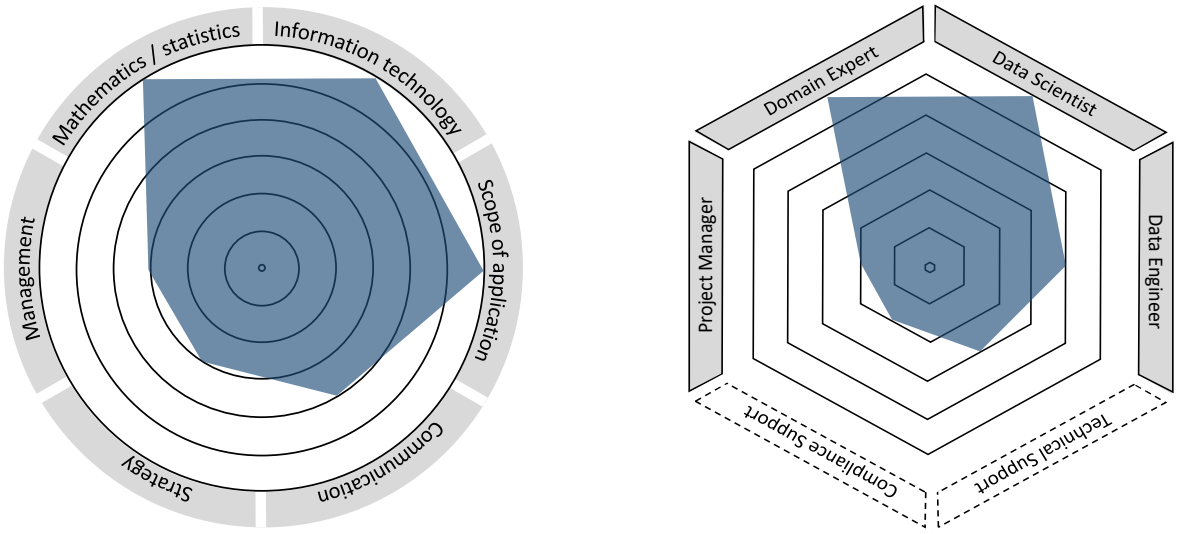

Figure 2. Competence profile (left) and role profile (right) for the analysis phase.

5 Conclusion

Data science is a field of research that involves examining aspects of the data-driven problem space, which can entail large amounts of data, heterogeneous and unstructured data, and data that are available in different formats. To address those challenges, data science builds upon interdisciplinarity—that is, the use of knowledge and approaches from different disciplines of research—including social science, cognitive science, information science, or mathematics. However, data science differs from related fields such as data mining and knowledge discovery by considering challenges that arise over the course of data science projects and that go beyond pure data analysis. Thus, data science involves more than the mere exploration of data-driven approaches. A holistic, more project-oriented approach to data science can extend the data-driven perspective to business and decision-making goals and to the understanding of underlying problems. It can also expand the problem space to include goal- and problem-oriented aspects.

To investigate the procedural character of data science projects in greater detail, we raised two questions in this paper. With RQ1, we sought to pinpoint the requirements placed upon data science process models from theoretical and practical perspectives. Among the results, our analysis has revealed four MRs that experts place on data science process models—horizontal completeness, vertical completeness, guidance, as well as reality and impact—and our survey of experts identified 17 requirements that can be subsumed under those four MRs. Our results show that related process models (i.e., the CRISP-DM, the TDSP, the KDD process, and the SEMMA process) particularly lack scientific methods and do not consider any economical, ecological, or social impacts. The requirements pertaining to vertical completeness, referring to the specification of roles and the exchange of information in teams, are also largely absent. In addition, none of the four models could fulfill all 17 identified requirements of a data science process model.

Next, with RQ2, we sought to determine how to conceptualize a data science process model that is aligned with the 17 requirements. In response, we presented a data science process model called DASC-PM, a novel framework that maps a data science project as a four-step process and situates it in terms of scientific procedures, areas of application, IT infrastructure, and related impacts. From the results of our survey, we also identified specific activities, interactions, and related competence and role profiles for all phases of the DASC-PM. Although limited space prevented us from presenting anything but the competence and role profiles for the analysis phase of a data science project, we plan to publish our survey findings for all phases of the DASC-PM in the future.

Our results stand to make important contributions to research and practice. For one, researchers can use the identified MRs to evaluate the completeness of process models for data science projects and to compare them with the requirements discussed. For another, practitioners as well as researchers can use the DASC-PM to structure data science projects in a phase-oriented way. Beyond that, it is possible to support the phases of the DASC-PM with competence and role profiles, as shown by the example of the analysis phase presented here. In that light, the DASC-PM provides a strong starting point for the standardized execution of data science projects.

To adequately assess the impact of our results, our study's limitations should be taken into account. First, both the collection of requirements placed upon data science process models and the conception of the DASC-PM itself were based on a survey of only 22 experts in data science: 9 professors as well as 13 practitioners and scientists. Although we are confident that our results are sound nevertheless, a larger sample might have facilitated the identification of additional requirements. Second, it is questionable whether a data science process model should be designed exclusively by experts in the first place. After all, it is possible that study participants without expert knowledge of data science can also provide interesting starting points for conceiving or revising such a model. Accordingly, the investigation of the requirements placed on data science process models should be supplemented by a larger sample in future research, one that also includes participants without expert status.

Moreover, models always represent an abstraction of reality. Accordingly, DASC-PM is only a simplified representation of the process of a data science project. Although the model is based both on a well-founded analysis of the theoretical background and the expertise of 22 scientists and practitioners, it does not claim to be comprehensive. The model constitutes a solid foundation for conducting data science projects by considering relevant data-, problem-, and goal-driven challenges. Although the model is empirically grounded, a summative empirical evaluation is still needed. Consequently, empirically evaluating whether the DASC-PM meets practical and scientific requirements constitutes a natural next step of research. The DASC-PM should not be considered as a finished deliverable, but more as a framework that can be continuously improved through scientific and practical discourse.

References

Alharthi, A., Krotov, V., and Bowman, M. 2017. "Addressing Barriers to Big Data," *Business Horizons* (60:3), pp. 285–292.

Azevedo, A., and Santos, M.F. 2008. "KDD SEMMA and CRISP-DM: A Parallel Overview", *Proc. Int'l Assoc. Development of the Information Soc. European Conf. Data Mining*, pp. 182-185.

Cao, L. 2017a. "Data Science: Challenges and Directions", *Communications of the ACM* (60:8), pp. 59-68.

Cao, L. 2017b."Data Science: A Comprehensive Overview", *ACM Computing Surveys*, (50:3), pp. 1-42.

Cao, L. 2018. "Data Science Discipline," in Data science Thinking. The next Scientific, Technological and Economic Revolution, L. Cao, Ed. Cham, Switzerland: Springer, pp. 129–160.

Carmichael, I., and Marron, J. S. 2018. "Data Science vs. Statistics: Two Cultures?," *Japanese Journal of Statistics and Data Science* (1:1)., pp. 117–138.

Chen, M., Mao, S. and Liu, Y. 2014. "Big Data: A Survey," *Mobile Networks and Applications* (19:2), pp. 171-209.

Conway, D., "The Data Science Venn Diagram." A*ccessed April 28 2020* [Online] Available: http://drewconway.com/zia/2013/ 3/26/the-data-science-venn-diagram.

Davenport, T. H., and Patil, D. J. 2012. "Data Scientist: The Sexiest Job of the 21st Century," Harvard Business Review (90:10), pp. 70-76.

Dhar, V. 2013. "Data Science and Prediction," *Communications of the ACM* (56:12), pp. 64–73.

Donoho, D. 2017. "50 Years of Data Science," *Journal of Computational and Graphical Statistics* (26:4), pp. 745–766.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. "From Data Mining to Knowledge Discovery in Databases: An Overview", *AI Magazine* (17:3), pp. 37-54.

Gandomi, A., Haider, M. 2015. "Beyond the Hype: Big data Concepts, Methods and Analytics", *International Iournal of Information Management* (35:2), pp. 137-144.

Gao, L., Song, J., Liu, X., Shao, J., Liu, J., and Shao, J. 2015. "Learning in High-Dimensional Multimedia Data: The State of the Art," *Multimedia Systems*, pp. 1–11.

Hendler, J. 2014. "Data Integration for Heterogenous Datasets", Big Data (2:4), pp. 205–215.

Hu, J., and Zhang, Y. 2017. "Discovering the Interdisciplinary Nature of Big Data Research through Social Network Analysis and Visualization," *Scientometrics* (112:1), pp. 91–109.

Huber, S., Wiemer, H., Schneider, D., and Ihlenfeldt, S. 2019. "DMME: Data Mining Methodology for Engineering Applications", *Procedia Cirp* (79), pp. 403-408.

IBM Corporation, "Analytics Solutions Unified Method," Datasheet. Mar 2016. A*ccessed April 28 2020* [Online]. Available: ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf

Kalgotra, P., and Sharda, R. 2016. "Progression Analysis of Signals: Extending CRISP-DM to Stream Analytics", *2016 IEEE International Conference on Big Data*, pp. 2880-2885.

March, S.T., Smith, G.F.: Design and Natural Science Research on Information Technology. *Decision Support Systems* (15), pp. 251–266 (1995)

Microsoft Corporation, "Team Data Science Process Documentation," Jan 2020. A*ccessed April 28 2020* [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process.

Naur, P. 1968. "'Datalogy '- the Science of Data and Data Processes", *IFIP Congress* (2), pp. 1383-1387.

Naur, P. 1974. Concise Survey of Computer Methods, New York:Petrocelli Books.

Oussous, A., Benjelloun, F.-Z., Ait Lahcen, A., and Belfkih, S. 2017. "Big Data Technologies: A Survey", *Journal of King Saud University-Computer and Information Sciences* (30:4), pp. 431–448.

Patil, D. 2011. Building Data Science Teams, Sebastopol, CA, USA:O'Reilly.

Piatetsky-Shapiro, G. (Ed.). 1993. Proceedings of Knowledge Discovery in Databases 1993. Papers from the AAAI Workshop, Technical Report WS-93-02. Menlo Park, Calif.: American Association for Artificial Intelligence, AAAI Press.

Pouyanfar, S., Yang, Y., Chen, S.-C., Shyu, M.-L., and Iyengar, S. S. 2018. "Multimedia Big Data Analytics: A Survey", *ACM Computing Surveys* (51:1), pp. 1-34.

SAS Institute Inc. "Introduction to SEMMA." A*ccessed April 28 2020* [Online] Available: https://documentation.sas.com/?docsetId=emref&docsetTarget=n061bzurmej4j3n1jnj8bbjjm1a2.htm&docsetVersion=14.3&locale=en.

SAS Institute Inc. 2003. "Data Mining Using SAS® Enterprise MinerTM: A Case Study Approach", Second Edition. A*ccessed November 02, 2020* [Online] Available: https://support.sas.com/documentation/onlinedoc/miner/casestudy_59123.pdf

van der Aalst, W. M. P. and Damiani, E. 2015. "Processes Meet Big Data: Connecting Data Science with Process Science ," IEEE Transactions on Services Computing (8:6), pp. 810–819.

van der Aalst, W. M. P. 2016. "Process Mining - Data Science in Action," Berlin, Germany: Springer.

Weihs, C. and Ickstadt, K. 2018. "Data Science: the Impact of Statistics," *International Journal of Data Science and Analytics* (6:3), pp. 189–194.

Wirth, R., Hipp, J. 2000. "CRISP-DM: Towards a Standard Process Model for Data Mining", *Proc. 4th Int. Conference on Practical Applications of Knowledge Discovery and Data mining*, pp. 29-39.

Wu, J., "Statistics = Data Science?" A*ccessed April 28 2020* [Online] Available: https://www2.isye.gatech.edu/~jeffwu/ presentations/ datascience.pdf.

Zschech, P., Fleißner, V., Baumgärtel, N., and Hilbert, A. 2018. "Data Science Skills and Enabling Enterprise Systems," *HMD Praxis der Wirtschaftsinformatik* (55:1), pp. 163-181.

Acknowledgements

The authors gratefully acknowledge the contributions of Ulrich Kerzel, Felix Welter, Maik Prothmann, Jens Passlick, Raphael Rissler, Alexander Gröschel, Michael Felderer, Dorothee Brauner, Philipp Gölzer, Harald Binder, Heiko Rohde and Nick Gehrke

Copyright